UNIVERSIDADE FEDERAL DE SÃO CARLOS CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

EDUARDO SCHNEIDER BUENO DE OLIVEIRA

CONTRIBUIÇÕES PARA MODELOS DE DIAGNÓSTICO COGNITIVO

Tese apresentada ao Departamento de Estatística – Des/UFSCar e ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre ou Doutor em Estatística - Programa Interinstitucional de Pós-Graduação em Estatística UFSCar-USP.

Orientador: Prof. Dr. Jorge Luis Bazán Guzmán

São Carlos
Abril de 2021

UNIVERSIDADE FEDERAL DE SÃO CARLOS CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

EDUARDO SCHNEIDER BUENO DE OLIVEIRA

CONTRIBUTIONS TO COGNITIVE DIAGNOSIS MODELS

Doctoral dissertation submitted to the Departamento de Estatística – Des/UFSCar and to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Joint Graduate Program in Statistics DEs-UFSCar/ICMC-USP. FINAL VERSION

Advisor: Prof. Dr. Jorge Luis Bazán Guzmán

São Carlos
April 2021

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Eduardo Schneider Bueno de Oliveira, realizada em 23/02/2021.

Comissão Julgadora:

Prof. Dr. Jorge Luis Bazán Guzmán (USP)

Prof. Dr. Caio Lucidius Naberezny Azevedo (UNICAMP)

Prof. Dr. Marcelo Andrade da Silva (ESALQ/USP)

Profa. Dra. Márcia D'Elia Branco (USP)

Prof. Dr. Luis Hilmar Valdivieso Serrano (PUC-Perú)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

Este trabalho é dedicado à todos aqueles
que, com sua presença nesse mundo
e jeito único de ser,
são partes especiais de minha vida.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida.

De maneira especial, após Deus, agradeço àqueles à quem muito amo e que estiveram mais próximos de mim durante esse período de doutoramento, dando aporte pessoal e muito carinho e cuidado: aos meus pais José Roberto e Sirlene, sempre tão presentes a cada dia, seja perto ou a distância, e à minha namorada Natália, sempre me escutando, me aconselhando, compartilhando momentos incríveis e me ajudando a ser uma pessoa melhor a cada dia. Vocês são partes essenciais do caminho trilhado.

À meus avós Milton e Yvone (*in memoriam*), pelo carinho, pelo cuidado e pela rica experiência compartilhada com a sabedoria de seus muitos anos vividos.

Aos meus primos Gilberto Jr, Amanda e Patrícia e à meu tio Gilberto, que foram parte importante de toda essa jornada, prestando apoio mesmo em momentos de maior dificuldade.

Aos amigos que sempre estiveram comigo durante todo o período de doutorado, compartilhando bons e maus momentos, com ajudas e conselhos muito preciosos, em especial a meu amigo de longa data, Murilo.

À família que me acolheu no período passado nos EUA, Juliana, Wally e Edi, os quais foram muito importantes para que esse período fosse o mais confortável possível. Também ao pequeno Nick, que tornou meus dias nos EUA muito mais alegres com sua energia e amabilidade.

A todos os demais amigos dessa jornada no exterior, que são muitos, mas em especial aos que mais tive contato no dia-a-dia, Luiz e Arman e a todos da Hope Lutheran Church, que me acolheram de maneira calorosa.

Ao professor Jorge Bazán, pela ajuda acadêmica e também pelo auxílio, principalmente durante o período passado nos EUA, em diversos momentos e situações. Também à professora Xiaojing Wang, a qual me recebeu na UCONN, permitindo essa experiência tão importante em meu desenvolvimento pessoal e profissional. Também aos professores componentes das bancas de avaliação do trabalho, desde a qualificação, pela disponibilidade e pelas dicas pertinentes para a melhoria constante.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, cabendo também agradecimento à CAPES.

Enfim, à todos os outros amigos, professores, familiares, todos com os quais tive contato

e me ajudaram em meu caminho, minha gratidão.

"Há o homem grande que faz com que todos
se sintam pequenos. Mas o verdadeiro grande homem
é aquele que faz com que todos se sintam grandes."
(G. K. Chesterton)

RESUMO

OLIVEIRA, E. S. B. **Contribuições para Modelos de Diagnóstico Cognitivo**. 2021. 146 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Modelos de Diagnóstico Cognitivo (MDCs) são modelos de variáveis latentes úteis para identificar o perfil de respondentes através de testes ou avaliações. Eles são usados principalmente em avaliações educacionais, mas também podem ser considerados para analisar outros tipos de variáveis latentes, incluindo traços de personalidade e outras áreas na psicometria, bem como qualquer tipo de dados que se enquadre em análises por meio de itens. Diferentemente dos modelos de Teoria de Resposta ao Item (TRI), nos quais a variável latente é contínua, em um MDC a variável latente é discreta, porém, as respostas podem ter os mais variados formatos. A proposta dessa pesquisa é contribuir para o estado da arte dos MDCs, preenchendo lacunas ainda existentes, com especial ênfase nos MDCs sob abordagem Bayesiana. Os capítulos dessa tese seguem uma sequência de construção de MDCs para diferentes tipos de variável resposta. Primeiramente, é mostrado um estudo colaborativo com o modelo DINA dicotômico, já presente na literatura, visando um melhor entendimento dos MDCs e mostrando a comparação de métodos de estimação já explorados com uma nova abordagem MCMC, por meio do algoritmo No-U-Turn Sampler (NUTS). São mostrados estudos de simulação e a metodologia é utilizada para uma aplicação na área de saúde mental. A seguir, considerando respostas contínuas, exploramos, sob abordagem Bayesiana, a extensão do modelo DINA para esse tipo de resposta (C-DINA), realizando um estudo de sensibilidade de prioris e avaliando o desempenho da metodologia por meio de estudos de simulação, bem como trazendo uma explicação mais detalhada da lógica da construção por trás de modelos dessa classe e mostrando uma aplicação relacionada à percepção de risco. Na sequência, propomos um MDC inédito, para respostas limitadas no intervalo unitário (B-DINA), explicitando os detalhes de sua formulação e estimação, sob abordagem Bayesiana, avaliando a recuperação de parâmetros da metodologia de estimação proposta por meio de estudo de simulação e também mostrando o potencial de seu uso em uma aplicação para dados sócio-demográficos. Por fim, propomos novas distribuições de probabilidade para variáveis aleatórias limitadas no intervalo unitário, com desenvolvimento de modelos de regressão quantílica com efeitos mistos, realização de estudos de simulação e uma aplicação para dados de pobreza extrema. Os diversos estudos de simulação e aplicações ao longo do texto mostram que as propostas trazem bons resultados e tem potencial de uso por pesquisadores de diversas áreas, com os códigos utilizados para a estimação dos parâmetros tornados disponíveis.

Palavras-chave: Variáveis Latentes, Modelos de Diagnóstico Cognitivo, Estatística Bayesiana, Respostas Dicotômicas, Respostas Contínuas, Respostas Limitadas.

ABSTRACT

OLIVEIRA, E. S. B. **Contributions to Cognitive Diagnosis Models**. 2021. 146 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Cognitive Diagnostic Models (CDMs) are latent variable models which are useful for identifying the profile of respondents through tests or assessments. They are mainly used in educational assessments, but can also be considered to analyze other types of latent variables, including personality traits and other areas in psychometrics, as well as any type of data that fits in item analysis. Unlike the Item Response Theory (IRT) models, in which the latent variable is continuous, in an CDM the latent variable is discrete, however, the responses can have multiple formats. The purpose of this research is to contribute to the CDMs state of the art, filling gaps that still exist, with special emphasis on the CDMs under a Bayesian approach. The chapters of this thesis follow a sequence of construction of CDMs for different types of response variables. First, a collaborative study with the dichotomous DINA model, already present in the literature, is shown, aiming at a better understanding of the CDMs and showing a comparison of estimation methods already explored with a new MCMC method, through the No-U-Turn Sampler algorithm (NUTS). Simulation studies are shown and the methodology is used for an application in the mental health area. Next, considering continuous responses, we develop the extension of the DINA model for this type of response (C-DINA), under a Bayesian approach, carrying out a priors sensitivity study and evaluating the performance of the methodology through simulation studies, as well as providing a more detailed explanation of the construction logic behind models of this class and showing an application related to risk perception. Then, we propose a CDM for limited responses in the unit interval (B-DINA), which is unprecedented in the literature, explaining the details of its formulation and estimation, under a Bayesian approach, evaluating the recovery of parameters of the proposed estimation methodology through a simulation study and also showing the potential of its use in an application for social-demographic data. Finally, we propose new probability distributions for random variables limited to the unit interval, with the development of quantile regression for mixed-effects models, carrying out simulation studies and an application for extreme poverty data. The different simulation and application studies throughout the text show that the proposals bring good results and have the potential to be used by researchers from different areas, with the codes used to estimate the parameters also made available.

Keywords: Latent Variables, Cognitive Diagnosis Models, Bayesian Statistics, Dichotomous Responses, Continuous Responses, Bounded Responses.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Modelos de Diagnóstico Cognitivo (MDCs) são modelos de variável latente (BARTHOLOMEW; KNOTT; MOUSTAKI, 2011) que são usualmente utilizados para identificar o perfil de respondentes à testes em avaliações por meio de questionários (TORRE, 2009; GEORGE; ROBITZSCH, 2015). Eles são mais utilizados e bastante úteis na área de avaliações educacionais, mas também podem ser considerados para analisar outros tipos de variáveis latentes, incluindo traços psicológicos e de personalidade, bem como serem utilizados para analisar dados de indicadores sociais, não necessariamente sendo limitados ao uso de análise de respostas de questionários por pessoas. Tais aplicações diversificadas serão exploradas ao longo dos capítulos desse trabalho.

Diferentemente dos modelos de Teoria de Resposta ao Item (TRI), nos quais há um traço latente é contínuo, em um MDC temos um atributo latente discreto. Uma análise por meio de um MDC é capaz de trazer um perfil de atributos para cada um dos respondentes e adicionalmente traz a porcentagem de respondentes que possuem os atributos avaliados no teste, permitindo a avaliação tanto individual quanto da população respondente. Por meio das análises feitas por um MDC, é possível providenciar aos respondentes um *feedback* específico, com suas forças e fraquezas, o que faz com que os MDCs possam ir além do simples ranqueamento de indivíduos em relação a um traço latente.

Ao contrário de outros modelos de variável latente já mais presentes e consolidados há tempos na literatura, como os modelos de TRI, os MDC vem recebendo mais atenção nas últimas décadas, havendo ainda grande espaço para desenvolvimento. O uso prático de um MDC requer considerável esforço computacional, sendo que a maior parte dos modelos desenvolvidos nessa classe são do século vinte e um, bem como o modelo DINA (JUNKER; SIJTSMA, 2001) (do inglês *deterministic inputs, noisy "and" gate*) , DINO (TEMPLIN; HENSON, 2006) (do inglês *deterministic inputs, noisy "or" gate*), GDM (DAVIER, 2008) (do inglês *general diagnosis model*), G-DINA (TORRE, 2011a) (do inglês *generalized DINA*), entre outros, com incrementos

ocorrendo após o desenvolvimento do modelo base.

Sob uma perspectiva Bayesiana, o desenvolvimento de MDCs é ainda mais recente, havendo dificuldade de encontrar literatura mais antiga sob uma abordagem Bayesiana, mas com notáveis avanços nos últimos anos, bem como pode ser visto em Culpepper (2015), Zhan (2017), Chen *et al.* (2017) e em avanços trazidos por essa tese, alguns dos quais já estão publicados, como em Silva *et al.* (2018).

Considerando o estado da arte, a principal proposta dessa tese é contribuir na área de Modelos de Diagnóstico Cognitivo, trazendo uma nova abordagem Bayesiana para o modelo dicotômico e estendendo os MDCs Bayesianos para o caso de respostas contínuas na reta e de respostas contínuas no intervalo unitário. Também há alguns estudos adicionais relacionados a outros tópicos em Estatística, com a proposta de uma nova distribuição no intervalo unitário, a qual pode ser utilizada em pesquisas futuras para trazer ainda mais avanço aos MDCs, bem como a outras áreas do saber.

No capítulo 2, mais informações são dadas sobre os Modelos de Diagnóstico Cognitivo, com especial ênfase no modelo DINA dicotômico, o qual é apresentado em detalhes, baseado no trabalho realizado em colaboração, já publicado em Silva *et al.* (2018). Esse modelo é um dos mais presentes na literatura de MDC, principalmente sob abordagem frequentista, mas também com desenvolvimentos iniciais na abordagem Bayesiana em Culpepper (2015). No nosso trabalho, apresentamos a proposta do uso do *No-U-Turn Sampler* (HOFFMAN; GELMAN, 2014) para a estimação dos parâmetros do modelo, realizando comparações com as metodologias de estimação já anteriormente presentes na literatura. Estudos de simulações são feitos, a fim de avaliar a recuperação de parâmetros e a análise de dados reais relacionados à depressão também é apresentada. Além de publicado em Silva *et al.* (2018), esse trabalho também foi apresentado e consta como resumo nos anais dos eventos V CONBRATRI - 5º Congresso Brasileiro de Teoria de Resposta ao Item, Campinas, 2016, com o título *Bayesian approach to the DINA model using No-U-Turn Hamiltonian Monte Carlo* e no *NextGen: Data Science Day, New Haven, 2018*, com o título *An Application of DINA model to Beck Depression Inventory data*.

O capítulo 3 apresenta uma abordagem Bayesiana para um modelo DINA de respostas contínuas, a qual ainda não era encontrada na literatura, havendo até então apenas um modelo frequentista em Minchen, Torre e Liu (2017). O desenvolvimento é apresentado em detalhes sobre a formulação Bayesiana e são realizados estudos de simulação para avaliar a recuperação de parâmetros, bem como uma aplicação relacionada a dados de percepção de risco é apresentada. Esse trabalho será submetido a um periódico especializado, além disso, ele foi apresentado no *3rd International Conference on Econometrics and Statistics, Taichung City, Taiwan, 2019*, com o título *Bayesian analysis of cognitive diagnostic models for continuous response data*, com resumo nos anais do evento, e também fez parte da sessão *Advances in Modeling and Computation for Latent Variable Models* do *33rd New England Statistics Symposium, Hartford, CT, USA, 2019*.

O capítulo 4 traz a proposta de um modelo DINA, sob abordagem Bayesiana, para variáveis contínuas limitadas, com ênfase no intervalo unitário. Esse modelo é inédito na literatura, ainda não sendo encontrado nem sob a abordagem Bayesiana e nem sob abordagem frequentista. São apresentados os detalhes da formulação Bayesiana, bem como um estudo de simulação para recuperação de parâmetros e uma aplicação a dados reais referentes a indicadores sociais dos municípios da região Sudeste do Brasil. Esse trabalho está em fase final de preparação do documento para ser submetido a um periódico especializado.

Com o intuito de estudar novas distribuições para variáveis de resposta limitada no intervalo unitário, no capítulo 5, duas novas distribuições de probabilidade para o intervalo unitário semi-fechado são apresentadas, bem como suas versões quantílicas. Em adição, também é desenvolvida a regressão paramétrica mista, baseada nessas distribuições, sob uma abordagem Bayesiana. O capítulo ainda apresenta aplicações a dados reais, levando em conta dados sobre pobreza no Peru e sobre testes educacionais aplicados a estudantes peruanos. Esse trabalho já está submetido no Journal of Applied Statistics, aguardando revisão e já foi apresentado no *VI Workshop on Probabilistic and Statistical Methods, São Carlos, 2018*, tendo o título *New Gompertz Based distributions to skewed bounded responses*, estando publicado como resumo nos anais do evento.

CAPÍTULO

2

MODELO DINA: ESTIMAÇÃO E APLICAÇÃO

O modelo DINA é um Modelo de Diagnóstico Cognitivo bastante popular em psicologia e psicometria, sendo utilizado para identificar o perfil de respondentes a testes com respeito a um conjunto de atributos ou habilidades latentes. O trabalho apresentado nesse capítulo se baseia em uma colaboração que propõe um método de estimação para o modelo DINA através do algoritmo No-U-Turn Sampler (NUTS), uma extensão do método Monte Carlo Hamiltoniano (HMC). Um estudo de simulação é realizado, com o propósito de avaliar a recuperação dos parâmetros e a eficiência desse novo método MCMC e compará-lo com outros métodos Bayesianos, os algoritmos Metropolis Hastings (MH) e Gibbs Sampling (GS), e um método frequentista, com o algoritmo de Maximização da Expectativa (EM). Os resultados indicam que o algoritmo NUTS quando aplicado ao modelo DINA, recupera devidamente todos os parâmetros, sendo mais eficiente e preciso do que os outros métodos utilizados na comparação, já presentes na literatura. A metodologia é utilizada para uma aplicação na área de saúde mental, ilustrando um novo método de classificação para respondentes do inventário de depressão de Beck. A aplicação traz resultados interessantes e mostra o potencial da utilização desse tipo de modelo para auxiliar no processo de diagnóstico médico.

O conteúdo desse capítulo é parte de uma colaboração com demais pesquisadores, a qual está publicada em Silva *et al.* (2018).

2.1 Introdução

MDCs são ferramentas úteis em psicometria para identificar o perfil de respondentes à questionários, ou o nível no qual eles possuem um determinado conjunto de atributos latentes que estão relacionados a uma variável latente. Essa variável latente pode ser uma habilidade cognitiva (habilidade matemática, por exemplo), um traço psicológico ou uma atitude.

Especificamente, como dito em George e Robitzsch (2015), MDCs são uma classe de modelos de variáveis latentes discretas, as quais se utilizam das respostas a itens para avaliar se o respondente possui características básicas relacionadas aos traços latentes que são cobertos pelos itens. Por exemplo, em uma avaliação educacional de habilidades cognitivas os desenvolvedores do teste podem mapear atributos necessários para responder corretamente a cada uma das questões desse teste; esse mapeamento é chamado de Matriz **Q**.

Embora a maior parte da pesquisa envolvendo MDCs seja conduzida em avaliações educacionais, eles também podem ser utilizados para avaliar outro tipo de variáveis latentes, incluindo traços de personalidade. Templin e Henson (2006) se utilizou de um MDC em um contexto clínico, analisando perfis patológicos relacionados a apostas; seu estudo foi capaz de estimar a porcentagem dos examinados que tinham perfis comportamentais ligados a tal patologia, bem como providenciar padrões comportamentais para cada indivíduo. Esses perfis foram úteis para análises clínicas posteriores, utilizando pontos de corte em escores relacionados à patologias ligadas a apostas.

Os MDCs também já foram utilizados em aplicações ligadas ao diagnóstico de desordens psicológicas (JAEGER *et al.*, 2006; TEMPLIN; HENSON, 2006) e desordens mentais, utilizando escores de itens obtidos através de instrumentos de mensuração de aspectos clínicos (TORRE; ARK; ROSSI, 2015). Nesse contexto, os MDCs auxiliam a avaliar perfis de desordem psicológica para cada um dos indivíduos, bem como identificar o quão prevalentes essas desordens são na população.

Diferentes versões para modelos na classe dos MDC dicotômicos podem ser vistos em George e Robitzsch (2015). Os modelos DINA Junker e Sijtsma (2001), são um dos mais populares e mais amplamente explorados na literatura de MDC, em testes de diagnóstico cognitivo utilizados para variados fins, devido à sua parcimônia e interpretabilidade. Dimitrov (2007) afirma que os modelos DINA atuam como um modelo conjuntivo, no sentido de que os respondentes precisam de todos os atributos que são requeridos para responder corretamente a um item sem chutar.

A estimação dos parâmetros de MDCs em aplicações reais é muitas vezes desafiadora, devido à múltiplos fatores (a quantidade de questões em um teste, o tamanho amostral, o número de atributos relacionados aos traços latentes, os graus de correlação entre os atributos, entre outros). Há na literatura a proposta de uso de métodos frequentistas (DAVIER, 2008; CHEN; XIN, 2012) e Bayesianos para a estimação dos parâmetros do modelo, sendo que, por vezes

podem ocorrer problemas para atingir a convergência quando se diz respeito à MDCs complexos. Nesse capítulo, baseado no artigo supracitado, é introduzida uma alternativa mais acurada para a estimação sob abordagem Bayesiana para os parâmetros do modelo DINA.

Alguns pesquisadores se utilizaram de algoritmos de Monte Carlo via Cadeias de Markov (MCMC) para estimar os parâmetros do modelo DINA e de suas extensões, através dos métodos de Metropolis Hastings (MH) ou do Gibbs Sampling (GS) (TORRE, 2009; JUNKER; SIJTSMA, 2001; TORRE; DOUGLAS, 2004; HUANG; WANG, 2014; CULPEPPER, 2015). Alguns algoritmos em softwares de estimação Bayesianas foram desenvolvidos para estimar os parâmetros do modelo, bem como pacotes no R (TEAM, 2017) como o pacote "dina" (CULPEPPER, 2015), o qual utiliza o GS, e o pacote "CDM" (GEORGE *et al.*, 2016), que utiliza o algoritmo de Maximização da Expectativa (EM) para a estimação sob abordagem frequentista. A estimação utilizando algoritmos de MH podem ser feita utilizando-se de softwares e pacotes como WinBUGS (LUNN *et al.*, 2000), OpenBUGS (THOMAS, 2008), JAGS (PLUMMER, 2003) ou PROC MCMC, do SAS (INC, 2009). Entretanto, a estimação utilizando o *No-U-Turn Sampler* (NUTS) não era encontrada na literatura até então.

Estudos recentes introduziram o algoritmo NUTS (HOFFMAN; GELMAN, 2014), o qual é uma extensão do algoritmo *Hamiltonian Monte Carlo* (HMC) (DUANE *et al.*, 1987; NEAL, 1994; NEAL, 2011), como uma boa alternativa ao MH ou ao GS. Hoffman e Gelman (2014), Granty *et al.* (2016) e Nugroho e Morimoto (2015) dizem que os algoritmos de HMC e o NUTS são capazes de trazer resultados mais acurados quando comparados a outros métodos MCMC, para diversos modelos estatísticos. Nesse capítulo, o uso do NUTS no contexto de MDCs é explorado, a fim de verificar se, para o modelo DINA, também há benefícios trazidos por esse algoritmo. Também é mostrada uma aplicação com dados relacionados à área de saúde mental, com um novo método de classificação para respondentes submetidos ao Inventário de Depressão de Beck (BECK *et al.*, 1961)

Como motivação para aplicação, temos que patologias psicológicas como a depressão são extremamente prejudiciais à qualidade de vida de quem as possui e é de suma importância que o diagnóstico seja rápido e acurado, a fim de evitar maiores complicações. Jaeger *et al.* (2006), Templin e Henson (2006) e Torre, Ark e Rossi (2015) aplicaram MDCs para o diagnóstico de pacientes baseados em suas respostas à instrumentos para avaliação da saúde mental. Nesse estudo, nós exploramos o uso do modelo DINA e do algoritmo NUTS em um contexto clínico, analisando dados de uma amostra de participantes que responderam a um questionário sobre a depressão. O interesse consiste em estimar a porcentagem de examinados que mostram cada um dos padrões comportamentais e também em encontrar um padrão comportamental individual para cada pessoa. Esses perfis são construídos utilizando descrições clínicas sob categorias da depressão baseadas na literatura já existente, como será mostrado mais adiante.

A fim de aplicar qualquer MDC, primeiramente é preciso identificar os atributos que estão por trás da construção do questionário e a relação entre cada uma das questões e esses atributos.

Tal mapeamento é chamado de Matriz **Q** e faz um link entre os itens (linhas) e os atributos (colunas). Diversas alternativas podem ser utilizados para identificar os atributos e construir a Matriz **Q**, bem como consultar experts na área de estudo e se utilizar de procedimentos estatísticos (LIU; HUGGINS-MANLEY; BRADSHAW, 2017; CHEN *et al.*, 2015). Nesse capítulo nós utilizamos uma abordagem que se baseia em resultados previamente obtidos e publicados na literatura Fragoso e Cúri (2013), onde é utilizada a TRI para avaliar respostas ao Inventário de Depressão de Beck, sendo identificadas duas dimensões para a depressão. Assim, em nosso trabalho, nós mapeamos os vinte e um itens do questionário, considerando dois atributos (dimensões) para a depressão: o cognitivo e o somático afetivo. Essa abordagem é explicada com mais detalhes ao longo do trabalho.

O principal objetivo desse capítulo é a proposta de um novo método de estimação para os parâmetros do modelo DINA, avaliando o desempenho desse algoritmo em relação à outras abordagens Bayesianas e frequentistas já presentes na literatura, bem como o desenvolvimento de uma aplicação mostrando a possibilidade do uso da metodologia proposta para a classificação de respondentes em um contexto clínico. Um estudo comparativo entre o NUTS e outros métodos de estimação é apresentado, tanto em cenários já presentes em simulações em outros trabalhos encontrados na literatura quanto em novos cenários simulados. Uma vez realizadas as simulações, a aplicação em um contexto clínico ilustra o potencial de uso dessa metodologia para questões práticas.

Esse capítulo é organizado da seguinte maneira: Na Seção 2.1 é apresentada uma revisão da literatura sobre o modelo DINA e os intuitos desse trabalho. Na Seção 2.2 o modelo DINA é apresentado em detalhe, bem como procedimentos como a ”marginalização” dos parâmetros discretos, a fim de se utilizar da técnica proposta de estimação dos parâmetros. Na Seção 2.3 são apresentadas informações sobre a estimação dos parâmetros sob abordagens Bayesianas e frequentistas, incluindo o algoritmo proposto nesse trabalho. Na Seção 2.4 são mostrados estudos de simulação, comparando a estimação com o uso do algoritmo NUTS com os resultados dos métodos de MH e GS e com a abordagem frequentista, com a comparação da recuperação de parâmetros e da eficiência dos métodos. Na Seção 2.5 é conduzido um estudo de dados reais, aplicando o modelo DINA com uso do NUTS para o diagnóstico da depressão. Finalmente, na Seção 2.6 são apresentadas conclusões, comentários e sugestões para pesquisas futuras.

2.2 O modelo DINA

Para o modelo DINA aqui apresentado é considerada uma situação na qual um teste com J itens dicotômicos é aplicado à N indivíduos, a fim de classificar cada um deles com base em seu domínio ou não domínio de K atributos (dimensões ou habilidades), previamente definidos. Uma vez que um indivíduo pode ou não possuir cada um dos atributos, existem $C = 2^K$ possíveis perfis de atributo, os quais são denotados pelo vetor $\boldsymbol{\alpha}_c = (\alpha_{c1}, \alpha_{c2}, \ldots, \alpha_{cK})'$, no qual α_{ck} é

igual a um se o perfil c possui o atributo k e é igual a zero, caso contrário. Quando lidamos com um vetor de atributos para o indivíduo i, denotamos esse vetor por $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})'$, com $i = 1, 2, \ldots, N$. Os atributos necessários para que haja sucesso no item j são armazenados no vetor $\mathbf{q}_j = (q_{j1}, q_{j2}, \ldots, q_{jK})'$, com $j = 1, 2, \ldots, J$, onde q_{jk} é 1 se o item j requer o atributo k e zero, caso contrário. A Matriz $\mathbf{Q}$ para o modelo DINA é definida por $\mathbf{Q} = (\mathbf{q}_1', \mathbf{q}_2', \ldots, \mathbf{q}_J')'$, essa matriz precisa ser previamente definida pelo pesquisador.

Consideramos que para cada indivíduo i e cada item j existe uma variável aleatória Y_{ij}, a qual assume o valor 0 para uma resposta incorreta (ou negativa em relação ao que se afirma naquele item) e 1 para uma resposta correta (ou positiva), como segue:

$$Y_{ij} | \boldsymbol{\alpha}_i, \boldsymbol{\Omega}_j \sim \text{Bernoulli}(P_{ij}), \tag{2.1}$$

onde $\boldsymbol{\Omega}_j = (g_j, s_j)$ é o vetor linha de parâmetros do item j, g_j (do inglês *guessing*, "chutar") é a probabilidade de um indivíduo que não possui os requisitos necessários para o item j respondê-lo corretamente (ou, considerando qualquer contexto, a probabilidade de sucesso no item j, dado que os atributos por ele mensurados não estão presentes no indivíduo i) e s_j (do inglês *slipping*, "escorregar") é a probabilidade de um indivíduo que possui os atributos necessários ao item j não respondê-lo corretamente (ou, considerando qualquer contexto, a probabilidade de não sucesso no item j, dado que os atributos por ele mensurados estão presentes no indivíduo i). P_{ij} é a probabilidade de sucesso, isto é, no contexto de um questionário, é a probabilidade de que o indivíduo i responda corretamente ao item j. Esse parâmetro de sucesso é definido por

$$P_{ij} = \text{P}(Y_{ij} = 1 | \boldsymbol{\alpha}_i, \boldsymbol{\Omega}_j) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}, \tag{2.2}$$

onde η_{ij} é a resposta ideal do indivíduo i para o item j. Definindo $\mathbb{1}(\cdot)$ como a função indicadora, temos que

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} = \mathbb{1}(\boldsymbol{\alpha}_i' \mathbf{q}_j = \mathbf{q}_j' \mathbf{q}_j)$$

.

É importante ressaltar que, para que a Equação 2.2 seja monôtoma com respeito a α_i, é necessário que $1 - s_j > g_j$ (JUNKER; SIJTSMA, 2001), isto é, $s_j + g_j < 1$, à qual chamamos de "restrição de monotonicidade" e será utilizada de maneira explícita na estimação pelo GS, como proposta por Culpepper (2015) e na modelagem utilizando o NUTS, como proposto nesse trabalho. Também é esperado que $s_j < 1 - g_j$, uma vez que um indivíduo que possui todos os atributos avaliados para um item, pela lógica, deve responder esse item mais corretamente que aqueles que não o possuem.

Note que η_{ij} age como uma função binária, assumindo o valor 1 se e somente se o indivíduo i possui todos os atributos requeridos pelo item j. Em outras palavras, η_{ij} faz a função do que se chama no nome do modelo DINA de *"and" gate*, e é interpretado como a resposta

ideal do indivíduo i ao item j. Sendo assim, assumimos que $\boldsymbol{\eta}_i = (\eta_{i1}, \eta_{i2}, \ldots, \eta_{iJ})'$ é o vetor de respostas ideais para o indivíduo i.

Os parâmetros de *guessing* e *slipping* são definidos utilizando probabilidades condicionais, de modo que

$$g_j = \mathrm{P}(Y_{ij} = 1 | \eta_{ij} = 0) \text{ e } s_j = \mathrm{P}(Y_{ij} = 0 | \eta_{ij} = 1).$$

Como pode ser visto na Equação (2.1), o modelo DINA nada mais é do que a distribuição de Y_{ij} condicionada a um vetor de atributos $\boldsymbol{\alpha}_i$ e um vetor de parâmetros de item $\boldsymbol{\Omega}_j$. Seja $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$ um vetor de respostas J dimensional do indivíduo i para cada um dos itens e $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \ldots, \boldsymbol{\Omega}_J)$ um vetor contendo os parâmetros de *guessing* e *slipping* para os J itens. Então, a probabilidade de observar $\mathbf{Y}_i$ condicionada a um determinado perfil c, considerando que as respostas aos itens são condicionalmente independentes dado $\boldsymbol{\alpha}_i$ é

$$\mathrm{P}(\mathbf{Y}_i | \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\Omega}) = \prod_{j=1}^{J} P_{cj}^{Y_{ij}} (1 - P_{cj})^{1 - Y_{ij}}, \tag{2.3}$$

onde $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c$ significa que o indivíduo i pertence ao perfil de atributos c e $P_{cj} = (1 - s_j)^{\eta_{cj}} g_j^{1 - \eta_{cj}}$.

Agora, seja $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_C)'$ um vetor C dimensional, com probabilidade marginal de um indivíduo pertencer ao perfil c dada por $\pi_c = \mathrm{P}(\boldsymbol{\alpha}_c) = \mathrm{P}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c | \boldsymbol{\pi})$. Obviamente, temos que $0 \leq \pi_c \leq 1$ e $\sum_{c=1}^{C} \pi_c = 1$.

Considerando (2.3), a probabilidade de observar $\mathbf{Y}_i$ condicionada aos parâmetros de item $\boldsymbol{\Omega}$ e ao vetor de probabilidades $\boldsymbol{\pi}$, dada pela marginalização do vetor $\boldsymbol{\alpha}_i$ é dada por

$$\mathrm{P}(\mathbf{Y}_i | \boldsymbol{\Omega}, \boldsymbol{\pi}) = \sum_{c=1}^{C} \pi_c \mathrm{P}(\mathbf{Y}_i | \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\Omega}). \tag{2.4}$$

A função de verossimilhança para uma amostra de N respondentes e J itens é dada por

$$\ell(\mathbf{Y} | \boldsymbol{\Omega}, \boldsymbol{\pi}) = \prod_{i=1}^{N} \sum_{c=1}^{C} \pi_c \mathrm{P}(\mathbf{Y}_i | \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\Omega}),$$

onde a matriz $\mathbf{Y} = [Y_{ij}]_{N \times J}$ representa as respostas de N indivíduos a um instrumento de mensuração com J itens.

2.3 Estimação do modelo DINA

Essa seção apresenta quatro diferentes maneiras de estimar os parâmetros do modelo DINA. Nosso principal objetivo é introduzir a estimação por meio do algoritmo NUTS, comparando os resultados obtidos com três métodos de estimação cujo uso já existia na literatura do DINA (uma explicação rápida sobre isso pode ser vista no Apêndice A): O algoritmo de

maximização da expectativa (EM) (DEMPSTER; LAIRD; RUBIN, 1977), o algoritmo de MH (METROPOLIS *et al.*, 1953; HASTINGS, 1970) e o GS (GEMAN; GEMAN, 1984; GELFAND; SMITH, 1990). A primeira parte dessa seção apresenta uma formulação Bayesiana do modelo DINA e a segunda parte descreve o algoritmo NUTS.

Modelo DINA sob uma perspectiva Bayesiana

Em contraste com a abordagem frequentista, na inferência Bayesiana os parâmetros de interesse são variáveis aleatórias, sendo necessária a especificação de distribuições de probabilidades *a priori* para os parâmetros, refletindo possíveis conhecimentos sobre seu comportamento e a incerteza em suas estimativas. A escolha das distribuições de probabilidade *a priori* é um importante passo na estatística Bayesiana, uma vez que as distribuições de probabilidade *a posteriori* serão influenciadas tanto pela amostra quanto pela priori. A estimação Bayesiana é baseada na distribuição *a posteriori*, que para o modelo DINA é dada por

$$p(\boldsymbol{\Omega}, \boldsymbol{\pi}|\mathbf{Y}) \propto \ell(\mathbf{Y}|\boldsymbol{\Omega}, \boldsymbol{\pi}) p(\boldsymbol{\Omega}) p(\boldsymbol{\pi}),$$

onde $p(\boldsymbol{\Omega})$ é a priori para os parâmetros de item e $p(\boldsymbol{\pi})$ é a priori para os perfis de atributo.

A formulação Bayesiana para o modelo DINA é apresentada em Culpepper (2015), considerando as seguintes distribuições *a priori*

$$\boldsymbol{\alpha}_i|\boldsymbol{\pi} \sim \text{Categórica}(\pi_1, \pi_2, \ldots, \pi_C), \tag{2.5}$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\delta}_0), \boldsymbol{\delta}_0 = (\delta_{01}, \delta_{02}, \ldots, \delta_{0C}), \tag{2.6}$$

$$s_j \sim \text{Beta}(a_s, b_s), \tag{2.7}$$

$$g_j \sim \text{Beta}(a_g, b_g). \tag{2.8}$$

A priori em (2.5) indica uma distribuição Categórica para o perfil de atributos de cada indivíduo i, com a probabilidade de pertencer a determinado atributo c sendo π_c. Ao considerar probabilidades iguais de pertencimento a cada atributo, essa priori será não informativa, correspondendo a uma distribuição uniforme discreta. A priori em (2.6) assume uma distribuição Dirichlet para o vetor de probabilidades $\boldsymbol{\pi}$, com o vetor de hiperparâmetros $\boldsymbol{\delta}_0 = (\delta_{01}, \delta_{02}, \ldots, \delta_{0C})$. É possível que pesquisadores definam valores para $\boldsymbol{\delta}_0$ de acordo com conhecimentos já adquiridos sobre os perfis de atributos, mas caso não haja informações prévias sobre $\boldsymbol{\pi}$, uma priori não informativa pode ser obtida ao definir $\delta_{0c} = 1$ para todo c.

As prioris em (2.7) e em (2.8) indicam uma distribuição Beta para s_j e g_j, com parâmetros a_s, b_s e a_g, b_g, respectivamente. Note que também podem ser obtidas prioris não informativas

ao considerar $a_s = b_s = a_g = b_g = 1$, o que faz com que a distribuição seja uniforme contínua. Culpepper (2015) impõe explicitamente a restrição de monotonicidade para os parâmetros de item ao tomar o produto das densidades das distribuições Beta como

$$p(\mathbf{\Omega}_j) = p(s_j, g_j) \propto s_j^{a_s-1}(1-s_j)^{b_s-1} g_j^{a_g-1}(1-g_j)^{b_g-1} \mathbb{1}((s_j, g_j) \in \mathscr{P}),$$

onde $\mathscr{P} = \{(s,g) : 0 \leq s+g < 1, 0 \leq s < 1, 0 \leq g < 1\}$. Utilizando essa restrição e considerando os parâmetros da distribuição Beta como 1, é obtida uma priori uniforme bivariada linearmente truncada para os parâmetros de item.

Culpepper (2015) introduziu uma formulação Bayesiana com distribuições condicionais completas para a estimação dos parâmetros do modelo DINA, apresentando alguns benefícios da formulação proposta. Essa formulação permite o uso do GS, ao apresentar condicionais completas que são analiticamente tratáveis tanto para os parâmetros de item quanto para os de indivíduos. Além disso, ela também permite que a restrição de monotonicidade $0 \leq g_j + s_j < 1$ seja explicitamente aplicada, o que faz com que a convergência seja mais rápida do que ao se utilizar softwares como WinBUGS e OpenBUGS, os quais não o fazem. Um outro benefício é o desenvolvimento de um algoritmo em C++ que está disponível no pacote "dina" do R, o qual é utilizado nesse trabalho.

Algoritmo No-U-Turn Hamiltonian Monte Carlo

Alguns métodos de MCMC amplamente conhecidos na estimação de parâmetros dos mais diversos modelos estatísticos, bem como o MH e o GS tem tendência a explorar o espaço paramétrico utilizando passeios aleatórios ineficientes, o que pode aumentar o número de iterações necessárias para atingir a convergência e a autocorrelação entre os valores gerados. Uma metodologia alternativa que vem sendo mais explorada recentemente são os métodos conhecidos HMC, os quais vem ganhando destaque para a estimação de parâmetros utilizando inferência Bayesiana. HMC é um método de MCMC que se utiliza de dinâmicas Hamiltonianas para construir cadeias de Markov e usualmente exigem um menor número de iterações para atingir a convergência. Dinâmicas Hamiltonianas podem ser utilizadas para descrever como um objeto se move em um sistema, bem como uma bola posta em uma rampa sem atrito a qual, conforme desce, converte energia potencial em energia cinética até que atinge o ponto mais inferior da rampa, a partir do qual precisa subir e converter energia cinética em energia potencial.

O movimento do objeto é descrito por sua localização $\mathbf{x}$ e momento $\mathbf{p}$ em determinado tempo t.. Para cada $\mathbf{x}$ existe uma energial potencial associada $U(\mathbf{x})$ e para cada $\mathbf{p}$ existe uma energia cinética $K(\mathbf{p})$. A energia total do sistema é definida como a soma da energia potencial e da cinética, isto é

$$\mathscr{H}(\mathbf{x}, \mathbf{p}) = U(\mathbf{x}) + K(\mathbf{p}),$$

onde $\mathscr{H}(\mathbf{x},\mathbf{p})$ é chamado de Hamiltoniano e é constante uma vez que o sistema é fechado. No exemplo da bola na rampa, o Hamiltoniano é constante porque não existe atrito.

O HMC utiliza as variáveis de localização $\mathbf{x}$ para representar as variáveis de interesse (parâmetros do modelo) e as variáveis de momento $\mathbf{p}$ como variáveis adicionais que permitem com que as dinâmicas Hamiltonianas operem. Dois parâmetros de especificação do método são definidos pelo usuário: o número de passos L e o tamanho de cada passo δ. Sendo assim, para cada iteração, o método HMC gera L pontos $(\mathbf{x},\mathbf{p})$ com o tamanho do passo δ, conforme a Figura 1.

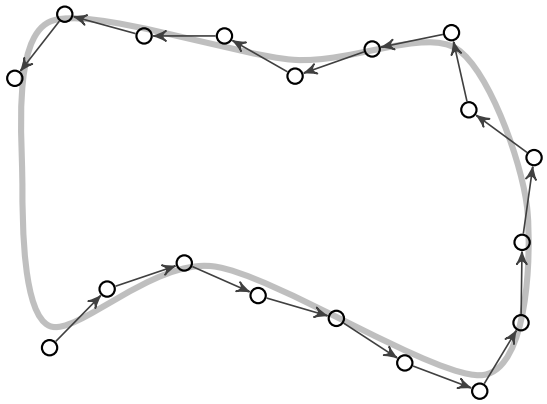


Figura 1 – Ilustração da trajetória do método HMC. L passos são construídos com tamanho de passo δ ao redor de $\mathscr{H}(\mathbf{x},\mathbf{p})$ a fim de obter o próximo estado da cadeia. O último passo gerado é escolhido para ser o próximo estado da cadeia de Markov com probabilidade de aceitação similar à probabilidade de aceitação de Metropolis.

Uma descrição mais detalhada do HMC pode ser vista no Apêndice A e em Duane *et al.* (1987), Neal (1994) e Neal (2011).

De acordo com Hoffman e Gelman (2014), a performance do HMC é bastante sensível à escolha dos dois parâmetros supracitados. A escolha errada dos parâmetros pode causar problemas como a inacurácia na simulação, alta taxa de rejeição ou o cálculo de mais passos que o necessário. A fim de eliminar a necessidade de escolher o parâmetro L manualmente, Hoffman e Gelman (2014) propôs o algoritmo No-U-Turn Sampler (NUTS), o qual é uma extensão do algoritmo HMC. Para cada iteração, o algoritmo NUTS constrói uma árvore com nós que são compostas de subárvores advindas de um processo recursivo, tal que uma subárvore com o dobro dos nós da árvore anterior é criada na mesma iteração em uma direção aleatória (para frente ou para trás). Em outras palavras, a subárvore j $(j = 0,1,2,\ldots)$ é gerada com 2^j nós na direção $v_j \sim \text{Uniform}(\{-1,+1\})$, indo para trás se $v_j = -1$ e para frente se $v_j = +1$. A Figura 2 mostra um exemplo de construção dessas subárvores

O k-ésimo nó $(\mathbf{x}_{(k)},\mathbf{p}_{(k)})$, para $k = 1,2,\ldots,2^j$, da j-ésima subárvore é selecionado com probabilidade

$$\alpha = \frac{\mathbb{1}\left(u \leq \exp\{-U(\mathbf{x}_{(k)}) - K(\mathbf{p}_{(k)})\}\right)}{n'},$$

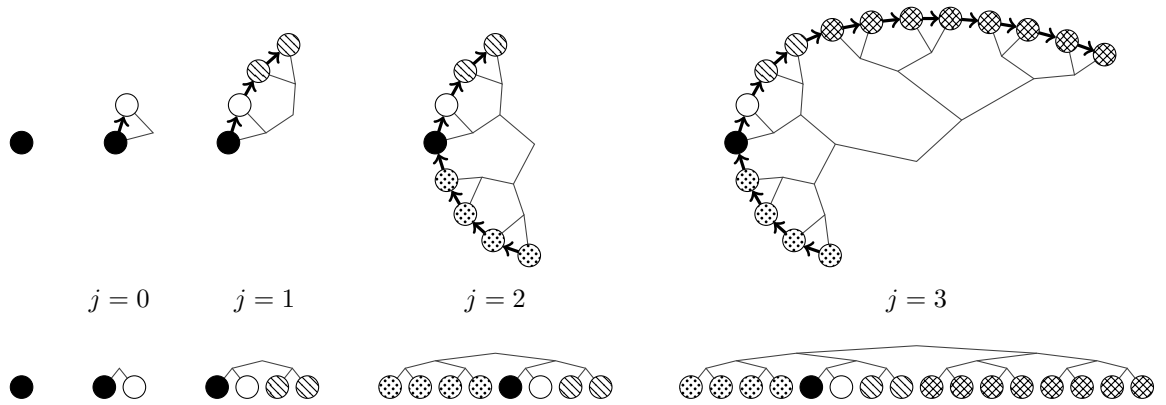


Figura 2 – Exemplo de construção de uma árvore binária similar a (HOFFMAN; GELMAN, 2014). Em cada passo o algoritmo dobra os nós, escolhendo aleatoriamente a direção. Nessa figura, as direções escolhidas para as quatro dobras foram frente (nó branco), frente (nó com linhas), trás (nó pontilhado) e frente (nós hachurados).

onde $n' = \sum_{h=1}^{k} \mathbb{1}\left(u \leq \exp\{-U(\mathbf{x}_{(h)}) - K(\mathbf{p}_{(h)})\}\right)$, i.e., n' é o número de nós na j-ésima subárvore até então, tal que $\mathbb{1}\left(u \leq \exp\{-U(\mathbf{x}_{(h)}) - K(\mathbf{p}_{(h)})\}\right) = 1$ para $h = 1, 2, \ldots, k$. Ao final da construção da J-ésima árvore, um nó pertencente à ela é escolhido como candidato para um novo estado da cadeia, com probabilidade $\dfrac{n'}{n}$, onde

$$n' = \sum_{h=1}^{2^J} \mathbb{1}\left(u \leq \exp\{-U(\mathbf{x}_{(h)}) - K(\mathbf{p}_{(h)})\}\right)$$

e

$$n = \sum_{j=0}^{J} \sum_{h=1}^{2^j} \mathbb{1}\left(u \leq \exp\{-U(\mathbf{x}_{(h)}) - K(\mathbf{p}_{(h)})\}\right).$$

Uma árvore é construída até que um extremo da trajetória comece a retroceder, voltando-se para o outro extremo. Assim, o candidato atual é o novo estado da cadeia.

O algoritmo NUTS pode ser utilizado via Stan, um software que se tornou disponível em 2012, sendo que para esse trabalho foi utilizada a versão 2.6. O Stan é um software livre e de código C++ aberto, cujo objetivo principal é realizar amostras de um modelo estatístico Bayesiano utilizando o algoritmo NUTS. Na prática, o Stan é similar ao BUGS (LUNN *et al.*, 2000) e ao JAGS (PLUMMER, 2003) nos quais o usuário escreve um modelo Bayesiano em um código conveniente. Entretanto, não é possível traduzir diretamente códigos de BUGS ou JAGS para Stan, uma vez que o NUTS e o HMC não permitem a amostragem direta de parâmetros discretos. Para modelos que envolvem parâmetros discretos é necessário realizar a marginalização desses parâmetros, como fazemos em (2.4) para o modelo DINA, a fim de escrever o código em Stan. O código em Stan utilizado nesse estudo está disponível no Apêndice B.

2.4 Estudo de simulação

Nessa seção um estudo de simulação é conduzido, a fim de avaliar a recuperação de parâmetros através do algoritmo NUTS e compará-la com outros métodos MCMC e com a estimação de máxima verossimilhança, por meio do algoritmo EM. O estudo de simulação é inspirado nos utilizados em (TORRE, 2009), (TORRE; DOUGLAS, 2004), (CULPEPPER, 2015) e (HUEBNER; WANG, 2011), principalmente no que diz respeito à Matriz $\mathbf{Q}$, a qual é a mesma que em Torre e Douglas (2004), com o número de dimensões $K = 5$.

Para o estudo de simulação, são considerados oito diferentes cenários, incluindo alguns que ainda não eram encontrados na literatura sobre o modelo DINA. Para o número de indivíduos temos que $N \in \{500, 1000\}$ e para o número de itens $J \in \{15, 30\}$. Em relação aos parâmetros do modelo, as configurações selecionadas foram divididas em dois casos, sendo que o primeiro tem $s = g = 0.2$ e uma distribuição chamada de *Flat* (o nome original proposto foi mantido por trazer melhor a ideia, mas em tradução livre poderíamos chamar de achatada, i.e., a probabilidade de pertencer a cada possível perfil é igual para todos os perfis) o segundo com $s = g = .1$ e uma distribuição chamada de *High* (crescente, i.e., a probabilidade de um indivíduo pertencer a um grupo com três ou mais dimensões é o dobro da de pertencer a grupos com menos de três dimensões). A distribuição das probabilidades para cada os vetores atributos $\boldsymbol{\alpha}_i$, de acordo com o número de atributos k de cada vetor, pode ser visto na Tabela 1.

		Número de atributos (k)					
		0	1	2	3	4	5
Distribuição *flat*	$P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c \vert k) = \boldsymbol{\pi}_c \vert k$	0.03125	0.03125	0.03125	0.03125	0.03125	0.03125
Distribuição *high*	$P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c \vert k) = \boldsymbol{\pi}_c \vert k$	0.021	0.021	0.021	0.042	0.042	0.042

Tabela 1 – Probabilidade de pertencer às classes de vetores de atributo $\boldsymbol{\alpha}_i$ de acordo com o número de atributos possuído pelos membros daquela classe.

De modo a avaliar o desempenho da estimação proposta pelo NUTS com aquelas que já existiam na literatura, três diferentes abordagens são consideradas:

1. Algoritmo EM + classificação EAP, utilizando o pacote "CDM" no software R (abordagem frequentista);
2. Algoritmo MH, pelo OpenBUGS;
3. Algoritmo GS, utilizando o pacote "dina" no R.

Para todos os métodos são geradas $R = 10$ réplicas para cada um dos oito cenários. Na abordagem Bayesiana, para cada cenário, são utilizadas 4000 iterações, descartando as primeiras 2000. Há diversas maneiras de realizar o diagnóstico da convergência em métodos MCMC. Uma dessas maneiras é a estatística de Gelman-Rubin (GELMAN; RUBIN, 1992), a qual compara o comportamento de uma cadeia com outras cadeias inicializadas automaticamente.

Essa comparação é feita estimando o $\hat{R}$, uma medida da razão da variância média das amostras em cada cadeia com a variância combinada entre todas. $\hat{R}$ será igual à 1 caso haja um equilíbrio entre essas variâncias e será maior que 1 se não houver convergência. A fim de verificar a convergência do algoritmo NUTS, utilizamos a estatística de Gelman Rubin (GELMAN; RUBIN, 1992) e análises gráficas.

Análise da recuperação de parâmetros

A fim de comparar o desempenho dos métodos de estimação em relação à recuperação dos parâmetros, são utilizadas duas estatísticas. Para calcular essas estatísticas nas abordagens Bayesianas, a média *a posteriori* foi utilizada como a estimativa dos parâmetros, comparando as estimativas com os valores reais em cada cenário simulado.

A primeira estatística é o RMSE (*square root of the mean square error*, raíz do erro quadrático médio), dada por

$$\text{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\widehat{\vartheta}_{lr} - \vartheta_l)^2},$$

onde ϑ_l é um elemento de $(\blacksquare, \boldsymbol{\pi})$ e l é um índice (j or c) e $\widehat{\vartheta}_{lr}$ é a estimativa obtida na réplica r, $r = 1, 2, \ldots, R$. A segunda estatística é o AVRB (*absolute value of the relative bias*, valor absoluto do viés relativo), dado por

$$\text{AVRB} = \frac{|\hat{\text{Vies}}|}{|\vartheta_l|},$$

onde $\hat{\text{Vies}} = (\widehat{\overline{\vartheta}}_l - \vartheta_l)$, com $\widehat{\overline{\vartheta}}_l = \frac{1}{R}\sum_{r=1}^{R}\widehat{\vartheta}_{lr}$.

Nessas análises, são mostradas a recuperação de parâmetros para os parâmetros dos itens g, s e também para a recuperação relacionada aos parâmetros dos indivíduos, por meio da probabilidade de pertença aos vetores de atributos $\boldsymbol{\pi}$. De acordo com (CULPEPPER, 2015), um RMSE maior que 0.04 não é desejável. O viés também é uma estatística útil a fim de comparar o quão distante as estimativas estão do verdadeiro parâmetro, sendo que o AVRB nos permite comparar parâmetros com escalas diferentes e maiores valores de AVRB indicam estimativas mais enviesadas.

A Tabela 2 mostra os resultados do RMSE para os parâmetros, considerando os quatro métodos abordados. Em relação aos parâmetros dos itens, o NUTS teve o menor valor de RMSE, junto com o GS, indicando que essas duas abordagens trouxeram os resultados com maior acurácia. A abordagem frequentista também traz bons resultados. Para a estimativa via MH, há valores altos de RMSE considerando o parâmetro g, para o qual o valor de RMSE fica acima de 0.10 em dois casos. Para os parâmetros s os valores de RMSE foram mais altos do que para g, em geral, o que também ocorre em Torre (2009) e Culpepper (2015). Ao considerarmos o cenário com a distribuição *Flat*, $s = 0.2$, $J = 15$ e $N = 500$, nota-se que o método com menor

valor de RMSE foi o NUTS. Para os demais cenários também houve valores de RMSE maiores que 0.04 para o MH, mas não para os demais métodos.

Para o parâmetro de indivíduos π, não houve problemas referentes ao RMSE para nenhum dos métodos, mas, mesmo em relação à esse parâmetro, os resultados são um pouco melhores para o NUTS e o GS quando comparados à abordagem frequentista e ao MH. Também é interessante notar que o valor de RMSE reduz com o o aumento do número de itens J, com o mesmo ocorrendo para o número de indivíduos N.

Em geral, as distribuições *Flat* trouxeram mais problemas do que as *High*, o que também foi observado por Culpepper (2015). O aumento no número de itens (J) faz com que os valores de RMSE sejam menores para g, o que também ocorre para s na maior parte dos casos, exceto para $s = .1$ *High* e $N = 500$. O aumento no número de respondentes N, em geral, também contribui para que as estimativas tenham maior acurácia, porém, mesmo para $N = 1000$ há casos com RMSE maior que 0.04 para o MH.

Tabela 2 – Resultados do estudo de simulação: RMSE utilizando diferentes métodos de estimação para o modelo DINA

Cenário	N	J	g				s				π			
			ML+EAP	MH	GS	NUTS	ML+EAP	MH	GS	NUTS	ML+EAP	MH	GS	NUTS
.2 Flat	500	15	.0374	.0584	.0336	.0335	.0505	.1121	.0472	.0469	.0155	.0126	.0121	.0120
		30	.0249	.0250	.0250	.0249	.0391	.0388	.0384	.0385	.0111	.0102	.0102	.0102
	1000	15	.0256	.0777	.0265	.0263	.0325	.1357	.0323	.0322	.0107	.0122	.0102	.0101
		30	.0171	.0172	.0172	.0172	.0300	.0294	.0295	.0294	.0069	.0071	.0071	.0071
.1 High	500	15	.0226	.1516	.0223	.0226	.0246	.2015	.0245	.0246	.0081	.0108	.0075	.0075
		30	.0176	.0180	.0179	.0179	.0254	.0255	.0256	.0256	.0054	.0050	.0050	.0050
	1000	15	.0152	.1496	.0156	.0155	.0169	.1963	.0173	.0172	.0061	.0087	.0058	.0058
		30	.0114	.0115	.0115	.0115	.0168	.0171	.0171	.0170	.0045	.0044	.0044	.0044

Os resultados para o AVRB são mostrados na Figura 3, com uma escala baseada no quadrado do AVRB para facilitar a comparação entre os cenários. Tais resultados confirmam que o algoritmo MH traz os piores resultados tanto para os parâmetros de indivíduos g e s, quanto para o de itens π. Além disso, o algoritmo NUTS continua sendo preciso para todos os parâmetros avaliados. Também é interessante notar que, em geral, o AVRB do parâmetro π é maior que aqueles para g e s.

Uma comparação da eficiência dos métodos MCMC

Nessa seção os algoritmos MCMC são comparados utilizando a abordagem de Sahu (2002) e Girolami e Calderhead (2011). Para cada cenário e método de estimação MCMC são calculados o tamanho efetivo da amostra (*effective sample size* (ESS)), definido em Gelman *et al.* (2014) como o número de amostras da posteriori, B, dividido pelo parâmetro de autocorrelação

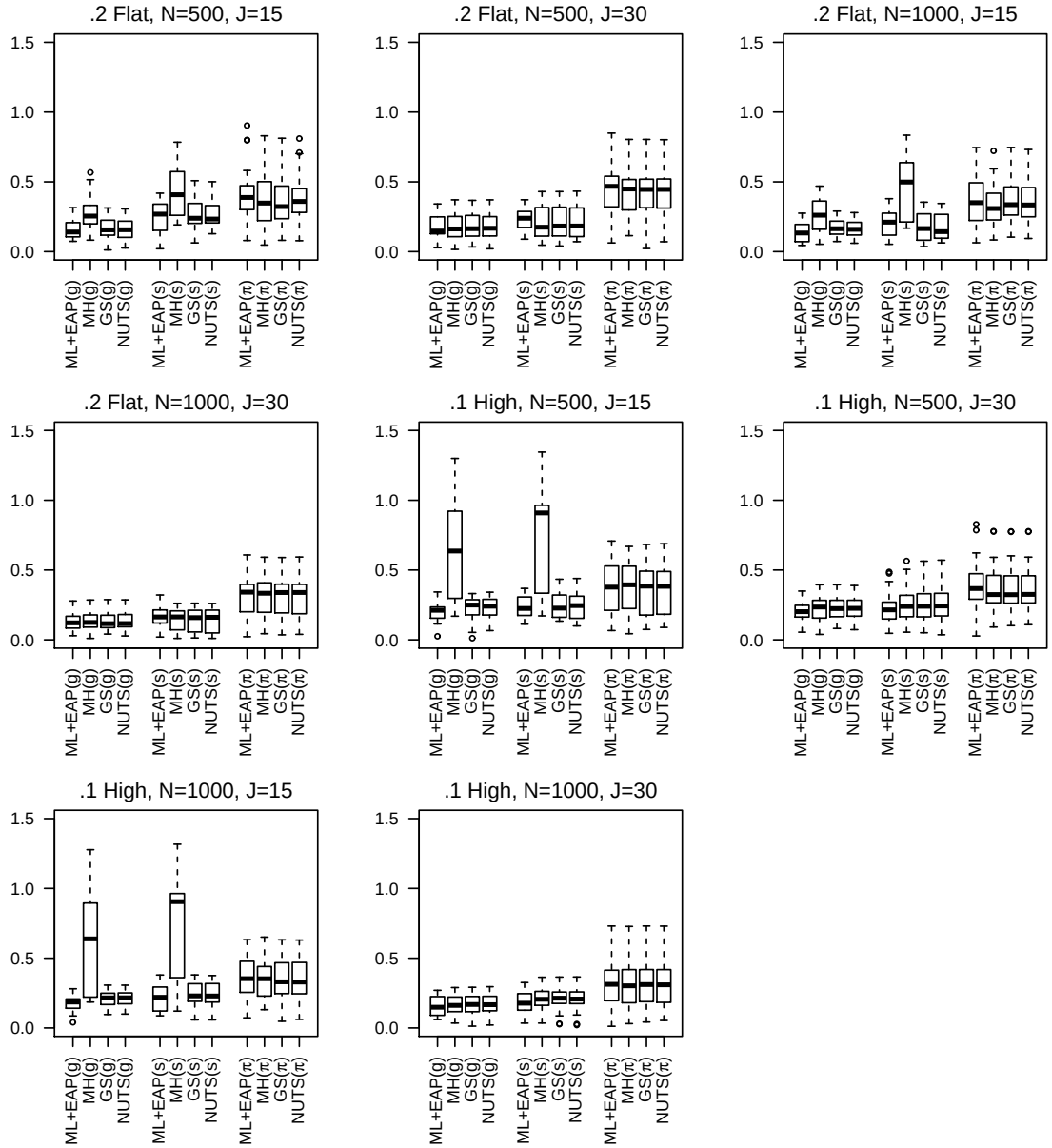


Figura 3 – Raiz quadrada dos valores de AVRB para os parâmetros g, s e π para os quatro métodos de estimação avaliados no modelo DINA.

temporal $\gamma = 1 + 2\sum_{k=1}^{\infty} \rho_k$, onde ρ_k é a autocorrelação amostral monótoma no *lag* k. O ESS é o número de amostras efetivamente independentes advindos da distribuição da posteriori, isso é, $ESS = \gamma/B$, podendo ser utilizado para informar sobre o grau de precisão obtido a partir das simulações. O ESS também pode ser normalizado pelo tempo requerido para que cada método MCMC seja rodado, resultando em NESS $= \dfrac{100 \times \text{ESS}}{s}$, onde s é o tempo em segundos. O pacote "coda" (PLUMMER *et al.*, 2006), no R, foi utilizado para obter os valores de ESS com base nas saídas dos métodos MCMC. O tempo computacional, a média de ESS e de NESS ao longo das réplicas são mostrados na Tabela 3.

Tabela 3 – Tempo computacional, média de ESS e de NESS para cada cenário e método MCMC.

Configurações	.2, Flat				.1, High			
N	500		1000		500		1000	
J	15	30	15	30	15	30	15	30
Tempo computacional								
MH	1977.83	6951.57	7722.97	34360.63	1938.03	7537.36	7629.43	33434.19
GS	37.65	84.45	83.96	165.57	39.28	89.57	85.53	164.20
NUTS	2429.57	4595.33	5511.71	9424.82	1433.77	3488.37	3373.07	6902.05
Média de ESS								
MH	228.43	283.48	269.47	302.61	257.72	270.05	280.30	320.72
GS	1061.14	1616.28	1447.71	1763.01	1661.18	1325.81	1822.58	1342.26
NUTS	1842.05	2142.78	1947.81	2057.88	1914.54	1457.38	1966.12	1420.78
Média de NESS								
MH	11.61	4.13	3.54	0.89	13.33	3.70	3.72	1.00
GS	2816.02	1963.57	1789.16	1101.93	4317.53	1552.92	2222.46	853.94
NUTS	77.36	48.97	38.52	23.31	134.09	46.21	64.10	23.37

Como podemos ver, o MH foi o método mais lento entre todos, seguido pelo NUTS. Considerando os cenários estudados, as distribuições *Flat* e *High* não apresentaram resultados muito diferentes, com exceção do NUTS, para o qual as distribuições *Flat* são consideravelmente mais lentas que as *High*. Em geral, o tempo computacional praticamente dobra para $J = 30$ itens quando comparado ao para $J = 15$ itens. O aumento no número de indivíduos também trouxe um aumento de tempo computacional considerável, para todos os métodos. Em termos de ESS, o algoritmo NUTS tem melhores resultados que o MH e o GS para todos os casos. Apesar disso, devido à velocidade do GS, o NESS é maior para ele do que para os demais métodos. No caso quando $N = 1000$ e $J = 30$ a diferença entre os valores de ESS para o GS e o NUTS diminui, mas, mesmo para esse cenário, o NUTS ainda traz melhores resultados.

Analisando os valores de ESS nota-se que o NUTS foi melhor (maiores valores de ESS) para todos os cenários, seguido pelo GS. Quando avaliamos os valores de NESS, é possível notar que o GS tem maiores valores para todos os cenários, seguido do NUTS, devido ao tempo computacional. Uma conclusão que podemos obter desses resultados é que, se não levarmos o tempo em conta, ao utilizarmos a mesma quantidade de iterações, o algoritmo NUTS tem resultados melhores que os demais. Entretanto, o algoritmo NUTS, através do software STAN, é bastante lento se comparado ao GS através do pacote "dina". Com futuros avanços na computação

espera-se que o tempo computacional do NUTS diminua, fazendo o tempo computacional um fator menos relevante.

A fim de comparar o NUTS e o HMC, também foram obtidos tempos computacionais, valores de ESSE e NESS para o cenário com distribuição .1 *High*, $N = 500$ e $J = 15$, utilizando o software STAN para ambos os casos. O algoritmo HMC é 5 vezes mais lento que o NUTS, tem valor de ESS igual à 2310.514 e NESS de 19.021. Isso indica que utilizar o HMC tem menor eficiência que a utilização do NUTS, com um tempo computacional também maior. Essa grande diferença no tempo de execução entre os dois métodos pode ser explicada pelo fato de que o HMC possui parâmetros setados inicialmente, enquanto o NUTS para automaticamente quando nota que os passos estão voltando ao mesmo local.

2.5 Aplicação: Análise da depressão usando o modelo DINA

Dados

Os dados utilizados nesse trabalho foram providenciados pelo Dr. Teng Chei-Tung, do Hospital das Clínicas, Faculdade de Medicina, Universidade de São Paulo. Eles advêm da resposta de $1,111$ estudantes universitários ao Inventário da Depressão de Beck (BDI; (BECK *et al.*, 1961)). O BDI é provavelmente o questionário mais comum sobre depressão e foi traduzido para diversos idiomas e validado em vários países. No Brasil Gorenstein *et al.* (1999) e YP. e Gorenstein (2013) trazem contribuições importantes nesse sentido. Os dados coletados foram validados utilizando o Alfa de Cronbach e a correlação total entre itens (FOX, 2010). A consistência interna considerando o Alfa de Cronbach é de 0.84 e a análise clássica de itens reportou uma média de correlação total entre itens de 0.43, indicando que os dados são confiáveis.

O BDI é utilizado para se obter um diagnóstico sobre a depressão, baseado em respostas para 21 questões de múltipla escolha em um inventário de autorrelato. Os itens avaliam dimensões da depressão (sintomas e atitudes), com intensidades variando desde neutra até um nível máximo de severidade, sendo ranqueadas entre 0 e 3. Um método usual para a avaliação da depressão é sugerido em Kendall *et al.* (1987), classificando os indivíduos em não depressivos (score de BDI $0-15$), disfóricos (score de BDI $16-20$) ou depressivos (score de BDI $21-63$), simplesmente somando os valores das respostas para cada item. Para o cálculo do score tradicional a escala original das respostas foi utilizada e para aplicar o modelo DINA, os dados são dicotomizados, sendo que o valor 0 foi atribuído às respostas iguais à zero e o valor 1 foi atribuído às respostas positivas (1, 2 ou 3).

Antes de seguir com a aplicação, é importante ter em mente que esse é um exemplo ilustrativo. A intenção é mostrar uma possível aplicação do modelo DINA para esse tipo de dados e os benefícios que ela pode trazer para sua análise. Para usos práticos da metodologia,

são necessários estudos mais aprofundados e discussões feitas por equipes interdisciplinares de pesquisadores, envolvendo não apenas estatísticos, mas também especialistas em áreas como a Psicologia e a Psiquiatria.

Especificação da Matriz Q

Para prosseguir com nossa abordagem do diagnóstico de depressão através da classificação dos respondentes por meio do modelo DINA, primeiro é necessário definir uma matriz **Q**. Diversas alternativas podem ser utilizadas, sendo que a literatura aponta, por exemplo, possibilidades como a consulta de especialistas na área de estudo ou a construção por meio de procedimentos estatísticos e algoritmos (LIU; HUGGINS-MANLEY; BRADSHAW, 2017; CHEN *et al.*, 2015; TORRE, 2008). Como levantado por Liu, Huggins-Manley e Bradshaw (2017), a definição correta da Matriz **Q** é um passo importante para que a classificação dos respondentes seja correta, impactando na acurácia e na confiabilidade das análises, sendo fortemente aconselhável que cada atributo possua ao menos um item que seja exclusivo para a mensuração do mesmo, não mensurando nenhum outro atributo simultaneamente (CHIU; DOUGLAS; LI, 2009; DECARLO, 2011), pois, caso isso não ocorra, as probabilidades *a posteriori* atribuidas à cada classe serão fortemente influenciadas pelas probabilidades *a priori*.

Para contruir nossa Matriz **Q** nessa aplicação, foi utilizado um estudo prévio sobre questionários de BDI utilizando Teoria de Resposta ao Item, considerando os resultados de Fragoso e Cúri (2013), os quais identificam duas dimensões relacionadas à depressão e definem como os itens do BDI são organizados entre essas dimensões, como pode ser visto na Figura 4. Em nossa abordagem, os 21 itens são mapeados considerando as dimensões identificadas no estudo citado, de forma que temos $K = 2$ dimensões: cognitiva (α_1) e somática afetiva (α_2). Assim, o número de possíveis perfis de atributos é $C = 4$. Com isso, consideramos para a construção da Matriz **Q** as classificações dos itens que podem ser vistas no diagrama de Venn em 4, baseado em Fragoso e Cúri (2013).

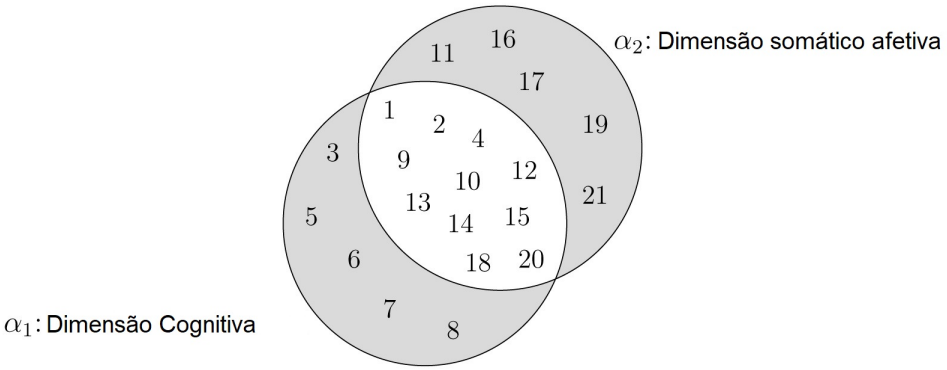


Figura 4 – Os Itens 3 e de 5 à 8 avaliam primariamente aspectos cognitivos da depressão, enquanto itens 11, 16, 17, 19, 21 avaliam primariamente aspectos somático-afetivos. Os demais itens avaliam ambas as dimensões de maneira balanceada.

Estimativas dos parâmetros de item

Os parâmetros do modelo foram estimados utilizando o algoritmo NUTS. São realizadas 4000 iterações, descartando a primeira metade. As prioris usadas nessa aplicação são baseadas na formulação Bayesiana do modelo DINA em (CULPEPPER, 2015) e apresentadas nas Equações (3.8) à (3.11).

Nessa aplicação, os parâmetros *guessing* são interpretados como a probabilidade de um indivíduo responder positivamente uma questão que aborde uma dimensão da depressão mesmo que ele(a) não possua, de fato, aquela dimensão. Os parâmetros *slipping* podem ser interpretados como a probabilidade de uma resposta negativa a uma característica da depressão mesmo que o(a) respondente possua aquela característica. Um indivíduo classificado na(s) dimensão(ões) cognitiva e/ou somático afetiva deve ser entendido como alguém que possui problemas relacionados àquela dimensão na qual foi classificado. De acordo com o modelo, se o/a respondente não é classificado em um desses aspectos, ele/ela é considerado(a) como sendo não depressivo(a).

As estimativas dos parâmetros dos itens e a especificação da Matriz $\mathbf{Q}$ podem ser vistas na Tabela 4. Os parâmetros *guessing* foram maiores que 0.3 para quatro dentre os vinte e um itens, mas não foram maiores que 0.5 para nenhum deles. Os parâmetros *slipping* foram maiores que 0.3 para treze itens dentre os vinte e um, com valores maiores que 0.5 para cinco itens. Isso é, de acordo com o modelo DINA, houve cinco itens do BDI tais que a probabilidade de um indivíduo expressamente dizer que ele/ela não possui os atributos avaliados naquele item, mesmo que o tenha, é maior que 50%. Há alguns possíveis motivos para isso ocorrer, tais quais, por essas questões serem mais sensíveis, os respondentes não admitirem que tem as características descritas naqueles itens, ou, eles de fato não possuírem aqueles sintomas, uma vez que são mais extremos. O Item (1) possui o maior valor de *guessing* ($\widehat{g}_1 = 0.469$) e o Item (9) o menor ($\widehat{g}_9 = 0.032$). Para os parâmetros de *slipping* o Item (1) é o menor ($\widehat{s}_1 = 0.102$) e o Item (19) o maior ($\widehat{s}_{19} = 0.850$). Itens com valores altos de g_j e s_j trazem menos informações para o diagnóstico quando avaliados pelo modelo DINA.

Altos valores de parâmetros de *slipping* e baixos valores de parâmetros de *guessing* mostram que a probabilidade de um indivíduo que tenha sido classificado como possuindo as características avaliadas por dado item responderem a ele negativamente é maior do que a probabilidade daqueles que não possuem ao menos uma daquelas características o responderem positivamente. No item a respeito de perda de peso (Item 19), por exemplo, 85% dos respondentes que foram avaliados como possuindo problemas somáticos-afetivos responderam negativamente.

A Figura 5 mostra intervalos de incerteza com base na posteriori dos parâmetros *guessing* (g) e *slipping* (s), com níveis de 95%. Em geral, os intervalos para os parâmetros *guessing* foram menores. Além disso, nenhum dos intervalos se destacou negativamente.

Tabela 4 – Matriz **Q** para a dominância dos aspectos da depressão avaliados pelo BDI e estimativas MCMC das probabilidades de g_j e s_j.

Item	**Q** (dimensões)		$\widehat{g}$		$\widehat{s}$	
	α_1	α_2	Média	dp	Média	dp
1. Tristeza	1	1	0.468	0.020	0.101	0.017
2. Pessimismo	1	1	0.189	0.017	0.334	0.026
3. Senso de fracasso	1	0	0.045	0.012	0.444	0.024
4. Falta de satisfação	1	1	0.309	0.020	0.163	0.020
5. Sentimentos de culpa	1	0	0.039	0.011	0.423	0.024
6. Senso de punição	1	0	0.115	0.017	0.453	0.023
7. Aversão a si mesmo	1	0	0.242	0.022	0.158	0.019
8. Auto acusação	1	0	0.422	0.023	0.163	0.017
9. Desejos suicidas	1	1	0.032	0.008	0.694	0.023
10. Crises de choro	1	1	0.142	0.014	0.502	0.026
11. Irritabilidade	0	1	0.283	0.024	0.279	0.023
12. Retraimento Social	1	1	0.210	0.017	0.394	0.025
13. Indecisão	1	1	0.205	0.016	0.320	0.025
14. Distorção da imagem do corpo	1	1	0.222	0.017	0.458	0.025
15. Inibição do trabalho	1	1	0.259	0.018	0.206	0.023
16. Distúrbio do sono	0	1	0.262	0.028	0.288	0.022
17. Fadiga	0	1	0.348	0.030	0.162	0.019
18. Perda de apetite	1	1	0.178	0.016	0.560	0.026
19. Perda de peso	0	1	0.062	0.012	0.851	0.016
20. Preocupação somática	1	1	0.223	0.016	0.518	0.026
21. Perda de libido	0	1	0.109	0.017	0.645	0.022

dp: desvio padrão.

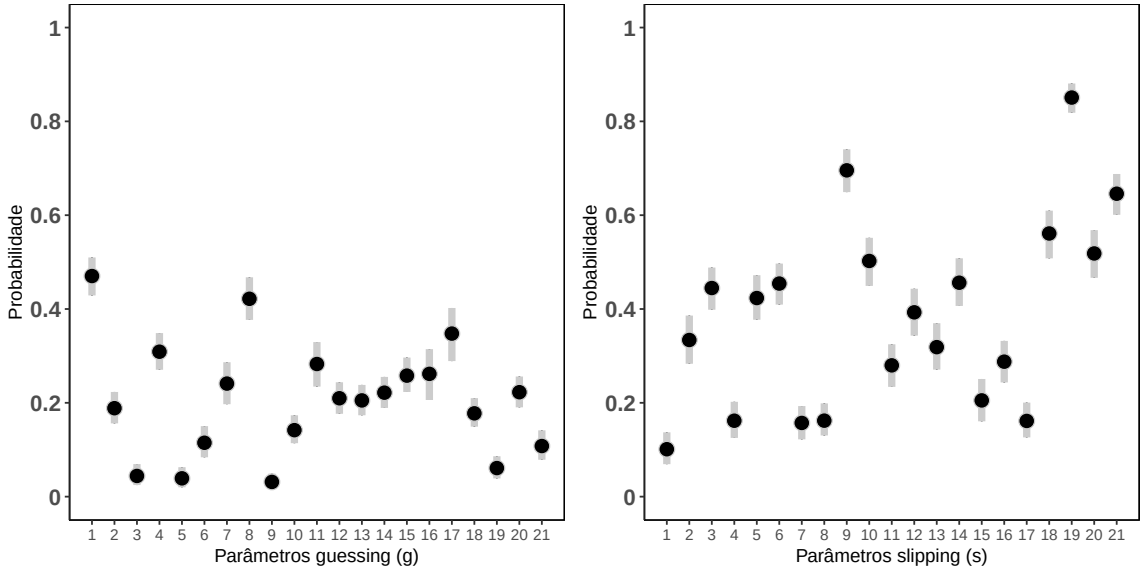


Figura 5 – Intervalos de incerteza *a posteriori* para os parâmetros *guessing* (g) e *slipping* (s) com nível de 95%.

Estimativas de perfis

A Tabela 5 mostra os resultados para as classes latentes relacionadas aos perfis de atritutos e suas respectivas probabilidades de ocorrência, utilizando o modelo DINA para os dados na aplicação.

Tabela 5 – Diagnóstico de depressão utilizando a Classificação *a posteriori* dos respondentes ao BDI ($n = 1111$)

c		Dimensões		$\widehat{\pi}$	
		α_1	α_2	Média	DP
1	(Não depressivo(a))	0	0	0.363	0.024
2	(Sintomático(a) para a dimensão cognitiva)	1	0	0.124	0.016
3	(Sintomático(a) para a dimensão somático afetiva)	0	1	0.124	0.021
4	(Sintomático(a) para ambas as dimensões)	1	1	0.389	0.019

DP: Desvio Padrão.

As maiores probabilidades são encontradas nas classes latentes nas quais os(as) respondentes não possuem nenhuma das características avaliadas ($\hat{\pi}_1 = 0.362$) ou quando ambas as características estão presentes ($\hat{\pi}_4 = 0.389$). Respondentes classificados(as) na classe latente $\boldsymbol{\alpha} = (0,0)$ são aqueles que não possuem problemas diagnosticados relacionados nem à dimensão cognitiva e nem à somático afetiva. Aqueles(as) classificados(as) na classe $\boldsymbol{\alpha} = (1,0)$ são diagnosticados com problemas relacionados à aspectos cognitivos mas não à aspectos somáticos afetivos, o contrário do que ocorre com a classe $\boldsymbol{\alpha} = (0,1)$, a qual indica problemas somáticos afetivos mas não cognitivos. A classificação em $\boldsymbol{\alpha} = (1,1)$ indica que o(a) respondente possui problemas relacionados à ambas as dimensões.

Em resumo, 36.2% dos respondentes são classificados no perfil não depressivo, sem nenhuma das dimensões presentes, 12.4% mostraram sintomas relacionados unicamente à dimensão cognitiva, 12.5% unicamente à dimensão somático afetiva e 38.9% em ambos os sintomas, com tais classificações sendo úteis para que os cuidados adequados sejam tomados para cada perfil.

Comparação do diagnóstico da abordagem tradicional e do modelo DINA

Também é interessante avaliar como as classes latentes são distribuídas considerando resultados obtidos através da abordagem tradicional para os scores do BDI. Os respondentes foram agrupados, seguindo a sugestão de Kendall *et al.* (1987), em não depressivos (score BDI $0-15$), disfóricos (score BDI $16-20$), e depressivos (score BDI $21-63$). A Tabela 6 mostra que todos os indivíduos no grupo de depressivos foram classificados no perfil 4 ($\boldsymbol{\alpha} = (1,1)$), isto é, o modelo DINA encontrou evidência de que todos os respondentes que, de acordo com a classificação tradicional do BDI, são classificados como depressivos, possuem sintomas tanto

para problemas cognitivos quanto para somático afetivos. Para respondentes classificados como disfóricos pela classificação tradicional, o modelo DINA encontrou evidências de que 95.61% tiveram sintomas em ambas as dimensões, sendo que 4.39% apresentaram apenas problemas relacionados aos aspectos cognitivos (perfil 2, $\boldsymbol{\alpha} = (1,0)$). No grupo dos não depressivos, segundo a classificação tradicional, a maioria foi classificada como não possuindo sintomas em nenhuma das dimensões avaliadas pela abordagem sob o modelo DINA (perfil 1, $\boldsymbol{\alpha} = (0,0)$), porém, um número considerável de respondentes desse grupo mostraram evidências de ao menos algum dos sintomas.

Tabela 6 – Distribuição de $\boldsymbol{\alpha}$ por grupos da avaliação tradicional do BDI.

Diagnóstico proposto pelo modelo DINA	Grupos de acordo com a proposta tradicional		
	Depressivos	Disfóricos	Não Depressivos
Não Depressivos	0(0%)	0(0%)	442(51.64%)
Sintomático(a) para a dimensão cognitiva	0(0%)	5(4.39%)	116(13.55%)
Sintomático(a) para a dimensão somático afetiva	0(0%)	0(0%)	106(12.38%)
Sintomático(a) para ambas as dimensões	141(100%)	109(95.61%)	192(22.43%)

A abordagem utilizando o modelo DINA nessa aplicação, a qual considera tanto os aspectos cognitivos quanto os somáticos afetivos da depressão, obtidos através da TRI, traz resultados que podem ser interpretados de maneira similar àqueles da classificação tradicional de scores BDI, no que diz respeito aos respondentes classificados como depressivos. No entanto, essa nova abordagem tem alguns resultados diferentes bastante interessantes para aqueles que são classificados como não depressivos sob a abordagem tradicional, podendo ser útil na classificação de indivíduos como parte do diagnóstico de depressão.

Entretanto, é notável que utilizar essa abordagem pode superestimar os resultados quanto à depressão, principalmente por causa da dicotomização utilizada, que faz com que todas as respostas positivas tenham o mesmo peso no diagnóstico final, independentemente do nível identificado pelos respondentes. Novamente, como foi enfatizado anteriormente, nosso exemplo com itens do BDI não é uma proposta clínica direta para uso, mas tem a intenção de mostrar que o modelo DINA pode ser ajustado para diferentes tipos de dados e motivar estudos mais aprofundados sobre as possibilidades trazidas por essa metodologia.

2.6 Conclusões finais e discussão

Nesse capítulo o uso do algoritmo NUTS para a estimação dos parâmetros do modelo DINA, sob uma abordagem Bayesiana, foi investigado. Inicialmente, o método foi estudado via simulações e, então, foi utilizado para identificar o perfil de pacientes com respeito a duas dimensões da depressão, utilizando os scores BID como parte do diagnóstico de depressão. Nosso estudo indicou que o NUTS tem resultados bastante bons na estimativa dos parâmetros do modelo DINA. Quando comparado a outros métodos já em uso, tanto sob abordagem Frequentista quanto Bayesiana, o NUTS teve resultados melhores ou, ao menos, similares no estudo de recuperação,

nos diversos cenários estudados. Quanto aos parâmetros *slipping*, o valor do RMSE é um pouco menor do que aquele apresentado em Culpepper (2015). Entretanto, o tempo necessário para o algoritmo NUTS é maior que o exigido por outras abordagens. Nós não podemos afirmar que isso se dá unicamente pela estrutura mais complexa do NUTS ou se o software STAN também é um fator de influência no tempo computacional.

O uso do algoritmo NUTS para o modelo DINA é recomendado, sendo que os resultados na recuperação de parâmetros mostraram que ele é bastante efetivo e traz resultados confiáveis. Considerando o quanto a capacidade computacional aumenta continuamente, em estudos futuros sobre MDC, é possível focar na investigação sobre o alto tempo computacional do NUTS, bem como checar o impacto de seu uso pelo software STAN. Também é possível desenvolver abordagens para outros MDCs com estimação via NUTS.

Com o devido cuidado quanto à sua extensão, a aplicação mostrou que o uso das duas dimensões em (FRAGOSO; CÚRI, 2013) para o diagnóstico da depressão pode ser bastante útil. Os resultados indicaram que utilizar mais de uma dimensão para a caracterização da depressão torna possível identificar pessoas que tenham dificuldades específicas em alguma das dimensões, as quais poderiam ser identificadas como não depressivas se a abordagem fosse mais generalista, trazendo um diagnóstico mais refinado, que pode ser útil para tratamentos mais bem guiados. Considerando o potencial observado ao se utilizar o modelo DINA na aplicação, uma possível abordagem para pesquisas futuras é o uso de um modelo DINA politômico para o BDI, considerando as quatro categorias originais $(0,1,2,3)$, bem como o estudo de testes relacionados à psicologia em outras escalas, como uma escala contínua no intervalo unitário.

O intuito desse trabalho é primariamente metodológico, com a aplicação sendo mais ilustrativa do que tendo a intenção de propor um método de diagnóstico a ser utilizado clinicamente para a detecção da depressão. Nós buscamos mostrar que o modelo DINA tem resultados bastante interessantes para dados relacionados à questionários nessa área do saber, mas, vale ressaltar novamente, é importante ter clareza de que para utilizar essa abordagem em prática seria necessário que uma equipe multidisciplinar estivesse envolvida e os estudos fossem amplamente validados. A aplicação nesse trabalho é uma ilustração inicial, a qual pode servir como motivação para estudos mais aprofundados, dado o potencial demonstrado.

CAPÍTULO

3

MODELO DINA DE RESPOSTA CONTÍNUA

Na literatura atual de modelos de variáveis latentes, bastante esforço foi realizado no desenvolvimento de Modelos de Diagnóstico Cognitivo para avaliações com itens dicotômicos e politômicos. Mais recentemente, a discussão sobre o uso de respostas contínuas foi trazida a tona. Entretanto, até o momento não há uma abordagem Bayesiana para a classe dos MDCs considerando respostas contínuas. O trabalho apresentado nesse capítulo traz a proposta de um esquema Bayesiano para o modelo DINA contínuo (C-DINA), assim contribuindo com o estado da arte com o desenvolvimendo de uma metodologia de estimação Bayesiana para tal modelo. Um estudo de sensibilidade de prioris é realizado, bem como um estudo de simulação para a recuperação de parâmetros, a fim de examinar a performance da abordagem Bayesiana. Os resultados mostram que a estimação tem bons resultados. O modelo proposto é, por fim, utilizado em uma aplicação para respostas sobre percepção de risco para atividades relacionadas à saúde e à exposição a determinadas situações externas, por meio de um questionário aplicado a uma amostra de indivíduos. Esta aplicação exemplifica o potencial de utilização do modelo proposto como um método de classificação em uma circunstância na qual temos diversas variáveis contínuas e sabemos quais os atributos de interesse utilizados para avaliação.

O trabalho desse capítulo será submetido a um periódico especializado.

3.1 Introdução

Como visto no Capítulo 2, inicialmente os MDCs utilizavam respostas dicotômicas para cada item, assumindo valor 0, se a resposta é incorreta ou discordante, dependendo do contexto do questionário, ou 1, se ela é correta ou concordante. Entretanto, a resposta para os itens de um teste podem assumir diferentes valores, não necessariamente dicotômicos. Por exemplo, as respostas podem ser discretas, porém politômicas ou até mesmo contínuas.

Nos MDCs, apesar de o traço latente ser discreto e dicotômico, as respostas podem assumir qualquer formato, não necessariamente sendo dicotômicas e nem mesmo discretas. Nos últimos anos, cada vez mais esforços vem sendo realizados na modelagem de respostas não dicotômicas. Para respostas discretas, porém politômicas, há diversos modelos desenvolvidos tanto em relação à TRI quanto aos MDCs. Na modelagem via TRI, detalhes podem ser encontrados em Nering e Ostini (2011) e para os MDCs há alguns artigos que discutem sobre esse tipo de resposta, como Chen e Torre (2013), Ma e Torre (2016), Tu *et al.* (2017).

Um teste com itens politômicos pode ser interessante, mas seu planejamento por vezes traz algumas dificuldades como, por exemplo, ter que determinar o número de categorias na resposta (III, 1980; PRESTON; COLMAN, 2000). O aumento no número de categorias, em alguns contextos, pode trazer análises mais informativas. Entretanto, um número muito grande de categorias pode trazer problemas de estimação, devido ao grande número de parâmetros que precisam ser estimados, levando ao questionamento se a utilização de respostas contínuas não poderia ser mais interessante em alguns casos.

Uma alternativa às respostas discretas é o uso de respostas no formato contínuo, sendo que, para esse tipo de respostas, há diversas possibilidades. Por exemplo, é possível pedir aos respondentes para que reportem seu nível de concordância com algumas afirmações, marcando um ponto em uma linha horizontal. Essa é uma escala visual e a medida da resposta é dada pela distância do ponto marcado ao canto esquerdo do segmento. Outra possibilidade é pedir aos respondentes para escolher um valor contínuo, por exemplo, entre 0% e 100%. Miller (1956) argumenta que esse tipo de avaliação traz maior riqueza de detalhes às análises do que a utilização de escalas com uma quantidade de valores pré estipulada. Respostas contínuas também aparecem na avaliação de intensidade da dor (MORIN; BUSHNELL, 1998) e em testes de personalidade (FERRANDO, 2001).

Outra fonte de respostas contínuas é o tempo de respostas à determinados itens, sendo que esse tipo de medida tende a ser cada vez mais comum, com o aumento de testes realizados em computadores. Há diversas aplicações relevantes referentes a tempos de resposta, por exemplo, detectar respostas aberrantes (LINDEN, 2008), avaliar a rapidez da resposta em diferentes contextos (LINDEN; XIONG, 2013; LINDEN, 2007), guiar a seleção de itens em testes adaptativos (FAN *et al.*, 2012), entre outras. Além disso, a modelagem dos tempos de resposta também foi aplicada para melhorar o processo de estimação de parâmetros para variáveis latentes (MENG;

TAO; CHANG, 2015) e (WANG; SAHA; DEY, 2016) e para classificar habilidades (SIE *et al.*, 2015).

Para a modelagem de respostas contínuas por meio da TRI, o Modelo de Resposta ao Item Beta foi proposto por Noel e Dauvier (2007). No que diz respeito aos MDCs para respostas contínuas, até pouco tempo atrás, não haviam trabalhos na literatura, sendo que uma primeira abordagem foi feita por Minchen, Torre e Liu (2017), os quais desenvolveram um modelo DINA contínuo (C-DINA), ainda havendo bastante espaço para avanços. Em aplicações práticas, a estimação dos parâmetros do modelo C-DINA é bastante desafiadora, uma vez que é afetada por diversos fatores, como o número de itens, o tamanho da amostra de respondentes e o número de habilidades avaliadas, sendo de grande valia o desenvolvimento de novas metodologias e o estudo dessas questões.

Como já foi ressaltado anteriormente, para modelos com respostas dicotômicas, há tanto abordagens frequentistas quanto Bayesianas para o DINA. Entretanto, para o modelo C-DINA, não há estudos sob a abordagem Bayesiana.

O principal objetivo desse capítulo é a proposta de uma abordagem Bayesiana para o modelo C-DINA, além de desenvolver uma explicação mais didática sobre a lógica por trás da construção do modelo, o que pode ser bastante útil para um melhor entendimento e não aparece usualmente na literatura, nem mesmo para modelos dicotômicos.

O capítulo está organizado da seguinte maneira. Na Seção 3.2 são apresentados detalhes sobre o modelo C-DINA. Na Seção3.3 apresentamos detalhes sobre o procedimento de estimação dos parâmetros sob a abordagem Bayesiana. A Seção 3.4 mostra estudos de simulação, avaliando algumas prioris e também mostrando informações sobre o sucesso na recuperação de parâmetros. Na Seção 3.5 um estudo com dados reais é conduzido, considerando dados sobre a percpeção de risco. Por fim, uma discussão e conclusão, bem como sugestões para futuras pesquisas, aparecem em 3.6.

3.2 Modelo DINA Contínuo (C-DINA)

Nesta seção, além de introduzir formalmente o modelo C-DINA, buscamos preencher algumas lacunas deixadas pela literatura na explicação desse modelo, com ênfase na sua interpretação. As explicações intuitivas sobre o modelo C-DINA, podem ser aplicadas, guardadas as devidas proporções, a outros MDCs.

Considere i como um índice para respondentes, com $i = 1,\ldots,N$, j um índice para itens, com $j = 1,\ldots,J$, e k um índice para atributos, com $k = 1,\ldots,K$. Para o modelo C-DINA, $Y_{ij} > 0$ é uma variável aleatória contínua, que representa a resposta do i-ésimo indivíduo ao j-ésimo item.

O vetor de atributos $\boldsymbol{\alpha}_i$, a matriz $\boldsymbol{Q}$ e os η_{ij} são construídos da mesma maneira que a

explicitada para o modelo DINA, na Seção 2.2.

Uma vez que Y_{ij} é uma medida contínua com valores positivos, a distribuição Log-Normal é uma candidata a ser considerada para a modelagem das respostas no C-DINA. Então, temos que Y_{ij} é condicional a η_{ij} e podemos escrever $[Y_{ij} \mid \eta_{ij} = \eta] \sim \text{LN}(\mu_{j\eta}, \sigma^2_{j\eta})$, com $\eta \in \{0,1\}$, onde $\text{LN}(\cdot,\cdot)$ denota a distribuição Log-Normal com $\mu_{j\eta}$ sendo relacionado à locação (a mediana é dada por $\exp\mu$) e $\sigma^2_{j\eta}$ relacionado também à dispersão. Denotando $f_{j\eta}(y_{ij})$ como a função de densidade de probabilidade (fdp) de uma distribuição Log-Normal dado $\eta_{ij} = \eta$, temos que

$$f_{j\eta}(y_{ij}) = \frac{1}{y_{ij}\sqrt{2\pi\sigma^2_{j\eta}}} \exp\left[-\frac{(log\ y_{ij} - \mu_{j\eta})^2}{2\sigma^2_{j\eta}}\right]. \tag{3.1}$$

Note que qualquer outra distribuição contínua que se adeque bem aos dados poderia ser utilizada como $f_{j\eta}(\cdot)$, não necessariamente a Log-Normal como na Equação (3.1).

Sob a abordagem frequentista, por mais que a alteração na distribuição de $f_{j\eta}(\cdot)$ na Equação (3.1) possa não trazer maiores problemas de interpretação, Minchen, Torre e Liu (2017) ressaltam que "seria necessária a realização de um trabalho substancial para que a estimação por máxima verossimilhança marginalizada seja refeita, bem como os algoritmos de erro padrão". Entretanto, sob a abordagem Bayesiana aqui proposta, a qual depende das amostras obtidas via MCMC da distribuição *a posteriori*, a revisão da distribuição em $f_{j\eta}(\cdot)$, em termos práticos, passa a ser mais fácil. Assim, em comparação com a abordagem frequentista, a Bayesiana possui a vantagem de ser bastante adaptável, permitindo aos pesquisadores a adoção do nosso processo de inferência para implementar o C-DINA com diferentes distribuições, dependendo da circunstância considerada. Além disso, para a abordagem frequentista, atualmente, não há pacotes ou programas implementados em algum software que permitam a fácil utilização em novas aplicações. Pensando no uso futuro de nossa proposta, o código em JAGS de nosso modelo está disponível no Apêndice D, tornando o uso do C-DINA por meio de nossa proposta Bayesiana fácil para outros pesquisadores.

Vale ressaltar que, tanto nesse capítulo quanto no seguinte, o JAGS foi utilizado devido ao STAN ter se mostrado bastante lento computacionalmente, para ambos os casos, inviabilizando seu uso.

Seja $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_J)'$ o vetor de parâmetros dos itens do modelo, onde $\boldsymbol{\Omega}_j = (\mu_{j0}, \sigma^2_{j0}, \mu_{j1}, \sigma^2_{j1})'$, para $j = 1, \cdots, J$ denota os parâmetros do j-ésimo item. Dado $\boldsymbol{\alpha}_i$ e $\boldsymbol{\Omega}_j$, a resposta Y_{ij} no modelo C-DINA será proveniente de uma mistura, uma vez que possui diferentes valores de parâmetros para os grupos $\eta = 0$ e $\eta = 1$, podendo ser representada por meio de uma fdp $f_j(y_{ij})$ com a seguinte configuração

$$[Y_{ij} \mid \boldsymbol{\alpha}_i, \boldsymbol{\Omega}_j] \sim f_j(y_{ij}) = \left[f_{j0}(y_{ij})\right]^{1-\eta_{ij}} \left[f_{j1}(y_{ij})\right]^{\eta_{ij}}, \tag{3.2}$$

$$= (1-\eta_{ij})\left[f_{j0}(y_{ij})\right] + \eta_{ij}\left[f_{j1}(y_{ij})\right]. \tag{3.3}$$

Em nosso modelo C-DINA proposto, é assumida independência entre os itens e também

entre os respondentes. Assim, a distribuição condicional de $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iJ})'$ dado $\boldsymbol{\alpha}_i$ e todos os parâmetros de item, pode ser escrita como

$$p(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\Omega}) = \prod_{j=1}^{J} f_j(y_{ij}) = \prod_{j=1}^{J} \left[f_{j0}(y_{ij})\right]^{1-\eta_{cj}} \left[f_{j1}(y_{ij})\right]^{\eta_{cj}}. \tag{3.4}$$

onde explicitamente expressamos $f_{j0}(y_{ij})$ e $f_{j1}(y_{ij})$ usando a Equação (3.1).

Finalmente, podemos definir $\pi_c = \mathrm{P}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c \mid \boldsymbol{\pi})$ como a probabilidade de pertencer à classe c, onde $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_C)'$ é um vetor C dimensional, $0 \leq \pi_c \leq 1$ e $\sum_{c=1}^{C} \pi_c = 1$. Então, pela lei da probabilidade total, a fdp de $\boldsymbol{y}_i$ dados todos os parâmetros de item $\boldsymbol{\Omega}$ e a classe latente de probabilidade $\boldsymbol{\pi}$ é

$$p(\boldsymbol{y}_i \mid \boldsymbol{\Omega}, \boldsymbol{\pi}) = \sum_{c=1}^{C} \pi_c p(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\Omega}). \tag{3.5}$$

Por fim, assumindo a distribuição Log-Normal para $f_{j\eta}(\cdot)$, como mostrado na Equação (3.1), a verossimilhança pode ser explicitamente escrita como

$$\begin{aligned} \mathscr{L}(\boldsymbol{\Omega}, \boldsymbol{\pi} \mid \boldsymbol{y}) = \prod_{i=1}^{N} \sum_{c=1}^{C} \pi_c \Bigg\{ \prod_{j=1}^{J} \left[\frac{\sqrt{\tau_{j0}}}{y_{ij}\sqrt{2\pi}} \exp\left[\frac{-\tau_{j0}(\log y_{ij} - \mu_{j0})^2}{2} \right] \right]^{1-\eta_{cj}} \times \\ \left[\frac{\sqrt{\tau_{j1}}}{y_{ij}\sqrt{2\pi}} \exp\left[\frac{-\tau_{j1}(\log y_{ij} - \mu_{j1})^2}{2} \right] \right]^{\eta_{cj}} \Bigg\}, \end{aligned} \tag{3.6}$$

onde $\boldsymbol{y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N)'$, $\tau_{j0} = 1/\sigma_{j0}^2$, e $\tau_{j1} = 1/\sigma_{j1}^2$. τ_{j0} e τ_{j1}, sendo essa uma transformação bastante utilizada em métodos MCMC. Minchen, Torre e Liu (2017) utilizou uma abordagem por máxima verossimilhança para estimar os parâmetros $\boldsymbol{\Omega}$ na Expressão (3.6). Na Seção 3.3, é apresentada a formulação Bayesiana para a estimação desses parâmetros, após uma explicação de maneira mais didática da interpretação da construção do modelo C-DINA.

Interpretação da construção do modelo C-DINA

Na literatura atual, há um espaço a ser preenchido sobre a explicação da lógica por trás da construção de MDCs, havendo falta de uma explicação mais detalhada. Pensando nisso, o modelo C-DINA que aqui é formulado sob uma perspectiva Bayesiana, é apresentado com maiores detalhes sobre sua lógica de construção, o que pode ser útil para um entendimento mais didático por pesquisadores de diversas áreas. O modelo C-DINA é um modelo com mais de um nível, sendo essa seção dedicada a trazer uma interpretação sobre a construção do modelo nesses diferentes níveis. Sem perda de generalidade, assumimos a distribuição Log-Normal para $f_{j\eta}(\cdot)$, como na Equação (3.1) para o desenvolvimento da explicação.

Nível 1

Para o desenvolvimento do modelo C-DINA, como já foi comentado anteriormente, considera-se o cenário no qual um teste com $j = 1, \ldots, J$ itens, mensurando $k = 1, \ldots K$ atributos

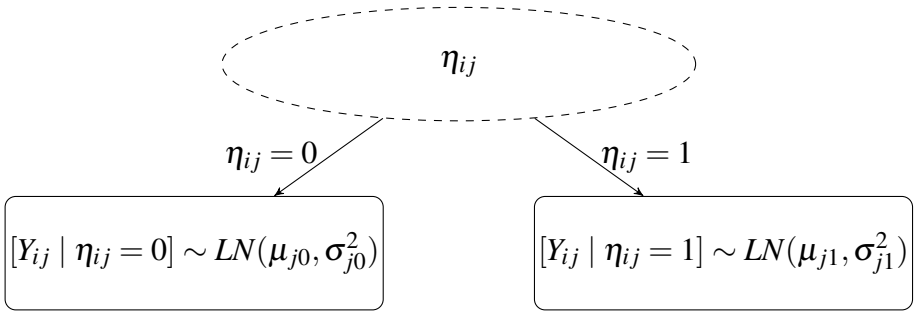


Figura 6 – Estrutura do Nível 1 para o modelo C-DINA, onde o formato elíptico indica variável latente e o formato retangular mostra dados observados.

latentes pré especificados é realizado por $i = 1, \ldots, N$ respondentes. Cada um dos indivíduos dá uma resposta contínua à cada item. Os dados coletados formarão o vetor de respostas $\boldsymbol{y}$, com dimensões $N \times J$ e cada elemento y_{ij} sendo contínuo e positivo.

Como mostrado na Figura 6, para o primeiro nível do modelo C-DINA, faz sentido particionar os respondentes em dois grupos, o que faz com que tenhamos um modelo de misturas, bem como ocorre nos modelos de perfis latentes (OBERSKI, 2016). Porém, diferentemente dos modelos de perfis latentes, nos quais o intuito é realizar o agrupamento (*clusterização*), apenas separando os componentes de cada grupo, sem se preocupar diretamente com a interpretação e natureza dos mesmos, para um MDC a origem dos grupos e sua interpretabilidade é importante em sua construção.

Um dos grupos será composto pelos respondentes que possuem todos os atributos requeridos pelo j-ésimo item, os quais farão parte do grupo $\eta_{ij} = 1$. Já os demais respondentes, que não possuem ao menos um dos atributos requeridos, irão compor o grupo $\eta_{ij} = 0$. Assim sendo, a resposta contínua dos indivíduos para cada um dos itens será condicional à η_{ij}. Para levar isso em conta, a fim de modelar as respostas, precisamos espeficiar distribuições contínuas com diferentes parâmetros para cada grupo, como mostrado na Figura 6.

No Nível 1, se desejamos estimar os parâmetros μ_{j0}, μ_{j1}, σ^2_{j0}, σ^2_{j1} para cada item, de acordo com a Equação (3.6), é necessária informação sobre a variável de resposta latente η_{ij}, o que conduz à discussão sobre o segundo nível de hierarquia na construção do modelo C-DINA.

Nível 2

Uma vez que as quantidades no primeiro nível foram definidas, podemos ir para o segundo, pensando sobre como computar η_{ij}. Seguindo a lógica de construção do modelo C-DINA, cada item irá avaliar um conjunto de dimensões e um respondente deve ter domínio de todas as dimensões avaliadas no item para ter uma boa performance.

Para modelar se um respondente possui determinada dimensão, introduzimos o vetor $\boldsymbol{\alpha}_i$, o qual é composto por quantidades binárias α_{ik}. Cada valor de $\boldsymbol{\alpha}_i$ terá uma probabilidade de

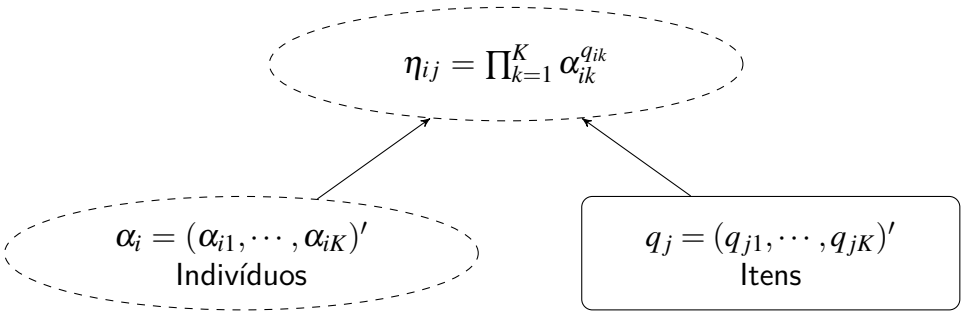


Figura 7 – Estrutura do Nível 2 do modelo C-DINA, onde formato elíptico indica variável latente e formato retangular dados observados.

ocorrer, levando à 2^K diferentes perfis de atributos. Assim, podemos definir $\pi_c = \mathrm{P}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c)$, com $\sum_{c=1}^{C} \pi_c = 1$, onde $\boldsymbol{\alpha}_c$ é um dos possíveis $C = 2^K$ perfis de atributo.

Um conceito importante do modelo C-DINA, o qual apareceu anteriormente para falar sobre os possíveis perfis de atributo, é que cada item avalia um conjunto de $K \geq 2$ dimensões de traços latentes. Ao longo de todo o teste há ao menos duas dimensões de traços latentes, sendo que, cada um dos itens avalia uma ou mais dessas dimensões. Cada dimensão precisa ser avaliada por ao menos um item. Essa informação é escrita na Matriz $\boldsymbol{Q}$, a qual possui J linhas e K colunas.

A Figura 7 mostra a hierarquia do modelo C-DINA no Nível 2. O valor de η_{ij} no Nível 1 é determinado tanto pelo perfil individual $\boldsymbol{\alpha}_i$ quanto pelo atributo de item $\boldsymbol{q}_j$, como mostrado na Figura 7. Também podemos computar a probabilidade de $\eta_{ij} = 1$, a qual é dada por

$$\mathrm{P}(\eta_{ij} = 1) = \prod_{\{k:q_{jk}=1\} \cap \{1,\cdots,K\}} \mathrm{P}(\alpha_{ik} = 1). \tag{3.7}$$

3.3 Estimação Bayesiana do modelo C-DINA

A expressão na Equação (3.3) e a Figura 6 tornam fácil de ver que a distribuição dos y_{ij} no modelo C-DINA é um caso especial de uma distribuição de mistura. Como indicado por Jasra, Holmes e Stephens (2005), um dos maiores desafios de se utilizar uma abordagem Bayesiana para misturas é a não identificabilidade dos componentes da mistura, a qual leva ao problema chamado de *label switching*.

Esse problema é discutido na literatura de modelos de mistura (STEPHENS, 2000; CASSIDAY; CHO; HARRING, 2020), também sendo incluidos na discussão de trabalhos sobre TRI (LI *et al.*, 2009). O *label switching* pode ocorrer tanto dentro das cadeias quanto entre cadeias, quando há permutação entre os parâmetros de cada componente da mistura. Como levantado por (STEPHENS, 2000), a raiz de sua existência é que a verossimilhança é a

mesma para as permutações do vetor de parâmetros e, caso não houver informações *a priori* que distinguam entre os componentes da mistura, com uma distribuição a priori igual para tais componentes, a posteriori será igualmente simétrica, podendo trazer problemas quando a estimação de componentes individuais da mistura é realizada, sendo mais comum para casos nos quais não há grande discriminação entre os grupos de cada componente da mistura.

A maneira de lidar com esse problema é variada, havendo várias propostas de acordo com cada diferente situação na literatura. Nesse capítulo, seguimos a ideia de (LUNN *et al.*, 2012), para distribuições contínuas, a qual aponta que uma possível solução é a de reparametrizar o modelo considerando que $\mu_{j1} = \mu_{j0} + \lambda_j$, com $\lambda_j > 0$. Note que essa parametrização, considerando a distribuição Log-Normal pressupõe que a mediana ($\exp\mu$) dos dados dos indivíduos pertencentes ao grupo $\eta = 1$ é maior que a dos que pertencem ao grupo $\eta = 0$, o que faz sentido pela lógica de construção do modelo C-DINA.

Para facilitar a computação, utilizaremos $\tau_{j\eta} = 1/\sigma^2_{j\eta}$ ao invés de $\sigma^2_{j\eta}$ para cada um dos grupos η. O conjunto de parâmetros desconhecidos do $j-$ésimo item é então dado por $\boldsymbol{\Omega}_j = (\mu_{j0}, \lambda_j, \tau_{j0}, \tau_{j1})'$.

A seguir, são discutidas as especificações de priori e a estimação de parâmetros é apresentada em detalhes.

Especificação de Prioris

A especificação das distribuições *a priori* é um passo chave para proceder com a abordagem Bayesiana. Considerando as características dos parâmetros que compõem o modelo C-DINA, podemos distingui-los entre "parâmetros dos indivíduos" ($\boldsymbol{\alpha}$ e $\boldsymbol{\pi}$) e "parâmetros dos itens" ($\boldsymbol{\Omega}$). Ambos os tipos de parâmetros dependem do grupo latente η, sendo apropriado dizer que tais parâmetros são parâmetros de "indivíduo $\times$ grupo latente"e de "item $\times$ grupo latente.

Parâmetros dos indivíduos

Em primeiro lugar, vamos discutir sobre a escolha das distribuições *a priori* para os parâmetros dos indivíduos.

Dado $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_C)'$, com $\sum_{c=1}^{C} \pi_c = 1$, uma distribuição categórica pode ser escolhida para cada vetor $\boldsymbol{\alpha}_i$, isto é

$$[\boldsymbol{\alpha}_i|\boldsymbol{\pi}] \overset{ind}{\sim} \text{Categorica}(\boldsymbol{\pi}),\ i = 1, \ldots, N, \tag{3.8}$$

parâmetro tal que mensura a probabilidade de um indivíduo específico pertencer a um dos diversos perfis de atributos. Se assumirmos os hiperparâmetros $\pi_1 = \pi_2 = \ldots = \pi_C = 1/C$ na Equação (3.8), a priori para $[\boldsymbol{\alpha}_i|\boldsymbol{\pi}]$ se torna não informativa. Entretanto, uma hiperpriori é definida para $\boldsymbol{\pi}$.

Podemos utilizar a distribuição de Dirichlet como segue,

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\delta}_0),\ \boldsymbol{\delta}_0 = (\delta_{01}, \delta_{02}, \ldots, \delta_{0C})'. \tag{3.9}$$

Se $\delta_{01} = \delta_{02} = \ldots = \delta_{0C} = 1$, a hiperpriori de $\boldsymbol{\pi}$ será não informativa.

Para $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N)$ assumimos independência entre os parâmetros dos indivíduos $\boldsymbol{\alpha}_i$ dado $\boldsymbol{\pi}$, de modo que a distribuição conjunta para $\boldsymbol{\alpha}$ e $\boldsymbol{\pi}$ é dada por

$$p(\boldsymbol{\alpha}, \boldsymbol{\pi}) = \left[\prod_{i=1}^{N} p(\boldsymbol{\alpha_i} \mid \boldsymbol{\pi})\right] p(\boldsymbol{\pi}) = \left(\prod_{i=1}^{N}\prod_{c=1}^{C} \pi_c^{\mathbb{1}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c)}\right)\left(\frac{1}{B(\boldsymbol{\delta}_0)}\prod_{c=1}^{C} \pi_c^{\delta_{0c}-1}\right) \tag{3.10}$$

onde $B(\boldsymbol{\delta}_0) = \prod_{i=1}^{C}\Gamma(\delta_{0i})/\Gamma(\sum_{i=1}^{C}\delta_{0i})$.

Parâmetros dos itens

Vamos discutir agora as escolhas das prioris para os parâmetros dos itens, $\boldsymbol{\Omega}_j = (\mu_{j0}, \lambda_j, \tau_{j0}, \tau_{j1})'$, para $j = 1, \ldots, J$.

Primeiramente, uma escolha natural para a priori de μ_{j0} é uma distribuição Normal, isto é

$$\mu_{j0} \sim \text{Normal}(\mu_{\mu_0}, \tau_{\mu_0}^{-1}),$$

onde μ_{μ_0} é a média e τ_{μ_0} é um parâmetro de precisão. Usualmente, é razoável assumir que $\mu_{\mu_0} = 0$, com τ_{μ_0} podendo ser fixado ou ter um hiperparâmetro adicionado. O estudo de simulação na Seção 3.4 mostra a consideração de uma hiperpriori para τ_{μ_0}, i.e., $\tau_{\mu_0} \sim \text{Gamma}(c_{\mu_0}, d_{\mu_0})$, a qual tem melhores resultados considerando um cenário de (MINCHEN; TORRE; LIU, 2017), sendo essa especificação adotada tanto para as simulações quanto para a aplicação.

De acordo com nossa suposição, temos que $\lambda_j > 0$, portanto, uma priori Gama é uma boa escolha para λ_j,

$$\lambda_j \sim \text{Gama}(a_\lambda, b_\lambda),\ j = 1, \ldots, J, \tag{3.11}$$

Onde $a_\lambda = b_\lambda = 0.01$ faz com que a priori seja menos informativa. Note que a definição com parâmetros de forma e proporção para a distribuição Gama é utilizada.

Para os parâmetros τ_{j0} e τ_{j1} em $\boldsymbol{\Omega}_j$, a distribuição escolhida deve ter o domínio na reta positiva. Uma boa candidata é, novamente, a distribuição Gama. Considerando isso, propomos o uso de

$$\tau_{j0} \sim \text{Gamma}(c_0, d_0),\ \text{e}\ \tau_{j1} \sim \text{Gamma}(c_1, d_1),\ j = 1, \ldots, J, \tag{3.12}$$

onde $c_0 = d_0 = 0.01$ e $c_1 = d_1 = 0.01$ são escolhas que tornam a priori menos informativa.

Por fim, assumimos independência entre os parâmetros de item para cada item $j = 1,\ldots,J$, com uma distribuição conjunta de $\boldsymbol{\Omega}$ e τ_{μ_0} dada por

$$\begin{aligned} p(\boldsymbol{\Omega},\tau_{\mu_0}) &= \left[\prod_{j=1}^{J} p(\lambda_j)p(\tau_{j0})p(\tau_{j1})p(\mu_{j0}\,|\tau_{\mu_0})\right] p(\tau_{\mu_0}) \\ &= \left\{\prod_{j=1}^{J} \frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)}\lambda_j^{a_\lambda-1}\exp(-b_\lambda\lambda_j) \times \frac{d_0^{c_0}}{\Gamma(c_0)}\left(\tau_{j0}\right)^{c_0-1}\exp\left(-d_0\tau_{j0}\right)\right. \\ &\times \left.\frac{d_1^{c_1}}{\Gamma(c_1)}\left(\tau_{j1}\right)^{c_1-1}\exp\left(-d_1\tau_{j1}\right) \times \sqrt{\frac{\tau_{\mu_0}}{2\pi}}\exp\left[-\frac{\tau_{\mu_0}\mu_{j0}^2}{2}\right]\right\} \\ &\times \frac{d_{\mu_0}^{c_{\mu_0}}}{\Gamma(c_{\mu_0})}\left(\tau_{\mu_0}\right)^{c_{\mu_0}-1}\exp\left(-d_{\mu_0}\tau_{\mu_0}\right). \end{aligned} \tag{3.13}$$

Estimação dos Parâmetros

Para sumarizar, o modelo C-DINA pode ser escrito de maneira hierárquica como

$$\begin{aligned} y_{ij} \mid \boldsymbol{\alpha}_i,\boldsymbol{\Omega}_j,\eta_{ij} &\sim \eta_{ij}\text{LN}(\mu_{j0}+\lambda_j,\tau_{j1}^{-1}) + (1-\eta_{ij})\text{LN}(\mu_{j0},\tau_{j0}^{-1}), \\ \eta_{ij} &= \prod_{k=1}^{K}\alpha_{ik}^{q_{jk}} = \mathbb{1}(\boldsymbol{\alpha}_i'\mathbf{q}_j = \mathbf{q}_j'\mathbf{q}_j), \\ \boldsymbol{\alpha}_i|\boldsymbol{\pi} &\sim \text{Categorical}(\boldsymbol{\pi}), \\ \boldsymbol{\pi} &\sim \text{Dirichlet}(\boldsymbol{\delta}_0), \\ \mu_{j0} &\sim \text{Normal}(0,\tau_{\mu_0}^{-1}), \\ \lambda_j &\sim \text{Gama}(a_\lambda,b_\lambda) \\ \tau_{j0} &\sim \text{Gama}(c_0,d_0) \\ \tau_{j1} &\sim \text{Gama}(c_1,d_1) \\ \tau_{\mu_0} &\sim \text{Gama}(c_{\mu_0},d_{\mu_0}), \end{aligned} \tag{3.14}$$

onde $\boldsymbol{\Omega}_j = (\mu_{j0},\lambda_j,\tau_{j0},\tau_{j1})'$, $\boldsymbol{\pi} = (\pi_1,\pi_2,\ldots,\pi_C)'$ e $\boldsymbol{\delta}_0 = (\delta_{01},\delta_{02},\ldots,\delta_{0C})'$. Embora a verossimilhança (3.6) tenha marginalizado $\boldsymbol{\alpha}$, a sumarização nessa verossimilhança não é conveniente para o desenvolvimento de um algoritmo MCMC. Portanto, incluímos o vetor latente $\boldsymbol{\alpha}_i$ para cada indivíduo na distribuição *a posteriori* do modelo C-DINA, tendo assim

$$\begin{aligned} & p(\boldsymbol{\Omega},\tau_{\mu_0},\boldsymbol{\alpha},\boldsymbol{\pi}\mid\mathbf{y}) \\ \propto\ & \mathscr{L}(\mathbf{y}|\boldsymbol{\Omega},\tau_{\mu_0},\boldsymbol{\alpha}) \times p(\boldsymbol{\Omega},\tau_{\mu_0}) \times p(\boldsymbol{\alpha}\mid\boldsymbol{\pi})p(\boldsymbol{\pi}) \\ \propto\ & \left\{\prod_{i=1}^{N}\prod_{j=1}^{J}\left[\frac{\sqrt{\tau_{j0}}}{y_{ij}\sqrt{2\pi}}\exp\left[\frac{-\tau_{j0}(\log y_{ij}-\mu_{j0})^2}{2}\right]\right]^{1-\eta_{ij}}\right. \\ \times\ & \left.\left[\frac{\sqrt{\tau_{j1}}}{y_{ij}\sqrt{2\pi}}\exp\left[\frac{-\tau_{j1}(\log y_{ij}-\mu_{j0}-\lambda_j)^2}{2}\right]\right]^{\eta_{ij}}\right\} \times p(\boldsymbol{\Omega},\tau_{\mu_0}) \times p(\boldsymbol{\alpha},\boldsymbol{\pi}), \end{aligned} \tag{3.15}$$

onde $p(\boldsymbol{\alpha},\boldsymbol{\pi})$ são definidos na Equação (3.10) e $p(\boldsymbol{\Omega},\tau_{\mu_0})$ são definidos na Equação (3.13).

O modelo foi ajustado utilizando o código escrito em JAGS (PLUMMER, 2015) em D, o qual é utilizado por meio de sua interface no software R "R2jags" (SU; YAJIMA, 2012).

A posteriori conjunta (3.15) não possui forma fechada, mas é possível encontrar distribuições conhecidas para as condicionais completas de cada parâmetro desconhecido, exceto por λ_j. Por mais que o JAGS tenha sido utilizado, não sendo necessário explicitar tais condicionais completas para a realização da estimação por meio do mesmo, no Apêndice C são mostrados resultados referentes às derivações dessas condicionais completas e um esquema do processo de estimação. Como a maior parte das distribuições condicionais completas possuem formas conhecidas, com exceção de λ_j, um método de amostrador de Gibbs pode ser utilizado a fim de obter amostras da distribuição *a posteriori* para a maior parte dos parâmetros do modelo C-DINA, enquanto para os λ_j, pode ser utilizado o *slice sampling* (NEAL, 2003).

Para a simulação e a aplicação, são retiradas 50.000 amostras para o modelo C-DINA, utilizando o esquema de MCMC ilustrado no Apêndice C, descartando as primeiras 10.000 iterações. A convergência do algoritmo MCMC é checada por meio da avaliação gráfica e do critério de Gelman-Rubin (GELMAN; RUBIN *et al.*, 1992), o qual é próximo a um para todos os parâmetros, indicando que as cadeias convergiram.

3.4 Estudos de Simulação

Nessa seção, são discutidas algumas questões sobre uma análise de sensibilidade para a priori do hiperparâmetro τ_{μ_0} no modelo proposto. Além disso, um estudo de simulação para examinar a performance do esquema MCMC proposto para a estimação dos parâmetros do modelo é realizado.

Um estudo de sensibilidade

A escolha do hiperparâmetro τ_{μ_0} é importante em relação à priori que foi atribuída ao parâmetro de item $\mu_{j0},\ j = 1,\ldots,J$. Nessa subseção, conduzimos um estudo de simulação a fim de comparar os resultados se um valor for fixado para τ_{μ_0} ou se uma hiperpriori for considerada. Mais especificamente, consideramos dois cenários para τ_{μ_0}, i.e.,

- Priori 1 (P1): $\mu_{j0} \sim \text{Normal}(0, \tau_{\mu_0}^{-1})$ com $\tau_{\mu_0} = 0.01$.
- Priori 2 (P2): $\mu_{j0} \sim \text{Normal}(0, \tau_{\mu_0}^{-1})$ e $\tau_{\mu_0} \sim \text{Gama}(0.01, 0.01)$.

Considerando nosso modelo hierárquico proposto, dado em (3.14), assumimos o cenário onde $N = 250$, $J = 30$ e $K = 5$ a fim de realizar esse estudo inicial. Os parâmetros de item são os mesmos para cada $j = 1,\ldots,J$, com $\mu_{j0} = 0$, $\lambda_j = 1$ (i.e., $\mu_{j1} = 1$), $\tau_{j0} = \sigma_{j0} = 1$ e $\tau_{j1} = 1/(\sigma_{j1}^2) = 1/0.230^2$. A Matriz $\boldsymbol{Q}$ utilizada em nosso estudo pode ser vista em 8, sendo

Tabela 7 – Média dos critérios de comparação e proporção de escolhas para a Priori P2 do parâmetro μ_{j0} ao longo das 100 réplicas

	Priori	EAIC	EBIC	DIC	WAIC	Tempo (seg.)
Média	$P1$	20039.15	21454.77	19531.78	19484.03	7384.88
	$P2$	20002.98	21418.61	19526.97	19477.46	9376.59
Porcentagem	$P2 < P1$	87%	87%	86%	97%	

similar à especificada em (MINCHEN; TORRE; LIU, 2017). Por fim, presumimos que $\boldsymbol{\pi} = (1/32, \ldots, 1/32)'$, isto é, a probabilidade de pertencer a cada um dos perfis de atributo é a mesma.

São consideradas 100 réplicas, geradas com os critérios definidos acima, mas com diferentes sementes aleatórias. Consideramos $a_\lambda = b_\lambda = 0.01$, $c_0 = d_0 = 0.01$ e $c_1 = d_1 = 0.01$ para as prioris no esquema em (3.14) e então o modelo C-DINA é ajustado sob a abordagem Bayesiana proposta, considerando as propostas de P1 e P2. Na Tabela 7 é apresentada a comparação dos resultados entre ambos os casos, utilizando o critério de informação de desvio (DIC) (SPIEGELHALTER *et al.*, 2002), o critério de informação de Akaike esperado (EAIC) e o Critério de informação Bayesiana esperado (EBIC) (GELMAN *et al.*, 2014), bem como o critério de informação de Watanabe (WAIC) (WATANABE, 2010a). Menores valores para esses critérios trazem indícios de que aquela configuração tem melhores resultados.

Na Tabela 7, são apresentados os valores da média de cada critério entre as 100 réplicas, para os cenários P1 e P2. Além disso, também reportamos o tempo computacional (em segundos) para cada cenário e a frequência em que o valor de cada critério é menor para o cenário P2 em relação ao cenário P1. Considerando os diferentes critérios apresentados na Tabela 7, o melhor ajuste é obtido quando P2 é considerado. A Tabela 7 mostra que a priori P2 foi escolhida mais frequentemente ao longo das réplicas considerando todos os critérios avaliados. Esses resultados são favoráveis ao uso de uma hiperpriori para o parâmetro de precisão τ_{μ_0} do parâmetro de item μ_{j0}. Assim sendo, iremos considerar essa especificação no restante do trabalho.

Estudo de Recuperação

Nessa subseção, um estudo de simulação é realizado, a fim de avaliar a recuperação de parâmetros do modelo C-DINA, utilizando nossa abordagem Bayesiana proposta em 3.3. Na simulação são considerados diferentes cenários para o número de itens (J), o número de respondentes (N) e a discriminação dos itens. O desenho dos cenários se inspira em Minchen, Torre e Liu (2017), os quais utilizaram a abordagem frequentista para inferir sobre o modelo C-DINA. Entretanto, para investigar os benefícios de nossa abordagem Bayesiana, consideramos cenários com menor número de respondentes que os utilizados naquele estudo.

Em nosso estudo, consideramos dois casos para os respondentes, com $N = 250$ e $N = 500$. Além disso, assumimos $J = 15$ ou $J = 30$ itens. A Tabela 8 mostra a Matriz $\boldsymbol{Q}$ utilizada na simulação para $J = 30$. Para os cenários com $J = 15$, são utilizados os itens de número 1, 2, 3, 4,

5, 11, 14, 15, 18, 20, 21, 23, 26, 27 e 30. Em ambos os casos há 5 atributos avaliados no teste, com os itens entre 1 e 10 exigindo um atributo, os itens entre 11 e 20 requisitando dois atributos e os itens entre 21 e 30 requerindo três atributos.

Também são utilizados três tipos de discriminação para os itens no teste, com discriminação baixa, média ou alta. Uma definição mais rigorosa da discriminação é baseada em medidas de similaridade ou dissimilaridade entre os grupos $\eta = 1$ e $\eta = 0$. Os parâmetros para tais grupos foram definidos em Minchen, Torre e Liu (2017), utilizando medidas gráficas e a divergência de Kullback-Leibler para avaliar a discirimanção.

Nesse trabalho, consideramos os mesmos valores de parâmetros propostos por Minchen, Torre e Liu (2017) para cada grupo, porém, além de considerarmos a divergência de Kullback–Leibler (KL), $D_{KL}(P,Q) = \int_{\mathscr{X}} f_P(x) \log \frac{f_P(x)}{f_Q(x)} dx$; calculamos outras distâncias pertinentes, primeiramente a distância de Kolmogorov-Smirnov (KS): $D_{KS}(P,Q) = \sup_{x \in \mathscr{X}} \mid F_P(x) - F_Q(x) \mid$ e a distância de Hellinger (H): $d_H^2(P,Q) = \int_{\mathscr{X}} \sqrt{f_P(x)} \sqrt{f_Q(x)} dx$, em que P e Q representam duas distribuições, com $f_P(\cdot)$ e $f_Q(\cdot)$ sendo as fdps e $F_P(\cdot)$ e $F_Q(\cdot)$ as fdas, $|\cdot|$ denota o valor absoluto e $\mathscr{X}$ é o domínio de x.

Maiores valores de KL, KS e H indicam maior discriminação entre duas distribuições comparadas. Os valores de KL estão no conjunto $[0,\infty)$, enquanto os de KS e H estão em $[0,1]$, possuindo simetria e sendo mais facilmente interpretáveis. De acordo com a distribuição de Hellinger, definimos um parâmetro de *discriminação* dos itens por $m_j = d_H^2(\eta = 0, \eta = 1) = \int_0^\infty \sqrt{f_{j0}(y_{ij})} \sqrt{f_{j1}(y_{ij})} dy_{ij}$, onde $m_j \in [0,1]$. Uma partição ruim dos grupos pode ocorrer com m_j abaixo de 0.25, indicando que o j-ésimo item possui discriminação baixa; se m_j tem um valor em torno de 0.5 o item tem discriminação média; se m_j tem um valor por volta de 0.75 ele tem alta discriminação.

Três configurações de parâmetros são consideradas para o j-ésimo item, i.e., $(\mu_{j0}, \mu_{j1}, \sigma_{j0}^2, \sigma_{j1}^2)' = (\mu_0, \mu_1, \sigma_0^2, \sigma_1^2)'$, com $(0, 0.5, 1, 0.397^2)'$, $(0, 1, 1, 0.230^2)'$ e $(0, 2, 1, 0.082^2)'$ como escolhas para $(\mu_0, \mu_1, \sigma_0^2, \sigma_1^2)'$. Os valores correspondentes de KL, KS e m_j para essas três escolhas são $KL = 2.54, KS = 0.40, m_j = 0.21$; $KL = 16.93, KS = 0.68, m_j = 0.48$; e $KL = 368.80, KS = 0.96, m_j = 0.85$, respectivamente, o que representa discriminação baixa, média e alta dos itens. Note que, embora os parâmetros τ_{j0} e τ_{j1} sejam utilizados na computação, a fim de facilitar a comparação de parâmetros com trabalhos prévios, será reportada a recuperação para σ_{j0} e σ_{j1} nos resultados sumarizados.

Portanto, há 12 diferentes cenários em nosso estudo, os quais são utilizados para examinar a abordagem Bayesiana proposta. A fim de gerar os dados, utilizamos os mesmos valores de $\boldsymbol{\pi}$ referidos anteriormente. Primeiramente, simulamos as classes de atributo de cada indivíduo por meio de uma distribuição multinomial de tamanho N e probabilidade $1/32$ de pertença à cada uma das 32 classes. Então, considerando os perfis gerados para cada indivíduo e a Matriz $\boldsymbol{Q}$ na Tabela 8, podemos obter o valor de η_{ij} para cada indivíduo e item. Por fim, as respostas y_{ij} são geradas a partir da distribuição Log-Normal, i.e., $LN(\mu_{j\eta}, \tau_{j\eta}^{-1})$. Os valores gerados de y_{ij} são

Tabela 8 – Matriz $\boldsymbol{Q}$ para o estudo de simulação

	Atributo						Atributo				
Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0	16	0	1	0	1	0
2	0	1	0	0	0	17	0	1	0	0	1
3	0	0	1	0	0	18	0	0	1	1	0
4	0	0	0	1	0	19	0	0	1	0	1
5	0	0	0	0	1	20	0	0	0	1	1
6	1	0	0	0	0	21	1	1	1	0	0
7	0	1	0	0	0	22	1	1	0	1	0
8	0	0	1	0	0	23	1	1	0	0	1
9	0	0	0	1	0	24	1	0	1	1	0
10	0	0	0	0	1	25	1	0	1	0	1
11	1	1	0	0	0	26	1	0	0	1	1
12	1	0	1	0	0	27	0	1	1	1	0
13	1	0	0	1	0	28	0	1	1	0	1
14	1	0	0	0	1	29	0	1	0	1	1
15	0	1	1	0	0	30	0	0	1	1	1

considerados como as respostas observadas, enquanto todos os parâmetros são tratados como desconhecidos em nosso estudo.

A fim de avaliar a recuperação de parâmetros, são realizadas 100 réplicas para cada cenário, levando à 12×100 conjunto de dados, considerando todos os cenários. Para analisar os dados simulados, seguimos o que foi discutido anteriormente, atribuindo uma hiperpriori para o parâmetro de precisão τ_{μ_0} em $\mu_{j0} \sim \text{Normal}(0, \tau_{\mu_0}^{-1})$, i.e., $\tau_{\mu_0} \sim \text{Gama}(0.01, 0.01)$, com as demais prioris também seguindo o que foi definido na subseção anterior.

Com a finalidade de avaliar a performance da recuperação dos parâmetros de item, é computado o valor da raiz do erro quadrático médio (RMSE) e também calculamos o viés absoluto, dado por Viés $= \widehat{\overline{\vartheta}}_l - \vartheta_l$, com $\widehat{\overline{\vartheta}}_l = \sum_{r=1}^{R} \widehat{\vartheta}_{lr}/R$. Como temos $(\mu_{j0}, \mu_{j1}, \sigma_{j0}^2, \sigma_{j1}^2)' = (\mu_0, \mu_1, \sigma_0^2, \sigma_1^2)'$ para $j = 1, \cdots, J$, a fim de facilitar a sumarização, primeiro computamos os valores de RMSE e do viés para cada item e então reportamos suas médias de acordo com o grupo dos parâmetros de item μ_0, μ_1, σ_0^2 e σ_1^2, respectivamente. Os resultados para μ_0, μ_1, σ_0^2 e σ_1^2 são mostrados da Tabela 9 até a Tabela 12.

Pelos resultados da Tabela 9 à Tabela 12, nota-se que o incremento no número de itens e no número de respondentes reduz o viés e o RMSE para os parâmetros de item. Além disso, a discriminação dos itens tem grande influência em ambas as medidas. Para os parâmetros μ_1 e σ_1^2 (parâmetros de item correspondentes ao grupo $\eta = 1$) é interessante notar que o grupo de itens que exigem mais atributos tem valores de RMSE e viés maiores que aqueles que exigem menos atributos, principalmente quando o tamanho da amostra é pequeno e a discriminação dos itens é baixa. Isso sugere que estimar os parâmetros de item para $\eta = 1$ de maneira precisa é mais

Tabela 9 – Performance para a recuperação do parâmetro μ_0 ao longo das 100 réplicas

			Grupos de itens pelo número de atributos requeridos					
			Um		Dois		Três	
J	Discriminação	*N*	RMSE	Viés	RMSE	Viés	RMSE	Viés
15	Baixa	250	0.052	0.03	0.045	0.02	0.040	0.03
		500	0.030	0.01	0.019	0.01	0.020	0.01
	Média	250	0.018	0.00	0.018	0.00	0.016	0.00
		500	0.016	0.01	0.014	0.00	0.015	0.00
	Alta	250	0.017	0.00	0.019	0.00	0.016	0.00
		500	0.016	0.00	0.014	0.00	0.014	0.00
30	Baixa	250	0.020	0.01	0.018	0.01	0.019	0.01
		500	0.012	0.01	0.010	0.00	0.009	0.00
	Média	250	0.011	0.00	0.010	0.00	0.011	0.00
		500	0.009	0.00	0.009	0.00	0.008	0.00
	Alta	250	0.011	0.00	0.010	0.00	0.011	0.00
		500	0.009	0.00	0.009	0.00	0.008	0.00

Tabela 10 – Performance para a recuperação do parâmetro μ_1 ao longo das 100 réplicas

			Grupos de itens pelo número de atributos requeridos					
			Um		Dois		Três	
J	Discriminação	*N*	RMSE	Viés	RMSE	Viés	RMSE	Viés
15	Baixa	250	0.037	-0.02	0.097	-0.07	0.217	-0.19
		500	0.017	0.00	0.034	-0.01	0.068	-0.04
	Média	250	0.010	0.00	0.014	0.00	0.021	0.00
		500	0.007	0.00	0.011	0.00	0.015	0.00
	Alta	250	0.004	0.00	0.004	0.00	0.007	0.00
		500	0.002	0.00	0.004	0.00	0.005	0.00
30	Baixa	250	0.016	-0.01	0.028	-0.01	0.080	-0.06
		500	0.009	0.00	0.014	-0.01	0.021	-0.01
	Média	250	0.007	0.00	0.009	0.00	0.014	0.00
		500	0.005	0.00	0.007	0.00	0.009	0.00
	Alta	250	0.002	0.00	0.003	0.00	0.005	0.00
		500	0.002	0.00	0.002	0.00	0.003	0.00

Tabela 11 – Performance para a recuperação do parâmetro σ_0 ao longo das 100 réplicas

			Grupos de itens pelo número de atributos requeridos					
			Um		Dois		Três	
J	Discriminação	N	RMSE	Viés	RMSE	Viés	RMSE	Viés
15	Baixa	250	0.043	0.01	0.027	0.00	0.023	0.00
		500	0.027	0.00	0.018	0.00	0.016	0.00
	Média	250	0.034	0.01	0.026	0.00	0.024	0.00
		500	0.021	0.00	0.016	0.00	0.015	0.00
	Alta	250	0.033	0.01	0.025	0.00	0.024	0.00
		500	0.020	0.00	0.016	0.00	0.015	0.00
30	Baixa	250	0.023	0.00	0.017	0.00	0.020	0.01
		500	0.016	0.00	0.013	0.00	0.014	0.00
	Média	250	0.022	0.00	0.017	0.00	0.020	0.01
		500	0.016	0.00	0.013	0.01	0.014	0.01
	Alta	250	0.022	0.00	0.017	0.00	0.020	0.01
		500	0.016	0.00	0.013	0.01	0.014	0.01

Tabela 12 – Performance para a recuperação do parâmetro σ_1 ao longo das 100 réplicas

			Grupos de itens pelo número de atributos requeridos					
			Um		Dois		Três	
J	Discriminação	N	RMSE	Viés	RMSE	Viés	RMSE	Viés
15	Baixa	250	0.038	0.02	0.071	0.05	0.699	0.19
		500	0.020	0.00	0.030	0.01	0.055	0.03
	Média	250	0.009	0.00	0.012	0.00	0.021	0.01
		500	0.005	0.00	0.008	0.00	0.010	0.00
	Alta	250	0.003	0.00	0.004	0.00	0.009	0.01
		500	0.002	0.00	0.003	0.00	0.004	0.00
30	Baixa	250	0.012	0.00	0.023	0.01	0.052	0.04
		500	0.008	0.00	0.012	0.00	0.017	0.01
	Média	250	0.005	0.00	0.008	0.00	0.014	0.01
		500	0.004	0.00	0.005	0.00	0.007	0.00
	Alta	250	0.002	0.00	0.004	0.00	0.008	0.01
		500	0.001	0.00	0.002	0.00	0.004	0.00

difícil quando são exigidos mais atributos.

Da Tabela 9 à Tabela 12 notamos que os maiores valores de RMSE são encontrados quando $N = 250$, $J = 15$ e a discriminação dos itens é baixa. Isso confirma que é interessante haver testes com maior número de itens e respondentes a fim de utilizar a metodologia do C-DINA e, principalmente, itens com boa discriminação devem ser elaborados.

A fim de avaliar os parâmetros dos indivíduos, é necessário o uso de outra medida sem ser o RMSE e o viés, uma vez que os vetores de atributos dos respondentes $\boldsymbol{\alpha}_i = (\alpha_{i1}, \cdots, \alpha_{ik})'$ são compostos por quantidades binárias. Para esses parâmetros, nós utilizamos a acurácia simples (SA) como a porcentagem de binários α_{ik}s que foram corretamente estimados, isto é, $SA = 100 \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}(\widetilde{\alpha}_{ik} = \alpha_{ik})/NK$, onde $\widetilde{\alpha}_{ik}$ é a moda *a posteriori* das estimativas obtidas nas amostras MCMC e α_{ik} é o valor verdadeiro. A medida SA tem um valor entre 0 e 100, onde $SA = 0$ indica que todos os parâmetros dos indivíduos foram estimados incorretamente e $SA = 100$ mostra que todos foram estimados corretamente. Uma alternativa é considerar a acurácia vetorial (VA), a qual quantifica a proporção de vetores $\boldsymbol{\alpha}_i$s que foram estimados corretamente, i.e., $VA = 100 \sum_{i=1}^{N} \mathbb{1}(\widetilde{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i)/N$ com $\widetilde{\boldsymbol{\alpha}}_i = (\widetilde{\alpha}_{i1}, \cdots, \widetilde{\alpha}_{ik})'$. A medida VA também tem valor entre 0 e 100 mas, agora, é avaliado se todos os valores α_{ik}, $k = 1, \ldots, K$ são estimados simultaneamente corretamente para o i-ésimo indivíduo. Assim, $VA = 0$ indica que todos os $\boldsymbol{\alpha}_i$'s são estimados incorretamente e $VA = 100$ significa que todos os $\boldsymbol{\alpha}_i$'s são estimados corretamente. Tais valores são obtidos para cada uma das réplicas, com a média dos valores considerando as 100 réplicas sendo consideradas ao reportar os resultados da simulação.

Na Tabela 13 são apresentados os resultados de SA e VA para os parâmetros dos indivíduos $\boldsymbol{\alpha}_i$. Os maiores valores, tanto para SA quanto para VA, são encontrados quando o número de itens e de participantes são maiores, bem como quando a Discriminação aumenta. Em geral, os valores de SA são melhores que os de VA. Isso ocorre devido ao fato de que, ao computarmos o VA, é necessário que todos os parâmetros dos indivíduos sejam corretamente estimados para o i-ésimo indivíduo. Da Tabela 13, podemos ver que a maior acurácia é encontrada quando $J = 30$, há discriminação alta e $N = 500$, enquanto $J = 15$, discriminação baixa e $N = 250$ traz os piores resultados de VA. Entretanto, mesmo para $J = 15$ e $N = 250$, quando os itens possuem boa discriminação, a recuperação melhora, com os valores de VA praticamente triplicando para discriminação alta em relação à baixa. Isso sugere que a discriminação dos itens tem um impacto crucial, também para os parâmetros dos indivíduos, evidenciando mais uma vez a importância do desenvolvimento dos itens no desenho dos testes. Além disso, aumentar o número de itens para $J = 30$ traz um ganho considerável para os valores de VA, com resultados consideravelmente bons até mesmo para discriminação baixa no que diz respeito ao SA. Isso nos mostra que a fim de ter bons resultados para os parâmetros dos indivíduos, também é interessante desenvolver testes que não tenham tão poucos itens. Por outro lado, pode-se notar que o número de respondentes não tem tanto impacto na performance da recuperação de parâmetros dos indivíduos. Ao compararmos os resultados de $N = 500$ na Tabela 13 com os resultados correspondentes em

(MINCHEN; TORRE; LIU, 2017), encontramos resultados um pouco maiores para SA e VA que aqueles encontrados na abordagem frequentista.

Para concluir, nossa estimação Bayesiana do modelo C-DINA fornece uma recuperação de parâmetros adequada, tanto para os parâmetros dos itens quanto para os dos indivíduos, em nossa simulação. A seguir, essa abordagem será utilizada em uma aplicação em dados de percepção de risco.

Tabela 13 – Performance para a recuperação do parâmetro $\boldsymbol{\alpha}$ ao longo das 100 réplicas

J	Discriminação	N	SA	VA
15	Baixa	250	76.48	33.53
		500	77.91	37.35
	Média	250	91.94	72.34
		500	92.21	73.40
	Alta	250	99.39	97.47
		500	99.42	97.61
30	Baixa	250	88.67	63.43
		500	89.46	66.16
	Média	250	98.52	94.43
		500	98.51	94.52
	Alta	250	99.99	99.96
		500	99.99	99.97

3.5 Aplicação: Dados sobre percepção de risco

Nessa seção será apresentada uma análise de dados reais, sobre percepção de risco (CARLSTROM; WOODWARD; PALMER, 2000). O conjunto de dados consiste de 611 participantes, os quais respondem a um questionário no qual é pedido que eles avaliem o risco percebido para diferentes atividades. Os respondentes são requisitados à atribuir um valor numérico, entre 0 e 100 (100 para o maior risco), para mensurar o risco relacionado à 22 atividades financeiras e relacionadas à saúde. Em nossa análise, apenas 14 dessas atividades foram consideradas (excluindo aquelas que não são comuns para homens e mulheres), resultando nos seguintes itens: 1) APPL: utilizar eletrodomésticos; 2) BIKE: andar de bicicleta por uma milha por dia em uma área urbana; 3) BRCA: testar positivo para um gene que predispõe ao câncer de mama; 4) CAR: dirigir um automóvel por 10 milhas por dia em uma área urbana; 5) DIAB: testar positivo para um gene que predispõe a ter diabetes; 6) DOC: trabalhar como um médico familiar em uma área rural; 7) FLU: tomar a vacina de gripe anualmente; 8) GREY: testar positivo para um gene que predispõe ao branqueamento prematuro do cabelo; 9) HEART: testar positivo para um gene que predispõe doença cardíaca; 10) NUC: viver próximo a uma estação de energia nuclear; 11) PLANE: voar em aviões comerciais todo mês; 12) POOL: nadar em uma piscina pública todas as semanas; 13) SWAT: trabalhar como membro da polícia; 14) XRAY: receber

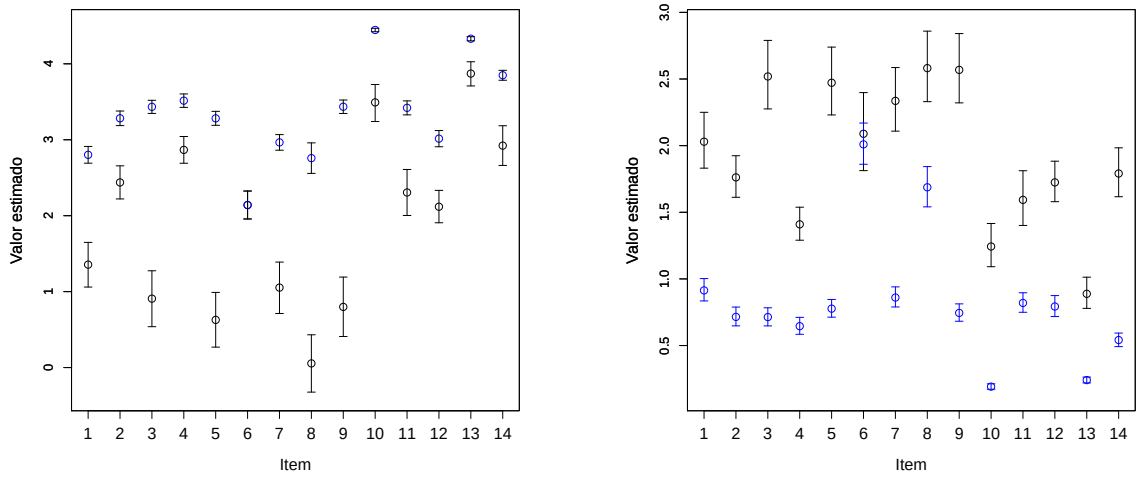
um diagnóstico por meio de exame de raio x a cada 6 meses. Para proceder com a análise, os indivíduos com respostas faltantes para alguma atividade foram removidos, resultando em 481 participantes analisados. Os valores de percepção de risco iguais à zero foram transformados para 0.1, permitindo o uso da distribuição Log Normal.

Para propor uma Matriz $\boldsymbol{Q}$ associada aos itens do teste, precisamos definir diferentes dimensões ou atributos para os riscos avaliados. Ao revisar a literatura, a classificação de riscos em atributos internos e externos é comum, em diferentes áreas do conhecimento (DUFFY, 2003; BABIAK; WOLFE, 2009; MENGUC; AUH; OZANNE, 2010). Com essa ideia, avaliando o conteúdo de cada um dos itens do questionário, a Matriz $\boldsymbol{Q}$ para os dados sobre percepção de risco foi construída, sendo apresentada nas Colunas 2 e 3 da Tabela 14. Os riscos foram divididos como relacionados à atributos *externos* (ou não pessoais) (α_1), nos quais os riscos dependem do ambiente circundando a pessoa e *internos* (ou pessoais) (α_2), nos quais o risco depende da pessoa. Por exemplo, HEART foi definido como avaliando apenas um atributo interno, uma vez que diz respeito unicamente à pessoa; NUC foi considerado um item relacionado a um atributo externo, uma vez que retrata uma situação cujo risco é trazido por viver perto de uma estação de energia nuclear, a qual traz risco independente de características pessoais.

Tabela 14 – Matriz **Q** do conjunto de dados de percepção de risco, estimativas pelas médias à posteriori para os parâmetros dos itens, e valores de medidas de distância e discriminação dos itens.

Item	α_1	α_2	$\hat{\mu}_0$	$\hat{\sigma}_0$	$\hat{\mu}_1$	$\hat{\sigma}_1$	KL	KS	m_j	λ_j
1. APPL	0	1	1.359	2.030	2.802	0.913	2.42	0.45	0.21	1.443
2. BIKE	1	1	2.436	1.762	3.283	0.715	2.34	0.38	0.20	0.846
3. BRCA	0	1	0.905	2.520	3.434	0.714	10.75	0.65	**0.41**	**2.529**
4. CAR	1	1	2.868	1.410	3.515	0.646	1.60	0.35	0.16	0.645
5. DIAB	0	1	0.631	2.472	3.282	0.777	9.22	0.66	**0.41**	**2.605**
6. DOC	1	0	2.134	2.090	2.144	2.007	0.00	0.01	0.00	0.009
7. FLU	0	1	1.046	2.336	2.966	0.860	4.68	0.54	0.30	1.920
8. GREY	0	1	0.056	2.583	2.758	1.688	1.53	0.49	0.20	**2.702**
9. HEART	0	1	0.807	2.569	3.435	0.745	10.43	0.66	**0.41**	**2.628**
10. NUC	1	0	3.507	1.243	4.446	0.192	30.65	0.65	**0.51**	0.939
11. PLANE	1	0	2.308	1.593	3.421	0.819	1.65	0.42	0.18	1.113
12. POOL	1	1	2.116	1.724	3.015	0.793	1.73	0.37	0.17	0.899
13. SWAT	1	0	3.868	0.888	4.331	0.241	6.85	0.48	0.33	0.463
14. XRAY	0	1	2.923	1.790	3.850	0.541	5.24	0.46	0.29	0.927

Uma vez definida a Matriz **Q** podemos aplicar o modelo C-DINA ao conjunto de dados, o que foi feito seguindo nossa abordagem Bayesiana apresentada na Seção 3.3, com as mesmas prioris empregadas no estudo de recuperação. Da Coluna 4 à Coluna 7 da Tabela 14 são mostrados os resultados para as estimativas com base na média à posteriori dos parâmetros de item dos grupos latentes $\eta = 0$ e $\eta = 1$, respectivamente. Também são mostradas as medidas de discriminação KL, KS e m_j, a fim de avaliar como cada item diferencia os dois grupos latentes, os quais podem ser vistos da Coluna 8 à Coluna 10. Por fim, reportamos a diferença entre os parâmetros de locação dos dois grupos, i.e., $\lambda_j = \mu_{j1} - \mu_{j0}$, na Coluna 11. Podemos ver que BRCA (item 3), DIAB (item 5) e HEART (item 9) são os itens com maiores valores de m_j's e λ_j's simultanemante. De acordo com a partição proposta por valores de m_j, esses itens possuem



(a) Parâmetros μ

(b) Parâmetros σ

Figura 8 – Intervalos HPD dos parâmetros de item. Linhas pretas representam grupo $\eta = 0$ e azuis grupo $\eta = 1$.

média discriminação. Para o item NUC (item 10) vemos um valor relativamente baixo para λ_j, mas, mesmo assim o valor de m_j mostra boa discriminação para o item. Em oposição, o item GREY (item 8) possui o maior valor de λ_j entre todos os itens, mas não possui um alto valor de m_j. Entre todos os itens, é interessante notar que DOC (Item 6) possui $m_j = 0.00$ e $\lambda_j = 0.01$, indicando que, de acordo com nosso modelo C-DINA proposto, os respondentes que estão no grupo $\eta = 1$, não avaliam "trabalhar como um médico de família em uma área rural" de maneira diferente daqueles que estão no grupo $\eta = 1$, os quais não veem condições externas como apresentando risco.

Além disso, é interessante avaliar as estimativas intervalares, as quais podem ser vistas, por meio dos intervalos de máxima densidade à posteriori (HPD) ao nível de 95% (BOX; TIAO, 2011; CHEN; SHAO, 1999) para todos os parâmetros de item μ_{j0}, μ_{j1}, σ_{j0} e σ_{j1} na Figura 8. Nota-se que não há intersecção entre os intervalos dos parâmetros do grupo $\eta = 0$ e $\eta = 1$, com exceção do item 6, o qual não distingue entre o comportamento dos respondentes classificados em cada grupo.

Na Tabela 15 são sumarizados os resultados para as estimativas pelas médias *a posteriori* e os intervalos HPD 95%, para a probabilidade de pertencer à cada perfil c (π_c). É claro que o perfil com a maior probabilidade estimada é o $(1,1)$, i.e., o perfil no qual os indivíduos veem ambas as condições internas e externas avaliadas no estudo como tendo riscos consideráveis. O perfil com a segunda maior probabilidade estimada é $(1,0)$, indicando que há mais respondentes que veem a exposição externa do que a interna como representando riscos consideráveis.

Assim sendo, a aplicação do modelo C-DINA com duas dimensões pode fornecer um diagnóstico mais refinado sobre a percepção de risco, o que, por sua vez, pode trazer uma melhor

Tabela 15 – Classificação: probabilidade de pertença à cada um dos perfis de atributo.

c perfis de atributo		Dimensões		$\widehat{\pi}_c$	
		α_1	α_2	Média	HPD (95%)
1	(não vêem riscos consideráveis)	0	0	0.12	$(0.089, 0.157)$
2	(vêem condições externas como mais arriscadas)	1	0	0.27	$(0.230, 0.319)$
3	(vêem condições internas como mais arriscadas)	0	1	0.14	$(0.103, 0.172)$
4	(vêem ambas as condições como arriscadas)	1	1	0.47	$(0.419, 0.513)$

orientação para a classificação de pessoas em diferentes grupos.

3.6 Discussão

Nesse capítulo, uma abordagem Bayesiana para a estimação do modelo C-DINA foi proposta, por meio de um método MCMC. Foram dados detalhes sobre a interpretação dos parâmetros e a lógica da construção do modelo C-DINA, bem como foram propostas medidas que facilitem na avaliação da discriminação dos itens nos grupos latentes.

O estudo de simulação mostrou que a abordagem Bayesiana proposta traz bons resultados na recuperação dos parâmetros. É bastante interessante notar que o número de respondentes e o número de itens são pertinentes para a correta recuperação de parâmetros, sendo que o aumento de ambos auxiliam na estimação. Ainda é notável que algo ainda mais importante que o número de itens, é que os itens tenham boa discriminção entre os grupos η, principalmente para a estimação dos parâmetros dos indivíduos. Quando itens com alta discriminação são considerados, os resultados são bastante bons, mesmo para valores menores de respondentes e de itens.

A aplicação mostra que há potencial uso de MDCs para dados que vão além do contexto educacional, podendo tal utilização ser expandida para outros tipos de questionário, bem como os de percepção de riscos.

Para futuros trabalhos, a abordagem Bayesiana aqui desenvolvida pode ser estendida para outras estruturas de modelos, gerando versões contínuas do modelo DINO (*deterministic input noisy output "or"gate*), por exemplo. Também há diversas possibilidades no avanço do nosso modelo proposto. Por exemplo, pode ser considerado o cenário no qual haja dependência entre itens no C-DINA, similar ao que (BRADLOW; WAINER; WANG, 1999) fez para modelos *testlet* de TRI e (ZHAN *et al.*, 2015) fez para MDCs. Ainda há a possibilidade de levar em conta a velocidade de resposta aos itens, como feito por (OSHIMA, 1994) para a TRI, o que pode ajudar em melhor estimação para os parâmetros dos indivíduos. Considerando dados relacionados ao tempo para modelos C-DINA, ainda podem ser consideradas questões relacionadas à dados censurados. Por fim, de maneira similar ao que (FOX; ENTINK; LINDEN, 2007) propõe para modelos TRI e 2018 propuseram para um MDC, nossa abordagem pode ser interessante para

estender a discussão no tópico de análises de respostas dicotômicas e tempos de resposta, simultaneamente, com o desenvolvimento de um modelo conjunto DINA e C-DINA.

CAPÍTULO

4

MODELO DINA DE RESPOSTA LIMITADA

Nos últimos anos, os MDCs vem ganhando espaço considerável na literatura. Diferentes métodos já foram considerados, levando em conta a diversidade existente dentre as possíveis maneiras de se mensurar respostas. Inicialmente desenvolvidos para respostas dicotômicas, hoje há modelos que avaliam respostas politômicas, ordenadas e também contínuas. No trabalho apresentado nesse capítulo, há a contribuição com a literatura por meio do desenvolvimento de um MDC para respostas contínuas limitadas, sob abordagem Bayesiana, preenchendo um espaço ainda existente até então. Respostas contínuas limitadas aparecem principalmente quando o interesse está em probabilidades, proporções ou taxas. O trabalho apresentado nesse capítulo propõe um inédito modelo DINA para respostas contínuas limitadas (B-DINA), sob abordagem Bayesiana. Um estudo de simulação é conduzido, a fim de avaliar a recuperação de parâmetros, mostrando bons resultados para tamanhos amostrais adequados. Uma aplicação para dados sócio demográficos de municípios na região Sudeste do Brasil é feita, elucidando o grande potencial de tais metodologias e mostrando que sua utilização pode ir além de aplicações psicométricas ou educacionais.

O conteúdo desse capítulo será submetido a um periódico especializado.

4.1 Introdução

Nos capítulos anteriores já exploramos os modelos DINA dicotômico e também o modelo C-DINA, o qual considera variáveis respostas contínuas. Entretanto, não há na literatura nenhum modelo específico para respostas limitadas. O desenvolvimento de um modelo considerando esse tipo de resposta é um avanço interessante para a literatura sobre MDC.

Diversas aplicações, em diferentes áreas do conhecimento, também poderiam se beneficiar de um modelo para respostas contínuas limitadas, em questionários realizados com respostas dentro de um intervalo pré estipulado ou até mesmo com a análise de outros tipos de dados, bem como o de indicadores sociais, nos quais o indivíduo por ser tomado, por exemplo, como um determinado município e os itens serem compostos por diferentes indicadores que avaliem determinadas dimensões.

O principal propósito desse capítulo é propor um primeiro modelo DINA para respostas limitadas (B-DINA), por meio de uma abordagem Bayesiana, considerando a distribuição Beta com a parametrização proposta por Ferrari e Cribari-Neto (2004a), garantindo assim a interpretabilidade do modelo resultante.

O capítulo está organizado da seguinte maneira: Na Seção 4.2 são apresentados detalhes do modelo B-DINA. Na Seção 4.3 mostramos a estimação Bayesiana do modelo proposto, com detalhes sobre as prioris e o método MCMC utilizado. Na Seção 4.4 é apresentado um estudo de simulação, avaliando a recuperação de parâmetros. Na Seção 4.5 dados reais relacionados à indicadores sociais são analisados e, por fim, na Seção 4.7 há uma breve discussão com considerações finais sobre o capítulo e sugestão para trabalhos futuros.

4.2 Modelos DINA Limitados

Nessa Seção, o modelo DINA limitado (B-DINA) é proposto. Grande parte das quantidades presentes no B-DINA são compartilhadas com o C-DINA, já tendo sido apresentadas na Seção 3.2, com a alteração ocorrendo em relação à variável resposta. Um modelo limitado pode ser proposto em qualquer intervalo genérico $[a,b]$, porém, sem perda de generalidade para propostas alternativas, focamos no caso $(0,1)$, com uma distribuição conveniente. Exemplos desse tipo de variável podem ser encontrados em diversas pesquisas em diferentes áreas do conhecimento, bem como em Anand e Sen (1994), Carlstrom, Woodward e Palmer (2000), Bijur, Silver e Gallagher (2001), Walraven *et al.* (2011), Noel e Dauvier (2007).

Especificação do modelo B-DINA

Para o modelo B-DINA, assim como para os modelos anteriormente apresentados, há $i = 1,\ldots,N$ respondentes à um questionário com $j = 1,\ldots,J$ itens, o qual avalia $k = 1,\ldots,K$ atributos (ou dimensões/habilidades).

A maior parte das quantidades, bem como q_{jk}, α_{ik} e η_{ij} são compartilhadas com o modelo C-DINA, com a grande diferença sendo no que diz respeito à Y_{ij}, que tem outra característica para modelo B-DINA. Tais quantidades também estão presentes no modelo DINA usual e uma explicação mais detalhada das quantidades comuns a esses modelos não será feita nessa seção, para evitar que esse processo se torne repetitivo.

Porém, é importante definir em detalhes a variável resposta, para a qual temos que:

- $Y_{ij} \in (0,1)$ é uma variável aleatória limitada, representando as respostas observadas do i-ésimo respondente para o j-ésimo item, no intervalo unitário. O vetor de respostas do indivíduo i é representado por $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})'$, sendo tais quantidades observadas.

Novamente, mesmo que a resposta Y_{ij} não seja dicotômica, a variável latente α_{ij} continua sendo, indicando se o indivíduo possui ou não os atributos avaliados, bem como η_{ij}, que recebe $\eta_{ij} = 0$, para aqueles respondentes que não possuem ao mínimo um dos atributos requeridos para o item j e $\eta_{ij} = 1$ para os que possuem.

Uma vez que agora Y_{ij} é uma variável contínua limitada no intervalo unitário, a distribuição Beta pode ser considerada para modelar as respostas no modelo B-DINA. Essa é uma escolha natural e tem sido satisfatoriamente utilizada na modelagem de dados limitados nos últimos anos, em diferentes campos de pesquisa (GUPTA; NADARAJAH, 2004; CHYLACK *et al.*, 2009; KOPP; GERIKE; AXHAUSEN, 2015; NOEL; DAUVIER, 2007). Por questões de interepretabilidade, é importante considerar uma parametrização da distribuição que tenha um parâmetro de localização, sendo que, possuir um parâmetro relacionado à dispersão/precisão também é interessante. A distribuição Beta aqui considerada segue a proposta de Ferrari e Cribari-Neto (2004a), a qual traz grandes benefícios para a interpretabilidade do modelo.

Assim, dado que Y_{ij} é condicional à η_{ij}, temos que $[Y_{ij} \mid \eta_{ij} = \eta] \sim \text{Beta}(\mu_{j\eta}, \phi_{j\eta})$, onde $\mu_{j\eta}$ é a média da distribuição e $\phi_{j\eta}$ é um parâmetro de precisão.

Denotando $f_{j\eta}(y_{ij})$ como a fdp da distribuição Beta dado $\eta_{ij} = \eta$, temos então que

$$f_{j\eta}(y_{ij}) = \frac{\Gamma(\phi_{j\eta})}{\Gamma(\mu_{j\eta}\phi_{j\eta})\Gamma((1-\mu_{j\eta})\phi_{j\eta})} y_{ij}^{\mu_{j\eta}\phi_{j\eta}-1}(1-y_{ij})^{(1-\mu_{j\eta})\phi_{j\eta}-1}. \tag{4.1}$$

$\eta \in \{0,1\}$, onde $\text{Beta}(\mu,\phi)$ denota a distribuição Beta, com $0 < y_{ij} < 1$, $0 < \mu_{j\eta} < 1$ e $\phi > 0$.

Note que qualquer distribuição contínua limitada no intervalo unitário pode ser considerada como $f_{j\eta}(\cdot)$ para futuras pesquisas, com possibilidades de extensão do modelo proposto para outras distribuições que não a Beta.

Agora, tome $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_J)'$ como o vetor de parâmetros dos itens no modelo, com $\boldsymbol{\Omega}_j = (\mu_{j0}, \phi_{j0}, \mu_{j1}, \phi_{j1})'$ para $j = 1, \cdots, J$ denotando os parâmetros para o j-ésimo item. Dado

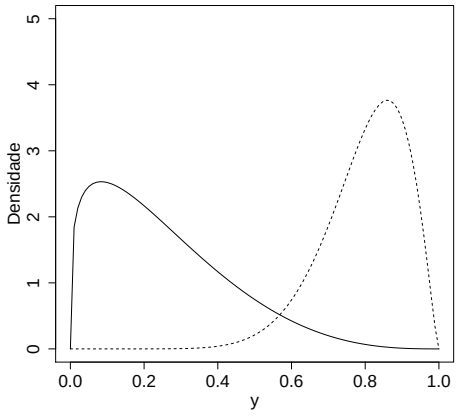


Figura 9 – Função densidade de probabilidade de um item bem construído. $\eta = 0$ linha cheia e $\eta = 1$ linha hachurada.

$\boldsymbol{\alpha}_i$ e $\boldsymbol{\Omega}_j$, a resposta Y_{ij} tem a seguinte distribuição:

$$[Y_{ij} \mid \boldsymbol{\alpha}_i, \boldsymbol{\Omega}_j] \sim f_j(y_{ij}) = \left[f_{j0}(y_{ij})\right]^{1-\eta_{ij}} \left[f_{j1}(y_{ij})\right]^{\eta_{ij}}, \tag{4.2}$$

$$= (1-\eta_{ij}) \left[f_{j0}(y_{ij})\right] + \eta_{ij} \left[f_{j1}(y_{ij})\right]. \tag{4.3}$$

Então, dado $\boldsymbol{\alpha}_i$ e $\boldsymbol{\Omega}_j$, a fda é $\mathrm{P}(Y_{ij} \leq y | \boldsymbol{\alpha}_i, \boldsymbol{\Omega}_j) = \int_0^y f_j(y_{ij}) dy_{ij}$.

Genericamente, para itens bem construídos, esperamos que a distribuição dos grupos $\eta = 0$ e $\eta = 1$ siga um padrão que permita ao modelo identificar diferentes densidades para cada grupo, como pode ser visto na Figura 9. Na Seção 4.4 são dados mais detalhes sobre as discriminações utilizadas no estudo de simulação desse capítulo, a fim de avaliar a recuperação de parâmetros, mostrando como a boa construção dos itens é essencial na separação dos públicos.

Em nosso modelo B-DINA proposto, os itens e os respondentes são considerados independentes. Então, a distribuição condicional de $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iJ})'$ dado $\boldsymbol{\alpha}_i$ e todos os parâmetros dos itens pode ser facilmente escrita como

$$p(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\Omega}) = \prod_{j=1}^{J} f_j(y_{ij}) = \prod_{j=1}^{J} \left[f_{j0}(y_{ij})\right]^{1-\eta_{cj}} \left[f_{j1}(y_{ij})\right]^{\eta_{cj}}, \tag{4.4}$$

onde podemos explicitar $f_{j0}(y_{ij})$ e $f_{j1}(y_{ij})$ usando a Equação (4.1).

Também definimos $\pi_c = \mathrm{P}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c \mid \boldsymbol{\pi})$ como a probabilidade de pertencer à classe c, onde $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_C)'$ é um vetor C dimensional, $0 \leq \pi_c \leq 1$ e $\sum_{c=1}^{C} \pi_c = 1$. Novamente temos que a fdp de $\boldsymbol{y}_i$ dados todos os parâmetros $\boldsymbol{\Omega}$ e todas as classes latentes de probabilidade $\boldsymbol{\pi}$ é

$$p(\boldsymbol{y}_i \mid \boldsymbol{\Omega}, \boldsymbol{\pi}) = \sum_{c=1}^{C} \pi_c p(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\Omega}). \tag{4.5}$$

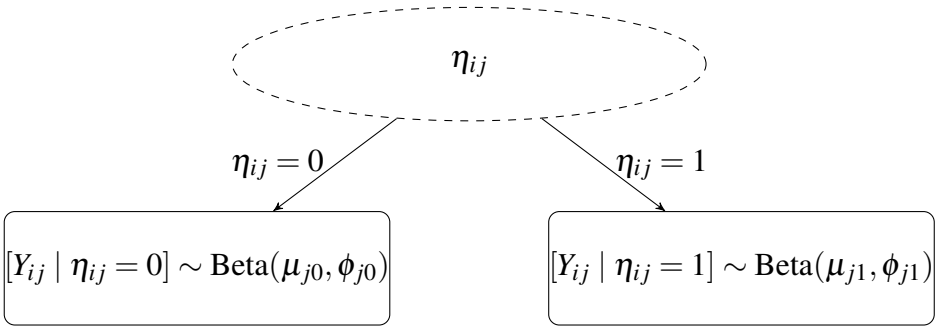


Figura 10 – Estrutura do Nível 1 para o modelo B-DINA, onde 'formato elíptico' indica uma variável latente e 'formato retangular' indica dados observados.

Assumindo a distribuição Beta para $f_{j\eta}(\cdot)$, como mostrado na Equação (4.1), a verossimilhança pode ser explicitamente escrita como

$$L(\boldsymbol{Y}|\boldsymbol{\Omega},\boldsymbol{\pi}) = \prod_{i=1}^{N}\sum_{c=1}^{C}\pi_c \prod_{j=1}^{J}\left[\frac{\Gamma(\phi_{j0})}{\Gamma(\mu_{j0}\phi_{j0})\Gamma((1-\mu_{j0})\phi_{j0})} y_{ij}^{\mu_{j0}\phi_{j0}-1}(1-y_{ij})^{(1-\mu_{j0})\phi_{j0}-1}\right]^{1-\eta_{cj}} \left[\frac{\Gamma(\phi_{j1})}{\Gamma(\mu_{j1}\phi_{j1})\Gamma((1-\mu_{j1})\phi_{j1})} y_{ij}^{\mu_{j1}\phi_{j1}-1}(1-y_{ij})^{(1-\mu_{j1})\phi_{j1}-1}\right]^{\eta_{cj}} \tag{4.6}$$

onde $\boldsymbol{y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N)'$. Na Seção 4.3, uma proposta com abordagem Bayesiana para a estimação de todos os parâmetros é apresentada.

Interpretação da construção do modelo B-DINA

Mais uma vez, a interpretação por trás da construção do modelo B-DINA é bastante similar ao caso do C-DINA, que pode ser visto na subseção 3.2. O modelo B-DINA também possui os mesmos dois níveis que foram explicados em detalhe anteriormente, havendo modificação na variável resposta. Portanto, a interpretação dos dois níveis não será explicitada nesse capítulo passo a passo, sendo apenas mostradas as diferenças.

No Nível 1, para o modelo B-DINA, as respostas observadas, como já ressaltado, são contínuas limitadas, sendo que para nossa proposta a distribuição Beta é considerada para a modelagem. Com isso, a estrutura do modelo será como apresentada na Figura 10.

Nesse nível, precisamos de informações sobre a variável latente η_{ij}, assim como ocorria para o modelo C-DINA, porém, agora os parâmetros do primeiro nível a serem estimados são μ_{j0}, μ_{j1}, ϕ_{j0}, ϕ_{j1}, para cada item. O segundo nível da hierarquia segue a mesma estrutura antes mostrada no C-DINA, com alteração apenas nesses parâmetros, sendo a explicação mais detalhada da lógica desse nível demonstrada no capítulo anterior.

4.3 Estimação Bayesiana do modelo B-DINA

Antes de seguirmos com a especificação das distribuições a *priori* e demais desenvolvimentos da estimação Bayesiana, é importante ressaltar que as expressões na Equação (4.1) mostram que o modelo B-DINA é um caso particular de um modelo de mistura. Assim como no caso do C-DINA, um problema que pode acontecer é o chamado *label switching*, caso não forem tomadas medidas adequada.

Para o modelo B-DINA, também é razoável assumir que $\mu_{j1} > \mu_{j0}$, uma vez que a média do grupo $\eta = 1$ deve ser maior do que a do grupo $\eta = 0$, por construção. Porém, mesmo que seja feita uma transformação para que o valor de μ_{j0} esteja na reta (digamos μ^*_{j0}) e possa ser adicionado a um λ_j positivo para que a soma de ambas as quantidades resulte na transformação de μ_{j1} (digamos $\mu^*_{j1} = \mu^*_{j0} + \lambda_j$) e, após essas quantidades serem estimadas, elas sofram uma nova transformação para voltar para sua forma original, o truque usado na Seção 3.3 não irá funcionar, uma vez que essas transformações não levariam em conta a natureza limitada de μ_{j1} que está limitado ao intervalo $(0,1)$. Para o caso do modelo B-DINA, uma maneira de corrigir o problema de *label switching* é desenvolver o código do modelo de tal maneira que não seja possível ter $\mu_{j1} < \mu_{j0}$. A abordagem utilizada para garantir isso é a de forçar a ordenação dos parâmetros μ, similar ao que é visto em Gonçalves, Dias e Soares (2018) para um modelo de mistura de TRI. Com isso, teremos uma ordenação de μ_{j0} e μ_{j1} em ordem crescente, assim evitando a possível bimodalidade. Vale ressaltar que também foram testadas outras maneiras de resolver o problema, bem como definir as prioris de maneira que μ_{j1} seja sempre maior que μ_{j0}, considerando, por exemplo, uma priori $\text{Unif}(\mu_{j0}, 1)$ para μ_{j1}. Entretanto, ao testar essa abordagem, os resultados para a recuperação de parâmetros foram piores, bem como os valores de WAIC maiores. Portanto, no restante do trabalho a abordagem da ordenação será seguida, como pode ser visto no código JAGS no Apêndice.

Especificação de Prioris

Para o modelo B-DINA, assim como para outros MDCs em geral, há parâmetros dos indivíduos e parâmetros dos itens. Nessa seção serão discutidas as distribuições a *priori* para ambos os tipos de parâmetros. Note que ambos os tipos de parâmetros são condicionados ao grupo latente η.

Parâmetros dos indivíduos

Para $\boldsymbol{\alpha}_i|\boldsymbol{\pi}$ a distribuição *a priori* precisa levar em conta a possibilidade de pertença de um determinado indivíduo à cada perfil de atributos, dada a probabilidade de pertença àquele perfil na população. A distribuição Categórica é, novamente, escolhida, levando a

$$\boldsymbol{\alpha}_i|\boldsymbol{\pi} \sim \text{Categorica}(\pi_1, \pi_2, \ldots, \pi_C), \tag{4.7}$$

com $\sum_{i=1}^{C} \pi_i = 1$. Assumindo que os hiperparâmetros sejam $\pi_1 = \pi_2 = \ldots = \pi_C = 1/C$, temos uma distribuição *a priori* uniforme, portanto, não informativa.

Entretanto, novamente, utilizamos uma hiperpriori para $\boldsymbol{\pi}$. Tal hiperpriori deve considerar a probabilidade de um indivíduo aletório de pertencer à cada um dos C diferentes perfis de atributo. Uma possível escolha é a distribuição Dirichlet

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\delta}_0), \boldsymbol{\delta}_0 = (\delta_{01}, \delta_{02}, \ldots, \delta_{0C}), \tag{4.8}$$

escolher $\delta_{01} = \delta_{02} = \ldots = \delta_{0C} = 1$ leva a hiperpriori para π a ser não informativa.

Assumindo independência para os parâmetros individuais $\boldsymbol{\alpha}_i$, $i = 1, \ldots, N$, dado π, a distribuição conjunta de $\boldsymbol{\alpha}$ e π é dada por

$$p(\boldsymbol{\alpha}, \boldsymbol{\pi}) = \left[\prod_{i=1}^{N} p(\boldsymbol{\alpha_i} \mid \boldsymbol{\pi})\right] p(\boldsymbol{\pi}) = \left(\prod_{i=1}^{N}\prod_{c=1}^{C} \pi_c^{\mathbb{1}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c)}\right)\left(\frac{1}{B(\boldsymbol{\delta}_0)}\prod_{c=1}^{C} \pi_c^{\delta_{0c}-1}\right) \tag{4.9}$$

com $B(\boldsymbol{\delta}_0) = \prod_{i=1}^{C} \Gamma(\delta_{0i}) / \Gamma(\sum_{i=1}^{C} \delta_{0i})$.

Parâmetros dos itens

Agora, passamos à discussão das prioris dos parâmetros dos itens, levando em conta a natureza de cada um dos parâmetros.

Os parâmetros de média estão localizados no intervalo $(0, 1)$, sendo que uma boa escolha é a distribuição Beta. Temos que

$$\mu_{j0} \sim \text{Beta}(\alpha_{\mu_0}, \beta_{\mu_0}), j = 1, \ldots, J, \tag{4.10}$$

$$\mu_{j1} \sim \text{Beta}(\alpha_{\mu_1}, \beta_{\mu_1}), j = 1, \ldots, J, \tag{4.11}$$

Note que $\alpha_{\mu_0} = \beta_{\mu_0} = 1$ e $\alpha_{\mu_1} = \beta_{\mu_1} = 1$ fazem com que as prioris sejam vagas.

Para os parâmetros de precisão, uma distribuição com domínio na reta positiva deve ser considerada. Uma boa escolha é a distribuição Gama. Temos que

$$\phi_{j0} \sim \text{Gama}(\alpha_{\phi_0}, \beta_{\phi_0}), j = 1, \ldots, J, \tag{4.12}$$

$$\phi_{j1} \sim \text{Gama}(\alpha_{\phi_1}, \beta_{\phi_1}), j = 1, \ldots, J, \tag{4.13}$$

Tomando $\alpha_{\phi_0} = \beta_{\phi_0} = 0.01$ a priori será menos informativa para ϕ_{j0}, bem como $\alpha_{\phi_1} = \beta_{\phi_1} = 0.01$ o faz para ϕ_{j1}.

Assumindo independência entre os parâmetros dos itens, para cada item $j = 1,\ldots,J$, temos que a distribuição conjunta é dada por

$$\begin{aligned} p(\boldsymbol{\Omega}) &= \prod_{j=1}^{J}\left[p(\mu_{j0})p(\mu_{j1})p(\phi_{j0})p(\phi_{j1})\right] \\ &= \left(\frac{\Gamma(\alpha_{\mu_0}+\beta_{\mu_0})\Gamma(\alpha_{\mu_1}+\beta_{\mu_1})}{\Gamma(\alpha_{\mu_0})\Gamma(\beta_{\mu_0})\Gamma(\alpha_{\mu_1})\Gamma(\beta_{\mu_1})}\right)^{J}\left(\frac{\beta_{\phi_0}^{\alpha_{\phi_0}}\beta_{\phi_1}^{\alpha_{\phi_1}}\phi_{j0}^{\alpha_{\phi_0}-1}\phi_{j1}^{\alpha_{\phi_1}-1}}{\Gamma(\alpha_{\phi_0})\Gamma(\alpha_{\phi_1})}\right)^{J} \\ &\times \prod_{j=1}^{J}\mu_{j0}^{\alpha_{\mu_0}-1}(1-\mu_{j0})^{\beta_{\mu_0}-1}\mu_{j1}^{\alpha_{\mu_1}-1}(1-\mu_{j1})^{\beta_{\mu_1}-1}\exp\left[-\left(\beta_{\phi_0}\phi_{j0}+\beta_{\phi_1}\phi_{j1}\right)\right] \end{aligned} \tag{4.14}$$

Estimação dos parâmetros

O modelo B-DINA pode ser sumarizado hierarquicamente como segue

$$\begin{aligned} y_{ij} \mid \boldsymbol{\alpha}_i, \boldsymbol{\Omega}_j, \eta_{ij} &\sim \eta_{ij}\text{Beta}(\mu_{j1},\phi_{j1}) + (1-\eta_{ij})\text{Beta}(\mu_{j0},\phi_{j0}), \qquad (4.15) \\ \eta_{ij} &= \prod_{k=1}^{K}\alpha_{ik}^{q_{jk}} = \mathbb{1}(\boldsymbol{\alpha}_i'\mathbf{q}_j = \mathbf{q}_j'\mathbf{q}_j), \\ \boldsymbol{\alpha}_i|\boldsymbol{\pi} &\sim \text{Categorica}(\boldsymbol{\pi}), \\ \boldsymbol{\pi} &\sim \text{Dirichlet}(\boldsymbol{\delta}_0), \\ \mu_{j0} &\sim \text{Beta}(\alpha_{\mu_0},\beta_{\mu_0}), \\ \mu_{j1} &\sim \text{Beta}(\alpha_{\mu_1},\beta_{\mu_1}), \\ \phi_{j0} &\sim \text{Gamma}(\alpha_{\phi_0},\beta_{\phi_0}) \\ \phi_{j1} &\sim \text{Gamma}(\alpha_{\phi_1},\beta_{\phi_1}) \end{aligned} \tag{4.16}$$

$\boldsymbol{\Omega}_j = (\mu_{j0},\mu_{j1},\phi_{j0},\phi_{j1})$, com $\boldsymbol{\pi} = (\pi_1,\pi_2,\ldots,\pi_C)'$ e $\boldsymbol{\delta}_0 = (\delta_{01},\delta_{02},\ldots,\delta_{0C})'$ Incluindo o vetor latente $\boldsymbol{\alpha}_i$ na distribuição conjunta da posteriori do modelo B-DINA temos que

$$\begin{aligned} & p(\boldsymbol{\Omega},\boldsymbol{\alpha},\boldsymbol{\pi} \mid \boldsymbol{y}) \\ \propto\ & \mathscr{L}(\boldsymbol{y}|\boldsymbol{\Omega},\boldsymbol{\alpha}) \times p(\boldsymbol{\Omega}) \times p(\boldsymbol{\alpha} \mid \boldsymbol{\pi})p(\boldsymbol{\pi}) \\ \propto\ & \left\{\prod_{i=1}^{N}\prod_{j=1}^{J}\left[\frac{\Gamma(\phi_{j0})}{\Gamma(\mu_{j0}\phi_{j0})\Gamma((1-\mu_{j0})\phi_{j0})}y_{ij}^{\mu_{j0}\phi_{j0}-1}(1-y_{ij})^{(1-\mu_{j0})\phi_{j0}-1}\right]^{1-\eta_{ij}}\right. \\ \times\ & \left.\left[\frac{\Gamma(\phi_{j1})}{\Gamma((\mu_{j1})\phi_{j1})\Gamma((1-\mu_{j1})\phi_{j1})}y_{ij}^{\mu_{j1}\phi_{j1}-1}(1-y_{ij})^{(1-\mu_{j1})\phi_{j1}-1}\right]^{\eta_{ij}}\right\} \\ \times\ & p(\boldsymbol{\Omega}) \times p(\boldsymbol{\alpha},\boldsymbol{\pi}), \end{aligned} \tag{4.17}$$

onde $p(\boldsymbol{\alpha}, \boldsymbol{\pi})$ é definido na Equação (4.9) e $p(\boldsymbol{\Omega})$ é definido na Equação (4.14).

A distribuição *a posteriori* é analiticamente intratável, e também há problemas para obter distribuições condicionais completas que possuam uma distribuição conhecida. Entretanto, há diversas maneiras de estimar os parâmetros do modelo B-DINA por meio de uma abordagem Bayesiana. Assim como no capítulo anterior, fazemos uso do software JAGS (PLUMMER, 2015), por meio de sua interface no R, "R2jags". Uma vez que não há forma fechada para as distribuições condicionais completas dos parâmetros desconhecidos, o software se vale do *slice sampling* para obter as amostras da posteriori (NEAL, 2003). Não são apresentados detalhes do algoritmo, mas, em resumo, considerando a fdp $f(x)$ ou uma expressão proporcional à essa função, ele consiste nos passos mostrados no Algoritmo 1.

Algoritmo 1 – *Slice sampling*

1: Tome um valor inicial x_0, com $f(x_0) > 0$
2: Considere o intervalo $(0, f(x_0))$, amostre um valor real y uniformemente, dentro desse intervalo
3: Trace uma linha horizontal, criando uma fatia (*slice sampling*), sob a curva na posição y
4: Encontre um intervalo $I = (L, R)$ ao redor do ponto x_0, contendo toda a fatia ou o máximo possível
5: Amostre x_1 da parte da fatia dentro do intervalo I
6: Volte ao passo 2

Tanto para a simulação quanto para a aplicação, as amostras da posteriori do B-DINA são retiradas considerando duas cadeias, com $25,000$ iterações cada, descartando as primeiras $5,000$. Essas cadeias foram rodadas em paralelo, a fim de diminuir o tempo computacional. A convergência MCMC foi checada por meio de análises gráficas e pelo critério de Gelman-Rubin, com estatísticas próximas a 1 indicando convergência.

Checagem preditiva a posteriori

Com o intuito de verificar o ajuste do modelo aos dados em aplicações realizadas por meio da abordagem Bayesiana, uma das possibilidades é a de utilizar a distribuição preditiva *a posteriori*.

Como dito em Gelman *et al.* (2013), valores p Bayesianos podem ser obtidos para comparar os dados observados à distribuição preditiva a posteriori, avaliando a probabilidade de que réplicas simuladas da posteriori sejam mais extremas do que os dados observados, utilizando uma estatística de teste $T(\boldsymbol{y})$ pertinente, a qual também pode ser entendida como uma medida de discrepância $D(\boldsymbol{y})$ (MENG *et al.*, 1994). Podemos definir o valor p Bayesiano como:

$$p_B = P\left(T(\boldsymbol{y}^{rep}, \boldsymbol{\vartheta}) \geq T(\boldsymbol{y}^{obs}, \boldsymbol{\vartheta}))|\boldsymbol{y}^{obs}\right), \tag{4.18}$$

onde essa probabilidade é calculada levando em conta a distribuição *a posteriori* dos parâmetros $\boldsymbol{\vartheta}$, a distribuição preditiva *a posteriori* de $\boldsymbol{y}^{rep}$, sendo que $p(\boldsymbol{y}^{rep}|\boldsymbol{y}^{obs}) = \int p(\boldsymbol{y}^{rep}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|\boldsymbol{y}^{obs})d\boldsymbol{\vartheta}$.

Temos ainda que $\boldsymbol{y}^{obs}$ representa os valores observados e T é uma estatística de teste conveniente para cada caso.

Na prática, $\boldsymbol{y}^{rep}$ pode ser obtido por meio de valores simulados para cada réplica a partir das estimativas *a posteriori* dos parâmetros. Para o nosso caso, podemos seguir os seguintes passos antes de calcular o valor p Bayesiano:

- **Passo 1:** Gerar as estimativas *a posteriori* para os parâmetros dos itens $\mu_{j0}, \mu_{j1}, \phi_{j0}, \phi_{j1}$, para $j = 1, 2, \ldots, J$.
- **Passo 2:** Gerar as estimativas *a posteriori* para os parâmetros dos indivíduos $\boldsymbol{\alpha}_i$, $i = 1, 2, \ldots, N$.
- **Passo 3:** Utilizando os valores das estimativas dos dois primeiros passos, simular a matriz de respostas $\boldsymbol{y}^{rep}$, com valores simulados para cada indivíduo i e item j.

Uma vez que esses passos sejam executados R vezes, teremos R réplicas dos dados simulados, ou seja, $\boldsymbol{y}^{rep}_{(1)}, \boldsymbol{y}^{rep}_{(2)}, \ldots, \boldsymbol{y}^{rep}_{(R)}$. Como estamos lidando com dados contínuos no intervalo unitário, podemos calcular o valor p Bayesiano avaliando a quantidade de vezes, ao longo das réplicas, que a média obtida nas réplicas é maior do que a média real dos dados, por exemplo, utilizando assim a média como $T(\boldsymbol{y})$, assim como citado no primeiro exemplo de Gelman *et al.* (2013). Esse procedimento pode ser feito para cada item j $(j = 1, 2, \ldots, J)$, sendo que, para cada um dos itens, teremos as seguintes quantidades, em cada réplica

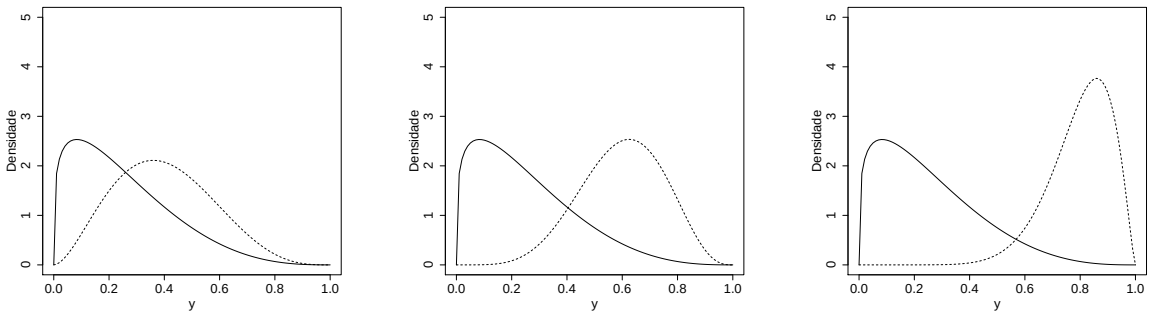
$$\bar{Y}^{obs}_{j} = \sum_{i=1}^{N} \frac{Y^{obs}_{i}}{N}, \quad \bar{Y}^{rep}_{j(r)} = \sum_{i=1}^{N} \frac{Y^{rep}_{i(r)}}{N},$$

para $r = 1, \ldots, R$.

Com isso, podemos calcular o valor p Bayesiano utilizando a média a posteriori de cada réplica, da seguinte maneira:

$$p_B = \frac{\sum_{r=1}^{R} \mathbb{1}\left(\bar{Y}^{rep}_{j(r)} \geq \bar{Y}^{obs}_{j}\right)}{R} \tag{4.19}$$

Uma vez obtido o valor p Bayesiano, podemos verificar como o ajuste para cada item se comporta em relação à adequabilidade aos dados reais. Quanto melhor o modelo se adeque aos dados, mais próximo de 0.5 deve ser p_B (GELMAN *et al.*, 2013). Gelman *et al.* (2013) citam como uma possibilidade para avaliar o bom ajuste dos dados considerar como bons valores p aqueles que estejam entre 0.05 e 0.95, enquanto há autores, como Sahu (2002), que sugerem uma abordagem mais conservadora, o intervalo entre 0.10 e 0.90, para modelos de TRI.



(a) Baixa Discriminação, $KS = 0.35, m_j = 0.11$

(b) Média Discriminação, $KS = 0.70, m_j = 0.42$

(c) Alta Discriminação, $KS = 0.90, m_j = 0.75$

Figura 11 – Densidade por discriminação dos itens. $\eta = 0$ linha cheia e $\eta = 1$ linha hachurada.

4.4 Estudo de Simulação

Nessa Seção um estudo de recuperação de parâmetros é conduzido, considerando diferentes cenários baseados na discriminação dos itens (baixa, média ou alta), o número de respondentes (N) e o número de itens (J).

Discriminação dos itens no modelo B-DINA

Como pode ser visto na Figura 6, no primeiro nível são propostas duas distribuições, uma para $\eta = 0$ e outra para $\eta = 1$. A fim de definir a discriminação dos itens podemos utilizar medidas de distância. Aqui consideramos os valores de KS e m_j, previamente definidos.

Os cenários da simulação possuem valores $(\mu_{j0}, \mu_{j1}, \phi_{j0}, \phi_{j1})' = (\mu_0, \mu_1, \phi_0, \phi_1)'$ para os diferentes itens, onde $(\mu_0, \mu_1, \phi_0, \phi_1)'$ é igual a $(0.25, 0.4, 5, 7)'$ para o caso de Baixa discriminação, igual a $(0.25, 0.6, 5, 10)'$ para Média discriminação e $(0.25, 0.8, 5, 12)'$ representa a Alta discriminação. Os valores de KS e m_j são $KS = 0.35, m_j = 0.11$; $KS = 0.70, m_j = 0.42$; e $KS = 0.90, m_j = 0.75$, respectivamente. A Figura 11 mostra a diferença entre as densidades para os grupos em cada discriminação, tornando mais visual como as curvas se comportam para cada η em cada um dos casos.

Cenários da Simulação

O estudo de simulação foi construído considerando dezoito diferentes cenários. Para o número de respondentes, temos três casos $N = 250$, $N = 500$ ou $N = 1000$. Também são considerados diferentes números dos itens, com $J = 15$ ou $J = 30$. A Tabela 16 mostra a Matriz $\boldsymbol{Q}$ utilizada no estudo de simulação, a qual considera $K = 5$ dimensões mensuradas pelo teste, de maneira similar à construção da matriz para o modelo C-DINA.

Para gerar os dados simulados, primeiramente são geradas as classes dos respondentes,

Tabela 16 – Matriz $\boldsymbol{Q}$ para o estudo de simulação.

	Atributo						Atributo				
Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0	16	0	1	0	1	0
2	0	1	0	0	0	17	0	1	0	0	1
3	0	0	1	0	0	18	0	0	1	1	0
4	0	0	0	1	0	19	0	0	1	0	1
5	0	0	0	0	1	20	0	0	0	1	1
6	1	0	0	0	0	21	1	1	1	0	0
7	0	1	0	0	0	22	1	1	0	1	0
8	0	0	1	0	0	23	1	1	0	0	1
9	0	0	0	1	0	24	1	0	1	1	0
10	0	0	0	0	1	25	1	0	1	0	1
11	1	1	0	0	0	26	1	0	0	1	1
12	1	0	1	0	0	27	0	1	1	1	0
13	1	0	0	1	0	28	0	1	1	0	1
14	1	0	0	0	1	29	0	1	0	1	1
15	0	1	1	0	0	30	0	0	1	1	1

por meio de uma distribuição multinomial com tamanho igual à N e probabilidade igual à $1/32$ (note que, como $K = 5$ e $C = 2^K = 32$, isso faz com que a probabilidade de pertencer a cada um dos perfis de atributos/dimensões seja a mesma. Então, baseado nas classes geradas e na Matriz $\boldsymbol{Q}$ na Tabela 16, é possível obter os valores de η_{ij} para cada respondente e item. Por fim, as respostas y_{ij} podem ser geradas a partir de uma distribuição Beta$(\mu_{j\eta}, \phi_{j\eta})$. Os valores gerados para as respostas são então considerados como os valores verdadeiros observados e, baseado nesses valores, é possível ajustar um modelo B-DINA com parâmetros desconhecidos, a fim de avaliar o quão boa é a recuperação de parâmetros para cada um dos cenários considerados.

Estudo de Recuperação

Nessa seção a recuperação de parâmetros é mostrada, considerando 100 réplicas para cada um dos cenários propostos. As distribuições *a priori* são especificadas como $\mu_{j0} \sim$ Beta$(1,1)$, $\mu_{j1} \sim$ Beta$(1,1)$, $\phi_{j0} \sim$ Gamma$(0.01, 0.01)$, $\phi_{j1} \sim$ Gamma$(0.01, 0.01)$, para cada um dos $j = 1, \ldots, J$ itens.

Para avaliar a performance do modelo ao longo das réplicas, para os parâmetros dos itens, além do RMSE, já definido anteriormente, também foi considerado o viés relativo (VR), ao invés do viés absoluto, devido à grande diferença de escala entre os parâmetros. Temos que $\hat{\text{VR}} = \frac{(\hat{\bar{\vartheta}}_l - \vartheta_l)}{\vartheta_l}$, com $\hat{\bar{\vartheta}}_l = \frac{1}{R}\sum_{r=1}^{R} \hat{\vartheta}_{lr}$. Para os parâmetros dos indivíduoss foram consideradas as mesmas medidas de acurácia utilizadas para a avaliação do modelo C-DINA, ou seja, a acurácia simples (SA) e a acurácia vetorial (VA).

Resultados

Os resultados para a recuperação dos parâmetros dos itens são mostrados da Tabela 17 até a Tabela 20. No que diz respeito à recuperação de parâmetros, é notável que o aumento no número de respondentes tem um grande impacto tanto no VR quanto no RMSE, para todos os cenários quando as outras características são fixadas, com melhores resultados obtidos para valores maiores de N. Quanto ao número de itens, é interessante notar que para $J = 30$, em geral, a recuperação de parâmetros também é melhor. Um dos resultados mais interessantes é o que diz respeito à discriminação dos itens. Para questionários com itens com discriminação média, um tamanho amostral de $N = 250$ trouxe resultados piores, sendo que é indicado ter ao menos $N = 500$ respondentes. A recuperação de parâmetros é bastante sensível em relação à discriminação dos itens, mostrando que um questionário/avaliação bem elaborado(a), com itens que descriminam bem entre respondentes, é essencial para o ajuste do modelo. Também é notável que os valores de RMSE para os parâmetros ϕ_0 e ϕ_1 são consideravelmente maiores que aqueles para μ_0 e μ_1, o que é esperado devido à escala dos parâmetros, mas não há grandes problemas para os cenários com melhor discriminação entre itens.

Baseado nesse resultado, podemos concluir que o modelo performa bem para recuperar os parâmetros dos itens para os cenários com discriminação Alta, para todas as combinações de números de respondentes e itens testados. Para os cenários com baixa discriminação a recuperação dos parâmetros não é muito boa, principalmente quando $N = 250$, com destaque para os parâmetros ϕ. Também foram encontrados bons resultados para os cenários com Média discriminação, mesmo para $N \leq 500$, mas é interessante notar que os resultados melhoram com o aumento do tamanho amostral. Também é interessante notar que tanto para o VR quanto para o RMSE, os parâmetros para o grupo $\eta = 1$ tem resultados maiores, principalmente para discriminação baixa ou média.

Na Tabela 21, são mostrados os resultados para os parâmetros dos indivíduoss ($\boldsymbol{\alpha}_i$'s). Podemos ver que a influência do número de itens é bastante forte no que diz respeito à recuperação desses parâmetros. Para os cenários com discriminação baixa, os resultados de AV são bastante ruins, principalmente quando $J = 15$, mostrando, novamente, a importância da construção adequada dos itens. Para os casos com itens de Média discriminação e Alta discriminação, o aumento no número de itens ajuda bastante na recuperação correta dos parâmetros dos indivíduoss, com resultados bastante melhores quando $J = 30$. O aumento no número dos respondentes não tem tanto impacto na recuperação dos parâmetros dos indivíduoss quanto foi verificado que tem nos parâmetros dos itens, para os cenários de Média discriminação e Alta discriminação, porém, para Baixa discriminação, o efeito de aumentar o tamanho amostral é mais destacado.

Para concluir, comparado à outras propostas, como o modelo C-DINA, o B-DINA é mais sensível em relação ao número de itens, de respondentes e principalmente em relação à discriminação dos itens. Principalmente para os parâmetros dos indivíduoss, que se mostraram ainda mais sensíveis, é de suma importância ter itens construídos adequadamente, que discriminem

bem entre os respondentes, bem como o número de itens J não deve ser muito pequeno. Nossa proposta Bayesiana para a estimação dos parâmetros do modelo B-DINA traz uma recuperação adequada de todos os parâmetros para questionários bem construídos. A seguir, nosso modelo Bayesiano será utilizado para uma aplicação do B-DINA em dados reais.

Tabela 17 – Performance para a recuperação do parâmetro μ_0 ao longo das 100 réplicas

			Grupos de itens pelo número de atributos requeridos					
			Um		Dois		Três	
J	Discriminação	N	RMSE	VR	RMSE	VR	RMSE	VR
15	Baixa	250	0.016	0.05	0.010	0.03	0.008	0.03
		500	0.011	0.03	0.008	0.02	0.006	0.02
		1000	0.008	0.02	0.006	0.01	0.004	0.01
	Média	250	0.009	0.02	0.006	0.01	0.005	0.01
		500	0.008	0.01	0.004	0.01	0.004	0.01
		1000	0.005	0.01	0.003	0.00	0.003	0.00
	Alta	250	0.008	0.01	0.006	0.00	0.005	0.00
		500	0.005	0.00	0.004	0.00	0.004	0.00
		1000	0.004	0.00	0.003	0.00	0.002	0.00
30	Baixa	250	0.011	0.03	0.008	0.02	0.006	0.02
		500	0.006	0.01	0.005	0.01	0.004	0.01
		1000	0.003	0.01	0.002	0.01	0.001	0.01
	Média	250	0.006	0.01	0.004	0.00	0.004	0.00
		500	0.004	0.00	0.003	0.00	0.003	0.00
		1000	0.002	0.00	0.002	0.00	0.002	0.00
	Alta	250	0.005	0.01	0.004	0.00	0.004	0.00
		500	0.003	0.00	0.003	0.00	0.002	0.00
		1000	0.003	0.00	0.002	0.00	0.002	0.00

4.5 Aplicação 1: Dados sociais da região Sudeste do Brasil

Nessa aplicação, dados sobre indicadores sociais da região Sudeste do Brasil são analisados. O conjunto de dados desse estudo vem do censo de 2010, estando disponível no site http://www.atlasbrasil.org.br/consulta (consultado em 2 de Outubro de 2020). Há dados disponíveis para todos os municípios brasileiros, porém o foco de nosso estudo são aqueles da região Sudeste, composta por São Paulo, Rio de Janeiro, Minas Gerais e Espírito Santo. O tamanho da base analisada é de 1668 municípios, com $J = 19$ itens, os quais avaliam três dimensões diferentes: a econômica, a de saúde e a de educação.

Todos os indicadores considerados como itens possuem valores no intervalo unitário. Em alguns casos há municípios com indicadores iguais à 1 (para porcentagem de domicílios com

Tabela 18 – Performance para a recuperação do parâmetro μ_1 ao longo das 100 réplicas

			Grupos de itens pelo número de atributos requeridos					
			Um		Dois		Três	
J	Discriminação	*N*	RMSE	VR	RMSE	VR	RMSE	VR
15	Baixa	250	0.013	-0.01	0.021	0.02	0.041	0.08
		500	0.010	0.00	0.015	0.01	0.022	0.03
		1000	0.007	0.00	0.010	0.00	0.015	0.02
	Média	250	0.008	0.00	0.010	0.00	0.015	0.00
		500	0.005	0.00	0.006	0.00	0.010	0.00
		1000	0.004	0.00	0.005	0.00	0.007	0.00
	Alta	250	0.005	0.00	0.007	0.00	0.009	0.00
		500	0.004	0.00	0.004	0.00	0.007	0.00
		1000	0.002	0.00	0.003	0.00	0.004	0.00
30	Baixa	250	0.008	0.01	0.015	0.02	0.021	0.03
		500	0.005	0.00	0.008	0.00	0.011	0.01
		1000	0.003	0.00	0.004	0.00	0.008	0.01
	Média	250	0.004	0.00	0.005	0.00	0.008	0.00
		500	0.003	0.00	0.004	0.00	0.006	0.00
		1000	0.002	0.00	0.003	0.00	0.004	0.00
	Alta	250	0.003	0.00	0.004	0.00	0.007	0.00
		500	0.002	0.00	0.003	0.00	0.005	0.00
		1000	0.002	0.00	0.002	0.00	0.003	0.00

banheiro e água corrente, por exemplo, há 51 municípios que tem o valor 1), sendo que nesses casos houve uma transformação para 0.995, para garantir que todas as observações estejam no intervalo aberto $(0,1)$. Valores maiores para cada item indicam que o município tem melhores condições em relação às dimensões avaliadas e valores menores indicam piores condições. Note que há alguns indicadores que, em sua forma original, possuem uma interpretação contrária à proposta, com valores maiores indicando piores situações do município. Quando isso acontece, consideramos nessa aplicação os itens como 1 menos o indicador, assegurando que todos os itens sejam maiores para municípios que possuem bons resultados para as dimensões avaliadas. Os itens são 1) DEP: 1 - porcentagem da população dependente (menor que 15 anos ou maior que 65); 2)S60: probabilidade de sobrevivência até os 60 anos; 3) ANA: 1 - taxa de analfabetismo; 4) MED: porcentagem de maiores de idade que completaram o ensino médio; 5) GINI: 1 - indicador de Gini; 6) PPOB: 1 - porcentagem de pessoas vulneráveis à pobreza; 7) SMIN: 1 - porcentagem de maiores de idade que ganham menos que um salário mínimo; 8) DESE: 1 - porcentagem de desempregados; 9) BAGUA: porcentagem de pessoas que moram em domicílios com banheiros e água corrente; 10) LIXO: porcentagem de pessoas que moram em domicílios com coleta de lixo; 11) LUZ: porcentagem de pessoas que vivem em domicílios com eletricidade; 12) SANI: 1 - porcentagem de pessoas que vivem em domicílios sem saneamento básico; 13) EFPO: 1 - porcentagem de pessoas que vivem em domicílios vulneráveis à pobreza e nos quais ninguém

Tabela 19 – Performance para a recuperação do parâmetro ϕ_0 ao longo das 100 réplicas

			Grupos de itens pelo número de atributos requeridos					
			Um		Dois		Três	
J	Discriminação	N	RMSE	VR	RMSE	VR	RMSE	VR
15	Baixa	250	0.572	0.06	0.252	0.02	0.219	0.01
		500	0.409	0.04	0.224	0.01	0.147	0.00
		1000	0.255	0.01	0.138	0.00	0.128	0.01
	Média	250	0.380	0.02	0.237	0.01	0.219	0.01
		500	0.260	0.00	0.181	0.00	0.156	0.01
		1000	0.162	-0.01	0.123	0.00	0.098	0.00
	Alta	250	0.300	0.00	0.251	0.02	0.193	0.00
		500	0.236	0.01	0.150	0.00	0.151	0.01
		1000	0.160	0.00	0.118	0.00	0.105	0.00
30	Baixa	250	0.289	0.03	0.154	0.01	0.154	0.01
		500	0.201	0.02	0.133	0.01	0.116	0.00
		1000	0.085	0.00	0.067	0.00	0.058	0.00
	Média	250	0.212	0.02	0.179	0.01	0.148	0.01
		500	0.142	0.01	0.120	0.01	0.110	0.00
		1000	0.101	0.01	0.087	0.00	0.082	0.00
	Alta	250	0.218	0.02	0.159	0.01	0.138	0.00
		500	0.142	0.01	0.109	0.01	0.099	0.00
		1000	0.107	0.01	0.064	0.00	0.075	0.01

completou o ensino fundamental; 14) CRIA: 1 - porcentagem de mulheres de 10 a 17 anos que tiveram crianças; 15) CHEF: 1 - porcentagem de mães que são chefes de família sem ter completado o ensino fundamental e tem filhos abaixo dos 15 anos; 16) ESTR: 1 - porcentagem de pessoas de 15 a 24 anos que nem estudam nem trabalham; 17) IDHE: dimensão de educação do IDH; 18) IDHS: dimensão de saúde do IDH; 19) IDHR: dimensão de renda do IDH. A Matriz $\boldsymbol{Q}$, a qual foi especificada considerando as principais dimensões de índices como o IDH, analisando empiricamente quais dessas dimensões são avaliadas por cada um dos indicadores do estudo, pode ser vista em 22, sendo definidas como econômicas (α_1), de saúde (α_2) e de educação (α_3).

Os resultados para os parâmetros estimados e as medidas de distância também são mostrados na Tabela 22, com os resultados dos parâmetros ϕ apresentados como a raíz quadrada. Podemos ver que a maior parte das questões possui, ao menos, uma diferenciação consideravelmente boa entre os grupos η, exceto pelos itens de número 8 e 12, com o item de número 5 também tendo uma diferenciação menor. É interessante notar que dois desses três itens são relacionados à dimensão econômica e um deles à de saúde, com itens relacionados à educação tendo boa diferenciação entre os grupos para todos os itens. O item 14, relacionado à saúde, o qual considera as mulheres menores de idade que tem filhos é um dos que possuem pior diferenciação entre os grupos, tendo valores altos tanto para μ_0 quanto para μ_1 e, também, possuindo valores

Tabela 20 – Performance para a recuperação do parâmetro ϕ_1 ao longo das 100 réplicas

			Grupos de itens pelo número de atributos requeridos					
			Um		Dois		Três	
J	Discriminação	*N*	RMSE	VR	RMSE	VR	RMSE	VR
15	Baixa	250	1.307	0.14	4.555	0.53	11.338	1.43
		500	0.712	0.08	1.811	0.21	4.699	0.53
		1000	0.440	0.04	0.853	0.07	1.872	0.19
	Média	250	0.735	0.02	1.058	0.04	1.857	0.12
		500	0.470	0.01	0.804	0.02	1.214	0.06
		1000	0.326	0.00	0.477	0.01	0.672	0.02
	Alta	250	0.825	0.03	1.048	0.02	1.572	0.05
		500	0.579	0.01	0.799	0.03	1.176	0.03
		1000	0.359	0.00	0.464	0.00	0.690	0.01
30	Baixa	250	0.688	0.07	1.848	0.21	5.194	0.62
		500	0.372	0.03	0.694	0.07	1.508	0.17
		1000	0.140	0.02	0.238	0.04	0.379	0.05
	Média	250	0.445	0.02	0.743	0.04	1.107	0.07
		500	0.355	0.01	0.533	0.02	0.678	0.04
		1000	0.200	0.00	0.264	0.01	0.498	0.02
	Alta	250	0.490	0.01	0.885	0.04	1.324	0.07
		500	0.404	0.02	0.522	0.02	0.832	0.04
		1000	0.256	0.00	0.371	0.01	0.439	0.01

Tabela 21 – Performance para a recuperação do parâmetro $\boldsymbol{\alpha}$ ao longo das 100 réplicas

J	Discriminação	*N*	SA	VA
15	Baixa	250	68.84	20.51
		500	69.63	21.41
		1000	70.06	22.49
	Média	250	90.78	67.41
		500	90.95	67.88
		1000	91.25	68.95
	Alta	250	98.13	91.81
		500	98.19	92.08
		1000	98.22	92.20
30	Baixa	250	78.15	36.76
		500	79.39	39.75
		1000	79.94	41.37
	Média	250	97.72	90.88
		500	97.91	91.53
		1000	97.89	91.50
	Alta	250	99.88	99.44
		500	99.88	99.43
		1000	99.89	99.43

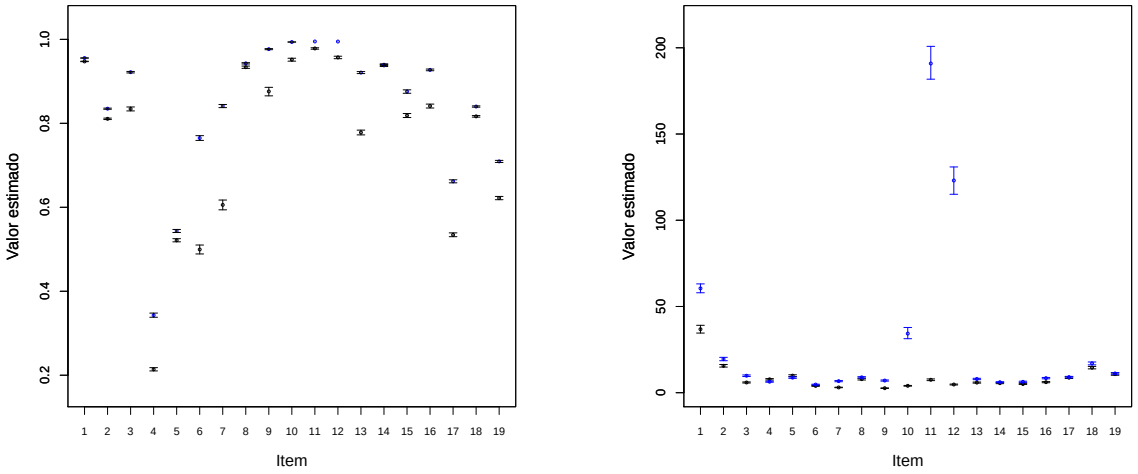
não tão grandes para ϕ_0 e ϕ_1. Também é bastante interessante notar que a diferenciação dos itens é influenciada tanto pelos valores de μ quanto pelos valores dos parâmetros ϕ. Por exemplo, o item 12, relacionado à saúde, o qual diz respeito ao saneamento básico é o item com maior diferenciação, tanto considerando o KS quanto o m_j, possuindo valores altos tanto para μ_0 quanto para μ_1, o que resulta em baixos valores de λ_j, no entanto, como o valor de ϕ_1 é bastante grande, a diferenciação do item é boa. É importante salientar que menores valores de ϕ fazem com que a distribuição seja menos dispersa e maiores valores fazem com que a distribuição se concentre ao redor da média. Para esse item, especificamente, os respondentes no grupo $\eta = 1$ tem uma precisão bastante alta (ou seja, dispersão bastante pequena), fazendo com que a distribuição para $\eta = 1$ seja mais concentrada. O mesmo acontece para outros itens com discriminação alta entre os grupos. Porém, também há casos nos quais itens com boa discriminação entre os grupos possuem maiores valores de λ_j e não tem valores tão altos para os parâmetros ϕ, bem como ocorre no item 6, relacionado à vulnerabilidade a pobreza.

Tabela 22 – Matriz **Q** para os aspectos sociais avaliados, estimativas dos parâmetros dos itens e medidas de distância e discriminação.

Item	α_1	α_2	α_3	$\hat{\mu}_0$	$\sqrt{\hat{\phi}_0}$	$\hat{\mu}_1$	$\sqrt{\hat{\phi}_1}$	KS	m_j	λ_j
1. DEP	1	0	0	0.947	36.823	0.955	60.512	**0.62**	**0.33**	0.008
2. S60	0	1	0	0.811	15.612	0.835	19.532	0.42	0.14	0.024
3. ANA	1	0	1	0.834	5.935	0.922	9.847	**0.69**	**0.40**	0.087
4. MED	1	0	1	0.214	7.769	0.343	6.373	**0.69**	**0.41**	0.129
5. GINI	1	0	0	0.521	9.889	0.543	8.739	0.17	0.03	0.022
6. PPOB	1	0	0	0.500	3.939	0.765	4.727	**0.78**	**0.53**	0.265
7. SMIN	1	0	0	0.606	3.048	0.841	6.701	**0.75**	**0.49**	0.235
8. DESE	1	0	0	0.933	7.867	0.943	8.924	0.12	0.01	0.009
9. BAGUA	1	1	0	0.876	2.627	0.977	7.054	0.56	0.27	0.101
10. LIXO	1	1	0	0.952	4.006	0.994	34.377	**0.71**	**0.43**	**0.042**
11. LUZ	1	0	0	0.979	7.505	0.995	190.939	**0.80**	**0.57**	**0.017**
12. SANI	0	1	0	0.958	4.756	0.995	123.062	**0.85**	**0.57**	0.038
13. EFPO	1	0	1	0.778	5.869	0.921	7.951	**0.84**	**0.63**	0.143
14. CRIA	0	1	1	0.938	5.533	0.940	6.108	0.02	0.00	0.002
15. CHEF	0	0	1	0.818	5.035	0.876	6.320	0.34	0.10	0.057
16. ESTR	1	0	1	0.841	6.113	0.927	8.453	**0.67**	**0.38**	0.086
17. IDHE	0	0	1	0.534	8.705	0.662	9.035	**0.75**	**0.49**	0.127
18. IDHS	0	1	0	0.817	14.530	0.840	16.996	0.37	0.11	0.023
19. IDHR	1	0	0	0.622	10.766	0.709	11.168	**0.69**	**0.41**	0.087

Também podemos avaliar as estimativas intervalares, por meio dos intervalos HPD ao nível de 95% para todos os parâmetros de item μ_{j0}, μ_{j1}, ϕ_{j0} e ϕ_{j1}. A Figura 12 mostra esses intervalos, sendo interessante ressaltar que, para os parâmetros μ, há comprimentos pequenos para diversos itens e que não há intersecção para a maioria dos itens, com exceção do item 14, apontando na mesma direção do que as medidas de discriminação.

A Figura 13 mostra os valores p Bayesianos, com base na média *a posteriori*, para cada item considerado na aplicação. É interessante notar que os valores estão próximos à 0.5, com todos os itens apresentando valor p dentro do intervalo $(0.1, 0.9)$, indicando bom ajuste dos



(a) Parâmetros μ (b) Raiz dos parâmetros ϕ

Figura 12 – Intervalos HPD dos parâmetros de item. Linhas pretas representam grupo $\eta = 0$ e azuis grupo $\eta = 1$.

dados por meio do B-DINA, sendo que o item com valor mais distante de 0.5 é o item 10, o qual apresenta valor p de 0.290, ainda assim mostrando bom ajuste aos dados.

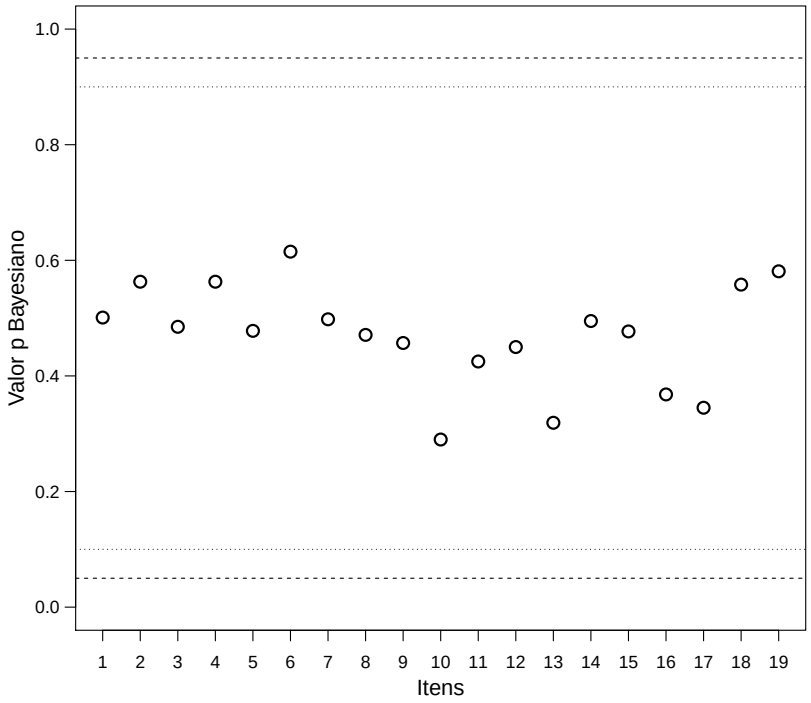


Figura 13 – Valor p Bayesiano por item (modelo B-DINA) para dados sócio-econômicos na região Sudeste do Brasil, com base na média à posteriori.

A Tabela 23 mostra as estimativas das médias da posteriori para a probabilidade de pertencer à cada perfil de atributo, bem como os intervalos HPD. É interessante notar que,

apesar de todos os perfis possuírem ao menos um atributo dentro deles, os dois perfis com maiores proporções são aqueles nos quais os municípios possuem bons resultados para todas as dimensões avaliadas (c= 8, 40.5% dos municípios) ou não possuem bons resultados para nenhuma das dimensões (c= 1, 27.9% dos municípios). Esses resultados são interessantes, uma vez que mostram que a maior parte dos municípios analisados que possuem problemas, tem resultados que indicam que tais problemas se relacionam à todas as dimensões avaliadas e, mesmo para aqueles municípios que não possuem problemas em todas as dimensões, por vezes há mais de uma dimensão com maus resultados. Considerando os outros perfis, aquele com a maior proporção é $c = 6$, o qual representa ter bons resultados para as dimensões econômicas e de educação, mas não para as de saúde, contendo cerca de 15.7% dos municípios.

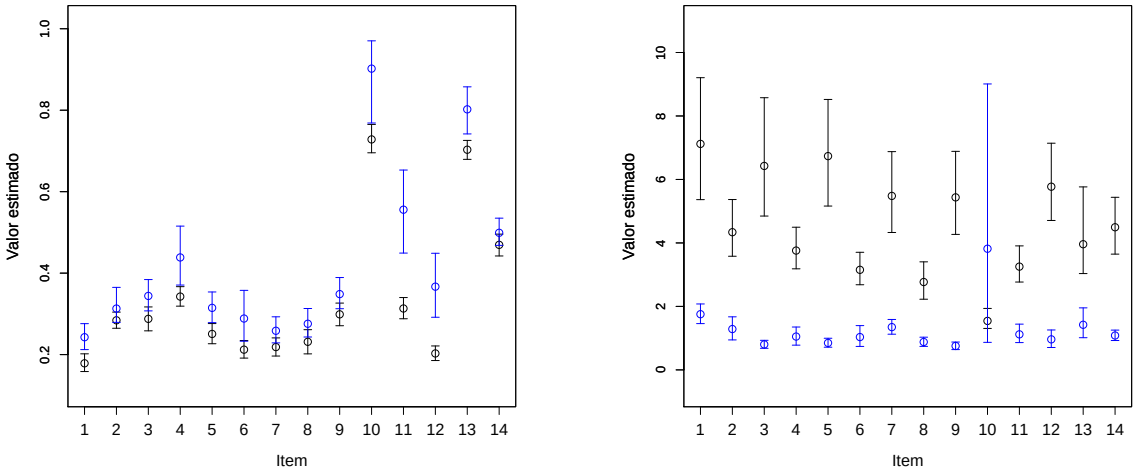
Tabela 23 – Probabilidade de pertencer a cada classe de perfil de atributos para as três dimensões avaliadas.

c perfis de atributo	Dimensões			$\hat{\pi}_c$	
	α_1	α_2	α_3	Média	HPD (95%)
1	0	0	0	0.279	(0.256, 0.302)
2	1	0	0	0.059	(0.045, 0.075)
3	0	1	0	0.029	(0.018, 0.040)
4	0	0	1	0.010	(0.005, 0.017)
5	1	1	0	0.055	(0.043, 0.068)
6	1	0	1	0.157	(0.139, 0.176)
7	0	1	1	0.004	(0.001, 0.009)
8	1	1	1	0.407	(0.407, 0.432)

4.6 Aplicação 2: Dados sobre percepção de risco

Para a segunda aplicação são considerados os mesmos dados que na Seção 3.5. Como agora estamos no intervalo $(0,1)$, as respostas utilizadas naquela aplicação são divididas por 100, sendo que respostas iguais à 1 tem o valor transformado para 0.999.

A Matriz $\boldsymbol{Q}$ é a mesma considerada na aplicação do C-DINA, com os resultados considerando o B-DINA sendo mostrados na Tabela 24, com base na média *a posteriori* dos parâmetros de item dos grupos latentes $\eta = 0$ e $\eta = 1$, bem como as medidas de discriminação KS e m_j, além da diferença entre os parâmetros de locação dos dois grupos, $\lambda_j = \mu_{j1} - \mu_{j0}$. Podemos notar que BRCA (item 3), DIAB (item 5) e HEART (item 9) continuam se destacando em relação aos outros itens no que diz respeito aos valores de m_j. Ainda é notavel que, no caso da aplicação por meio do B-DINA, há alguns itens que se destacam quanto ao KS, mas não quanto ao m_j, bem como PLANE (Item 11) e SWAT (Item 13). O item DOC (Item 6) continua sendo aquele que possui menor valor tanto de m_j quanto de KS. Vale ressaltar que nenhum item foi considerado como possuindo alta discriminação ao utilizar o B-DINA, assim como ocorria no caso do ajuste



(a) Parâmetros μ (b) Parâmetros ϕ

Figura 14 – Intervalos HPD dos parâmetros de item. Linhas pretas representam grupo $\eta = 0$ e azuis grupo $\eta = 1$.

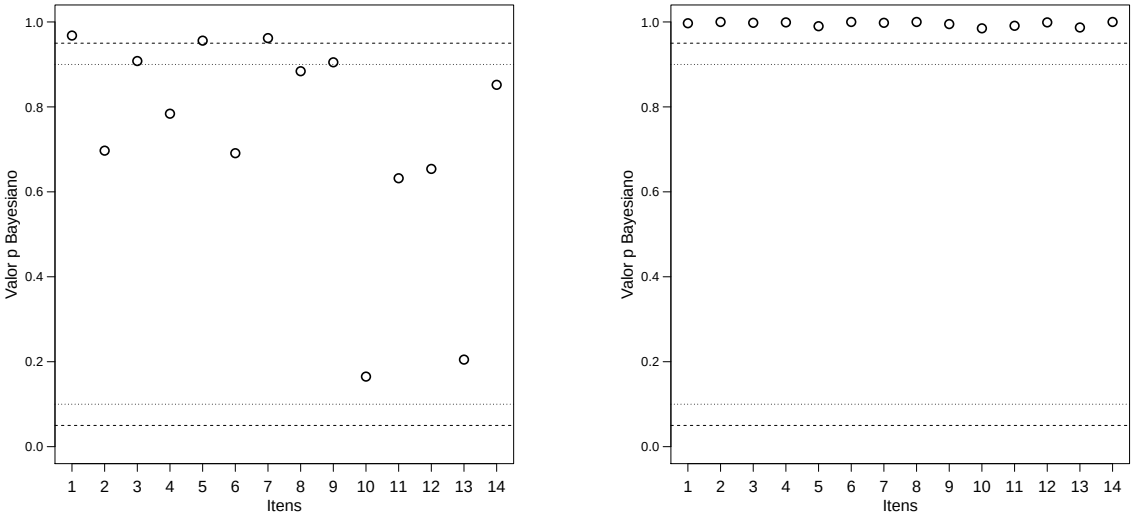
pelo C-DINA, porém, agora vemos que a discriminação dos itens é considerada como baixa para todos os itens.

Tabela 24 – Matriz **Q** do conjunto de dados de percepção de risco, estimativas pelas médias à posteriori para os parâmetros dos itens, e valores de medidas de distância e discriminação dos itens para o modelo B-DINA.

Item	α_1	α_2	$\hat{\mu}_0$	$\hat{\phi}_0$	$\hat{\mu}_1$	$\hat{\phi}_1$	KS	m_j	λ_j
1. APPL	0	1	0.179	7.121	0.243	1.755	0.17	0.12	0.064
2. BIKE	1	1	0.284	4.339	0.313	1.287	0.19	0.11	0.029
3. BRCA	0	1	0.288	6.426	0.344	0.802	0.30	**0.26**	0.056
4. CAR	1	1	0.342	3.760	0.439	1.051	0.23	0.12	0.096
5. DIAB	0	1	0.251	6.736	0.315	0.850	0.30	**0.25**	0.064
6. DOC	1	0	0.212	3.153	0.289	1.035	0.15	0.08	0.076
7. FLU	0	1	0.219	5.482	0.259	1.348	0.22	0.13	0.040
8. GREY	0	1	0.231	2.769	0.276	0.879	0.19	0.10	0.044
9. HEART	0	1	0.299	5.436	0.349	0.754	0.30	**0.25**	0.050
10. NUC	1	0	0.728	1.543	0.902	3.818	0.29	0.09	**0.174**
11. PLANE	1	0	0.313	3.254	0.556	1.120	**0.37**	0.15	0.243
12. POOL	1	1	0.203	5.773	0.367	0.962	0.29	0.19	0.164
13. SWAT	1	0	0.703	3.960	0.802	1.422	**0.37**	0.15	0.099
14. XRAY	0	1	0.469	4.494	0.499	1.084	0.22	0.15	0.030

As estimativas intervalares, considerando os intervalos HPD de 95%, podem ser vistas na Figura 14. Nota-se que há intersecção entre os intervalos dos parâmetros do grupo $\eta = 0$ e $\eta = 1$ para os itens 2, 3, 6, 7, 8, 9 e 14.

É interessante notar, na Figura 15, que os itens são melhor ajustados pelo B-DINA do que eram pelo C-DINA. Ao utilizarmos o C-DINA para os dados, os valores p Bayesiano estão



(a) B-DINA (b) C-DINA

Figura 15 – Valor p Bayesiano por item (modelo B-DINA e modelo C-DINA) para dados de percepção de risco, com base na média à posteriori.

acima de 0.95, para todos os itens. Já para o B-DINA, há alguns items com valores acima de 0.90 (Item 1, 0.968; Item 3, 0.908; Item 5, 0.956; Item 7, 0.962; Item 9, 0.905) e apenas os itens $1, 5$ e 7 apresentam valor p acima de 0.95.

Em relação aos parâmetros dos indivíduos, a Tabela 25 mostra que há consideráveis diferenças em relação aos resultados obtidos pela aplicação do C-DINA. O B-DINA indica que o perfil de atributos mais prevalente é o primeiro, no qual os indivíduos não vêem riscos consideráveis, ao contrário do que o C-DINA indicava, para o qual o perfil que via ambas as condições como arriscadas era mais prevalente. Esses resultados discordantes mostram a importância de que o modelo se adeque bem aos dados, e, além disso, também resalta o quão fundamental é que os itens sejam bem formulados, possuindo boa discriminação entre os grupos, principalmente no que diz respeito à estimativa dos parâmetros dos indivíduos.

Tabela 25 – Classificação: probabilidade de pertença à cada um dos perfis de atributo para o modelo B-DINA.

c perfis de atributo		Dimensões		$\widehat{\pi}_c$	
		α_1	α_2	Média	HPD (95%)
1	(não vêem riscos consideráveis)	0	0	0.48	$(0.406, 0.546)$
2	(vêem condições externas como mais arriscadas)	1	0	0.02	$(0.003, 0.051)$
3	(vêem condições internas como mais arriscadas)	0	1	0.30	$(0.219, 0.372)$
4	(vêem ambas as condições como arriscadas)	1	1	0.20	$(0.135, 0.304)$

Os resultados dessa aplicação indicam que o uso do B-DINA traz respostas com melhor ajuste para os dados do que o do C-DINA, segundo o valor p Bayesiano. No entanto, nota-se

que a discriminação entre os itens não é tão boa quanto a que obtivemos para a os dados sobre indicadores sócio-econômicos na região Sudeste. A diferenciação entre os grupos propostos é menos evidente na Aplicação 2, referente ao questionário sobre percepção de riscos do que na aplicação mostrada na Seção 4.5. Isso pode trazer problemas, principalmente em relação ao ajuste dos parâmetros dos indivíduos, como também pode ser visto nos estudos de simulação para ambos os modelos.

Com isso, temos indicativos de que seria interessante explorar esses dados novamente, no futuro, por meio de MDCs desenvolvidos com outras distribuições de probabilidade, as quais possam assumir valores nos extremos, ou seja, permitam trabalhar no intervalo fechado $[0, 1]$, o que permitiria uma análise mais aprofundada sobre tais dados, dando mais insumos em relação ao ajuste e até mesmo à análise sobre a adequabilidade desse questionário para a separação entre os dois grupos propostos.

4.7 Discussão

A proposta do B-DINA permite o uso de um Modelo de Diagnóstico Cognitivo para diferentes tipos de resposta além daqueles que já se encontram na literatura. Esse trabalho é uma primeira abordagem do uso desse tipo de modelo para variáveis contínuas limitadas, trazendo a importância de haver tais modelos para esse tipo de dado.

É bastante notável o quão sensível o B-DINA é em relação à qualidade das questões e, também, que questionários com baixo número de respondentes e poucos itens fazem com que a estimação dos parâmetros do modelo seja mais difícil, principalmente para parâmetros dos indivíduos.

A primeira aplicação mostra o potencial de utilizção dos MDCs, os quais, além de serem úteis em situações variadas que fujam do escopo educacional, como comentado nos capítulos anteriores, também vai além dos questionários respondidos por pessoas. Esses modelos podem ser aplicados em diversas realidades que possam ser mensuradas por meio de itens, tais quais municípios que possuem diferentes indicadores, relacionados à aspectos distintos.

A segunda aplicação mostra que ainda há espaço para o desenvolvimento de novos MDCs para respostas limitadas, considerando outras distribuições, desde que elas também permitam o ajuste e a interpretação dos parâmetros. Trabalhos que considerem ditribuições que se adequem ao intervalo fechado $[0, 1]$ seriam de particular interesse.

Trabalhos futuros podem explorar mais a fundo os casos de baixo número de respondentes. A diminuição no tempo computacional também seria uma contribuição interessante em trabalhos futuros.

Considerando a abordagem Bayesiana, no futuro, outros métodos MCMC podem ser explorados. Um estudo considerando o impacto de diferentes quantidades de dimensões K

também seria interessante.

Outro trabalho interessante seria a avaliação do impacto de uma boa especificação da Matriz **Q** para modelos DINA de resposta contínua limitada.

CAPÍTULO

5

NOVAS DISTRIBUIÇÕES DE RESPOSTA LIMITADA

Quando estamos interessados em saber como covariáveis impactam diferentes níveis da variável resposta, modelos de regressão quantílica podem ser bastante úteis. O uso de variáveis contínuas limitadas para tais modelos é bastante comum quando os dados contêm porcentagem, proporções ou taxas. No trabalho apresentado nesse capítulo, novos modelos paramétricos de regressão quantílica com efeitos mistos são propostos, considerando variáveis respostas limitadas com caudas pesadas. Tendo a distribuição de Gompertz como base, são derivadas duas novas distribuições bi-paramétricas com suporte limitado, bem como suas versões quantílicas. A estimação para os parâmetros é realizada por meio de uma abordagem Bayesiana. Resultados de estudos de simulação são reportados, mostrando que os modelos e métodos inferenciais propostos trazem boa recuperação dos parâmetros. Além disso, um conjunto de dados relacionado à pobreza extrema no Peru é analisado, utilizando modelos de regressão quantílica com efeitos fixos e aleatórios. Os modelos propostos nesse capítulo enriquecem o ferramental de alternativas para a modelagem e análise de dados limitados.

Por mais que alguns dos desenvolvimentos não tenham relação direta com o tópico de MDC, as ideias presentes nessa parte da tese podem ser futuramente utilizadas, por exemplo, para propor novos modelos para respostas limitadas dentro da classe dos MDCs, algo que, conforme ressaltado no capítulo anterior, pode ser bastante interessante no avanço do estado da arte dos MDCs.

O trabalho desenvolvido nesse capítulo também será submetido a um periódico especializado.

5.1 Introdução

Diversos modelos de regressão para variáveis limitadas foram introduzidos ao longo dos últimos anos, como pode ser visto em Ferrari e Cribari-Neto (2004b), Lemonte e Bazán (2016), Smithson e Shou (2017) e Migliorati *et al.* (2018), entre outros, e, em particular, modelos com efeitos mistos para respostas limitadas também foram introduzidos na literatura (QIU; SONG; TAN, 2008; FIGUEROA-ZÚÑIGA; ARELLANO-VALLE; FERRARI, 2013). Tais modelos podem ser úteis, por exemplo, para analisar dados de medidas repetidas ou dados agrupados, mas não estão restritos à tais casos. A popularidade desses modelos pode ser explicada pela flexibilidade existente para modelar a correlação entre os sujeitos e lidar tanto com dados balanceados quanto desbalanceados. Usualmente os autores focam na relação entre algumas covariáveis e a média condicional de uma variável resposta limitada, dados alguns efeitos fixos e aleatórios relacionados às covariáveis.

Modelos de regressão quantílica com efeitos mistos (GERACI; BOTTAI, 2014; YU; LU; STANDER, 2003)) podem ter grande utilidade na modelagem da relação entre covariáveis e os quantis condicionais da variável resposta dada tais covariáveis e efeitos fixos e/ou aleatórios, também tendo a capacidade de trazer uma imagem mais completa e complexa da distribuição condicional da variável resposta.

Com motivação em um conjunto de dados sobre pobreza extrema no Peru, duas novas distribuições são propostas para dados limitados, derivadas da distribuição Gompertz (LENART, 2014), a qual possui expressões simples para a fda e a para a obtenção de funções de quantil, facilitando a formulação de um modelo paramétrico de regressão quantílica.

O principal objetivo desse capítulo é propor novas distribuições e modelos de regressão quantílica com efeitos mistos baseados nessas distribuições, as quais podem modelar dados assimétricos e com caudas pesadas. Os modelos propostos são convenientes para variáveis resposta assimétricas e concentradas em um dos extremos do intervalo, podendo também lidar com dados aumentados seja em 0 ou 1.

O capítulo está organizado da seguinte maneira: Na Seção 5.2 as distribuições propostas nesse trabalho e uma reparametrização com base nos quantis é introduzida, a fim de formular os modelos de regressão quantílica com efeitos mistos. Na Seção 5.3 uma abordagem Bayesiana é proposta para a estimação dos parâmetros do modelo. Na Seção 5.4 são apresentados resultados para um estudo de simulação e na Seção 5.5 dados reais sobre pobreza extrema no Peru são analisados. Por fim, considerações finais são feitas na Seção 5.6.

5.2 Novas distribuições para dados limitados

As novas distribuições para variáveis limitadas são construídas com base na distribuição de Gompertz (LENART, 2014).

Definição 5.2.1. *Se $X \sim Gompertz(a,c)$, então a fdp e a fda são dadas, respectivamente, por $f_X(x|a,c) = c\exp(ax)\exp\{-c[\exp(ax)-1]/a\}$ e $F_X(x|a,c) = 1-\exp\{-c[\exp(ax)-1]/a\}$, para $x \geq 0$, onde $a \neq 0$ é um parâmetro de forma e $c > 0$ um parâmetro de locação.*

Da Definição 5.2.1, a fdp de duas distribuições assimétricas com suporte limitado pode ser derivado por meio de transformações não lineares. Podemos definir a variável aleatória $Y = \exp(-X)$, com suporte no intervalo $(0,1]$ quando $a > 0$, a qual segue a distribuição LogGompertz (daqui em diante LG), com fdp e fda dadas por

$$f_{\mathrm{LG}}(y|a,c) = cy^{-a-1}\exp[-c(y^{-a}-1)/a] \quad \text{e} \quad F_{\mathrm{LG}}(y|a,c) = \exp[-c(y^{-a}-1)/a]. \tag{5.1}$$

Por outro lado, se Y segue a distribuição LG, a variável aleatória $1-Y$ segue a distribuição Complementar LogGompertz (daqui para frente CLG), com suporte no intervalo $[0,1)$ quando $a > 0$. A fdp e a fda da distribuição CLG são denotadas por $f_{\mathrm{CLG}}(y|a,c)$ e $F_{\mathrm{CLG}}(y|a,c)$, respectivamente.

De (5.1) temos que $\lim_{y\downarrow 0} f_{\mathrm{LG}}(y|a,c) = 0$, $f_{\mathrm{LG}}(1|a,c) = c$, $f_{\mathrm{CLG}}(0|a,c) = c$ e $\lim_{y\uparrow 1} f_{\mathrm{CLG}}(y|a,c) = 0$. As distribuições LG e CLG propostas podem acomodar caudas pesadas à direita ou à esquerda. Enfatizamos que a função de densidade de distribuições bem como a Beta, por exemplo, ou são 0 ou vão para ∞ nos extremos do intervalo unitário.

Como pode ser visto em Tsodikov, Ibrahim e Yakovlev (2003), se $X \sim \mathrm{Gompertz}(a,c)$ e $a < 0$, então $\lim_{x\to\infty} F_X(x|a,c) = \exp(c/a)$, de tal maneira que a distribuição é defeituosa. Assim sendo, se $Y \sim \mathrm{LG}(a,c)$ e $a < 0$, então $y = 0$ tem probabilidade $\exp(c/a)$. Por outro lado, se $Y \sim \mathrm{CLG}(a,c)$ e $a < 0$, então $y = 1$ tem probabilidade $\exp(c/a)$. Dessa maneira, as distribuições LG e CLG, no caso em que $a < 0$, podem acomodar dados aumentados para 0 e 1, respectivamente, se tornando fechadas em ambos os extremos do intervalo unitário. As duas novas distribuições podem ainda ser estendidas para qualquer intervalo (A_1, A_2), com $A_1 < A_2$. Uma vez que ambas as distribuições são bastante conectadas, a apresentação do capítulo irá focar na distribuição LG.

Algumas propriedades de nossa distribuição bi paramétrica também estão presentes em algumas outras distribuições para dados limitados na literatura. Por exemplo, a distribuição bi paramétrica em Jodrá (2020) se baseia na curva de Gompertz deslocada, enquanto o modelo em Migliorati *et al.* (2018) é construído com base na distribuição *Flexible* Beta de quatro parâmetros e o modelo em Ghitany *et al.* (2019) é baseado na distribuição Gaussiana Inversa. No entanto, assim como ocorre com as distribuições Beta e Simplex (QIU; SONG; TAN, 2008), as funções quantílicas em Jodrá (2020), Migliorati *et al.* (2018) e Ghitany *et al.* (2019) não são simples, fazendo com que essas distribuições não sejam convenientes para a proposta de modelos de regressão quantílica, como o que propomos nesse capítulo. Além disso, essas distribuições não têm a flexibilidade de ajustar dados no intervalo unitário fechado.

Outra distribuição que tem propriedades similares à proposta dessa capítulo é a Unit-Gompertz (MAZUCHELI; MENEZES; DEY, 2019). Entretanto, em Mazucheli, Menezes e Dey

(2019) uma definição e notação diferentes da distribuição de Gompertz são utilizadas, com o caso com valores negativos para o parâmetro a sendo desconsiderado. Além disso, não há a proposta de modelos de regressão e nem de alguma abordagem Bayesiana.

Reparametrização

Podemos encontrar uma expressão simples para a função quantílica da distribuição LG, dada por $\kappa(q) = F_{\text{LG}}^{-1}(q|a,c) = [1 - a\log(q)/c]^{-1/a}$, para $0 < q < 1$. Em particular, a mediana é dada por $\kappa(0.5) = [1 + \log(2)a/c]^{-1/a}$.

A fim de propor um modelo de regressão quantílica, para q fixado, a distribuição pode ser reparametrizada em termos do q-ésimo quantil $\kappa(q)$ e do parâmetro a. Assim, obtemos uma estrutura mais apropriada para a regressão, substituindo $c = a\log(q)/(1-\kappa^{-a})$ em (5.1) como a nova parametrização. Nesse caso, q é assumido conhecido e o espaço paramétrico de $(\kappa, a)^T$ é dado por $((0,1) \times \mathbb{R} - \{0\})^T$.

A fdp e a fda da distribuição LG reparametrizada são dadas, respectivamente, por

$$f_{\text{LG}}(y|\kappa,a,q) = [a\log(q)/(1-\kappa^{-a})]y^{-a-1}\exp\{[-\log(q)/(1-\kappa^{-a})](y^{-a}-1)\}$$

e $F_{\text{LG}}(y|\kappa,a,q) = \exp\{[-\log(q)/(1-\kappa^{-a})](y^{-a}-1)\}$.

Iremos adotar as notações $Y \sim \text{LG}(\kappa,a,q)$ e $Y \sim \text{CLG}(\kappa,a,q)$, com o parâmetro quantílico $0 < \kappa < 1$ associado à uma probabilidade fixa de interesse q e $a \neq 0$ como parâmetro de locação (LENART, 2014), para o caso sem dados aumentados. Para $Y \sim LG(\kappa,a,q)$, se $b < a+1$, onde $b = a\log(q)/(1-\kappa^{-a})$, a moda y_0 é dada por $y_0 = [(a+1)/(ac)]^{-1/a}$, com $c = \log(q)/(1-\kappa^{-a})$. Se $b \geq a+1$, $y_0 = 1$.

A Figura 16 mostra a fdp da versão reparametrizada da distribuição LG para diferentes valores de κ e a. São considerados o primeiro decil, a mediana e o último decil. Quando κ é fixado, notamos que a de fato é um parâmetro que controla a forma da distribuição. Para valores maiores de a também se observa menor dispersão, com a distribuição sendo mais concentrada em torno da moda. Por outro lado, quando a é fixado, nota-se que κ age como um parâmetro que controla a locação da distribuição. Por exemplo, para maiores valores de κ a moda tende a se mover para a direita. Uma vez que κ é o q-ésimo quantil de Y, ele pode ser interpretado como um parâmetro de locação.

Também é interessante notar que a distribuição LG é assimétrica à direita e é capaz de ajustar dados com respostas iguais à 1. Note que esse ainda não é o caso com dados aumentados no extremo do intervalo unitário, como em Ospina e Ferrari (2012) e Harris e Zhao (2007), as quais levam em conta o caso de dados aumentados, mas não consideram situações de comportamento assimétrico com valores assumidos nos extremos, isso é, os casos nos quais a função não decai para 0 (decaindo para valores positivos ao lado direito) ou para 1 (decaindo para valores menores

do lado esquerdo). A distribuição LG acomoda esse caso e, também, pode se ajustar a diferentes tipos de dados, sendo bastante flexível. Quando $a > 0$, tanto dados assimétricos à direita com suporte em $(0,1]$ ou $(0,1)$ podem ser ajustados por meio da distribuição LG. Além disso, quando $a < 0$, como pode ser visto na Seção 5.2, a probabilidade para $y = 0$ é positiva, sendo possível ajustar dados que sejam tanto assimétricos à direita quanto zero aumentados, com suporte em $[0,1]$ ou $[0,1)$, fazendo com que essa distribuição tenha utilidade em vários casos diferentes.

A respeito da interpretação dos parâmetros, os momentos da distribuição também são úteis. Apesar de não ser possível uma avaliação analítica dos momentos, é possível os avaliar por meio de análises gráficas. Para a distribuição LG, o r-ésimo momento, $r = 1,2,\ldots$ é dado por

$$E\left[Y^r\right] = ac\int_0^1 y^{n-a-1}\exp\left[-c(y^{-a}-1)\right]dy, \quad c = \frac{\log(q)}{1-k^{-a}.} \tag{5.2}$$

Para solucionar as integrais na Equação (5.2), pode ser utilizada a integração numérica, a fim de avaliar momentos importantes como $E[Y]$ e $E[Y^2]$. Com tais momentos, também é possível avaliar a variância $Var(Y)$ e a precisão $\tau(Y) = 1/Var(Y)$. Na Figura 17a pode-se ver que o valor esperado é bastante relacionado com o parâmetro κ, dando outro forte argumento sobre o fato de κ ser um parâmetro de locação, aumentando quando a média da variável resposta aumenta. No que diz respeito ao parâmetro a, a Figura 17b mostra que a precisão aumenta quando o parâmetro a tem valores maiores, trazendo evidências de que esse parâmetro possui relações com a dispersão e, uma vez que valores maiores de a tornam a distribuição mais precisa, a é inversamente proporcional à dispersão. Ainda nessa Figura, é possível ver que menores valores de κ potencializam a precisão conforme a se torna maior.

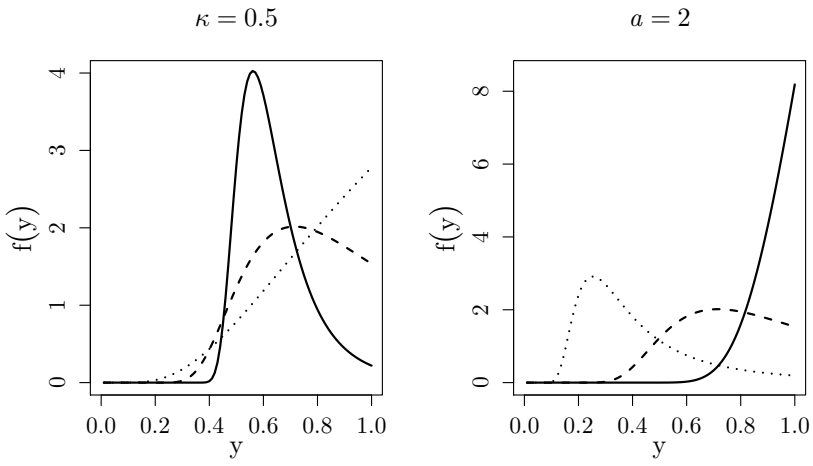
5.3 Modelos mistos

Para introduzir os modelos de regressão com efeitos mistos, que chamaremos de modelos mistos, é necessário primeiro definir algumas quantidades. Tome $\boldsymbol{Y}_i = (y_{i1},\ldots,y_{in_i})^T$ como o vetor de n_i respostas no intervalo unitário para a i-ésima unidade amostral. Para cada unidade amostral, nós introduzimos os efeitos aleatórios (WU, 2009) a fim de levar em conta a dependência entre as medidas que se repetem dentro da unidade amostral. Usando essa abordagem, é possível modelar tanto a variabilidade entre unidades diferentes como dentro das unidades.

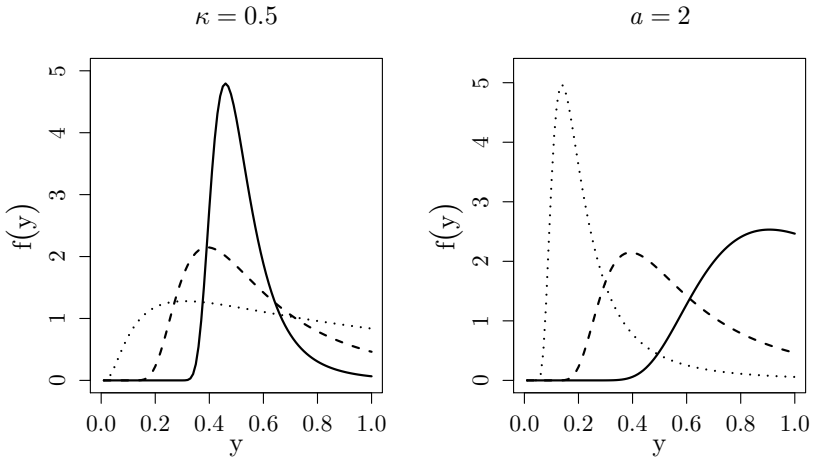
Os modelos de regressão mista baseada nos parâmetros de locação (quantil) e de forma são especificados como

$$\begin{aligned}
&y_{ij} \overset{\text{indep.}}{\sim} \text{LG}(\kappa_{ij}, a_{ij}, q) \quad \text{ou} \quad y_{ij} \overset{\text{indep.}}{\sim} \text{CLG}(\kappa_{ij}, a_{ij}, q),\\
&g_1(\kappa_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i, \quad g_2(a_{ij}) = -\boldsymbol{w}_{ij}^T\boldsymbol{\delta} - \boldsymbol{h}_{ij}^T\boldsymbol{d}_i,\\
&\boldsymbol{b}_i \overset{\text{indep.}}{\sim} \text{N}_p(\boldsymbol{0},\boldsymbol{\Sigma}_b) \quad \text{e} \quad \boldsymbol{d}_i \overset{\text{indep.}}{\sim} \text{N}_r(\boldsymbol{0},\boldsymbol{\Sigma}_d),
\end{aligned} \tag{5.3}$$

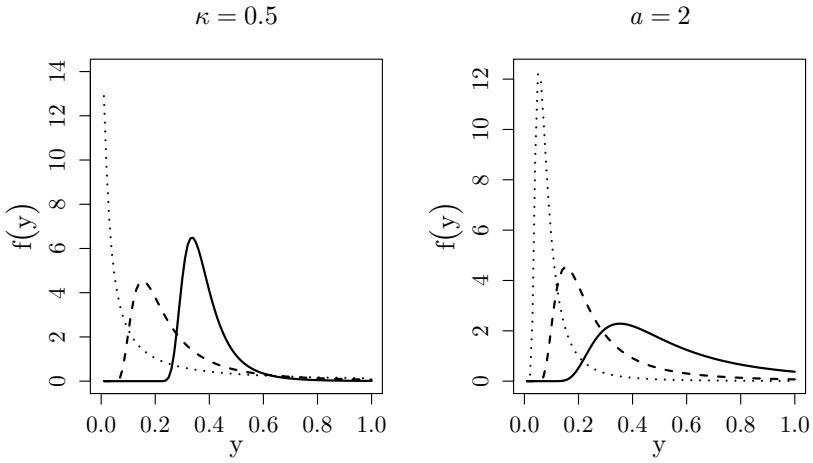
para $j = 1,\ldots,n_i$ e $i = 1,\ldots,n$, onde $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_k)^T$ é o vetor de efeitos fixos dos coeficientes da regressão associados ao parâmetro de locação e $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_l)^T$ é o vetor de efeitos fixos



(a) Quantil 0.1

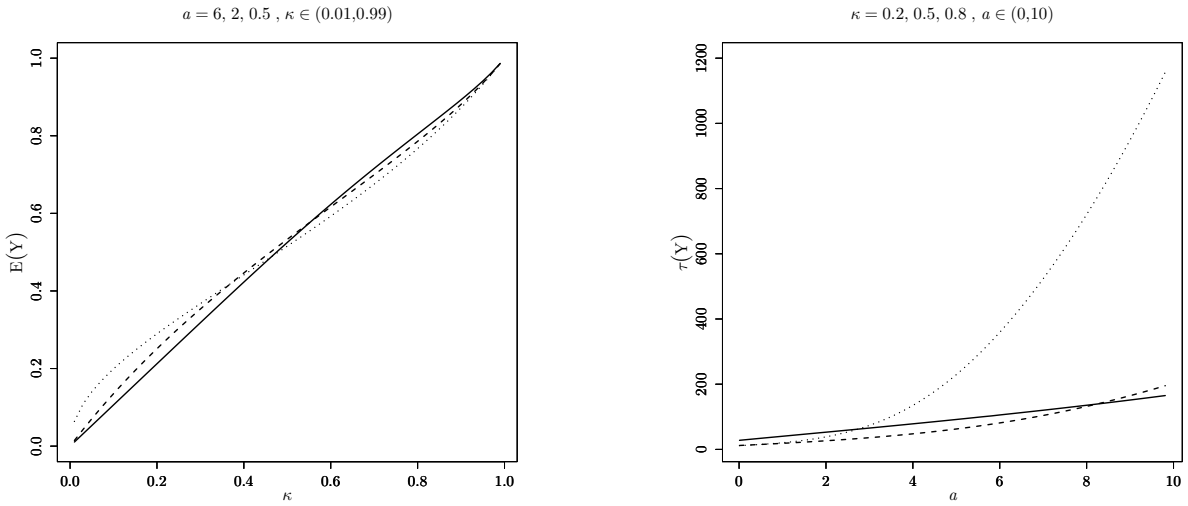


(b) Quantil 0.5



(c) Quantil 0.9

Figura 16 – fdp da distribuição LG para diferentes valores de κ e a. Painel esquerdo: $\kappa = 0.5$ e diferentes valores de a: 6 (linha sólida), 2 (linha tracejada), e 0.5 (linha pontilhada). Painel direito: $a = 2$ e diferentes valores de κ: 0.8 (linha sólida), 0.5 (linha tracejada), e 0.2 (linha pontilhada).



(a) (b)

Figura 17 – Média da LG (a) para diferentes valores de κ e a, com $q = 0.5$. Valores fixados de a variando κ; a: 6 (linha sólida), 2 (linha tracejada), e 0.5 (linha pontilhada) e dispersão (b) para diferentes valores de κ e a, com $q = 0.5$. Valores fixos de κ variando a; κ: 0.8 (linha sólida), 0.5 (linha tracejada), e 0.2 (linha pontilhada).

dos coeficientes da regressão associados ao parâmetro de forma. Os efeitos aleatórios dos parâmetros de locação e de forma são denotados, respectivamente, por $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{ip})^T$ e $\boldsymbol{d}_i = (d_{i1}, \ldots, d_{ir})^T$. Em (5.3), $N_s(\boldsymbol{m}, \boldsymbol{\Sigma})$ denota uma distribuição Normal s-variada com vetor de média $\boldsymbol{m}$ e matriz de covariância definida positiva $\boldsymbol{\Sigma}$. Uma vez que a dispersão diminui quando o valor de a aumenta, tomamos o sinal negativo no preditor linear para facilitar a interpretação dos coeficientes. Além disso, $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijk})^T$, $\boldsymbol{w}_{ij} = (w_{ij1}, \ldots, w_{ijl})^T$, $\boldsymbol{z}_{ij} = (z_{ij1}, \ldots, z_{ijp})^T$ e $\boldsymbol{h}_{ij} = (h_{ij1}, \ldots, h_{ijr})^T$ são vetores de covariáveis, os quais não precisam ser necessariamente idênticos e podem se sobrepor. Por fim, $q \in (0,1)$ é uma probabilidade fixada, correspondente ao quantil de interesse, e, para a função de ligação $g_1(\cdot)$ e $g_2(\cdot)$ em (5.3) assumimos as funções logística e logarítmica, respectivamente.

Sob a parametrização na subseção 5.2, a função aumentada de verossimilhança do modelo é dada por

$$L(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{d}; \boldsymbol{Y}) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} f(y_{ij} | \kappa_{ij}, a_{ij}) \phi_p(\boldsymbol{b}_i | \boldsymbol{0}, \boldsymbol{\Sigma}_b) \phi_r(\boldsymbol{d}_i | \boldsymbol{0}, \boldsymbol{\Sigma}_d) \tag{5.4}$$

onde $f(\cdot)$ e κ denotam a fdp e o parâmetro do quantil para as distribuições LG e CLG, $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_n^T)^T$, $\boldsymbol{d} = (\boldsymbol{d}_1^T, \ldots, \boldsymbol{d}_n^T)^T$ e $\boldsymbol{\theta}$ encapsulam $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, $\boldsymbol{\Sigma}_b$ e $\boldsymbol{\Sigma}_d$. Além disso, $\text{logito}(\kappa_{ij}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{z}_{ij}^T \boldsymbol{b}_i$, $\log(a_{ij}) = -\boldsymbol{w}_{ij}^T \boldsymbol{\delta} - \boldsymbol{h}_{ij}^T \boldsymbol{d}_i$ e $\phi_s(\cdot | \boldsymbol{m}, \boldsymbol{S})$ denota a fdp da distribuição $N_s(\boldsymbol{m}, \boldsymbol{\Sigma})$. O modelo apresentado nessa seção é um modelo geral, que pode ser chamado de modelo de regressão quantílica com efeitos mistos, sendo que modelos particulares podem ser obtidos, com efeitos fixos, efeitos aleatórios ou ambos os efeitos, fixos e aleatórios.

Inferência Bayesiana

A função de verossimilhança para o modelo de regressão quantílica com efeitos mistos é dada em (5.4). A distribuição *a posteriori* aumentada de $\boldsymbol{\theta}$, $\boldsymbol{b}$, e $\boldsymbol{d}$, denotada por $p(\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{d}|\boldsymbol{Y})$, é $p(\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{d}|\boldsymbol{Y}) \propto L(\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{d};\boldsymbol{Y})p(\boldsymbol{\theta})$, onde $p(\boldsymbol{\theta})$ denota a distribuição *a priori* de $\boldsymbol{\theta}$. Para completar a especificação Bayesiana do modelo, nós também assumimos que os elementos do vetor de parâmetros são, *a priori*, independentes de tal forma que

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta})p(\boldsymbol{\delta})p(\boldsymbol{\Sigma}_b)p(\boldsymbol{\Sigma}_d). \tag{5.5}$$

Distribuições Normais Multivariadas são propostas como prioris para os efeitos fixos, isto é, $\boldsymbol{\beta} \sim \mathrm{N}_k(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ e $\boldsymbol{\delta} \sim \mathrm{N}_l(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{\delta}})$, onde $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ e $\boldsymbol{\Sigma}_{\boldsymbol{\delta}}$ são matrizes positivas definidas. Para as matrizes de covariâncias dos efeitos aleatórios, distribuições Wishart Invertidas (IW, do inglês *Inverse Wishart*) podem ser adotadas, com $\boldsymbol{\Sigma}_b \sim \mathrm{IW}_p(\psi_b,\boldsymbol{\Psi}_b)$ e $\boldsymbol{\Sigma}_d \sim \mathrm{IW}_r(\psi_d,\boldsymbol{\Psi}_d)$. Os hiperparâmetros são ψ_b, $\boldsymbol{\Psi}_b$, ψ_d e $\boldsymbol{\Psi}_d$, onde $\boldsymbol{\Psi}_b$ e $\boldsymbol{\Psi}_d$ também são matrizes definidas positivas. A média da distribuição $IW_p(\psi_b,\boldsymbol{\Psi}_b)$ é $\boldsymbol{\Psi}_b/(\psi_b - p - 1)$, onde p é a dimensão de $\boldsymbol{\Sigma}_b$.

Podemos obter a distribuição *a posteriori* combinando a função de verossimilhança em (5.4) com as prioris em (5.5), resultando em

$$\begin{aligned} p(\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{d}|\boldsymbol{Y}) \propto & \prod_{i=1}^{n}\prod_{j=1}^{n_i} f(y_{ij}|\kappa_{ij},a_{ij})\phi_p(\boldsymbol{b}_i|\mathbf{0},\boldsymbol{\Sigma}_b)\phi_r(\boldsymbol{d}_i|\mathbf{0},\boldsymbol{\Sigma}_d)\phi_k(\boldsymbol{\beta}|\mathbf{0},\boldsymbol{\Sigma}_b)\phi_l(\boldsymbol{\delta}|\mathbf{0},\boldsymbol{\Sigma}_d) \\ & \times f_{IW}(\boldsymbol{\Sigma}_b|\psi_b,\boldsymbol{\Psi}_b)f_{IW}(\boldsymbol{\Sigma}_d|\psi_d,\boldsymbol{\Psi}_d), \end{aligned} \tag{5.6}$$

onde $f_{IW}(\cdot|\psi,\boldsymbol{\Psi})$ denota a fdp da distribuição Wishart Invertida.

A distribuição *a posteriori* em (5.6) não é analiticamente tratável. Entretanto, uma aproximação pode ser obtida por meio de métodos de Monte Carlo via Cadeias de Markov (MCMC), obtendo amostras da posteriori. A implementação desse capítulo se vale do pacote "RStan" (Stan Development Team, 2014) no R, utilizando o algoritmo NUTS (HOFFMAN; GELMAN, 2014). Os códigos se encontram no Apêndice.

Comparação de modelos

Há diversos critérios que podem ser utilizados para comparar como diferentes modelos se ajustam a um conjunto de dados, auxiliando na decisão de qual dos modelos testados é melhor para cada ocasião. Nesse capítulo adotamos o critério de informação de Watanabe-Akaike (WAIC) (WATANABE, 2010b). O WAIC pode ser visto como uma aproximação do critério de validação cruzada (GELMAN; HWANG; VEHTARI, 2014). Dadas as estimativas a posterior, $\widehat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{\theta}}^T,\widehat{\boldsymbol{b}}^T,\widehat{\boldsymbol{d}}^T)^T$, o valor de WAIC é computado como $\mathrm{WAIC} = -2\mathrm{lppd} + 2p_{\mathrm{WAIC}}$, onde $\mathrm{lppd} = \sum_{i=1}^{n}\log\int p(y_i|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|y)d\boldsymbol{\vartheta}$ é o logaritmo da densidade preditiva (lppd, do inglês *log pointwise predictive density* e a penalização $p_{\mathrm{WAIC}} = \sum_{i=1}^{n}\mathrm{Var}_{\mathrm{post}}\big(\log[p(y_i|\boldsymbol{\vartheta})]\big)$, com $\mathrm{Var}_{\mathrm{post}}$ sendo a variância dos termos individuais no logaritmo da densidade preditiva, termo que é

considerado como uma aproximação do número de parâmetros irrestritos e não informados, sendo computada a partir da saída das amostras MCMC. Para um conjunto de modelos candidatos, aquele como o menor valor de WAIC é considerado como aquele que melhor ajusta os dados segundo esse critério.

Os valores de WAIC foram computados por meio do pacote "loo" no R (VEHTARI; GELMAN; GABRY, 2017). Merkle, Furr e Rabe-Hesketh (2018) aponta que a log-verossimilhança condicional é uma boa escolha se o intuito do modelo é específico para os grupos (ou clusters) observados, enquanto a log-verossimilhança marginal deve ser utilizada caso a intenção seja generalizar as conclusões do modelo para outros grupos fora daqueles constantes nos dados analisados. Em nosso caso, a proposta é fazer a inferência unicamente para os grupos contidos no conjunto de dados, de tal forma que a função de log-verossimilhança condicional foi utilizada para computar o WAIC para os modelos mistos.

5.4 Estudo de Simulação

Nessa seção, são conduzidos estudos de simulação a fim de avaliar a estimação dos parâmetros para o modelo LG sob diferentes cenários. Primeiramente, são considerados cenários onde o parâmetro a é maior do que 0, sendo também analisados posteriormente cenários com $a < 0$, levando ao caso de dados simulados zero-aumentados.

Dados no intervalo (0,1)

Nessa seção um primeiro estudo de simulação é realizado, a fim de verificar a recuperação de parâmetros de nosso modelo e comparar sua performance com outras propostas, com base nos critérios de comparação de modelos previamente mostrados, para o caso quando $a > 0$. Para tal, consideramos um modelo de regressão com covariáveis e um parâmetro de quantil, simulando então a variável resposta no intervalo $(0,1)$ sob um modelo com $y_i|\boldsymbol{\beta}, a \sim \text{LG}(\kappa_i, a, q)$, onde $\text{logito}(\kappa_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$, $i = 1, \ldots, n$, $\log(a) = -\delta_0$ e utilizando como parâmetros $\boldsymbol{\beta} = (1, 4, -3)^T$ e $-\delta_0 = \log(2)$, com $q \in \{0.1, 0.5, 0.9\}$. Para as covariáveis, são gerados n covariáveis independentes com $x_{ik} \sim \text{Uniform}(0,1)$, $k = 2, 3$ e $x_{i1} = 1$ para o intercepto. O estudo de simulação compreende nove cenários e 100 réplicas geradas para cada um dos cenários.

Uma vez que os dados simulados são obtidos, os modelos de regressão para o quantil 0.1, para a mediana e para o quantil 0.9 da variável resposta no intervalo unitário aberto com distribuição LG foram ajustados. Isto é, os modelos são ajustados considerando $y_i \overset{\text{indep.}}{\sim} \text{LG}(\kappa_i, a, q)$, para $i = 1, \ldots, n$, com $q = 0.1$, $q = 0.5$ e $q = 0.9$ e n sendo o número de observações para cada cenário.

Adicionalmente ao modelo LG, foram ajustados modelos alternativos, seguindo a distri-

buição Kumaraswamy (BAYES; BAZÁN; CASTRO, 2017), onde

$$f_{\text{Kumaraswamy}}(y|\kappa,\varphi) = \frac{\log(1-q)\varphi}{\log(1-\exp^{-\varphi})\log(\kappa)} y^{-\frac{\varphi}{\log(\kappa)}-1}\left[1-y^{-}\frac{\varphi}{\log(\kappa)}\right]^{\frac{\log(1-q)}{(1-\exp^{-\varphi})}-1},$$

com $y \in (0,1), \kappa \in (0,1), \varphi > 0$ a distribuição Beta, utilizando a parametrização proposta por Ferrari e Cribari-Neto (2004b), onde

$$f_{\text{Beta}}(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},$$

com $y \in (0,1), \mu \in (0,1), \phi > 0$ e a distribuição Beta Retangular (BRr) (BAYES; BAZÁN; GARCÍA, 2012), onde

$$f_{\text{BRr}}(y|\mu,\phi) = \alpha(1-|2\gamma-1|) + (1-\alpha(1-|2\gamma-1|)) f_{\text{Beta}}\left(y \;\middle|\; \frac{\gamma-0.5\alpha(1-|2\gamma-1|)}{1-\alpha(1-|2\gamma-1|)}, \phi\right),$$

com $y \in (0,1), \gamma \in [0,1], \alpha \in [0,1], \phi > 0$.

Foram ajustados modelos de regressão quantílica Kumaraswamy (BAYES; BAZÁN; CASTRO, 2017), com $q = 0.1$, $q = 0.5$ e $q = 0.9$, respectivamente, onde $y_i \overset{\text{indep.}}{\sim} \text{Kumaraswamy}(\kappa_i, \varphi, q)$, com $\text{logito}(\kappa_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$, e $\log(\varphi) = -\delta_0$, tendo como prioris $\boldsymbol{\beta} \sim \text{N}_3(\mathbf{0}, 10^4 \boldsymbol{I}_3)$ e $\delta_0 \sim \text{N}(0, 10^4)$. Além disso, modelos de regressão beta sobre a média são ajustados, utilizando a parametrização proposta por Ferrari e Cribari-Neto (2004b), i.e., $y_i \overset{\text{indep.}}{\sim} \text{Beta}(\mu_i, \phi)$, com $\text{logito}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ e $\phi = \exp(-\delta_0)$, tendo como prioris $\boldsymbol{\beta} \sim \text{N}_3(\mathbf{0}, 10^4 \boldsymbol{I}_3)$ e $\phi \sim \text{Gama Inversa}(0.01, 0.01)$ como em Figueroa-Zúñiga, Arellano-Valle e Ferrari (2013). Por fim, são ajustados modelos de regressão Beta Retangular (BRr) (BAYES; BAZÁN; GARCÍA, 2012), isto é, $y_i \overset{\text{indep.}}{\sim} \text{BRr}(\gamma_i, \phi, \alpha)$, com $\text{logito}(\gamma_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ e $\log(\phi) = -\delta_0$ tendo prioris $\boldsymbol{\beta} \sim \text{N}_3(\mathbf{0}, 10^4 \boldsymbol{I}_3)$, $\delta_0 \sim \text{N}(0, 10^4)$ e $\alpha \sim \text{Unif}(0,1)$, similar ao que se encontra em Bayes, Bazán e García (2012). Uma vez que os modelos Beta e Beta Retangular são ajustados sobre a média, os seus resultados são mais comparados com o caso quando $q = 0.5$ em relação aos com outros quantis, de tal maneira que eles são ajustados para os dados gerados com $q = 0.5$ e comparados aos resultados dos modelos LG e Kumaraswamy de regressão na mediana.

Para as regressões LG, as distribuições *a priori* foram especificadas como $\boldsymbol{\beta} \sim \text{N}_3(\mathbf{0}, 10^4 \boldsymbol{I}_3)$ e $\delta_0 \sim \text{N}(0, 10^4)$, onde $\boldsymbol{I}_3$ denota a a matriz unitária 3×3. Essa priori foi testada contra a priori Gama Inversa para $q = 0.5$ e $n = 500$. A média do WAIC ao longo das 100 replicas foi menor utilizando a priori Log-Normal (-1019.5 versus -1007.7) e ela também foi escolhida mais frequentemente (57 vs 43 replicas), de maneira similar ao que ocorre em Bayes, Bazán e Castro (2017) para o modelo de regressão quantílica Kumaraswamy.

Para todos os modelos, foram utilizados códigos em Stan, assim como feito para as distribuições LG e CLG. Para cada uma das réplicas, foram realizadas 2000 iterações, descartando as primeiras 1000. A convergência foi checada por meio da estatística de Gelman-Rubin (GELMAN; RUBIN *et al.*, 1992) e por inspeção gráfica.

Em geral, na Tabela 26, estimativas *a posteriori* menos viesadas e mais precisas são encontradas quando o modelo de regressão LG é ajustado, para todos os valores de q, como esperávamos. A única exceção é a raiz do erro quadrático médio (RMSE) para β_0, quando $q = 0.5$ da regressão Beta sobre a média, mas, mesmo nesse caso, os valores do desvio padrão e do RMSE são próximos quando a regressão LG na mediana é ajustada.

Para os modelos de regressão LG na mediana, a probabilidade de cobertura (CP, do inglês *coverage probability*) dos intervalos de credibilidade bilaterais a nível de 95% variam entre 0.89 e 0.98 para os componentes de $\boldsymbol{\beta}$, enquanto para todos os demais modelos de regressão, CP pode ser até mesmo igual a 0.00. Para o parâmetro a, os resultados da recuperação para os modelos de regressão LG são adequados para todos os cenários, mas tem resultados um pouco inferiores quando $q = 0.1$ e $n = 100$.

Os melhores resultados, como esperávamos, são encontrados quando o modelo é ajustado seguindo a distribuição da geração de dados, sendo que modelos mal especificados são amplamente superados pelos modelos de regressão LG. Esses resultados mostram que, em geral, o método de estimação proposta recupera adequadamente os parâmetros e também mostra que o modelo de regressão Beta sobre a média, o modelo de regressão beta Retangular sobre a média e o modelo Kumaraswamy de regressão quantílica podem ser inadequados para esse tipo de dados, com cauda mais pesada.

A Tabela 27 mostra as médias de WAIC e p_{WAIC} , bem como o número de vezes que cada modelo é selecionado usando o WAIC para cada uma das 100 réplicas. É notável que o WAIC seleciona o verdadeiro modelo de regressão em todas as réplicas geradas para $q = 0.1$ e $q = 0.9$. Para $q = 0.5$ é interessante notar que o aumento no tamanho amostral faz com que o modelo LG seja selecionado mais frequentemente pelo critério WAIC. Além disso, a média dos valores de WAIC diminuem quando o tamanho amostral aumenta e os valores médios de p_{WAIC} estão bem mais próximos do número de parâmetros quando o modelo verdadeiro (LG) é ajustado. No geral, concluímos que o WAIC é um bom critério de seleção para os modelos sob os cenários em nosso estudo.

Dados no intervalo [0,1)

Nessa subseção outro estudo de simulação é conduzido, agora para o caso no qual o parâmetro a assume valor negativo. Nesse caso, os dados gerados aleatoriamente com base em uma distribuição LG terão valores iguais à 0, com probabilidade $p(y = 0) = \exp(c/a)$, como dito na Seção 5.2. Novamente, considerados um modelo de regressão com covariáveis no parâmetro κ e então a variável resposta é simulada. Agora, os valores simulados estão no intervalo $[0,1)$, sendo gerados sob um modelo com $y_i|\boldsymbol{\beta}, a \sim \text{LG}(\kappa_i, a, q)$, sendo $\text{logito}(\kappa_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$, $i = 1, \ldots, n$ e $\log(-a) = \delta_0$.

Nesse caso, os dados são zero-aumentados e o aumento nos valores iguais à zero no

Tabela 26 – Resumos da posteriori para diferentes modelos para os dados gerados sob o modelo de regressão LG com a positivo (Est: média da média da posteriori, DP: média do desvio padrão da posteriori, CP: probabilidade de cobertura do intervalo de credibilidade bilateral 95% e RMSE: raiz do erro quadrático médio da média da posteriori).

				Modelo															
				Beta				Beta Retangular				Kumaraswamy				LG			
	n		Valor verdadeiro	Est	DP	CP	RMSE	Est	DP	CP	RMSE	Est	DP	CP	RMSE	Est	DP	CP	RMSE
$q=0.5$	100	β_0	1	0.944	0.214	0.93	0.220	0.819	0.249	0.85	0.307	-0.293	0.169	0.00	1.304	1.077	0.224	0.94	0.236
		β_1	4	2.966	0.348	0.10	1.090	3.143	0.329	0.27	0.917	3.460	0.346	0.87	0.641	3.900	0.319	0.98	0.332
		β_2	-3	-2.277	0.317	0.33	0.789	-2.331	0.310	0.34	0.737	0.734	0.087	0.00	3.735	-3.034	0.274	0.97	0.275
		a	2													1.959	0.334	0.96	0.336
	200	β_0	1	0.933	0.155	0.91	0.168	0.834	0.181	0.73	0.245	-0.195	0.125	0.00	1.201	1.003	0.179	0.94	0.179
		β_1	4	2.973	0.246	0.01	1.056	3.155	0.217	0.03	0.872	3.296	0.230	0.35	0.740	4.007	0.223	0.95	0.223
		β_2	-3	-2.236	0.226	0.08	1.056	-2.332	0.228	0.15	0.706	0.711	0.058	0.00	3.712	-3.002	0.206	0.97	0.206
		a	2													1.986	0.251	0.96	0.251
	500	β_0	1	0.924	0.093	0.88	0.120	0.836	0.010	0.61	0.192	-0.160	0.080	0.00	1.162	1.023	0.118	0.92	0.120
		β_1	4	2.928	0.172	0.00	1.086	3.111	0.155	0.00	0.903	3.425	0.157	0.09	0.596	3.961	0.152	0.89	0.156
		β_2	-3	-2.179	0.129	0.00	0.831	-2.293	0.117	0.00	0.716	0.675	0.034	0.00	3.676	-2.997	0.144	0.95	0.144
		a	2													1.958	0.189	0.93	0.194
$q=0.1$	100	β_0	1									-1.401	0.231	0.00	2.412	0.966	0.255	0.86	0.256
		β_1	4									4.601	0.454	0.88	0.753	4.078	0.405	0.87	0.410
		β_2	-3									2.119	0.101	0.00	5.120	-3.078	0.344	0.87	0.351
		a	2													1.604	0.808	0.83	0.899
	200	β_0	1									-1.253	0.163	0.00	2.259	0.992	0.171	0.94	0.171
		β_1	4									4.341	0.300	0.90	0.453	4.027	0.218	0.95	0.218
		β_2	-3									2.109	0.070	0.00	5.110	-3.016	0.219	0.93	0.220
		a	2													1.964	0.440	0.94	0.441
	500	β_0	1									-1.207	0.109	0.00	2.210	1.000	0.095	0.98	0.095
		β_1	4									4.424	0.203	0.53	0.469	4.022	0.140	0.94	0.141
		β_2	-3									2.081	0.043	0.00	5.081	-3.013	0.131	0.95	0.131
		a	2													2.003	0.328	0.95	0.328
$q=0.9$	100	β_0	1									0.430	0.239	0.38	0.617	0.996	0.256	0.95	0.256
		β_1	4									2.328	0.246	0.00	1.690	3.953	0.351	0.97	0.353
		β_2	-3									0.349	0.088	0.00	3.350	-2.984	0.306	0.94	0.306
		a	2													2.032	0.195	0.98	0.198
	200	β_0	1									0.547	0.177	0.32	0.486	0.978	0.180	0.97	0.181
		β_1	4									2.211	0.162	0.00	1.796	4.015	0.223	0.98	0.223
		β_2	-3									0.312	0.054	0.00	3.312	-3.008	0.223	0.95	0.223
		a	2													2.046	0.128	0.98	0.136
	500	β_0	1									0.640	0.102	0.12	0.374	0.982	0.118	0.98	0.119
		β_1	4									2.330	0.107	0.00	1.673	4.029	0.149	0.93	0.151
		β_2	-3									0.270	0.029	0.00	3.270	-3.016	0.130	0.95	0.130
		a	2													2.037	0.106	0.94	0.112

conjunto de dados faz com que a estimação dos parâmetros seja mais difícil, uma vez que os dados se tornam bastante esparsos. Dito isso, a fim de avaliar a robustez da estimação para diferentes números de zeros, são avaliados múltiplos cenários, gerando conjuntos de dados simulados sob diferentes condições. Para os casos com menos zeros, os parâmetros são definidos como $\boldsymbol{\beta} = (1,4,-3)^T$ e $\delta_0 = \log(2)$, com $q \in \{0.1, 0.5, 0.9\}$. Para os casos com mais valores iguais à zero, tomamos com parâmetros $\boldsymbol{\beta} = (-1,-4,3)^T$ e $\delta_0 = \log(2)$. Quinze diferentes cenários foram considerados para cada um dos dois casos que levam em conta diferentes quantidades de zeros e um cenário extra com valor amostral maior de tamanho n foi incluído para $q = 0.9$, com 100 réplicas para cada cenário.

Uma vez que os dados simulados são gerados, modelos de regressão para os quantis 0.1, para a mediana e para o quantil 0.9 da variável resposta no intervalo $[0,1)$ são ajustados. Isto é, temos que $y_i \overset{\text{indep.}}{\sim} \text{LG}(\kappa_i, a, q)$, para $i = 1, \ldots, n$ com $q = 0.1$, $q = 0.5$ e $q = 0.9$. A distribuição *a priori* é especificada como $\boldsymbol{\beta} \sim \text{N}_3(\mathbf{0}, 10^4 \boldsymbol{I}_3)$ e $\delta_0 \sim \text{N}(0, 10^4)$. São realizadas 2000 iterações, descartando as primeiras 1000 para cada réplica.

A Tabela 28 mostra que, no caso com menos zeros, para $q = 0.5$, a probabilidade

Tabela 27 – Média de p_{WAIC}, WAIC e frequência de seleção baseada no WAIC para as 100 réplicas geradas sob o modelo de regressão quantílica (Beta: modelo de regressão Beta sobre a média, Beta Retangular: modelo de regressão Beta Retangular sobre a média, LG: regressão quantílica LG e Kumaraswamy: regressão quantílica Kumaraswamy)

	n	Modelo	p_{WAIC}	WAIC	Frequência
$q = 0.5$	100	Beta	4.6	-137.4	5
		Beta Retangular	5.4	-140.9	6
		Kumaraswamy	2.8	-87.6	0
		LG	3.7	-163.0	89
	200	Beta	4.8	-303.7	2
		Beta Retangular	5.5	-310.5	3
		Kumaraswamy	2.7	-213.2	0
		LG	3.8	-355.0	95
	500	Beta	5.1	-866.5	0
		Beta Retangular	5.9	-883.8	1
		Kumaraswamy	2.9	-652.6	0
		LG	4.0	-1019.5	99
$q = 0.1$	100	Kumaraswamy	2.9	-195.3	0
		LG	3.6	-265.1	100
	200	Kumaraswamy	2.8	-451.7	0
		LG	3.8	-586.8	100
	500	Kumaraswamy	3.1	-1298.7	0
		LG	4.0	-1636.6	100
$q = 0.9$	100	Kumaraswamy	3.1	-36.8	0
		LG	3.8	-126.0	100
	200	Kumaraswamy	3.0	-75.2	0
		LG	3.9	-246.2	100
	500	Kumaraswamy	3.0	-211.6	0
		LG	4.0	-634.0	100

de cobertura dos intervalos de credibilidade bilaterais de 95% variam entre 0.90 e 0.97, com estimativas próximas aos valores reais. No entanto, é interessante notar que embora a recuperação de parâmetro seja adequada para todos os valores de n para os componentes de $\boldsymbol{\beta}$, ela se torna melhor com o incremento em n, com estimativas ainda mais próximas aos valores reais e menores valores de DP e RMSE para $n = 500$ ou maior.

Quando diferentes valores de q são considerados, a recuperação dos componentes dos parâmetros $\boldsymbol{\beta}$ se torna pior, principalmente para $n = 100$. Considerando $q = 0.1$, a recuperação de parâmetros foi adequada para $n = 200$ ou maior, mas não foi para $n = 100$, com estimativas distantes dos valores reais (tome, por exemplo, β_0, o quai possui valor real igual à 1 e valor estimado igual a -0.257). Para $q = 0.9$, a recuperação de parâmetros não é adequada para $n = 100$ e mesmo para $n = 200$ e $n = 500$ os valores recuperados não estão tão próximos aos

reais. Para $n = 2000$ a recuperação é adequada. Também vale notar que para o parâmetro a a recuperação é boa, mesmo para valores menores de n.

A Tabela 29 mostra resultados para o caso quando há mais valores iguais à 0 para a variável resposta. Como esperado, a estimação dos parâmetros se torna mais desafiadora, uma vez que os dados são bastante esparsos e concentrados em 0, principalmente para valores maiores de q (note que para $q = 0.9$ a proporção de respostas iguais à 0 é próxima de 88%. Para todos os valores de q, a recuperação dos componentes de $\boldsymbol{\beta}$ não é adequada para $n = 500$ ou menor. Para $q = 0.1$ e $q = 0.5$, quando $n = 2000$, os resultados começam a estar mais próximos dos valores reais, entretanto, resultados adequados são obtidos apenas para $n = 5000$ ou maior, mostrando que a estimação para cenários com mais valores no limite do intervalo unitário é, de fato, mais difícil. Para $q = 0.9$, as respostas se tornem extremamente esparsas, com apenas cerca de 10% dos valores diferentes de 0. Nesse caso, assim como também esperado, a recuperação dos componentes de $\boldsymbol{\beta}$ (as quais não são mostradas) não trazem bons resultados, mesmo quando os valores de n aumentam. É interessante notar que, mais uma vez, a recuperação para o parâmetro a é adequada mesmo para valores menores de n.

Tabela 28 – Resumos da posteriori para diferentes modelos para os dados gerados sob o modelo de regressão LG com a negativo, cenário como menos zeros (Est: média da média da posteriori, DP: média do desvio padrão da posteriori, CP: probabilidade de cobertura do intervalo de credibilidade bilateral 95% e RMSE: raiz do erro quadrático médio da média da posteriori).

	n	Proporção de zeros		Valor verdadeiro	Est	DP	CP	RMSE
$q = 0.1$	100	0.0236	β_0	1	-0.132	8.059	0.91	8.138
			β_1	4	4.043	4.982	0.89	4.982
			β_2	-3	-3.985	5.736	0.91	5.782
			a	-2	-2.074	0.710	0.93	0.711
	200	0.0205	β_0	1	0.926	0.324	0.95	0.331
			β_1	4	4.135	0.502	0.93	0.517
			β_2	-3	-3.067	0.430	0.93	0.433
			a	-2	-1.960	0.448	0.92	0.449
	500	0.0183	β_0	1	0.984	0.162	0.99	0.162
			β_1	4	4.045	0.267	0.94	0.270
			β_2	-3	-3.022	0.251	0.94	0.251
			a	-2	-1.974	0.277	0.91	0.279
	2000	0.0172	β_0	1	1.001	0.092	0.97	0.092
			β_1	4	4.009	0.144	0.94	0.144
			β_2	-3	-3.012	0.125	0.93	0.125
			a	-2	-1.992	0.141	0.91	0.141
	5000	0.0171	β_0	1	1.009	0.063	0.92	0.063
			β_1	4	3.991	0.092	0.92	0.092
			β_2	-3	-2.999	0.075	0.97	0.075
			a	-2	-2.011	0.077	0.96	0.077
$q = 0.5$	100	0.2255	β_0	1	0.787	0.559	0.93	0.598
			β_1	4	4.393	0.895	0.90	0.977
			β_2	-3	-3.295	0.810	0.91	0.862
			a	-2	-1.975	0.307	0.95	0.308
	200	0.2027	β_0	1	0.884	0.354	0.94	0.375
			β_1	4	4.214	0.532	0.93	0.573
			β_2	-3	-3.122	0.446	0.93	0.463
			a	-2	-1.959	0.229	0.98	0.233
	500	0.1835	β_0	1	0.956	0.176	0.98	0.181
			β_1	4	4.088	0.284	0.97	0.297
			β_2	-3	-3.053	0.259	0.95	0.264
			a	-2	-1.975	0.156	0.94	0.158
	2000	0.1755	β_0	1	0.997	0.100	0.97	0.100
			β_1	4	4.016	0.158	0.94	0.158
			β_2	-3	-3.018	0.130	0.94	0.130
			a	-2	-2.001	0.078	0.93	0.078
	5000	0.1735	β_0	1	1.011	0.068	0.93	0.068
			β_1	4	3.988	0.101	0.91	0.101
			β_2	-3	-2.996	0.082	0.91	0.082
			a	-2	-2.013	0.049	0.96	0.050
$q = 0.9$	100	0.7138	β_0	1	-1.687	10.267	0.90	10.613
			β_1	4	5.820	5.281	0.85	5.586
			β_2	-3	-4.339	3.727	0.92	3.960
			a	-2	-2.008	0.389	0.96	0.389
	200	0.6855	β_0	1	0.643	0.674	0.96	0.763
			β_1	4	4.515	0.907	0.93	1.043
			β_2	-3	-3.316	0.753	0.94	0.817
			a	-2	-2.004	0.283	0.97	0.283
	500	0.6446	β_0	1	0.856	0.351	0.95	0.379
			β_1	4	4.219	0.508	0.91	0.553
			β_2	-3	-3.144	0.391	0.96	0.416
			a	-2	-2.005	0.208	0.90	0.208
	2000	0.6308	β_0	1	0.998	0.166	0.96	0.166
			β_1	4	4.015	0.221	0.93	0.221
			β_2	-3	-3.021	0.184	0.93	0.185
			a	-2	-2.010	0.089	0.94	0.089
	5000	0.6290	β_0	1	1.021	0.109	0.93	0.110
			β_1	4	3.978	0.148	0.93	0.149
			β_2	-3	-2.995	0.121	0.91	0.121
			a	-2	-2.012	0.053	0.93	0.054

Tabela 29 – Resumos da posteriori para diferentes modelos para os dados gerados sob o modelo de regressão LG com a negativo, cenário como mais zeros (Est: média da média da posteriori, DP: média do desvio padrão da posteriori, CP: probabilidade de cobertura do intervalo de credibilidade bilateral 95% e RMSE: raiz do erro quadrático médio da média da posteriori).

	n	Proporção de zeros		Valor verdadeiro	Est	DP	CP	RMSE
$q = 0.1$	100	0.0705	β_0	-1	-89.691	18.563	0.91	90.594
			β_1	-4	-37.340	9.960	0.99	34.782
			β_2	3	-38.080	15.222	1.00	43.783
			a	-2	-2.328	0.280	0.85	0.430
	200	0.0709	β_0	-1	-87.350	22.729	0.86	89.262
			β_1	-4	-37.512	11.003	0.98	35.255
			β_2	3	-36.453	15.294	0.98	42.286
			a	-2	-2.245	0.220	0.76	0.328
	500	0.0736	β_0	-1	-40.138	45.301	0.85	59.694
			β_1	-4	-22.597	20.942	0.93	27.929
			β_2	3	-12.442	22.923	0.93	27.544
			a	-2	-2.097	0.180	0.77	0.204
	2000	0.0767	β_0	-1	-1.208	0.450	0.95	0.493
			β_1	-4	-4.288	0.794	0.94	0.841
			β_2	3	3.260	0.551	0.96	0.607
			a	-2	-2.002	0.064	0.95	0.064
	5000	0.0772	β_0	-1	-1.019	0.271	0.97	0.271
			β_1	-4	-4.033	0.421	0.96	0.421
			β_2	3	3.013	0.338	0.96	0.338
			a	-2	-2.004	0.038	0.94	0.038
$q = 0.5$	100	0.4315	β_0	-1	-88.391	20.824	0.91	89.813
			β_1	-4	-39.844	15.804	0.96	39.141
			β_2	3	-39.226	20.506	0.96	46.897
			a	-2	-2.045	0.283	0.97	0.285
	200	0.4423	β_0	-1	-89.965	24.171	0.79	92.159
			β_1	-4	-35.221	20.821	0.97	42.901
			β_2	3	-35.741	18.521	0.97	0.463
			a	-2	-2.035	0.194	0.97	0.196
	500	0.4468	β_0	-1	-57.005	45.228	0.76	71.845
			β_1	-4	-28.863	18.522	0.95	30.949
			β_2	3	-19.497	23.002	0.93	32.092
			a	-2	-2.000	0.137	0.96	0.137
	2000	0.4483	β_0	-1	-1.337	0.669	0.95	0.749
			β_1	-4	-4.487	1.110	0.92	1.212
			β_2	3	3.403	0.797	0.96	0.896
			a	-2	-2.000	0.072	0.92	0.072
	5000	0.4476	β_0	-1	-1.035	0.338	0.94	0.340
			β_1	-4	-4.051	0.549	0.95	0.551
			β_2	3	3.030	0.415	0.95	0.416
			a	-2	-2.009	0.044	0.95	0.045

5.5 Aplicação: Dados sobre pobreza extrema no Peru

Nessa seção, são analisados dados sobre a pobreza extrema no Peru. Estamos interessados na relação entre a pobreza extrema (variável resposta) e o índice de desenvolvimento humano (IDH; covariável). O conjunto de dados possui 195 províncias peruanas, agrupadas em 25 departamentos. Departamentos são subdivisões políticas de segundo nível as quais são utilizadas para análises mais detalhadas.

Esse conjunto de dados é proposto pelas informações sobre pobreza extrema em Sociales (2010) e sobre IDH em Perú (2009), os quais também descrevem as metodologias para computar a proporção de pobreza extrema e o IDH. A mediana da pobreza extrema é 0.195, com média 0.217, mínimo 0.0004 e máximo 0.701, Para o IDH, a mediana é 0.570, com média igual a 0.578, mínimo 0.484 e máximo 0.684. A mediana do número de províncias por departamento é de 10, com mínimo igual a 1 (departamento de Callao) e máximo igual a 20 (departamento de Ancash). Nesse caso, temos a variável resposta no intervalo $(0,1)$, de tal forma que uma regressão Beta sobre a média pode ser uma possível escolha. Para esse conjunto de dados, foram ajustados modelos de regressão quantílica LG e CLG, para a mediana da resposta como modelos alternativos, comparando os resultados com um modelo de regressão Beta na média e um modelo Kumaraswamy na mediana.

Modelo de regressão quantílica de locação

São ajustados modelos de regressão quantílica LG e CLG, com $q = 0.5$ (mediana), um modelo de regressão quantílica Kumaraswamy, também com $q = 0.5$ e um modelo de regressão Beta sobre a média. Ou seja, os modelos ajustados são $y_i \overset{\text{indep.}}{\sim} \text{LG}(\kappa_i, a, 0.5)$, $y_i \overset{\text{indep.}}{\sim} \text{CLG}(\kappa_i, a, 0.5)$, $y_i \overset{\text{indep.}}{\sim} \text{Kumaraswamy}(\kappa_i, \varphi, 0.5)$, com $\text{logito}(\kappa_i) = \beta_0 + \beta_1 \text{IDH}_i$, $\log(a) = -\delta_0$, $\log(\varphi) = -\delta_0$ e $y_i \overset{\text{indep.}}{\sim} \text{Beta}(\mu_i, \phi)$, $\text{logito}(\mu_i) = \beta_0 + \beta_1 \text{IDH}_i$ e $\phi = \exp(-\delta_0)$ para $i = 1, \ldots, 195$, com μ_i correspondendo à média e ϕ sendo um parâmetro de dispersão.

As distribuições *a priori* utilizadas são $\boldsymbol{\beta} \sim \text{N}_2(\mathbf{0}, 10^4 \boldsymbol{I}_2)$, $\delta_0 \sim \text{N}(0, 10^4)$ e, para o caso da regressão Beta, $\phi \sim$ Gama Inversa$(0.01, 0.01)$. A priori para o parâmetro a é justificada uma vez que os dados estão no intervalo $(0,1)$, com valores positivos tanto para a LG quanto para a CLG para tal parâmetro. Para realizar a inferência Bayesiana, são geradas 2000 amostras MCMC, descartando as primeiras 1000. A convergência foi checada utilizando *traceplots* e o critério de Gelman-Rubin (GELMAN; RUBIN *et al.*, 1992). A Tabela 30 mostra resumos dos resultados obtidos a partir da posteriori para os modelos ajustados, bem como valores de WAIC e de tempo computacional, em segundos. Os resultados para o modelo LG foram omitidos, uma vez que o modelo CLG trouxe melhor ajuste e ambos os modelos são complementares. O computador utilizado para o ajuste tem um processador Intel i7-8550U e 8GB de memória RAM DDR4. Com respeito à seleção de modelos baseada no WAIC, escolhemos o modelo de regressão quantílica CLG como o modelo que trouxe o melhor ajuste.

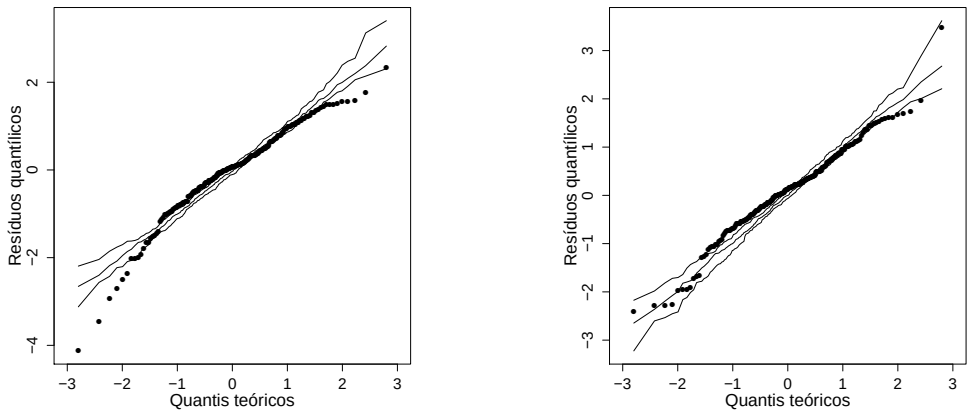
Tabela 30 – Resumos da posteriori para os modelos de regressão quantílica ajustados e para o modelo Beta de regressão na média (Est: média da média da posteriori, DP: média do desvio padrão da posteriori, IC: intervalo de credibilidade bilateral 95% para os dados de pobreza extrema.

Modelo	Parâmetro	Est	DP	95% IC
Beta	β_0	9.69	0.58	(8.60, 10.80)
	β_{IDH}	-19.27	1.03	(-21.25, -17.33)
	ϕ	15.25	1.52	(12.16, 18.30)
	WAIC	-439.7		
	Time (sec)	59.2		
Kumaraswamy	β_0	11.50	0.67	(10.18, 12.90)
	β_{IDH}	-22.63	1.19	(-25.16, -20.36)
	φ	3.54	0.20	(3.13, 3.94)
	WAIC	-443.1		
	Time (sec)	43.4		
CLG	β_0	13.76	0.76	(12.31, 15.21)
	β_{IDH}	-26.64	1.35	(-29.23, -24.01)
	a	4.14	0.41	(3.36, 4.92)
	WAIC	-445.8		
	Time (sec)	43.1		

Adicionalmente, também utilizamos os resíduos quantílicos normalizados (DUNN; SMYTH, 1996; PEREIRA, 2019) para checar a adequabilidade do ajuste de cada modelo. Tais resíduos são definidos como $r_{Qi} = \Phi^{-1}\big(F(y_i;\widehat{\boldsymbol{\theta}}_i)\big)$, $i = 1,\ldots,n = 195$, onde $\Phi(\cdot)$ é a fda da distribuição normal padrão e $F(\cdot)$ é a fda do modelo ajustado (em nosso caso, Beta e CLG são mostrados), enquanto $\widehat{\boldsymbol{\theta}}_i$ é computado com a média *a posteriori* dos parâmetros. Para a regressão CLG, por exemplo, $\widehat{\boldsymbol{\theta}}_i = (\widehat{\kappa}_i, \widehat{a})^T$, com $\text{logito}(\widehat{\kappa}_i) = \widehat{\beta}_0 + \widehat{\beta}_1 \text{IDH}_i$ e $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{a})^T$ denotam a média *a posteriori* dos parâmetros do modelo. A parte da incerteza que ocorre no processo de estimação dos parâmetros, para um modelo bem ajustado, r_{Qi}, $i = 1,\ldots,n$ constitui uma amostra aleatória que segue aproximadamente a distribuição normal padrão. Envelopes simulados (ATKINSON, 1985) são úteis para verificar a qualidade do ajuste por meio de resíduos, sendo utilizados para a análise dos resíduos quantílicos normalizados.

A Figura 18 mostra os envelopes para tais resíduos. Os limites dos envelopes e das linhas tracejadas correspondem aos quantis 0.025, 0.975 e 0.5 dos resíduos *a posteriori* computados da saída MCMC. Mais uma vez, a regressão quantílica CLG na mediana se adequa melhor aos dados (ver também a Tabela 30), sendo escolhido para prosseguir as análises.

O modelo de regressão LG na mediana é o escolhido como o modelo que melhor explica a relação entre o IDH e uma medida central de pobreza extrema no Peru. A Tabela 30 mostra que há uma forte relação entre o IDH e a pobreza extrema nas províncias peruanas avaliadas, com províncias com maiores valores de IDH tendendo a ter proporções menores de pobreza extrema.



(a) (b)

Figura 18 – Resíduos quantílicos normalizados com envelopes para os modelos Beta (a) e CLG (b) para os dados de pobreza extrema.

Nessa abordagem, consideramos modelos com $q = 0.5$ pelo fato de que um modelo na mediana é mais comparável ao modelo de regressão Beta sobre a média e também pela interpretabilidade dos resultados, os quais são adequados para explicar uma medida de tendência central para a proporção de pobreza extrema. Se o interesse do pesquisador for em outros quantis de interesse, então modelos LG e CLG com diferentes valores de q poderiam ser propostos. Para mostrar que isso seria possível, ajustamos modelos com $q = 0.1$ e $q = 0.9$. Os resultados de WAIC são mostrados na Tabela 31. Note que, para os casos de ambos os quantis, o modelo CLG também é melhor ajustado que o LG, com menores valores de WAIC sendo encontrados.

Tabela 31 – Resumos da posteriori para os modelos de regressão quantílica para q=0.1 e q=0.9 (Est: média da média da posteriori, DP: média do desvio padrão da posteriori, IC: intervalo de credibilidade bilateral 95% para os dados de pobreza extrema.

	Model	Parâmetro	Est	DP	95% CI
$q = 0.1$	LG	β_0	15.58	1.3	$(12.83, 18.10)$
		β_{IDH}	-32.66	2.27	(-37.04, -28.08)
		a	0.59	0.05	$(0.49, 0.69)$
		WAIC	-264.2		
		Time (sec.)	41.9		
	CLG	β_0	16.25	1.06	$(14.27, 16.97)$
		β_{IDH}	-33.60	1.78	(-37.07, -30.29)
		a	4.31	0.36	$(3.64, 5.00)$
		WAIC	-454.8		
		Time (sec.)	45.4		
$q = 0.9$	LG	β_0	14.57	1.22	$(12.14, 16.92)$
		β_{IDH}	-24.04	2.22	(-28.31, -19.50)
		a	0.58	0.05	$(0.48, 0.68)$
		WAIC	-265.7		
		Time (sec.)	45.1		
	CLG	β_0	11.56	0.64	$(10.30, 12.87)$
		β_{IDH}	-21.42	1.15	(-23.75, -19.18)
		a	4.31	0.41	$(4.06, 5.10)$
		WAIC	-443.6		
		Time (sec)	45.3		

Modelos de regressão quantílica com efeitos mistos

Uma vez que as províncias no Peru são organizadas em departamentos, elas podem, naturalmente, formar alguns grupos (ou clusters). Portanto, nessa seção são ajustados modelos com um efeito aleatório a fim de capturar a dependência dentro dos departamentos. Quatro diferentes modelos são ajustados, baseados nos modelos construídos anteriormente (modelo de efeitos fixos na mediana, aqui nomeado de Modelo I) e modelos alternativos incluindo efeitos aleatórios, com respostas que seguem uma distribuição CLG. A mediana da proporção de pobreza extrema é tomada novamente como característica de interesse e os seguintes quatro modelos são ajustados:

Modelo I. Modelo na mediana: $y_i \overset{\text{indep.}}{\sim} \text{CLG}(\kappa_i, a_i, 0.5)$ e $\text{logito}(\kappa_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 \text{IDH}_i$, $\log(a_i) = -\delta_0$,

Modelo II. Modelo na mediana e de forma: $y_i \overset{\text{indep.}}{\sim} \text{CLG}(\kappa_i, a_i, 0.5)$ e $\text{logito}(\kappa_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 \text{IDH}_i$, $\log(a_i) = -\boldsymbol{w}_i^T \boldsymbol{\delta} = -(\delta_0 + \delta_1 \text{IDH}_i)$, para $i = 1, \ldots, 195$ províncias,

Modelo III. Modelo na mediana de intercepto aleatório: $y_{ij}|b_i \overset{\text{indep.}}{\sim} \text{CLG}(\kappa_{ij}, a_{ij}, 0.5)$, $\text{logito}(\kappa_{ij}) =$

Tabela 32 – Comparação dos modelos propostos para os dados de pobreza extrema por meio do critério WAIC (p: número de parâmetros)

	Modelo	p	p_{WAIC}	WAIC	Tempo (seg)
Efeitos fixos	Modelo I	3	4.2	-445.8	45.8
	Modelo II	4	4.1	-472.6	58.4
Efeitos mistos	Modelo III	28	22.9	-572.4	64.6
	Modelo IV	54	30.1	-608.1	115.4

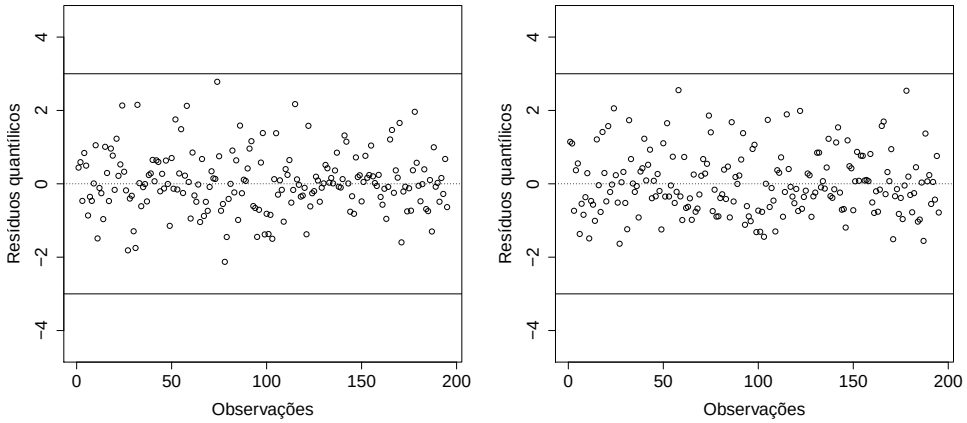
$\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_i = \beta_0 + \beta_1 \text{IDH}_{ij} + b_i$, $\log(a_{ij}) = -\delta_0$ e $b_i \sim \text{N}(0, \sigma_b^2)$,

Modelo IV. Modelo na mediana e de forma de interceptos aleatórios: $y_{ij}|b_i, d_i \overset{\text{indep.}}{\sim} \text{CLG}(\kappa_{ij}, a_{ij}, 0.5)$, $\text{logito}(\kappa_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_i = \beta_0 + \beta_1 \text{IDH}_{ij} + b_i$, $\log(a_{ij}) = -\boldsymbol{w}_{ij}^T\boldsymbol{\delta} - d_i = -(\delta_0 + \delta_1 + \text{IDH}_{ij} + d_i)$, $b_i \sim \text{N}(0, \sigma_b^2)$ e $d_i \sim \text{N}(0, \sigma_d^2)$, para $j = 1, \ldots, n_i$ províncias e $i = 1, \ldots, 25$ departmentos.

Os modelos I e II são modelos de regressão com efeitos fixos, onde as covariáveis tem um efeito na mediana e, simultaneamente, na mediana e no parâmetro de forma, respectivamente, sem considerar efeitos aleatórios. Os modelos III e IV são obtidos a partir dos modelos I e II, adicionando um efeito aleatório no intercepto, a fim de capturar a dependência dentro dos departamentos, primeiramente para a mediana e depois simultaneamante para a mediana e o parâmetro de forma. As distribuições *a priori* adotadas são $\boldsymbol{\beta} \sim \text{N}_2(\boldsymbol{0}, 10^4\boldsymbol{I}_2)$, $\delta_1 \sim \text{N}(0, 10^4)$ (modelos I e III), e $\boldsymbol{\delta} \sim \text{N}_2(\boldsymbol{0}, 10^4\boldsymbol{I}_2)$ para os coeficientes de regressão. Além disso, para σ_b^2 e σ_d^2 são adotadas prioris Gama Inversa, com parâmetros $(0.01, 0.01)$, tornando as prioris vagas assim como em Bayes, Bazán e Castro (2017). Mais uma vez, são realizadas 2000 iterações, descartando as primeiras 1000, através do NUTS, com a convergência sendo checada por meio de análises gráficas e a estatística de Gelman-Rubin, que é próxima a 1 para todos os parâmetros, indicando convergência.

Os quatro modelos foram ajustados e o WAIC foi computado por meio da saídas do MCMC, a log-verossimilhança condicional é utilizada para a obtenção dos resultados. O WAIC para os quatro modelos é mostrado na Tabela 32. Podemos ver que adicionar efeitos aleatórios faz com que o ajuste seja melhor (os modelos III e IV tem menores valores de WAIC que I e II. O menor valor para os WAIC é obtido para o modelo IV, seguido bastante de perto pelo modelo III. Apesar de o tempo computacional aumentar com os efeitos aleatórios, esse incremento não faz com que o proceso se torne demasiadamente lento e, mesmo o modelo IV, levou menos de dois minutos para ser ajustado.

A seguir foram analisados os resíduos dos modelos mistos. A Figura 19 mostra os resultados para os modelos III e IV. É possível ver que todos os resíduos estão no intervalo $(-3, 3)$, mostrando que não há valores discrepantes nos resíduos. Ambos os modelos III e IV



(a) Modelo III (b) Modelo IV

Figura 19 – Valores dos resíduos quantílicos normalizados para dados de pobreza extrema.

trazem bom ajuste aos dados, mas os resultados mostrados a seguir se baseiam no modelo III, apesar do modelo IV trazer menor valor de WAIC, uma vez que os intervalos de credibilidade para os interceptos aleatórios no parâmetro de forma têm o 0 incluso para todos os departamentos no modelo IV, indicando que esse efeito não é necessário ao ajuste.

Os resumos da posteriori para o modelo III são mostrados na Tabela 33. Como já observado na aplicação anterior, o IDH é fortemente relacionado à proporção de pobreza extrema. Incluindo os efeitos aleatórios b_i, obtemos um modelo de interceptos aleatórios, tornando a análise dos dados mais rica. Podemos ver, por exemplo, que o departamento de Apurímac tem o maior valor de b_i, indicando que as províncias nesse departamento tendem a ter maiores proporções de pobreza extrema. Por outro lado, o departamento com menores valores de pobreza extrema é, de longe, Ica. Dentre os departamentos estudados, temos que dois deles, os de Ica e Ucayali possuem proporções significativamente menores de pobreza extrema em relações às demais, não tendo o zero incluso no intervalo de credibilidade. Por outro lado, há oito departamentos com proporções significativamente maiores de pobreza extrema, os quais são os de Apurímac, Huánuco, Loreto, Huancavelica, Amazonas, Pasco, La e Cusco, em ordem decrescente.

5.6 Considerações finais

Nesse capítulo adicional, duas novas distribuições para dados limitados e seus respectivos modelos de regressão quantílica paramétrica são propostos. Reparametrizações das distribuições com base em um dado quantil e um parâmetro de forma nos permitem fazer uma ligação entre qualquer quantil da distribuição e um conjunto de covariáveis. Essas distribuições podem ser úteis em diversos casos, inclusive naqueles que tem um comportamento assimétrico e, também,

Tabela 33 – Resumos da posteriori para o modelo III (Est: média da média da posteriori, DP: média do desvio padrão da posteriori, IC: intervalo de credibilidade bilateral 95% para os dados de pobreza extrema

		Est	DP	95% IC
Efeitos fixos da mediana ($\boldsymbol{\beta}$)	Intercepto	9.37	0.76	(7.85, 10.88)
	IDH	-19.37	1.26	(-21.88, -16.79)
Efeitos fixos de forma ($\boldsymbol{\delta}$)	Intercept	-2.16	0.09	(-2.32, -1.98)
Efeitos aleatórios ($\boldsymbol{b}$)	Ica	-2.80	0.87	(-4.35, -0.96)
	Ucayali	-0.96	0.38	(-1.74, -0.21)
	Tumbes	-0.69	0.52	(-1.72, 0.28)
	Callao	-0.49	0.83	(-2.16, 1.16)
	Madre	-0.44	0.48	(-1.40, 0.49)
	Lambayeque	-0.43	0.42	(-1.26, 0.39)
	Ancash	-0.4	0.24	(-0.85, 0.09)
	Tacna	-0.37	0.38	(-1.08, 0.38)
	Lima	-0.33	0.31	(-0.95, 0.32)
	Moquegua	-0.29	0.46	(-1.12, 0.58)
	Junín	-0.17	0.29	(-0.69, 0.40)
	Arequipa	-0.11	0.31	(-0.71, 0.49)
	Piura	-0.09	0.26	(-0.61, 0.43)
	San	0.24	0.26	(-0.25, 0.80)
	Cajamarca	0.3	0.24	(-0.15, 0.77)
	Ayacucho	0.45	0.24	(0.00, 0.95)
	Puno	0.45	0.23	(0.00, 0.92)
	Cusco	0.57	0.24	(0.14, 1.06)
	La	0.61	0.24	(0.16, 1.08)
	Pasco	0.69	0.30	(0.14, 1.29)
	Amazonas	0.70	0.25	(0.21, 1.19)
	Huancavelica	0.77	0.24	(0.33, 1.22)
	Loreto	0.81	0.23	(0.38, 1.27)
	Huánuco	0.87	0.22	(0.44, 1.31)
	Apurímac	1.28	0.23	(0.83, 1.76)
	σ_b^2	0.87	0.42	(0.31, 1.89)

em casos nos quais há valores concentrados no extremo do intervalo unitário, sendo assim distribuições com bom ajuste para variáveis de caudas pesadas.

A respeito dos modelos com valores negativos para o parâmetro a, principalmente no caso quando há muitos valores no limite do intervalo unitário, o fato de os dados serem esparsos faz com que maiores valores amostrais sejam necessários para a estimação dos parâmetros.

A inferência se baseia em uma abordagem Bayesiana com prioris próprias (e vagas). Uma vez que as distribuições a posterior não permitem tratamento analítico, métodos de MCMC são de suma importância. O estudo de simulação mostra que o WAIC é um bom critério para auxiliar na escolha entre diferentes modelos com distribuições distintas. Um conjunto de dados sobre pobreza extrema é analisado considerando um dos modelos propostos.

Para trabalhos futuros, há a possibilidade de diferentes funções de ligação em (5.3) serem exploradas, inclusive funções assimétricas. Modelos com uma componente espacial também podem ser extensões interessantes, bem como para dados censurados. Como em Chakraborty (2003) também há o interesse no desenvolvimento de modelos multivariados.

Mais voltado ao cenário de Modelos de Diagnóstico Cognitivo, as distribuições aqui propostas podem ser consideradas para trabalhos futuros de proposta de modelos B-DINA com distribuições alternativas.

CAPÍTULO

6

DISCUSSÃO E CONCLUSÃO

Nesse capítulo são sumarizadas as contribuições do presente trabalho, produções resultantes do mesmo e uma breve discussão em relação à perspectivas futuras, em linha com o que é apresentado ao longo dos capítulos, porém ampliando a discussão para as intersecções e relações entre os capítulos é apresentada.

6.1 Contribuições no estado da arte

A literatura dos Modelos de Diagnóstico Cognitivo ainda possui espaços para ser enriquecida com novas pesquisas, bem como as propostas nessa tese de Doutorado. Em suma, as maiores contribuições para essa área de pesquisa que foram realizadas nessa pesquisa foram:

- A colaboração em um trabalho que propôs uma nova abordagem Bayesiana para o modelo DINA tradicional (dicotômico), por meio de um algoritmo até então não explorado, bem como na exploração de uma aplicação em uma área diferente da educacional, na qual mais aparecem contribuições para MDCs, ressaltando a discussão para demais pesquisadores sobre a utilidade dessa técnica em diversas áreas do conhecimento, como a de saúde mental, a qual foi essencial para o melhor entendimento dos MDCs e para as propostas dos demais capítulos;

- A proposta de uma inédita versão Bayesiana para o modelo DINA para distribuição contínua, o qual somente existia em sua versão frequentista;

- O entendimento e a explicação mais didática da lógica por trás do modelo C-DINA, que ainda era algo em falta na literatura dos diversos MDCs existentes, que pode ser estendida para o modelo B-DINA e até mesmo para os demais modelos de diagnóstico cognitivo;

- A proposta de um modelo DINA inédito para dados limitados (B-DINA), sob abordagem Bayesiana, o qual ainda não era presente na literatura nem sob uma abordagem frequentista nem Bayesiana;
- Ainda dentro do modelo para respostas limitadas, uma aplicação em uma área do conhecimento diferente foi proposta, mostrando que, além de avaliação de questionários respondidos por pessoas, os modelos estudados também podem ser úteis para outras aplicações, bem como a análise de indicadores sociais;
- Para ambos os modelos propostos, o C-DINA sob abordagem Bayesiana e o B-DINA, a disponibilização de códigos que possibilitam à pesquisadores utilizarem as metodologias para seus próprios estudos futuros
- Motivado pelo estudo de respostas limitadas, a proposta de uma nova distribuição, bem como sua versão quantílica e regressão mista, que, além da importância em si mesmo de um novo modelo e da regressão utilizando tal modelo, pode ser útil para trabalhos futuros relacionados à modelos de diagnóstico cognitivo para dados limitados, bem como para demais aplicações.

6.2 Produção

Trabalhos para eventos

- Apresentação de Pôster e resumo publicado: V CONBRATRI - 5º Congresso Brasileiro de Teoria da Resposta ao Item, 2016. *Bayesian approach to the DINA model using No-U-Turn Hamiltonian Monte Carlo.*
- Mesa Redonda: Semana da Estatística, UFSCar, 2016. Vida Acadêmica. Vida Acadêmica.
- Apresentação de Pôster e resumo publicado: 5th Workshop on Probabilistic and Statistical Methods, 2017. *Bayesian estimation of DINA model using no-U-turn Hamiltonian Monte Carlo.*
- Palestra: Programa de Verão em Estatística, ICMC USP, 2017. Modelos de Diagnóstico cognitivo com ênfase no modelo DINA. Palestra dada durante o minicurso Tópicos em Modelos de Resposta ao Item.
- Entrevista: Metodologia PBL é aprimorada nas aulas de Estatística, 2017. Disponível em: http://www.cemeai.icmc.usp.br/projetos/item/597-metodologia-pbl-e-aprimorada-nas-aulas-de-estatistica.
- Apresentação de Pôster e resumo publicado: 6th Workshop on Probabilistic and Statistical Methods, 2018. *New Gompertz based distributions to skewed bounded responses.*

- Apresentação de Pôster e resumo publicado: NextGen: Data Science Day, Yale, 2018. *An Application of DINA model to Beck Depression Inventory data.*
- Apresentação de Mini-Curso: 63ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria (RBras), 2018. *Uma introdução aos Modelos de Diagnóstico Cognitivo utilizando o software Stan.*
- Resumo publicado: 3rd International Conference on Econometrics and Statistics, Taichung City, Taiwan, 2019. *Bayesian analysis of cognitive diagnostic models for continuous response data*

Artigos Publicados

da Silva MA, de Oliveira ESB, von Davier AA, Bazán JL. Estimating the DINA model parameters using the No-U-Turn Sampler. *Biometrical Journal*. 2017;0:1–17. https//-doi.org/10.1002/bimj.201600225

Artigos a serem submetidos

Artigo referente ao C-DINA, atualmente com o seguinte nome e composição: de Oliveira E. S. B., Wang, X., Bazán J. L. A classification model for continuous response: identifying risk perception groups on health-related activities,

Também será submetido um artigo, em fase de finalização, do modelo B-DINA proposto, dos mesmos autores do artigo do C-DINA.

Artigo referente à LG/CLG, atualmente com o seguinte nome e composição: de Oliveira E. S. B., de Castro, M., Bayes, C. L., Bazán J. L. Bayesian parametric quantile models for heavy tailed bounded variables.

6.3 Possibilidades de trabalhos futuros

Ao longo dos capítulos foram escritas algumas considerações sobre possibilidades de trabalhos futuros específicos à cada um dos mesmos, porém, vale ressaltar algumas das possibilidades, bem como conexões entre elas.

Tanto para o modelo C-DINA quanto para o B-DINA, podem ser exploradas diversas abordagens para a estimação de parâmetros, similar ao que pode ser visto no trabalho do modelo DINA dicotômico.

O modelo B-DINA tem amplo espaço para ser explorado, uma vez que essa é uma primeira abordagem na literatura. No futuro pode ser proposta uma versão frequentista para o modelo, com elaboração e explicação de seus detalhes. Também podem ser desenvolvidos modelos considerando outras distribuições de probabilidade, o que pode ser feito, por exemplo,

em conexão com o Capítulo 5 da tese ou com distribuições correlatas à utilizada nesse trabalho, como a Beta inflacionada (OSPINA; FERRARI, 2012).

Também há a possibilidade de se explorar, principalmente para os modelos C-DINA e B-DINA propostos no trabalho, o uso de covariáveis observadas, a fim de enriquecer a análise, como pode ser visto em Park, Xing e Lee (2018), bem como explorar novas formas de resposta.

Outra possibilidade é explorar diferentes formatos de respostas, como feito nesse trabalho para o modelo DINA, considerando modelos mais generalistas, como o G-DINA (TORRE, 2011b), amplificando ainda mais a discussão sobre o uso de variáveis respostas não dicotômicas para MDCs.

Para as distribuições LG e CLG propostas, as regressões quantílicas podem ser exploradas também sob uma abordagem frequentista, bem como outros métodos MCMC para a estimação dos parâmetros podem ser estudados.

REFERÊNCIAS

ANAND, S.; SEN, A. **Human Development Index: Methodology and Measurement.** [S.l.], 1994. Citado na página 70.

ATKINSON, A. C. **Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis**. [S.l.], 1985. Citado na página 110.

BABIAK, K.; WOLFE, R. Determinants of corporate social responsibility in professional sport: Internal and external factors. **Journal of Sport Management**, v. 23, n. 6, p. 717–742, 2009. Citado na página 65.

BARTHOLOMEW, D. J.; KNOTT, M.; MOUSTAKI, I. **Latent variable models and factor analysis: A unified approach**. [S.l.]: John Wiley & Sons, 2011. v. 904. Citado na página 21.

BAYES, C. L.; BAZÁN, J. L.; CASTRO, M. D. A quantile parametric mixed regression model for bounded response variables. **Statistics and Its Interface**, International Press of Boston, v. 10, n. 3, p. 483–493, 2017. Citado nas páginas 102 e 113.

BAYES, C. L.; BAZÁN, J. L.; GARCÍA, C. A new robust regression model for proportions. **Bayesian Analysis**, International Society for Bayesian Analysis, v. 7, n. 4, p. 841–866, 2012. Citado na página 102.

BECK, A. T.; WARD, C. H.; MENDELSON, M.; ERBAUGH, J. An inventory for measuring depression. **Archives of General Psychiatry**, v. 4, p. 53–63, 1961. Citado nas páginas 27 e 40.

BIJUR, P. E.; SILVER, W.; GALLAGHER, E. J. Reliability of the visual analog scale for measurement of acute pain. **Academic emergency medicine**, Wiley Online Library, v. 8, n. 12, p. 1153–1157, 2001. Citado na página 70.

BOX, G. E.; TIAO, G. C. **Bayesian Inference in Statistical Analysis**. [S.l.]: John Wiley & Sons, 2011. v. 40. Citado na página 66.

BRADLOW, E. T.; WAINER, H.; WANG, X. A bayesian random effects model for testlets. **Psychometrika**, v. 64, n. 2, p. 153–168, 1999. Citado na página 67.

CARLSTROM, L. K.; WOODWARD, J. A.; PALMER, C. G. Evaluating the simplified conjoint expected risk model: Comparing the use of objective and subjective information. **Risk Analysis**, v. 20, n. 3, p. 385–392, 2000. Citado nas páginas 64 e 70.

CASSIDAY, K. R.; CHO, Y.; HARRING, J. R. A comparison of label switching algorithms in the context of growth mixture models. **Educational and Psychological Measurement**, SAGE Publications Sage CA: Los Angeles, CA, p. 0013164420970614, 2020. Citado na página 53.

CHAKRABORTY, B. On multivariate quantile regression. **Journal of statistical planning and inference**, Elsevier, v. 110, n. 1-2, p. 109–132, 2003. Citado na página 116.

CHEN, J.; TORRE, J. de la. A general cognitive diagnosis model for expert-defined polytomous attributes. **Applied Psychological Measurement**, Sage Publications Sage CA: Los Angeles, CA, v. 37, n. 6, p. 419–437, 2013. Citado na página 48.

CHEN, M.-H.; SHAO, Q.-M. Monte carlo estimation of bayesian credible and hpd intervals. **Journal of Computational and Graphical Statistics**, v. 8, n. 1, p. 69–92, 1999. Citado na página 66.

CHEN, P.; XIN, T. Online calibration methods for the DINA model with independent attibutes in cd-cat. **Psychometrika**, v. 77, n. 2, p. 201–222, 2012. Citado na página 26.

CHEN, Y.; CULPEPPER, S. A.; CHEN, Y.; DOUGLAS, J. Bayesian estimation of the DINA Q matrix. **Psychometrika**, Springer, p. 1–20, 2017. Citado na página 22.

CHEN, Y.; LIU, J.; XU, G.; YING, Z. Statistical analysis of Q-matrix based diagnostic classification models. **Journal of the American Statistical Association**, v. 110, p. 850–866, 2015. Citado nas páginas 28 e 41.

CHIU, C.-Y.; DOUGLAS, J. A.; LI, X. Cluster analysis for cognitive diagnosis: Theory and applications. **Psychometrika**, Springer, v. 74, n. 4, p. 633–665, 2009. Citado na página 41.

CHYLACK, L. T.; PETERSON, L. E.; FEIVESON, A. H.; WEAR, M. L.; MANUEL, F. K.; TUNG, W. H.; HARDY, D. S.; MARAK, L. J.; CUCINOTTA, F. A. Nasa study of cataract in astronauts (nasca). report 1: Cross-sectional study of the relationship of exposure to space radiation and risk of lens opacity. **Radiation research**, Allen Press, v. 172, n. 1, p. 10–20, 2009. Citado na página 71.

CULPEPPER, S. A. Bayesian estimation of the DINA model with gibbs sampling. **Journal of Educational and Behavioral Statistics**, v. 40, n. 5, p. 454–476, 2015. Disponível em: <http://www.hermanaguinis.com/pubs.html>. Citado nas páginas 22, 27, 29, 31, 32, 35, 36, 37, 42 e 46.

DAVIER, M. von. A general diagnostic model applied to language testing data. **British Journal of Mathematical and Statistical Psychology**, v. 61, n. 2, p. 287–307, 2008. Citado nas páginas 21 e 26.

DECARLO, L. T. On the analysis of fraction subtraction data: The dina model, classification, latent class sizes, and the q-matrix. **Applied Psychological Measurement**, Sage Publications Sage CA: Los Angeles, CA, v. 35, n. 1, p. 8–26, 2011. Citado na página 41.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the Royal Statistical Society**, v. 39, n. 1, p. 1–38, 1977. Citado nas páginas 31 e 131.

DIMITROV, D. M. Least squares distance method of cognitive validation and analysis for binary items using their item response theory parameters. **Applied Psychological Measurement**, v. 31, p. 367–387, 2007. Citado na página 26.

DUANE, A.; KENNEDY, A.; PENDLETON, B.; ROWETH, D. Hybrid monte carlo. **Physics Letters B**, v. 195, n. 2, p. 216–222, 1987. Citado nas páginas 27, 33 e 131.

DUFFY, D. L. Internal and external factors which affect customer loyalty. **Journal of Consumer Marketing**, v. 20, n. 5, p. 480–485, 2003. Citado na página 65.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. Citado na página 110.

FAN, Z.; WANG, C.; CHANG, H.-H.; DOUGLAS, J. Utilizing response time distributions for item selection in cat. **Journal of Educational and Behavioral Statistics**, v. 37, n. 5, p. 655–670, 2012. Citado na página 48.

FERRANDO, P. J. A nonlinear congeneric model for continuous item responses. **British Journal of Mathematical and Statistical Psychology**, Wiley Online Library, v. 54, n. 2, p. 293–313, 2001. Citado na página 48.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of applied statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004. Citado nas páginas 70 e 71.

FERRARI, S. L. P.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of Applied Statistics**, v. 31, p. 799–815, 2004. Citado nas páginas 94 e 102.

FIGUEROA-ZÚÑIGA, J. I.; ARELLANO-VALLE, R. B.; FERRARI, S. L. Mixed beta regression: A bayesian perspective. **Computational Statistics & Data Analysis**, Elsevier, v. 61, p. 137–147, 2013. Citado nas páginas 94 e 102.

FOX, J.-p. **Bayesian Item Response Modeling: Theory and Applications**. New York: Springer, 2010. Citado na página 40.

FOX, J.-P.; ENTINK, R. K.; LINDEN, W. J. van der. Modeling of responses and response times with the package cirt. **Journal of Statistical Software**, v. 20, n. 7, p. 1–14, 2007. Citado na página 67.

FRAGOSO, T. M.; CÚRI, M. Improving psychometric assessment of the beck depression inventory using multidimensional item response theory. **Biometrical Journal**, v. 55, p. 527–540, 2013. Citado nas páginas 28, 41 e 46.

GELFAND, A.; SMITH, A. Sampling-based approaches to calculating marginal densities. **Journal of the American Statistical Association**, v. 85, p. 398–409, 1990. Citado nas páginas 31 e 131.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; DUNSON, D. B.; VEHTARI, A.; RUBIN, D. B. **Bayesian data analysis**. [S.l.]: CRC press, 2013. Citado nas páginas 77 e 78.

____. **Bayesian Data Analysis**. 3rd. ed. New York: Chapman & Hall, 2014. Citado nas páginas 37 e 58.

GELMAN, A.; HWANG, J.; VEHTARI, A. Understanding predictive information criteria for Bayesian models. **Statistics and Computing**, v. 24, p. 997–1016, 2014. Citado na página 100.

GELMAN, A. *et al.* Two simple examples for understanding posterior p-values whose distributions are far from uniform. **Electronic Journal of Statistics**, The Institute of Mathematical Statistics and the Bernoulli Society, v. 7, p. 2595–2602, 2013. Citado na página 78.

GELMAN, A.; RUBIN, D. Inference from iterative simulation using multiple sequences. **Statistical Science**, v. 7, n. 4, p. 457–472, 1992. Citado nas páginas 35 e 36.

GELMAN, A.; RUBIN, D. B. *et al.* Inference from iterative simulation using multiple sequences. **Statistical Science**, v. 7, n. 4, p. 457–472, 1992. Citado nas páginas 57, 102 e 109.

GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distributions and the Bayesian restoration of images. **IEEE Trans. Pattern Anal. Mach. Intell**, v. 6, n. 6, p. 721–741, 1984. Citado nas páginas 31 e 131.

GEORGE, A. C.; ROBITZSCH, A. Cognitive diagnosis models in r: a didactic. **The quantitative methods for psychology**, v. 11, n. 3, p. 189–205, 2015. Citado nas páginas 21, 26 e 131.

GEORGE, A. C.; ROBITZSCH, A.; KIEFER, T.; GROSS, J.; ÜNLÜ, A. The R package CDM for cognitive diagnosis models. **Journal of Statistical Software**, v. 74, n. 2, p. 1–24, 2016. Citado nas páginas 27 e 132.

GERACI, M.; BOTTAI, M. Linear quantile mixed models. **Statistics and Computing**, v. 24, p. 461–479, 2014. Citado na página 94.

GHITANY, M.; MAZUCHELI, J.; MENEZES, A.; ALQALLAF, F. The unit-inverse gaussian distribution: A new alternative to two-parameter distributions on the unit interval. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 48, n. 14, p. 3423–3438, 2019. Citado na página 95.

GIROLAMI, M.; CALDERHEAD, B. Riemann manifold langevin and hamiltonian monte carlo methods. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 73, n. 2, p. 123–214, 2011. Citado na página 37.

GONÇALVES, F. B.; DIAS, B. da C. C.; SOARES, T. M. Bayesian item response model: a generalized approach for the abilities' distribution using mixtures. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 88, n. 5, p. 967–981, 2018. Citado na página 74.

GORENSTEIN, C.; ANDRADE, L.; FILHO, A. H. G. V.; TENG, C. T.; ARTES, R. Psychometric properties of the portuguese version of the beck depression inventory on brazilian college students. **Journal of Clinical Psychology**, v. 55, n. 5, p. 553–562, 1999. Citado na página 40.

GRANTY, R. L.; FURRZ, D. C.; CARPENTER, B.; GELMAN, A. Fitting bayesian item response models in stata and stan. corr. **arXiv preprint arXiv:1601.03443**, 2016. Citado na página 27.

GUPTA, A. K.; NADARAJAH, S. **Handbook of beta distribution and its applications**. [S.l.]: CRC press, 2004. Citado na página 71.

HARRIS, M. N.; ZHAO, X. A zero-inflated ordered probit model, with an application to modelling tobacco consumption. **Journal of Econometrics**, Elsevier, v. 141, n. 2, p. 1073–1099, 2007. Citado na página 96.

HASTINGS, W. K. Monte carlo sampling methods using Markov chains and their applications. **Biometrika**, v. 57, p. 97–109, 1970. Citado nas páginas 31 e 131.

HOFFMAN, M. D.; GELMAN, A. The No-U-Turn sampler: adaptively setting path lengths in hamiltonian monte carlo. **Journal of Machine Learning Research**, v. 15, n. 1, p. 1593–1623, 2014. Citado nas páginas 15, 22, 27, 33, 34, 100 e 131.

HUANG, H. Y.; WANG, W. C. The random-effect DINA model. **Journal of Educational Measurement**, v. 51, n. 1, p. 75–97, 2014. Citado na página 27.

HUEBNER, A.; WANG, C. A note on comparing examinee classification methods for cognitive diagnosis models. **Educational and Psychological Measurement**, v. 71, p. 407–419, 2011. Citado na página 35.

III, E. P. C. The optimal number of response alternatives for a scale: A review. **Journal of marketing research**, JSTOR, p. 407–422, 1980. Citado na página 48.

INC, S. I. **SAS/STAT® 9.2 User's Guide, Second Edition**. Cary, NC: SAS Institute Inc, 2009. Citado na página 27.

JAEGER, J.; TATSUOKA, C.; BERNS, S.; VARADI, F. Distinguishing neurocognitive functions using partially ordered classification models. **Schizophrenia Bulletin**, v. 32, p. 679–691, 2006. Citado nas páginas 26 e 27.

JASRA, A.; HOLMES, C. C.; STEPHENS, D. A. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. **Statistical Science**, p. 50–67, 2005. Citado na página 53.

JODRÁ, P. A bounded distribution derived from the shifted gompertz law. **Journal of King Saud University-Science**, Elsevier, v. 32, n. 1, p. 523–536, 2020. Citado na página 95.

JUNKER, B.; SIJTSMA, K. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. **Applied Psychological Measurement**, v. 25, p. 258–272, 2001. Citado nas páginas 21, 26, 27, 29 e 131.

KENDALL, P. C.; HOLLON, S. D.; BECK, A. T.; HAMMEN, C. L.; INGRAM, R. E. Issues and recommendations regarding the use of the beck depression inventory. **Cognitive Therapy and Research**, v. 11, p. 289–299, 1987. Citado nas páginas 40 e 44.

KOPP, J.; GERIKE, R.; AXHAUSEN, K. W. Do sharing people behave differently? an empirical evaluation of the distinctive mobility patterns of free-floating car-sharing members. **Transportation**, Springer, v. 42, n. 3, p. 449–469, 2015. Citado na página 71.

LEMONTE, A. J.; BAZÁN, J. L. New class of johnson distributions and its associated regression model for rates and proportions. **Biometrical Journal**, Wiley Online Library, v. 58, n. 4, p. 727–746, 2016. Citado na página 94.

LENART, A. The moments of the gompertz distribution and maximum likelihood estimation of its parameters. **Scandinavian Actuarial Journal**, Taylor & Francis, v. 2014, n. 3, p. 255–277, 2014. Citado nas páginas 94 e 96.

LI, F.; COHEN, A. S.; KIM, S.-H.; CHO, S.-J. Model selection methods for mixture dichotomous irt models. **Applied Psychological Measurement**, Sage Publications Sage CA: Los Angeles, CA, v. 33, n. 5, p. 353–373, 2009. Citado na página 53.

LINDEN, W. J. van der. A hierarchical framework for modeling speed and accuracy on test items. **Psychometrika**, v. 72, n. 3, p. 287, 2007. Citado na página 48.

____. Using response times for item selection in adaptive testing. **Journal of Educational and Behavioral Statistics**, v. 33, n. 1, p. 5–20, 2008. Citado na página 48.

LINDEN, W. J. van der; XIONG, X. Speededness and adaptive testing. **Journal of Educational and Behavioral Statistics**, v. 38, n. 4, p. 418–438, 2013. Citado na página 48.

LIU, R.; HUGGINS-MANLEY, A.; BRADSHAW, L. The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. **Educational and Psychological Measurement**, v. 77, n. 2, p. 220–240, 2017. Citado nas páginas 28 e 41.

LUNN, D.; JACKSON, C.; BEST, N.; SPIEGELHALTER, D.; THOMAS, A. **The BUGS book: A Practical Introduction to Bayesian Analysis**. [S.l.]: Chapman and Hall/CRC, 2012. Citado na página 54.

LUNN, D.; THOMAS, A.; BEST, N.; SPIEGELHALTER, D. Winbugs - a bayesian modelling framework: concepts, structure, and extensibility. **Statistics and Computing**, v. 10, p. 325–337, 2000. Citado nas páginas 27 e 34.

MA, W.; TORRE, J. A sequential cognitive diagnosis model for polytomous responses. **British Journal of Mathematical and Statistical Psychology**, Wiley Online Library, v. 69, n. 3, p. 253–275, 2016. Citado na página 48.

MAZUCHELI, J.; MENEZES, A. F.; DEY, S. Unit-gompertz distribution with applications. **Statistica**, v. 79, n. 1, p. 25–43, 2019. Citado nas páginas 95 e 96.

MENG, X.-B.; TAO, J.; CHANG, H.-H. A conditional joint modeling approach for locally dependent item responses and response times. **Journal of Educational Measurement**, v. 52, n. 1, p. 1–27, 2015. Citado na página 49.

MENG, X.-L. *et al.* Posterior predictive *p*-values. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 22, n. 3, p. 1142–1160, 1994. Citado na página 77.

MENGUC, B.; AUH, S.; OZANNE, L. The interactive effect of internal and external factors on a proactive environmental strategy and its influence on a firm's performance. **Journal of Business Ethics**, v. 94, n. 2, p. 279–298, 2010. Citado na página 65.

MERKLE, E.; FURR, D.; RABE-HESKETH, S. Bayesian model assessment: Use of conditional vs marginal likelihoods. **arXiv preprint arXiv:1802.04452**, 2018. Citado na página 101.

METROPOLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E. Equation of state calculations by fast computing machines. **The journal of chemical physics**, v. 21, p. 1087–1092, 1953. Citado nas páginas 31 e 131.

MIGLIORATI, S.; BRISCO, A. M. D.; ONGARO, A. *et al.* A new regression model for bounded responses. **Bayesian Analysis**, International Society for Bayesian Analysis, v. 13, n. 3, p. 845–872, 2018. Citado nas páginas 94 e 95.

MILLER, G. A. The magical number seven, plus or minus two: some limits on our capacity for processing information. **Psychological review**, American Psychological Association, v. 63, n. 2, p. 81, 1956. Citado na página 48.

MINCHEN, N. D.; TORRE, J. de la; LIU, Y. A cognitive diagnosis model for continuous response. **Journal of Educational and Behavioral Statistics**, SAGE Publications Sage CA: Los Angeles, CA, v. 42, n. 6, p. 651–677, 2017. Citado nas páginas 22, 49, 50, 51, 55, 58, 59 e 64.

MORIN, C.; BUSHNELL, M. Temporal and qualitative properties of cold pain and heat pain: a psychophysical study. **Pain**, Elsevier, v. 74, n. 1, p. 67–73, 1998. Citado na página 48.

NEAL, R. Mcmc using hamiltonian dynamics. In: BROOKS, I. S.; GELMAN, A.; J., G. L.; MENG, X.-l. (Ed.). **Handbook of Markov chain Monte Carlo**. Ch. 5: Chapman and Hall/CRC, 2011. p. 116–162. Citado nas páginas 27, 33 e 131.

NEAL, R. M. An improved acceptance procedure for the hybrid monte carlo algorithm. **Journal of Computational Physics**, v. 111, p. 194–203, 1994. Citado nas páginas 27, 33 e 131.

____. Slice sampling. **The Annals of Statistics**, v. 31, n. 3, p. 705–767, 2003. Citado nas páginas 57 e 77.

NERING, M. L.; OSTINI, R. **Handbook of polytomous item response theory models**. [S.l.]: Taylor & Francis, 2011. Citado na página 48.

NOEL, Y.; DAUVIER, B. A beta item response model for continuous bounded responses. **Applied Psychological Measurement**, Sage Publications Sage CA: Thousand Oaks, CA, v. 31, n. 1, p. 47–73, 2007. Citado nas páginas 49, 70 e 71.

NUGROHO, D.; MORIMOTO, T. Estimation of realized stochastic volatility models using hamiltonian monte carlo-based methods. **Computational Statistics**, v. 30, n. 2, p. 491–516, 2015. Citado na página 27.

OBERSKI, D. Mixture models: Latent profile and latent class analysis. In: **Modern statistical methods for HCI**. [S.l.]: Springer, 2016. p. 275–287. Citado na página 52.

OSHIMA, T. The effect of speededness on parameter estimation in item response theory. **Journal of Educational Measurement**, v. 31, n. 3, p. 200–219, 1994. Citado na página 67.

OSPINA, R.; FERRARI, S. L. A general class of zero-or-one inflated beta regression models. **Computational Statistics & Data Analysis**, Elsevier, v. 56, n. 6, p. 1609–1623, 2012. Citado nas páginas 96 e 120.

PARK, Y. S.; XING, K.; LEE, Y.-S. Explanatory cognitive diagnostic models: Incorporating latent and observed predictors. **Applied psychological measurement**, SAGE Publications Sage CA: Los Angeles, CA, v. 42, n. 5, p. 376–392, 2018. Citado na página 120.

PEREIRA, G. H. On quantile residuals in beta regression. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 48, n. 1, p. 302–316, 2019. Citado na página 110.

PERÚ, P. Informe sobre desarrollo humano perú 2009: Por una densidad del estado al servicio de la gente. **Lima: Programa de las Naciones Unidas para el Desarrollo**, 2009. Citado na página 109.

PLUMMER, M. JAGS: A program for analysis of bayesian graphical models using gibbs sampling. In: **Proceedings of the 3rd international workshop on distributed statistical computing**. Vienna, Austria: [s.n.], 2003. p. 125. Citado nas páginas 27 e 34.

____. Jags version 4.0.0 user manual. **See https://sourceforgenet/projects/mcmc-jags/files/Manuals/4x**, 2015. Citado nas páginas 57 e 77.

PLUMMER, M.; BEST, N.; COWLES, K.; VINES, K. Coda: Convergence diagnosis and output analysis for mcmc. **R News**, v. 6, n. 1, p. 7–11, 2006. Disponível em: <http://CRAN.R-project.org/doc/Rnews/>. Citado na página 39.

PRESTON, C. C.; COLMAN, A. M. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. **Acta psychologica**, Elsevier, v. 104, n. 1, p. 1–15, 2000. Citado na página 48.

QIU, Z.; SONG, P. X.-K.; TAN, M. Simplex mixed-effects models for longitudinal proportional data. **Scandinavian Journal of Statistics**, Wiley Online Library, v. 35, n. 4, p. 577–596, 2008. Citado nas páginas 94 e 95.

SAHU, S. K. Bayesian estimation and model choice in item response models. **Journal of Statistical Computation and Simulation**, v. 72, n. 3, p. 217–232, 2002. Citado nas páginas 37 e 78.

SIE, H.; FINKELMAN, M. D.; RILEY, B.; SMITS, N. Utilizing response times in computerized classification testing. **Applied psychological measurement**, v. 39, n. 5, p. 389–405, 2015. Citado na página 49.

SILVA, M. A. da; OLIVEIRA, E. S. de; DAVIER, A. A. von; BAZÁN, J. L. Estimating the dina model parameters using the no-u-turn sampler. **Biometrical Journal**, v. 60, n. 2, p. 352–368, 2018. Citado nas páginas 22 e 25.

SMITHSON, M.; SHOU, Y. Cdf-quantile distributions for modelling random variables on the unit interval. **British Journal of Mathematical and Statistical Psychology**, Wiley Online Library, v. 70, n. 3, p. 412–438, 2017. Citado na página 94.

SOCIALES, I. N. de Estadística e Informática (Peru). Dirección Técnica de Demografía e I. **Mapa de pobreza provincial y distrital 2009: el enfoque de la pobreza monetaria**. [S.l.]: Instituto Nacional de Estadística e Informática (Perú). Direción Técnica de . . . , 2010. Citado na página 109.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. V. D. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 64, n. 4, p. 583–639, 2002. Citado na página 58.

Stan Development Team. **RStan: the R interface to Stan, version 2.5.0**. 2014. Disponível em: <http://mc-stan.org/rstan.html>. Citado na página 100.

STEPHENS, M. Dealing with label switching in mixture models. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 62, n. 4, p. 795–809, 2000. Citado na página 53.

SU, Y.-S.; YAJIMA, M. R2jags: A package for running jags from r. **R package version 0.03-08, URL http://CRAN. R-project. org/package= R2jags**, 2012. Citado na página 57.

TEAM, R. C. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2017. Disponível em: <http://www.R-project.org/>. Citado na página 27.

TEMPLIN, J.; HENSON, R. Measurement of psychological disorders using cognitive diagnosis models. **Psychological Methods**, v. 11, n. 3, p. 287–305, 2006. Citado nas páginas 21, 26 e 27.

THOMAS, A. **OpenBUGS: Constructing MCMC Software**. John and Sons: Wiley, 2008. Citado na página 27.

TORRE, J. D. L. An empirically based method of q-matrix validation for the dina model: Development and applications. **Journal of educational measurement**, Wiley Online Library, v. 45, n. 4, p. 343–362, 2008. Citado na página 41.

____. The generalized DINA model framework. **Psychometrika**, Springer, v. 76, n. 2, p. 179–199, 2011. Citado na página 21.

____. The generalized dina model framework. **Psychometrika**, v. 76, n. 2, p. 179–199, 2011. Citado na página 120.

TORRE, J. de la. DINA model and parameter estimation: a didactic. **Journal of Educational and Behavioural Statistics**, v. 34, p. 115–130, 2009. Citado nas páginas 21, 27, 35, 36 e 131.

TORRE, J. de la; ARK, L. van der; ROSSI, G. Analysis of clinical data from cognitive diagnosis modeling framework. **Measurement and Evaluation in Counseling and Develpment**, p. 1–16, 2015. Citado nas páginas 26 e 27.

TORRE, J. de la; DOUGLAS, J. A. Higher-order latent trait models for cognitive dignosis. **Psychometrika**, v. 69, p. 333–353, 2004. Citado nas páginas 27, 35 e 131.

TSODIKOV, A.; IBRAHIM, J.; YAKOVLEV, A. Estimating cure rates from survival data: an alternative to two-component mixture models. **Journal of the American Statistical Association**, Taylor & Francis, v. 98, n. 464, p. 1063–1078, 2003. Citado na página 95.

TU, D.; ZHENG, C.; CAI, Y.; GAO, X.; WANG, D. A polytomous model of cognitive diagnostic assessment for graded data. **International Journal of Testing**, Taylor & Francis, p. 1–21, 2017. Citado na página 48.

VEHTARI, A.; GELMAN, A.; GABRY, J. Practical bayesian model evaluation using leave-one-out cross-validation and waic. **Statistics and computing**, Springer, v. 27, n. 5, p. 1413–1432, 2017. Citado na página 101.

WALRAVEN, C. V.; BENNETT, C.; JENNINGS, A.; AUSTIN, P. C.; FORSTER, A. J. Proportion of hospital readmissions deemed avoidable: a systematic review. **Cmaj**, Can Med Assoc, v. 183, n. 7, p. E391–E402, 2011. Citado na página 70.

WANG, X.; SAHA, A.; DEY, D. K. **Bayesian Analysis of Joint Modeling Response Times with Dynamic Latent Ability in Educational Testing**. [S.l.], 2016. Citado na página 49.

WATANABE, S. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. **Journal of Machine Learning Research**, v. 11, n. 12, p. 3571–3594, 2010. Citado na página 58.

____. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. **Journal of Machine Learning Research**, v. 11, p. 3571–3594, 2010. Citado na página 100.

WU, L. **Mixed Effects Models for Complex Data**. [S.l.]: Chapman and Hall/CRC, 2009. Citado na página 97.

YP., W.; GORENSTEIN, C. Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. **Revista Brasileira de Psiquiatria**, v. 35, n. 4, p. 416–431, 2013. Citado na página 40.

YU, K.; LU, Z.; STANDER, J. Quantile regression: applications and current research areas. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley Online Library, v. 52, n. 3, p. 331–350, 2003. Citado na página 94.

ZHAN, P. Using JAGS for bayesian cognitive diagnosis models: A tutorial. **arXiv preprint arXiv:1708.02632**, 2017. Citado na página 22.

ZHAN, P.; JIAO, H.; LIAO, D. Cognitive diagnosis modelling incorporating item response times. **British Journal of Mathematical and Statistical Psychology**, Wiley Online Library, v. 71, n. 2, p. 262–286, 2018. Citado na página 67.

ZHAN, P.; LI, X.; WANG, W.-C.; BIAN, Y.; WANG, L. The multidimensional testlet-effect cognitive diagnostic models. **Acta Psychologica Sinica**, v. 47, n. 5, p. 689–701, 2015. Citado na página 67.

APÊNDICE

A

MÉTODOS DE ESTIMAÇÃO DO MODELO DINA

Nesse material suplementar são apresentados os quatro métodos de estimação do modelo DINA (JUNKER; SIJTSMA, 2001; TORRE; DOUGLAS, 2004) utilizados no Capítulo 2. São apresentados os algoritmos de *Expectation Maximization* (EM) (DEMPSTER; LAIRD; RUBIN, 1977), de Metropolis Hastings (MH) (METROPOLIS *et al.*, 1953; HASTINGS, 1970), de Gibbs Sampling (GS) (GEMAN; GEMAN, 1984; GELFAND; SMITH, 1990), bem como o *No-U-Turn sampler* (NUTS) (HOFFMAN; GELMAN, 2014), com uma breve explicação do método de Monte Carlo Hamiltoniano (HMC) (DUANE *et al.*, 1987; NEAL, 1994; NEAL, 2011), da qual sentimos falta na literatura, sendo que esse método originou o NUTS.

Algoritmo EM

Quando não há a possibilidade de obter resultados analíticos para a estimação de máxima verossimilhança, uma das possibilidades de métodos computacionais é o algoritmo EM.

O algoritmo EM pode ser utilizado com base na máxima verossimilhança marginal, como em George e Robitzsch (2015) e Torre (2009). Antes da primeira iteração do algoritmo, é necessário esecolher parâmetros iniciais para a estimação de $[g,s]$ e $P(\boldsymbol{\alpha}_c)$. Uma vez que esses parâmetros são escolhidos, o algoritmo se alterna entre os passos E e M até convergir, utilizando um critério de parada que avalia se a convergência foi alcançada.

No passo E, a posteriori é dada por

$$\mathrm{P}(\boldsymbol{\alpha}_c|\mathbf{Y}_i) = \frac{\mathrm{P}(\mathbf{Y}_i|\boldsymbol{\alpha}_c)\mathrm{P}(\boldsymbol{\alpha}_c|\boldsymbol{\pi})}{\sum_{m=1}^{C}\mathrm{P}(\mathbf{Y}_i|\boldsymbol{\alpha}_m)\mathrm{P}(\boldsymbol{\alpha}_m|\boldsymbol{\pi})},$$

e tem dois tipos de valores esperados derivadas dela. O primeiro é o número esperado de respondentes classificados nos perfis $\boldsymbol{\alpha}_c$, para o item j, $c = (1,2,\ldots,C), j = (1,2,\ldots,J)$. O

outro é o número de respondentes classificados nos perfis $\boldsymbol{\alpha}_c$, porém com a restrição de que esses respondentes tiveram respostas positivas ao item j.

No passo M, tanto os parâmetros de item como o da distribuição dos perfis de atributos são atualizados, consecutivamente. Inicialmente, a primeira derivada da log-verossimilhança é definida como zero. A derivada pode ser escrita em termos dos dois valores obtidos no primeiro passo, permitindo assim que os parâmetros de item sejam atualizados. Depois disso, o número esperado de respondentes no perfil $\boldsymbol{\alpha}_c$ é determinado, sendo utilizado como base para a atualização da probabilidade de possuir cada atributo e da distribuição dos perfis de atributos.

Uma vez que a convergência é alcançada, é possível derivar nesse segundo estágio o perfil individual por meio do modelo estimado, utilizando a classificação de máxima verossimilhança (MLE), esperança a posteriori (EAP) ou o estimador máxima a posteriori (MAP).

O procedimento por meio do algoritmo EM está disponível no pacote "CDM" (GEORGE *et al.*, 2016) no R.

Metropolis Hastings

As distribuições condicionais completas dos parâmetros condicionados aos dados e outros parâmetros são dadas por

$$p(\boldsymbol{\alpha}_i|\mathbf{Y}_i,\boldsymbol{\pi},\boldsymbol{\Omega}) \propto p(\mathbf{Y}_i|\boldsymbol{\alpha}_i,\boldsymbol{\Omega})\cdot\pi_c, \tag{A.1}$$

$$p(\boldsymbol{\pi}|\mathbf{Y}_i,\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_n,\boldsymbol{\Omega}) \propto p(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_n|\boldsymbol{\pi})\cdot p(\boldsymbol{\pi}), \tag{A.2}$$

$$p(\boldsymbol{\Omega}_j|\mathbf{Y}_i,\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_n,\boldsymbol{\pi}) \propto \left[\prod_{i=1}^{n} p(Y_{ij}|\boldsymbol{\alpha}_i,\boldsymbol{\Omega}_j)\right]\cdot p(\boldsymbol{\Omega}_j). \tag{A.3}$$

A seguir é apresentado um esboço do algoritmo de MH para estimar os parâmetros do modelo DINA. Na iteração t temos:

1. Para $\boldsymbol{\alpha}$, retire uma amostra para os valores candidatos de $\alpha_{ik}^{\star}$ de uma Bernoulli(0.5) e aceite $\boldsymbol{\alpha}^{\star}$ com probabilidade

$$p\left(\boldsymbol{\alpha}^{(t-1)},\boldsymbol{\alpha}^{\star}\right) = \min\left\{\frac{p(\mathbf{Y}_i|\boldsymbol{\alpha}_i^{\star},\boldsymbol{\Omega}^{(t-1)})\cdot p(\boldsymbol{\alpha}_i^{\star})}{p(\mathbf{Y}_i|\boldsymbol{\alpha}_i^{(t-1)},\boldsymbol{\Omega}^{(t-1)})\cdot p(\boldsymbol{\alpha}_i^{(t-1)})},1\right\}. \tag{A.4}$$

2. Para $\boldsymbol{\pi}$, retire uma amostra para os valores candidatos de $\boldsymbol{\pi}^{\star}$ de uma Dirichlet$(\boldsymbol{\delta}_0)$ e aceite $\boldsymbol{\pi}^{\star}$ com probabilidade

$$p\left(\boldsymbol{\pi}^{(t-1)},\boldsymbol{\pi}^{\star}\right) = \min\left\{\frac{p(\boldsymbol{\alpha}_1^{(t)},\ldots,\boldsymbol{\alpha}_n^{(t)}|\boldsymbol{\pi}^{\star})\cdot p(\boldsymbol{\pi}^{\star})}{p(\boldsymbol{\alpha}_1^{(t)},\ldots,\boldsymbol{\alpha}_n^{(t)}|\boldsymbol{\pi}^{(t-1)})\cdot p(\boldsymbol{\pi}^{(t-1)})},1\right\}. \tag{A.5}$$

3. Para $\boldsymbol{\Omega}_j$, $j = 1,2,\ldots,J$, retire uma amostra dos valores candidatos de $\boldsymbol{\Omega}_j^\star = (s_j^\star, g_j^\star)$ por meio da construção do produto de densidades Beta para $s_j^\star$ e $g_j^\star$ com parâmetros a_s, b_s e a_g, b_g respectivamente, e aceite $\boldsymbol{\Omega}_j^\star$ com probabilidade

$$p\left(\boldsymbol{\Omega}_j^{(t-1)}, \boldsymbol{\Omega}_j^\star\right) = \min\left\{\frac{\left[\prod_{i=1}^n p(Y_{ij}|\boldsymbol{\alpha}_i^{(t)}, \boldsymbol{\Omega}_j^\star)\right] \cdot p(\boldsymbol{\Omega}_j^\star)}{\left[\prod_{i=1}^n p(Y_{ij}|\boldsymbol{\alpha}_i^{(t)}, \boldsymbol{\Omega}_j^{(t-1)})\right] \cdot p(\boldsymbol{\Omega}_j^{(t-1)})}, 1\right\}. \tag{A.6}$$

Gibbs sampling

Para implementar o GS para o modelo DINA, consideramos a seguinte distribuição conjunta a posteriori

$$p(\boldsymbol{\Omega}, \boldsymbol{\pi}|\mathbf{Y}) \propto \ell(\mathbf{Y}|\boldsymbol{\Omega}, \boldsymbol{\pi}) p(\boldsymbol{\Omega}) p(\boldsymbol{\pi}),$$

onde $\ell(\mathbf{Y}|\boldsymbol{\Omega}, \boldsymbol{\pi})$ é a função da verossimilhança, $p(\boldsymbol{\Omega})$ é a priori para os parâmetros de item e $p(\boldsymbol{\pi})$ é a distribuição a priori para os perfis de atributos.

As distribuições condicionais completas de $\boldsymbol{\alpha}_i$, $\boldsymbol{\pi}$ e $\boldsymbol{\Omega}_j$ podem ser derivadas em formas fechadas como

$$p(\boldsymbol{\alpha}_i|\mathbf{Y}_i, \boldsymbol{\pi}, \boldsymbol{\Omega}) = p(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c|\mathbf{Y}_i, \boldsymbol{\pi}, \boldsymbol{\Omega}) = \prod_{c=1}^C \pi_{ic}^{\mathbb{1}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c)}, \tag{A.7}$$

onde $\mathbb{1}(\cdot)$ denota a função indicadora,

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n) \propto \prod_c^C \pi_c^{N_c+1}, \tag{A.8}$$

onde $N_c = \sum_{i=1}^n \mathbb{1}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c)$,

$$p(\boldsymbol{\Omega}_j|\mathbf{Y}_i, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n) \propto s_j^{\tilde{a}_s - 1}(1-s_j)^{\tilde{b}_s - 1} g_j^{\tilde{a}_g - 1}(1-g_j)^{\tilde{b}_s - 1} \mathbb{1}((s_j, g_j) \in \mathscr{P}), \tag{A.9}$$

onde $\tilde{a}_s = a_s + \sum_{i|y_{ij}=0}^n \eta_{ij}$, $\tilde{b}_s = b_s + \sum_{i=1}^n \eta_{ij} - \sum_{i|y_{ij}=0}^n \eta_{ij}$, $\tilde{a}_g = a_g + \sum_{i|y_{ij}=1}^n (1-\eta_{ij})$,
$\tilde{b}_g = n + b_g - \sum_{i=1}^n \eta_{ij} - \sum_{i|y_{ij}=1}^n (1-\eta_{ij})$ e $\mathscr{P} = \{(s,g) : 0 \leq s+g < 1, 0 \leq s < 1, 0 \leq g < 1\}$.
Lembre que η_{ij} depende de $\boldsymbol{\alpha}_i$.

Assim, com valores iniciais $\boldsymbol{\alpha}_i$, $\boldsymbol{\pi}$ e $\boldsymbol{\Omega}_j$, podemos simular $\left(\boldsymbol{\alpha}_i^{(t)}, \boldsymbol{\pi}^{(t)}, \boldsymbol{\Omega}_j^{(t)}\right)$ por meio do amostrador de Gibbs retirando, iterativamente, amostras de suas respectivas distribuições condicionais completas especificadas em (A.7), (A.8) e (A.9). Para ir de $\left(\boldsymbol{\alpha}_i^{(t-1)}, \boldsymbol{\pi}^{(t-1)}, \boldsymbol{\Omega}_j^{(t-1)}\right)$ para $\left(\boldsymbol{\alpha}_i^{(t)}, \boldsymbol{\pi}^{(t)}, \boldsymbol{\Omega}_j^{(t)}\right)$, há três passos de transição

1. Amostre $\boldsymbol{\alpha}_i^{(t)} \sim p\left(\boldsymbol{\alpha}_i|\mathbf{Y}_i, \boldsymbol{\pi}^{(t-1)}, \boldsymbol{\Omega}^{(t-1)}\right)$;
2. Amostre $\boldsymbol{\pi}^{(t)} \sim p\left(\boldsymbol{\pi}|\boldsymbol{\alpha}_1^{(t)}, \ldots, \boldsymbol{\alpha}_n^{(t)}\right)$;

3. Amostre $\boldsymbol{\Omega}_j^{(t)} \sim p\left(\boldsymbol{\Omega}_j | \mathbf{Y}_i, \boldsymbol{\alpha}_1^{(t)}, \ldots, \boldsymbol{\alpha}_n^{(t)}\right)$.

Para amostrar $\boldsymbol{\Omega}_j = (s_j, g_j)$, temos a restrição de Monotonicidade $(0 \leq s_j + g_j < 1)$. Por isso, o passo 3 deve ser performado por meio de dois passos

3a. Amostre $g_j^{(t)} | s_j^{(t-1)} \sim \text{Beta}(\tilde{a}_g, \tilde{b}_g) \mathbb{1}(0 \leq g_j^{(t)} < 1 - s_j^{(t-1)})$;

3b. Amostre $s_j^{(t)} | g_j^{(t)} \sim \text{Beta}(\tilde{a}_g, \tilde{b}_g) \mathbb{1}(0 \leq s_j^{(t)} < 1 - g_j^{(t)})$.

Esse processo iterativo produz uma sequência de $\left(\boldsymbol{\alpha}_i^{(t)}, \boldsymbol{\pi}^{(t)}, \boldsymbol{\Omega}_j^{(t)}\right)$, $l = 0, \ldots, T$.. Para reduzir o efeito dos valores iniciais, as iterações iniciais são descartadas. As amostras das iterações seguintes são então utilizadas para sumarizar a densidade a posteriori dos parâmetros.

Breve explicação do algoritmo de Monte Carlo Hamiltoniano

Antes de apresentar o Monte Carlos Hamiltoniano, é interessante falar um pouco sobre a ideia por trás da dinâmica Hamiltoniana. A dinâmica Hamiltoniana pode ser usada para descrever o movimento de um objeto por meio de sua localização, dada por um vetor $\mathbf{x}$ e momento (massa vezes velocidade), dada por um vetor $\mathbf{p}$ em um instante t. As dimensões desses vetores correspondem às dimensões do problema considerado. Associada a cada localização do objeto, há uma energia potencial denotada por $U(\mathbf{x})$, e associada com cada momento do objeto há uma energia cinética denotada por $K(\mathbf{p})$. O Hamiltoniano é definido como $H(\mathbf{x}, \mathbf{p}) = U(\mathbf{x}) + K(\mathbf{p})$. Podemos determinar a localização e o momento de um objeto em qualquer tempo t por meio das derivadas parciais

$$\frac{\partial x_i}{\partial t} = \frac{\partial H}{\partial p_i} = \frac{\partial K(\mathbf{p})}{\partial p_i}, \tag{A.10}$$

$$\frac{\partial p_i}{\partial t} = -\frac{\partial H}{\partial x_i} = -\frac{\partial U(\mathbf{x})}{\partial x_i}, \tag{A.11}$$

para $i = 1, 2, \ldots, d$. De (A.10) e (A.11) podemos determinar a localização e o momento de um objeto em qualquer intervalo de tempo com curação T, começando no tempo t_0 e indo até o tempo $t = t_0 + T$.

A implementação computacional da dinâmica Hamiltoniana pode ser feita por meio de uma aproximação das equações através da discretização do tempo, dividindo o intervalo de tempo em pequenas fatias de tamanho δ. Essa aproximação se torna cada vez mais acurada quando o valor de δ diminui. A abordagem utilizada no Monte Carlo Hamiltoniano para realizar a discretização do tempo é conhecida como o método de *leap frog*. De maneira resumida, esse método consiste em três passos descritos a seguir.

1. Comece atualizando as variáveis de momento em um pequeno intervalo de tempo $\delta/2$

$$p_i(t + \delta/2) = p_i(t) - (\delta/2)\frac{\partial U}{\partial x_i(t)}. \tag{A.12}$$

2. Atualize as variáveis de posição no intervalo de tempo δ utilizando os novos valores das variáveis de momento.

$$x_i(t+\delta) = x_i(t) + \delta \frac{\partial K}{\partial p_i(t+\delta/2)}. \tag{A.13}$$

3. Complete a atualização das variáveis de momento no intervalo de tempo $\delta/2$

$$p_i(t+\delta) = p_i(t+\delta/2) - (\delta/2) \frac{\partial U}{\partial x_i(t+\delta)}. \tag{A.14}$$

A ideia do Monte Carlo Hamiltoniano é usar a dinâmica Hamiltoniana para amostrar valores de uma distribuição $p(\mathbf{x})$ por meio do conceito de distribuição canônica da mecânica estatística. Definimos a distribuição canônica para qualquer função de energia $E(\boldsymbol{\theta})$ sobre um conjunto de variáveis $\boldsymbol{\theta}$ como

$$p(\boldsymbol{\theta}) = \frac{1}{Z} e^{-E(\boldsymbol{\theta})}, \tag{A.15}$$

onde Z é uma constante de normalização. Como distribuições de probabilidade de qualquer escala podem ser amostradas por métodos MCMC e a função de energia na dinâmica Hamiltoniana é uma combinação de energias potenciais e cinéticas, i.e., $E(\boldsymbol{\theta}) = H(\mathbf{x}, \mathbf{p}) = U(\mathbf{x}) + K(\mathbf{p})$, podemos escrever as distribuições canônicas da função de energia para a dinâmica Hamiltoniana como segue

$$\begin{aligned} p(\mathbf{x}, \mathbf{p}) &\propto e^{-H(\mathbf{x},\mathbf{p})} \\ &= e^{-[U(\mathbf{x})+K(\mathbf{p})]} \\ &= e^{-U(\mathbf{x})} \cdot e^{-K(\mathbf{p})} \\ &\propto p(\mathbf{x}) \cdot p(\mathbf{p}). \end{aligned} \tag{A.16}$$

Note que $\mathbf{x}$ e $\mathbf{p}$ são independentes. A função da energia potencial é definida por

$$U(\mathbf{x}) = -\log\left[\pi(\mathbf{x}) \cdot \ell(\mathbf{x}|D)\right], \tag{A.17}$$

onde $\pi(\mathbf{x})$ é a densidade da priori e $\ell(\mathbf{x}|D)$ é a função de verossimilhança condicionada aos dados D. A energia cinética é definida como

$$K(\mathbf{p}) = \sum_{i=1}^{d} \frac{p_i^2}{2}. \tag{A.18}$$

As variáveis de localização $\mathbf{x}$ são utilizadas para representar as variáveis de interesse e as variáveis de momento $\mathbf{p}$ são utilizadas como variáveis auxiliares que permitem a dinâmica Hamiltoniana funcionar. Como $\mathbf{x}$ e $\mathbf{p}$ são independentes, podemos escolher qualquer distribuição para as variáveis auxiliares $\mathbf{p}$. A opção mais comum é utilizar a distribuição normal padrão.

O HMC se inicia a partir da simulação de valores $\mathbf{x}^{(0)}$ para a variável de interesse baseada nas distribuições a priori. Então, com uma estrutura de repetição, os valores de $\mathbf{p}_0$ são obtidos por meio de variáveis auxiliares de uma distribuição normal padrão. Para cada iteração t são

consideradas como variáveis de interesse os valores de $\mathbf{x}_0 = \mathbf{x}^{(t-1)}$, onde $\mathbf{x}^{(t-1)}$ vem da iteração $t-1$. Para a iteração $t = 1$, o valor considerado é inicialmente o valor simulado $\mathbf{x}^{(0)}$, isto é, $\mathbf{x}_0 = \mathbf{x}^{(0)}$. Então o método de *leap frog* é usado com os valores $\mathbf{x}_0$ e $\mathbf{p}_0$ para os passos L e tamanho de passo δ, para obter os valores propostos $\mathbf{x}^*$ e $\mathbf{p}^*$. Com esses valores propostos, a probabilidade de aceitação é calculada, sendo esse um processo análogo ao que ocorre no MH, dado por

$$\alpha = \min(1, \exp(-U(\mathbf{x}^*) + U(\mathbf{x}_0) - K(\mathbf{p}^*) + K(\mathbf{p}_0))). \tag{A.19}$$

Um número aleatório u é simuado de uma Uniforme$(0,1)$ e se $u \leq \alpha$, $\mathbf{x}^*$ é aceito e define o próximo estado da cadeia de Markov $\mathbf{x}^{(t)} = \mathbf{x}^*$. Caso contrário, $\mathbf{x}^{(t)} = \mathbf{x}^{(t-1)}$. Retorne ao início da estrutura de repetição T vezes.

O tamanho dos passos δ e o número de passos L são parâmetros definidos pelo usuário do HMC. Essa tarefa é difícil uma vez que requere conhecimento especializado. A escolha errada nos parâmetros pode causar problemas, como a Tabela 34 mostra.

Tabela 34 – Possíveis problemas causados pela escolha errada dos parâmetros do HMC

Parâmetros	excessivamente pequeno	excessivamente grande
δ	poder computacional desperdiçado (passos pequenos)	simulação não acurada e alta taxa de rejeição
L	comportamento de passeio aleatório e mistura lenta	trajetórias refazem seus passos

APÊNDICE

B

CÓDIGO EM STAN PARA ESTIMAÇÃO DO MODELO DINA

O código em STAN para a estimação do modelo DINA é apresentado a seguir

```
data {
  int<lower=1> N;
  int<lower=1> J;
  int<lower=1> K;
  int<lower=1> C;
  int<lower=0, upper=1> Y_sim[N, J];
  int<lower=0, upper=1> Q[J, K];
  int<lower=0, upper=1> As[C, K];
}
transformed data{
  vector[C] delta;
  delta = rep_vector(1, C);
}
parameters {
  simplex[C] pi;
  simplex[3] Omega[J];
}
transformed parameters {
  vector<lower=0, upper=1>[J] prob[C];
  vector<lower=0, upper=1>[J] eta[C];
  vector<lower=0, upper=1>[J] g;
  vector<lower=0, upper=1>[J] s;
```

```
  vector[C] lp[N];
  g = to_vector(Omega[, 1]);
  s = to_vector(Omega[, 2]);
  for (c in 1:C) {
    for (j in 1:J) {
  eta[c, j] = 1;
  for (k in 1:K)
    eta[c, j] = eta[c, j] * As[c, k] ^ Q[j, k];
      prob[c, j] = (1-s[j])^eta[c, j]*(g[j])^(1-eta[c, j]);
    }
  }
  for (i in 1:N) {
    for (c in 1:C) {
  lp[i, c] = log(pi[c]);
  for (j in 1:J)
    lp[i, c] = lp[i, c] + bernoulli_log(Y_sim[i, j], prob[c, j]);
}
  }
}
model {
  for (j in 1:J) {
    Omega[j, 1] ~ beta(1, 1);
    Omega[j, 2] ~ beta(1, 1);
  }
  pi ~ dirichlet(delta);
  for (i in 1:N)
    target+=(log_sum_exp(lp[i, 1:C]));
}
generated quantities {
  int<lower=1, upper=C> alpha_star[N];
  vector<lower=0, upper=1>[J] u;
  for (i in 1:N)
    alpha_star[i] = categorical_rng(softmax(lp[i]));
  u = 1-s;
}
```

APÊNDICE

C

ESQUEMA MCMC PARA O MODELO C-DINA

O esquema MCMC utilizado para retirar amostras da posteriori conjunta (3.15) é um esquema de *Gibbs sampling* com *slice sampling* para os parâmetros λ_js. Para o uso do *Gibbs sampling*, derivamos as distribuições condicionais completas para o modelo C-DINA, que podem ser vistas a seguir.

C.1 Parâmetros dos indivíduos

- Para cada parâmetro dos itens $\boldsymbol{\alpha}_i$, a distribuição condicional completa dado $\boldsymbol{\Omega}$, $\boldsymbol{\pi}$ e os dados $\boldsymbol{y}$ é uma distribuição categórica com probabilidade de pertença à cada categoria c sendo $\widetilde{\pi}_{ic}$, i.e.,

$$p(\boldsymbol{\alpha}_i \mid \boldsymbol{\Omega}, \boldsymbol{\pi}, \boldsymbol{y}_i) = \prod_{c=1}^{C} \widetilde{\pi}_{ic}^{\mathbb{1}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c)}, \tag{C.1}$$
$$\text{with } \widetilde{\pi}_{ic} = \frac{\pi_c p(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\Omega})}{\sum_{c=1}^{C} \pi_c p(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\Omega})},$$

  onde $p(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\Omega})$ é definido na Equação (3.4).

- Dado $\boldsymbol{\alpha}$, a distribuição condicional completa do parâmetro dos indivíduos $\boldsymbol{\pi}$ é dada por Dirichlet$(\widetilde{\boldsymbol{N}} + \boldsymbol{\delta}_0)$, isto é

$$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \propto \prod_{c=1}^{C} \pi_c^{\widetilde{N}_c + \delta_{0c}} \tag{C.2}$$

  com $\widetilde{\boldsymbol{N}}' = (\widetilde{N_1}, \ldots, \widetilde{N_C}) = (\sum_{i=1}^{N} \mathbb{1}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_1), \ldots, \sum_{i=1}^{N} \mathbb{1}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_C))$.

C.2 Parâmetros dos Itens

Para cada parâmetro dos itens $\boldsymbol{\Omega}_j = (\mu_{j0}, \lambda_j, \tau_{j0}, \tau_{j1})'$ e hiperparâmetro τ_{μ_0}, também é possível avaliar a distribuição condicional completa.

- Dado $\boldsymbol{y}$, $\boldsymbol{\alpha}$, τ_{μ_0}, λ_j, τ_{j0}, τ_{j1}, a distribuição condicional completa para μ_{j0} é uma distribuição Gaussiana, i.e.,

$$\left[\mu_{j0} \mid \boldsymbol{y}, \boldsymbol{\alpha}, \tau_{\mu_0}, \lambda_j, \tau_{j0}, \tau_{j1}\right] \sim \text{Normal}\left(\frac{B_j}{2A_j}, \frac{1}{A_j}\right) \tag{C.3}$$

  onde $A_j = (N - \sum_{i=1}^{N} \eta_{ij})\tau_{j0}/2 + \sum_{i=1}^{N} \eta_{ij}\tau_{j1}/2 + \tau_{\mu_0}/2$ e $B_j = (N - \sum_{i=1}^{N} \eta_{ij})\tau_{j0} \times \log(y_{ij}) + \sum_{i=1}^{N} \eta_{ij}\tau_{j1}(\log(y_{ij} - \lambda_j))$.

- Dado que μ_{j0} é conhecido, a distribuição condicional completa para τ_{μ_0} é uma distribuição Gama, a qual é dada por

$$\left[\tau_{\mu_0} \mid \mu_{j0}\right] \sim \text{Gama}\left(J/2 + c_{\mu_0}, \sum_{j=1}^{J} \mu_{j0}^2/2 + d_{\mu_0}\right) \tag{C.4}$$

- Assumindo $\boldsymbol{y}$, $\boldsymbol{\alpha}$, μ_{j0} e τ_{j1} conhecidos, a distribuição condicional completa de λ_j não possui forma fechada, mas é proporcional a

$$p(\lambda_j \mid \boldsymbol{y}, \boldsymbol{\alpha}, \mu_{j0}, \tau_{j1}) \quad \propto \quad \lambda_j^{a_\lambda - 1} \exp\left[-T_j^2\left(\lambda_j - \frac{W_j}{2T_j}\right)^2\right] \tag{C.5}$$

  onde $T_j = \sum_{i=1}^{N} \eta_{ij}\tau_{j1}/2$ and $W_j = 2T_j\left[\log(y_{ij}) - \mu_{j0}\right] - b_\lambda$.

- Dado $\boldsymbol{y}$, $\boldsymbol{\alpha}$ e μ_{j0}, a distribuição condicional completa de τ_{j0} é uma distribuição Gama, i.e.,

$$\left[\tau_{j0} \mid \boldsymbol{y}, \boldsymbol{\alpha}, \mu_{j0}\right] \sim \text{Gama}\left(\frac{N}{2} - \sum_{i=1}^{N} \frac{\eta_{ij}}{2} + c_0, \sum_{i=1}^{N} (1 - \eta_{ij}) \frac{(\log(y_{ij}) - \mu_{j0})^2}{2} + d_0\right) \tag{C.6}$$

- A distribuição condicional completa de τ_{j1} dado $\boldsymbol{y}$, $\boldsymbol{\alpha}$, λ_j e μ_{j0} também é uma Gama, a qual é

$$\left[\tau_{j1} \mid \boldsymbol{y}, \boldsymbol{\alpha}, \mu_{j0}, \lambda_j\right] \sim \text{Gama}\left(\sum_{i=1}^{N} \frac{\eta_{ij}}{2} + c_1, \sum_{i=1}^{N} \eta_{ij} \frac{(\log(y_{ij}) - \mu_{j0} - \lambda_j)^2}{2} + d_1\right) \tag{C.7}$$

Para implementar o *Gibbs sampling*, começamos em (C.1), com valores iniciais para $\boldsymbol{\pi}$, μ_{j0}s, λ_js, τ_{j0}s e τ_{j1}s, percorrendo até (C.7) até que haja convergência. Note que no Passo (C.5) será utilizado o *slice sampling* para amostrar λ_j de sua distribuição condicional completa.

APÊNDICE

D

CÓDIGO EM JAGS PARA ESTIMAÇÃO DO MODELO C-DINA

O código JAGS para o modelo C-DINA se encontra a seguir

```
model{
 for (i in 1:N) {
  for (j in 1:J) {
    for (k in 1:K) {w[i, j, k] <- pow(alpha[i, k], Q[j, k])}
    eta[i, j] <- prod(w[i, j, 1:K])
    u[i, j] <- ifelse(eta[i,j]==1, u0[j]+lambda[j], u0[j])
    sigma2[i, j] <- ifelse(eta[i,j]==1, sigma21[j], sigma20[j])
    tau[i, j] <- 1/sigma2[i, j]
    #verossimilhança
    Y[i, j] ~ dlnorm(u[i,j], tau[i,j])
    }

  #prioris para parâmetros de indivíduo
  for (k in 1:K) {alpha[i, k] <- all.patterns[c[i], k]}
    c[i] ~ dcat(pai[1:C])}
  pai[1:C] ~ ddirch(delta[1:C])

  #prioris para parâmetros de item
  taumu0 ~ dgamma(0.01,0.01)
  sigma2mu0 <- 1/taumu0
  for (j in 1:J) {
    u0[j] ~ dnorm(0, taumu0)
```

```
tau0[j] ~ dgamma(0.01,0.01)
sigma20[j] <- 1/tau0[j]
lambda[j] ~ dgamma(0.01, 0.01)
tau1[j] ~ dgamma(0.01,0.01)
sigma21[j] <- 1/tau1[j]
u1[j] <- u0[j] + lambda[j]
 }}
```

APÊNDICE

E

CÓDIGO EM JAGS PARA ESTIMAÇÃO DO MODELO B-DINA

O código em JAGS para a estimação do modelo B-DINA, seguindo as prioris estipuladas, é apresentado a seguir

```
model{
 for (i in 1:N) {
  for (j in 1:J) {
    for (k in 1:K) {w[i, j, k] <- pow(alpha[i, k], Q[j, k])}
    eta[i, j] <- prod(w[i, j, 1:K])
    u[i, j] <- ifelse(eta[i,j]==1, u1[j], u0[j])
    phif[i, j] <- ifelse(eta[i,j]==1, phi1[j], phi0[j])
    p[i,j] <- u[i,j]*phif[i,j]
    q[i,j] <- (1-u[i,j])*phif[i,j]
    #Likelihood
    Y[i,j] ~ dbeta(p[i, j], q[i, j])
    }

  #Prioris para parâmetros de indivíduo
  for (k in 1:K) {alpha[i, k] <- all.patterns[c[i], k]}
    c[i] ~ dcat(pai[1:C])}
  pai[1:C] ~ ddirch(delta[1:C])

  #Priors para parâmetros de item
 for (i in 1:2){
  for (j in 1:J) {
```

```
    muaux[i,j] ~ dbeta(1, 1)
    }
}
  for (j in 1:J){
     mu[1:2,j] <- sort(muaux[1:2,j])
     u1[j] <- mu[2,j]
     u0[j] <- mu[1,j]
     phi0[j] ~ dgamma(0.01,0.01)
     phi1[j] ~ dgamma(0.01,0.01)
}
}
```

APÊNDICE

F

CÓDIGO EM STAN PARA REGRESSÃO MISTA UTILIZANDO A DISTRIBUIÇÃO LG

O código para os modelos M4 sob distribuição LG são dados a seguir. Ao utilizar o código, os hiperparâmetros para as prioris precisam ser especificados, bem como o valor do quantil q de interesse. Para a distribuição CLG, basta uma alteração na Log-Verossimilhança

```
data {
    int<lower=0> n; // número de observações
    int<lower=0> M; // número de grupos
    real<lower=0,upper=1> y[n]; // variável resposta
    real x1[n]; // cováriavel
    real x2[n]; // cováriavel
    int<lower=0> id[n]; // id para o efeito aleatório
    real<lower=0,upper=1> q; // quantil
}
parameters {
    real delta0;
    real delta1;
    real delta2;
    real beta0;
    real beta1;
    real beta2;
    real<lower=0> sigma2b;
    real<lower=0> sigma2d;
    real bib[M];
    real bid[M];
```

```
}
transformed parameters {
    real<lower=0> sigmab;
    real<lower=0> sigmad;
    sigmab = sqrt(sigma2b);
    sigmad = sqrt(sigma2d);
}
model {
    real kappa[n];
    real phi[n];
    real a[n];
    real b[n];
    for(j in 1:M){
    bib[j] ~ normal(0,sigmab);
    bid[j] ~ normal(0,sigmad);
}
    beta0 ~ normal(0, 100);
    beta1 ~ normal(0, 100);
    beta2 ~ normal(0, 100);
    delta0 ~ normal(0, 100);
    delta1 ~ normal(0, 100);
    delta2 ~ normal(0, 100);
    sigma2b ~ inv_gamma(0.01, 0.01);
    sigma2d ~ inv_gamma(0.01, 0.01);
    for(i in 1:n){
        a[i] = -delta0 - delta1 * x1[i] - delta2*x2[i] - bid[id[i]];
        kappa[i] = inv_logit(beta0 + beta1 * x1[i] + beta2 * x2[i] + bib[id[i]]);
        b[i] = (a[i]*log(q))/(1-pow(kappa[i],-a[i]));
        target+= (log(b[i])+(-a[i]-1)*log(y[i])-(b[i]/a[i])*(pow(y[i],-a[i])-1));
}
}
```