UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Predição conforme para dados composicionais

Lucas Pereira do Amaral

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP
SÃO CARLOS

Lucas Pereira do Amaral

Predição conforme para dados composicionais

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *EXEMPLAR DE DEFESA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Gustavo Henrique de Araújo Pereira

USP – São Carlos
Janeiro de 2026

Amaral, Lucas Pereira do

Predição conforme para dados composicionais / Lucas Pereira do Amaral -- 2026.
69f.

Dissertação (Mestrado) - Universidade Federal de São Carlos, campus Lagoa do Sino, Buri
Orientador (a): Gustavo Henrique de Araújo Pereira
Banca Examinadora: Gustavo Henrique de Araújo Pereira, Helton Graziadei de Carvalho, Tiago Maia Magalhães
Bibliografia

1. Dados composicionais. 2. Predição conforme. 3. Regressão Dirichlet. I. Amaral, Lucas Pereira do. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Lissandra Pinhatelli de Britto - CRB/8 7539

UFSCar

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas

Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Lucas Pereira do Amaral, realizada em 24/02/2026.

Comissão Julgadora:

Prof. Dr. Gustavo Henrique de Araújo Pereira (UFSCar)

Prof. Dr. Helton Graziadei de Carvalho (UFSCar)

Prof. Dr. Tiago Maia Magalhães (UFJF)

O Relatório de Defesa produzido pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação.

Lucas Pereira do Amaral

Conformal prediction for compositional data

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Statistics

Advisor: Prof. Dr. Gustavo Henrique de Araújo Pereira

USP – São Carlos
January 2026

Este trabalho é dedicado a quem acredita na ciência e na educação,
a quem escolheu estar do lado certo da História,
a quem tem esperança de dias melhores,
a quem não fez da educação mercadoria.

AGRADECIMENTOS

Primeiramente, gostaria de agradecer à minha família pelo apoio incondicional. À minha mãe, pelo amor e carinho independentemente das minhas escolhas. Ao meu pai, pelo apoio na ida a São Carlos, pois foi quem embarcou comigo e me transmitiu a segurança de ter sempre um porto seguro para onde voltar. À minha irmã, pelo apoio, e ao meu irmão, pela companhia.

Ainda em relação à família, um agradecimento especial à Ana Caroline, meu romance e amor da minha vida, por ter permanecido ao meu lado mesmo nos momentos mais difíceis.

Ao meu orientador, Prof. Dr. Gustavo Henrique, obrigado por toda a paciência, compreensão e humanidade. A universidade certamente deveria ter mais professores como o senhor.

Ao meu coorientador não oficial, Dr. Thiago Ramos, obrigado por ter embarcado nesta aventura conosco, a UFSCar tem muita sorte em tê-lo no corpo docente.

Aos amigos da pós-graduação que me auxiliaram nos momentos mais árduos: Riquelme, Abderraman, Rafaela, Fernanda e Guilherme. Ao amigo Luben Miguel, que me ajudou na pesquisa atuando quase como um coorientador.

Aos professores do Programa Interinstitucional de Pós-Graduação em Estatística da USP/UFSCar, por proporcionarem disciplinas desafiadoras e transformadoras em minha jornada.

Aos membros da banca examinadora pelas excelentes sugestões.

Ao meu professor e amigo da Universidade Federal do Ceará, Dr. Gualberto Agamez, por ter me aconselhado a buscar desafios maiores e, diante destes, nunca me ter deixado dar por vencido.

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

“Eu não sei para onde estou indo, mas estou a caminho.”

- Roronoa Zoro

RESUMO

AMARAL, L. P. **Predição conforme para dados composicionais**. 2026. 69 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

Modelos de regressão Dirichlet são apropriados para dados composicionais, nos quais a variável resposta representa proporções que somam um. No entanto, ainda não existem métodos consolidados para a construção de conjuntos de predição válidos nesse contexto, especialmente levando em conta a geometria do espaço composicional. Neste trabalho, investigamos estratégias baseadas em predição conforme para a construção de regiões preditivas válidas em modelos de regressão Dirichlet. Avaliamos três abordagens distintas: um método baseado em resíduos quantílicos, uma construção aproximada de regiões de maior densidade (HDR) e uma adaptação da HDR aproximada via discretização por *grid* no simplex. A performance dos métodos foi analisada por meio de estudos de simulação em diferentes cenários, variando a complexidade do modelo, a dimensionalidade da resposta e a estrutura das covariáveis. Os resultados indicaram que a abordagem de aproximação da HDR apresenta boa robustez em termos de cobertura, enquanto a discretização por *grid* se mostrou eficaz para reduzir a sobrecobertura e a área da região de predição em relação ao método original. O método quantílico proporcionou regiões de previsão maiores em comparação com o método via *grid*, mantendo, ao mesmo tempo, uma cobertura adequada. As metodologias também foram aplicadas a dois conjuntos de dados reais: um sobre estágios do sono e outro sobre alocação de biomassa em plantas. Em ambos os casos, os métodos propostos demonstraram viabilidade prática e produziram interpretações coerentes dentro do espaço composicional. Por fim, discutimos possíveis extensões desse trabalho.

Palavras-chave: Dados composicionais, Predição conforme, Regressão Dirichlet, Regressão multivariada.

ABSTRACT

AMARAL, L. P. **Conformal prediction for compositional data**. 2026. 69 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

Dirichlet regression models are suitable for compositional data, in which the response variable represents proportions that sum to one. However, there are still no well-established methods for constructing valid prediction sets in this context, especially considering the geometry of the compositional space. In this work, we investigate conformal prediction-based strategies for constructing valid predictive regions in Dirichlet regression models. We evaluate three distinct approaches: a method based on quantile residuals, an approximate construction of highest density regions (HDR), and an adaptation of the approximate HDR using grid-based discretization over the simplex. The performance of the methods was analyzed through simulation studies under different scenarios, varying the model complexity, response dimensionality, and covariate structure. The results indicated that the HDR approximation approach exhibits good robustness in terms of coverage, while the grid discretization proved effective in reducing overcoverage and the area of the prediction region compared to the original method. The quantile method provided larger prediction regions compared to the grid method, while maintaining adequate coverage. The methodologies were also applied to two real datasets: one concerning sleep stages and another on biomass allocation in plants. In both cases, the proposed methods demonstrated practical feasibility and produced coherent interpretations within the compositional space. Finally, we discuss possible extensions of this work.

Keywords: Compositional data, Conformal prediction, Dirichlet regression, Multivariate regression.

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Dados composicionais são amplamente utilizados em diversas áreas quando o interesse está em modelar variáveis que expressam partes relativas de um total, como proporções ou frações que somam uma unidade. Essa abordagem tem sido utilizada, por exemplo, em problemas de microbiologia com dados de microbioma gerados por sequenciamento de alto desempenho (GLOOR *et al.*, 2017), na química para a construção de um índice de poluição de águas subterrâneas (OH *et al.*, 2024), e na geologia para o estudo da composição geoquímica de minérios (GOTTSCHALK, 2024).

Define-se um dado composicional pela formação de D números positivos $\mathbf{y} = (y_1, \ldots, y_D)^\top$ tais que $\sum_{j=1}^{D} y_j = 1$. Sendo o conjunto

$$\Delta^D = \{(y_1, \ldots, y_D) \in \mathbb{R}^D : \sum_{j=1}^{D} y_j = 1,\ y_j > 0,\ j = 1, \ldots, D\}$$

denominado de $D-1$-Simplex, $y_j > 0$ evidencia que é o Simplex aberto, não incluindo $y_j = 0$.

Dados composicionais podem ser analisados de diversas formas, mas devido às restrições que impõem, sua modelagem é mais complexa. No entanto, essa área tem evoluído significativamente ao longo dos anos, como detalhado em Alenazi (2023). Esses dados possuem um espaço amostral próprio, o Simplex (AITCHISON, 1982), e é comum a utilização da distribuição de Dirichlet (BARNDORFF-NIELSEN; JØRGENSEN, 1991) para modelá-los. Em análises de regressão, frequentemente se recorre aos modelos de regressão Dirichlet (HIJAZI; JERNIGAN, 2009), que podem ser vistos como uma generalização dos modelos de regressão beta (FERRARI; CRIBARI-NETO, 2004).

A modelagem por meio de modelos paramétricos costuma priorizar a interpretabilidade dos coeficientes e a verificação das suposições feitas sobre os dados, pois essas impactam diretamente a qualidade do ajuste. Ainda assim, a capacidade preditiva também deve ser considerada, já que pode ser de interesse do pesquisador realizar predições confiáveis.

Tarefas preditivas desempenham um papel central em diversos contextos aplicados, podendo envolver tanto predições pontuais quanto intervalares. Neste trabalho, o foco está na construção de predições intervalares para o caso de dados composicionais. Apesar da relevância prática desse tipo de dado, métodos para predição intervalar ainda são pouco explorados nesse contexto. No caso univariado, em que os dados estão restritos ao intervalo unitário $(0,1)$, já existem abordagens bem estabelecidas baseadas em reamostragem via *bootstrap*, utilizando resíduos como medida de erro preditivo (ESPINHEIRA; FERRARI; CRIBARI-NETO, 2014; CRIBARI-NETO; LIMA, 2021).

Recentemente, Wu, Leisen e Rubio (2025) propuseram métodos de predição intervalar voltados para dados limitados, dentro do arcabouço da predição conforme (VOVK; GAMMERMAN; SHAFER, 2005; SHAFER; VOVK, 2008; PAPADOPOULOS; GAMMERMAN; VOVK, 2008). Esses métodos oferecem uma abordagem para construir intervalos de predição com garantias em amostras finitas, assegurando validade marginal e validade condicional assintótica. Eles são livres de distribuição, agnósticos ao modelo e oferecem garantias de cobertura marginal em amostras finitas, sem depender de suposições fortes sobre os dados, apenas que sejam permutáveis. Essa suposição é frequentemente razoável e permite integrar a predição conforme a uma ampla variedade de modelos.

Uma componente principal da predição conforme é o escore de não conformidade, que mede o quão distante um ponto está do comportamento esperado segundo o modelo treinado. Diversas extensões dos métodos conforme vêm sendo desenvolvidas para modelos gerais de regressão (TIBSHIRANI *et al.*, 2019), bem como diferentes estratégias para construção dos escores de não conformidade em contextos variados (KATO; TAX; LOOG, 2023).

A heterocedasticidade, isto é, quando a variância da variável resposta depende das covariáveis, permanece um desafio na modelagem. Nos modelos de regressão Dirichlet, média e variância estão intrinsecamente ligadas, o que permite acomodar a variância de forma natural. Nesse contexto, diversas adaptações da predição conforme têm sido desenvolvidas para lidar com essa característica, incluindo a regressão quantílica conformalizada (ROMANO; PATTERSON; CANDES, 2019), a predição conforme normalizada (PAPADOPOULOS; GAMMERMAN; VOVK, 2008; LEI *et al.*, 2018; KATO; TAX; LOOG, 2023) e a predição conforme Mondrian (DEWOLF; BAETS; WAEGEMAN, 2025).

Uma abordagem baseada em modelo para a predição conforme em regressão distribucional foi apresentada por Chernozhukov, Wüthrich e Zhu (2021), utilizando a transformada integral de probabilidade para construir escores de não conformidade com base na distribuição condicional estimada da resposta. Essa abordagem garante, assintoticamente, a validade dos intervalos mesmo sob heterocedasticidade. Em paralelo, Barber, Candes e Ramdas (2023) propuseram um método de predição conforme por quantis ponderados, que oferece garantias em amostras finitas mesmo quando a permutabilidade não é válida.

À luz dessas considerações, neste trabalho objetivamos propor métodos de predição in-

tervalar para dados composicionais. Como o termo intervalos preditivos não faz muito sentido no contexto multivariado devido ao fato de termos a dependência de soma um entre as componentes, iremos nos referir a partir de agora como conjuntos ou regiões de predição. Com isso, propomos três novas abordagens para a construção de conjuntos preditivos, a primeira é baseada no resíduo quantílico (DUNN; SMYTH, 1996) das componentes marginais, a segunda na aproximação da região de maior densidade que contém o conjunto exato de predição e a terceira na utilização de um *grid* na região aproximada para busca numérica, com o objetivo de reduzir a sobrecobertura e o alargamento excessivo dos conjuntos de predição.

Com base nesse panorama, este trabalho propõe métodos de predição conforme para dados composicionais e está organizado da seguinte forma, o Capítulo 2 apresenta as parametrizações mais comuns da distribuição Dirichlet e a especificação do modelo de regressão associado, bem como outras abordagens alternativas para a modelagem de dados composicionais. No Capítulo 3, propõem-se os métodos de construção dos conjuntos de predição conforme para dados composicionais via regressão Dirichlet. O Capítulo 4 detalha os estudos de simulação, apresentando os resultados para os três métodos propostos, avaliados segundo cobertura empírica, área relativa média dos conjuntos construídos com relação ao simplex total e tempo de execução. No Capítulo 5, aplicamos os métodos a dois conjuntos de dados reais, sendo um deles aumentado artificialmente e comparamos o desempenho preditivo dos métodos. Por fim, o Capítulo 6 apresenta as considerações finais e sugestões para trabalhos futuros.

CAPÍTULO

2

MODELOS DE REGRESSÃO PARA DADOS COMPOSICIONAIS

Neste capítulo abordaremos os principais modelos relacionados à regressão para dados composicionais. Daremos ênfase na regressão Dirichlet, pois será o modelo predominante neste trabalho.

Inicialmente, na Seção 2.1 abordamos os fundamentos da distribuição Dirichlet, a qual possui duas parametrizações, conhecidas como comum e alternativa (MORAIS; THOMAS-AGNAN; SIMIONI, 2018). A alternativa, elencada na Seção 2.2, é a que nos permite modelar a média da resposta, uma estrutura bastante utilizada em análise de regressão, tendo em vista a interpretabilidade dos parâmetros.

Posteriormente, na Seção 2.3 introduzimos a estrutura de regressão a ser utilizada para a construção dos intervalos de predição conforme nos capítulos seguintes. Por último, elencamos abordagens alternativas à regressão Dirichlet na Seção 2.4, baseadas em transformações composicionais. Isso porque na presença de um conjunto de variáveis estritamente positivas, é comum normalizar os valores de cada observação para que passem a representar proporções de um todo (AITCHISON, 1982).

2.1 Distribuição Dirichlet

O vetor aleatório $(Y_1,\ldots,Y_D)$, com $Y_1+\cdots+Y_D=1$ segue uma distribuição Dirichlet de parâmetros $\lambda_1>0,\ \lambda_2>0,\ldots,\ \lambda_D>0$, com $Y_i>0$, se a função de densidade de $(Y_1,\ldots,Y_D)$ puder ser expressa na forma

$$f(\mathbf{y};\boldsymbol{\lambda})=\frac{\Gamma(\sum_{j=1}^{D}\lambda_j)}{\prod_{j=1}^{D}\Gamma(\lambda_j)}\prod_{j=1}^{D}y_j^{\lambda_j-1}, \tag{2.1}$$

em que $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_D)$, $\mathbf{y} = (y_1, \ldots, y_D)^\top \in \Delta^D$ e $\Gamma(\cdot)$ é a função Gama, definida por

$$\Gamma(t) = \int_0^\infty u^{t-1} e^{-u} du. \tag{2.2}$$

Um ponto importante é que, no caso particular $D = 2$, ela se reduz à distribuição beta, com parâmetros λ_1 e λ_2 (LIN, 2016). Se todos os parâmetros forem iguais ($\lambda_1 = \cdots = \lambda_D$), obtém-se a chamada Dirichlet simétrica, cuja densidade não privilegia nenhum componente e se distribui de forma simétrica sobre o simplex.

Essa parametrização em termos de parâmetros de forma é bastante utilizada para a distribuição Dirichlet. A partir dela, pode-se expressar esperança e variância, respectivamente, por

$$\mathrm{E}[Y_j] = \frac{\lambda_j}{\sum_{t=1}^D \lambda_t} \quad \text{e} \quad \mathrm{Var}(Y_j) = \frac{\lambda_j(\sum_{t\neq j}^D \lambda_t)}{(\sum_{t=1}^D \lambda_t)^2(\sum_{t=1}^D \lambda_t + 1)}.$$

Na Figura 2.1.1 é possível observar algumas formas que a densidade da Dirichlet assume para três componentes da resposta e diferentes valores para os parâmetros $(\lambda_1, \lambda_2, \lambda_3)$. É notório que a medida que os valores de $\boldsymbol{\lambda}$ aumentam, a variância diminui, indo ao encontro do que é observado na expressão da variância.

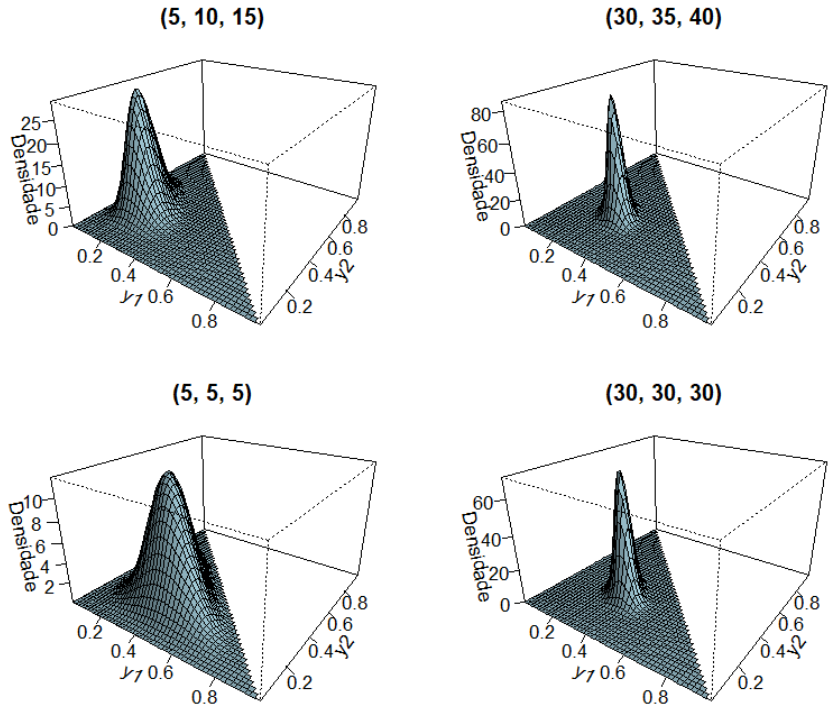

Figura 2.1.1 – Densidades da distribuição Dirichlet para diferentes valores dos parâmetros de forma λ e três componentes de resposta.

Uma discussão aprofundada sobre a distribuição Dirichlet, bem como a demonstração de suas propriedades teóricas, pode ser encontrada em Silva (2004).

2.2 Distribuição Dirichlet Alternativa

A seção anterior apresentou a fundamentação teórica da distribuição de Dirichlet em sua parametrização canônica. Entretanto, para problemas de regressão, é comum que sejam utilizadas estruturas que modelem a média de uma variável resposta juntamente com um parâmetro de precisão (FERRARI; CRIBARI-NETO, 2004), permitindo uma interpretabilidade mais natural dos parâmetros estimados.

Assim, seja $\phi = \lambda_1 + \cdots + \lambda_D$ e $\mu_j = \lambda_j/\phi$, com $j = 2, \ldots, D$ e $\mu_1 = 1 - \sum_{j=2}^{D} \mu_j$. A função de densidade com parametrização alternativa é dada por

$$f(\boldsymbol{y}; \mu_1, \ldots, \mu_D, \phi) = \frac{\Gamma(\phi)}{\prod_{j=1}^{D} \Gamma(\mu_j \phi)} \prod_{j=1}^{D} y_j^{\phi\mu_j - 1}, \tag{2.3}$$

em que $0 < \mu_j < 1$, $\forall j \in (1, \ldots, D)$ e $\phi > 0$.

Nesta reparametrização, a esperança e a variância são dadas, respectivamente, por

$$\mathrm{E}[Y_j] = \frac{\lambda_j}{\phi} = \frac{\mu_j \phi}{\phi} = \mu_j$$

e

$$\mathrm{Var}[Y_j] = \frac{\lambda_j(\phi - \lambda_j)}{\phi^2(\phi+1)} = \frac{\phi\mu_j(\phi - \phi\mu_j)}{\phi^2(\phi+1)} = \frac{\mu_j(1-\mu_j)}{\phi+1},$$

sendo ϕ então interpretado como um parâmetro de precisão, no sentido de que a medida que ϕ aumenta, para um dado μ_j fixo, a variância da resposta diminui.

Um aspecto estrutural importante da distribuição Dirichlet, e que se mantém sob sua versão regressiva, é a forma imposta para a dependência entre componentes da composição. Em particular, como as partes devem somar exatamente 1, a variação de um componente necessariamente compete com a variação dos demais, o que se reflete em covariâncias invariavelmente negativas entre componentes distintos. Na parametrização alternativa (μ, ϕ), tem-se, para $j \neq k$ (SILVA, 2004),

$$\begin{aligned} \mathrm{Cov}(Y_j, Y_k) &= -\frac{\alpha_j \alpha_k}{\left(\sum_{t=1}^{p} \alpha_t\right)^2 \left(\sum_{t=1}^{p} \alpha_t + 1\right)} \\ &= -\frac{\phi\mu_j \cdot \phi\mu_k}{\phi^2(\phi+1)} \\ &= -\frac{\mu_j \mu_k}{\phi+1} \end{aligned} \tag{2.4}$$

Pode-se notar que toda a estrutura de covariância é determinada apenas pelo vetor de médias μ e por um único parâmetro de precisão ϕ, restringindo o padrão de correlações admissíveis e implicando que aumentos em uma parte estejam sempre associados a reduções esperadas em outras. Essa característica, embora coerente com a restrição composicional, pode ser limitante em aplicações nas quais a dependência empírica entre componentes seja mais complexa.

2.3 Especificação do Modelo

Seja $\mathbf{Y}_1,\ldots,\mathbf{Y}_n$, $\mathbf{Y}_i = (Y_{i1},\ldots,\ Y_{iD})^\top$ variáveis aleatórias distribuídas segundo uma distribuição Dirichlet como definido em (2.3). Temos que os componentes sistemáticos do modelo de regressão Dirichlet logístico (PEREIRA; CAI, 2024) são expressos como

$$\begin{aligned} \log\left(\frac{\mu_{ij}}{\mu_{i1}}\right) &= x_{ij1}\beta_{j1} + x_{ij2}\beta_{j2} + \cdots + x_{ijp_j}\beta_{jp_j}, \quad 2 \leq j \leq D, \\ \log(\phi_i) &= d_{i1}\gamma_1 + d_{i2}\gamma_2 + \cdots + d_{ip_\phi}\gamma_{p_\phi}, \end{aligned} \tag{2.5}$$

em que $(x_{ij1}, x_{ij2}, \ldots, x_{ijp_j}, d_{i1}, d_{i2}, \ldots, d_{ip_\phi})^\top$ são covariáveis e $(\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp_j}, \gamma_1, \gamma_2, \ldots, \gamma_{p_\phi})^\top$ parâmetros desconhecidos. Note que o modelo (2.5) utiliza a função de ligação logito, assim como é feito na regressão logística multinomial (JR; LEMESHOW; STURDIVANT, 2013). Além disso, temos intercepto para $x_{ij1} = 1$

De forma análoga aos modelos de regressão com resposta multinomial, modelos com resposta Dirichlet também permitem o uso de diferentes funções de ligação (DAS; MUKHOPADHYAY, 2014; LI *et al.*, 2022). No entanto, a escolha da função de ligação logito para o vetor de médias e da logarítmica para o parâmetro de dispersão é particularmente vantajosa, devido à interpretabilidade direta dos parâmetros resultantes. A estimação dos parâmetros do modelo (2.5) é realizada por máxima verossimilhança, e seu ajuste pode ser implementado computacionalmente através do pacote `DirichletReg` (MAIER, 2014) do software R.

Outros resultados e mais detalhes acerca do modelo de regressão Dirichlet podem ser consultados em Melo *et al.* (2022). Nele, os autores observaram que, embora o estimador de máxima verossimilhança apresente normalidade assintótica em grandes amostras, o teste da Razão de Verossimilhanças (LR) tende a ser liberal e superestimar o tamanho do teste quando o tamanho amostral é reduzido. Para mitigar esse problema no modelo de regressão Dirichlet, os autores derivaram estatísticas de teste modificadas com distribuições nulas mais próximas da distribuição qui-quadrado. Outros resultados relevantes do estudo demonstram que o emprego de correções analíticas de Bartlett e métodos de reamostragem *bootstrap* garantem inferências consideravelmente mais acuradas, superando as limitações do teste LR convencional em diferentes cenários do modelo de regressão Dirichlet.

2.4 Modelos baseados em transformações

Quando a variável resposta é composicional uma alternativa natural à regressão Dirichlet consiste em transformar $\mathbf{Y}_i$ para um espaço euclidiano de dimensão $D-1$ por meio de transformações *log-ratio*. Aplica-se uma transformação $\mathbf{Z}_i = g(\mathbf{Y}_i) \in \mathbb{R}^{D-1}$. A partir disso, pode-se utilizar um modelo de regressão linear (AITCHISON, 2005), por exemplo, para $\mathbf{Z}_i$ condicionado às covariáveis, e em seguida reconverte-se ao simplex via g^{-1}. Essa estratégia permite empregar ferramentas estatísticas usuais no espaço transformado, ao mesmo tempo em

que garante que as predições finais pertençam ao simplex. Os modelos que utilizamos como alternativa a Dirichlet são advindos de transformações, as quais podem ser consultadas com mais detalhes em Greenacre *et al.* (2023).

2.4.1 Transformação Logaritmo da Razão Aditiva (ALR)

A transformação *log-ratio* aditiva (ALR) depende da seleção de uma casela de referência. Como há D opções possíveis para essa escolha, é essencial definir essa referência de maneira que atenda a um objetivo específico, seja ele estatístico ou relacionado ao conteúdo analisado. Fixando, por conveniência, a D-ésima componente como denominador, define-se para $\mathbf{Y}_i \in \Delta^D$:

$$\text{ALR}(\mathbf{Y}_i) \;=\; \left(\log\frac{Y_{i1}}{Y_{iD}},\, \log\frac{Y_{i2}}{Y_{iD}},\, \ldots,\, \log\frac{Y_{i(D-1)}}{Y_{iD}}\right)^{\top} \in \mathbb{R}^{D-1}. \tag{2.6}$$

A inversa que corresponde ao retorno ao simplex é obtida por

$$\text{ALR}^{-1}(\mathbf{Z}_i) \;=\; \frac{\big(\exp(Z_{i1}),\ldots,\exp(Z_{i(D-1)}),1\big)}{\sum_{k=1}^{D-1}\exp(Z_{ik})+1}. \tag{2.7}$$

A ALR é simples e amplamente utilizada, mas a interpretação dos parâmetros depende da casela de referência, e as log-razões requerem $Y_{ij} > 0$.

2.4.2 Transformação Logaritmo da Razão Centralizada (CLR)

A transformação logaritmo da razão centralizada utiliza como denominador a média geométrica das partes, o que a torna invariante a permutações das componentes. Seja $g(Y_i) = \left(\prod_{j=1}^{D} Y_{ij}\right)^{1/D}$. Define-se:

$$\text{CLR}(\mathbf{Y}_i) \;=\; \left(\log\frac{Y_{i1}}{g(Y_i)},\, \ldots,\, \log\frac{Y_{iD}}{g(Y_i)}\right)^{\top}.$$

A inversa é

$$\text{CLR}^{-1}(\mathbf{U}_i) \;=\; \frac{(\exp(U_{i1}),\ldots,\exp(U_{iD}))}{\sum_{k=1}^{D-1}\exp(U_{ik})}.$$

Note que $\sum_{j=1}^{D}\text{CLR}(Y_i)_j = 0$, de modo que as coordenadas CLR pertencem a um subespaço de dimensão $D-1$.

2.4.3 Transformação Isométrica do Logaritmo da Razão (ILR)

A transformação isométrica do logaritmo da razão (ILR) fornece uma representação em $\mathbb{R}^{D-1}$ baseada em uma base ortonormal do simplex, no sentido da geometria de Aitchison (PAWLOWSKY-GLAHN; EGOZCUE; TOLOSANA-DELGADO, 2007). Denote por $\Psi \in \mathbb{R}^{D\times(D-1)}$ uma matriz cujas colunas formam uma base ortonormal. Define-se:

$$\text{ILR}(\mathbf{Y}_i) \;=\; \text{CLR}((\mathbf{Y}_i)\Psi \;\in\; \mathbb{R}^{D-1}.$$

A inversa é dada por:

$$\text{ILR}^{-1}(\mathbf{Z}_i) \;=\; \frac{(\exp(\Psi\mathbf{Z}_i))}{\mathbf{1}^\top \exp(\Psi\mathbf{Z}_i)},$$

em que $\mathbf{1}$ é o vetor de uns no $\mathbb{R}^D$. A principal vantagem da ILR é que ela produz coordenadas em $\mathbb{R}^{D-1}$ com geometria euclidiana padrão, facilitando a estimação e a inferência em modelos de regressão.

2.4.4 Regressão com resposta composicional no espaço transformado

Seja $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{iD})^\top \in \Delta^D$, para $i = 1,\ldots,n$, uma variável resposta composicional. Considere uma transformação log-razão $g : \Delta^D \to \mathbb{R}^{D-1}$ (por exemplo, alr ou ilr) e defina o vetor transformado

$$\mathbf{Z}_i \;=\; g(\mathbf{Y}_i) \;=\; (Z_{i1},\ldots,Z_{i(D-1)})^\top \in \mathbb{R}^{D-1}.$$

A especificação de regressão no espaço transformado pode ser escrita de forma análoga a um modelo de regressão multivariada. Para $j = 1,\ldots,D-1$, considere o sistema de equações

$$z_{ij} \;=\; w_{i1}\beta_{j1} + w_{i2}\beta_{j2} + \cdots + w_{ip}\beta_{jp} + \varepsilon_{ij}, \qquad j = 1,\ldots,D-1, \tag{2.8}$$

em que $\mathbf{w}_i = (w_{i1},\ldots,w_{ip})^\top$ é o vetor de covariáveis, $\boldsymbol{\beta}_j = (\beta_{j1},\ldots,\beta_{jp})^\top$ é o vetor de parâmetros associado à j-ésima coordenada transformada e $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1},\ldots,\varepsilon_{i(D-1)})^\top$ é um vetor de erros com dependência entre coordenadas. Em notação matricial, (2.8) equivale a

$$\mathbf{z}_i \;=\; \mathbf{w}_i^\top \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{D-1}(0,\Sigma), \tag{2.9}$$

em que $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_{D-1}) \in \mathbb{R}^{p\times(D-1)}$ é a matriz de coeficientes e $\Sigma \in \mathbb{R}^{(D-1)\times(D-1)}$ captura a dependência entre os erros das coordenadas transformadas. Uma vez ajustado o modelo no espaço euclidiano, a predição para a composição é obtida por

$$\widehat{\mathbf{Y}}_i \;=\; g^{-1}(\widehat{\mathbf{Z}}_i),$$

garantindo $\widehat{\mathbf{Y}}_i \in \Delta^D$. Sob a suposição de normalidade em (2.9), a resposta $\mathbf{Y}_i$ induzida no simplex pertence à família normal logística, oferecendo maior flexibilidade para a estrutura de covariância do que a regressão Dirichlet, ao permitir uma matriz Σ não restrita.

CAPÍTULO

3

CONJUNTOS DE PREDIÇÃO CONFORME EM REGRESSÃO DIRICHLET

Uma prática comum em análise de regressão é o uso de um modelo ajustado para prever novas observações. Uma previsão pontual pode ser facilmente obtida usando o modelo ajustado, mas frequentemente também é desejável calcular uma região de predição com um determinado nível de cobertura. Neste capítulo, serão apresentados alguns métodos utilizados para a obtenção dos conjuntos preditivos via métodos de predição conforme.

Na Seção 3.1, são discutidos os aspectos iniciais da teoria de predição conforme (VOVK; GAMMERMAN; SHAFER, 2005; LEI; ROBINS; WASSERMAN, 2013; LEI; RINALDO; WASSERMAN, 2015; LEI *et al.*, 2018), bem como uma breve discussão sobre quais métodos para a construção desses conjuntos preditivos seriam interessantes e aplicáveis ao contexto de dados composicionais com a regressão Dirichlet.

Ademais, na Seção 3.2, adentra-se na principal contribuição trabalho. Primeiro, apresenta-se a construção de conjuntos preditivos via resíduo quantílico (DUNN; SMYTH, 1996), os quais são produzidos marginalmente para cada componente da resposta. Posteriormente, introduz-se o conceito de regiões de maior densidade (HDR) (HYNDMAN, 1996), o qual é imprescindível para a construção dos conjuntos preditivos baseados na log-verossimilhança negativa.

3.1 Predição conforme

Recentemente, os métodos de predição conforme (CP) têm sido amplamente empregados para a construção de regiões de predição sob hipóteses fracas (VOVK; GAMMERMAN; SHAFER, 2005; SHAFER; VOVK, 2008). Dada uma amostra de pares permutáveis (X_i, Y_i), $i = 1, \ldots, n$, a predição conforme é utilizada para produzir um conjunto preditivo $\mathscr{C}(\cdot)$ para um novo vetor de covariáveis X_{n+1}, com resposta desconhecida Y_{n+1}, de modo que $\mathbb{P}\big(Y_{n+1} \in \mathscr{C}(X_{n+1})\big) \geq 1 - \alpha$. A literatura oferece diversos algoritmos para a construção desses

conjuntos preditivos, incluindo *Full Conformal Prediction* (FCP) (VOVK; GAMMERMAN; SHAFER, 2005), *Split Conformal Prediction* (SCP) (PAPADOPOULOS; GAMMERMAN; VOVK, 2008; VOVK, 2012; LEI *et al.*, 2018) e Jackknife+ (BARBER *et al.*, 2021).

Neste trabalho, estamos interessados em predições para dados composicionais em amostras grandes. Nesses casos, o método SCP é mais conveniente e muito mais utilizado por ter um custo computacional muito inferior aos demais. Desta forma, consideraremos neste trabalho o método SCP. Este método particiona os dados disponíveis em um conjunto de treino ($\mathscr{D}_{\text{train}}$) e um conjunto de calibração ($\mathscr{D}_{\text{cal}}$). Enquanto o modelo é ajustado apenas em $\mathscr{D}_{\text{train}}$, essa divisão implica um *trade-off* de poder estatístico, podendo levar a conjuntos preditivos mais largos, evidenciando maior incerteza sobre a predição para aquele ponto. Porém, quando a amostra é pequena o SCP é inviável justamente por exigir essa divisão da base de dados em treino e calibração.

O componente central do CP é o escore de não conformidade, denotado por $s(\mathbf{X}, Y)$, definido como uma função $s : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$, sendo $\mathscr{X} \subseteq \mathbb{R}^p$ e $\mathscr{Y} \subseteq \mathbb{R}$ (ou no caso composicional, $\mathscr{Y} \subseteq \Delta^D$), que quantifica quão atípica é uma nova observação $(\mathbf{X}_{n+1}, Y_{n+1})$. Para um nível de significância $\alpha \in (0,1)$, a região preditiva do SCP ao nível $(1-\alpha)$, $\mathscr{C}(\mathbf{X}_{n+1})$, é, em geral, definida sobre o domínio de resposta $\mathscr{Y}$ como:

$$\mathscr{C}(\mathbf{x}_{n+1}) = \{y_{n+1} \in \mathscr{Y} : s(\mathbf{x}_{n+1}, y_{n+1}) \leq q_{1-\alpha}\},$$

em que $q_{1-\alpha}$ é o quantil empírico de ordem $\lceil (1-\alpha)(n_{\text{cal}}+1) \rceil$ dos escores de não conformidade calculados no conjunto de calibração $\mathscr{D}_{\text{cal}}$ e n_{cal} é o tamanho do conjunto de calibração. Essa região fornece uma garantia de cobertura marginal de pelo menos $1-\alpha$ em amostras finitas sob a hipótese de permutabilidade.

Em abordagens de CP, a escolha de $s(\mathbf{x}, y)$ é crítica, pois influencia diretamente a forma e a informatividade das regiões preditivas resultantes (ANGELOPOULOS; BATES, 2021; IZBICKI, 2025). Em regressão para respostas contínuas univariadas e não limitadas, uma escolha comum para $s(\mathbf{x}, y)$ é o resíduo ordinário absoluto, $|y - \widehat{\mathbb{E}}[Y|\mathbf{x}]|$ (LEI *et al.*, 2018), em que $\widehat{\mathbb{E}}[Y|\mathbf{x}]$ é a função de regressão estimada e ajustada com os dados de treino.

Entretanto, a natureza composicional dos dados e o uso da distribuição Dirichlet requerem escores que levem em conta, de forma natural, as propriedades da distribuição e a geometria intrínseca. Por essa razão, focamos na construção de regiões preditivas sobre o simplex Δ^D. Em particular, optamos por utilizar escores de não conformidade baseados no resíduo quantílico (DUNN; SMYTH, 1996) e na log-verossimilhança negativa, os quais são detalhados nas seções seguintes. O procedimento completo para a construção das regiões preditivas usando esses escores específicos é apresentado nos Algoritmos 1 e 2, que aparecem no Apêndice A.

3.2 Conjuntos de predição conforme para dados composicionais

3.2.1 Conjuntos de predição com o resíduo quantílico

O primeiro método CP proposto neste trabalho utiliza o resíduo quantílico. O resíduo quantílico, proposto por Dunn e Smyth (1996), é uma medida de uso simples e aplicável a diversos modelos de regressão. Quando os parâmetros do modelo são estimados de maneira consistente, esse resíduo apresenta, assintoticamente, distribuição normal padrão. Além disso, estudos indicam que, mesmo com amostras pequenas, sua distribuição costuma ser bem aproximada pela normal padrão em diferentes modelos de regressão (PEREIRA, 2019; LEMONTE; MORENO-ARENAS, 2019; FENG; LI; SADEGHPOUR, 2020). Por esse motivo o resíduo quantílico vem sendo bastante utilizado para o diagnóstico de modelos de regressão.

No caso da regressão Dirichlet, nossa primeira proposta é usar *split conformal* com o resíduo quantílico, utilizando que as distribuições de cada componente marginal de um vetor Dirichlet segue uma distribuição beta. Aqui como estamos usando a parametrização alternativa da distribuição Dirichlet, sua marginal é expressa em termos da parametrização da distribuição beta proposta por Ferrari e Cribari-Neto (2004). Em particular, para cada componente $j = 1, \ldots, D$, tem-se que, condicionalmente às covariáveis $\mathbf{X} = \mathbf{x}$,

$$Y_j \mid \mathbf{X} = \mathbf{x} \sim \text{Beta}\big(\mu_j(\mathbf{x}), \phi(\mathbf{x})\big),$$

em que $\mu_j(\mathbf{x}) \in (0,1)$ representa a média condicional do componente j e $\phi(\mathbf{x}) > 0$ é o parâmetro de precisão do modelo de regressão Dirichlet.

Dado o ajuste do modelo, obtemos as estimativas $\widehat{\mu}_j(\mathbf{x})$ e $\widehat{\phi}(\mathbf{x})$, que caracterizam a distribuição condicional de cada componente marginal. A partir dessa parametrização, define-se o escore de não conformidade baseado no resíduo quantílico. Para a observação i e o componente j, o resíduo quantílico é dado por

$$r_{ij}^{q} = \Phi^{-1}\Big\{F(y_{ij}; \widehat{\mu}_{ij}, \widehat{\phi}_i)\Big\}, \tag{3.1}$$

em que $\widehat{\mu}_{ij} = \widehat{\mu}_j(\mathbf{x}_i)$ e $\widehat{\phi}_i = \widehat{\phi}(\mathbf{x}_i)$. Aqui, $\Phi(\cdot)$ denota a função de distribuição acumulada da normal padrão e $F(\cdot)$ a função de distribuição acumulada da distribuição beta correspondente ao componente j.

Entretanto, é necessário que haja uma forma de termos uma região de predição em que todas as componentes estejam simultaneamente dentro dos seus respectivos intervalos. Para isso, poderíamos utilizar diversos escores, por exemplo, elipses de Mahalanobis (GHORBANI, 2019), mas, nesse caso, não seria possível realizar a inversão dos termos e construir os intervalos sem utilizar uma busca via *grid*. À luz dessas considerações, propomos como escore de não conformidade

$$s(\mathbf{x}, \mathbf{y}) = \max_j |r_{ij}^{q}|, \tag{3.2}$$

em que r_{ij}^q é dado por (3.1). No Algoritmo 1 descrevemos como é feito o passo a passo para realizar a implementação do método.

Note que o conjunto preditivo é, por definição, $\mathscr{C}(\mathbf{x}_{n+1}) = \{\mathbf{y} \in \Delta^D : s(\mathbf{x},\mathbf{y}) \leq q_{1-\alpha}\}$. Logo, ao utilizar o máximo do módulo do resíduo quantílico da $j-$ésima componente aplicamos uma calibração global, pois $\max_j |r_j^q| \leq q_{1-\alpha}$ garante que todas as componentes com escores menores em módulo, satisfaçam $|r_j^q| \leq q_{1-\alpha}$. Analisando componente a componente, isso equivale a

$$-q_{1-\alpha} \leq r_j^q \leq q_{1-\alpha} \quad \Leftrightarrow \quad \Phi(-q_{1-\alpha}) \leq F_j \leq \Phi(q_{1-\alpha}),$$

porque a função de distribuição acumulada (CDF) da normal padrão $\Phi(\cdot)$ é uma função monótona,ou seja, ela preserva a desigualdade. Então, seja $p_{\text{inf}} = \Phi(-q_{1-\alpha})$ e $p_{\text{sup}} = \Phi(q_{1-\alpha})$, obtemos intervalos marginais fechados na escala original da resposta, dados por

$$I_j = [F_j^{-1}(p_{\text{inf}}),\ F_j^{-1}(p_{\text{sup}})].$$

Logo,

$$\mathscr{C}(\mathbf{x}_{n+1}) = \{\mathbf{y} \in \Delta^D : y_j \in I_j(\mathbf{X}_{n+1})\ \forall\ j\}.$$

Exemplo para D = 3. Para fins didáticos, apresentamos um exemplo ilustrativo da construção do conjunto preditivo conformal pelo método do resíduo quantílico. Considere um modelo de regressão Dirichlet ajustado conforme a equação (2.5). Suponha que o parâmetro de precisão seja constante e dado por

$$\widehat{\phi}(\mathbf{x}) \equiv \widehat{\phi} = 10,$$

enquanto o vetor de médias condicionais

$$\widehat{\mu}(\mathbf{x}) = \big(\widehat{\mu}_1(\mathbf{x}), \widehat{\mu}_2(\mathbf{x}), \widehat{\mu}_3(\mathbf{x})\big)$$

varia com as covariáveis.

Suponha que o conjunto de calibração contenha $n_{\text{cal}} = 6$ observações $(\mathbf{x}_i, \mathbf{y}_i)$, usando $\widehat{\phi} = 10$, com médias condicionais estimadas e respostas observadas dadas por

i	$\widehat{\mu}(\mathbf{x}_i)$	$\mathbf{y}_i$
1	(0,40, 0,30, 0,30)	(0,30, 0,20, 0,50)
2	(0,38, 0,31, 0,31)	(0,45, 0,25, 0,30)
3	(0,25, 0,50, 0,25)	(0,20, 0,55, 0,25)
4	(0,33, 0,42, 0,25)	(0,35, 0,45, 0,20)
5	(0,46, 0,23, 0,31)	(0,50, 0,15, 0,35)
6	(0,23, 0,31, 0,46)	(0,18, 0,22, 0,60)

Para cada observação i e componente j, calculamos

$$U_{ij} = F_j(y_{ij} \mid \mathbf{x}_i) = F_{\text{Beta}(10\widehat{\mu}_{ij},\, 10(1-\widehat{\mu}_{ij}))}(y_{ij}),$$

isto é, a função de distribuição acumulada da marginal beta avaliada no valor observado. Os valores obtidos são, aproximadamente,

i	U_{i1}	U_{i2}	U_{i3}
1	0,2703	0,2618	0,9102
2	0,6877	0,3711	0,5083
3	0,4012	0,6214	0,5491
4	0,5820	0,5875	0,4012
5	0,6041	0,3049	0,6363
6	0,4014	0,2897	0,8127

Em seguida, aplica-se a transformação quantílica normal

$$r_{ij}^{q} = \Phi^{-1}(U_{ij}),$$

em que Φ^{-1} denota a função quantil da normal padrão. Assim, os resíduos quantílicos são, aproximadamente,

i	r_{i1}^{q}	r_{i2}^{q}	r_{i3}^{q}
1	$-0,612$	$-0,638$	1,342
2	0,489	$-0,329$	0,021
3	$-0,250$	0,309	0,124
4	0,207	0,221	$-0,250$
5	0,264	$-0,510$	0,348
6	$-0,250$	$-0,554$	0,888

O escore de não conformidade da observação i é definido por

$$s_i = \max_{j=1,2,3} |r_{ij}^{q}|,$$

o que produz

i	s_i
1	1,342
2	0,489
3	0,309
4	0,250
5	0,510
6	0,888

Ordenando os escores em ordem crescente, obtemos

$$s_{(1)} = 0,250, \quad s_{(2)} = 0,309, \quad s_{(3)} = 0,489, \quad s_{(4)} = 0,510, \quad s_{(5)} = 0,888, \quad s_{(6)} = 1,342.$$

Fixando $\alpha = 0,30$ e tendo $n_{\text{cal}} = 6$, o índice conformal é dado por

$$k = \lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil = \lceil 7 \times 0,7 \rceil = 5.$$

Logo, o quantil conformal é

$$\widehat{q}_{1-\alpha} = \widehat{q}_{0,70} = s_{(5)} = 0{,}888.$$

Suponha agora que, para o ponto de teste $\mathbf{x}_{n+1}$, o modelo retorne

$$\widehat{\mu}(\mathbf{x}_{n+1}) = (0{,}5,\, 0{,}3,\, 0{,}2).$$

Como $\widehat{\phi} = 10$, as distribuições marginais condicionais são

$$Y_1 \mid \mathbf{x}_{n+1} \sim \text{Beta}(5,5), \qquad Y_2 \mid \mathbf{x}_{n+1} \sim \text{Beta}(3,7), \qquad Y_3 \mid \mathbf{x}_{n+1} \sim \text{Beta}(2,8).$$

A condição

$$|r_j^q| \leq \widehat{q}_{0,70}$$

é equivalente a

$$-\widehat{q}_{0,70} \leq \Phi^{-1}(F_j(y_j)) \leq \widehat{q}_{0,70},$$

ou, aplicando Φ em todos os termos,

$$\Phi(-0{,}888) \leq F_j(y_j) \leq \Phi(0{,}888).$$

Como

$$\Phi(-0{,}888) \approx 0{,}1873 \qquad \text{e} \qquad \Phi(0{,}888) \approx 0{,}8127,$$

os intervalos marginais de predição são dados por

$$I_1 = \left[F^{-1}_{\text{Beta}(5,5)}(0{,}1873),\, F^{-1}_{\text{Beta}(5,5)}(0{,}8127)\right] \approx [0{,}359,\, 0{,}641],$$

$$I_2 = \left[F^{-1}_{\text{Beta}(3,7)}(0{,}1873),\, F^{-1}_{\text{Beta}(3,7)}(0{,}8127)\right] \approx [0{,}171,\, 0{,}425],$$

$$I_3 = \left[F^{-1}_{\text{Beta}(2,8)}(0{,}1873),\, F^{-1}_{\text{Beta}(2,8)}(0{,}8127)\right] \approx [0{,}089,\, 0{,}305].$$

Portanto, o conjunto de predição conformal associado ao ponto $\mathbf{x}_{n+1}$ é

$$\mathscr{C}(\mathbf{x}_{n+1}) = \left\{\mathbf{y} \in \Delta^3 : y_1 \in [0{,}359, 0{,}641],\, y_2 \in [0{,}171, 0{,}425],\, y_3 \in [0{,}089, 0{,}305]\right\}.$$

Note-se que esse conjunto corresponde à interseção entre o simplex

$$\Delta^3 = \left\{\mathbf{y} \in \mathbb{R}^3 :\; y_j > 0,\; \sum_{j=1}^{3} y_j = 1\right\}$$

e os intervalos marginais obtidos por inversão das distribuições acumuladas. Assim, a predição respeita simultaneamente a estrutura composicional da resposta e a calibração conforme induzida pelos resíduos quantílicos.

3.2.2 Regiões de maior densidade (HDR)

Uma forma natural de resumir a incerteza associada a uma variável aleatória multivariada é por meio da construção de regiões de confiança que, para um dado nível de probabilidade, ocupam o menor volume possível no espaço amostral. Tais conjuntos são denominados regiões de maior densidade ou, em inglês, *highest density regions* (HDR).

Definição 1 (Região de maior densidade; Hyndman (1996))**.** *Seja f a função densidade de probabilidade de uma variável aleatória $\mathbf{Y} \in \mathbb{R}^d$ contínua e possivelmente multivariada; a HDR de $100(1-\alpha)\%$ é definida como o subconjunto $\mathscr{C}(f_\alpha)$ do espaço amostral de $\mathbf{y}$ tal que:*

$$\mathscr{C}(f_\alpha) = \{\mathbf{y} : f(\mathbf{y}) \geq f_\alpha\}, \tag{3.3}$$

em que f_α é a maior constante tal que $P(\mathbf{Y} \in \mathscr{C}(f_\alpha)) \geq 1-\alpha$, com $\alpha \in (0,1)$.

Hyndman (1996) destaca uma propriedade fundamental das HDRs, entre todas as regiões com cobertura de probabilidade de pelo menos $1-\alpha$, a região $\mathscr{C}$ definida em (3.3) é a que possui o menor volume, com respeito à medida de Lebesgue. Geometricamente, a fronteira de uma HDR é um corte de nível da densidade $f(\mathbf{y})$.

Essa propriedade torna as HDRs particularmente atrativas para distribuições multimodais ou fortemente assimétricas, como é frequentemente o caso da distribuição Dirichlet no simplex. Diferente de regiões baseadas em momentos centrais ou intervalos simétricos, a HDR se adapta à geometria da densidade, capturando a probabilidade onde ela está mais concentrada.

A Figura 3.2.1 apresenta superfícies da densidade Dirichlet no simplex para duas escolhas de μ e dois valores de ϕ. A região em vermelho indica a HDR para um nível fixo de probabilidade $1-\alpha$, em que $\alpha = 0,10$. Comparando $\phi = 20$ com $\phi = 100$, observa-se que o aumento de ϕ torna a distribuição muito mais concentrada, o pico fica mais alto e estreito e a HDR se contrai, ocupando uma área menor do simplex. Isso reflete menor dispersão das composições em torno do centro, como também fora visualizado na Figura 2.1.1.

Pode-se notar que $\boldsymbol{\mu}$ controla a localização da maior probabilidade e o posicionamento da HDR. No caso simétrico, a massa se concentra longe dos extremos do simplex, gerando uma HDR centrada, no caso assimétrico, o pico se desloca em direção ao componente dominante (maior μ_j), e a HDR se aproxima da região do simplex onde esse componente é mais elevado. Assim, μ define onde estão as composições típicas, enquanto ϕ define quão apertada é a concentração ao redor delas.

No gráfico de contorno com HDR presente na Figura 3.2.2, o núcleo (cores mais intensas) indica a região de maior densidade, enquanto o contorno delimita a HDR, o conjunto de pontos do simplex com densidade acima de um limiar escolhido para conter $1-\alpha$ da probabilidade, fora desse contorno a densidade é menor e a massa restante é α.

No contexto de predição conforme, a conexão com HDRs é estabelecida através da escolha do escore de não-conformidade. Se definirmos o escore como a log-verossimilhança

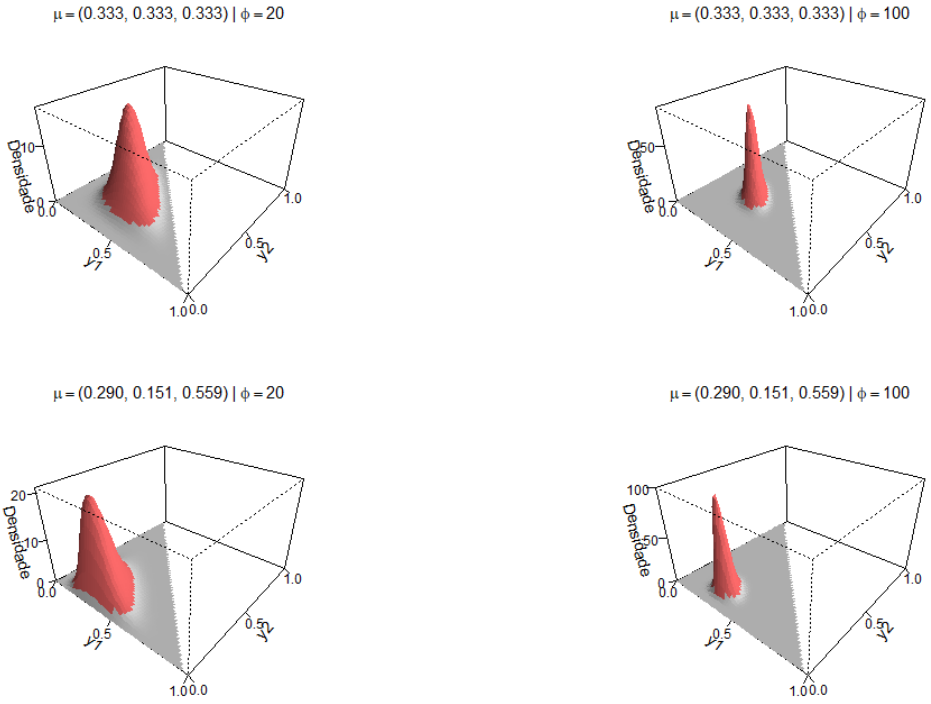

Figura 3.2.1 – Representação gráfica da HDR em 3D para três componentes e diferentes valores dos parâmetros da distribuição Dirichlet.

negativa, $s(\mathbf{x},\mathbf{y}) = -\log f(\mathbf{y}|\mathbf{x})$, então os conjuntos de predição gerados pelo método conforme consistirão nos pontos $\mathbf{y}$ onde a densidade $f(\mathbf{y}|\mathbf{x})$ excede um determinado limiar calibrado pelos dados, resultando, portanto, em estimadores válidos de regiões de maior densidade.

3.2.3 Predição conforme no simplex com aproximação da região de maior densidade

Como abordado na seção anterior, a região de maior densidade é o cerne do método e contém os pontos de maior probabilidade. Logo, é onde está o conjunto preditivo. Assim, propomos aproximar a região de maior densidade por meio de um politopo de pisos que a contém. Vale salientar que politopo é apenas a generalização do termo conhecido para polígonos e poliedros, mas para qualquer dimensão. Por exemplo, na segunda dimensão, um politopo é um poligono, já na terceira dimensão é um poliedro.

Chamamos de politopo de pisos o subconjunto do simplex obtido ao impor limites inferiores $y_i \geq \tau_i$. Isto é, $\mathscr{T}(\boldsymbol{\tau}) = \{\mathbf{y} \in \Delta^D : y_i \geq \tau_i,\ i = 1,\ldots,D\}$, que corresponde a um simplex truncado e será escolhido para conter a região exata $\mathscr{C}_{\text{exato}}(\mathbf{x})$.

Considere o modelo de regressão Dirichlet definido em (2.5). Suponha que, dado $\mathbf{x}$, o modelo fornece parâmetros de uma distribuição Dirichlet, em que a densidade é dada por (2.3). Por simplicidade, denotaremos $\mu_j(\mathbf{x}) = \mu_j$ e $\phi(\mathbf{x}) = \phi$, mas são dependentes de $\mathbf{x}$, visto que

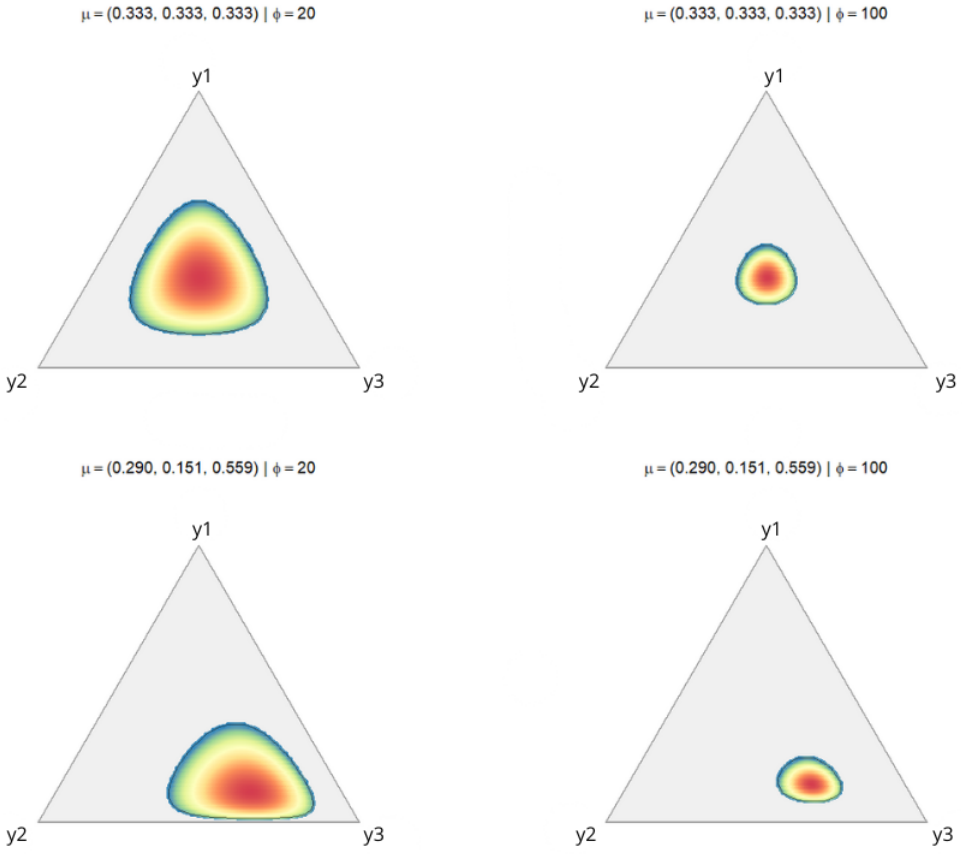

Figura 3.2.2 – Representação do gráfico de contorno para HDR com três componentes e diferentes valores dos parâmetros da distribuição Dirichlet.

serão utilizadas as estimativas do modelo. Assim, definimos o escore de não conformidade como,

$$s(\mathbf{x},\mathbf{y}) = -\log f(\mathbf{y} \mid \mathbf{x}) = -\log\Gamma(\widehat{\phi}) + \sum_{j=1}^{D} \log\Gamma(\widehat{\mu}_j\widehat{\phi}) - \sum_{j=1}^{D} (\widehat{\mu}_j\widehat{\phi} - 1)\log y_j,$$

como $s = -\log f$ é estritamente decrescente em f, o conjunto $\{y:\ s(\mathbf{x},\mathbf{y}) \leq c\}$ é um conjunto de nível da densidade (região de alta densidade) condicionada em $\mathbf{x}$.

No conjunto de calibração calculamos $s_i = s(\mathbf{x}_i, \mathbf{y}_i)$ e definimos o limiar conforme o quantil

$$\widehat{q}_{1-\alpha} \ = \ \text{ordem-}k \text{ de } \{s_i\}, \qquad k = \left\lceil (1-\alpha)\,(n_{\text{cal}}+1) \right\rceil.$$

Assim, após encontrar o quantil conforme, podemos inverter o escore de forma que é possível isolar $\sum_{j=1}^{D}(\widehat{\mu}_j\widehat{\phi} - 1)\log y_j$, que contém os componentes de resposta que estamos interessados para construir a aproximação. Para o ponto de teste $\mathbf{x}_{n+1}$, escreva $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}(\mathbf{x}_{n+1})$ e $\widehat{\phi} = \widehat{\phi}(\mathbf{x}_{n+1})$, logo

$$\widehat{t}_{1-\alpha} = -\widehat{q}_{1-\alpha} - \log\Gamma(\widehat{\phi}) + \sum_{j=1}^{D} \log\Gamma(\widehat{\mu}_j\widehat{\phi}), \qquad w_j = \widehat{\phi}(\mathbf{x}_{n+1})\widehat{\mu}_j(\mathbf{x}_{n+1}) - 1,$$

em que $\widehat{t}_{1-\alpha}$ depende de $\widehat{q}_{1-\alpha}$, de modo que o conjunto exato seria

$$\mathscr{C}_{\text{exato}}(\mathbf{X}_{n+1}) = \Big\{\mathbf{y} \in \Delta^D : \sum_{j=1}^{D} w_j \log y_j \geq \widehat{t}_{1-\alpha}\Big\}. \tag{3.4}$$

Em contraste com os escores baseados em resíduos quantílicos, o escore de não conformidade baseado na log-verossimilhança não admite uma inversão analítica em termos do vetor de respostas. Para contornar essa limitação, aproximamos a região HDR por meio de um politopo obtido a partir de limites inferiores definidos componente a componente. O processo para obter essa aproximação é dado na seção seguinte.

3.2.4 Aproximação da HDR: formulação via otimização

Fixe $\mathbf{x}_{n+1}$, $w_j = \phi\mu_j - 1$ e $W = \sum_{j=1}^{D} w_j$. O conjunto exato para o score $s(\mathbf{x},\mathbf{y}) = -\log f(\mathbf{y} \mid \mathbf{x})$ é

$$\mathscr{C}_{\text{exato}}(\mathbf{X}_{n+1}) = \Big\{\mathbf{y} \in \Delta^D : \sum_{j=1}^{D} w_j \log y_j \geq t_{1-\alpha}\Big\}, \qquad t_{1-\alpha} = -q_{1-\alpha} - \log\Gamma(\phi) + \sum_{j=1}^{D} \log\Gamma(\phi\mu_j).$$

No caso $w_j > 0$ para todo j, a função

$$g(\mathbf{y}) = \sum_{j=1}^{D} w_j \log y_j$$

é côncava no interior do simplex, pois $\log(\cdot)$ é côncava e a soma ponderada com pesos positivos preserva concavidade. Assim, a região exata é um conjunto convexo definido por

$$\mathscr{C}_{\text{exato}}(\mathbf{x}_{n+1}) = \{\mathbf{y} \in \Delta^D : \; g(\mathbf{y}) \geq t_{1-\alpha}\},$$

considerando $\mathbf{y}$ no interior do simplex ($y_j > 0$), pois o escore envolve $\log y_j$.

O limiar $g(\mathbf{y}) \geq t_{1-\alpha}$ tende a excluir pontos próximos às faces do simplex onde alguma componente é muito pequena, desde que $w_j > 0$ e o conjunto seja não vazio. De fato, como $\log y_j \to -\infty$ quando $y_j \downarrow 0$ e $w_j > 0$, valores muito pequenos de y_j reduzem fortemente o valor de $g(\mathbf{y}) = \sum_{j=1}^{D} w_j \log y_j$, violando a desigualdade $g(\mathbf{y}) \geq t_{1-\alpha}$.

Explorando essa intuição, construímos uma aproximação geométrica simples impondo limites inferiores $y_i \geq \tau_i$. Geometricamente, isso corresponde a deslocar (truncar) cada face do simplex $y_i = 0$ para a face paralela $y_i = \tau_i$, produzindo um simplex truncado que pode ser escolhido de modo a conter a região exata.

Para cada coordenada i (equivalentemente, para a face $y_i = 0$), o piso mínimo τ_i é dado pela solução do problema convexo

$$\min_{\mathbf{y}\in\mathbb{R}^D} \; y_i \quad \text{sujeito a} \quad \sum_{j=1}^{D} y_j = 1, \quad \sum_{j=1}^{D} w_j \log y_j \; \geq \; t_{1-\alpha}, \quad y_j \geq 0 \;\; (j = 1,\ldots,D). \tag{P_i}$$

Como P_i tem objetivo linear e conjunto factível convexo, trata-se de um problema convexo. Ademais, embora incluamos a restrição $y_j \geq 0$, a factibilidade de $\sum_{j=1}^{D} w_j \log y_j \geq t_{1-\alpha}$ com $w_j > 0$, implica $y_j > 0$ para todo ponto factível. Portanto, $\log y_j$ está bem definido no conjunto factível.

Definimos τ_i como o menor valor admissível da coordenada y_i dentro de $\mathscr{C}_{\text{exato}}(\mathbf{x}_{n+1})$. Assim, para todo $\mathbf{y} \in \mathscr{C}_{\text{exato}}(\mathbf{x}_{n+1})$ vale $y_i \geq \tau_i$, consequentemente,

$$\mathscr{C}_{\text{exato}}(\mathbf{x}_{n+1}) \subseteq \bigcap_{i=1}^{D} \{\mathbf{y} \in \Delta^D : \; y_i \geq \tau_i\} = \mathscr{T}(\tau).$$

Assumindo que existe ponto estritamente factível ($\mathbf{y} \in \Delta^D$ com $y_j > 0$ e $\sum_{j=1}^{D} w_j \log y_j > t_{1-\alpha}$), vale Slater e as condições de Karush-Kuhn-Tucker (KKT) (BERTSEKAS, 2009) são necessárias e suficientes. O Lagrangiano é

$$\mathscr{L}(\mathbf{y}, \rho, \theta, \delta) = y_i + \rho\Big(\sum_{j=1}^{D} y_j - 1\Big) + \theta\Big(t_{1-\alpha} - \sum_{j=1}^{D} w_j \log y_j\Big) - \sum_{j=1}^{D} \delta_j y_j,$$

com $\rho \in \mathbb{R}$ (restrição de soma), $\theta \geq 0$ (restrição de nível) e $\delta_j \geq 0$ ($y_j \geq 0$). As condições KKT incluem factibilidade primal e dual ($\theta \geq 0$, $\delta_j \geq 0$), complementaridade $\theta\,(t_{1-\alpha} - \sum_{j=1}^{D} w_j \log y_j) = 0$ e $\delta_j y_j = 0$, além da condição de estacionaridade, dada por

$$\frac{\partial \mathscr{L}}{\partial y_j} = \mathbf{1}_{\{j=i\}} + \rho - \theta\,\frac{w_j}{y_j} - \delta_j = 0, \quad j = 1, \ldots, D.$$

Sob a hipótese $w_j > 0$ para todo j e assumindo que o conjunto é não vazio, a restrição $g(\mathbf{y}) = \sum_{j=1}^{D} w_j \log y_j \geq t_{1-\alpha}$ implica $y_j > 0$ para todo ponto factível (pois $\log y_j \to -\infty$ quando $y_j \downarrow 0$). Logo, o ótimo de (P_i) é interior e, pela complementaridade $\delta_j y_j = 0$, segue que $\delta_j = 0$.

Assim,

$$\mathbf{1}_{\{j=i\}} + \rho = \theta\,\frac{w_j}{y_j} \qquad \Longrightarrow \qquad y_j = \begin{cases} \dfrac{\theta\, w_j}{\rho}, & j \neq i, \\[2ex] \dfrac{\theta\, w_i}{1+\rho}, & j = i, \end{cases}$$

com $\theta > 0$ e $\rho > 0$. De fato, como a restrição de nível é ativa no ótimo, segue $\theta > 0$ e como $y_j > 0$ e $w_j > 0$, das expressões $y_j = \theta w_j / \rho$ (para $j \neq i$) temos $\rho > 0$.

Impondo a restrição $\sum_{j=1}^{D} y_j = 1$, obtemos

$$\theta = \frac{1}{\frac{w_i}{1+\rho} + \frac{W - w_i}{\rho}}.$$

Além disso, no ótimo a restrição de nível é ativa, isto é, $\sum_{j=1}^{D} w_j \log y_j = t_{1-\alpha}$, caso contrário, seria possível reduzir ainda mais y_i mantendo factibilidade. Substituindo a expressão de y_j acima, chegamos à equação unidimensional em ρ:

$$F_i(\rho) = w_i \log \rho + (W - w_i)\log(1+\rho) - W \log\big(w_i \rho + (W - w_i)(1+\rho)\big) + \sum_{j=1}^{D} w_j \log w_j = t_{1-\alpha}.$$

Quando $w_j > 0$ para todo j, $F_i(\rho)$ é estritamente crescente em $(0,\infty)$, de modo que existe uma única raiz, a monotonicidade estrita pode ser verificada derivando $F_i(\rho)$, usando $w_j > 0$ e $W > 0$. Isso é exatamente a condição $w_j = \phi\mu_j - 1 > 0$ (isto é, $\phi\mu_j > 1$), que na parametrização original corresponde a $\lambda_j > 1$.

Dada a solução ρ, segue que

$$\theta = \frac{1}{\frac{w_i}{1+\rho} + \frac{W-w_i}{\rho}}, \qquad \tau_i = \frac{\theta w_i}{1+\rho},$$

e, portanto, o conjunto aproximado é

$$\mathscr{T}(\mathbf{x}_{n+1}) = \{\mathbf{y} \in \Delta^D : \; y_i \geq \tau_i, \; i = 1,\ldots,D\},$$

o qual contém $\mathscr{C}_{\text{exato}}(\mathbf{x}_{n+1})$ e, portanto, é conservadora, não reduz a cobertura marginal garantida pelo conformal, podendo induzir sobrecobertura. Na prática, ρ pode ser obtido por busca de raiz numérica, por exemplo, via `uniroot` da linguagem de programação R (R Core Team, 2022), que implementa o método de Brent (BRENT, 2013).

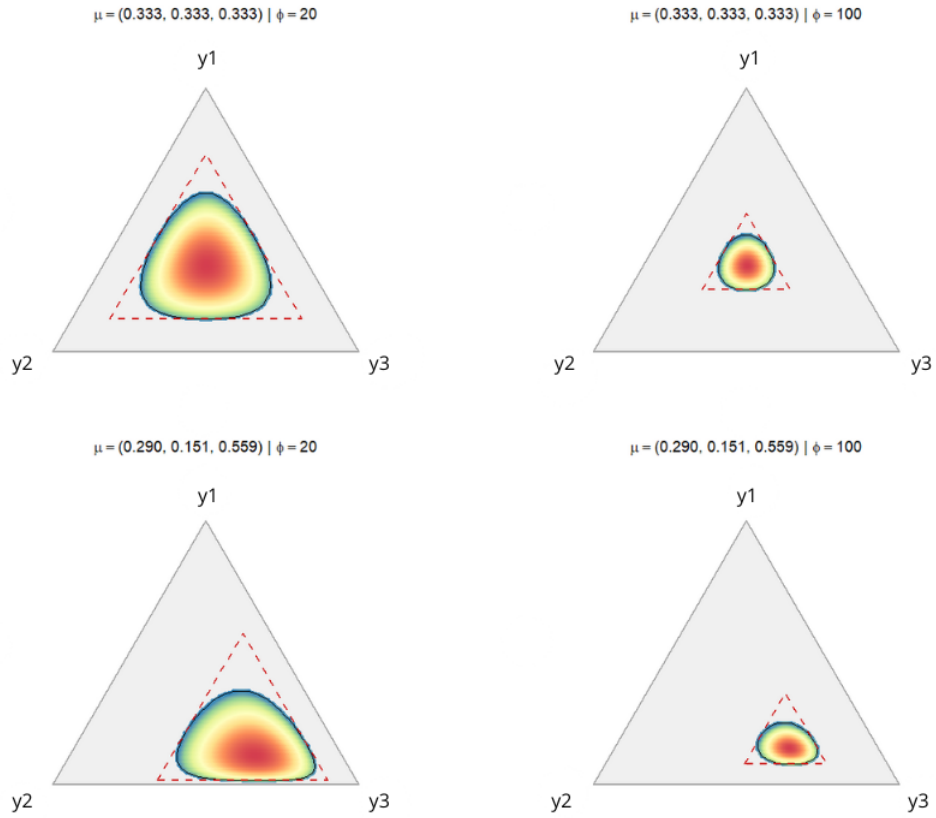

Figura 3.2.3 – Gráficos da HDR vs triângulo aproximado (vermelho) para diferentes valores dos parâmetros. A HDR é ilustrada pelo gradiente de cores dentro do triângulo vermelho pontilhado.

Entretanto, a aproximação da HDR para obter o conjunto de predição pode levar a uma notória sobrecobertura, a qual é visualmente observada na Figura 3.2.3, e os resultados empíricos podem ser avaliados no capítulo seguinte por meio das simulações.

Com base nisso, para mitigar a sobrecobertura e tornar a construção computacionalmente viável, restringimos a discretização ao politopo encontrado, evitando uma busca via *grid* no simplex inteiro para aproximar a HDR.

3.2.5 Aproximação refinada da HDR via grid

Uma solução inicial para aproximar a HDR e construir os conjuntos preditivos, seria a utilização de uma discretização no simplex, variando as componentes até encontrar pontos que são aceitos com base no critério conforme. Entretanto, tendo em vista que a Dirichlet é multivariada e possui restrição de soma, essa procura no simplex inteiro é inviável e exaustiva computacionalmente.

Apesar de a aproximação da HDR pelo politopo de pisos ser uma alternativa interessante, ela pode gerar sobrecobertura e conjuntos de predição maiores. Com isso, dado que o politopo de pisos constrói uma região paralela ao próprio simplex, o método permite que seja possível restringir a aproximação da HDR pelo *grid* dentro dessa região, que, por construção, é menor que o simplex inteiro, desde que $w_j > 0$.

Dado o vetor τ, definimos o *grid* $\mathscr{Y}^* \subset \mathscr{T}(\tau)$, em que $\mathscr{Y}^*$ são os valores de $\mathbf{y}$ discretizados dentro do politopo de pisos, fixando um passo $\delta > 0$ (nos estudos de simulação e nas aplicações utilizamos $\delta = 0{,}005$). Os pontos são gerados recursivamente para $k = 1, \ldots, D-1$ por

$$y_k \in \left\{ \tau_k + m\delta : \ m = 0, 1, 2, \ldots \right\} \cap \left[\tau_k, \ 1 - \sum_{j=k+1}^{D} \tau_j - \sum_{j=1}^{k-1} y_j \right], \tag{3.5}$$

m é o contador do número de pontos testados advindos da construção do grid. O último componente é fixado como

$$y_D = 1 - \sum_{j=1}^{D-1} y_j.$$

Por construção, mantemos apenas os pontos que satisfazem $y_D \geq \tau_D$, garantindo $\mathbf{y} \in \mathscr{T}(\tau) \subset \Delta^D$. O conjunto preditivo baseado em *grid* é então definido por

$$\mathscr{C}_{grid}(\mathbf{x}_{n+1}) = \{ \mathbf{y} \in \mathscr{Y}^* : \ g(\mathbf{y}) \geq t_{1-\alpha} \}.$$

3.2.6 Exemplo de obtenção do conjunto preditivo pelo método de aproximação da HDR para 3 componentes

Exemplo para D = 3. Para fins didáticos, ilustramos o método de construção do conjunto preditivo baseado na aproximação da região de maior densidade (HDR) por meio de um politopo de pisos. Considere um modelo de regressão Dirichlet ajustado conforme (2.5). Para um ponto genérico $\mathbf{x}$, o escore de não conformidade baseado na log-verossimilhança negativa é dado por

$$s(\mathbf{x}, \mathbf{y}) = -\log\Gamma(\widehat{\phi}) + \sum_{j=1}^{3} \log\Gamma(\widehat{\phi}\widehat{\mu}_j) - \sum_{j=1}^{3} (\widehat{\phi}\widehat{\mu}_j - 1) \log y_j.$$

Suponha agora que o conjunto de calibração tenha tamanho $n_{\text{cal}} = 6$ e que tomemos $\alpha = 0{,}30$. Para cada ponto $(\mathbf{x}_i, \mathbf{y}_i)$ do conjunto de calibração, calculamos

$$s_i = s(\mathbf{x}_i, \mathbf{y}_i) = -\log f(\mathbf{y}_i \mid \mathbf{x}_i).$$

Os valores obtidos são apresentados a seguir:

i	$\widehat{\mu}(\mathbf{x}_i)$	$\mathbf{y}_i$	$s_i = -\log f(\mathbf{y}_i \mid \mathbf{x}_i)$
4	(0,333, 0,417, 0,250)	(0,35, 0,45, 0,20)	−2,876
6	(0,231, 0,308, 0,462)	(0,18, 0,22, 0,60)	−2,625
3	(0,231, 0,462, 0,308)	(0,20, 0,55, 0,25)	−2,513
5	(0,462, 0,231, 0,308)	(0,50, 0,15, 0,35)	−2,236
2	(0,385, 0,308, 0,308)	(0,45, 0,25, 0,30)	−1,931
1	(0,333, 0,250, 0,417)	(0,30, 0,20, 0,50)	−1,727

Ordenando os escores $s_{(1)} \leq \cdots \leq s_{(6)}$, o índice conformal é

$$k = \lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil = \lceil 7 \cdot 0{,}7 \rceil = 5,$$

de modo que o quantil conformal é

$$\widehat{q}_{1-\alpha} = \widehat{q}_{0{,}70} = s_{(5)} \approx -1{,}931.$$

Considere agora o ponto de teste $\mathbf{x}_{n+1}$, para o qual o modelo retorna

$$\widehat{\mu}(\mathbf{x}_{n+1}) = (0{,}5,\ 0{,}3,\ 0{,}2)^{\top}, \qquad \widehat{\phi}(\mathbf{x}_{n+1}) = \widehat{\phi} = 10.$$

Assim, os pesos definidos por

$$w_j = \widehat{\phi}\widehat{\mu}_j - 1, \qquad j = 1, 2, 3,$$

são dados por

$$\mathbf{w} = (4,\ 2,\ 1), \qquad W = \sum_{j=1}^{3} w_j = 7.$$

Além disso,

$$t_{1-\alpha} = -\widehat{q}_{1-\alpha} - \log\Gamma(\widehat{\phi}) + \sum_{j=1}^{3} \log\Gamma(\widehat{\phi}\widehat{\mu}_j),$$

e, substituindo os valores do exemplo, obtemos

$$t_{1-\alpha} = -(-1{,}931) - \log\Gamma(10) + \log\Gamma(5) + \log\Gamma(3) + \log\Gamma(2) \approx -6{,}9996.$$

Portanto, o conjunto exato de predição é

$$\mathscr{C}_{\text{exato}}(\mathbf{x}_{n+1}) = \left\{ \mathbf{y} \in \Delta^3 : \sum_{j=1}^{3} w_j \log y_j \geq t_{1-\alpha} \right\},$$

isto é,

$$\mathscr{C}_{\text{exato}}(\mathbf{x}_{n+1}) = \left\{\mathbf{y} \in \Delta^3 : 4\log y_1 + 2\log y_2 + \log y_3 \geq -6{,}9996\right\}.$$

Como esse conjunto não admite, em geral, descrição analítica simples em termos de intervalos para as componentes, construímos uma aproximação por meio de um politopo de pisos. Para cada coordenada $i \in \{1,2,3\}$, definimos

$$\tau_i = \min_{\mathbf{y}\in\Delta^3} y_i \quad \text{sujeito a} \quad \sum_{j=1}^{3} w_j \log y_j \geq t_{1-\alpha}. \tag{P$_i$}$$

Por construção,

$$\mathscr{C}_{\text{exato}}(\mathbf{x}_{n+1}) \subseteq \bigcap_{i=1}^{3} \left\{\mathbf{y} \in \Delta^3 : y_i \geq \tau_i\right\} = \mathscr{T}(\tau),$$

em que $\mathscr{T}(\tau)$ é o politopo de pisos.

Como $w_j > 0$ para todo j, toda solução factível satisfaz $y_j > 0$, e podemos aplicar as condições de Karush–Kuhn–Tucker. Para cada i, obtemos a equação unidimensional

$$F_i(\rho) = w_i \log\rho + (W - w_i)\log(1+\rho) - W\log\Big(w_i\rho + (W-w_i)(1+\rho)\Big) + \sum_{j=1}^{3} w_j \log w_j = t_{1-\alpha},$$

cuja solução numérica fornece o piso ótimo correspondente. Em seguida,

$$\tau_i = \frac{w_i\rho}{w_i\rho + (W-w_i)(1+\rho)}.$$

Resolvendo numericamente os três problemas, obtemos

$$\tau \approx (0{,}4233,\ 0{,}1663,\ 0{,}0604)^{\top},$$

de modo que

$$\sum_{i=1}^{3} \tau_i \approx 0{,}6499.$$

Logo, o politopo de pisos é dado por

$$\mathscr{T}(\tau) = \left\{\mathbf{y} \in \Delta^3 : y_1 \geq 0{,}4233,\ y_2 \geq 0{,}1663,\ y_3 \geq 0{,}0604\right\},$$

o qual contém $\mathscr{C}_{\text{exato}}(\mathbf{x}_{n+1})$.

Dentro de $\mathscr{T}(\tau)$, os intervalos projetados para cada componente são

$$y_1 \in [\tau_1,\ 1-\tau_2-\tau_3],$$
$$y_2 \in [\tau_2,\ 1-\tau_1-\tau_3],$$
$$y_3 \in [\tau_3,\ 1-\tau_1-\tau_2].$$

Substituindo os valores obtidos,

$$y_1 \in [0{,}4233,\ 0{,}7733],$$
$$y_2 \in [0{,}1663,\ 0{,}5163],$$
$$y_3 \in [0{,}0604,\ 0{,}4104].$$

Observe que a amplitude projetada é a mesma para todas as componentes, sendo dada por

$$1 - \sum_{i=1}^{3} \tau_i \approx 0{,}3501.$$

De fato,

$$(1-\tau_2-\tau_3)-\tau_1 = (1-\tau_1-\tau_3)-\tau_2 = (1-\tau_1-\tau_2)-\tau_3 = 1-\sum_{i=1}^{3}\tau_i.$$

Por fim, o politopo $\mathscr{T}(\tau)$ pode ser discretizado por meio de um *grid*, restringindo a busca aos pontos do simplex truncado. Para isso, fixamos um passo de discretização $\delta > 0$. O parâmetro δ controla a resolução do *grid*: quanto menor for δ, mais fino será o conjunto de pontos gerado e, consequentemente, melhor será a aproximação numérica do conjunto preditivo.

Em cada coordenada, os candidatos são gerados no formato

$$y_k = \tau_k + m\delta, \qquad m \in \{0,1,2,\ldots\},$$

em que m é apenas um índice inteiro que conta quantos passos de tamanho δ foram dados a partir do piso τ_k.

Passo 1: gerar y_1. Como $y_2 \geq \tau_2$ e $y_3 \geq \tau_3$, o maior valor possível para y_1 dentro de $\mathscr{T}(\tau)$ é

$$1 - \tau_2 - \tau_3.$$

Logo,

$$y_1 \in [\tau_1,\ 1-\tau_2-\tau_3] = [0{,}4233,\ 1-0{,}1663-0{,}0604] = [0{,}4233,\ 0{,}7733].$$

Fixando, por exemplo, $\delta = 0{,}01$, geramos os candidatos

$$y_1 = 0{,}4233 + m(0{,}01), \qquad m = 0,1,\ldots,\left\lfloor \frac{0{,}7733-0{,}4233}{0{,}01} \right\rfloor = 35.$$

Assim, os valores possíveis de y_1 no *grid* são

$$0{,}4233,\ 0{,}4333,\ 0{,}4433,\ \ldots,\ 0{,}7733.$$

Passo 2: dado y_1, gerar y_2 e fechar y_3. Para cada valor fixado de y_1, o intervalo factível de y_2 é determinado pelas restrições

$$y_2 \geq \tau_2 \qquad \text{e} \qquad y_3 = 1 - y_1 - y_2 \geq \tau_3.$$

Da segunda desigualdade, segue que

$$y_2 \leq 1 - \tau_3 - y_1.$$

Portanto,

$$y_2 \in [\tau_2,\ 1 - \tau_3 - y_1].$$

Por exemplo, se escolhermos

$$y_1 = 0{,}4533,$$

então

$$y_2 \in [0{,}1663,\ 1 - 0{,}0604 - 0{,}4533] = [0{,}1663,\ 0{,}4863].$$

Com $\delta = 0{,}01$, geramos

$$y_2 = 0{,}1663 + \ell(0{,}01), \qquad \ell = 0, 1, \ldots, \left\lfloor \frac{0{,}4863 - 0{,}1663}{0{,}01} \right\rfloor = 32,$$

em que ℓ desempenha o mesmo papel de m, isto é, conta quantos passos de tamanho δ foram dados a partir de τ_2.

Para cada par (y_1, y_2) gerado dessa forma, definimos

$$y_3 = 1 - y_1 - y_2.$$

Como o valor máximo de y_2 foi construído justamente para garantir

$$y_3 \geq \tau_3,$$

segue que, por construção, cada ponto (y_1, y_2, y_3) obtido pertence ao politopo $\mathscr{T}(\tau)$.

Passo 3: formar o conjunto discretizado. Repetindo o Passo 2 para cada valor de y_1 gerado no Passo 1, obtemos o conjunto de pontos do *grid*

$$\mathscr{Y}^* \subset \mathscr{T}(\tau).$$

Em seguida, dentre esses pontos discretizados, mantemos apenas aqueles que satisfazem a desigualdade da região exata,

$$4\log y_1 + 2\log y_2 + \log y_3 \geq -6{,}9996.$$

Assim, o conjunto preditivo aproximado via *grid* é dado por

$$\mathscr{C}_{grid}(\mathbf{x}_{n+1}) = \{\mathbf{y} \in \mathscr{Y}^* : 4\log y_1 + 2\log y_2 + \log y_3 \geq -6{,}9996\}.$$

Desse modo, a discretização do politopo de pisos permite aproximar numericamente o conjunto preditivo sem perder a restrição composicional imposta pelo simplex.

CAPÍTULO

4

ESTUDOS DE SIMULAÇÃO

Neste capítulo, buscamos avaliar empiricamente o desempenho dos métodos propostos por meio de estudos de simulação. A utilização de simulações controladas permite analisar, de forma sistemática, o comportamento do método sob diferentes condições de geração dos dados, variando aspectos como a distribuição das covariáveis, a magnitude da variância da variável resposta, e a forma funcional da dispersão.

Os experimentos foram elaborados com o objetivo de verificar a capacidade dos métodos em manter níveis adequados de cobertura preditiva e em fornecer conjuntos de predição com área relativa razoável. Os cenários incluem variações tanto na estrutura média quanto na estrutura de dispersão do modelo, e consideram diferentes distribuições para as covariáveis, bem como diferentes níveis de complexidade na geração da variável resposta.

A seguir, na Seção 4.1, apresentamos em detalhes os cenários considerados, os parâmetros utilizados em cada caso, e os resultados obtidos em termos de cobertura, área relativa dos conjuntos de predição e tempo computacional. Na Seção 4.2 estão contidos os resultados das simulações, bem como os comentários acerca do desempenho dos métodos, tanto para o cenário de correta especificação como o de má especificação.

4.1 Estudos de simulação

Para verificar a performance do método proposto, conduzimos um estudo de simulação de Monte Carlo, no qual consideramos 1000 iterações. Em cada cenário geramos amostras de tamanho $n = 1000$, em que dividimos 70% para treino, 30% para calibração e utilizamos um novo ponto para teste. Em cada cenário, a cobertura do conjunto de predição foi avaliada a partir de um nível de cobertura nominal $(1-\alpha)$ de 90%. Para a utilização da discretização com o *grid*, utilizamos pontos uniformemente espaçados para cada componente por um passo $\delta = 0,005$.

Além disso, consideramos 3 componentes de resposta para as simulações. Assim, para

calcular as áreas relativas ao simplex total, utilizamos aproximação via Monte Carlo, gerando 20.000 pontos aleatoriamente dentro do simplex inteiro utilizando uma distribuição Dirichlet com todos os parâmetros iguais a 1 e verificamos, por meio do quantil conforme $q_{1-\alpha}$, quais pontos são aceitos. Por fim, retornamos a razão entre pontos aceitos e pontos totais testados. Para o caso do politopo de pisos, a área do conjunto gerado pode ser obtida facilmente pela área do triângulo equilátero, como abordado na Seção 3.2.6.

Em cada um dos cenários avaliados, nós geramos duas covariáveis, $x_{i22} = x_{i32} = d_{i2}$, $x_{i23} = x_{i33} = d_{i3}$ e $x_{i21} = x_{i31} = d_{i1} = 1$ para todo i. Todas as simulações foram realizadas utilizando a linguagem de programação R e o pacote `DirichletReg`. Os 5 primeiros cenários foram considerados por Pereira e Cai (2024). No cenário 1a, as covariáveis foram geradas da distribuição uniforme padrão e o valor dos parâmetros foi escolhido de uma forma que a média de μ_{i1}, μ_{i2} e μ_{i3} entre as n observações sejam as mesmas e a variância da variável de resposta é alta (ϕ próximo de 20). Neste cenário, $\beta_{21} = -0.3$, $\beta_{22} = 1.0$, $\beta_{23} = -0.5$, $\beta_{31} = -0.3$, $\beta_{32} = -0.5$, $\beta_{33} = 1.0$, $\gamma_1 = 3.0$, e $\gamma_2 = \gamma_3 = 0$.

Nos cenários 2a e 3a, nós mudamos o valor de alguns parâmetros para fazer a média das componentes da variável de resposta ficarem longe e muito longe uma da outra, respectivamente. No cenário 4a, nós consideramos que ϕ é uma função das covariáveis usando $\gamma_2 = 0.5$ e $\gamma_3 = -0.5$. No cenário seguinte, a primeira e a segunda covariáveis foram geradas da distribuição de Bernoulli com parâmetro 0.5 e da distribuição Gama com parâmetros 3 e 6 (média e variância como na distribuição uniforme padrão), respectivamente. Cenários 1b a 5b são similares aos cenários 1a a 5a, respectivamente, com $\gamma_1 = 4.6$ (ϕ próximo de 100 que corresponde a uma baixa variância da variável resposta).

Além disso, para diversificar com relação ao número de covariáveis preditoras, adicionamos os cenários 6a e 6b, nos quais passamos a utilizar 6 covariáveis preditoras geradas de uma distribuição uniforme padrão. Os valores para os coeficientes de regressão foram $(\beta_{21}, \beta_{22}, \ldots, \beta_{27}) = (0,\ 0.20,\ -0.20,\ 0.15,\ -0.15,\ 0.10,\ -0.10)^\top$, $(\beta_{31}, \beta_{32}, \ldots, \beta_{37}) = (0,\ -0.20,\ 0.20,\ -0.15,\ 0.15,\ -0.10,\ 0.10)^\top$, isto é, $\beta_3 = -\beta_2$. Também adotamos precisão constante de forma análoga aos cenários anteriores, com $\gamma_1 = 3,0$ e $\gamma_1 = 4,6$.

Também avaliamos a robustez dos métodos quando o modelo ajustado é Dirichlet, mas os dados são gerados por mecanismos que não seguem exatamente uma Dirichlet. Em cada iteração, geramos $p \in \{2,6\}$ covariáveis independentes $U(0,1)$. Consideramos dois tipos de má especificação, cada um avaliado com $p = 2$ e $p = 6$, e em dois níveis de variabilidade, totalizando 10 cenários com variabilidade alta e 10 cenários com variabilidade baixa.

Para os cenários 6a a 7b geramos a resposta como uma mistura de duas distribuições Dirichlet, mecanismo gerador que denominaremos de MixDir, com probabilidade 0,6 para o componente 1 e 0,4 para o componente 2. Os dois componentes possuem a mesma precisão ϕ, mas médias diferentes, construídas por dois preditores lineares, um para a segunda componente e outro para a terceira. Os coeficientes do componente 2 são simplesmente o oposto dos coeficientes

do componente 1. Para $p = 2$, os coeficientes do componente 1 são:

$$\beta_{21}^{(1)} = 0, \quad \beta_{22}^{(1)} = 1{,}20, \quad \beta_{23}^{(1)} = -1{,}20,$$
$$\beta_{31}^{(1)} = 0, \quad \beta_{32}^{(1)} = -1{,}20, \quad \beta_{33}^{(1)} = 1{,}20.$$

No componente 2, temos:

$$\beta_{21}^{(2)} = 0, \quad \beta_{22}^{(2)} = -1{,}20, \quad \beta_{23}^{(2)} = 1{,}20,$$
$$\beta_{31}^{(2)} = 0, \quad \beta_{32}^{(2)} = 1{,}20, \quad \beta_{33}^{(2)} = -1{,}20.$$

Para $p = 6$, os coeficientes do componente 1 são:

$$\beta_{21}^{(1)} = 0, \quad \beta_{22}^{(1)} = 0{,}60, \quad \beta_{23}^{(1)} = -0{,}60,$$
$$\beta_{24}^{(1)} = 0{,}60, \quad \beta_{25}^{(1)} = -0{,}60, \quad \beta_{26}^{(1)} = 0{,}60, \quad \beta_{27}^{(1)} = -0{,}60,$$
$$\beta_{31}^{(1)} = 0, \quad \beta_{32}^{(1)} = -0{,}60, \quad \beta_{33}^{(1)} = 0{,}60,$$
$$\beta_{34}^{(1)} = -0{,}60, \quad \beta_{35}^{(1)} = 0{,}60, \quad \beta_{36}^{(1)} = -0{,}60, \quad \beta_{37}^{(1)} = 0{,}60.$$

No componente 2, novamente, todos os coeficientes acima são multiplicados por -1. A precisão foi mantida constante em dois níveis, com $\gamma_1 \in \{3{,}0,\ 4{,}6\}$ e os demais parâmetros γ iguais a zero.

Para a segunda forma de má especificação, a resposta composicional é gerada por meio de um mecanismo baseado na distribuição normal logística, que denotamos por LogN. Especificamente, duas variáveis contínuas com distribuição normal bivariada são simuladas, com correlação fixa $\rho = 0{,}30$, e os valores são então transformados para o espaço do simplex por meio de uma inversão da transformação ALR em três partes como visto em (2.7). O resultado é uma composição com três componentes positivas que somam 1. Além disso, a utilização de ρ positivo, gerando correlação positiva vai de encontro ao que é observado na expressão da covariância dada (2.4), induzindo uma má especificação na estrutura de correlação.

Consideramos dois valores para o número de covariáveis, $p \in \{2, 6\}$. Para $p = 2$, os coeficientes associados aos dois preditores são $\beta_{21} = 0,\ \beta_{22} = 0{,}90,\ \beta_{23} = -0{,}90, \beta_{31} = 0,\ \beta_{32} = -0{,}90,\ \beta_{33} = 0{,}90$. Para $p = 6$, utilizamos: $\beta_{21} = 0,\ \beta_{22} = 0{,}45,\ \beta_{23} = -0{,}45,\ \beta_{24} = 0{,}45,\ \beta_{25} = -0{,}45,\ \beta_{26} = 0{,}45,\ \beta_{27} = -0{,}45,\ \beta_{31} = 0,\ \beta_{32} = -0{,}45,\ \beta_{33} = 0{,}45,\ \beta_{34} = -0{,}45,\ \beta_{35} = 0{,}45,\ \beta_{36} = -0{,}45,\ \beta_{37} = 0{,}45$.

A dispersão do mecanismo LogN é controlada por um parâmetro σ. No código, σ é calibrado numericamente, separadamente para $p = 2$ e $p = 6$, para produzir dois níveis de variabilidade comparáveis aos de um modelo Dirichlet de referência com $\phi \approx 20$ (alta variabilidade) e $\phi \approx 100$ (baixa variabilidade). O resumo dos cenários de simulação podem ser visualizados na Tabela 4.1.1.

Ademais, também consideramos dois últimos experimentos para avaliar a eficiência computacional do método que utiliza a aproximação da região de maior densidade com a busca

Tabela 4.1.1 – Resumo dos cenários de simulação, corretamente especificados e mal especificados.

Cenário	**Mecanismo gerador**	**Média de** (μ_1, μ_2, μ_3)			**Covariáveis**	p	ϕ
1a, 1b	Dirichlet	0.333	0.333	0.333	Uniforme	2	Constante
2a, 2b	Dirichlet	0.290	0.151	0.559	Uniforme	2	Constante
3a, 3b	Dirichlet	0.308	0.049	0.643	Uniforme	2	Constante
4a, 4b	Dirichlet	0.333	0.333	0.333	Uniforme	2	Variável
5a, 5b	Dirichlet	0.333	0.333	0.333	Bernoulli/Gama	2	Constante
6a, 6b	Dirichlet	0.333	0.333	0.333	Uniforme	6	Constante
7a, 7b	MixDir	0.333	0.333	0.333	Uniforme	2	Constante
8a, 8b	MixDir	0.333	0.333	0.333	Uniforme	6	Constante
9a, 9b	LogN	0.333	0.333	0.333	Uniforme	2	(σ) Constante
10a, 10b	LogN	0.333	0.333	0.333	Uniforme	6	(σ) Constante

via *grid* em relação ao uso do *grid* no simplex inteiro. Utilizamos dois tamanhos de δ diferentes, para o método com a aproximação da HDR e posterior busca, utilizamos $\delta = 0,02$, já para o método ingênuo, $\delta = 0,005$.

No primeiro cenário desse experimento adotamos a mesma estrutura dos cenários 1a e 1b. Entretanto, a medida que o número de componentes da resposta aumenta, espera-se que o custo computacional das buscas via *grid* aumente também. Assim, consideramos um caso de quatro componentes de resposta. Para isso, $\beta_{21} = -0.3, \beta_{22} = 1.0, \beta_{23} = -0.5, \beta_{31} = -0.3, \beta_{32} = -0.5, \beta_{33} = 1.0, \beta_{41} = -0.3, \beta_{42} = 0.5, \beta_{43} = 0.5, \gamma_1 = 3.0, \gamma_2 = \gamma_3 = \gamma_4 = 0$.

Essa configuração de parâmetros gera componentes com médias iguais à $0,2335$, $0,1816$, $0,2850$ e $0,2999$. As duas covariáveis utilizadas no modelo foram geradas de uma uniforme padrão, semelhante ao que fora feito nos cenários anteriores. Como são cenários de correta especificação, os resultados serão reportados na subseção referente aos mesmos.

4.2 Resultados

Nesta seção, apresentamos e discutimos os resultados do estudo de simulação para os cenários de correta especificação e má especificação. Em todas as configurações, avaliamos a cobertura empírica do conjunto preditivo sob o nível nominal de 90%, bem como a área relativa média do conjunto construído em relação ao simplex total e o tempo computacional médio por repetição. Para facilitar a leitura, ao longo do texto os três procedimentos serão referidos, respectivamente, como resíduo quantílico, HDR-aprox e HDR-aprox-grid.

4.2.1 Resultados: correta especificação

A Tabela 4.2.1 apresenta os resultados das simulações para os cenários de correta especificação, nos quais o mecanismo gerador dos dados coincide com o modelo Dirichlet utilizado no ajuste. Em todos os cenários, o nível nominal de cobertura considerado foi de 90%.

De forma geral, os três procedimentos avaliados conseguem controlar adequadamente a cobertura empírica, embora apresentem diferenças sistemáticas quanto ao grau de conservadorismo, ao tamanho dos conjuntos preditivos e ao custo computacional.

O método baseado em resíduos quantílicos apresenta cobertura empírica próxima ao nível nominal na maioria dos cenários, com valores tipicamente entre 89% e 92%. Pequenas oscilações em torno de 90% são observadas, sobretudo em configurações com maior assimetria entre as componentes da resposta, mas sem indicar perda sistemática do controle de cobertura. Esse comportamento é compatível com a construção do método, que combina intervalos marginais para formar o conjunto preditivo conjunto.

O método HDR-aprox, por sua vez, exibe um padrão claramente mais conservador. Em todos os cenários de correta especificação, a cobertura empírica supera o nível nominal, variando aproximadamente entre 93,5% e 96,1%. Esse aumento de cobertura decorre da aproximação da região de maior densidade por meio de um politopo definido por pisos, o que tende a incluir regiões adicionais do simplex. Como consequência direta, os conjuntos preditivos gerados por esse método apresentam áreas relativas substancialmente maiores, sobretudo nos cenários em que as médias das componentes da resposta estão mais afastadas entre si.

Já o método HDR-aprox-grid apresentou coberturas entre $89,3\%$ e $91,4\%$, atingindo o nível nominal especificado de 90%. Mesmo em cenários de maior variabilidade e assimetria o método mostrou que consegue conservar a cobertura. Logo, é notório que o método corrige a sobrecobertura advinda do método HDR-aprox, o que é essencial para se obter regiões mais eficientes.

No que se refere à área relativa dos conjuntos de predição, observa-se um padrão claro entre os métodos. O HDR-aprox produz os maiores conjuntos, refletindo seu caráter conservador. O método baseado em resíduos quantílicos gera conjuntos consideravelmente menores, porém, de forma consistente, com áreas relativas ligeiramente maiores do que aquelas obtidas pelo HDR-aprox-grid. Por sua vez, o HDR-aprox-grid apresenta as menores áreas relativas entre os métodos considerados em todos os cenários analisados.

Além disso, verifica-se o efeito esperado da variabilidade da resposta: nos cenários com menor dispersão (cenários "b", associados a maiores valores de ϕ), as áreas relativas diminuem de forma consistente para todos os métodos, indicando maior concentração da massa de probabilidade no simplex.

Quanto ao custo computacional, os métodos de resíduos quantílicos e HDR-aprox apresentam tempos médios de execução baixos e comparáveis, em geral inferiores a dois centésimos de segundo por repetição, sendo que o segundo desses métodos é ligeiramente mais rápido. Assim, eles são plenamente adequados para aplicações com grande número de iterações. Em contraste, o método HDR-aprox-grid apresenta um custo computacional mais elevado, com tempos médios de execução cerca de 10 vezes ao observado no método de resíduo quantílico. Ainda

assim, os tempos observados permanecem moderados no contexto das simulações realizadas.

Em síntese, sob correta especificação do modelo, os resultados indicam um trade-off claro entre os métodos. O procedimento baseado em resíduos quantílicos oferece boa cobertura com conjuntos relativamente compactos e baixo custo computacional, embora ligeiramente maiores do que os obtidos pelo HDR-aprox-grid. O HDR-aprox garante cobertura elevada, porém à custa de conjuntos preditivos amplos. Já o HDR-aprox-grid surge como uma alternativa intermediária, capaz de combinar cobertura próxima ao nível nominal com conjuntos preditivos mais eficientes do ponto de vista geométrico, ainda que com maior custo computacional.

Tabela 4.2.1 – Resultados da simulação para os cenários de correta especificação. As métricas reportadas são a cobertura empírica, a área relativa média e o tempo de execução.

Cenário	Método	Cobertura Empírica (%)	Área Relativa	Tempo Médio (s)
1a	Res. Quantílico	90,1	0,2433	0,0139
	HDR-aprox	94,7	0,3221	0,0106
	HDR-aprox-grid	90,3	0,2339	0,1532
1b	Res. Quantílico	89,5	0,0537	0,0175
	HDR-aprox	94,5	0,0802	0,0101
	HDR-aprox-grid	90,1	0,0509	0,1753
2a	Res. Quantílico	92,1	0,1985	0,0206
	HDR-aprox	94,4	0,2709	0,0123
	HDR-aprox-grid	91,4	0,1796	0,1923
2b	Res. Quantílico	89,9	0,0437	0,0169
	HDR-aprox	94,6	0,0674	0,0106
	HDR-aprox-grid	89,3	0,0405	0,1686
3a	Res. Quantílico	88,8	0,0962	0,0144
	HDR-aprox	96,1	0,6910	0,0081
	HDR-aprox-grid	89,4	0,0539	0,1387
3b	Res. Quantílico	91,1	0,0179	0,0143
	HDR-aprox	95,4	0,2075	0,0084
	HDR-aprox-grid	90,7	0,0137	0,1381
4a	Res. Quantílico	90,1	0,1924	0,0145
	HDR-aprox	95,5	0,3195	0,0085
	HDR-aprox-grid	89,8	0,1718	0,1388
4b	Res. Quantílico	89,4	0,0429	0,0143
	HDR-aprox	93,5	0,0667	0,0077
	HDR-aprox-grid	89,4	0,0398	0,1390
5a	Res. Quantílico	89,3	0,2320	0,0144
	HDR-aprox	94,6	0,3083	0,0093
	HDR-aprox-grid	89,6	0,2210	0,1425
5b	Res. Quantílico	90,1	0,0511	0,0150
	HDR-aprox	94,6	0,0770	0,0093
	HDR-aprox-grid	90,2	0,0483	0,1397
6a	Res. Quantílico	90,5	0,2537	0,0177
	HDR-aprox	95,0	0,3361	0,0104
	HDR-aprox-grid	90,1	0,2468	0,1891
6b	Res. Quantílico	90,1	0,0563	0,0164
	HDR-aprox	96,0	0,0838	0,0103
	HDR-aprox-grid	90,5	0,0538	0,1703

Por fim, a Tabela 4.2.2 apresenta os resultados comparativos entre o método HDR-aprox

com aproximação e discretização via *grid* e a abordagem mais ingênua baseada em busca exaustiva sobre o simplex (Simplex-grid), usada como referência. Para a discretização dentro do politopo, utilizamos $\delta \approx 0,02$ e para o método ingênuo $\delta \approx 0,005$.

Os dados mostram que, mesmo utilizando um *grid* bastante modesto, o método HDR-aprox-grid produziu áreas e volumes relativos menores em todos os cenários avaliados. No caso com três componentes, o tempo médio de execução foi inferior à 25% do necessário para o Simplex-grid, evidenciando uma vantagem significativa em termos de eficiência computacional.

A cobertura empírica, embora importante em outras comparações, não é o foco central aqui, pois, devido ao custo elevado da busca no simplex com 4 componentes e 1000 iterações, optou-se por reduzir esse número para 100 iterações apenas neste experimento.

Para quatro componentes, o custo computacional do Simplex-grid cresce de forma acentuada, enquanto o HDR-aprox-grid, embora também apresente aumento de tempo, permanece muito mais eficiente. O custo computacional do Simplex-grid nesses casos é mais de 50 vezes o do HDR-aprox-grid. Além disso, as áreas e volumes relativos são menores para o HDR aprox grid. Esses resultados justificam a preferência pelo HDR-aprox-grid em relação ao Simplex-grid.

Tabela 4.2.2 – Comparação de desempenho: HDR-aprox-grid vs. Simplex-grid (3 e 4 componentes).

ϕ	**Método**	**Cob. (%)**	**Área/Volume Relativa(o)**	**Tempo Médio (s)**
Cenário: 3 componentes				
≈ 20	HDR-aprox-grid	89,0	0,2160	0,0240
	Simplex-grid	89,0	0,2330	0,1100
≈ 100	HDR-aprox-grid	90,0	0,0433	0,0160
	Simplex-grid	90,0	0,0514	0,0993
Cenário: 4 componentes				
≈ 20	HDR-aprox-grid	87,0	0,1770	0,2930
	Simplex-grid	87,0	0,1970	15,5000
≈ 100	HDR-aprox-grid	95,0	0,0173	0,1050
	Simplex-grid	95,0	0,0216	16,4000

4.2.2 Resultados: má especificação

A Tabela 4.2.3 apresenta os resultados das simulações para os cenários de má especificação, nos quais o modelo ajustado assume uma distribuição Dirichlet, mas o mecanismo gerador dos dados não segue exatamente essa distribuição, contemplando tanto misturas de distribuições Dirichlet quanto um mecanismo baseado na normal logística. Assim como nos cenários de correta especificação, o nível nominal de cobertura considerado foi de 90%.

De maneira geral, observa-se que os métodos mantêm um desempenho satisfatório em termos de cobertura empírica mesmo sob má especificação do modelo, evidenciando a robustez

dos procedimentos conformais considerados. O método baseado em resíduos quantílicos apresenta coberturas empíricas bastante próximas ao nível nominal em todos os cenários analisados, tipicamente variando entre 89% e 92%. Esse comportamento sugere que, mesmo quando a distribuição assumida no ajuste está incorreta, o método consegue preservar o controle de cobertura, ainda que baseado em restrições marginais.

O método HDR-aprox continua a apresentar um padrão mais conservador sob má especificação. Em todos os cenários considerados, as coberturas empíricas superam o nível nominal, atingindo valores entre aproximadamente 93% e 96%. Esse resultado indica que a construção do conjunto preditivo a partir de uma aproximação da região de maior densidade mantém uma margem adicional de segurança mesmo quando a forma funcional da densidade está incorreta. No entanto, esse conservadorismo vem acompanhado, novamente, de conjuntos preditivos mais amplos.

Além disso, em relação ao método HDR-aprox-grid, observa-se que, assim como ocorreu nos cenários de correta especificação, o HDR-aprox-grid é sólido em termos de cobertura, variando entre $89,5\%$ e $91,3\%$, reduzindo a sobrecobertura do método original e atingindo a cobertura nominal especificada.

Esse comportamento se reflete claramente nas áreas relativas médias. Em comparação aos cenários de correta especificação, as áreas relativas são substancialmente maiores, sobretudo nos cenários baseados em mistura de Dirichlet (7a a 8b), o que era esperado dado o aumento da complexidade do mecanismo gerador. O método HDR-aprox produz sistematicamente as maiores áreas relativas, refletindo a inclusão de regiões adicionais do simplex decorrentes da aproximação por pisos em um contexto de má especificação.

Além disso, de forma geral, o HDR-aprox-grid produz áreas relativas menores do que as do método de resíduos quantílicos, mas para a mistura de Dirichlets isso se inverte nos casos 7b e 8b, diferentemente do que fora observado nos cenários de correta especificação, nos quais o método HDR-aprox-grid atingiu a cobertura nominal e produziu as menores áreas relativas para todas as situações avaliadas.

Assim como nos cenários corretamente especificados, observa-se também o efeito da variabilidade da resposta. Nos cenários com menor dispersão (cenários "b"), as áreas relativas são consistentemente menores para todos os métodos, enquanto as coberturas empíricas permanecem estáveis, indicando que a redução da variabilidade concentra a massa de probabilidade e resulta em conjuntos preditivos mais compactos, mesmo sob má especificação.

No que se refere ao custo computacional, o padrão observado é semelhante ao dos cenários anteriores. Os métodos de resíduos quantílicos e HDR-aprox apresentam tempos médios de execução baixos e comparáveis entre si, enquanto o HDR-aprox-grid incorre em um custo computacional substancialmente maior, devido à busca adicional no *grid*. Ainda assim, os tempos médios permanecem moderados e compatíveis com a realização de estudos de simulação

e aplicações práticas.

Em síntese, os resultados sob má especificação indicam que os métodos avaliados são robustos no sentido de manter cobertura empírica próxima ou superior ao nível nominal. O método de resíduos quantílicos oferece boa estabilidade e baixo custo computacional, o HDR-aprox garante cobertura elevada à custa de conjuntos mais amplos, e o HDR-aprox-grid novamente se destaca por produzir conjuntos preditivos mais eficientes em termos de área relativa, mantendo cobertura adequada mesmo quando o modelo ajustado não coincide com o mecanismo gerador dos dados.

Tabela 4.2.3 – Resultados da simulação para os cenários de má especificação. As métricas reportadas são a cobertura empírica, a área relativa média e o tempo médio de execução de cada método.

Cenário	Método	Cobertura Empírica (%)	Área Relativa	Tempo Médio (s)
7a	Res. Quantílico	90,4	0,5664	0,0172
	HDR-aprox	93,0	0,6696	0,0119
	HDR-aprox-grid	89,5	0,5583	0,1896
7b	Res. Quantílico	90,8	0,4068	0,0210
	HDR-aprox	94,8	0,5281	0,0131
	HDR-aprox-grid	91,1	0,4177	0,2263
8a	Res. Quantílico	89,6	0,4987	0,0195
	HDR-aprox	93,8	0,6029	0,0141
	HDR-aprox-grid	90,6	0,4902	0,2139
8b	Res. Quantílico	90,6	0,3336	0,0195
	HDR-aprox	94,0	0,4436	0,0123
	HDR-aprox-grid	90,3	0,3399	0,2279
9a	Res. Quantílico	91,6	0,2489	0,0183
	HDR-aprox	96,3	0,3264	0,0102
	HDR-aprox-grid	91,3	0,2379	0,1729
9b	Res. Quantílico	91,6	0,0567	0,0171
	HDR-aprox	96,1	0,0841	0,0112
	HDR-aprox-grid	90,9	0,0535	0,1699
10a	Res. Quantílico	90,2	0,2551	0,0221
	HDR-aprox	95,3	0,3345	0,0147
	HDR-aprox-grid	89,8	0,2448	0,2424
10b	Res. Quantílico	90,0	0,0548	0,0194
	HDR-aprox	95,8	0,0816	0,0149
	HDR-aprox-grid	90,2	0,0519	0,2109

CAPÍTULO

5

APLICAÇÃO

Para verificar o desempenho dos métodos em conjuntos de dados reais, utilizamos duas aplicações distintas. A Seção 5.1 apresenta a análise aplicada a dados de estágios do sono, enquanto a Seção 5.2 aborda o problema de alocação de biomassa. Assim como nas simulações, consideramos dois cenários: um com especificação correta do modelo e outro com má especificação.

Em ambos os casos, realizamos uma análise exploratória inicial, seguida do ajuste do modelo e da avaliação diagnóstica por meio de envelopes simulados, com base na abordagem de resíduos proposta por Pereira e Cai (2024).

O principal objetivo das aplicações é conduzir uma análise de regressão paramétrica aliada à tarefa preditiva, com foco na avaliação do desempenho dos métodos em contextos reais. Embora o estudo inferencial tradicional também seja contemplado, o foco não está em obter o melhor ajuste possível, mas sim em investigar o comportamento dos métodos propostos em cenários práticos. Os códigos e os dados das aplicações podem ser consultados em `https://github.com/LucAmaralDS/CP-dirichlet`.

5.1 Aplicação 1: estágios do sono

A distribuição do tempo total de sono entre os estágios N1, N2, N3 e REM pode ser entendida como uma composição, já que as parcelas precisam somar 100%. Em termos funcionais, costuma-se dar atenção especial aos estágios N3 e REM, porque eles estão associados, respectivamente, à recuperação fisiológica e a processos psicológicos importantes, por isso, espera-se que ocupem uma fração relevante do tempo dormido.

Do ponto de vista fisiológico, N1 e N2 são tipicamente classificados como sono leve (RITMALA-CASTREN *et al.*, 2015). O N3 corresponde ao sono profundo, fase em que ocorre maior restauração de músculos e tecidos (HUSSAIN *et al.*, 2022). Já o REM é reconhecido por

movimentos oculares rápidos e mudanças autonômicas, como respiração mais acelerada, e tem papel de destaque em aspectos ligados ao processamento emocional (TEMPESTA *et al.*, 2018). Como referência, González-Naranjo *et al.* (2019) reportam intervalos considerados usuais para o tempo relativo em cada estágio: 3–8% em N1, 45–55% em N2, 15–20% em N3 e 20–25% em REM.

Apesar dessa natureza composicional, é comum que estudos tratem as proporções dos estágios como se fossem independentes, conduzindo análises separadas para cada um deles (GONZÁLEZ-NARANJO *et al.*, 2019; MASKI *et al.*, 2021). Essa abordagem pode ignorar o fato de que aumentar a participação de um estágio necessariamente reduz a de pelo menos outro, já que o total é fixo.

Nesta aplicação, são utilizados os dados de (VEJE *et al.*, 2021), que avaliaram 42 adultos, 22 com encefalite transmitida por carrapato (TBE) e 20 controles saudáveis, com o objetivo de investigar possíveis impactos da doença na qualidade do sono. Aqui, adotamos um modelo de regressão de Dirichlet para examinar se a composição do tempo de sono entre os estágios se altera em função do tempo total dormido (TST) e da presença de TBE.

Como este trabalho também tem como objetivo avaliar os métodos propostos em cenários realistas, buscamos aplicar o SCP a um conjunto de dados reais. No entanto, não há disponível na literatura um conjunto com tamanho amostral suficientemente grande que seja bem ajustado com a regressão Dirichlet.

Diante disso, optamos por ampliar o conjunto de dados referente aos estágios do sono. Inicialmente, ajustamos o modelo corretamente utilizando as 42 observações originais e utilizando como covariável o TST (já que a covariável presença de TBE é não significante). Em seguida, realizamos uma reamostragem *bootstrap* da covariável. Para cada valor de covariável sorteada, temos associadas a ela uma média e uma precisão estimadas a partir do modelo ajustado.

A fim de compor um conjunto final com 1000 observações, utilizamos as 42 observações originais e geramos as 958 restantes da seguinte forma: sorteamos uma covariável com reposição, utilizamos os parâmetros estimados correspondentes e, a partir deles, geramos uma nova resposta sintética. Esse procedimento foi repetido até alcançar o total desejado de 1000 observações.

5.1.1 Análise exploratória

A Tabela 5.1.1 apresenta estatísticas descritivas das proporções de tempo em cada estágio do sono (N1, N2, N3 e REM), além do tempo total de sono (TST), considerando os 42 indivíduos do estudo. Observa-se que, em média, a maior parte do tempo é dedicada ao estágio N2 (média = 0,5336), em conformidade com os valores esperados pela literatura (GONZÁLEZ-NARANJO *et al.*, 2019). Já os estágios N3 e REM, funcionalmente mais relevantes, apresentam médias próximas do limite inferior dos intervalos de normalidade: 13,4% e 19,2%, respectivamente. O estágio N1, por sua vez, representa em média 14% do tempo, com variação considerável (de

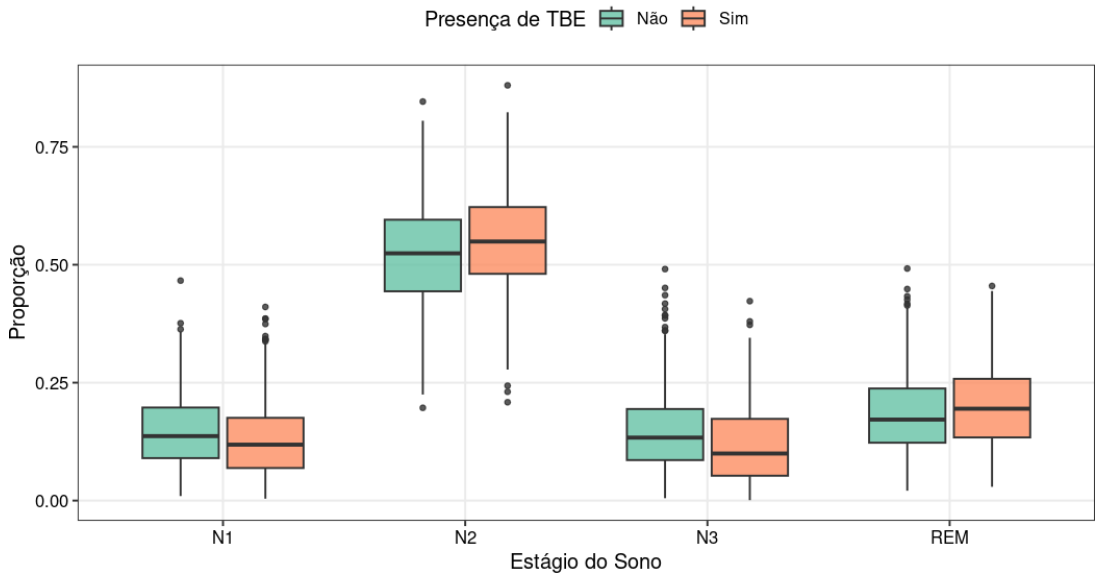

Figura 5.1.1 – Distribuição das proporções dos estágios do sono entre os grupos controle e TBE.

0,4% a 46,6%), o que pode sugerir instabilidade ou fragmentação do sono em alguns indivíduos.

Tabela 5.1.1 – Estatísticas Descritivas das Fases do Sono e Tempo Total de Sono

Variável	**Mín.**	**1º Q.**	**Mediana**	**Média**	**3º Q.**	**Máx.**
N1	0,0040	0,0788	0,1307	0,1403	0,1846	0,4663
N2	0,1967	0,4632	0,5335	0,5336	0,6102	0,8804
N3	0,0010	0,0707	0,1217	0,1341	0,1859	0,4907
REM	0,0211	0,1274	0,1814	0,1920	0,2490	0,4920
TST	3,877	5,622	6,613	6,481	7,217	8,692

A Figura 5.1.1 mostra a distribuição das proporções de cada estágio de sono por grupo. Nota-se que, para todos os estágios, não há muita diferença entre os boxplots para os indivíduos com ou sem TBE. Dessa forma, não parece que distribuição das proporções de tempo em cada estágio de sono varie em função da presença ou não de TBE

A Figura 5.1.2 ilustra a relação entre o tempo total de sono (TST) e a proporção de tempo gasto em cada estágio, separando os grupos controle (0) e com TBE (1). Observa-se uma tendência geral de aumento das proporções de N3 e REM com o aumento do TST, enquanto os estágios N1 e N2 tendem a diminuir, especialmente no grupo com TBE. Essa redistribuição sugere que, à medida que os indivíduos dormem por mais tempo, há uma transição favorável para os estágios mais restauradores do sono, especialmente no grupo acometido pela doença.

5.1.2 Ajuste do modelo

Para o ajuste do modelo consideramos o modelo final utilizado em Pereira e Cai (2024), no qual o vetor de médias da proporção de tempo gasta em cada estágio do sono e a precisão foram

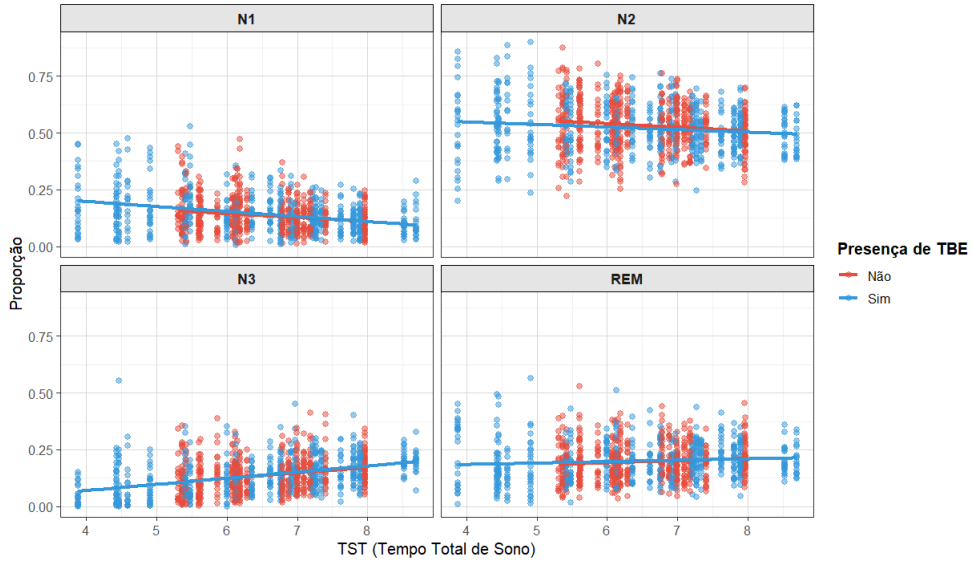

Figura 5.1.2 – Relação entre o tempo total de sono (TST) e a proporção de tempo em cada estágio do sono, por grupo (controle e TBE)

ambos modelados pelo tempo total de sono. Isso porque pelo teste da razão de verossimilhança os autores concluíram que tanto em adultos saudáveis como em adultos com TBE não houve diferença significativa na proporção gasta em cada estágio do sono.

A Tabela 5.1.2 apresenta as estimativas dos coeficientes do modelo de regressão de Dirichlet ajustado para a proporção de tempo gasto em cada estágio do sono, N2, N3 e REM, com o estágio N1 tomado como categoria de referência. O modelo inclui o tempo total de sono (TST) como covariável comum a todos os submodelos, além do submodelo de precisão (ϕ)

Para o submodelo do estágio N2, observa-se um coeficiente positivo para o TST, o que implica que um maior tempo total de sono está associado a um aumento na proporção relativa de N2 em relação ao estágio de referência. Estima-se que a razão entre o tempo médio no estágio N2 e o tempo médio no estágio N1 aumente 12,1% a cada aumento de uma hora no TST.

No submodelo do estágio N3, observa-se novamente um coeficiente positivo para TST, estima-se que a razão entre o tempo médio no estágio N3 e o tempo médio no estágio N1 aumente 31,6% a cada aumento de uma hora no TST. Esse resultado reforça a associação positiva entre o sono profundo e a duração total do sono.

Para o estágio REM, estima-se que a razão entre o tempo médio nesse estágio e o tempo médio no estágio N1 aumente em 20,4% a cada aumento de uma hora no TST. Indicando que períodos mais longos de sono estão associados a maiores proporções relativas de REM, um estágio essencial para o processamento emocional e consolidação da memória.

O submodelo de precisão (ϕ) traz um resultado interessante, o coeficiente de TST foi negativo, o que sugere que, à medida que o tempo total de sono aumenta, há uma leve redução na precisão da variável resposta, indicando maior variabilidade individual na composição do sono em períodos mais longos. Isso pode refletir diferenças interindividuais na arquitetura do

sono, sobretudo nos estágios mais profundos e nos períodos REM, à medida que a duração total aumenta.

Em conjunto, os resultados indicam que o tempo total de sono tem um papel determinante na composição dos estágios de sono, especialmente nos estágios N3 e REM, mais restauradores. A utilização da regressão de Dirichlet se mostra apropriada para captar a natureza proporcional e interdependente desses dados, revelando como variações no tempo total dormido afetam a distribuição relativa entre os estágios do sono.

Tabela 5.1.2 – Estimativas dos parâmetros e erros-padrão no modelo para a proporção de tempo gasto em cada estágio do sono.

Submodelo	Covariável	Estimativa	Erro Padrão	Exp(estim)
μ_{N2}	Intercepto	0,0800	0,2315	1,083
	TST	0,1146	0,0355	1,121
μ_{N3}	Intercepto	-1,7599	0,2487	0,172
	TST	0,2745	0,0379	1,316
μ_{REM}	Intercepto	-1,2929	0,2486	0,274
	TST	0,1857	0,0379	1,204
ϕ	Intercepto	0,8955	0,1294	2,449
	TST	-0,0329	0,0196	0,968

Buscando a completude da análise inferencial, realizamos uma breve análise de diagnóstico a partir do resíduo proposto e recomendado por Pereira e Cai (2024) para dados composicionais. A Figura 5.1.3 apresenta o gráfico de probabilidade normal com envelope simulado para esse resíduo. Nota-se que o gráfico não sugere inadequação do modelo de regressão de Dirichlet para a proporção de tempo gasto em cada estágio do sono, tendo o tempo total de sono como covariável em todos os submodelos.

5.1.3 Análise preditiva

A Tabela 5.1.3 apresenta a comparação entre três métodos de construção de conjuntos preditivos, Quantílico, HDR-aprox e HDR-aprox-grid, avaliados em relação ao volume relativo médio e à cobertura empírica. Utilizamos 70% para treino, 20% para calibração e 10% para teste e um nível de cobertura nominal igual a 90% Além disso, para reduzir a variabilidade dos resultados, repetimos esse processo de divisão 5 vezes.

Observa-se que o método quantílico atinge uma cobertura empírica de 89,6%, muito próxima do nível nominal adotado, com um volume relativo intermediário. Esse resultado indica que, embora o método consiga controlar a cobertura de forma adequada, ele o faz por meio de regiões relativamente amplas, refletindo a natureza marginal do procedimento, que não explora plenamente a estrutura de dependência composicional dos dados.

O método HDR-aprox, por sua vez, apresenta a maior cobertura empírica entre os três procedimentos (97,8%), superando de forma considerável o nível nominal. Esse comportamento

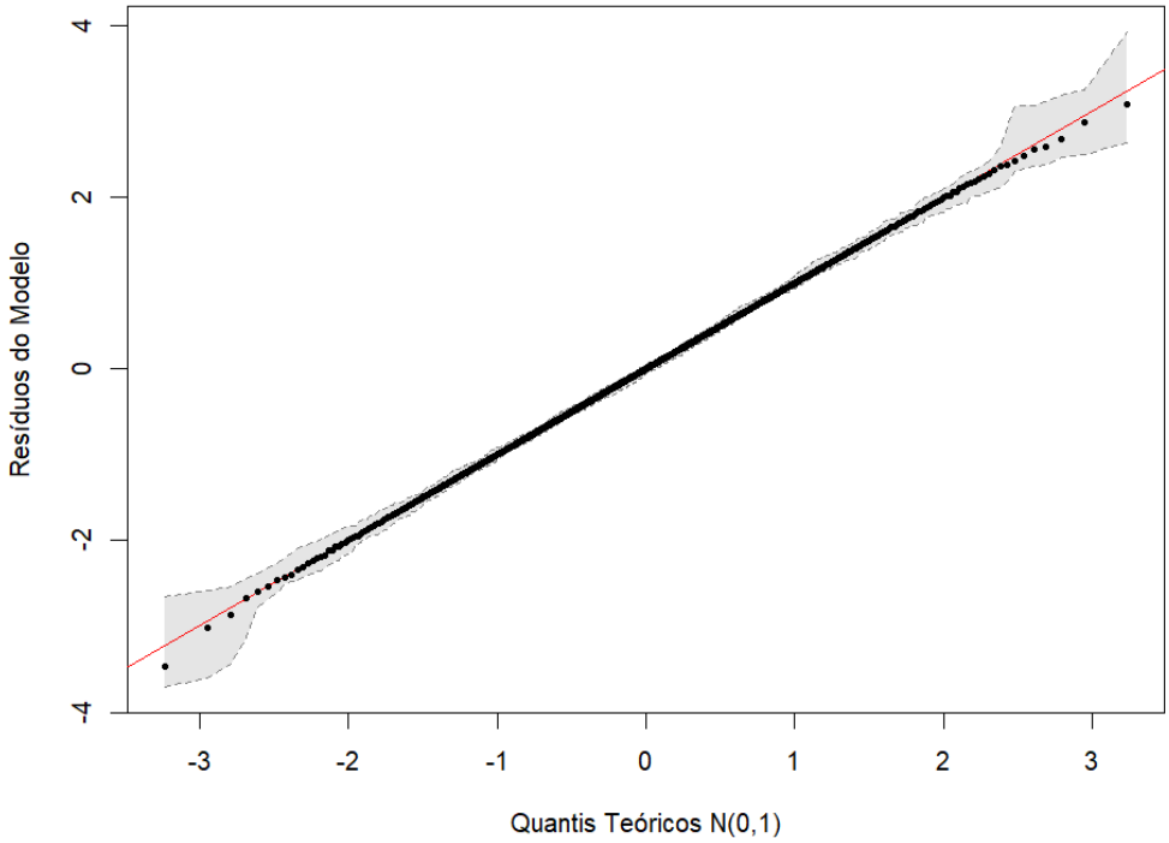

Figura 5.1.3 – Gráfico de probabilidade normal com envelope simulado para os resíduos.

indica um procedimento conservador, no qual a garantia de cobertura é obtida à custa de regiões preditivas mais volumosas. De fato, o volume relativo médio associado a esse método é substancialmente maior, sugerindo que a aproximação analítica da HDR tende a incluir porções adicionais do simplex que não seriam estritamente necessárias para atingir a cobertura desejada.

Em contraste, o método HDR-aprox-grid alcança uma cobertura empírica de 90,7%, maior em relação à do método quantílico, porém com o menor volume relativo médio entre os métodos avaliados. Esse resultado evidencia um ganho importante de eficiência, indicando que a combinação da aproximação da HDR com a verificação em grid físico permite construir regiões preditivas mais concentradas, sem comprometer o controle da cobertura. Em termos práticos, esse método produz conjuntos preditivos mais informativos, capturando de forma mais precisa as regiões de maior densidade da distribuição condicional no simplex.

De maneira geral, os resultados corroboram o *trade-off* clássico entre cobertura e volume das regiões de predição. Enquanto o HDR-aprox privilegia segurança em termos de cobertura, o HDR-aprox-grid se destaca por oferecer um equilíbrio entre cobertura e volume relativo. No contexto dos estágios do sono, em que a interpretação conjunta das proporções é fundamental, esse último método se mostra particularmente atrativo, pois produz regiões preditivas mais compactas e coerentes com a estrutura composicional dos dados.

Tabela 5.1.3 – Desempenho dos métodos preditivos em termos de cobertura empírica e volume médio relativo

Método	**Cobertura Empírica (%)**	**Volume relativo**
Quantílico	89,6	0,182
HDR-aprox	97,8	0,312
HDR-aprox-grid	90,7	0,141

5.2 Aplicação 2: alocação de biomassa

Para esta aplicação utilizamos o banco de dados apresentado em Douma e Weedon (2019). No contexto do problema, sabe-se que as plantas podem alocar biomassa de forma diferenciada entre folhas, caules, raízes e estruturas reprodutivas, seguindo trajetórias ontogenéticas que interagem com o clima predominante.

Diversas abordagens metodológicas são utilizadas para analisar os padrões de alocação resultantes, incluindo o cálculo das proporções ou frações de biomassa dos diferentes órgãos em determinado momento, bem como a análise alométrica de dados coletados entre espécies ou ao longo de períodos experimentais de crescimento.

O conjunto de dados analisado é proveniente de um experimento com duas espécies de plantas com diferentes velocidades de crescimento: *Deschampsia flexuosa* (crescimento lento) e *Holcus lanatus* (crescimento rápido). As plantas foram cultivadas sob dois níveis de fornecimento de nitrato (alto e baixo) por até 49 dias. Durante esse período, indivíduos foram colhidos em diferentes momentos para mensuração da biomassa alocada em folhas, caules e raízes, totalizando 500 observações.

As variáveis de resposta correspondem às proporções da biomassa total alocadas em folhas (LMF), caules (SMF) e raízes (RMF). Essas proporções foram modeladas simultaneamente por meio de regressão de Dirichlet, considerando como variáveis explicativas: espécie, nível de nitrato, tempo (incluindo termo quadrático), biomassa total e suas interações. A precisão do modelo também foi ajustada com base nessas variáveis.

5.2.1 Análise exploratória

A Figura 5.2.1 apresenta a evolução ao longo dos dias das proporções de biomassa alocadas em folhas (LMF), caules (SMF) e raízes (RMF) para as espécies *Deschampsia flexuosa* (representada em vermelho) e *Holcus lanatus* (em azul), cultivadas sob dois níveis de fornecimento de nitrato: alto (painel superior) e baixo (painel inferior).

No tratamento com alto nitrato, observa-se que *D. flexuosa* tende a manter valores mais elevados de alocação foliar em comparação com *H. lanatus*. Com o passar do tempo, ambas as espécies mostram uma leve redução nessa proporção. Em relação à alocação nos caules (SMF), as duas espécies apresentam padrões semelhantes, com proporções relativamente estáveis ao

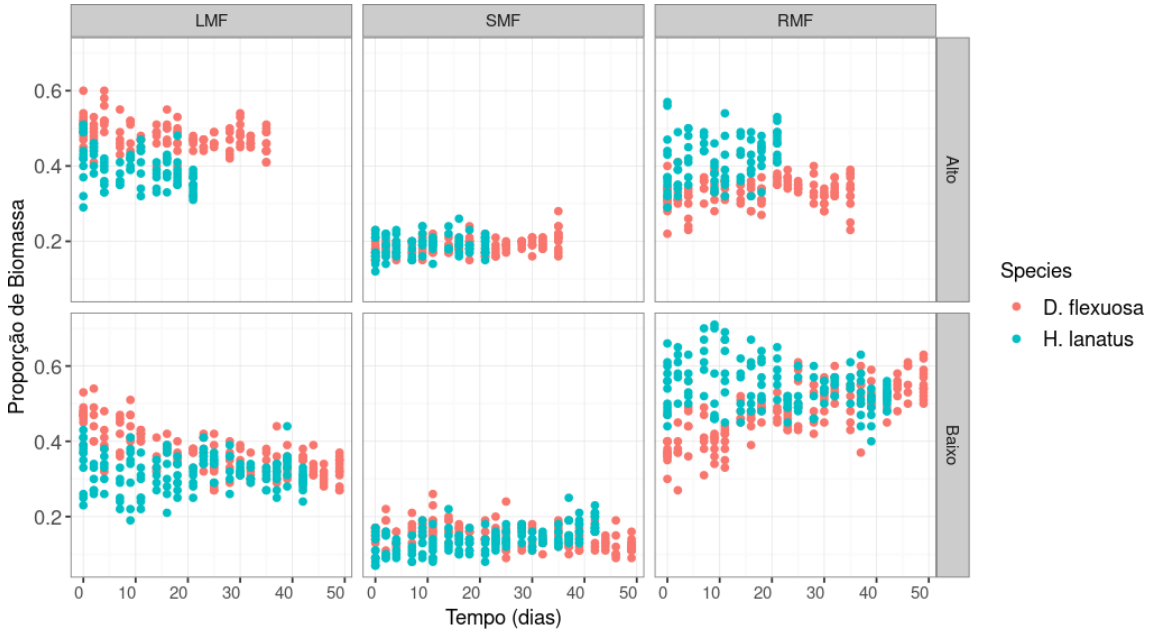

Figura 5.2.1 – Proporção de biomassa alocada em folhas (LMF), caules (SMF) e raízes (RMF) ao longo do tempo para as espécies Deschampsia flexuosa (vermelho) e Holcus lanatus (azul), cultivadas sob dois níveis de fornecimento de nitrato: alto (painel superior) e baixo (painel inferior).

longo do tempo e um discreto aumento nos últimos dias. Já na alocação radicular (RMF), *H. lanatus* apresenta uma proporção maior de biomassa em raízes e a média dessa proporção não parece variar muito ao longo do tempo. Já a espécie *D. flexuosa* mantém valores mais baixos e estáveis ao longo do período experimental.

Sob a condição de baixo fornecimento de nitrato, observa-se um comportamento distinto. A alocação foliar (LMF) mantém-se relativamente constante para ambas as espécies, com *D. flexuosa* apresentando ligeiramente maiores proporções. Na alocação de caules (SMF), os padrões permanecem similares aos do tratamento com alto nitrato, sem variações expressivas. Em contraste, a alocação em raízes (RMF) mostra diferenças mais pronunciadas entre as espécies. *H. lanatus* inicia o experimento com maior proporção de biomassa alocada em raízes, diferença que se reduz ao longo do tempo. Por outro lado, *D. flexuosa* apresenta um aumento gradual na alocação radicular ao longo dos 49 dias. Esses padrões sugerem que as duas espécies respondem de forma diferenciada à limitação de nitrogênio, particularmente na forma como distribuem sua biomassa entre raízes e parte aérea.

De modo geral, os resultados refletem uma interação clara entre espécie, tempo e condição nutricional, o que reforça a escolha por modelos que podem considerar essas interações complexas, como a regressão de Dirichlet adotada na análise. As diferenças de alocação radicular sob baixo fornecimento de nitrato indicam estratégias contrastantes de adaptação ao estresse nutricional, com *H. lanatus* inicialmente investindo mais em raízes, enquanto *D. flexuosa* ajusta sua alocação ao longo do tempo.

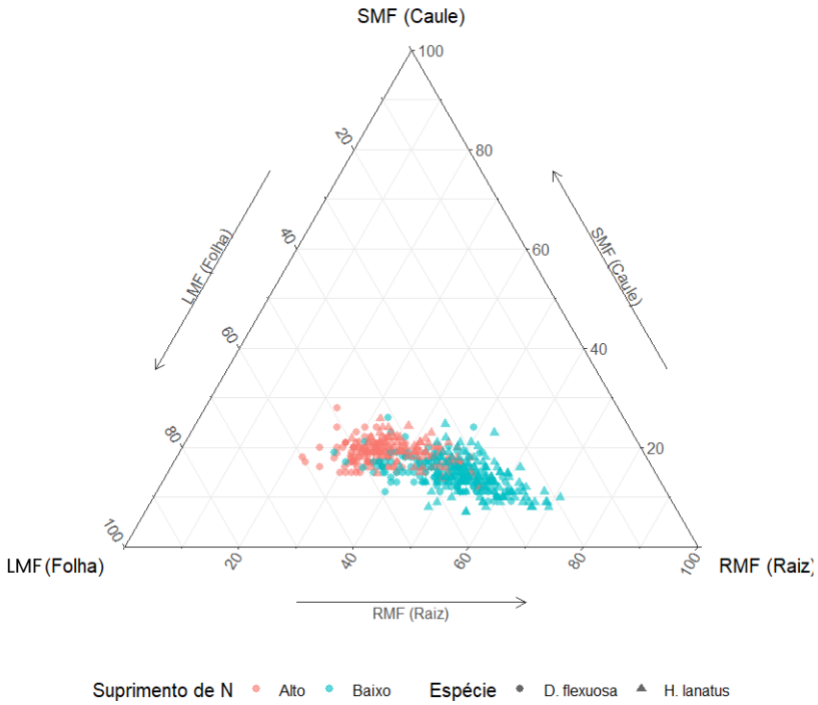

Figura 5.2.2 – Gráfico ternário representando as proporções de biomassa alocadas em folhas (LMF), caules (SMF) e raízes (RMF) para as espécies Deschampsia flexuosa (círculos) e Holcus lanatus (triângulos), cultivadas sob dois níveis de suprimento de nitrogênio (alto e baixo).

A Figura 5.2.2 apresenta um gráfico ternário que resume a alocação proporcional de biomassa entre folhas (LMF), caules (SMF) e raízes (RMF) para diferentes combinações de espécie e suprimento de nitrogênio. Cada ponto no gráfico representa um indivíduo, cuja posição é determinada pela proporção relativa de biomassa alocada em cada um dos três compartimentos, totalizando 100%. Os dados referem-se às espécies *Deschampsia flexuosa* (representada por círculos) e *Holcus lanatus* (representada por triângulos), cultivadas sob dois níveis de fornecimento de nitrogênio: alto (pontos em rosa) e baixo (pontos em azul).

Observa-se que os pontos se agrupam em regiões distintas do gráfico, refletindo variações marcantes nas estratégias de alocação de biomassa entre as espécies e entre os tratamentos. A espécie *H. lanatus*, especialmente sob baixo fornecimento de nitrogênio, concentra-se na região inferior direita do triângulo, sugerindo maior alocação de biomassa em raízes (RMF). Por outro lado, *D. flexuosa* tende a se concentrar mais próximo ao centro e à lateral esquerda do gráfico, o que indica uma distribuição mais equilibrada da biomassa, com maior proporção alocada em folhas (LMF) e menor investimento relativo em raízes.

Além disso, o efeito do suprimento de nitrogênio é evidente: indivíduos cultivados sob baixo nitrogênio (pontos azuis) tendem a deslocar-se para regiões com maior RMF, independentemente da espécie, enquanto os indivíduos sob alto nitrogênio (pontos rosa) se concentram em regiões com distribuição mais homogênea entre folhas, caules e raízes. Esses padrões reforçam a influência conjunta da identidade da espécie e das condições ambientais na definição das

estratégias de alocação de biomassa, evidenciando a interação entre genótipo e ambiente no crescimento vegetal.

5.2.2 Ajuste do modelo

A Tabela 5.2.1 apresenta os coeficientes estimados do modelo de regressão de Dirichlet para as proporções de biomassa alocadas em caule (SMF) e raiz (RMF), bem como para o modelo de precisão, utilizando a especificação dada em (2.5).

Utilizamos a mesma especificação de modelo utilizada em Douma e Weedon (2019), nos quais os autores consideraram interações complexas e dispersão variável. Para a seleção das interações e covariáveis, os autores se basearam no critério de informação de Akaike (AIC).

Tabela 5.2.1 – Estimativas dos parâmetros e erros-padrão no modelo de regressão de Dirichlet para as proporções de biomassa (SMF e RMF).

Submodelo	Covariável	Estimativa	Erro Padrão	Exp(estim)
μ_{SMF}	Intercepto	-1,1110	0,1532	0,329
(Caule)	Tempo^2	-0,0200	0,0139	0,980
	log(Biomassa Total)	0,0426	0,0314	1,044
	Espécie: *H. lanatus*	0,2140	0,0557	1,239
	Tratamento: Baixo N	0,0547	0,0369	1,056
	Tempo	0,0273	0,0429	1,028
	Espécie × Baixo N	-0,1616	0,0628	0,851
	Espécie × Tempo	-0,0163	0,0616	0,984
	Baixo N × Tempo	-0,0230	0,0336	0,977
	Esp. × N × Tempo	0,1363	0,0657	1,146
μ_{RMF}	Intercepto	-1,0179	0,1192	0,361
(Raiz)	Tempo^2	-0,0280	0,0103	0,972
	log(Biomassa Total)	0,1412	0,0243	1,152
	Espécie: *H. lanatus*	0,3442	0,0441	1,411
	Tratamento: Baixo N	0,6648	0,0286	1,944
	Tempo	-0,1293	0,0338	0,879
	Espécie × Baixo N	-0,0117	0,0485	0,988
	Espécie × Tempo	-0,0407	0,0495	0,960
	Baixo N × Tempo	0,2590	0,0263	1,296
	Esp. × N × Tempo	-0,2430	0,0511	0,784
ϕ	Intercepto	5,4220	0,0815	226,331
(Precisão)	Espécie: *H. lanatus*	-0,2625	0,0931	0,769
	Tempo	0,1180	0,0596	1,125
	Tratamento: Baixo N	-0,4637	0,0976	0,629
	Espécie × Tempo	0,2910	0,0933	1,338

Dado que estamos em um cenário de má especificação, não faz sentido interpretar os parâmetros, visto que isso pode levar a conclusões equivocadas nos efeitos das variáveis. Além disso, algumas interações complexas não permitem fácil interpretabilidade. Logo, temos um caso restrito essencialmente à previsão.

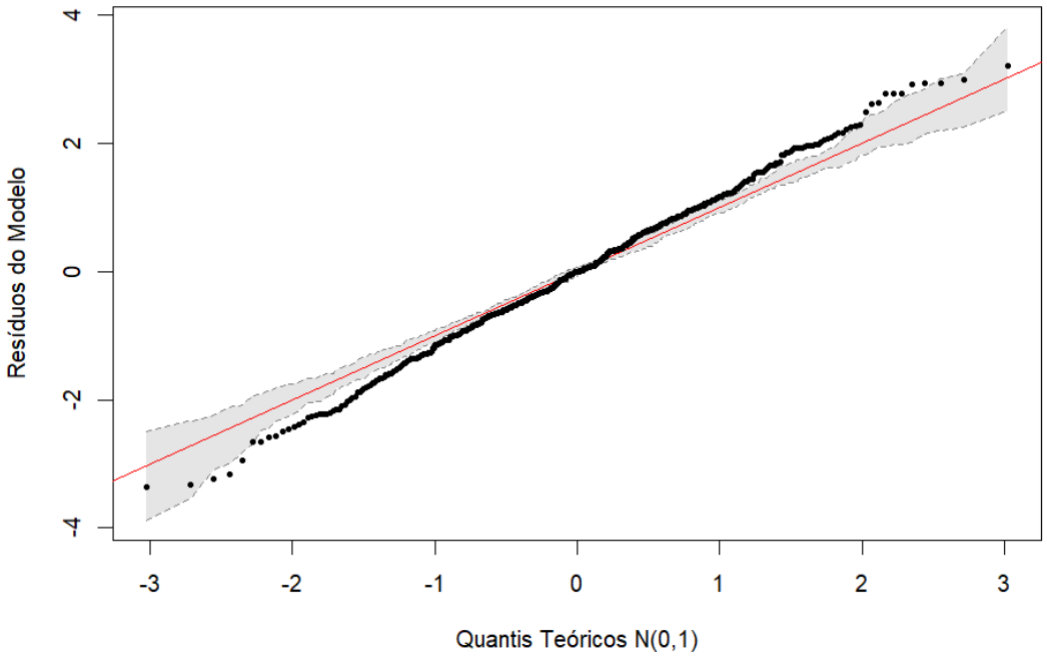

Figura 5.2.3 – Gráfico de probabilidade normal com envelope simulado para os resíduos.

Por fim, para averiguar se o modelo está incorretamente especificado de fato, utilizamos o gráfico de probabilidade normal com envelope simulado para os resíduos na Figura 5.2.3. Considerando o modelo final, o gráfico sugere inadequação do modelo de regressão de Dirichlet para a proporção de biomassa alocada em cada estrutura, já que há inúmeros pontos fora do envelope. A seguir iremos avaliar em termos preditivos essa aplicação em um contexto de má especificação.

5.2.3 Análise Preditiva

Ademais, com o intuito de avaliar o desempenho dos métodos desenvolvidos ao longo deste trabalho, realizamos uma análise dos conjuntos preditivos gerados a partir do banco de dados real. Essa etapa visa não apenas verificar a acurácia preditiva dos modelos, mas sobretudo compreender como as abordagens propostas se comportam diante da complexidade inerente a dados proporcionais e correlacionados, como é o caso da alocação de biomassa em folhas, caules e raízes.

A Tabela 5.2.2 apresenta o desempenho preditivo dos métodos avaliados para a aplicação de alocação de biomassa, em termos de cobertura empírica e área relativa média das regiões de predição. Diferentemente da aplicação anterior, trata-se aqui de um cenário deliberadamente mal especificado, no qual a interpretação inferencial dos parâmetros do modelo não é apropriada e a análise se concentra exclusivamente na qualidade preditiva dos conjuntos construídos.

O método quantílico apresenta cobertura empírica de 92,2%, ligeiramente acima do nível

nominal, associada a uma área relativa intermediária. Esse resultado sugere que, mesmo sob má especificação, o procedimento mantém um controle razoável da cobertura, embora às custas de regiões que não exploram plenamente a geometria do simplex e a dependência inerente entre as proporções de biomassa.

O método HDR-aprox novamente se mostra mais conservador, atingindo a maior cobertura empírica (94,8%), acompanhada da maior área relativa média entre os métodos considerados. Esse comportamento reflete a sensibilidade do procedimento à má especificação do modelo gerador, levando à construção de regiões preditivas mais amplas como forma de compensar incertezas adicionais na distribuição condicional estimada.

Em contraste, o método HDR-aprox-grid alcança uma cobertura empírica de 91,2%, bastante próxima do nível nominal, com a menor área relativa média observada. Esse resultado evidencia que, mesmo em um contexto no qual o modelo subjacente é inadequado, a combinação da aproximação da HDR com a verificação em grid físico é capaz de produzir regiões preditivas mais concentradas e eficientes, sem perda relevante no controle da cobertura.

Em conjunto, os resultados reforçam a robustez dos métodos baseados em SCP frente à má especificação do modelo. Em particular, o desempenho do HDR-aprox-grid destaca sua capacidade de adaptar-se a cenários complexos e não ideais, oferecendo um compromisso mais favorável entre acurácia probabilística e parcimônia geométrica. No contexto da alocação de biomassa, em que as relações entre compartimentos são fortemente interdependentes e moduladas por fatores ambientais e ontogenéticos, esse método se mostra especialmente adequado para fins preditivos.

Tabela 5.2.2 – Desempenho dos métodos preditivos em termos de cobertura empírica e área relativa média
.

Método	**Cobertura Empírica (%)**	**Área relativa**
Quantílico	92,2	0,0423
HDR-aprox	94,8	0,0613
HDR-aprox-grid	91,2	0,0370

Para uma melhor compreensão da forma e do comportamento dos conjuntos preditivos obtidos pelos diferentes métodos, realizamos uma análise visual das regiões de predição em exemplos individuais do conjunto de teste. Essa visualização é particularmente útil em contextos de dados composicionais, uma vez que permite observar diretamente a geometria das regiões no espaço do simplex ternário, bem como a posição relativa das observações em relação aos conjuntos gerados.

Os gráficos da Figura 5.2.4 mostram que o método baseado em resíduos quantílicos produz regiões que se baseiam em limites marginais independentes para cada componente da composição. Embora essa abordagem seja simples e computacionalmente eficiente, ela pode

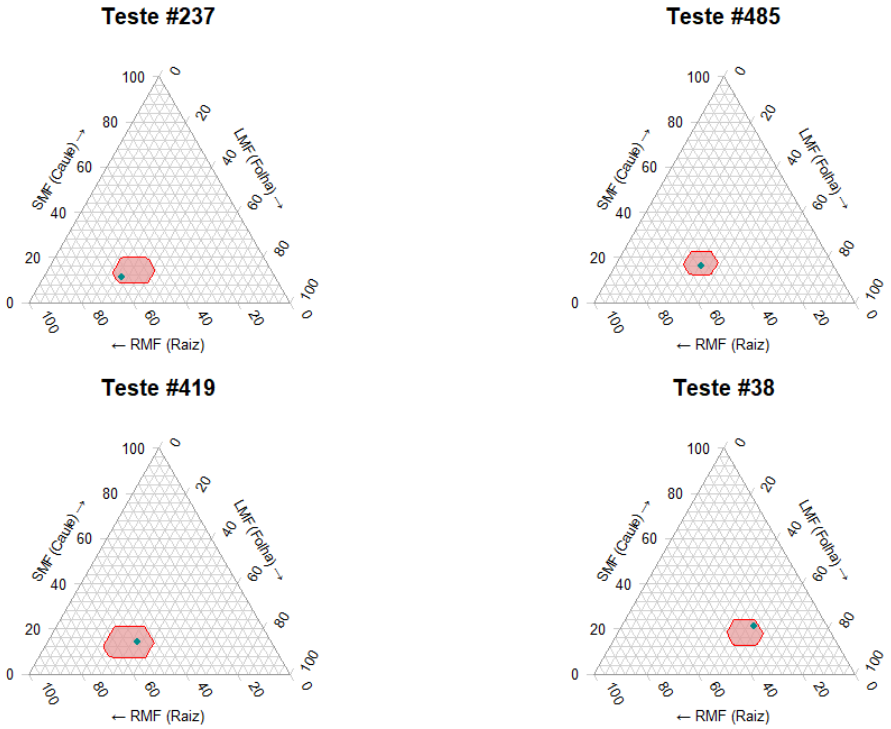

Figura 5.2.4 – Regiões de predição obtidas pelo método baseado em resíduos quantílicos para diferentes composições. As áreas em vermelho nos gráficos ternários representam as regiões de predição. Os pontos verdes correspondem às observações reais de cada teste.

gerar regiões preditivas que não respeitam a estrutura conjunta das variáveis, o que pode levar à inclusão de composições inviáveis.

Em contraste, os resultados da Figura 5.2.5 evidenciam que o método HDR-aprox-grid gera regiões mais adaptadas à forma da distribuição de Dirichlet ajustada, resultando em conjuntos com formas suaves e alinhadas às áreas de maior densidade.

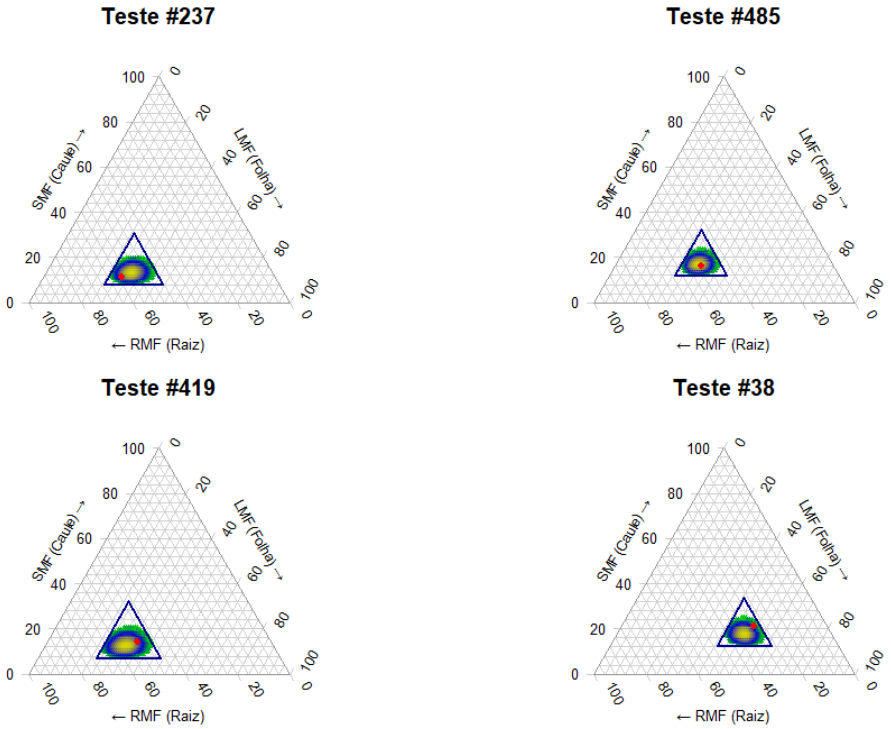

Figura 5.2.5 – Regiões de predição obtidas pelos métodos HDR-aprox e HDR-aprox-grid para diferentes composições. As áreas coloridas nos gráficos ternários indicam as regiões de maior densidade preditiva, onde os valores futuros têm maior probabilidade de ocorrer. Os pontos vermelhos representam as observações reais para cada teste. A HDR-aprox é representada pelo triângulo que envolve a aproximação da HDR via *grid* (HDR-aprox-grid) representada pelo gradiente de cores de azul para amarelo, indicando probabilidades mais altas a medida que a cor se aproxima do amarelo.

CAPÍTULO

6

CONSIDERAÇÕES FINAIS

Este trabalho propôs três abordagens para a construção de regiões de predição válidas em modelos de regressão Dirichlet, com foco em dados composicionais. Elas foram comparadas em estudos de simulação e avaliadas em duas aplicações, uma relacionada a dados de estágio de sono e a outra é no contexto de alocação de biomassa vegetal. As abordagens propostas foram as seguintes: o método baseado em resíduos quantílicos, a abordagem de aproximação de regiões de maior densidade (HDR-aprox) e a extensão HDR-aprox-grid, baseada em *grids*.

Os experimentos revelaram diferenças importantes entre as abordagens. O método baseado em resíduos quantílicos apresentou desempenho bastante estável, com regiões relativamente compactas e baixo tempo de execução, o que o torna atrativo quando se busca simplicidade e eficiência. Ainda assim, por depender de uma construção essencialmente marginal, seu comportamento pode se tornar mais sensível em configurações em que a distribuição no simplex é mais irregular, o que limita sua robustez em alguns cenários.

A abordagem HDR-aprox, por sua vez, mostrou-se sistematicamente mais conservadora. Ao aproximar a região de alta densidade por meio do politopo de pisos, ela tende a incorporar uma margem adicional de segurança, resultando em conjuntos preditivos mais amplos. Esse padrão sugere maior robustez, especialmente quando o mecanismo gerador se afasta do modelo ajustado, mas ao custo de regiões menos informativas do ponto de vista geométrico.

Nesse contexto, o HDR-aprox-grid se destacou como a alternativa mais eficiente em termos de tamanho do conjunto, preservando cobertura adequada. Ao restringir a busca ao interior do politopo e discretizar a região de forma compatível com a geometria da densidade prevista, o método reduz a expansão desnecessária observada no HDR-aprox e, geralmente, também supera o procedimento quantílico em área relativa. Em contrapartida, essa melhoria vem acompanhada de maior custo computacional, associado à avaliação no grid. Apesar disso, a comparação com a discretização ingênua no simplex inteiro evidencia que explorar o politopo de pisos é crucial para viabilizar a construção por *grids* em dimensões maiores, evitando o

crescimento explosivo do custo da busca exaustiva.

Em síntese, os experimentos indicam que o resíduo quantílico fornece uma solução rápida e estável, o HDR-aprox privilegia robustez com maior conservadorismo, e o HDR-aprox-grid oferece um compromisso particularmente atraente quando se deseja reduzir o tamanho do conjunto mantendo validade preditiva.

Além disso, as representações gráficas em diagramas ternários permitiram observar como as regiões preditivas se comportam geometricamente. Isso trouxe evidências visuais de que as regiões baseadas em HDR seguem de forma mais fiel o contorno das densidades previstas, enquanto a abordagem baseada no resíduo quantílico produz regiões mais simétricas e regulares, que menos alinhadas à forma real das distribuições.

Como perspectivas para trabalhos futuros, vislumbram-se diversas direções promissoras que podem complementar e expandir os resultados aqui obtidos. Primeiramente, tem-se o desenvolvimento de métodos para pequenas amostras. Apesar de o método *full conformal* ser uma alternativa natural, a busca via *grid* aliada a muitos ajustes de modelos torna inviável a utilização desse método para esse contexto. Ademais, pode-se desenvolver um trabalho semelhante a este para os casos em que a amostra de dados composicionais contenha valores iguais a zero. Por fim, pode-se desenvolver um trabalho envolvendo predição conforme para dados composicionais, mas usando modelos transformados, os quais diferentemente da distribuição Dirichlet, permitem a captura de correlações positivas, por exemplo.

REFERÊNCIAS

AITCHISON, J. The statistical analysis of compositional data. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 44, n. 2, p. 139–160, 1982. Citado nas páginas 11 e 14.

____. A concise guide to compositional data analysis. In: **Proceedings of the 2nd Compositional Data Analysis Workshop (CoDaWork 2005)**. Girona, Spain: University of Girona, 2005. Citado na página 17.

ALENAZI, A. A review of compositional data analysis and recent advances. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 52, n. 16, p. 5535–5567, 2023. Citado na página 11.

ANGELOPOULOS, A. N.; BATES, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. **arXiv preprint arXiv:2107.07511**, 2021. Citado na página 21.

BARBER, R. F.; CANDES, E. J.; RAMDAS. Conformal prediction beyond exchangeability. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 51, n. 2, p. 816–845, 2023. Citado na página 12.

BARBER, R. F.; CANDES, E. J.; RAMDAS, A.; TIBSHIRANI, R. J. Predictive inference with the jackknife+. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 49, n. 1, p. 486–507, 2021. Citado na página 21.

BARNDORFF-NIELSEN, O. E.; JØRGENSEN, B. Some parametric models on the simplex. **Journal of Multivariate Analysis**, Elsevier, v. 39, n. 1, p. 106–116, 1991. Citado na página 11.

BERTSEKAS, D. **Convex optimization theory**. [S.l.]: Athena Scientific, 2009. v. 1. Citado na página 30.

BRENT, R. P. **Algorithms for minimization without derivatives**. [S.l.]: Courier Corporation, 2013. Citado na página 31.

CHERNOZHUKOV, V.; WÜTHRICH, K.; ZHU, Y. Distributional conformal prediction. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 118, n. 48, p. e2107794118, 2021. Citado na página 12.

CRIBARI-NETO, F.; LIMA, F. P. Resampling-based prediction intervals in beta regressions under correct and incorrect model specification. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 50, n. 5, p. 1398–1416, 2021. Citado na página 12.

DAS, I.; MUKHOPADHYAY, S. On generalized multinomial models and joint percentile estimation. **Journal of Statistical Planning and Inference**, Elsevier, v. 145, p. 190–203, 2014. Citado na página 17.

DEWOLF, N.; BAETS, B. D.; WAEGEMAN, W. Conditional validity of heteroskedastic conformal regression. **Information and Inference: A Journal of the IMA**, Oxford University Press, v. 14, n. 2, p. iaaf013, 2025. Citado na página 12.

DOUMA, J. C.; WEEDON, J. T. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and dirichlet regression. **Methods in Ecology and Evolution**, Wiley Online Library, v. 10, n. 9, p. 1412–1430, 2019. Citado nas páginas 53 e 56.

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 nas páginas 13, 20, 21 e 22.

ESPINHEIRA, P. L.; FERRARI, S. L.; CRIBARI-NETO, F. Bootstrap prediction intervals in beta regressions. **Computational Statistics**, Springer, v. 29, p. 1263–1277, 2014. Citado na página 12.

FENG, C.; LI, L.; SADEGHPOUR, A. A comparison of residual diagnosis tools for diagnosing regression models for count data. **BMC Medical Research Methodology**, BioMed Central, v. 20, n. 1, p. 175, 2020. Citado na página 22.

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 11, 16 e 22.

GHORBANI, H. Mahalanobis distance and its application for detecting multivariate outliers. **Facta Universitatis, Series: Mathematics and Informatics**, p. 583–595, 2019. Citado na página 22.

GLOOR, G. B.; MACKLAIM, J. M.; PAWLOWSKY-GLAHN, V.; EGOZCUE, J. J. Microbiome datasets are compositional: and this is not optional. **Frontiers in microbiology**, Frontiers Media SA, v. 8, p. 2224, 2017. Citado na página 11.

GONZÁLEZ-NARANJO, J. E.; ALFONSO-ALFONSO, M.; GRASS-FERNANDEZ, D.; MORALES-CHACÓN, L. M.; PEDROSO-IBÁÑEZ, I.; FE, Y. Ricardo-de la; PADRÓN-SÁNCHEZ, A. Analysis of sleep macrostructure in patients diagnosed with parkinson's disease. **Behavioral Sciences**, MDPI, v. 9, n. 1, p. 6, 2019. Citado na página 48.

GOTTSCHALK, A. **Mapping potential mineralization: A geostatistical study of B-Horizon Soil Samples from the Rajapalot-Rompas project area in northern Finland**. Dissertação (Mestrado) — A. Gottschalk, 2024. Citado na página 11.

GREENACRE, M.; GRUNSKY, E.; BACON-SHONE, J.; ERB, I.; QUINN, T. Aitchison's compositional data analysis 40 years on: A reappraisal. **Statistical Science**, Institute of Mathematical Statistics, v. 38, n. 3, p. 386–410, 2023. Citado na página 18.

HIJAZI, R. H.; JERNIGAN, R. W. Modelling compositional data using dirichlet regression models. **Journal of Applied Probability & Statistics**, v. 4, n. 1, p. 77–91, 2009. Citado na página 11.

HUSSAIN, I.; HOSSAIN, M. A.; JANY, R.; BARI, M. A.; UDDIN, M.; KAMAL, A. R. M.; KU, Y.; KIM, J.-S. Quantitative evaluation of eeg-biomarkers for prediction of sleep stages. **Sensors**, MDPI, v. 22, n. 8, p. 3079, 2022. Citado na página 47.

HYNDMAN, R. J. Computing and graphing highest density regions. **The American Statistician**, Taylor & Francis, v. 50, n. 2, p. 120–126, 1996. Citado nas páginas 20 e 26.

IZBICKI, R. **Machine Learning Beyond Point Predictions: Uncertainty Quantification**. 1st. ed. [S.l.: s.n.], 2025. 260 p. ISBN 978-65-01-20272-3. Citado na página 21.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. [S.l.]: John Wiley & Sons, 2013. Citado na página 17.

KATO, Y.; TAX, D. M.; LOOG, M. A review of nonconformity measures for conformal prediction in regression. In: **Conformal and probabilistic prediction with applications**. [S.l.]: PMLR, 2023. p. 369–383. Citado na página 12.

LEI, J.; G'SELL, M.; RINALDO, A.; TIBSHIRANI, R. J.; WASSERMAN, L. Distribution-free predictive inference for regression. **Journal of the American Statistical Association**, Taylor & Francis, v. 113, n. 523, p. 1094–1111, 2018. Citado nas páginas 12, 20 e 21.

LEI, J.; RINALDO, A.; WASSERMAN, L. A conformal prediction approach to explore functional data. **Annals of Mathematics and Artificial Intelligence**, Springer, v. 74, p. 29–43, 2015. Citado na página 20.

LEI, J.; ROBINS, J.; WASSERMAN, L. Distribution-free prediction sets. **Journal of the American Statistical Association**, Taylor & Francis, v. 108, n. 501, p. 278–287, 2013. Citado na página 20.

LEMONTE, A. J.; MORENO-ARENAS, G. On residuals in generalized Johnson SB regressions. **Applied Mathematical Modelling**, Elsevier, v. 67, p. 62–73, 2019. Citado na página 22.

LI, D.; AL-MAHAMDA, M. F.; SONG, Y.; FENG, S.; SZE, N. N. An alternate crash severity multicategory modeling approach with asymmetric property. **Analytic methods in accident research**, Elsevier, v. 35, p. 100218, 2022. Citado na página 17.

LIN, J. On the dirichlet distribution. **Department of Mathematics and Statistics, Queens University**, v. 40, 2016. Citado na página 15.

MAIER, M. J. **DirichletReg: Dirichlet Regression for Compositional Data in R**. [S.l.], 2014. (Research Report Series, 125). Citado na página 17.

MASKI, K. P.; COLCLASURE, A.; LITTLE, E.; STEINHART, E.; SCAMMELL, T. E.; NAVIDI, W.; BEHN, C. D. Stability of nocturnal wake and sleep stages defines central nervous system disorders of hypersomnolence. **Sleep**, Oxford University Press US, v. 44, n. 7, p. zsab021, 2021. Citado na página 48.

MELO, T. F.; VARGAS, T. M.; LEMONTE, A. J.; MORENO-ARENAS, G. Higher-order asymptotic refinements in the multivariate dirichlet regression model. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 51, n. 1, p. 53–71, 2022. Citado na página 17.

MORAIS, J.; THOMAS-AGNAN, C.; SIMIONI, M. Using compositional and dirichlet models for market share regression. **Journal of Applied Statistics**, Taylor & Francis, v. 45, n. 9, p. 1670–1689, 2018. Citado na página 14.

OH, J.; KIM, K.-H.; KIM, H.-R.; PARK, S.; YUN, S.-T. Using isometric log-ratio in compositional data analysis for developing a groundwater pollution index. **Scientific reports**, Nature Publishing Group UK London, v. 14, n. 1, p. 12196, 2024. Citado na página 11.

PAPADOPOULOS, H.; GAMMERMAN, A.; VOVK, V. Normalized nonconformity measures for regression conformal prediction. In: **Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)**. [S.l.: s.n.], 2008. p. 64–69. Citado nas páginas 12 e 21.

PAWLOWSKY-GLAHN, V.; EGOZCUE, J. J.; TOLOSANA-DELGADO, R. Lecture notes on compositional data analysis. **Universitat de Girona, Girona**, 2007. Citado na página 18.

PEREIRA, G. H.; CAI, J. A class of bootstrap based residuals for compositional data. **arXiv preprint arXiv:2403.13544**, 2024. Citado nas páginas 17, 38, 47, 49 e 51.

PEREIRA, G. H. A. 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 22.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>. Citado na página 31.

RITMALA-CASTREN, M.; VIRTANEN, I.; LEIVO, S.; KAUKONEN, K.-M.; LEINO-KILPI, H. Sleep and nursing care activities in an intensive care unit. **Nursing & health sciences**, Wiley Online Library, v. 17, n. 3, p. 354–361, 2015. Citado na página 47.

ROMANO, Y.; PATTERSON, E.; CANDES, E. Conformalized quantile regression. **Advances in Neural Information Processing Systems**, v. 32, 2019. Citado na página 12.

SHAFER, G.; VOVK, V. A tutorial on conformal prediction. **Journal of Machine Learning Research**, v. 9, n. 3, 2008. Citado nas páginas 12 e 20.

SILVA, T. F. d. N. M. d. **Estimação do posto da matriz dos parâmetros do modelo de regressão Dirichlet**. Dissertação (Dissertação (Mestrado em Estatística)) — Universidade Federal de Pernambuco, Recife, 2004. Citado nas páginas 15 e 16.

TEMPESTA, D.; SOCCI, V.; GENNARO, L. D.; FERRARA, M. Sleep and emotional processing. **Sleep medicine reviews**, Elsevier, v. 40, p. 183–195, 2018. Citado na página 48.

TIBSHIRANI, R. J.; BARBER, R. F.; CANDES, E.; RAMDAS, A. Conformal prediction under covariate shift. **Advances in Neural Information Processing Systems**, v. 32, 2019. Citado na página 12.

VEJE, M.; STUDAHL, M.; THUNSTRÖM, E.; STENTOFT, E.; NOLSKOG, P.; CELIK, Y.; PEKER, Y. Sleep architecture, obstructive sleep apnea and functional outcomes in adults with a history of tick-borne encephalitis. **PLoS One**, Public Library of Science San Francisco, CA USA, v. 16, n. 2, p. e0246767, 2021. Citado na página 48.

VOVK, V. Conditional validity of inductive conformal predictors. In: PMLR. **Asian Conference on Machine Learning**. [S.l.], 2012. p. 475–490. Citado na página 21.

VOVK, V.; GAMMERMAN, A.; SHAFER, G. **Algorithmic learning in a random world**. [S.l.]: Springer, 2005. v. 29. Citado nas páginas 12, 20 e 21.

WU, Z.; LEISEN, F.; RUBIO, F. J. Conformalized regression for continuous bounded outcomes. **arXiv preprint arXiv:2507.14023**, 2025. Citado na página 12.

APÊNDICE

A

ALGORITMOS

Algoritmo 1 – Conjunto de Predição *Split Conformal* para Regressão Dirichlet

Entrada: Base $\mathscr{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ com $\mathbf{y}_i \in \Delta^D$; partição em índices de treino I_1 e calibração I_2 (com $n_{\text{cal}} = |I_2|$); especificação do modelo em (2.5); nível de significância $\alpha \in (0,1)$.
Saída: Conjunto de predição conforme $\mathscr{C}_{\text{split}}(\mathbf{x}_{n+1})$ para uma nova covariável $\mathbf{x}_{n+1}$.

1. **Divisão:** Divida os dados em treino I_1 e calibração I_2.
2. **Ajuste:** Usando I_1, ajuste a regressão Dirichlet de (2.5) e obtenha os parâmetros $\widehat{\mu}_i = \widehat{\mu}(\mathbf{x}_i)$ e $\widehat{\phi}_i = \widehat{\phi}(\mathbf{x}_i)$.
3. **Escores de calibração:** Para cada $i \in I_2$:
   (a) Para cada componente $j = 1, \ldots, D$, compute
   $$U_{ij} \;=\; F\big(y_{ij}; \widehat{\mu}_{ij}, \widehat{\phi}_i\big),$$
   em que F é a CDF da beta marginal induzida pela Dirichlet ajustada em x_i para o componente j.
   (b) Transforme U_{ij} via normal padrão: $r_{ij}^{q} = \Phi^{-1}(U_{ij})$.
   (c) Defina o escore de não-conformidade: $S_i = \max_j |r_j^q|$.
4. **Quantil conforme:** Seja $S_{(1)} \leq \cdots \leq S_{(n_{\text{cal}})}$ a ordenação crescente dos $\{S_i : i \in I_2\}$. Defina
   $$\widehat{q}_{1-\alpha} = S_{(\lceil (n_{\text{cal}}+1)(1-\alpha) \rceil)}.$$
5. **Predição para $\mathbf{X}_{n+1}$:**
   (a) Calcule $p_{\text{inf}} = \Phi(-\widehat{q}_{1-\alpha})$ e $p_{\text{sup}} = \Phi(\widehat{q}_{1-\alpha})$.
   (b) Obtenha $\widehat{\boldsymbol{\mu}}_{n+1} = \widehat{\boldsymbol{\mu}}(\mathbf{x}_{n+1})$ e $\widehat{\phi}_{n+1} = \widehat{\phi}(\mathbf{x}_{n+1})$, para cada componente j, construa o intervalo marginal
   $$I_j(\mathbf{x}_{n+1}) = \Big[F^{-1}_{\text{Beta}_j}(p_{\text{inf}}) \;,\; F^{-1}_{\text{Beta}_j}(p_{\text{sup}})\Big].$$
   (c) Retorne o conjunto de predição cartesiano restrito ao simplex:
   $$\mathscr{C}_{\text{split}}(\mathbf{x}_{n+1}) = \Big\{ \mathbf{y} \in \Delta^D : \; y_j \in I_j(\mathbf{x}_{n+1}) \;\; \forall\, j = 1, \ldots, D \Big\}.$$

Algoritmo 2 – Conjunto de Predição *Split Conformal* com Aproximação da HDR

Entrada: Base $\mathscr{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ com $y_i \in \Delta^D$; partição em índices de treino I_1 e calibração I_2 (com $n_{\text{cal}} = |I_2|$); especificação do modelo Dirichlet $\lambda_j(\mathbf{x}) = \phi(\mathbf{x})\mu_j(\mathbf{x})$; nível $\alpha \in (0,1)$.
Saída: Conjunto de predição conforme $\mathscr{T}(\mathbf{x}_{n+1})$ para uma nova covariável $\mathbf{x}_{n+1}$.

1. **Divisão:** Divida os dados em treino I_1 e calibração I_2.

2. **Ajuste do modelo:** Usando I_1, ajuste a regressão Dirichlet e obtenha os preditores

$$\widehat{\boldsymbol{\mu}}(\mathbf{x}), \qquad \widehat{\phi}(\mathbf{x}).$$

3. **Escores de calibração (score de verossimilhança):** Para cada $i \in I_2$, defina

$$s_i \;=\; -\log f\big(\mathbf{y}_i \mid \mathbf{x}_i; \widehat{\boldsymbol{\mu}}(\mathbf{x}_i), \widehat{\phi}(\boldsymbol{x_i})\big) \;=\; -\log\Gamma(\widehat{\phi}_i) + \sum_{j=1}^{D} \log\Gamma(\widehat{\mu}_{ij}\widehat{\phi}_i) - \sum_{j=1}^{D} (\widehat{\mu}_{ij}\widehat{\phi}_i - 1)\log y_{ij}.$$

4. **Quantil conforme:** Seja $s_{(1)} \le \cdots \le s_{(n_2)}$ a ordenação crescente dos $\{s_i : i \in I_2\}$. Defina

$$\widehat{q}_{1-\alpha} \;=\; s_{(\lceil (n_{\text{cal}}+1)(1-\alpha) \rceil)}.$$

5. **Parâmetros no ponto de teste x_{n+1}:** Calcule

$$\widehat{\boldsymbol{\mu}}_* = \widehat{\boldsymbol{\mu}}(\mathbf{x}_{n+1}), \quad \widehat{\phi}_* = \widehat{\phi}(\mathbf{x}_{n+1}),$$

$$\widehat{t}_{1-\alpha} \;=\; -\widehat{q}_{1-\alpha} - \log\Gamma(\widehat{\phi}_*) + \sum_{j=1}^{D} \log\Gamma(\widehat{\mu}_{*j}\widehat{\phi}_*), \qquad w_j \;=\; \widehat{\phi}_*\widehat{\mu}_{*j} - 1.$$

6. **Pisos mínimos via otimização 1D (um i por vez):** Para cada $i \in \{1,\ldots,D\}$, resolva o problema convexo

$$\min_{y \in \mathbb{R}^D} y_i \quad \text{sujeito a} \quad \sum_j y_j = 1, \;\; \sum_j w_j \log y_j \ge t_{1-\alpha}, \;\; y_j \ge 0,$$

obtendo τ_i. Proceda em duas etapas:

   (a) **Forma fechada (caso interior):** Se $w_j > 0$ para todo j, use as KKT para obter

$$\theta = \frac{1}{\frac{w_i}{1+\rho} + \frac{W - w_i}{\rho}}, \qquad \tau_i = \frac{\theta\, w_i}{1+\rho}, \qquad W = \sum_{j=1}^{D} w_j,$$

   em que $\rho > 0$ é a única raiz da equação unidimensional

$$F_i(\rho) \;=\; w_i \log\rho + (W - w_i)\log(1+\rho) - W\log\big(w_i\rho + (W-w_i)(1+\rho)\big) + \sum_{j=1}^{D} w_j \log w_j - \widehat{t}_{1-\alpha} \;=\; 0.$$

   (b) **Fallback (caso limite/negativo):** Se algum $w_j \le 0$ (ou a busca falhar), $\tau_i = 0$.

7. **Conjunto aproximado:** Defina o politopo de pisos

$$\mathscr{T}(\mathbf{x}_{n+1}) \;=\; \Big\{ \mathbf{y} \in \Delta^D : \; y_i \ge \tau_i, \;\; i = 1,\ldots,D \Big\}.$$

8. **Saída:** Retorne $\mathscr{T}(\mathbf{x}_{n+1})$ como o conjunto de predição para $\mathbf{y}_{n+1}$.


ICMC
USP
SÃO CARLOS