UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
BACHARELADO EM ESTATÍSTICA

Jaqueline Lamas da Silva

Misturas Finitas de Modelos Parcialmente Lineares: Uma abordagem via P-splines para estimação das componentes não paramétricas

Juiz de Fora
2025

Jaqueline Lamas da Silva

Misturas Finitas de Modelos Parcialmente Lineares: Uma abordagem via P-splines para estimação das componentes não paramétricas

Trabalho de conclusão de curso apresentado ao Bacharelado em Estatística da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Bacharel em Estatística.

Orientador: Prof.ª Dr.ª Camila Borelli Zeller

Coorientador: Prof. Dr. Clécio da Silva Ferreira

Juiz de Fora

2025

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Silva, Jaqueline Lamas da.
Misturas Finitas de Modelos Parcialmente Lineares : Uma abordagem via P-splines para estimação das componentes não paramétricas / Jaqueline Lamas da Silva. – 2025.
99 f. : il.

Orientador: Camila Borelli Zeller
Coorientador: Clécio da Silva Ferreira
Trabalho de conclusão de Curso (graduação) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Bacharelado em Estatística, 2025.

1. misturas finitas. 2. modelos parcialmente lineares. 3. P-splines. 4. algoritmo EM. I. Zeller, Camila Borelli, orient. II. Ferreira, Clécio da Silva, coorient. III. Título.

Jaqueline Lamas da Silva

Misturas Finitas de Modelos Parcialmente Lineares: Uma abordagem via P-splines para estimação das componentes não paramétricas

Trabalho de conclusão de curso apresentado ao Bacharelado em Estatística da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Bacharel em Estatística.

Aprovada em 25 de agosto de 2025

BANCA EXAMINADORA

Prof.ª Dr.ª Camila Borelli Zeller - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Clécio da Silva Ferreira - Coorientador
Universidade Federal de Juiz de Juiz de Fora

Prof. Dr. Ronaldo Rocha Bastos
Universidade Federal de Juiz de Fora

Prof. Dr. Tiago Maia Magalhães
Universidade Federal de Juiz de Fora

Dedico este trabalho ao meu pai, Vardeli, que nos deixou muito cedo.

AGRADECIMENTOS

Agradeço à minha mãe, Valdineia, por todo o cuidado e apoio. Agradeço aos meus irmãos Aline e Gil pelo companheirismo ao longo dos anos e, claro, por todas as caronas.

Agradeço a todos os professores que contribuíram para minha jornada acadêmica até aqui. Agradeço especialmente à Professora Camila, que aceitou me orientar no desenvolvimento deste trabalho, sou imensamente grata por toda a ajuda e por todos os ensinamentos. Gostaria também de expressar meus agradecimentos ao Professor Clécio, cuja orientação em trabalhos anteriores me proporcionou uma base essencial para a condução desta pesquisa. Agradeço aos professores Ronaldo e Tiago por aceitarem participar da avaliação deste trabalho e por contribuírem para sua melhoria.

Agradeço também aos meus familiares, amigos e colegas que, de diversas formas, me ajudaram a viver o presente.

Não posso deixar de agradecer ao programa de financiamento do CNPq e aos programas de bolsas e auxílios financeiros da UFJF que fizeram possível minha permanência na universidade.

Meus sinceros agradecimentos a todos.

Os seres humanos podem ansiar pela certeza absoluta; [...] Mas a história da ciência [...] ensina que o máximo que podemos esperar é um aperfeiçoamento sucessivo de nosso entendimento, um aprendizado por meio de nossos erros, uma abordagem assintótica do Universo, mas com a condição de que a certeza absoluta sempre nos escapará. (SAGAN, Carl. *O mundo assombrado pelos demônios.* 1995, p.46)

RESUMO

Os modelos de regressão, consolidados ao longo do desenvolvimento estatístico, permanecem como ferramenta primordial para investigar relações entre preditores e desfechos. Em sua formulação clássica, esse tipo de modelo assume que as observações são provenientes de uma única população homogênea. No entanto, na prática, características não observadas podem gerar comportamentos distintos entre subgrupos de observações. Em tais circunstâncias, podemos utilizar modelos de misturas de regressão para incorporar essa heterogeneidade ao modelo, estimando não apenas os parâmetros específicos de cada componente da mistura (subgrupo), mas também as probabilidades a posteriori de cada observação pertencer a cada componente, as quais podem ser utilizadas em contextos de classificação e clusterização (ou agrupamento) no âmbito de aprendizagem supervisionada e não supervisionada, respectivamente. No presente trabalho, estudamos misturas de modelos parcialmente lineares com a adoção de P-splines para estimação das componentes não paramétricas. Nossa configuração permite que diferentes covariáveis lineares e não-lineares componham a estrutura semiparamétrica de cada grupo. Os estimadores de máxima verossimilhança penalizada foram obtidos através de um algoritmo do tipo EM, enquanto os erros padrão foram calculados via matriz de informação empírica. Para a seleção dos parâmetros de suavização das curvas e do número de grupos, utilizamos o critério de informação Bayesiano (BIC). A metodologia proposta foi avaliada através de estudos de simulação e por meio de aplicação a dados reais. Além disso, este trabalho também propõe uma contribuição metodológica inédita: a avaliação da qualidade do ajuste dos modelos de mistura por meio da construção de envelopes simulados baseados nos resíduos quantílicos.

Palavras-chave: misturas finitas; modelos parcialmente lineares, P-splines; algoritmo EM.

ABSTRACT

Regression models, well-established in statistical development, remain a primary tool for investigating relationships between predictors and outcomes. In their classical formulation, these models assume that observations come from a single homogeneous population. However, in practice, unobserved characteristics may lead to distinct behaviors among observation subgroups. In such circumstances, we can employ regression mixture models to incorporate this heterogeneity, estimating not only the specific parameters of each mixture component (subgroup) but also the posterior probabilities of each observation belonging to each component - which can be used in classification and clustering contexts, in the framework of supervised and unsupervised learning, respectively. In this work, we study mixtures of partially linear models using P-splines for estimating the nonparametric components. Our configuration allows different linear and nonlinear covariates to compose the semiparametric structure of each group. The penalized maximum likelihood estimators were obtained through an EM-type algorithm, while standard errors were calculated via the empirical information matrix. For selecting the curve smoothing parameters and the number of groups, we used the Bayesian Information Criterion (BIC). The proposed methodology was evaluated through simulation studies and real data applications. In addition, this study introduces a novel methodological contribution: the assessment of goodness-of-fit of mixture models through simulated envelopes constructed from randomized quantile residuals.

Keywords: finite mixtures; partially linear models; P-splines; EM algorithm.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Os modelos baseados em misturas finitas constituem uma ferramenta estatística poderosa. Sua flexibilidade estrutural permite a modelagem de fenômenos diversos, o que justifica tanto sua ampla adoção em aplicações práticas quanto seu contínuo desenvolvimento teórico.

> Campos nos quais os modelos de misturas têm sido aplicados com sucesso incluem astronomia, biologia, genética, medicina, psiquiatria, economia, engenharia e marketing, entre muitos outros campos das ciências biológicas, físicas e sociais. Nessas aplicações, os modelos de misturas finita fundamentam diversas técnicas em áreas principais da estatística, incluindo análises de cluster e de classe latente, análise discriminante, análise de imagens e análise de sobrevivência, além de seu papel mais direto na análise de dados e inferência de modelos descritivos para distribuições. (MCLACHLAN; PEEL, 2000, p. 1, tradução nossa).

Segundo Montgomery et al. (2012), os modelos de regressão consolidaram-se como uma das principais ferramentas para investigar relações entre variáveis explicativas (preditoras) e variáveis resposta (desfechos). Em sua formulação clássica, esses modelos assumem que todas as observações pertencem a uma única população homogênea. No entanto, essa suposição frequentemente se mostra irrealista, uma vez que características não observadas podem gerar comportamentos distintos entre diferentes subgrupos da população. Considere, por exemplo, um cenário em que uma determinada covariável apresenta uma relação positiva com a variável resposta em um subgrupo, mas negativa em outro. Nessas situações, os modelos de misturas de regressão tornam-se particularmente úteis, pois permitem incorporar explicitamente a heterogeneidade estrutural dos dados ao modelo. Por meio dessa abordagem, é possível estimar tanto os parâmetros específicos de cada componente da mistura (ou subgrupo latente) quanto as probabilidades a posteriori de cada observação pertencer a cada componente, as quais podem ser utilizadas em contextos de classificação ou agrupamento.

Conforme destacado por Yao e Xiang (2024), como as suposições dos modelos paramétricos, como a linearidade, podem não ser satisfeitas, tanto do ponto de vista teórico quanto prático, diversos modelos de misturas finitas de regressão semiparamétricos têm sido propostos nos últimos anos, demonstrando desempenho superior em várias situações. Além disso, diferentes estudos têm considerado a presença de covariáveis relacionadas de forma não linear com a resposta, como nos trabalhos de Zhang e Pan (2020), Skhosana et al. (2023) e Hwang et al. (2025). O primeiro artigo propõe misturas de modelos aditivos parcialmente lineares, em que as curvas são modeladas por meio da técnica SBK (Spline-Based Kernel). O segundo trabalho propõe as misturas de modelos parcialmente lineares, utilizando a abordagem de verossimilhança local com kernel para a estimação

dos parâmetros, apresentando ainda soluções para o problema de label switching e para a estimação de curvas suaves. Por fim, o terceiro estudo apresenta misturas aplicadas a modelos parcialmente lineares e igualmente recorre à metodologia de verossimilhança local com kernel para o ajuste dos componentes.

Modelos de misturas de regressão têm se revelado particularmente eficazes na análise de conjuntos de dados contemporâneos, os quais frequentemente apresentam elevada complexidade e heterogeneidade. Tais modelos são especialmente valiosos em contextos onde estruturas latentes e regimes distintos tendem a emergir, como nas áreas médica, econômica e ambiental. No âmbito da saúde, por exemplo, pacientes com perfis clínicos aparentemente semelhantes podem responder de maneira distinta a um mesmo tratamento, em razão de fatores genéticos ou ambientais não observados. Neste trabalho, consideramos uma aplicação com dados reais na área médica, inspirada no estudo de Reaven e Miller (1979), cujos detalhes serão apresentados no Capítulo 5. No campo econômico, dados extraídos do banco STARS do Banco Mundial, compostos por PIB real, capital humano, estoque de capital e educação para 82 países no período de 1960 a 1987, revelam a possibilidade de identificar grupos de países com dinâmicas de crescimento distintas, como demonstrado por Duffy e Papageorgiou (2000) e posteriormente revisitado por Zhang e Pan (2020). De forma semelhante, Skhosana et al. (2023) analisaram a relação entre consumo de energia, urbanização e emissões de CO_2, utilizando dados da base Our World in Data para países da OCDE entre 1990 e 2019, identificando regimes com comportamentos distintos e relações não lineares entre as variáveis. Em outro estudo, Skhosana et al. (2022) examinaram dados sobre emissões per capita de CO_2 e Produto Nacional Bruto per capita para 145 países no ano de 1992, evidenciando agrupamentos latentes de países com trajetórias diferenciadas de desenvolvimento econômico.

Inspirados por essa literatura, propomos o uso de misturas de modelos parcialmente lineares com uma estrutura semiparamétrica flexível, distinguindo-nos dos trabalhos existentes ao empregar P-splines para a estimação das componentes não paramétricas. Nossa formulação permite que diferentes covariáveis, tanto lineares quanto não lineares, sejam incorporadas de forma específica à estrutura de cada componente da mistura, proporcionando maior flexibilidade na modelagem e melhor adequação a cenários com heterogeneidade entre subgrupos latentes.

Além da formulação de um modelo de mistura parcialmente linear com estrutura semiparamétrica flexível, utilizando P-splines para a estimação das componentes não paramétricas, este trabalho também propõe uma contribuição metodológica inédita: a avaliação da qualidade do ajuste dos modelos de mistura por meio da construção de envelopes simulados baseados nos resíduos quantílicos. Essa abordagem diagnóstica, usualmente aplicada em modelos clássicos, é aqui estendida ao contexto de modelos de mistura, oferecendo uma nova ferramenta para a verificação gráfica da adequação do modelo aos dados. Tal proposta contribui para o aprimoramento das técnicas de diagnóstico em

modelos com estruturas complexas e latentes, ampliando as possibilidades de análise e validação em aplicações reais.

1.1 OBJETIVOS

O principal objetivo desta pesquisa é investigar o uso de misturas de modelos parcialmente lineares com estimação baseada em P-splines, onde buscamos contribuir para a etapa de avaliação do ajuste por meio da proposta de construção de envelopes simulados. Para tanto, estabelecemos as seguintes metas: (i) formular o modelo proposto com base na estrutura de mistura semiparamétrica; (ii) desenvolver um procedimento de estimação via máxima verossimilhança e computar os erros padrão dos estimadores; (iii) avaliar o desempenho da abordagem proposta por meio de estudos de simulação; (iv) demonstrar a aplicabilidade da metodologia por meio de análises com dados reais; e (v) verificar a qualidade do ajuste dos modelos propostos, na aplicação em dados reais, por meio da construção de envelopes simulados baseados nos resíduos quantílicos.

1.2 ORGANIZAÇÃO DO TRABALHO

Este trabalho está estruturado da seguinte maneira: o Capítulo 2 explora a metodologia das componentes não paramétricas, enquanto o Capítulo 3 apresenta a estrutura do modelo proposto, abordando também os processos de estimação, incluindo a descrição do algoritmo desenvolvido, a obtenção dos erros padrão aproximados por meio da matriz empírica, os critérios utilizados e o procedimento de construção dos envelopes simulados. O Capítulo 4 é dedicado à avaliação do desempenho da abordagem proposta por meio de estudos de simulação. Já o Capítulo 5 apresenta os resultados das aplicações em dados reais. Finalmente, o último capítulo traz uma síntese das conclusões e das principais contribuições deste estudo.

2 MODELAGEM DA COMPONENTE NÃO PARAMÉTRICA

Neste capítulo, apresenta-se a fundamentação teórica relacionada à representação das curvas envolvidas no modelo que será formalmente definido no Capítulo 3. No contexto da regressão não paramétrica (Green e Silverman, 1994) considera-se uma variável resposta, relacionada a uma covariável de forma não linear,

$$\mathbf{Y} = g(\mathbf{t}) + \boldsymbol{\varepsilon}, \tag{2.1}$$

em que $Y = (Y_1, \cdots, Y_n)$ é um vetor coluna de dimensões $n \times 1$ que contém os valores da variável resposta para cada uma das n observações; $\boldsymbol{t} = (t_1, \ldots, t_n)^\intercal$ é o vetor de preditores; $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbb{I}_{\boldsymbol{n}})$ e $\mathbb{I}_{\boldsymbol{n}}$ é uma matriz identidade de ordem n. A função g representa uma curva de forma desconhecida, cuja estimativa pode ser obtida por diferentes métodos, como kernel smoothers, smoothing splines, loess, entre outros. Neste trabalho, adota-se a metodologia proposta por Eilers e Marx (1996), que combina B-splines (DE BOOR, 1978) com uma penalização aplicada aos coeficientes associados às bases. Tal técnica é amplamente empregada em contextos não paramétricos, destacando-se por sua eficiência e simplicidade de implementação.

2.1 B-SPLINES

As bases B-spline, utilizadas na construção da curva, consistem em segmentos de polinômios conectados por pontos denominados nós, de forma contínua, o que assegura transições suaves entre os trechos. As bases definidas a partir de polinômios de grau m são obtidas por meio de uma definição recursiva, baseada na divisão do intervalo $[a, b]$ da covariável t por k nós (ξ). A definição recursiva é dada por:

- B-Splines de grau *0*:

$$B_j^{(0)}(t) = \begin{cases} 1, \text{ se } \xi_j \leq t \leq \xi_{j+1} \\ 0, \text{ caso contrário} \end{cases}$$

- B-Splines de grau m:

$$B_j^{(m)}(t) = \frac{t - \xi_j}{\xi_{t+m} - \xi_j} B_j^{(m-1)}(t) + \frac{\xi_{j+m+1} - t}{\xi_{j+m+1} - \xi_{j+1}} B_{j+1}^{(m-1)}(t).$$

Diversas metodologias podem ser empregadas para o posicionamento dos nós. Neste trabalho, os nós serão posicionados de forma igualmente espaçada. Na prática, define-se o número de nós k, de modo que se obtenha um conjunto $q = k - m - 1$ bases. Uma curva suave pode então ser construída a partir de uma combinação linear de B-splines de grau 3, conforme a expressão:

$$g(t_i) = \sum_{j=1}^{q} \gamma_j B_j^{(3)}(t_i) = \mathbf{n}_i^\top \boldsymbol{\gamma},$$

em que $\mathbf{n_i} = \left[B_1^{(3)}(t_i), ..., B_q^{(3)}(t_i)\right]^\intercal$ é a i-ésima linha da matriz $\boldsymbol{N}$, proveniente da aplicação da covariável t_i nas q bases e, $\boldsymbol{\gamma} = \left[\gamma_1, ..., \gamma_q\right]^\intercal$ é o vetor de parâmetros a serem estimados. Dessa forma, ao considerar a matriz $\boldsymbol{N} = \left[\mathbf{n_1}^\intercal, ..., \mathbf{n_n}^\intercal\right]^\intercal$, de dimensão $n \times q$, tem-se:

$$g(\boldsymbol{t}) = \boldsymbol{N}\boldsymbol{\gamma}.$$

Ao substituir essa representação no modelo definido em (2.1), é possível obter os estimadores de máxima verossimilhança para $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\intercal, \sigma^2)^\intercal$ de forma direta, como para um regressão múltipla. No entanto, aplicar essa metodologia apenas com a escolha do número de nós, sem nenhuma restrição adicional, pode levar à ocorrência de sobreajuste (*overfitting*). Para contornar este desafio, a próxima seção aborda a aplicação de uma penalização na log-verossimilhança, com o objetivo de controlar a rugosidade da curva

2.2 CONTROLE DA RUGOSIDADE DA CURVA

Green e Silverman (1994) propõem, como medida da rugosidade (*roughness*) de uma curva $g(t)$, assumida como duas vezes continuamente diferenciável e definida no intervalo $[a, b]$, a integral do quadrado de sua segunda derivada:

$$\int_a^b \left[g''(t)\right]^2 dt. \tag{2.2}$$

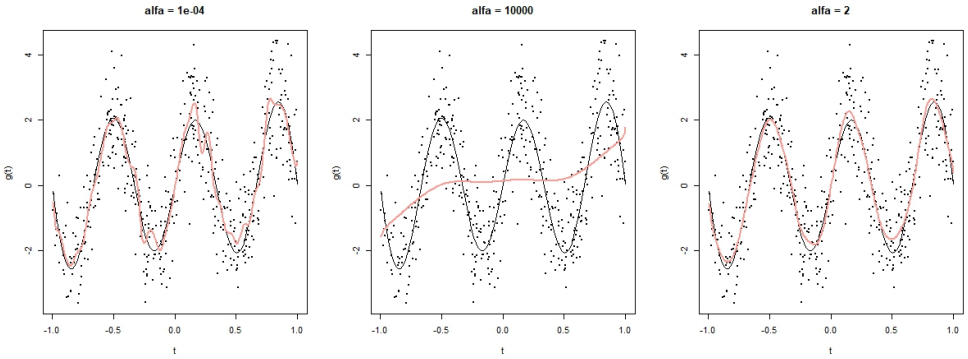
Se $g(t)$ for uma função suave, essa medida assume valores pequenos. Por outro lado, caso a curva apresente grande variação local, ou seja, seja muito rugosa, o valor da medida será elevado. Dessa forma, uma abordagem para estimar curvas bem comportadas, penalizando aquelas excessivamente irregulares, consiste em subtrair da log-verossimilhança ($\ell(\boldsymbol{\theta})$) um termo proporcional a essa medida:

$$\ell_p(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \frac{\alpha}{2}\int_a^b \left[g''(t)\right]^2 dt,$$

em que $\ell_p(\boldsymbol{\theta})$ representa a log-verossimilhança penalizada e α é um coeficiente de afinação (*tuning*). Quando $\alpha = 0$, não há penalização, permitindo que a curva interpole os pontos observados. Por outro lado, à medida que $\alpha \to \infty$, a função $g(t)$ tende a ser restringida a um comportamento linear. A Figura 1 ilustra estimativas da mesma curva sob diferentes níveis de penalização: insuficiente (painel à esquerda), excessiva (painel do meio) e ideal (painel à direita).

A proposta de Eilers e Marx (1996) consite em basear a penalização em diferenças finitas dos coeficientes adjacentes dos B-splines. Assim, o problema terá uma redução na dimensionalidade, de n (número de observações) para q (número de bases B-splines). A penalização baseada em diferenças de ordem 2 é uma boa aproximação discreta da penalização representada pela medida da equação (2.2). Assim, a log-verossimilhança penalizada pode ser reescrita como:

Figura 1 – Curvas estimadas sobre a verdadeira para diferentes coeficientes de suavização.



Fonte: Elaboração própria (2025).

$$\ell_p(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \frac{\alpha}{2}\boldsymbol{\gamma}^\intercal \boldsymbol{K}\boldsymbol{\gamma}, \tag{2.3}$$

em que $\boldsymbol{K} = \boldsymbol{D}^\intercal\boldsymbol{D}$ e $\boldsymbol{D}$ é a matriz de diferenças de ordem, definida por:

$$\boldsymbol{D} = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & -2 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 & 1 & -2 & 1 & 0 \\ \vdots & \vdots & \cdots & \ddots & 0 & 0 & 1 & -2 & 1 \end{bmatrix}.$$

O parâmetro de suavização α pode ser selecionado por meio de validação cruzada do tipo K-Fold (Hastie e Tibshirani, 2013) ou via algum critério de seleção, como o Critério de Informação Bayesiano ou BIC (Schawarz, 1978). Seguindo Ferreira et al. (2022) adotamos o BIC, definido como:

$$BIC(\alpha) = -2 \cdot \ell_p(\boldsymbol{\theta}) + log(n) \cdot p^*, \tag{2.4}$$

em que p^* é número de graus de liberdade efetivos do modelo.

O número de graus de liberdade efetivo da curva, denotado por df_N, é dado pelo traço da matriz $\boldsymbol{P}$, em que $\boldsymbol{P}$ é a matriz de projeção (ou matriz hat), tal que $\hat{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{y}$. O valor de df_N deve ser, no máximo, igual à dimensão das bases B-spline (q). Além disso, à medida que a penalização aumenta, os graus de liberdade efetivos da curva diminuem,

tendendo a 2 quando $\alpha \to \infty$. No contexto de regressão não paramétrica, temos que $\hat{g}(\boldsymbol{t}) = \boldsymbol{N}\hat{\boldsymbol{\gamma}}$, com $\hat{\boldsymbol{\gamma}} = (\boldsymbol{N}^\intercal\boldsymbol{N} + \alpha\ \hat{\sigma^2}\boldsymbol{K})^{-1}\boldsymbol{N}^\intercal\boldsymbol{Y}$. Dessa forma, a matriz de projeção é dada por: $\boldsymbol{P} = \boldsymbol{N}(\boldsymbol{N}^\intercal\boldsymbol{N} + \alpha\ \hat{\sigma^2}\boldsymbol{K})^{-1}\boldsymbol{N}^\intercal$ e o número de graus de liberdade efetivo da curva é:

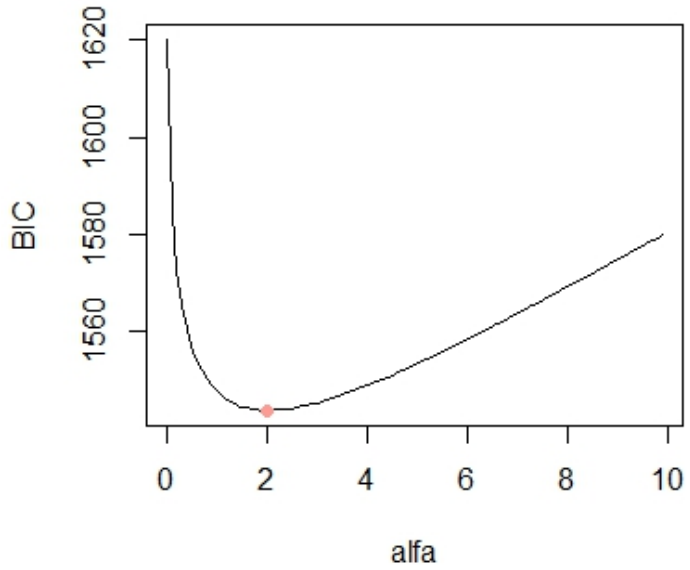
$$df_{\mathbf{N}} = tr\Big(\mathbf{N}(\mathbf{N}^\intercal\mathbf{N} + \alpha\ \hat{\sigma^2}\mathbf{K})^{-1}\mathbf{N}^\intercal\Big). \tag{2.5}$$

Com isso, tem-se $p^* = df_N + 1$.

Levando em consideração que, quanto menor o valor do BIC, melhor é o ajuste do modelo aos dados em questão, buscamos o valor de α que minimiza a expressão apresentada na Equação (2.4).

A Figura 2 apresenta o gráfico do BIC para um conjunto de valores de α, referentes à mesma curva ilustrada nos gráficos da Figura 1. Observa-se que, para essa curva específica, o valor $\alpha = 2$ resulta na penalização considerada ótima.

Figura 2 – Gráfico do BIC em função dos valores de α.



Fonte: Elaboração própria (2025).

3 MISTURAS DE MODELOS PARCIALMENTE LINEARES

Engle et al. (1986) introduziram o modelo parcialmente linear para modelar populações homogêneas, no qual a variável resposta $\mathbf{Y}$ é representada como uma combinação linear de covariáveis específicas $\mathbf{x} \in \mathbb{R}^m$ e uma função não paramétrica desconhecida de uma covariável adicional $t \in \mathbb{R}$, resultando na forma:

$$\mathbf{Y} = \mathbf{x}^\top \boldsymbol{\beta} + g(t) + \varepsilon,$$

em que ε é um termo de erro com média zero e variância finita, e $g(\cdot)$ é uma função não paramétrica que captura relações não lineares. Esse modelo combina a interpretabilidade da parte linear com a flexibilidade da parte não paramétrica para representar diferentes padrões nos dados. A separação entre as covariáveis $\mathbf{x}$ e t pode ser feita com base em conhecimento prévio ou por meio de análise exploratória, como gráficos de dispersão ou testes estatísticos.

A partir dessa formulação, e inspirados nos trabalhos de Hunter e Young (2012), Zhang e Pan (2020), Skhosana et al. (2023) e Hwang et al. (2024), estendemos o modelo parcialmente linear para o contexto de populações heterogêneas, por meio da estrutura de modelos de mistura. Nesse novo cenário, assumimos que cada subpopulação (ou componente da mistura) pode apresentar sua própria relação parcialmente linear, permitindo diferentes efeitos lineares e não lineares entre os grupos latentes. Alem disso, diferentemente dos trabalhos mencionados, nossa abordagem também contempla explicitamente a presença de um intercepto na parte paramétrica do modelo, além de adotar a representação da função não paramétrica $g(\cdot)$ por meio de bases B-splines. Essa escolha proporciona maior flexibilidade e controle na estimação das curvas suaves associadas a cada grupo, conforme a teoria apresentada no Capítulo 2.

3.1 MODELO PROPOSTO

Para cada observação i, introduzimos um vetor latente $\boldsymbol{Z_i} = (Z_{i1}, ..., Z_{iG})^\intercal$, com distribuição multinomial:

$$\boldsymbol{Z_i} = (Z_{i1}, ..., Z_{iG})^\intercal \sim Multinomial(1, p_1, p_2, ..., p_G),$$

em que $P(Z_{ij} = 1) = p_j$ e $\sum_{j=1}^{G} p_j = 1$. Isso significa que

$$Z_{ij} = \begin{cases} 1, \text{ se a i-ésima observação pertence ao grupo j} \\ 0, \text{ caso contrário.} \end{cases}$$

Condicionalmente a $Z_{ij} = 1$, a variável resposta Y_i segue um modelo semiparamétrico parcialmente linear da forma

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_j + g_j(t_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2),$$

em que $\mathbf{x_i} \in \mathbb{R}^m$ é o vetor de covariáveis lineares, desenhada para permitir a inclusão do intercepto; $\boldsymbol{\beta}_j \in \mathbb{R}^m$ é o vetor de coeficientes lineares para o grupo j; $t_i \in \mathbb{R}$ é a covariável com efeito não linear e $g_j(\cdot)$ é uma função não paramétrica específica do grupo j.

Para representar $g_j(t)$, utilizamos bases B-splines, de modo que o modelo pode ser reescrito como

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_j + \mathbf{n}_i^\top \boldsymbol{\gamma}_j + \varepsilon_{ij},$$

em que $\mathbf{n}_i \in \mathbb{R}^q$ é o vetor de avaliação das bases B-splines em t_i; $\boldsymbol{\gamma}_j \in \mathbb{R}^q$ é o vetor de coeficientes da parte não paramétrica para o grupo j e $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$.

A abordagem permite que as matrizes $\mathbf{X} \in \mathbb{R}^{n\times m}$, $\mathbf{N} \in \mathbb{R}^{n\times q}$ e a matriz de penalização $\mathbf{K} \in \mathbb{R}^{q\times q}$ sejam diferentes entre os grupos. De forma que $\mathbf{x}_i^\top$ representa, na verdade, $\mathbf{x}_{i,j}^\top$ a i-ésima linha da matriz $\mathbf{X}_j \in \mathbb{R}^{n\times m_j}$. O mesmo vale para a matriz $\mathbf{N}$, cuja versão por grupo é denotada por $\mathbf{N}_j \in \mathbb{R}^{n\times q_j}$.

Condicionalmente a $Z_{ij} = 1$, a variável resposta Y_i segue um modelo semiparamétrico parcialmente linear específico do grupo j, da forma:

$$Y_i = \mathbf{x}_{i,j}^\top \boldsymbol{\beta}_j + \mathbf{n}_{i,j}^\top \boldsymbol{\gamma}_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2),$$

em que:

- $\mathbf{x}_{i,j} \in \mathbb{R}^{m_j}$ é o vetor de covariáveis lineares da observação i no grupo j, correspondente à i-ésima linha da matriz $\mathbf{X}_j \in \mathbb{R}^{n\times m_j}$;
- $\boldsymbol{\beta}_j \in \mathbb{R}^{m_j}$ é o vetor de coeficientes lineares para o grupo j;
- $\mathbf{n}_{i,j} \in \mathbb{R}^{q_j}$ é o vetor de avaliação das bases B-splines em t_i para o grupo j, correspondente à i-ésima linha da matriz $\mathbf{N}_j \in \mathbb{R}^{n\times q_j}$;
- $\boldsymbol{\gamma}_j \in \mathbb{R}^{q_j}$ é o vetor de coeficientes da parte não paramétrica para o grupo j;
- σ_j^2 é a variância do erro associada ao grupo j.

Essa formulação permite que tanto o número quanto a natureza das covariáveis (lineares e não lineares) variem entre os grupos, conferindo maior flexibilidade ao modelo para capturar diferentes estruturas presentes nos dados.

3.2 REPRESENTAÇÃO HIERÁRQUICA

A representação hierárquica, por meio da variável latente $\mathbf{Z}_i$, desempenha um papel fundamental na aplicação do Algoritmo EM, pois permite reformular o problema de inferência como um caso de dados incompletos, em que $\mathbf{Z}_i$ é a parte não observada. Quando $Z_{ij} = 1$, temos que:

$$Y_i \mid Z_{ij} = 1 \sim \mathcal{N}(\mu_{ij}, \sigma_j^2), \quad \text{com} \quad \mu_{ij} = \mathbf{x}_{i,j}^\top \boldsymbol{\beta_j} + \mathbf{n}_{i,j}^\top \boldsymbol{\gamma_j}.$$

A densidade condicional é dada por:

$$f_{Y_i|\mathbf{Z}_i}\left(y_i \mid \mathbf{Z}_i = (z_{i1}, \ldots, z_{iG})\right) = \prod_{j=1}^{G} \phi(y_i; \mu_{ij}, \sigma_j^2)^{z_{ij}},$$

em que $\phi(\cdot; \mu, \sigma^2)$ é a densidade da normal univariada. A distribuição conjunta dos dados completos $(Y_i, \mathbf{Z}_i)$ é o produto da condicional com a distribuição multinomial de $\mathbf{Z}_i$.

3.3 LOG-VEROSSIMILHANÇA PENALIZADA

A densidade marginal de Y_i é a mistura:

$$f_{Y_i}(y_i \mid \boldsymbol{\theta}) = \sum_{j=1}^{G} p_j \, \phi(y_i; \mu_{ij}, \sigma_j^2),$$

em que $\boldsymbol{\theta}$ representa o vetor completo de parâmetros do modelo, definido como

$$\boldsymbol{\theta} = \left(p_1, \ldots, p_G, \boldsymbol{\beta_1}^\top, \ldots, \boldsymbol{\beta_G}^\top, \boldsymbol{\gamma_1}^\top, \ldots, \boldsymbol{\gamma_G}^\top, \sigma_1^2, \ldots, \sigma_G^2\right)^\top.$$

A log-verossimilhança penalizada dos dados observados $\mathbf{y} = (y_1, \ldots, y_n)^\top$ é:

$$\ell_p(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{G} p_j \, \phi(y_i; \mu_{ij}, \sigma_j^2)\right) - \frac{1}{2}\sum_{j=1}^{G} \alpha_j \boldsymbol{\gamma_j}^\top \boldsymbol{K_j} \boldsymbol{\gamma_j},$$

em que o segundo termo penaliza a complexidade da parte não linear, controlado pelos parâmetros α_j.

Os dados completos são definidos por $\mathbf{y}_c = (\mathbf{y}^\top, \mathbf{z}^\top)^\top$, em que $\mathbf{z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_n^\top)^\top$. A log-verossimilhança penalizada completa é:

$$\ell_{cp}(\boldsymbol{\theta} \mid \mathbf{y}_c) = c + \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij} \log(p_j) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij} \log(\sigma_j^2) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \frac{z_{ij}}{\sigma_j^2}(y_i - \mu_{ij})^2 - \frac{1}{2}\sum_{j=1}^{G} \alpha_j \boldsymbol{\gamma_j}^\top \mathbf{K_j} \boldsymbol{\gamma_j},$$

em que c é uma constante que não depende de $\boldsymbol{\theta}$.

3.4 ALGORITMO EM

No decorrer desta seção, apresentaremos um algoritmo do tipo EM, desenvolvido com o objetivo de obter os estimadores de máxima verossimilhança dos parâmetros do modelo proposto. O algoritmo EM foi originalmente introduzido por Dempster et al. (1977) e consiste em duas etapas principais: a etapa de esperança (passo E) e a etapa de maximização (passo M). Esse tipo de algoritmo é amplamente utilizado no contexto

de dados incompletos. Como discutido anteriormente, estamos justamente nesse cenário, pois o modelo envolve uma variável latente que indica a qual grupo cada observação pertence. Portanto, a presença dessa variável não observável justifica o uso do algoritmo EM. Com base na representação hierárquica apresentada na Seção 3.2, podemos construir o procedimento de estimação conforme descrito a seguir.

3.4.1 Etapa E

Nesta etapa do algoritmo EM, calculamos o valor esperado da função de log-verossimilhança penalizada dos dados completos, condicionada aos dados observados e às estimativas atuais dos parâmetros. Isso corresponde à função Q, definida como:

$$E\Big[\ell_{cp}(\boldsymbol{\theta}|\boldsymbol{y_c})|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(k)}\Big] = Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k)}) = c + \sum_{i=1}^{n}\sum_{j=1}^{G} \hat{z}_{ij}^{(k)} \cdot \log(p_j) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \hat{z}_{ij}^{(k)} \cdot \log(\sigma_j^2) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \frac{\hat{z}_{ij}^{(k)}}{\sigma_j^2} \cdot (y_i - \mu_{ij})^2 - \frac{1}{2}\sum_{j=1}^{G} \alpha_j \boldsymbol{\gamma}_j^{\intercal} \boldsymbol{K_j} \boldsymbol{\gamma}_j.$$

em que c é uma constante irrelevante para a maximização, pois não depende dos parâmetros de interesse. Agrupando os termos e reescrevendo em notação matricial, temos:

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k)}) = c &+ \sum_{i=1}^{n}\sum_{j=1}^{G} \hat{z}_{ij}^{(k)} \cdot \log(p_j) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \hat{z}_{ij}^{(k)} \cdot \log(\sigma_j^2) \\ &- \frac{1}{2}\sum_{j=1}^{G} \frac{1}{\sigma_j^2} \cdot (\boldsymbol{Y} - \mathbf{X_j}\boldsymbol{\beta}_j - \mathbf{N_j}\boldsymbol{\gamma}_j \ )^{\intercal} \hat{\mathbf{H}}_j^{(k)} (\boldsymbol{Y} - \mathbf{X_j}\boldsymbol{\beta}_j - \mathbf{N_j}\boldsymbol{\gamma}_j \ ) \\ &- \frac{1}{2}\sum_{j=1}^{G} \alpha_j \boldsymbol{\gamma}_j^{\intercal} \mathbf{K_j} \boldsymbol{\gamma}_j, \end{aligned} \tag{3.1}$$

em que $\hat{\mathbf{H}}_j^{(k)}$ é uma matriz diagonal contendo os valores $\hat{z}_{1j}^{(k)}$, ... , $\hat{z}_{nj}^{(k)}$.

Dessa forma, para que a função Q esteja completamente especificada, é necessário calcular o valor esperado da variável latente Z_i, dado o valor observado y_i e as estimativas atuais dos parâmetros, denotadas por $\hat{\boldsymbol{\theta}}^{(k)}$. O vetor de expectativas condicionais é dado por:

$$\begin{aligned} E\Big[\boldsymbol{Z}_i|y_i, \hat{\boldsymbol{\theta}}^{(k)}\Big] &= \Big[E(Z_{i1}|y_i, \hat{\boldsymbol{\theta}}^{(k)}), \ \dots \ , E(Z_{iG}|y_i, \hat{\boldsymbol{\theta}}^{(k)})\Big] \\ &= [\hat{z_{i1}}^{(k)}, \ \dots \ , \hat{z_{iG}}^{(k)}], \end{aligned}$$

em que $\hat{z}_{ij}^{(k)}$ representa a probabilidade estimada de que a i-ésima observação pertença ao grupo j, na k-ésima iteração. Utilizando a distribuição condicional $\boldsymbol{Z_i} \mid Y_i = y_i$ e aplicando as propriedades da esperança condicional, obtemos:

$$\hat{z}_{ij}^{(k)} = \frac{\hat{p}_j^{(k)} \phi(y_i, \hat{\mu_{ij}}^{(k)}, \hat{\sigma_j^2}^{(k)})}{\sum_{g=1}^{G} \hat{p}_g^{(k)} \phi(y_i, \hat{\mu_{ig}}^{(k)}, \hat{\sigma_g^2}^{(k)})}. \tag{3.2}$$

3.4.2 Etapa M

A seguir, são apresentadas as expressões utilizadas para a obtenção das estimativas dos parâmetros na iteração $(k+1)$, de modo a maximizar a função Q definida na equação (3.1):

$$\hat{\boldsymbol{p}}_j^{(k+1)} = \sum_{i=1}^{n} \frac{\hat{z}_{ij}^{(k)}}{n};$$

$$\hat{\boldsymbol{\beta}}_j^{(k+1)} = \left(\mathbf{X_j}^{\intercal}\hat{\mathbf{H}}_j^{(k)}\mathbf{X_j}\right)^{-1}\mathbf{X_j}^{\intercal}\hat{\mathbf{H}}_j^{(k)}\left(\mathbf{Y} - \mathbf{N_j}\hat{\boldsymbol{\gamma}}_j^{(k)}\right);$$

$$\hat{\boldsymbol{\gamma}}_j^{(k+1)} = \left(\mathbf{N_j}^{\intercal}\hat{\mathbf{H}}_j^{(k)}\mathbf{N_j} + \alpha_j\hat{\sigma_j^2}^{(k)}\mathbf{K_j}\right)^{-1}\mathbf{N_j}^{\intercal}\hat{\mathbf{H}}_j^{(k)}\left(\mathbf{Y} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_j^{(k)}\right);$$

$$\hat{\sigma^2}_j^{(k+1)} = \frac{\left(\mathbf{Y} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_j^{(k)} - \mathbf{N_j}\hat{\boldsymbol{\gamma}}_j^{(k)}\right)^{\intercal}\hat{\mathbf{H}}_j^{(k)}\left(\mathbf{Y} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_j^{(k)} - \mathbf{N_j}\hat{\boldsymbol{\gamma}}_j^{(k)}\right)}{\sum_{i=1}^{n}\hat{z}_{ij}^{(k)}}.$$

Um dos principais desafios no ajuste de modelos complexos é encontrar um equilíbrio entre a incorporação de informações adicionais e o aumento do número de parâmetros. No caso específico de modelos de mistura, esse desafio torna-se ainda mais evidente, uma vez que a inclusão de novas componentes implica um crescimento significativo na quantidade de parâmetros a serem estimados. Para lidar com esse problema, diversas estratégias têm sido propostas com o objetivo de aumentar a parcimônia dos modelos, mesmo diante da complexidade crescente. Neste trabalho, exploramos uma alternativa que permite reduzir o número de parâmetros ao assumir que todos os grupos possuem a mesma variabilidade. Nessa configuração, em vez de estimar uma variância distinta σ_j^2 para cada grupo, estimamos uma única variância comum σ^2. Consequentemente, a atualização desse parâmetro na Etapa M do algoritmo assume a seguinte forma:

$$\hat{\sigma^2}^{(k+1)} = \frac{\sum_{j=1}^{G}\left(\mathbf{Y} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_j^{(k)} - \mathbf{N_j}\hat{\boldsymbol{\gamma}}_j^{(k)}\right)^{\top}\hat{\mathbf{H}}_j^{(k)}\left(\mathbf{Y} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_j^{(k)} - \mathbf{N_j}\hat{\boldsymbol{\gamma}}_j^{(k)}\right)}{n},$$

em que $\hat{\mathbf{H}}_j^{(k)}$ é a matriz de ponderação atualizada com base nas probabilidades a posteriori $\hat{z}_{ij}^{(k)}$. Essa simplificação reduz o número de parâmetros relacionados à variância de G para apenas 1, contribuindo significativamente para a parcimônia do modelo. A avaliação do modelo que melhor se ajusta aos dados, seja com restrição ou não na variância, pode ser realizada por meio do critério BIC, conforme descrito na Seção 3.6. Outra alternativa seria aplicar um teste assintótico, cuja hipótese nula é a igualdade das variabilidades dos grupos, ou seja, $H_0 : \sigma_1^2 = \cdots = \sigma_G^2$.

3.4.3 Classificação ou Agrupamento das Observações

A etapa de atribuição das observações aos grupos pode ser interpretada como *classificação* ou *agrupamento*, a depender da existência (ou não) de conhecimento prévio sobre os rótulos dos grupos:

- Quando os **rótulos verdadeiros são conhecidos** (como em estudos de simulação), a tarefa se assemelha a uma *classificação supervisionada*, ainda que o modelo tenha sido ajustado de forma não supervisionada.

- Quando os **rótulos são desconhecidos** (como em aplicações reais), trata-se de um *agrupamento* (*clustering*), típico de contextos de *aprendizado não supervisionado*.

Após a estimação dos parâmetros do modelo, as observações são associadas aos grupos com base nos valores esperados das variáveis latentes Z_i, ou seja, nas probabilidades $\hat{z}_{ij}^{(k)}$ obtidas na Etapa E (Seção **3.4.1**). A regra de decisão adotada consiste em alocar cada observação i ao grupo j cuja probabilidade a posteriori estimada seja máxima:

$$j^* = \arg\max_j \hat{z}_{ij}^{(k)}. \tag{3.3}$$

Dessa forma, cada observação é atribuída ao grupo cuja componente do modelo apresenta maior verossimilhança condicional com os dados observados, resultando em uma partição final do conjunto de dados.

3.4.4 Valores Iniciais

Valores iniciais são necessários para implementar o algoritmo proposto. Devido à presença de múltiplos máximos locais na função de log-verossimilhança em modelos de misturas finitas, é recomendável executar o algoritmo EM a partir de uma variedade de valores iniciais, a fim de verificar se a estimativa de máxima verossimilhança global foi alcançada. Nesse contexto, propomos uma abordagem simples para a obtenção de valores iniciais, conforme descrito a seguir:

- Os dados são particionados em G grupos pelo algoritmo de classificação k-Means (Hartigan e Wong, 1979);

- Os rótulos dos grupos são redefinidos em ordem decrescente de média da variável resposta isto é, o rótulo 1 para a maior média, o rótulo 2 para a segunda maior, e assim por diante;

- Calcula-se a proporção de observações pertencentes ao grupo j como valor inicial de $p_j^{(0)}$, $j = 1, ..., G$;

- Para cada grupo, com base no agrupamento inicial, temos:

  - Estimativa inicial dos coeficientes do componente paramétrico (mínimos quadrados):

    $\hat{\boldsymbol{\beta}}_{\mathbf{j}}^{(0)} = (\mathbf{X_j}^\intercal \mathbf{H}_j \mathbf{X_j})^{-1} \mathbf{X_j}^\intercal \mathbf{H}_j \mathbf{Y}$, onde $\mathbf{H}_j$ é uma matriz diagonal composta por zeros e uns, indicando se a observação pertence ao grupo j;

  - Estimativa inicial da variância do erro:
    $\hat{\sigma_j^{2(0)}} = \frac{(\mathbf{Y} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\mathbf{0})})^{\intercal}\mathbf{H}_j(\mathbf{Y} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\mathbf{0})})}{n_j}$ sendo nj o número de observações atribuídas ao grupo j.
  - Estimativa inicial dos coeficientes do componente não paramétrico (modelo semiparamétrico):
    $\hat{\boldsymbol{\gamma}}_{\boldsymbol{j}}^{(0)} = (\mathbf{N_j}^{\intercal}\mathbf{H}_j\mathbf{N_j} + \alpha_j\hat{\sigma_j^{2(0)}}\mathbf{K_j})\mathbf{N_j}^{\intercal}\mathbf{H}_j(\mathbf{Y_j} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\mathbf{0})})$.

- A seguir, atualizamos a estimativa de $\hat{\sigma_j^{2(0)}}$ considerando o novo resíduo e, em seguida, recalculamos $\hat{\boldsymbol{\gamma}}_{\boldsymbol{j}}^{(0)}$ com a nova variância:

  - Atualização da variância do erro:
    $\hat{\sigma_j^{2(0)}} = \frac{(\mathbf{Y} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\mathbf{0})} - \mathbf{N_j}\hat{\boldsymbol{\gamma}}_{\boldsymbol{j}}^{(0)})^{\intercal}\mathbf{H}_j(\mathbf{Y} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\mathbf{0})} - \mathbf{N_j}\hat{\boldsymbol{\gamma}}_{\boldsymbol{j}}^{(0)})}{n_j}$.
  - Reestimativa dos coeficientes do componente não paramétrico com a nova variância:
    $\hat{\boldsymbol{\gamma}}_{\boldsymbol{j}}^{(0)} = (\mathbf{N_j}^{\intercal}\mathbf{H}_j\mathbf{N_j} + \alpha_j\hat{\sigma_j^{2(0)}}\mathbf{K_j})\mathbf{N_j}^{\intercal}\mathbf{H}_j(\mathbf{Y} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\mathbf{0})})$.

Observação: esses dois últimos passos são opcionais, uma vez que os parâmetros serão atualizados normalmente na próxima iteração do algoritmo. Importante destacar que a eficácia dessa estratégia foi analisada por meio dos estudos de simulação discutidos no Capítulo 4.

Ao assumir que todos os grupos possuem a mesma variabilidade, temos que

$$\hat{\sigma^{2(0)}} = \frac{\sum_{j=1}^{G}\left(\mathbf{Y} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_j^{(0)}\right)^{\top}\mathbf{H}_j\left(\mathbf{Y} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_j^{(0)}\right)}{n},$$

em que $\mathbf{H}_j$ é a matriz diagonal de pesos associada ao grupo j, e $\hat{\boldsymbol{\beta}}_j^{(0)}$ é a estimativa inicial dos coeficientes para aquele grupo.

3.4.5 Critério de Parada

Assim como a escolha dos valores iniciais, algoritmos iterativos também requerem um critério de parada. Ou seja, as iterações são repetidas até que uma regra de convergência adequada seja satisfeita. No nosso algoritmo, adotamos o seguinte critério: as etapas E e M serão repetidas enquanto for satisfeita a seguinte condição

$$\left|\frac{\ell_p(\hat{\boldsymbol{\theta}}^{(k+1)})}{\ell_p(\hat{\boldsymbol{\theta}}^{(k)})} - 1\right| \geq 10^{-5}. \tag{3.4}$$

Ou seja, o algoritmo continuará sendo executado até que a diferença relativa entre os valores da função log-verossimilhança em duas iterações consecutivas seja menor que 10^{-5}.

Quando isso ocorrer, consideramos que houve convergência e as estimativas dos parâmetros foram obtidas. Além disso, estabelecemos um limite máximo de 5000 iterações, de modo que, caso o algoritmo não convirja antes disso, a execução será interrompida.

A implementação do algoritmo foi realizada na linguagem R (R CORE TEAM, 2024). Os códigos utilizados encontram-se disponíveis e podem ser fornecidos mediante solicitação por e-mail para camila.zeller@ufjf.br ou jaquelinelamasdasilva@gmail.com.

3.5 MATRIZ DE INFORMAÇÃO EMPÍRICA

Nesta seção, descrevemos o procedimento para obtenção dos erros padrão associados às estimativas de máxima verossimilhança penalizada do modelo proposto. Com base no resultado assintótico para estimadores de máxima verossimilhança (EMV), temos que:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, [MI_F(\boldsymbol{\theta})]^{-1}).$$

Inspirados no trabalho de Ferreira et. al. (2022) que tratam da estimação dos erros padrão dos estimadores obtidos via máxima verossimilhança penalizada com o uso da matriz de informação empírica, adotamos como aproximação da matriz de informação de Fisher avaliada em $\widehat{\boldsymbol{\theta}}$ a matriz de informação empírica, utilizando o mesmo procedimento descrito por Lin (2010). Dessa forma, os erros padrão associados às estimativas de máxima verossimilhança penalizada do modelo proposto foram obtidos com base nessa aproximação, conforme detalhado a seguir. A partir da log-verossimilhança penalizada completa correspondente à i-ésima observação, definida por:

$$\begin{aligned}\ell_{ci}(\boldsymbol{\theta}) = c + \sum_{j=1}^{G} z_{ij} \cdot log(p_j) - \frac{1}{2}\sum_{j=1}^{G} z_{ij} \cdot log(\sigma_j^2) \\ - \frac{1}{2}\sum_{j=1}^{G} \frac{z_{ij}}{\sigma_j^2} \cdot (y_i - \mathbf{x}_{i,j}^{\intercal}\boldsymbol{\beta}_j - \mathbf{n}_{i,j}^{\intercal}\boldsymbol{\gamma}_j)^2 - \frac{1}{2n}\sum_{j=1}^{G} \alpha_j \boldsymbol{\gamma}_j^{\intercal}\mathbf{K_j}\boldsymbol{\gamma}_j,\end{aligned}$$

calculamos a função escore e, em seguida, obtemos seu valor esperado condicional aos dados observados e aos parâmetros estimados:

$$\hat{\boldsymbol{s}}_i = E\left[\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\middle| y_i, \hat{\boldsymbol{\theta}}\right].$$

Dessa forma, temos:

$$\hat{\boldsymbol{s}}_i^{\intercal} = (\hat{s}_{i,p_1}, ..., \hat{s}_{i,p_{G-1}}, \hat{\boldsymbol{s}}_{\boldsymbol{i},\boldsymbol{\beta_1}}{}^{\intercal}, ..., \hat{\boldsymbol{s}}_{\boldsymbol{i},\boldsymbol{\beta_G}}{}^{\intercal}, \hat{\boldsymbol{s}}_{\boldsymbol{i},\boldsymbol{\gamma_1}}{}^{\intercal}, ..., \hat{\boldsymbol{s}}_{\boldsymbol{i},\boldsymbol{\gamma_G}}{}^{\intercal}, \hat{s}_{i,\sigma_1^2}, ..., \hat{s}_{i,\sigma_G^2}), \text{ em que}$$

$$\hat{s}_{i,p_j} = E\left[\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \boldsymbol{p_j}}\middle| y_i, \hat{\boldsymbol{\theta}}\right] = \frac{\hat{z_{ij}}}{\hat{p}_j} - \frac{\hat{z_{ig}}}{\hat{p}_g};$$

$$\hat{\boldsymbol{s}}_{\boldsymbol{i},\boldsymbol{\beta_j}} = E\left[\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta_j}}\middle| y_i, \hat{\boldsymbol{\theta}}\right] = \frac{\hat{z_{ij}}}{\hat{\sigma_j^2}}(y_i - \mathbf{x}_{i,j}^\intercal \hat{\boldsymbol{\beta}}_j - \mathbf{n}_{i,j}^\intercal \hat{\boldsymbol{\gamma}}_j)\mathbf{x}_{i,j};$$

$$\hat{\boldsymbol{s}}_{\boldsymbol{i},\boldsymbol{\gamma_j}} = E\left[\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma_j}}\middle| y_i, \hat{\boldsymbol{\theta}}\right] = \frac{\hat{z_{ij}}}{\hat{\sigma_j^2}}(y_i - \mathbf{x}_{i,j}^\intercal \hat{\boldsymbol{\beta}}_j - \mathbf{n}_{i,j}^\intercal \hat{\boldsymbol{\gamma}}_j)\mathbf{n}_{i,j} - \frac{\alpha_j}{n}\mathbf{K_j}\hat{\boldsymbol{\gamma}}_j;$$

$$\hat{s}_{i,\sigma_j^2} = E\left[\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \sigma_j^2}\middle| y_i, \hat{\boldsymbol{\theta}}\right] = -\frac{\hat{z_{ij}}}{2\hat{\sigma_j^2}} + \frac{\hat{z_{ij}}}{2(\hat{\sigma_j^2})^2}(y_i - \mathbf{x}_{i,j}^\intercal \hat{\boldsymbol{\beta}}_j - \mathbf{n}_{i,j}^\intercal \hat{\boldsymbol{\gamma}}_j)^2.$$

Com base nessas quantidades, a matriz de informação empírica é definida como:

$$MI_{emp}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \hat{\boldsymbol{s}}_i \hat{\boldsymbol{s}}_i^\intercal.$$

Essa matriz constitui uma aproximação da matriz de informação de Fisher $MI_F(\boldsymbol{\theta})$. Portanto, os erros padrão foram obtidos a partir da seguinte expressão:

$$\text{sd.emp}(\hat{\boldsymbol{\theta}}) = \sqrt{\text{diag}\left([MI_{emp}(\hat{\boldsymbol{\theta}})]^{-1}\right)}.$$

3.6 CRITÉRIO DE INFORMAÇÃO BAYESIANO (BIC)

No Capítulo 2, foi apresentada a definição do Critério de Informação Bayesiano (BIC), conforme a Equação (2.4). Assim como em Zeller et al. (2018), o critério será adotado para selecionar o número de componentes G. No contexto de modelos de misturas parcialmente lineares, o cálculo do BIC exige a log-verossimilhança dos dados observados penalizada, além dos graus de liberdade efetivos p^*.

Os graus de liberdade efetivos associados às curvas podem ser obtidos a partir do traço da matriz de projeção, ou seja, $df_j = \text{tr}(\mathbf{P}_j)$. A seguir, descreve-se o procedimento utilizado para a obtenção da matriz P_j:

1. Considere o estimador:

$$\hat{\boldsymbol{\gamma}}_j = \left(\mathbf{N}_j^\top \mathbf{H}_j \mathbf{N}_j + \alpha_j \hat{\sigma}_j^2 \mathbf{K}_j\right)^{-1} \mathbf{N}_j^\top \mathbf{H}_j \left(\mathbf{Y} - \mathbf{X}_j \hat{\boldsymbol{\beta}}_j\right);$$

2. Substitua $\hat{\boldsymbol{\beta}}_j$ pela expressão:

$$\hat{\boldsymbol{\beta}}_j = \left(\mathbf{X}_j^\top \mathbf{H}_j \mathbf{X}_j\right)^{-1} \mathbf{X}_j^\top \mathbf{H}_j \left(\mathbf{Y} - \mathbf{N}_j \hat{\boldsymbol{\gamma}}_j\right),$$

   na equação de $\hat{\boldsymbol{\gamma}}_j$;

3. Após manipulações algébricas, obtém-se a matriz P_j tal que:

$$\boldsymbol{N}_j \hat{\boldsymbol{\gamma}}_j = \boldsymbol{P}_j \boldsymbol{Y}.$$

Com base nisso, o número total de graus de liberdade efetivos do modelo é dado por:

$$p^* = (G-1) + \sum_{j=1}^{G} m_j + \sum_{j=1}^{G} df_j + s,$$

em que G representa o número de componentes (ou grupos) da mistura; m_j é o número de parâmetros lineares do componente j; df_j é o grau de liberdade efetivo associado à curva do componente j e s é o número de parâmetros de escala σ_j^2 estimados.

No *modelo parcimonioso*, assume-se uma variância comum entre os grupos ($\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_G^2$), implicando que apenas um parâmetro de escala é estimado, ou seja, $s = 1$.

Nas aplicações em dados reais, considera-se um conjunto de valores para o número de componentes G. Para cada valor considerado, busca-se os valores ótimos dos parâmetros de suavização α_j, de forma a minimizar o BIC. Essa otimização é realizada por meio da função `optim`, do software R (R CORE TEAM, 2024), utilizando o método de Byrd et al. (1995). O número ótimo de grupos é, então, aquele que resulta no menor valor do BIC entre os modelos avaliados.

3.7 RESÍDUOS QUANTÍLICOS ALEATORIZADOS (*Randomized Quantile Residuals*)

De acordo com Dunn e Smyth (1996), seja $F(y|\mu, \phi)$ a função de distribuição acumulada associada à densidade $f(y|\mu, \phi)$. Quando F é contínua, é possível definir os *resíduos quantílicos aleatorizados* da seguinte forma:

$$r_{q,i} = \Phi^{-1}\left(F(y_i|\hat{\mu}, \hat{\phi})\right),$$

em que y_i é a observação da variável Y_i, e $\Phi(\cdot)$ representa a função de distribuição acumulada da normal padrão. Sob estimação consistente dos parâmetros, os resíduos $r_{q,i}$ convergem em distribuição para uma normal padrão $\mathcal{N}(0, 1)$.

Conforme discutido no Capítulo 3, dado que $Z_{ij} = 1$, assume-se que:

$$Y_i \sim \mathcal{N}(\mu_{ij}, \sigma_j^2),$$

de acordo com a classificação fornecida pelo algoritmo EM. Assim, o resíduo quantílico para a observação i pode ser obtido por:

$$r_{q,i} = \Phi^{-1}\left(\Phi(y_i|\hat{\mu}_{ij}, \hat{\sigma}_j^2)\right).$$

Os resíduos quantílicos aleatorizados podem ser empregados na construção de envelopes simulados para fins de checagem da adequação do modelo. Essa abordagem fornece uma ferramenta diagnóstica útil para identificar possíveis desvios sistemáticos, outliers ou falhas no ajuste do modelo.

3.8 ENVELOPE SIMULADO

A seguir, apresentamos os detalhes do procedimento para a construção do *gráfico de envelope simulado*, baseado nos resíduos quantílicos aleatorizados, aplicado ao modelo proposto.

3.8.1 Procedimento de Construção do Envelope

1. **Ajuste do modelo**: Estime os parâmetros $\hat{\boldsymbol{\theta}}$ do modelo e calcule os resíduos quantílicos $r_{q,i}$, para $i = 1, \ldots, n$.

2. **Bootstrap paramétrico** (com B replicações, $b = 1, \ldots, B$):

   a) Para cada replicação, gere uma nova amostra de respostas $\boldsymbol{Y}^{(b)}$ a partir da distribuição do modelo ajustado:

   - Gere $\boldsymbol{Z}_i \sim \text{Multinomial}(1; \hat{p}_1, \ldots, \hat{p}_G)$;
   - Condicional a $Z_{ij} = 1$, gere $Y_i^{(b)} \sim \mathcal{N}(\hat{\mu}_{ij}, \hat{\sigma}_j^2)$, onde $\hat{\mu}_{ij} = \mathbf{x}_{i,j}^\top \hat{\boldsymbol{\beta}}_j + \mathbf{n}_{i,j}^\top \hat{\boldsymbol{\gamma}}_j$.

   b) Com os novos dados $\boldsymbol{Y}^{(b)}$, reestime os parâmetros $\hat{\boldsymbol{\theta}}^{(b)}$ e calcule os novos resíduos quantílicos $r_{q,i}^{(b)}$.

   c) Ordene os resíduos $\mathbf{r}^{(b)} = (r_{q,1}^{(b)}, r_{q,2}^{(b)}, \ldots, r_{q,n}^{(b)})$ em ordem crescente.

   d) Armazene esse vetor ordenado como a coluna b de uma matriz $\mathbf{A}$, de dimensão $n \times B$.

3. **Construção dos envelopes** após as B replicações:

   - **Faixa inferior**: para cada linha da matriz $\mathbf{A}$, calcule o quantil inferior $\alpha/2$;
   - **Faixa superior**: calcule o quantil superior $1 - \alpha/2$;
   - **Linha média**: calcule a média de cada linha de $\mathbf{A}$.

O gráfico final do envelope simulado é construído comparando os resíduos observados $r_{i,q}$ com as faixas simuladas. Desvios sistemáticos em relação ao envelope podem indicar inadequações no modelo ajustado. Importante destacar que esta proposta constitui uma contribuição original deste trabalho para a avaliação do ajuste de modelos no contexto de misturas, oferecendo uma ferramenta diagnóstica robusta para a checagem da adequação do modelo.

4 ESTUDOS DE SIMULAÇÃO

Este capítulo apresenta diversos estudos de simulação de Monte Carlo com o objetivo de examinar as propriedades assintóticas dos estimadores de máxima verossimilhança sob diferentes cenários e avaliar o desempenho de agrupamento. O primeiro estudo investiga como a acurácia e a variabilidade das estimativas são influenciadas por diferentes tamanhos amostrais ($n = 100, 300, 500, 1000, 2000$). O segundo experimento também avalia o desempenho do modelo proposto em termos de classificação.

4.1 ASPECTOS COMPUTACIONAIS

Modelos aditivos requerem a imposição de restrições sobre as curvas para resolver problemas de não-identificabilidade do modelo, uma vez que combinações infinitas entre o intercepto e deslocamentos nas funções podem resultar no mesmo valor predito. Durante simulações iniciais, observou-se que as estimativas da curva em um dos grupos apresentavam viés sistemático. Diante disso, optou-se por utilizar, na implementação das matrizes $\boldsymbol{N}$ e $\boldsymbol{K}$, a função *smoothCon*, do pacote *mgcv*. Essa função possui uma opção para embutir uma restrição de identificabilidade nas bases, a qual impõe que:

$$\sum_{i}^{n} g(t_i) = 0. \tag{4.1}$$

Essa restrição já foi aplicada em trabalhos como Hunter e Young (2012) e não altera o formato da função, apenas realiza um deslocamento vertical para que a curva resultante tenha média zero. Para mais detalhes, consultar Wood (2017). O valor padrão dessa função define uma base com dimensão $q = 10$, o qual foi mantido nas simulações. Para as aplicações com dados reais, apresentadas no próximo capítulo, adotamos valores mais adequados conforme o comportamento observado em cada caso, ou seja, pela inspeção visual dos gráficos, verificamos que as curvas não apresentavam alta complexidade e por isso, o número de nós pôde ser reduzido, o que também implicou na diminuição da dimensão das matrizes $\boldsymbol{N}$ e $\boldsymbol{K}$.

4.2 SIMULAÇÃO 1

No decorrer desta seção, discutiremos os resultados dos estudos de simulação realizados com o objetivo de avaliar o algoritmo de estimação proposto. Com foco na capacidade de recuperação dos parâmetros, analisamos cinco cenários de mistura entre dois modelos parcialmente lineares. Um resumo desses cenários é apresentado na Tabela 1. Os três primeiros cenários correspondem, respectivamente, a grupos bem, pouco e mal separados, adotando as mesmas covariáveis para ambos os grupos. No quarto cenário, mantemos uma separação moderada entre os grupos, porém com covariáveis lineares

distintas. Por fim, no quinto cenário, tanto as covariáveis lineares quanto as não lineares diferem entre os grupos.

Tabela 1 – Resumo dos cenários considerados nos estudos de simulação.

Cenário	**Separação**	**Covariáveis lineares**	**Covariáveis não lineares**
C1	Alta	Iguais	Iguais
C2	Moderada	Iguais	Iguais
C3	Baixa	Iguais	Iguais
C4	Moderada	Diferentes	Iguais
C5	Moderada	Diferentes	Diferentes

Adicionalmente, investigamos as propriedades assintóticas dos estimadores de máxima verossimilhança, considerando 500 réplicas para cada combinação de cenário e tamanho amostral. A consistência dos estimadores dos parâmetros dos modelos, sob os diversos cenários estudados, será avaliada por meio de tabelas que apresentam as médias, os desvios padrão e os erros padrão calculados a partir da matriz de informação empírica, para cada um dos dois grupos, com base nas 500 estimativas obtidas pelo algoritmo EM para cada tamanho amostral considerado. Além disso, boxplots dessas estimativas serão utilizados como recurso visual complementar.

A performance do método para a estimação da função $g(t)$ é avaliada por meio do erro quadrático médio, conhecido como ASE (*average squared error*). Para complementar essa avaliação quantitativa, utilizamos também recursos visuais, como gráficos que comparam as curvas estimadas com as curvas verdadeiras. O ASE mede o quão distante as curvas estimadas estão da curva verdadeira. Para a $k-$ésima curva, essa medida é calculada pela fórmula:

$$\text{ASE}_k = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{g}^{(k)}(t_i) - g(t_i)\right]^2, \quad k = 1, \ldots, 500. \tag{4.2}$$

Os parâmetros de suavização (α_j), embora usualmente selecionados por critérios de otimização em aplicações com dados reais, foram fixados em valores pré-definidos nas simulações, a fim de reduzir o custo computacional. Além disso, para garantir que as curvas satisfaçam a restrição (4.1), elas foram especificadas de modo que sua integral seja zero no intervalo em que cada covariável correspondente é gerada.

Durante a rotina de geração das amostras, verifica-se a capacidade do algoritmo em estimar os parâmetros, bem como a invertibilidade da matriz de informação. Amostras que resultam em falhas no processo de estimação, isto é, quando o algoritmo interrompe por atingir o número máximo de iterações sem convergir segundo o critério de parada estabelecido ou que apresentam matriz de informação singular são descartadas, sendo geradas novas amostras até que essas condições sejam satisfeitas.

Para definir os cenários, considere o seguinte modelo geral:

$$\begin{cases} Y_i = \mathbf{x}_{i,1}^\top \boldsymbol{\beta}_1 + g_1(t_{i1}) + \varepsilon_1, & \text{se } Z_{i1} = 1, \\ Y_i = \mathbf{x}_{i,2}^\top \boldsymbol{\beta}_2 + g_2(t_{i2}) + \varepsilon_2, & \text{se } Z_{i2} = 1, \end{cases}$$

em que Z_{ij} é o componente indicador de Y_i, com

$$P(Z_{ij} = 1) = p_j, \quad j = 1, 2, \quad i = 1, \ldots, n;$$

$$\varepsilon_1 \sim \mathcal{N}(0, \sigma_1^2) \quad \text{e} \quad \varepsilon_2 \sim \mathcal{N}(0, \sigma_2^2);$$

$\mathbf{x}_{i,j}^\top = (1, x_{i1,j}, \ldots, x_{i(m_j-1),j})$ representa a i-ésima linha da matriz $\mathbf{X}$ do grupo j, correspondente às covariáveis lineares; t_{ij} é a covariável que apresenta relação não linear com a resposta; e $\boldsymbol{\beta}_j$ é um vetor de dimensão m_j. Dessa forma, com o modelo definido, os cenários variam de acordo com os parâmetros, as curvas e as distribuições das covariáveis. As seções seguintes apresentam cada cenário e seus respectivos resultados.

4.2.1 Cenário 1 (C1)

Neste cenário de simulação, os dados foram gerados de forma a garantir uma separação clara entre dois grupos distintos. A proporção de indivíduos em cada grupo foi definida como $p_1 = 0.35$ para o Grupo 1, o que implica que os 65% restantes pertencem ao Grupo 2. Essa escolha já introduz uma diferença inicial na composição dos dados.

Os vetores de coeficientes associados às covariáveis também foram definidos de maneira a reforçar essa separação. Para o Grupo 1, o vetor de coeficientes é $\boldsymbol{\beta}_1 = (8,\ 4,\ 6)^\top$, enquanto para o Grupo 2 temos $\boldsymbol{\beta}_2 = (-2,\ -3,\ -5)^\top$. Esses vetores não apenas apresentam valores distintos, mas também possuem sinais opostos, indicando que os efeitos das covariáveis sobre a variável resposta ocorrem em direções contrárias entre os dois grupos. Essa diferença acentuada nos coeficientes dos modelos reforça a distinção entre os grupos em termos de estrutura de regressão.

Além das componentes lineares, cada grupo também possui uma função específica do tempo, incorporando efeitos não lineares na modelagem. Para o Grupo 1, a função do tempo é $g_1(t) = \cos(\pi t)$, enquanto para o Grupo 2 ela é $g_2(t) = -2\cos(\pi t)$. Observa-se aqui não apenas uma inversão de sinal, mas também uma diferença de amplitude, com o Grupo 2 apresentando uma oscilação mais intensa. Essa característica faz com que, ao longo do tempo, as respostas dos dois grupos sigam trajetórias bastante distintas, mesmo quando submetidas à mesma covariável.

A variabilidade das observações dentro de cada grupo também foi diferenciada. O Grupo 1 foi definido com $\sigma_1^2 = 0.1$, ao passo que o Grupo 2 apresenta uma variância maior, $\sigma_2^2 = 0.2$. Essa diferença implica que, além de responderem de maneira diferente às covariáveis e ao tempo, os indivíduos do Grupo 2 apresentam maior dispersão em suas respostas. Ambos os grupos compartilham o mesmo valor para um parâmetro adicional α,

com $\alpha_1 = \alpha_2 = 0.1$, o que sugere que esse parâmetro de regularização não está envolvido diretamente na separação dos grupos.

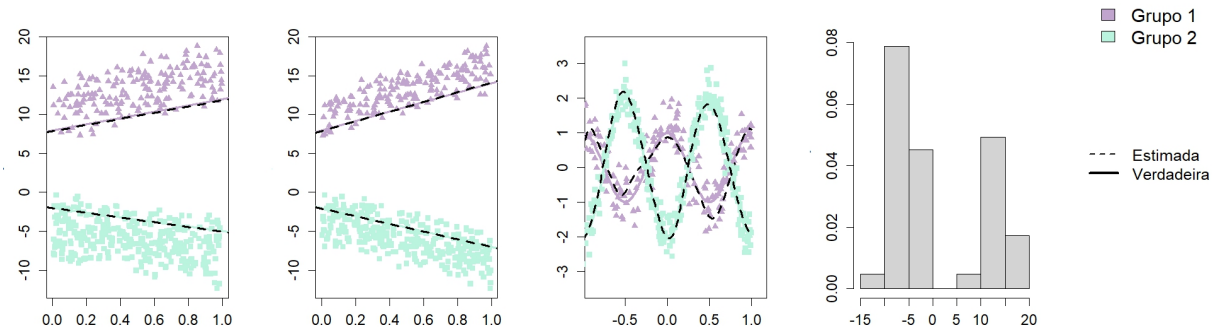
As covariáveis foram geradas de maneira totalmente aleatória e independente. Para cada indivíduo $i \in \{1, \ldots, n\}$, e para cada $j = 1, 2$, as covariáveis $x_{im,j}$, com $m = 1, 2$, foram amostradas a partir de uma distribuição uniforme no intervalo $(0, 1)$, ou seja, $x_{im,j} \sim \text{Uniforme}(0, 1)$. Os tempos de observação t_{ij} foram também gerados aleatoriamente, com $t_{ij} \sim \text{Uniforme}(-1, 1)$, garantindo variação dentro de um intervalo simétrico em torno de zero.

Essa combinação de diferenças marcantes entre os grupos, em proporção, efeitos das covariáveis, função do tempo e variância, foi projetada para garantir que os grupos fossem completamente separáveis, facilitando a identificação de padrões distintos e o desenvolvimento de métodos de classificação ou agrupamento com alto desempenho, por exemplo, como veremos no próximo estudo de simulação (Seção 4.3). A Figura 3 apresenta os gráficos de dispersão da variável resposta y em função das covariáveis x_1 e x_2 e dos resíduos não paramétricos em função de t, com as respectivas retas e curvas ajustadas, para uma amostra de tamanho $n = 500$. Dado que $Z_{ij} = 1$, o resíduo não paramétrico é definido como a diferença entre a resposta observada e a parte explicada pelas componentes lineares, ou seja:

$$\text{res.np}_i = Y_i - \mathbf{x}_{i,j}^\top \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}.$$

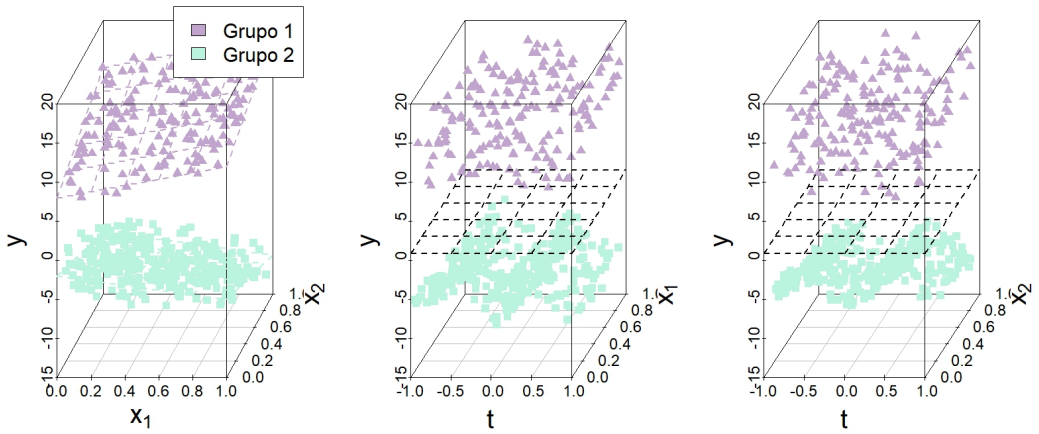
Os pontos apresentados nos gráficos das curvas estimadas correspondem a esses resíduos não paramétricos, calculados individualmente para cada observação. Eles representam, portanto, a parte da resposta que não é explicada pelas covariáveis lineares, e que é modelada pela função não paramétrica associada ao tempo. Adicionalmente, a Figura 4 apresenta os gráficos de dispersão em 3D da variável resposta y, em função das covariáveis, para uma amostra de tamanho $n = 500$. Observa-se que os grupos estão bem separados, refletindo a diferença nas estruturas de geração dos dados.

Figura 3 – Gráficos de dispersão da variável resposta y em função das covariáveis (primeiro painel à esquerda: x_1, segundo painel: x_2) e dos resíduos não paramétricos em função de t (terceiro painel), com as respectivas retas e curvas ajustadas, acompanhados de um histograma da densidade de y (quarto painel) para uma amostra de tamanho $n = 500$.



Fonte: Elaboração própria (2025).

Figura 4 – Gráficos de dispersão em 3D da variável resposta y, para uma amostra de tamanho $n = 500$. No painel à esquerda, as covariáveis x_1 e x_2 formam o plano da base; no painel central, t e x_1; e, no painel à direita, t e x_2.



Fonte: Elaboração própria (2025).

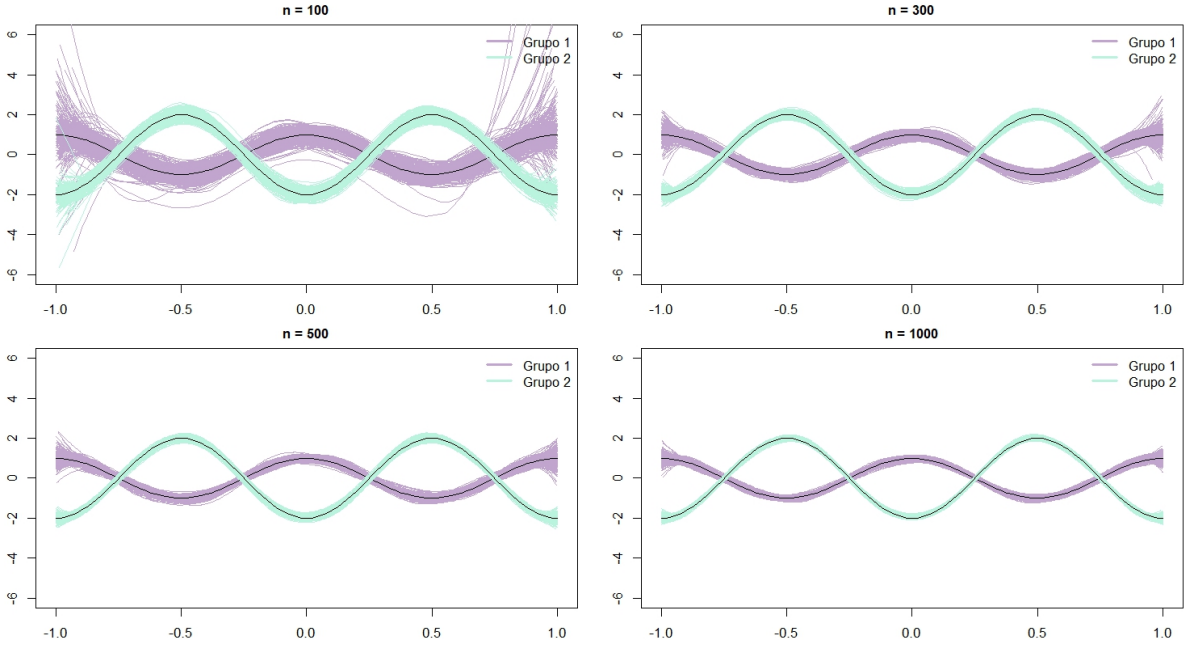
Na Tabela 2, observa-se que o aumento no tamanho da amostra conduz a estimativas mais próximas dos valores reais dos parâmetros. Além disso, o método de estimação do erro padrão com base na matriz de informação empírica produz valores relativamente próximos aos desvios padrão observados (SD), o que indica consistência na estimação dos erros padrão sob o modelo proposto. Na Figura 5, nota-se que as curvas estimadas se aproximam da curva verdadeira à medida que o tamanho da amostra aumenta, o que também pode ser verificado na Figura 6. Seguindo Ferreira et al. (2022), no gráfico do ASE representamos a

Tabela 2 – Resultados da simulação para o Cenário 1: valor verdadeiro do parâmetro, seguido da média, do desvio padrão (*sd*) e do erro padrão calculado pela matriz de informação empírica (*sd*.emp) das 500 estimativas obtidas pelo algoritmo EM, para cada tamanho de amostra considerado.

		n=100			n=300			n=500			n=1000			n=2000		
	θ	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp
p_1	0.35	0.349	0.048	0.048	0.350	0.028	0.027	0.349	0.022	0.021	0.350	0.015	0.015	0.350	0.010	0.011
β_{10}	8.00	8.004	0.268	2.145	7.981	0.134	0.131	8.000	0.099	0.097	7.997	0.068	0.066	7.997	0.046	0.046
β_{11}	4.00	4.000	0.310	0.367	4.018	0.167	0.169	3.999	0.121	0.127	4.002	0.082	0.086	4.002	0.062	0.060
β_{12}	6.00	6.001	0.345	0.371	6.014	0.165	0.168	5.998	0.124	0.126	6.002	0.082	0.086	6.003	0.056	0.060
β_{20}	-2.00	-1.996	0.179	0.146	-2.004	0.099	0.064	-1.996	0.081	0.048	-2.000	0.054	0.034	-2.001	0.039	0.024
β_{21}	-3.00	-3.002	0.147	0.164	-2.998	0.080	0.083	-3.001	0.063	0.063	-2.997	0.043	0.044	-3.000	0.029	0.031
β_{22}	-5.00	-4.999	0.154	0.163	-5.001	0.079	0.083	-5.003	0.061	0.063	-5.001	0.045	0.044	-4.997	0.032	0.031
σ_1^2	0.20	0.135	0.040	0.048	0.177	0.027	0.029	0.187	0.022	0.022	0.194	0.016	0.016	0.196	0.011	0.011
σ_2^2	0.10	0.083	0.017	0.018	0.094	9.45×10^{-03}	0.010	0.097	7.87×10^{-03}	8.02×10^{-03}	0.099	5.37×10^{-03}	5.68×10^{-03}	0.100	3.94×10^{-03}	3.98×10^{-03}

Fonte: Elaboração própria (2025).

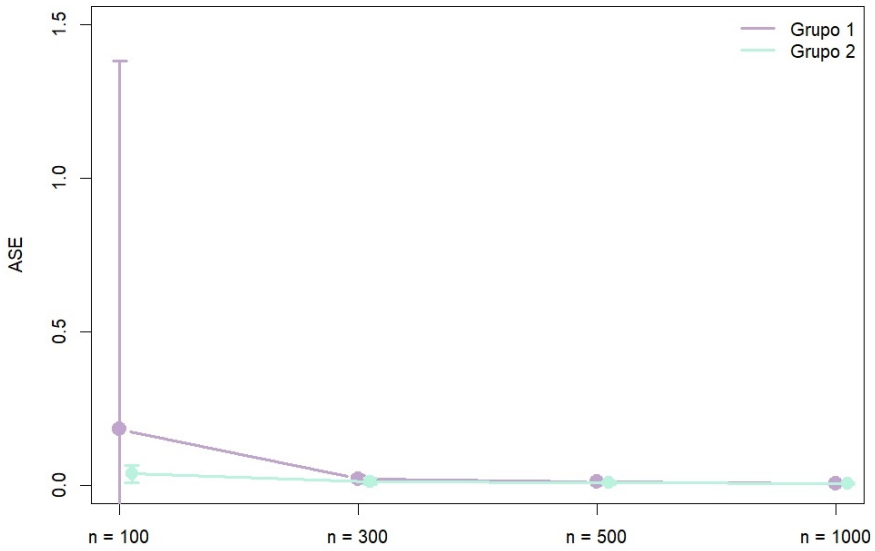
Figura 5 – Gráficos das componentes não paramétricas com 500 réplicas. Curvas ajustadas (linhas representadas nas cores roxa e azul) e curvas verdadeiras (linhas pretas) sob o Cenário 1, para diferentes tamanhos de amostra.



Fonte: Elaboração própria (2025).

média ± 1 desvio padrão dessas medidas, definidas na equação (4.2), para cada tamanho amostral. O Grupo 1 (representado na cor roxa) apresenta menor representatividade na amostra e maior variância, o que está de acordo com as estimativas menos precisas obtidas para esse grupo. As Figuras 7 e 8 apresentam os boxplots das estimativas dos parâmetros do modelo sob o Cenário 1. À medida que o tamanho da amostra aumenta, os boxplots tornam-se mais concentrados em torno dos valores verdadeiros dos parâmetros,

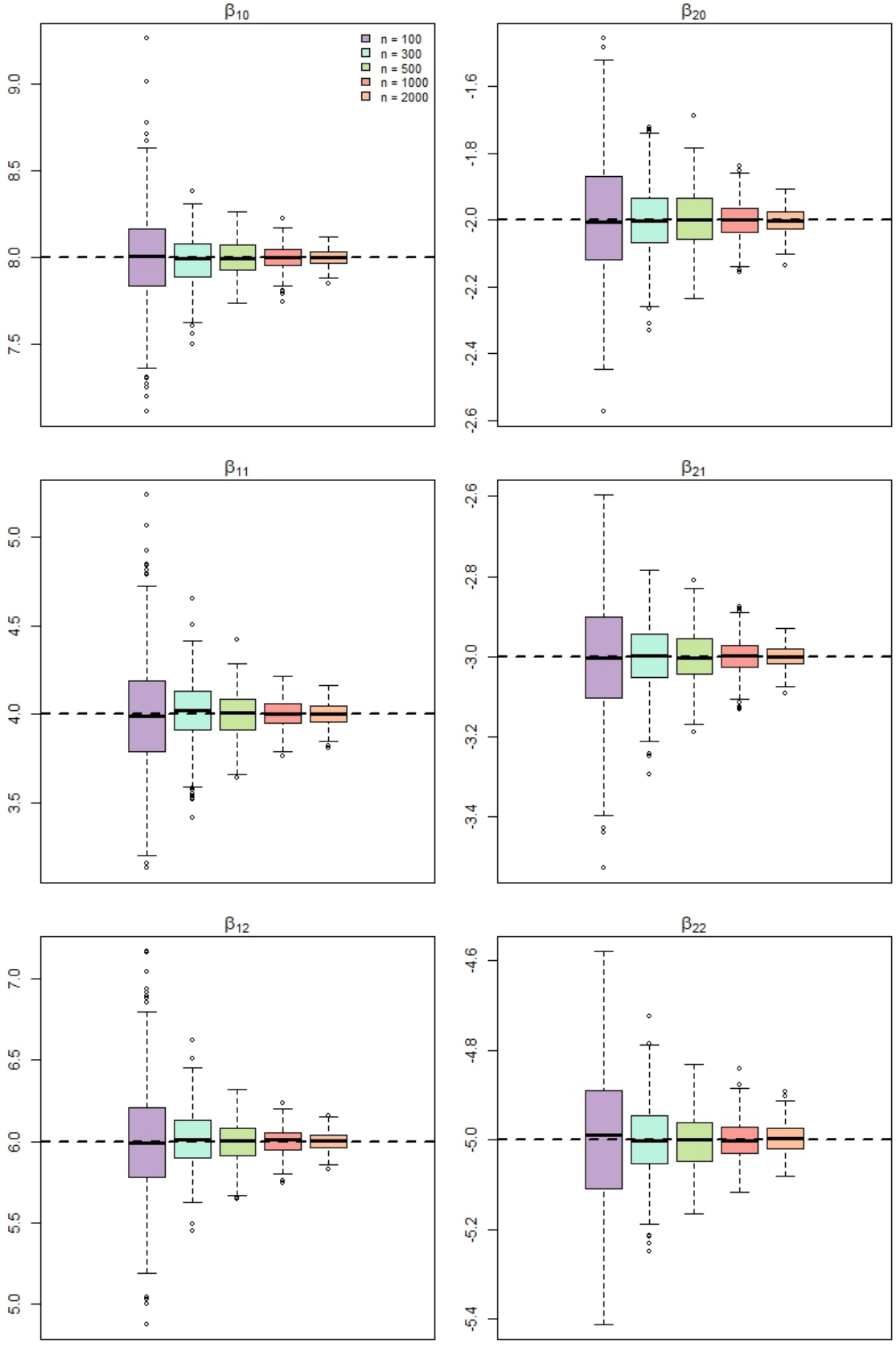
Figura 6 – Média ± um desvio padrão dos ASEs das curvas, sob o Cenário 1, para cada tamanho de amostra.



Fonte: Elaboração própria (2025).

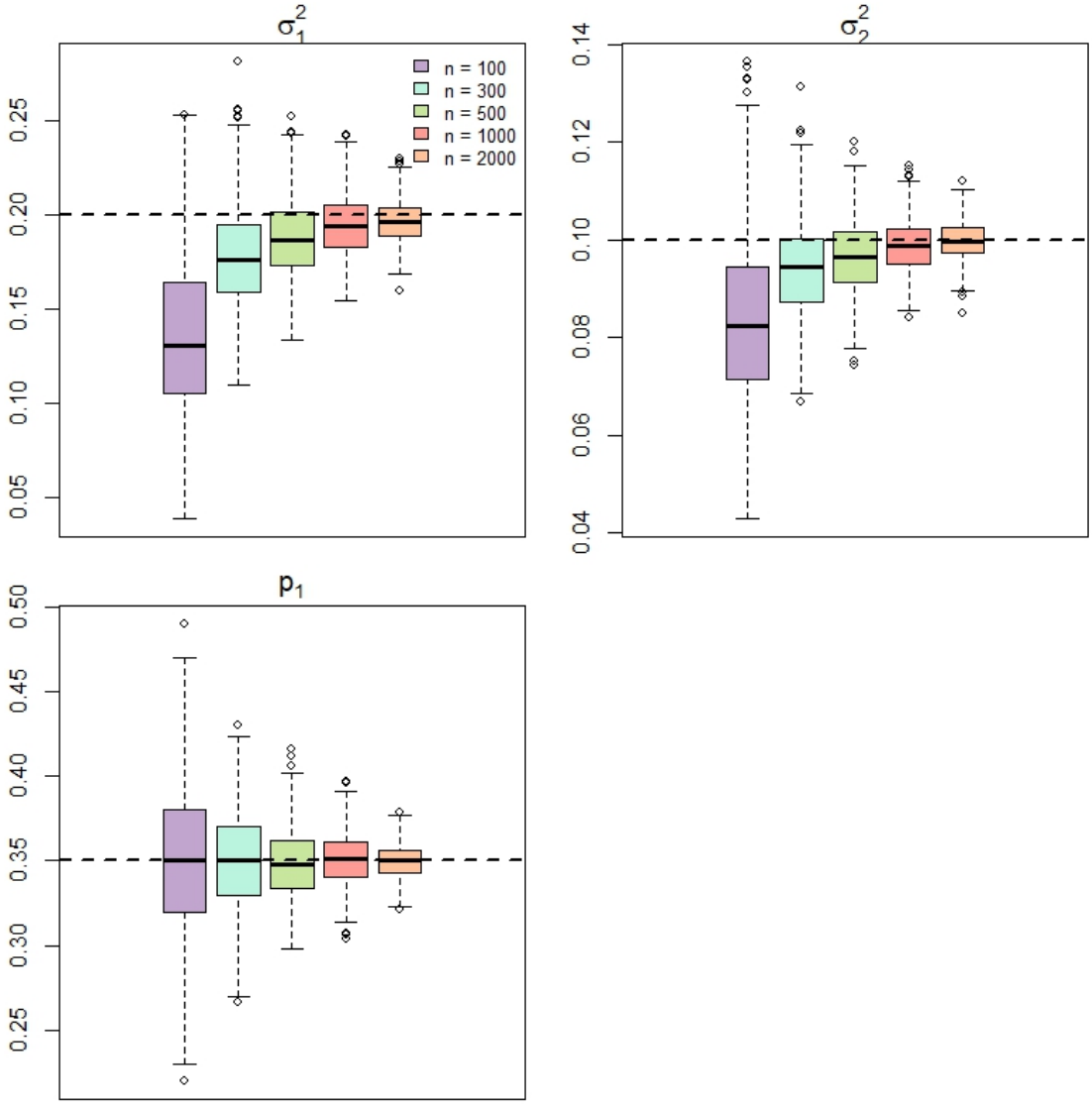
indicando uma melhora na precisão das estimativas. De forma geral, tanto o viés quanto a variabilidade das estimativas dos parâmetros diminuem com o aumento do tamanho amostral, o que está de acordo com as propriedades assintóticas bem estabelecidas do método de máxima verossimilhança.

Figura 7 – Boxplots das estimativas de β_{10}, β_{20}, β_{11}, β_{21}, β_{12}, β_{22} para o modelo proposto sob o Cenário 1. A linha tracejada indica o valor verdadeiro do parâmetro.



Fonte: Elaboração própria (2025).

Figura 8 – Boxplots das estimativas de σ_1^2, σ_2^2 e p_1 para o modelo proposto sob o Cenário 1. A linha tracejada indica o valor verdadeiro do parâmetro.



Fonte: Elaboração própria (2025).

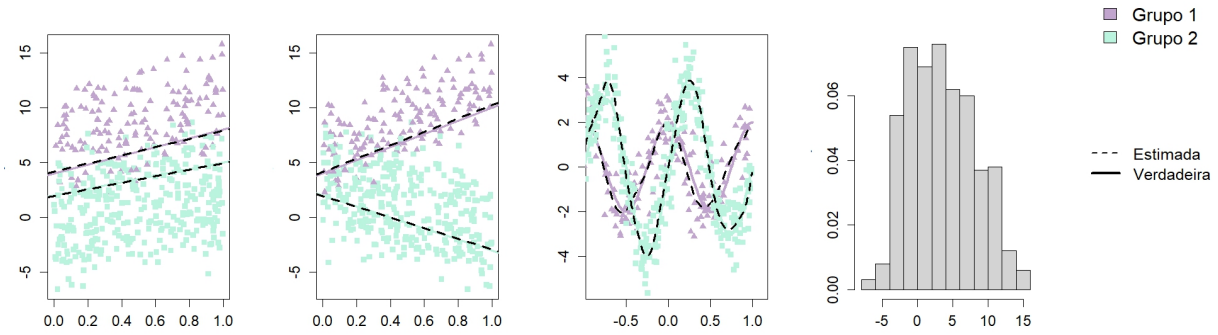
4.2.2 Cenário 2 (C2)

Neste cenário, os parâmetros foram definidos de modo a garantir uma separação moderada entre os grupos. Os valores adotados foram: $p_1 = 0.35$; os vetores de coeficientes $\boldsymbol{\beta}_1 = (4, 4, 6)$ e $\boldsymbol{\beta}_2 = (2, 3, -5)$; variâncias iguais para ambos os grupos, $\sigma_1^2 = \sigma_2^2 = 1$; funções não lineares definidas como $g_1(t) = 2\cos(\pi t)$ e $g_2(t) = 4\sin(\pi t)\exp(-0.5t^2)$; além dos parâmetros de suavização α_1 e α_2, ambos fixados em 0.1. As covariáveis foram geradas de acordo com distribuições uniformes: $x_{im,j} \sim \text{Uniforme}(0, 1)$ e $t_{ij} \sim \text{Uniforme}(-1, 1)$, para $m = 1, 2$, $j = 1, 2$ e $i = 1, \ldots, n$.

A parcimônia do modelo refere-se à busca por uma formulação suficientemente simples, capaz de evitar complexidades desnecessárias sem comprometer a capacidade de representar adequadamente a estrutura dos dados. Nesse contexto, também foi considerada a aplicação de uma versão mais parcimoniosa do algoritmo, que estima uma única variância comum a todos os grupos (*4.2.2.1*). Tal simplificação pode ser vantajosa quando há evidências ou suposições prévias de que os grupos apresentam variabilidades semelhantes, pois permite a redução do número de parâmetros estimados e do custo computacional, mantendo a qualidade do ajuste.

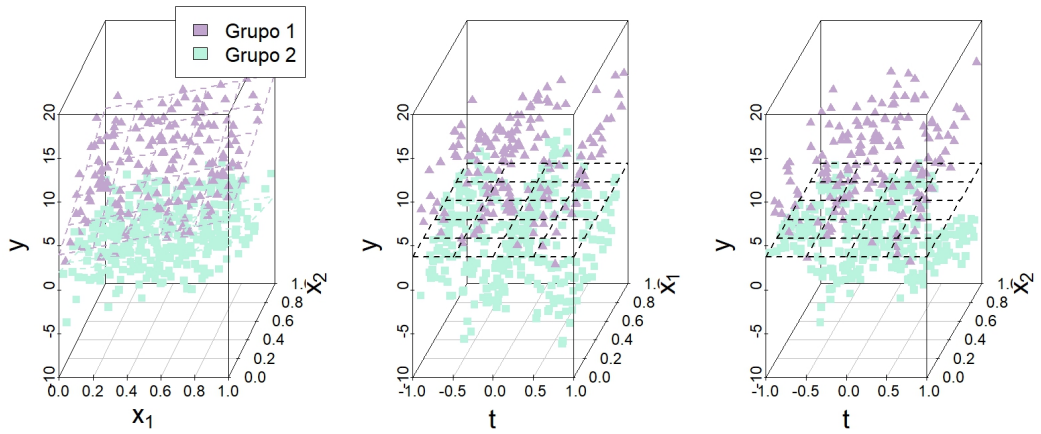
As Figuras 9 e 10 indicam uma leve sobreposição entre os grupos, evidenciando as distinções nas estruturas de geração dos dados adotadas. Os resultados apresentados na Tabela 3 e nas Figuras 11 a 14 indicam que, em comparação com o Cenário 1, a presença de uma leve sobreposição entre os grupos exige um tamanho amostral maior para se alcançar o mesmo nível de precisão nas estimativas. Ainda assim, nesse cenário, as propriedades assintóticas, em particular, a consistência dos estimadores, foram verificadas, reforçando a robustez do método mesmo sob condições moderadas de separação entre os grupos.

Figura 9 – Gráficos de dispersão da variável resposta y em função das covariáveis (primeiro painel à esquerda: x_1, segundo painel: x_2) e dos resíduos não paramétricos em função de t (terceiro painel), com as respectivas retas e curvas ajustadas, acompanhados de um histograma da densidade de y (quarto painel) para uma amostra de tamanho $n = 500$.



Fonte: Elaboração própria (2025).

Figura 10 – Gráficos de dispersão em 3D da variável resposta y, para uma amostra de tamanho $n = 500$. No painel à esquerda, as covariáveis x_1 e x_2 formam o plano da base; no painel central, t e x_1; e, no painel à direita, t e x_2.



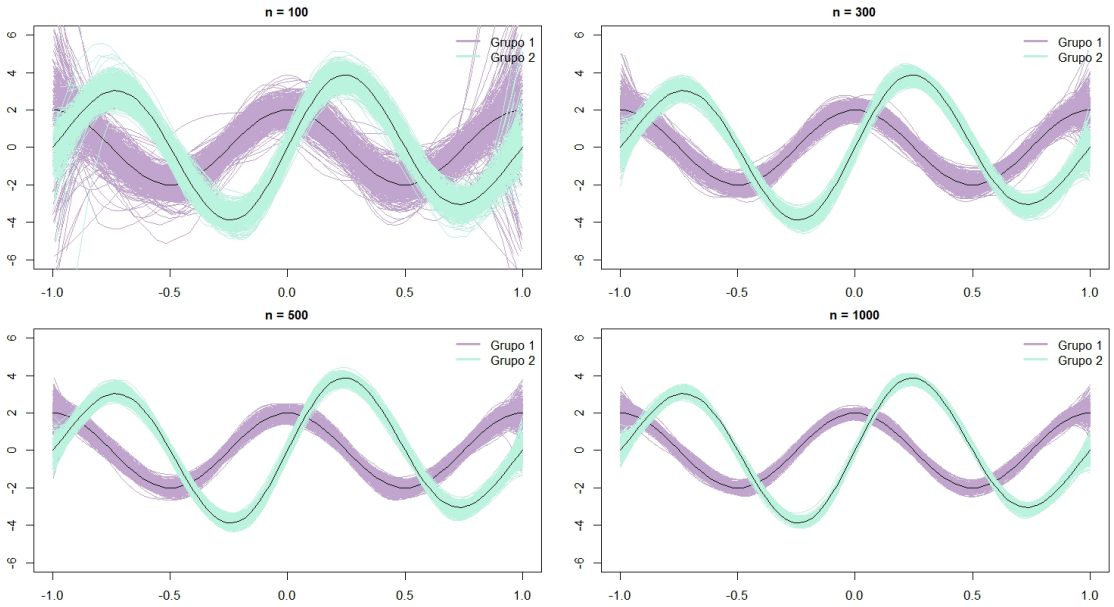
Fonte: Elaboração própria (2025).

Tabela 3 – Resultados da simulação para o Cenário 2: valor verdadeiro do parâmetro, seguido da média, do desvio padrão (*sd*) e do erro padrão calculado pela matriz de informação empírica (*sd*.emp) das 500 estimativas obtidas pelo algoritmo EM, para cada tamanho de amostra considerado.

		n=100			n=300			n=500			n=1000			n=2000		
	θ	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp
p_1	0.35	0.352	0.050	0.053	0.350	0.030	0.029	0.349	0.022	0.023	0.350	0.016	0.016	0.350	0.011	0.011
β_{10}	4.00	4.003	0.653	1.342	4.001	0.320	0.319	4.005	0.243	0.236	3.999	0.171	0.161	4.008	0.117	0.113
β_{11}	4.00	3.981	0.839	0.830	3.999	0.379	0.393	4.000	0.275	0.294	4.003	0.192	0.201	3.994	0.140	0.140
β_{12}	6.00	5.948	0.769	0.851	5.975	0.404	0.405	5.995	0.300	0.304	5.990	0.203	0.209	5.990	0.140	0.146
β_{20}	2.00	1.968	0.467	0.466	2.005	0.246	0.213	1.992	0.207	0.161	1.985	0.137	0.112	2.007	0.098	0.079
β_{21}	3.00	2.972	0.510	0.517	2.977	0.265	0.271	2.998	0.201	0.205	3.011	0.143	0.143	2.988	0.092	0.101
β_{22}	-5.00	-4.973	0.512	0.532	-4.991	0.263	0.277	-4.993	0.213	0.209	-4.989	0.146	0.146	-4.999	0.101	0.102
σ_1^2	1.00	0.671	0.246	0.256	0.870	0.146	0.152	0.919	0.116	0.117	0.964	0.082	0.083	0.985	0.057	0.058
σ_2^2	1.00	0.799	0.168	0.185	0.939	0.107	0.109	0.959	0.084	0.084	0.983	0.059	0.059	0.998	0.041	0.042

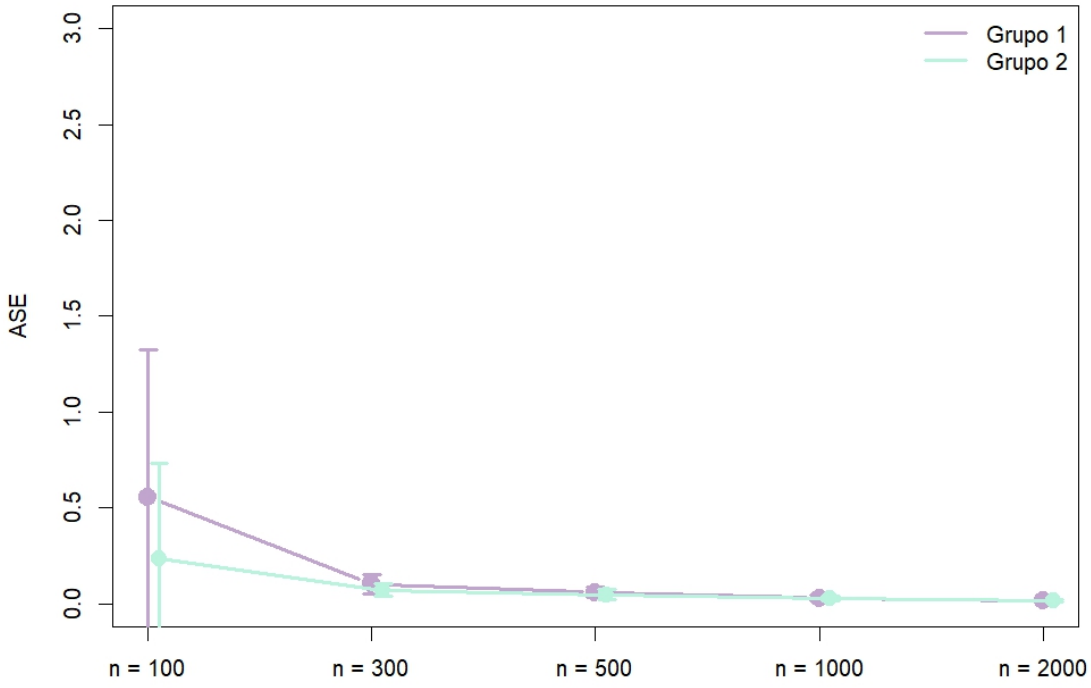
Fonte: Elaboração própria (2025).

Figura 11 – Gráficos das componentes não paramétricas com 500 réplicas. Curvas ajustadas (linhas representadas nas cores roxa e azul) e curvas verdadeiras (linhas pretas) sob o Cenário 2, para diferentes tamanhos de amostra.



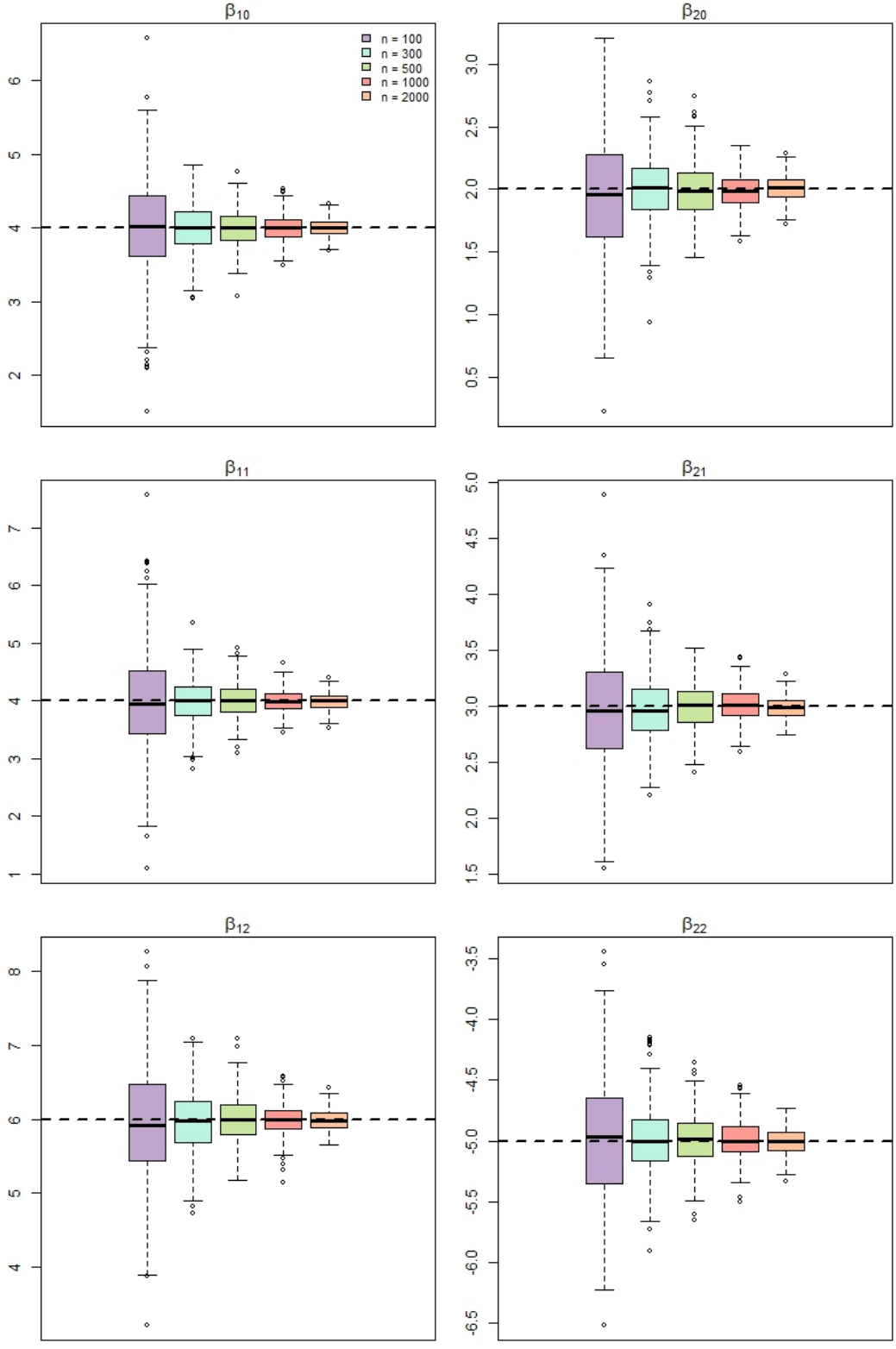
Fonte: Elaboração própria (2025).

Figura 12 – Média ± um desvio padrão dos ASEs das curvas, sob o Cenário 2, para cada tamanho de amostra.



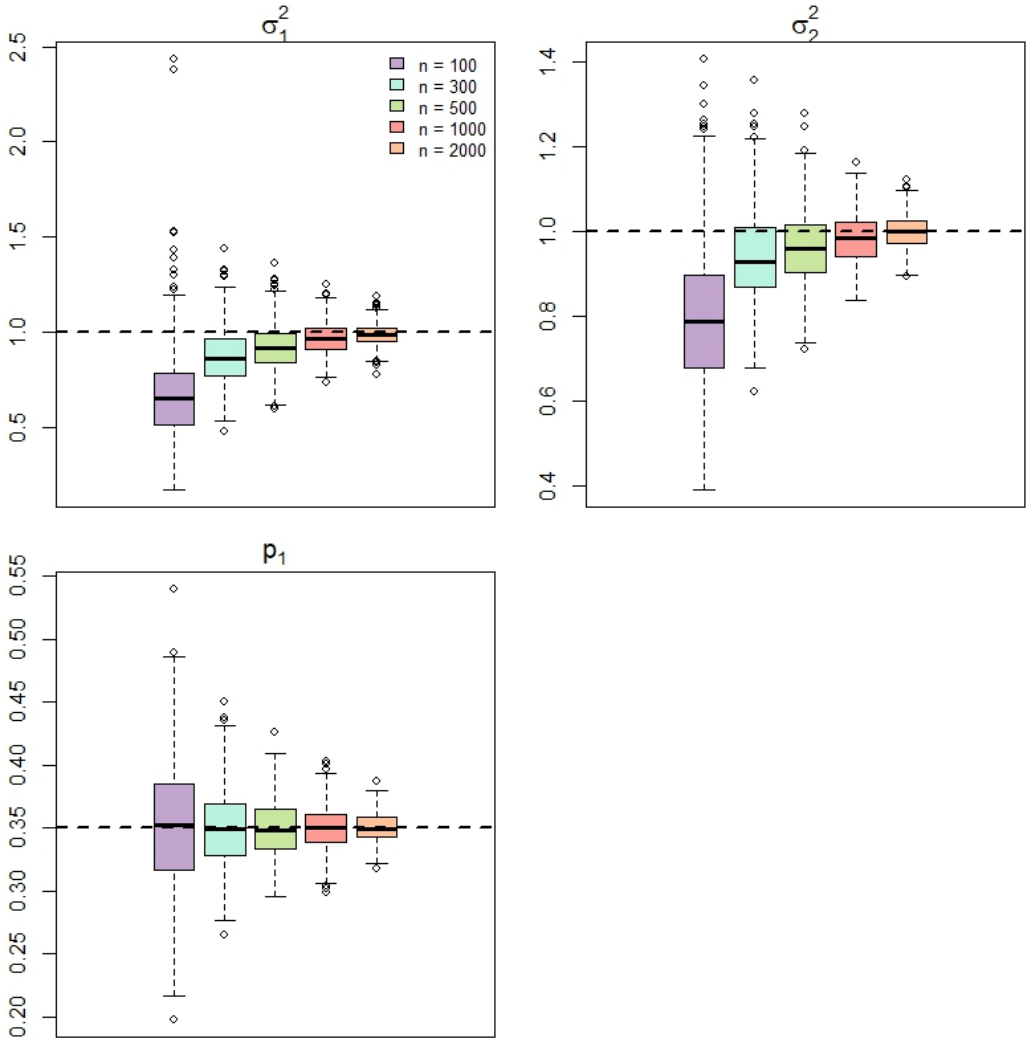
Fonte: Elaboração própria (2025).

Figura 13 – Boxplots das estimativas de β_{10}, β_{20}, β_{11}, β_{21}, β_{12}, β_{22} para o modelo proposto sob o Cenário 2. A linha tracejada indica o valor verdadeiro do parâmetro.



Fonte: Elaboração própria (2025).

Figura 14 – Boxplots das estimativas de σ^2 e p_1 para o modelo proposto sob o Cenário 2. A linha tracejada indica o valor verdadeiro do parâmetro.



Fonte: Elaboração própria (2025).

4.2.2.1 Cenário 2 Parcimonioso

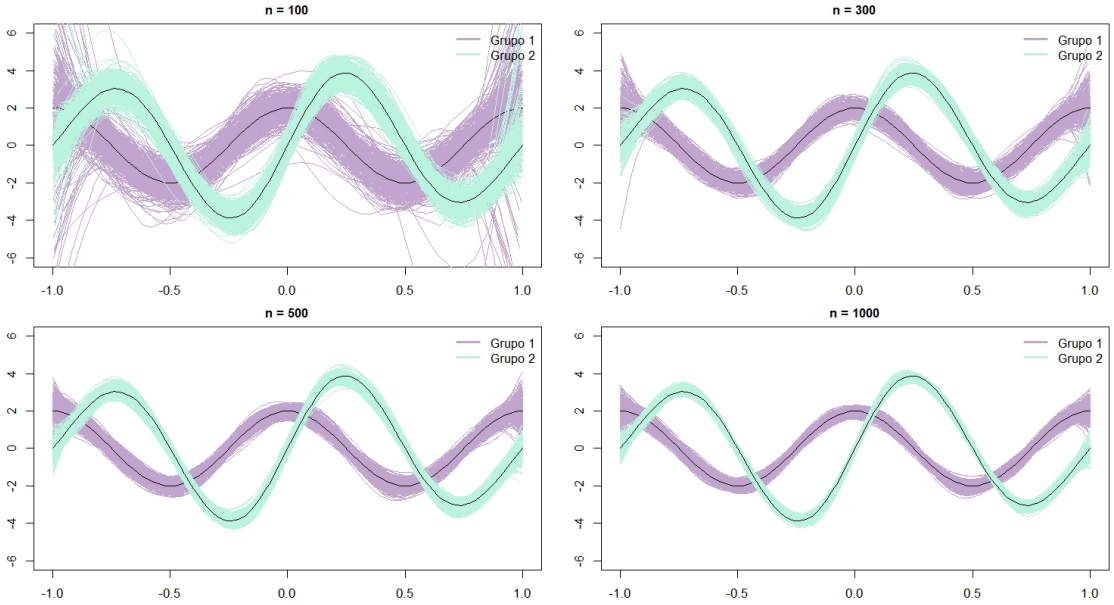
Sem perda de generalidade, neste cenário, consideramos $n = 100, 300, 500$ e 1000. Com base nos resultados apresentados na Tabela 4 e nas Figuras 15 a 18, observa-se que o algoritmo, em sua configuração parcimoniosa, apresenta desempenho comparável ao da versão completa em termos de qualidade das estimativas. A principal diferença entre as duas abordagens está no tempo de execução, sendo que a versão parcimoniosa demanda menos operações computacionais, o que a torna mais eficiente do ponto de vista computacional.

Tabela 4 – Resultados da simulação para o Cenário 2 parcimonioso: valor verdadeiro do parâmetro, seguido da média, do desvio padrão (*sd*) e do erro padrão calculado pela matriz de informação empírica (*sd*.emp) das 500 estimativas obtidas pelo algoritmo EM, para cada tamanho de amostra considerado.

		n=100			n=300			n=500			n=1000		
	θ	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp
p_1	0.35	0.354	0.048	0.053	0.352	0.028	0.029	0.351	0.023	0.023	0.351	0.016	0.016
β_{10}	4.00	3.923	0.632	1.362	4.011	0.305	0.331	4.009	0.234	0.240	4.011	0.174	0.164
β_{11}	4.00	4.038	0.755	0.954	4.017	0.375	0.408	3.988	0.282	0.300	3.997	0.206	0.204
β_{12}	6.00	6.022	0.810	0.957	5.974	0.405	0.423	5.987	0.283	0.311	5.985	0.205	0.211
β_{20}	2.00	1.961	0.462	0.402	2.012	0.250	0.210	2.001	0.194	0.159	1.998	0.128	0.112
β_{21}	3.00	3.046	0.513	0.497	2.986	0.260	0.268	2.998	0.194	0.203	3.005	0.136	0.142
β_{22}	-5.00	-5.005	0.510	0.511	-4.998	0.281	0.272	-4.989	0.212	0.207	-5.008	0.138	0.145
σ^2	1.00	0.768	0.172	0.173	0.916	0.078	0.106	0.952	0.063	0.083	0.980	0.048	0.059

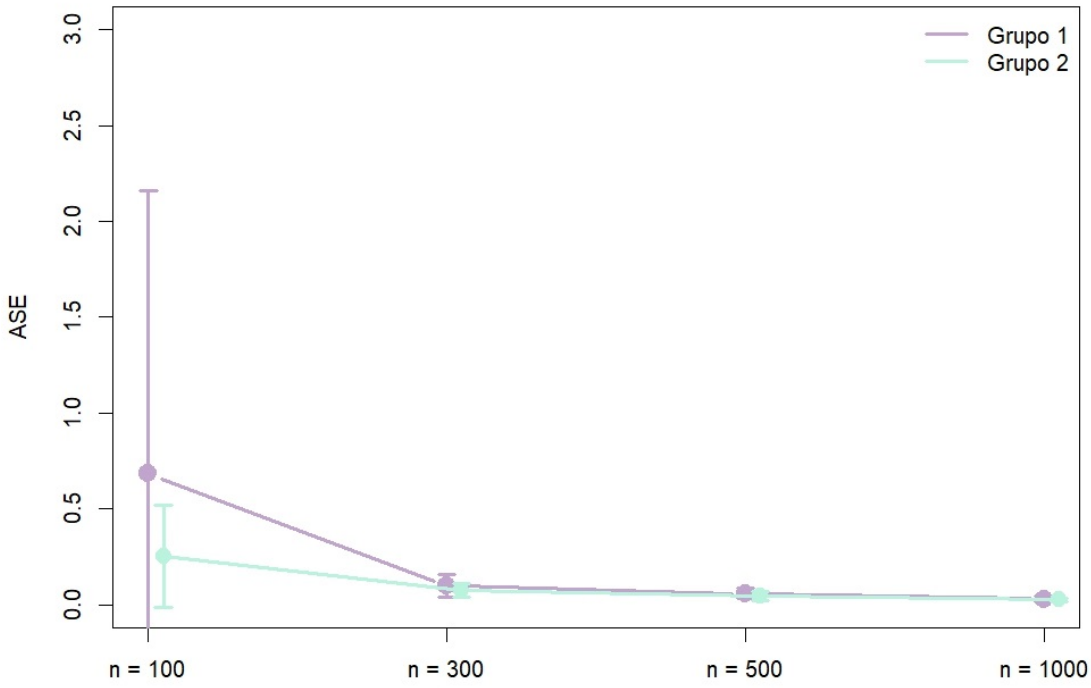
Fonte: Elaboração própria (2025).

Figura 15 – Gráficos das componentes não paramétricas com 500 réplicas. Curvas ajustadas (linhas representadas nas cores roxa e azul) e curvas verdadeiras (linhas pretas) sob o Cenário 2 parcimonioso, para diferentes tamanhos de amostra.



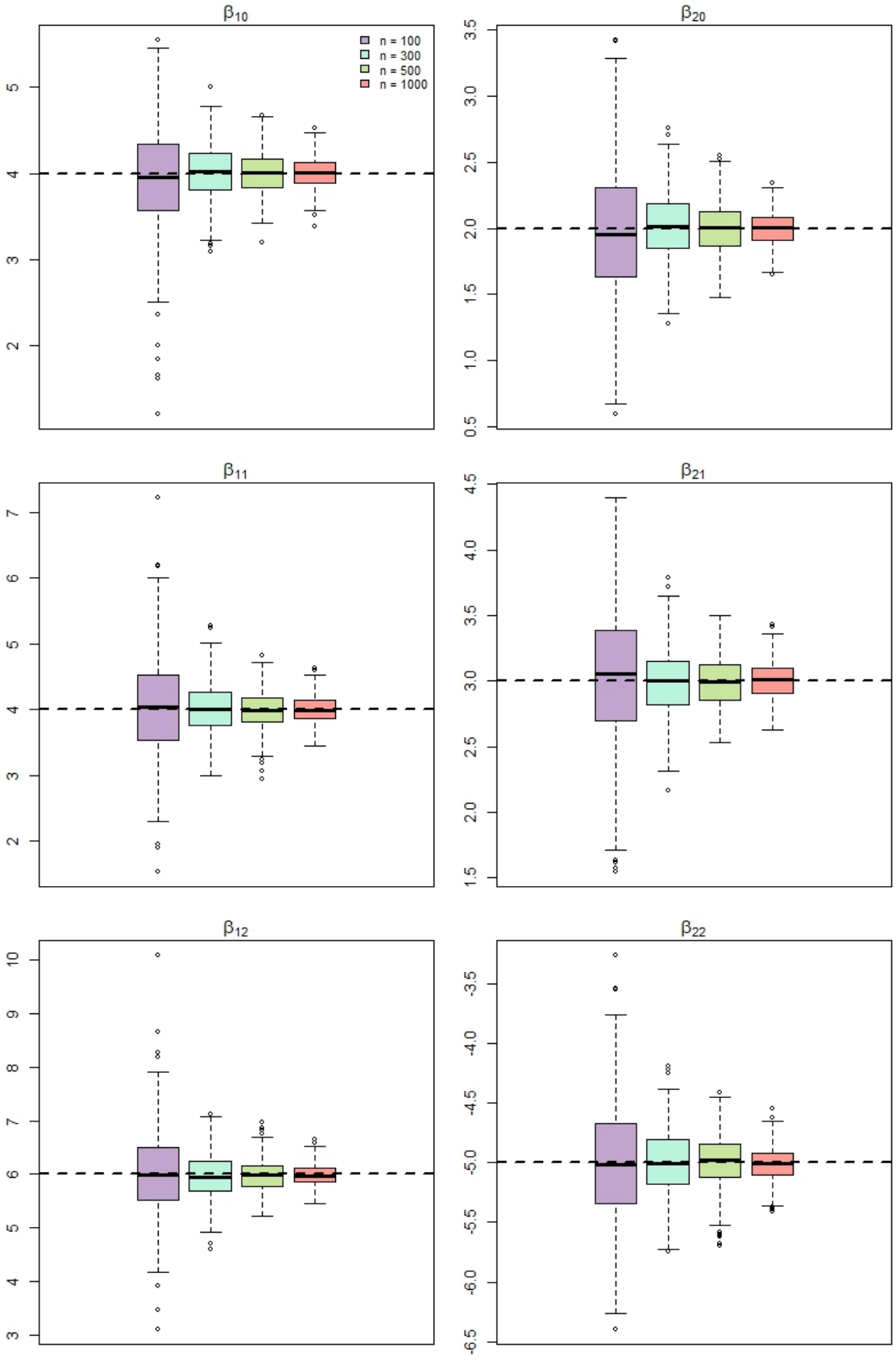
Fonte: Elaboração própria (2025).

Figura 16 – Média ± um desvio padrão dos ASEs das curvas, sob o Cenário 2 parcimonioso, para cada tamanho de amostra.



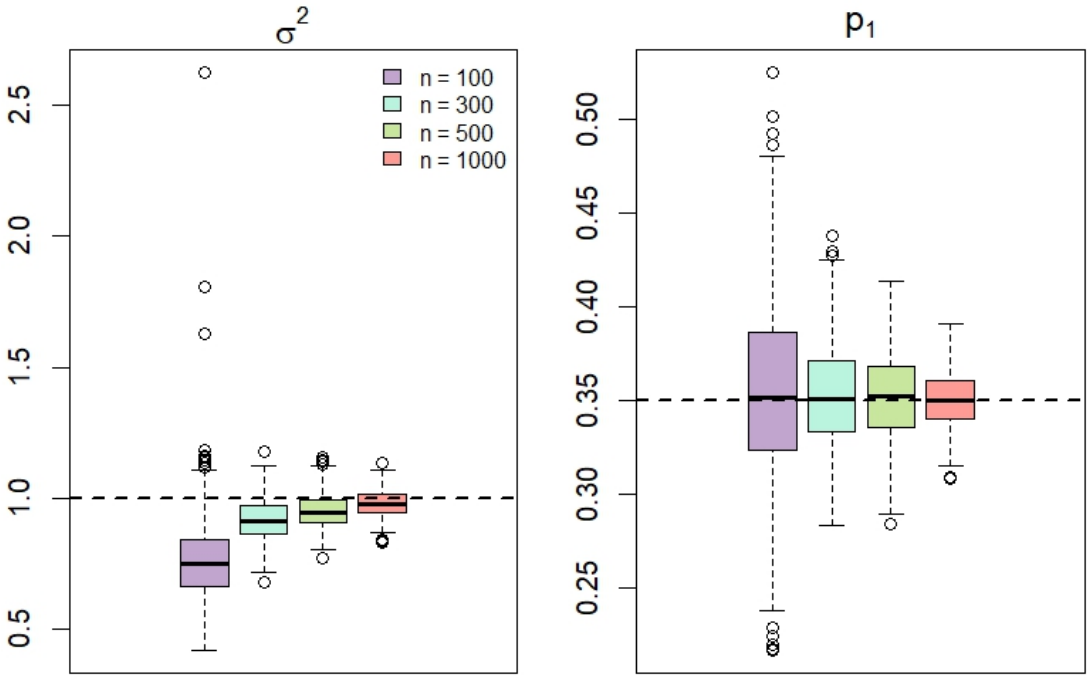
Fonte: Elaboração própria (2025).

Figura 17 – Boxplots das estimativas de β_{10}, β_{20}, β_{11}, β_{21}, β_{12}, β_{22} para o modelo proposto sob o Cenário 2 parcimonioso. A linha tracejada indica o valor verdadeiro do parâmetro.



Fonte: Elaboração própria (2025).

Figura 18 – Boxplots das estimativas de σ^2 e p_1 para o modelo proposto sob o Cenário 2 parcimonioso. A linha tracejada indica o valor verdadeiro do parâmetro.



Fonte: Elaboração própria (2025).

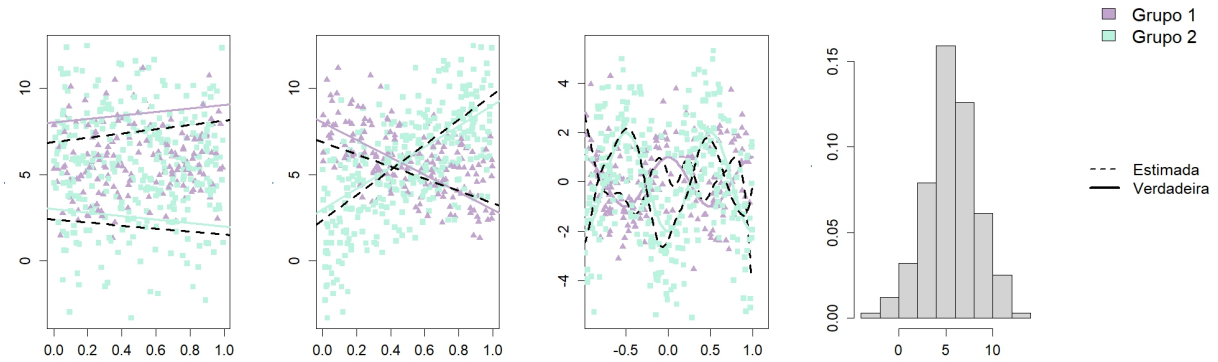
Os resultados apresentados mostram que o algoritmo, em sua configuração parcimoniosa, apresenta as propriedades assintóticas realizando um número menor operações.

4.2.3 Cenário 3 (C3)

Neste cenário, os parâmetros foram definidos de modo a gerar uma alta sobreposição entre os grupos, caracterizando grupos mal separados. Os valores adotados foram: $p_1 = 0.35$; os vetores de coeficientes $\boldsymbol{\beta}_1 = (8, 1, -5)^\top$ e $\boldsymbol{\beta}_2 = (3, -1, 6)^\top$; variâncias distintas, com $\sigma_1^2 = 2$ e $\sigma_2^2 = 4$; funções não lineares definidas como $g_1(t) = \cos(\pi t)$ e $g_2(t) = -2\cos(\pi t)$; e os parâmetros de suavização α_1 e α_2 fixados em 0.1. As covariáveis foram geradas conforme distribuições uniformes: $x_{im,j}$ no intervalo $(0, 1)$ e t_{ij} no intervalo $(-1, 1)$, para $m = 1, 2$; $j = 1, 2$; e $i = 1, \ldots, n$.

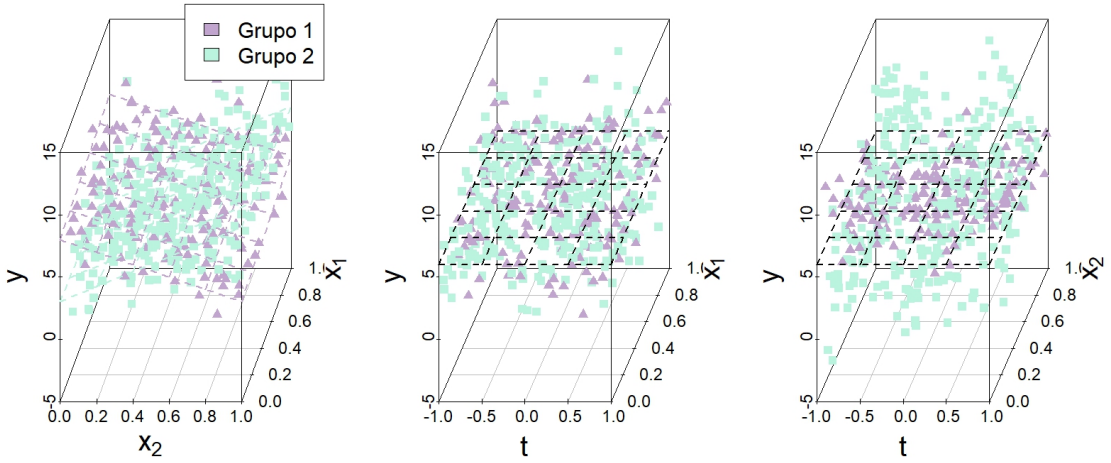
As Figuras 19 e 20 revelam uma considerável sobreposição entre os grupos, destacando as diferenças nas estruturas utilizadas para a geração dos dados. Neste cenário, devido à alta sobreposição dos grupos, o algoritmo proposto apresentou maior dificuldade na recuperação dos valores verdadeiros dos parâmetros (veja a Tabela 5 e as Figuras 21 a 24). Mesmo com amostras de tamanho 5000, as estimativas ainda não apresentam boa precisão. Neste cenário, a convergência parece ser mais lenta; porém, percebe-se que, à medida que o tamanho amostral aumenta, há menor viés e menor variabilidade. No estudo de simulação 2 (Seção 4.3), apresentaremos algumas métricas relacionadas à classificação, que ajudarão a esclarecer o motivo dessa dificuldade.

Figura 19 – Gráficos de dispersão da variável resposta y em função das covariáveis (primeiro painel à esquerda: x_1, segundo painel: x_2) e dos resíduos não paramétricos em função de t (terceiro painel), com as respectivas retas e curvas ajustadas, acompanhados de um histograma da densidade de y (quarto painel) para uma amostra de tamanho $n = 500$.



Fonte: Elaboração própria (2025).

Figura 20 – Gráficos de dispersão em 3D da variável resposta y, para uma amostra de tamanho $n = 500$. No painel à esquerda, as covariáveis x_1 e x_2 formam o plano da base; no painel central, t e x_1; e, no painel à direita, t e x_2.



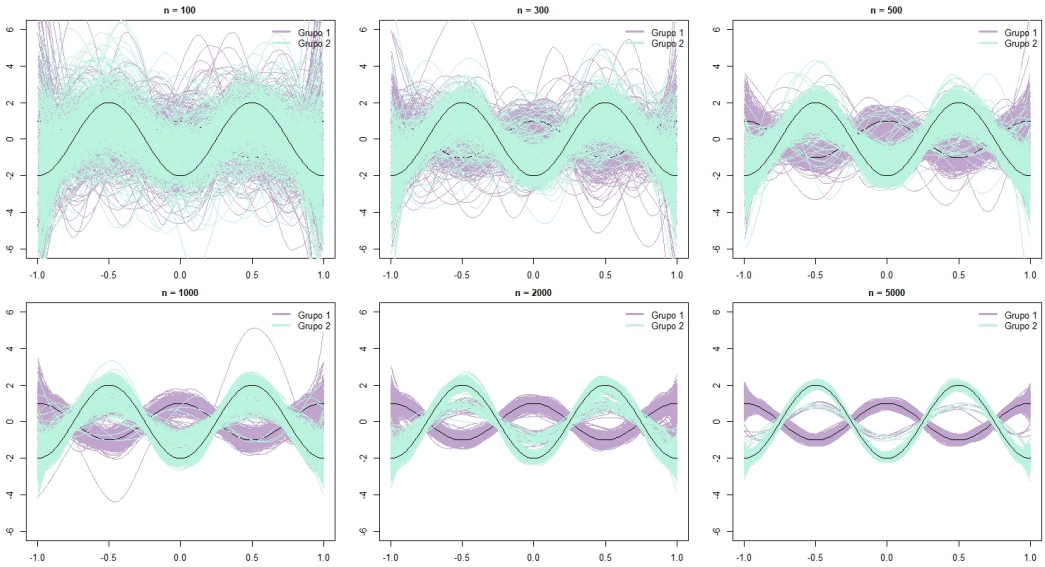
Fonte: Elaboração própria (2025).

Tabela 5 – Resultados da simulação para o Cenário 3: valor verdadeiro do parâmetro, seguido da média, do desvio padrão (*sd*) e do erro padrão calculado pela matriz de informação empírica (*sd*.emp) das 500 estimativas obtidas pelo algoritmo EM, para cada tamanho de amostra considerado.

		n=100			n=300			n=500			n=1000			n=2000			n=5000		
	θ	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp
p_1	0.35	0.478	0.201	0.084	0.426	0.190	0.078	0.391	0.126	0.073	0.380	0.072	0.050	0.366	0.038	0.027	0.357	0.016	0.013
β_{10}	8.00	6.964	1.754	1.113	7.132	1.468	0.594	7.517	1.018	0.512	7.712	0.714	0.359	7.895	0.388	0.226	7.961	0.176	0.131
β_{11}	1.00	-0.181	1.852	1.092	0.351	1.266	0.685	0.605	0.876	0.559	0.755	0.652	0.386	0.895	0.396	0.263	0.978	0.195	0.163
β_{12}	-5.00	1.334	2.616	1.053	-0.866	3.559	0.679	-2.474	3.314	0.571	-3.627	2.686	0.404	-4.507	1.608	0.281	-4.878	0.603	0.176
β_{20}	3.00	2.833	1.460	1.463	3.164	1.207	0.583	3.070	0.749	0.461	3.021	0.504	0.320	2.963	0.209	0.201	2.972	0.116	0.116
β_{21}	-1.00	-0.377	1.640	1.073	-0.608	0.968	0.670	-0.746	0.721	0.500	-0.868	0.450	0.341	-0.943	0.282	0.232	-1.006	0.146	0.143
β_{22}	6.00	2.677	2.052	1.041	3.661	2.588	0.670	4.599	2.081	0.523	5.281	1.704	0.369	5.844	0.957	0.257	6.034	0.356	0.161
σ_1^2	2.00	1.992	1.515	0.712	2.460	1.493	0.633	2.426	1.226	0.555	2.345	0.912	0.372	2.158	0.592	0.218	2.045	0.243	0.120
σ_2^2	4.00	2.147	1.390	0.745	3.427	1.260	0.680	3.755	0.788	0.559	3.867	0.370	0.373	3.929	0.210	0.227	3.942	0.127	0.129

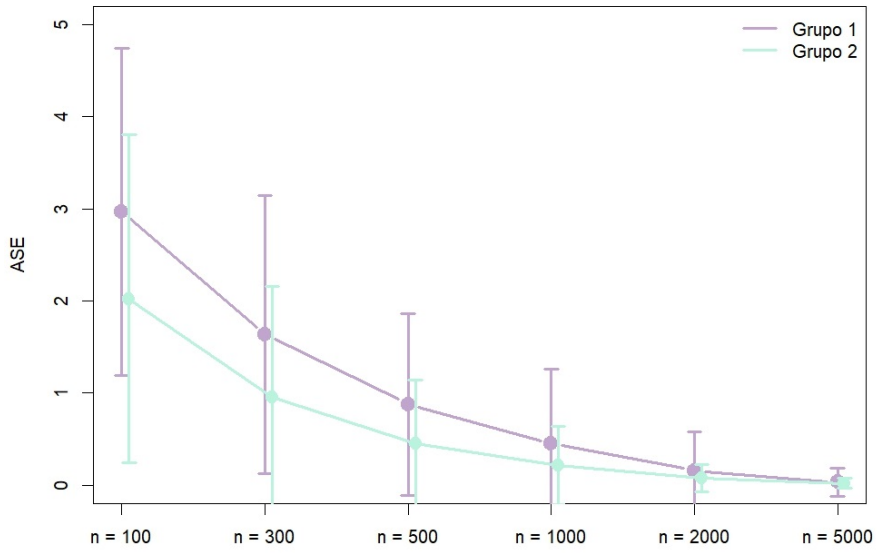
Fonte: Elaboração própria (2025).

Figura 21 – Gráficos das componentes não paramétricas com 500 réplicas. Curvas ajustadas (linhas representadas nas cores roxa e azul) e curvas verdadeiras (linhas pretas) sob o Cenário 3, para diferentes tamanhos de amostra.



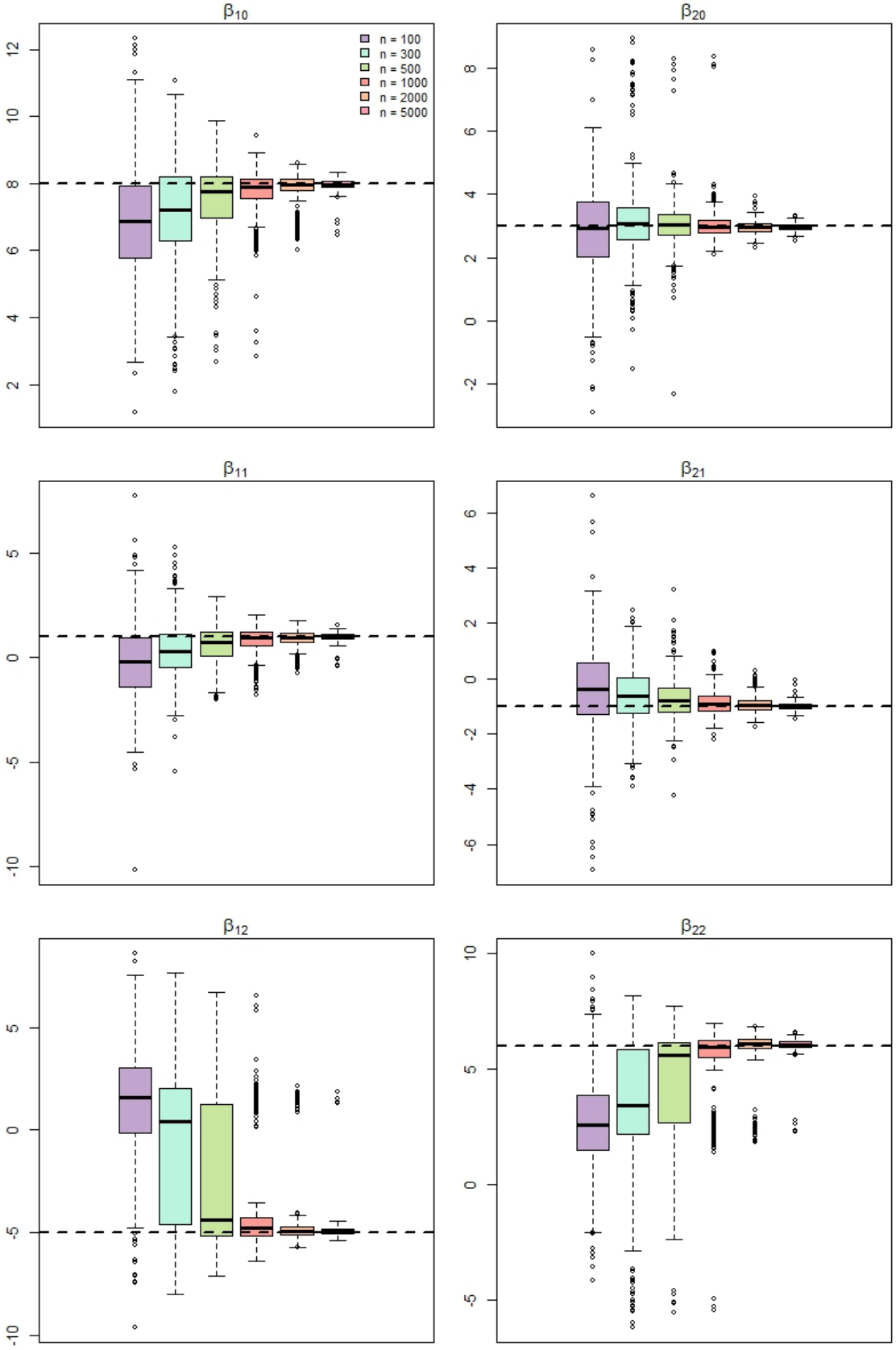
Fonte: Elaboração própria (2025).

Figura 22 – Média ± um desvio padrão dos ASEs das curvas, sob o Cenário 3, para cada tamanho de amostra.



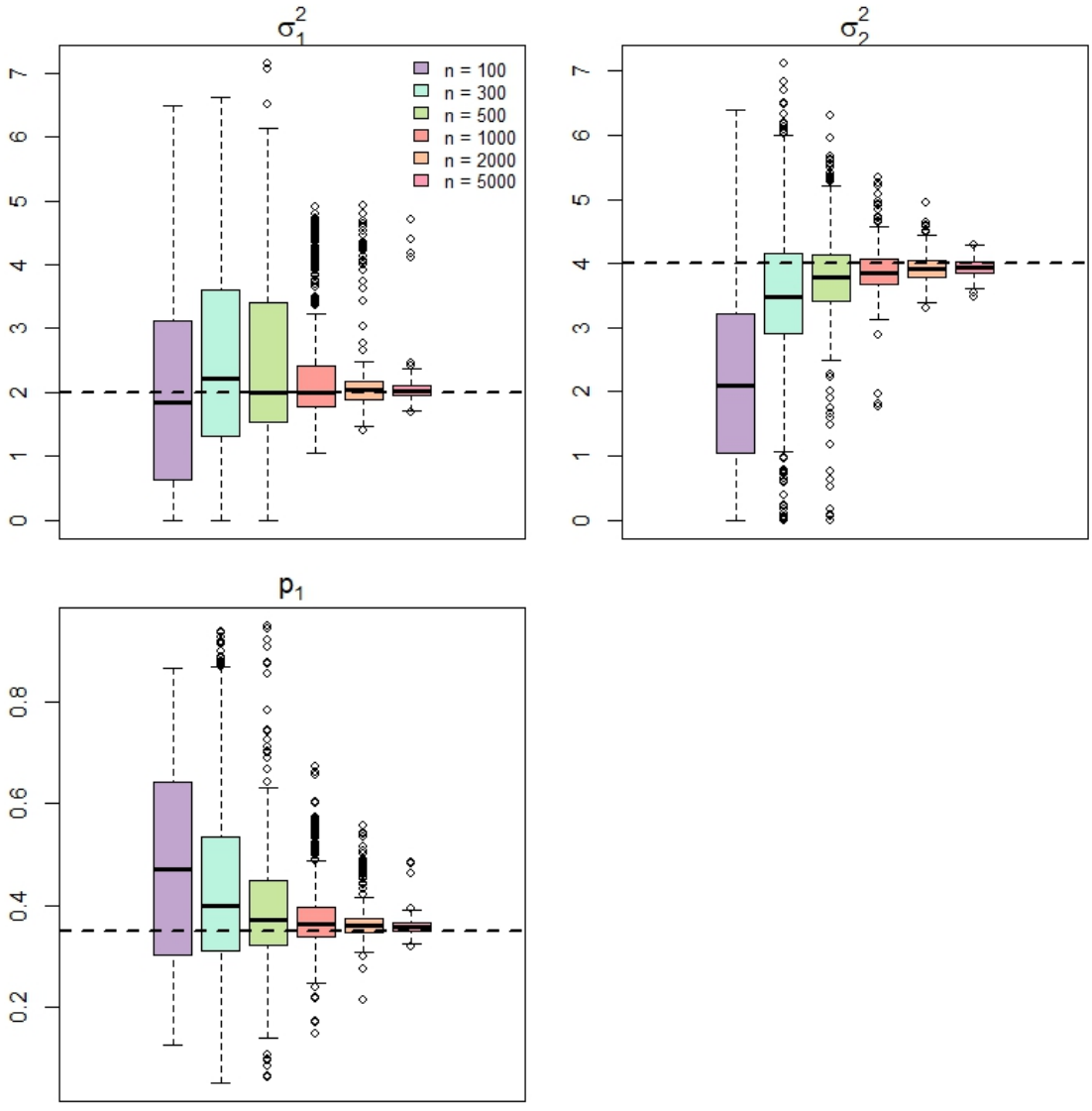
Fonte: Elaboração própria (2025).

Figura 23 – Boxplots das estimativas de β_{10}, β_{20}, β_{11}, β_{21}, β_{12}, β_{22} para o modelo proposto sob o Cenário 3. A linha tracejada indica o valor verdadeiro do parâmetro.



Fonte: Elaboração própria (2025).

Figura 24 – Boxplots das estimativas de σ_1^2, σ_2^2 e p_1 para o modelo proposto sob o Cenário 3. A linha tracejada indica o valor verdadeiro do parâmetro.



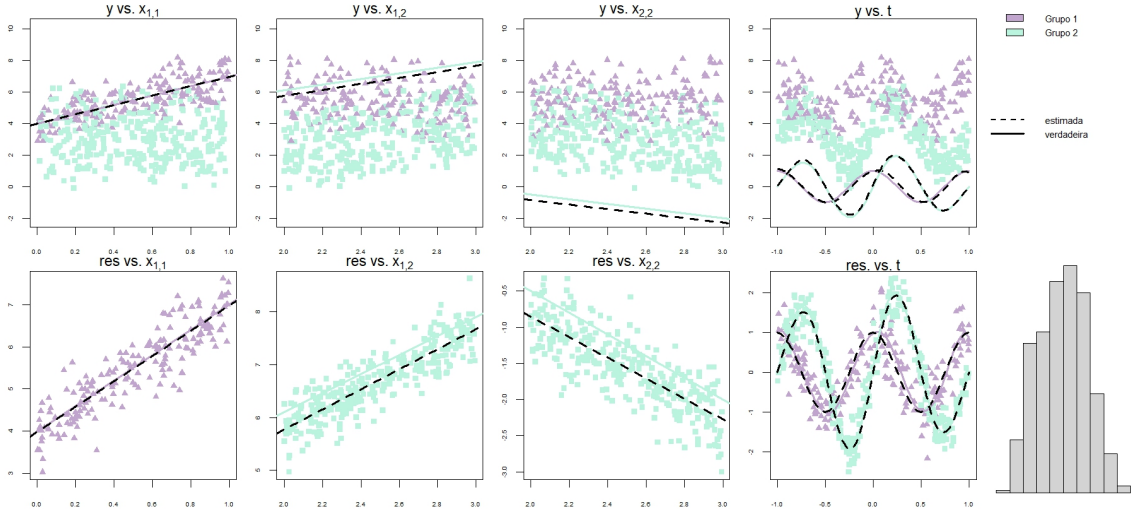
Fonte: Elaboração própria (2025).

4.2.4 Cenário 4 (C4)

No cenário 4, os parâmetros foram definidos de forma que os grupos estivessem separados, porém com algumas interseções, ou seja, moderamente separados. A proporção do primeiro grupo é $p_1 = 0.35$, os vetores de coeficientes são $\boldsymbol{\beta}_1 = (4, 3)^\top$ e $\boldsymbol{\beta}_2 = (2.5, 1.8, -1.5)^\top$, e as variâncias são $\sigma_1^2 = 0.2$ e $\sigma_2^2 = 0.1$. As funções $g_1(t)$ e $g_2(t)$ foram definidas como $g_1(t) = \cos(\pi t)$ e $g_2(t) = 2\sin(\pi t)\exp(-0.5t^2)$, respectivamente, enquanto os parâmetros de regularização são iguais para os dois grupos, com $\alpha_1 = \alpha_2 = 0.1$. As covariáveis foram geradas de acordo com as seguintes distribuições: $x_{i1,1} \sim \text{Uniforme}(0, 1)$, $x_{i1,2} \sim \text{Uniforme}(2, 3)$, $x_{i2,2} \sim \text{Uniforme}(2, 3)$ e $t_{ij} \sim \text{Uniforme}(-1, 1)$, para $i = 1, \ldots, n$ e $j = 1, 2$. É importante destacar que os vetores $\boldsymbol{\beta}$ possuem dimensões diferentes entre os grupos.

As Figuras 25 e 26 indicam uma leve sobreposição entre os grupos, evidenciando as distinções nas estruturas de geração dos dados adotadas, especialmente devido às dimensões diferentes dos vetores de coeficientes $\boldsymbol{\beta}$ e às distribuições distintas das covariáveis lineares. Os resultados apresentados na Tabela 6 e nas Figuras 27 a 30 revelam que, apesar da separação moderada e da heterogeneidade nas covariáveis, o algoritmo performou de forma satisfatória, com estimativas apresentando boa precisão mesmo para tamanhos amostrais menores (como $n = 100$). Observa-se que o aumento no tamanho da amostra leva a médias das estimativas mais próximas dos valores verdadeiros, com redução no viés e na variabilidade. Na Figura 27, as curvas estimadas se aproximam da verdadeira à medida que n aumenta, o que é corroborado pela Figura 28, onde a média dos ASEs diminui rapidamente, com variabilidade reduzida, refletindo bom desempenho na recuperação das componentes não paramétricas. Os boxplots nas Figuras 29 e 30 mostram maior concentração em torno dos valores verdadeiros com amostras maiores, confirmando a redução no viés e na variabilidade, alinhado às propriedades assintóticas do método de máxima verossimilhança, mesmo com estruturas semiparamétricas diferenciadas entre grupos.

Figura 25 – Gráficos de dispersão da variável resposta y em função das covariáveis (primeira linha) e dos resíduos referentes a cada covariável (segunda linha), com as respectivas retas e curvas ajustadas, acompanhados de um histograma da densidade de y para uma amostra de tamanho $n = 500$.



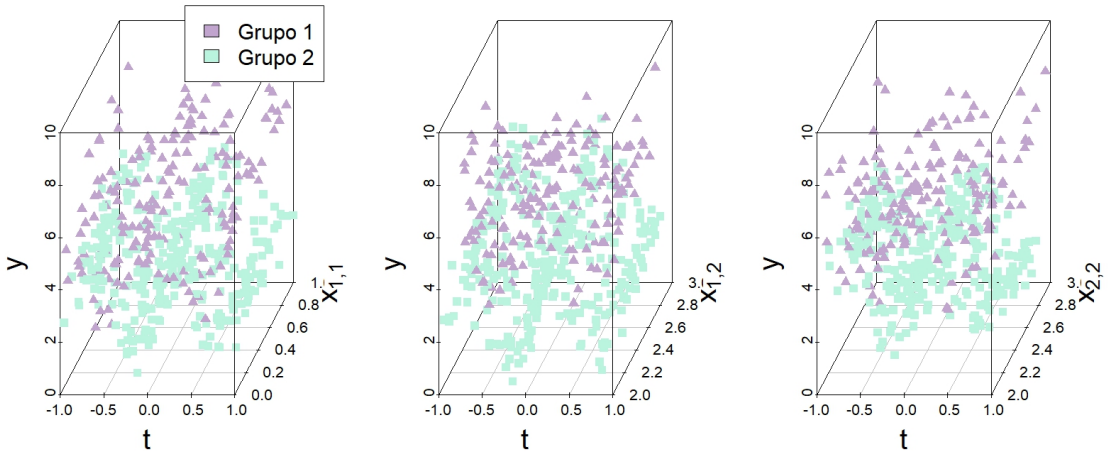
Fonte: Elaboração própria (2025).

Tabela 6 – Resultados da simulação para o Cenário 4: valor verdadeiro do parâmetro, seguido da média, do desvio padrão (sd) e do erro padrão calculado pela matriz de informação empírica (sd.emp) das 500 estimativas obtidas pelo algoritmo EM, para cada tamanho de amostra considerado.

		n=100			n=300			n=500			n=1000		
	θ	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp
p_1	0.35	0.362	0.063	0.054	0.352	0.030	0.031	0.349	0.024	0.024	0.350	0.016	0.017
β_{10}	4.00	4.003	0.261	0.524	3.993	0.112	0.111	4.000	0.088	0.081	3.999	0.061	0.056
β_{11}	3.00	2.934	0.437	0.364	3.008	0.173	0.180	3.004	0.133	0.135	3.005	0.092	0.093
β_{20}	2.50	2.499	0.661	0.861	2.474	0.325	0.315	2.500	0.243	0.238	2.489	0.168	0.166
β_{21}	1.80	1.793	0.179	0.165	1.804	0.085	0.089	1.801	0.069	0.067	1.803	0.045	0.047
β_{22}	-1.50	-1.508	0.169	0.166	-1.494	0.089	0.089	-1.502	0.064	0.067	-1.498	0.047	0.047
σ_1^2	0.20	0.145	0.078	0.051	0.177	0.028	0.032	0.185	0.021	0.024	0.193	0.017	0.017
σ_2^2	0.10	0.078	0.018	0.019	0.095	0.012	0.012	0.098	8.43×10^{-03}	8.89×10^{-03}	0.100	6.44×10^{-03}	6.28×10^{-03}

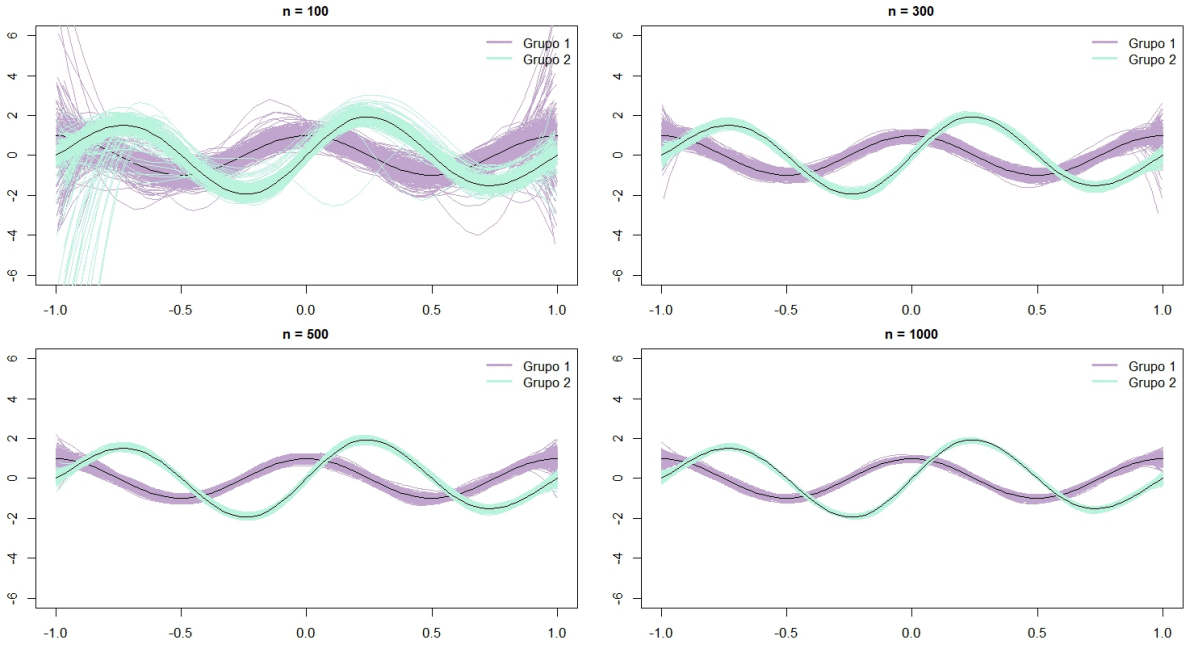
Fonte: Elaboração própria (2025).

Figura 26 – Gráficos de dispersão em 3D da variável resposta y, para uma amostra de tamanho $n = 500$. No painel à esquerda, as covariáveis t e $x_{1,1}$ formam o plano da base; no painel central, t e $x_{1,2}$; e, no painel à direita, t e $x_{2,2}$.



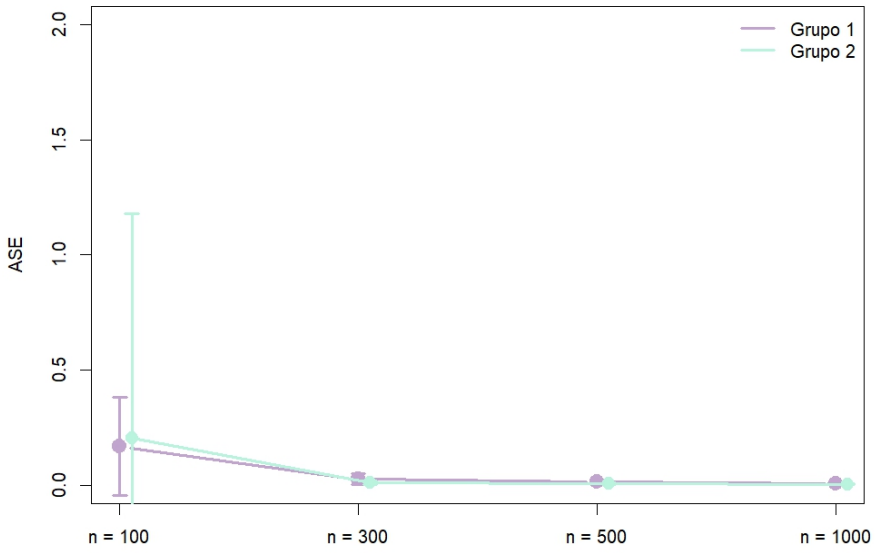
Fonte: Elaboração própria (2025).

Figura 27 – Gráficos das componentes não paramétricas com 500 réplicas. Curvas ajustadas (linhas representadas nas cores roxa e azul) e curvas verdadeiras (linhas pretas) sob o Cenário 4, para diferentes tamanhos de amostra.



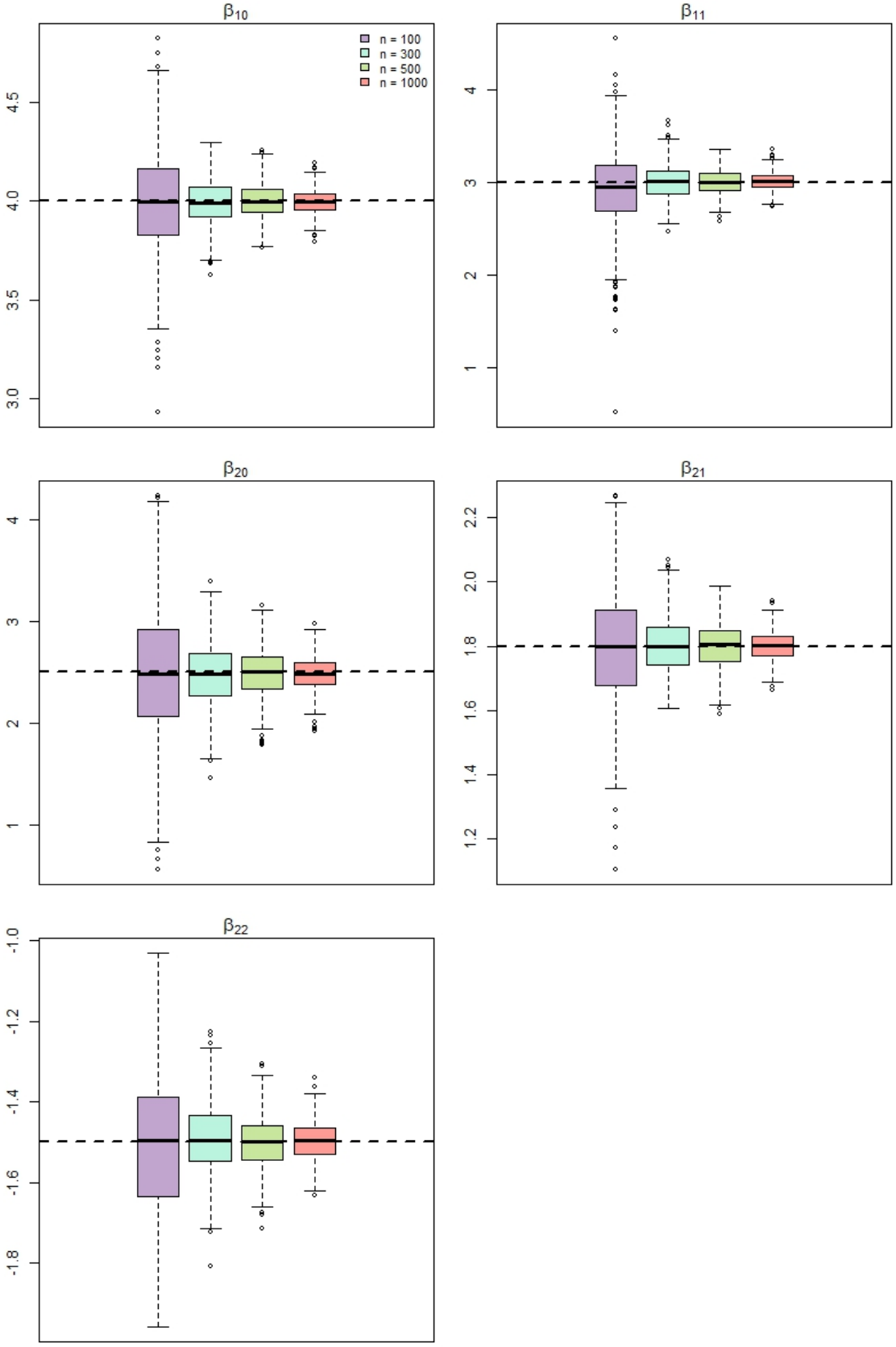
Fonte: Elaboração própria (2025).

Figura 28 – Média ± um desvio padrão dos ASEs das curvas, sob o Cenário 4, para cada tamanho de amostra.



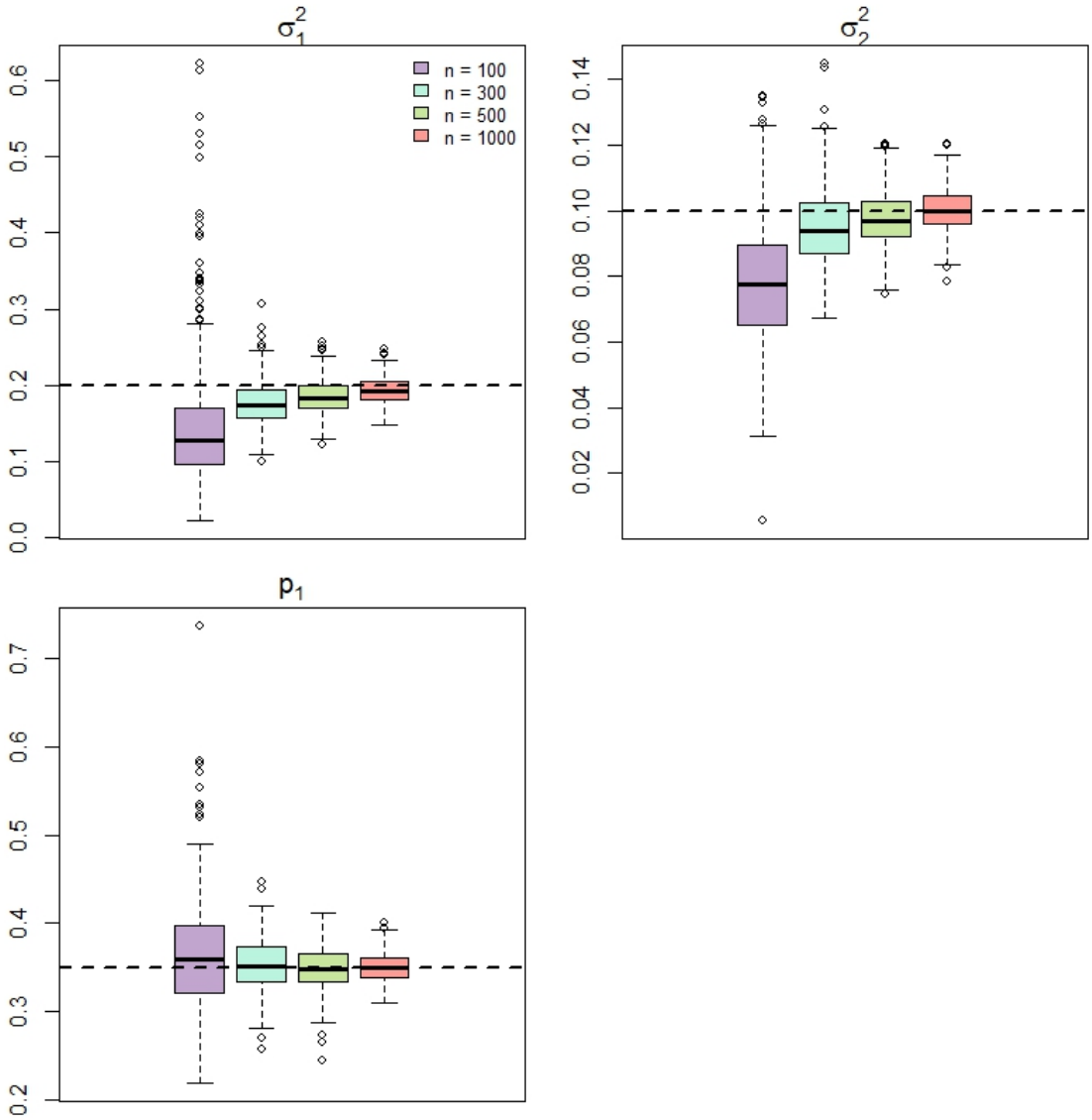
Fonte: Elaboração própria (2025).

Figura 29 – Boxplots das estimativas de β_{10}, β_{11}, β_{20}, β_{21}, β_{22} para o modelo proposto sob o Cenário 4. A linha tracejada indica o valor verdadeiro do parâmetro.



Fonte: Elaboração própria (2025).

Figura 30 – Boxplots das estimativas de σ_1^2, σ_2^2 e p_1 para o modelo proposto sob o Cenário 4. A linha tracejada indica o valor verdadeiro do parâmetro.



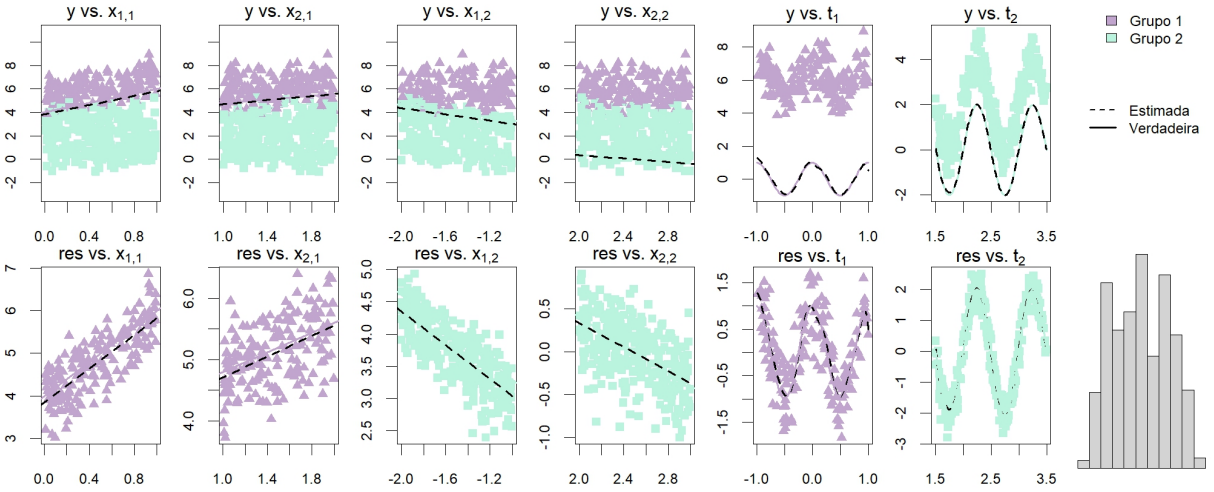
Fonte: Elaboração própria (2025).

4.2.5 Cenário 5 (C5)

No cenário 5, os parâmetros foram definidos de forma que os grupos estivessem moderamente separados também. A proporção do primeiro grupo é $p_1 = 0.35$, os vetores de coeficientes são $\boldsymbol{\beta}_1 = (4, 2, 0.8)^\top$ e $\boldsymbol{\beta}_2 = (2, -1.3, -0.8)^\top$, e as variâncias são $\sigma_1^2 = 0.2$ e $\sigma_2^2 = 0.1$. As funções $g_1(t)$ e $g_2(t)$ foram definidas como $g_1(t) = \cos(2\pi t)$ e $g_2(t) = 2\sin(2\pi t)$, respectivamente, enquanto os parâmetros de suavização são iguais para os dois grupos, com $\alpha_1 = \alpha_2 = 1$. As covariáveis foram geradas de acordo com as seguintes distribuições: $x_{i1,1} \sim \text{Uniforme}(0, 1)$, $x_{i2,1} \sim \text{Uniforme}(1, 2)$, $x_{i1,2} \sim \text{Uniforme}(-2, -1)$, $x_{i2,2} \sim \text{Uniforme}(2, 3)$, $t_{i1} \sim \text{Uniforme}(-1, 1)$ e $t_{i2} \sim \text{Uniforme}(1.5, 3.5)$, para $i = 1, \ldots, n$ e $j = 1, 2$. Observe que, entre os grupos, as covariáveis lineares e as curvas estão definidas em intervalos diferentes.

As Figuras 31 e 32 revelam uma sobreposição moderada entre os grupos, destacando as diferenças nas estruturas de geração dos dados, particularmente pelos intervalos distintos das covariáveis lineares e não lineares. Com base na Tabela 7 e nas Figuras 33 a 36, observa-se um desempenho robusto do algoritmo, com estimativas consistentes apesar da heterogeneidade nos intervalos das variáveis. O aumento no tamanho amostral resulta em médias mais próximas dos valores reais, acompanhado de redução no viés e na variabilidade. A Figura 33 ilustra as curvas estimadas convergindo para as verdadeiras em escalas diferentes de t, enquanto a Figura 34 mostra a média dos ASEs diminuindo com n, com desvio padrão reduzido, indicando boa recuperação das funções não paramétricas mesmo em domínios disjuntos. Os boxplots nas Figuras 35 e 36 exibem maior concentração ao redor dos valores verdadeiros à medida que n cresce, reforçando a consistência assintótica e a capacidade do método em lidar com covariáveis e curvas definidas em intervalos variados entre grupos, sem comprometer a precisão geral.

Figura 31 – Gráficos de dispersão da variável resposta y em função das covariáveis (primeira linha) e dos resíduos referentes a cada covariável (segunda linha), com as respectivas retas e curvas ajustadas, acompanhados de um histograma da densidade de y para uma amostra de tamanho $n = 500$.



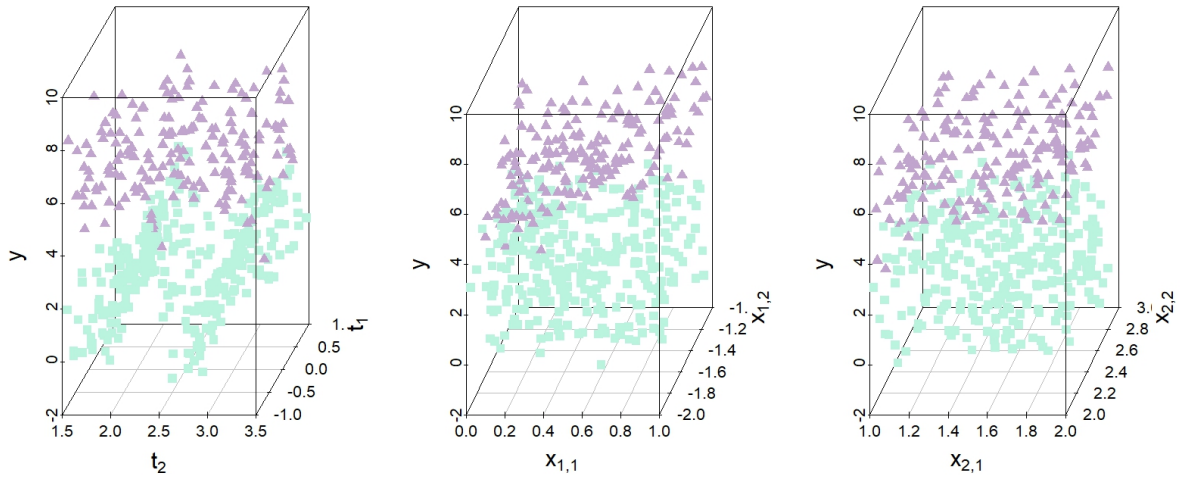
Fonte: Elaboração própria (2025).

Tabela 7 – Resultados da simulação para o Cenário 5: valor verdadeiro do parâmetro, seguido da média, do desvio padrão (*sd*) e do erro padrão calculado pela matriz de informação empírica (*sd*.emp) das 500 estimativas obtidas pelo algoritmo EM, para cada tamanho de amostra considerado.

		n=100			n=300			n=500			n=1000		
	θ	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp	$\hat{\theta}$	sd	sd.emp
p_1	0.35	0.350	0.051	0.052	0.349	0.028	0.028	0.351	0.021	0.022	0.349	0.016	0.015
β_{10}	4.00	4.041	0.539	1.140	3.980	0.280	0.283	4.022	0.216	0.206	4.006	0.139	0.142
β_{11}	2.00	1.992	0.330	0.384	2.006	0.160	0.173	2.001	0.119	0.127	1.992	0.089	0.088
β_{12}	0.80	0.769	0.347	0.383	0.811	0.170	0.173	0.785	0.133	0.127	0.800	0.085	0.088
β_{20}	2.00	2.009	0.469	0.483	1.994	0.260	0.249	2.010	0.195	0.187	2.016	0.135	0.13
β_{21}	-1.30	-1.299	0.148	0.161	-1.300	0.088	0.085	-1.297	0.066	0.064	-1.299	0.045	0.044
β_{22}	-0.80	-0.798	0.147	0.162	-0.796	0.083	0.085	-0.802	0.063	0.064	-0.805	0.043	0.044
σ_1^2	0.20	0.133	0.043	0.051	0.176	0.027	0.030	0.184	0.022	0.022	0.193	0.015	0.016
σ_2^2	0.10	0.078	0.016	0.018	0.094	0.010	0.011	0.097	7.97×10^{-03}	8.21×10^{-03}	0.098	5.67×10^{-03}	5.71×10^{-03}

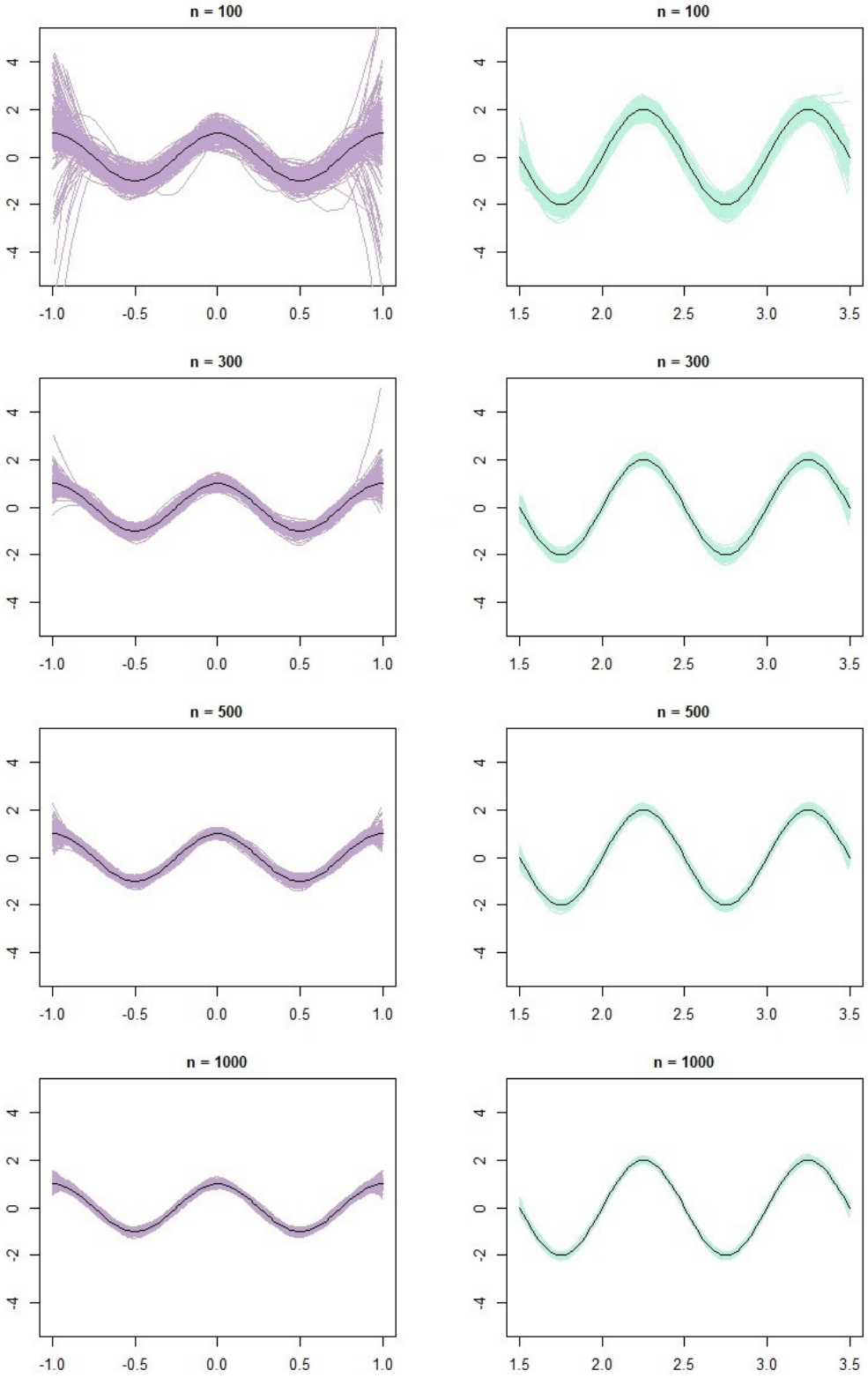
Fonte: Elaboração própria (2025).

Figura 32 – Gráficos de dispersão em 3D da variável resposta y, para uma amostra de tamanho $n = 500$. No painel à esquerda, as covariáveis t_1 e t_2 formam o plano da base; no painel central, $x_{1,1}$ e $x_{1,2}$; e, no painel à direita, $x_{2,1}$ e $x_{2,2}$.



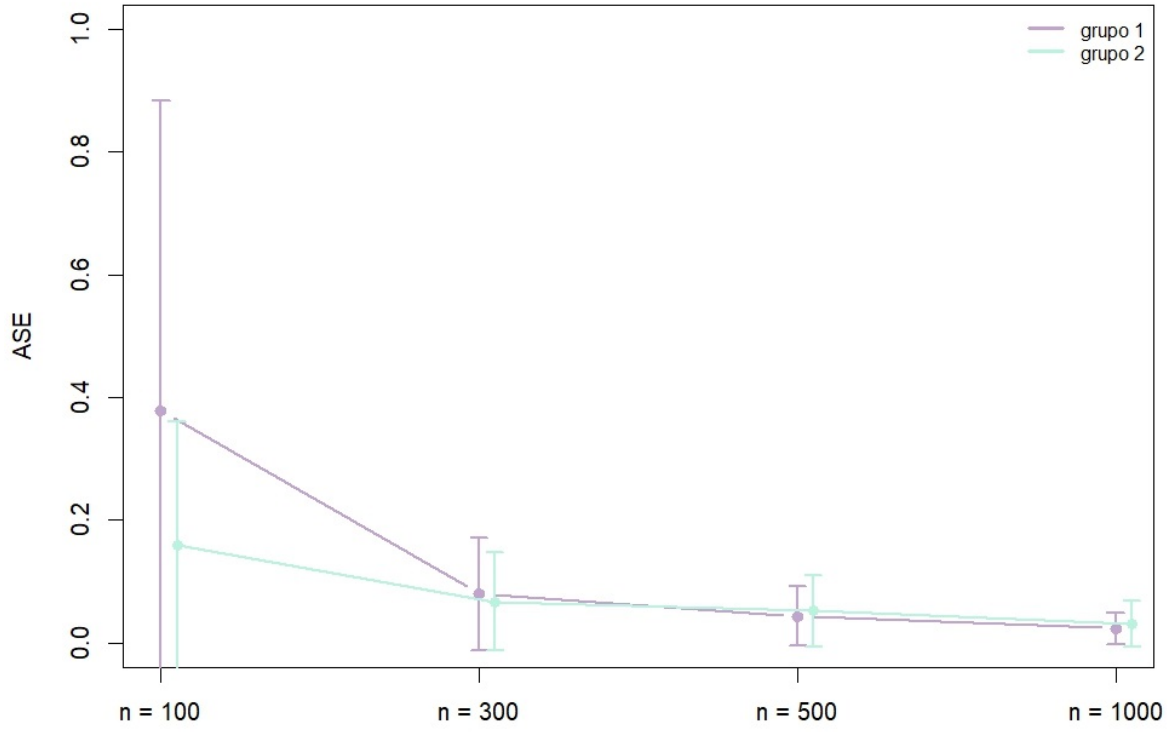
Fonte: Elaboração própria (2025).

Figura 33 – Gráficos das componentes não paramétricas com 500 réplicas. Curvas ajustadas (linhas representadas nas cores roxa e azul) e curvas verdadeiras (linhas pretas) sob o Cenário 5, para diferentes tamanhos de amostra.



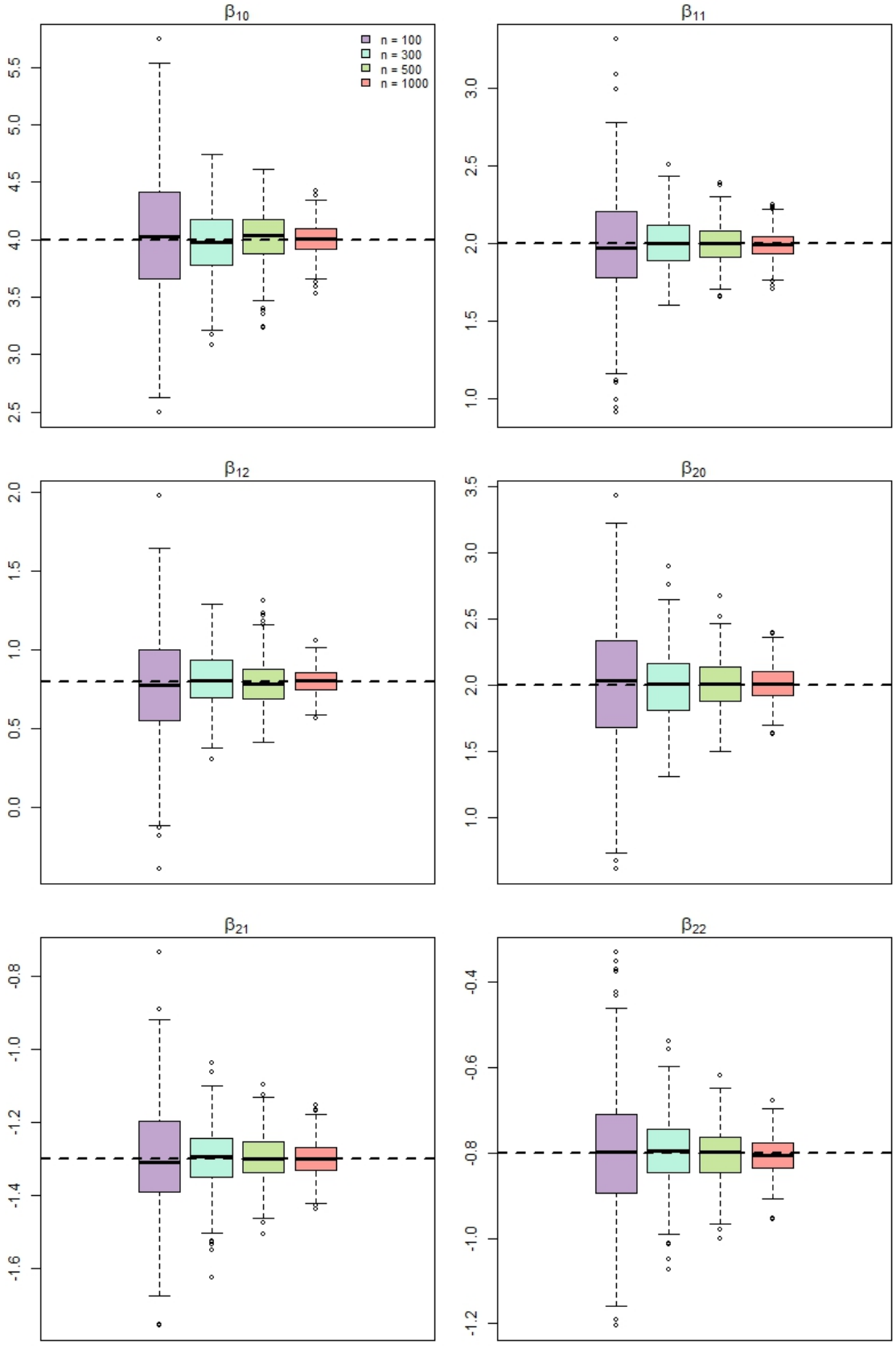
Fonte: Elaboração própria (2025).

Figura 34 – Média ± um desvio padrão dos ASEs das curvas, sob o Cenário 5, para cada tamanho de amostra.



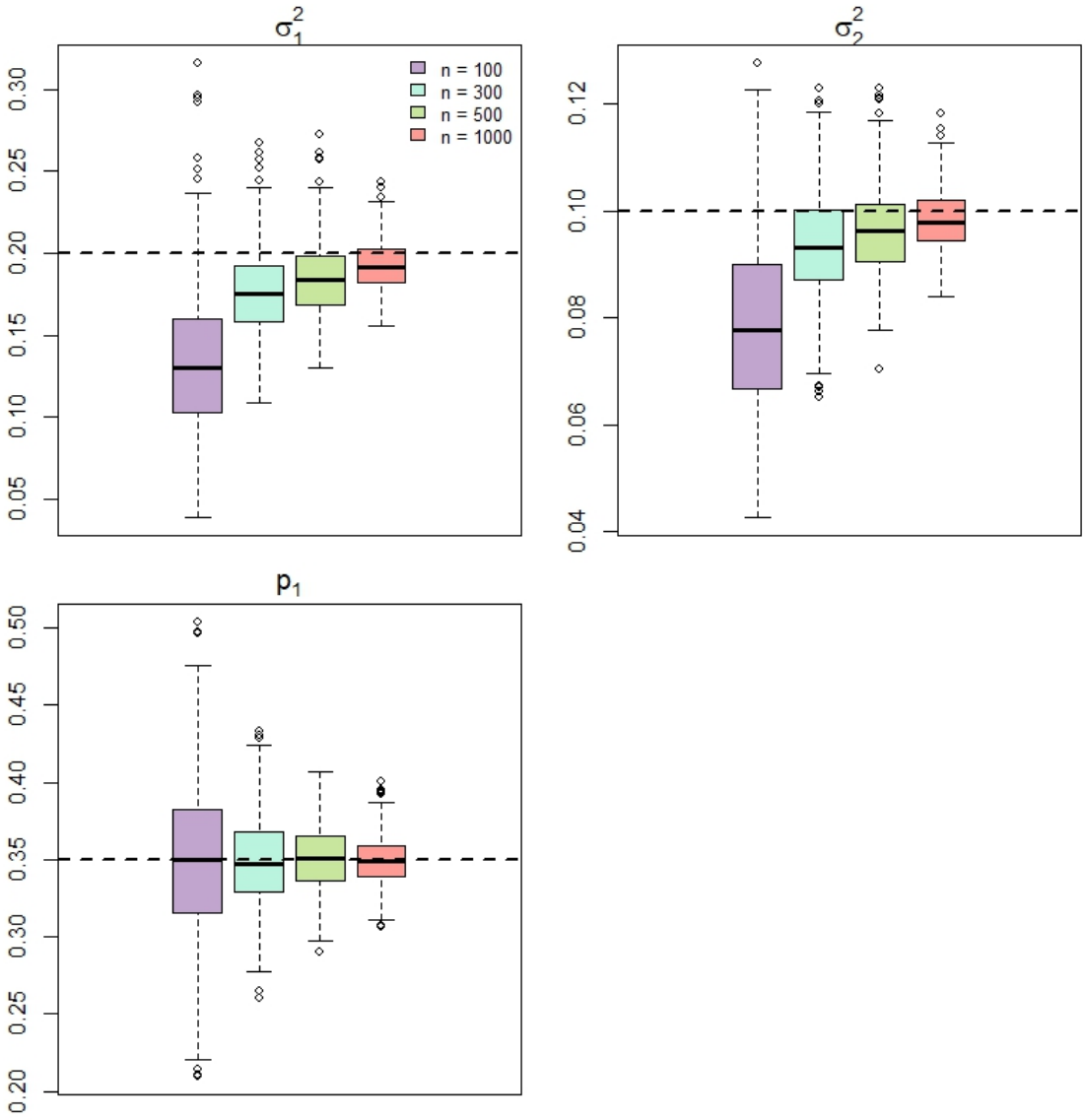
Fonte: Elaboração própria (2025).

Figura 35 – Boxplots das estimativas de β_{10}, β_{11}, β_{12}, β_{20}, β_{21}, β_{22} para o modelo proposto sob o Cenário 5. A linha tracejada indica o valor verdadeiro do parâmetro.



Fonte: Elaboração própria (2025).

Figura 36 – Boxplots das estimativas de σ_1^2, σ_2^2 e p_1 para o modelo proposto sob o Cenário 5. A linha tracejada indica o valor verdadeiro do parâmetro.



Fonte: Elaboração própria (2025).

4.3 SIMULAÇÃO 2

Nesta seção, investigamos a capacidade do modelo proposto em agrupar as observações, ou seja, alocá-las em grupos que são semelhantes em certo sentido. Sabemos que cada ponto de dado pertence a uma das g populações heterogêneas, mas não sabemos como discriminá-las. A modelagem por meio de modelos de mistura permite o agrupamento dos dados em termos da probabilidade estimada (a posteriori) de que um ponto específico pertença a um determinado grupo. Conforme descrito na Seção **3.4.3** utilizamos os valores de $\hat{z}_{ij}$ (probabilidades a posteriori) para definir, de acordo com o algoritmo EM (etapa E, em especial), a qual grupo cada observação pertence. Assim, consideramos que a observação i foi alocada ao grupo j quando sua probabilidade a posteriori é maior para esse grupo. Na Tabela 8, apresentamos algumas medidas para os cenários estudados na Seção 4.2; para fins de comparação, consideramos o maior tamanho amostral que os cenários têm em comum.

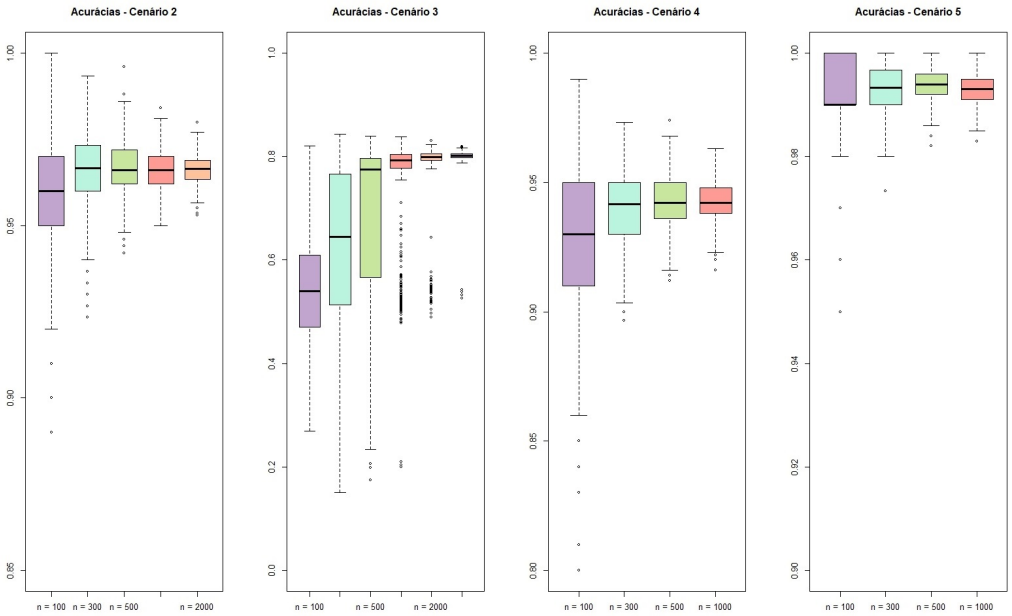
No primeiro cenário, com grupos totalmente separados, o algoritmo classifica corretamente todas as observações. Os cenários 2, 4 e 5, todos com grupos moderadamente separados, apresentam ótimo desempenho, com acurácias médias acima de 0.925 para $n = 100$ e acima de 0.943 para $n = 1000$, indicando robustez na classificação mesmo sob sobreposições moderadas. Já o cenário 3, caracterizado por alta sobreposição entre os grupos, apresenta desempenho inferior, com acurácia média de aproximadamente 0.541 para $n = 100$, melhorando para 0.744 com $n = 1000$. Esses resultados refletem a maior dificuldade do modelo em distinguir os grupos nesse cenário. Observa-se ainda que o aumento do tamanho amostral reduz a variabilidade (menor desvio padrão) e melhora, ainda que modestamente, a acurácia na maioria dos cenários.

A Figura 37 ilustra a distribuição das acurácias para cada tamanho amostral nos diferentes cenários, destacando que, dependendo do grau de sobreposição entre os grupos, os desempenhos variam consideravelmente. Em cenários mal separados, como o cenário 3, a acurácia atinge no máximo 80%, enquanto em cenários com sobreposição leve, o algoritmo consegue classificar corretamente mais de 90% das observações. Esses resultados reforçam que o modelo proposto, juntamente com o algoritmo EM desenvolvido, é eficaz para tarefas de classificação em contextos de heterogeneidade leve a moderada.

Tabela 8 – Alocações corretas das 500 réplicas de cada cenário (Média das alocações corretas ($\bar{x}$), Desvio padrão das alocações corretas (SD) e Média das acurárias ($\bar{x}_{ac}$)

	n=100			n=1000		
Cenário	$\bar{x}$	SD	$\bar{x}_{ac}$	$\bar{x}$	SD	$\bar{x}_{ac}$
C1	100.000	0.000	1.000	1000.000	0.000	1.000
C2	95.962	1.985	0.960	965.942	5.968	0.966
C3	54.134	10.031	0.541	744.254	111.560	0.744
C4	92.466	3.963	0.925	942.602	7.790	0.943
C5	99.242	0.893	0.992	993.222	2.722	0.993

Figura 37 – Distribuição das acurácias das classificações para cada tamanho de amostra dos cenários.



Fonte: Elaboração própria (2025).

5 APLICAÇÕES

Neste capítulo, aplicamos a metodologia de misturas finitas de modelos parcialmente lineares, com estimação via P-splines, a conjuntos de dados reais, demonstrando sua utilidade prática em cenários heterogêneos. Especificamente, analisamos dados sobre diabetes, inspirados no estudo de Reaven e Miller (1979), os quais envolvem medidas de tolerância à glicose em 145 indivíduos classificados em tipos de diabetes manifesta e química (Seção 5.1). Em seguida, examinamos o conjunto *Prestige*, composto por 102 observações relativas ao prestígio das ocupações no Canadá, considerando variáveis como educação média e renda, inspirado em abordagens como as de Hwang, Seo e Oh (2025) (Seção 5.2). Por fim, exploramos o conjunto *Boston Housing*, formado por 506 setores censitários de Boston, focando nas relações entre preços de moradias e fatores socioeconômicos e ambientais, conforme estudos como os de Ibacache-Pulgar e Paula (2013) e Lopes (2025) (Seção 5.3). Essas aplicações evidenciam a flexibilidade do modelo para capturar subgrupos latentes e efeitos não lineares, com verificação do ajuste realizada por meio de envelopes simulados baseados em resíduos quantílicos.

5.1 ESTUDO SOBRE DIABETES

Inspirados no trabalho de Reaven e Miller (1979), que investigou as relações entre dois tipos de diabetes, diabetes mellitus manifesta (Overt) e diabetes química (Chemical), o conjunto de dados analisado compreende 145 indivíduos submetidos a testes de tolerância à glicose. As variáveis metabólicas coletadas incluem:

- SSPG (*Steady State Plasma Glucose Response*): medida de resistência à insulina, adotada como variável resposta neste estudo;
- Glutest: área sob a curva de glicose plasmática durante o teste, modelada como efeito linear neste estudo;
- Instest: área sob a curva de insulina plasmática, tratada como componente não paramétrico, dada a possível relação não linear com a variável resposta;

Em Reaven e Miller (1979), os indivíduos foram classificados clinicamente em três grupos: *Normal* (76), *Chemical* (36) e *Overt* (33), com base em critérios médicos que consideram apenas aspectos parciais do metabolismo. No entanto, uma análise alternativa baseada em agrupamento, utilizando as técnicas de Friedman e Rubin (1967), redesenhou essa estrutura latente a partir das variáveis SSPG, Glutest e Instest, redistribuindo os indivíduos nos grupos Normal (84), Chemical (35) e Overt (26). Essa discrepância entre as classificações ressalta a potencial inadequação de critérios clínicos isolados para capturar padrões metabólicos complexos.

Revisitamos esse conjunto de dados e o analisamos no contexto de misturas de modelos parcialmente lineares. O conjunto de dados está disponível publicamente sob o título *Glucose Tolerance and Diabetes Development* na plataforma Kaggle, o que possibilita a replicabilidade dos resultados.

Na Figura 38, apresentamos o histograma da variável SSPG, observe que ela apresenta uma assimetria à direita. Com base nos diagramas de dispersão tridimensionais apresentados na Figura 39, realizados como parte da análise exploratória, observa-se uma associação não linear entre a variável *Instest* e a resistência à insulina (*SSPG*), enquanto a variável *Glutest* apresenta uma relação aproximadamente linear com *SSPG*. Diante dessas evidências empíricas e considerando a classificação clínica dos pacientes em três grupos, optou-se por ajustar um modelo de mistura finita com três componentes, utilizando modelos parcialmente lineares (PLM) para capturar tanto os efeitos lineares quanto os não lineares nas diferentes subpopulações latentes.

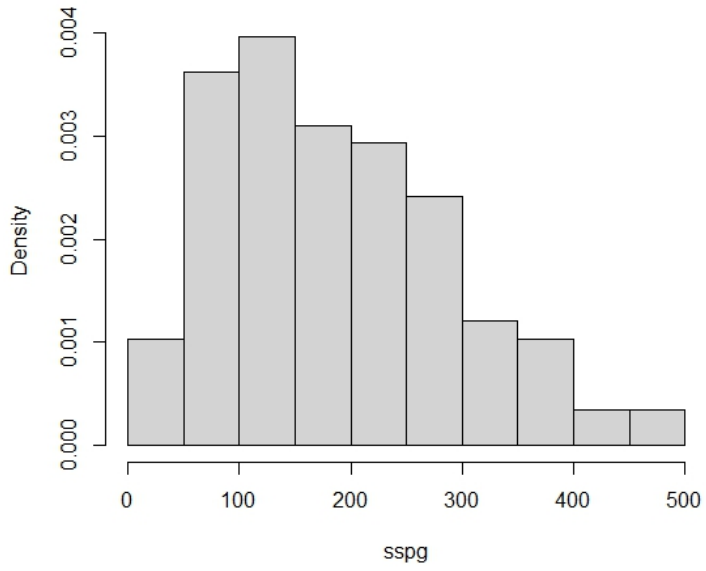
O modelo considerado, condicional à atribuição do indivíduo i ao grupo j (isto é, $Z_{ij} = 1$), é definido por:

$$\text{SSPG}_i = \beta_{1j} \cdot \text{Glutest}_i + g_j(\text{Instest}_i) + \epsilon_{ij},$$

em que $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$, para $i = 1, \ldots, 145$ e $j = 1, 2, 3$, correspondendo aos três subgrupos da mistura. Para a construção das matrizes de base **N** e de penalização **K**, foram utilizados $k = 5$ nós, visando garantir flexibilidade adequada no ajuste dos efeitos não lineares.

Dividiu-se o conjunto de dados em dois subconjuntos: treinamento e teste. Para a etapa de treinamento, foram sorteadas 116 observações (80% do total), distribuídas em 28 indivíduos do grupo *Overt*, 29 do grupo *Chemical* e 59 do grupo *Normal*, com o objetivo de ajustar o modelo. As 29 observações restantes foram reservadas para a etapa de teste, sendo utilizadas para avaliar a capacidade de classificação do modelo proposto.

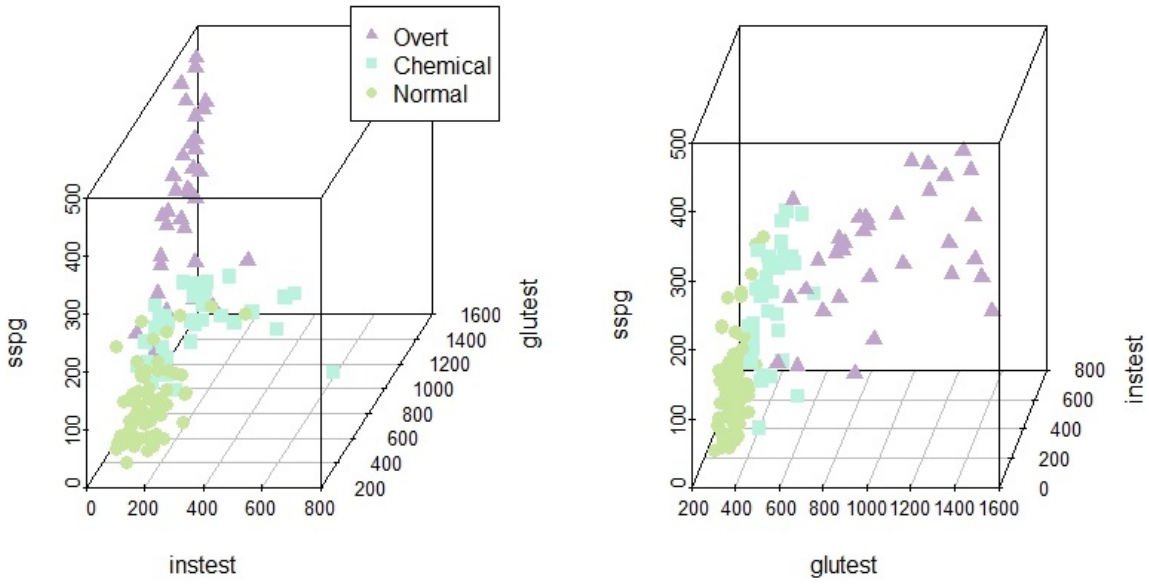
Figura 38 – Histograma de densidade da variável SSPG.



Fonte: Elaboração própria (2025).

O ajuste do modelo na base de treinamento foi realizado utilizando o algoritmo EM penalizado. Na Tabela 9 são apresentadas as estimativas dos parâmetros do modelo ajustado, incluindo as proporções dos grupos, os coeficientes lineares e as variâncias, juntamente com os erros padrão e os intervalos de confiança de 95% obtidos via *Bootstrap*. Para este conjunto de dados, a matriz de informação empírica revelou-se não invertível. Além disso, devido ao tamanho limitado da amostra, não é possível recorrer aos resultados assintóticos. Como alternativa à estimação dos erros padrão por meio da matriz de informação, foi adotado o procedimento de *Bootstrap* paramétrico, semelhante ao utilizado na construção dos envelopes simulados. A partir das reamostragens, novas estimativas

Figura 39 – Diagrama de dispersão 3D da SSPG pelas demais variáveis.



Fonte: Elaboração própria (2025).

Tabela 9 – Estimativas dos parâmetros do modelo, erros padrão obtidos via bootstrap (EP.boot) e intervalos de confiança bootstrap de 95%.

	Estimativa	EP.boot	LI	LS
p_1 (Overt)	0.060	0.161	0.043	0.567
p_2 (Chemical)	0.444	0.160	0.212	0.844
β_{11} (Overt)	1.894	0.336	0.793	2.245
β_{12} (Chemical)	1.728	0.217	1.226	2.136
β_{13} (Normal)	1.437	0.247	0.867	1.869
σ_1^2 (Overt)	90.326	449.047	0.065	$1.542 \times 10^{+03}$
σ_2^2 (Chemical)	915.439	777.625	224.131	$3.008 \times 10^{+03}$
σ_3^2 (Normal)	632.276	293.438	0.003	$9.329 \times 10^{+02}$
α_1 (Overt)	1.060	-	-	-
α_2 (Chemical)	2.960	-	-	-
α_3 (Normal)	0.001	-	-	-

dos parâmetros são obtidas, e suas variâncias empíricas são então calculadas. O desvio padrão das estimativas ao longo das $B = 500$ replicações é utilizado como medida de incerteza. Nesse sentido, o intervalo de confiança baseado nos percentis *Bootstrap* foi

obtido pelos quantis 2.5% e 97.5% das 500 réplicas ordenadas. Para uma apresentação detalhada da técnica, recomenda-se a leitura do capítulo 13 do livro de Efron e Tibshirani (1994). Especificamente, para as funções de suavização estimadas g_j, calculadas como $\hat{g}_j(t) = N_j\hat{\gamma}_j$, o erro padrão associado é dado por:

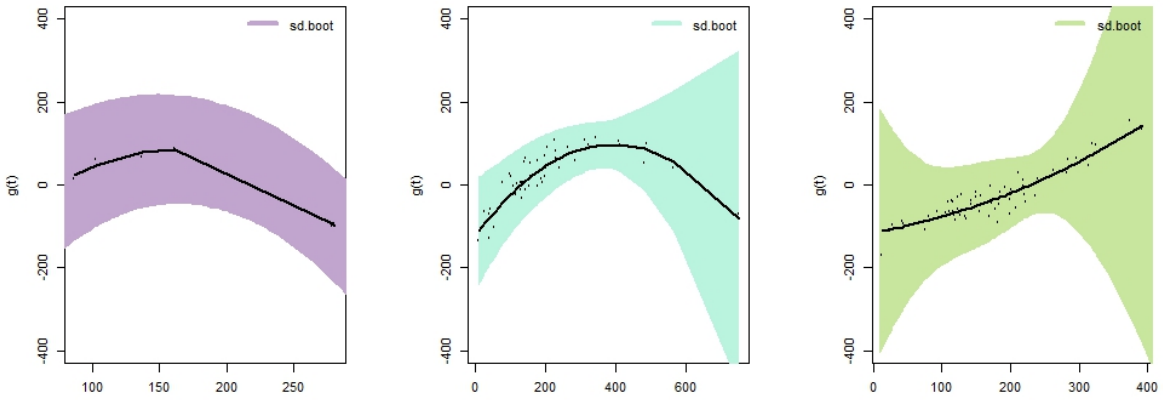
$$\text{ep}_{\text{curva},j}(t) = \sqrt{N_j \cdot \text{Cov}(\hat{\gamma}_j) \cdot N_j^\top}.$$

As bandas de confiança pontuais foram construídas considerando a forma:

$$\hat{g}_j(t) \pm 1{,}96 \cdot \text{ep}_{\text{curva},j}(t),$$

permitindo uma avaliação visual da incerteza associada às curvas estimadas para cada subgrupo.

Figura 40 – Curvas estimadas sobre os pontos referentes aos resíduos não paramétricos de cada grupo (painel à esquerda: Overt, painel central: Chemical e painel à direita: Normal).



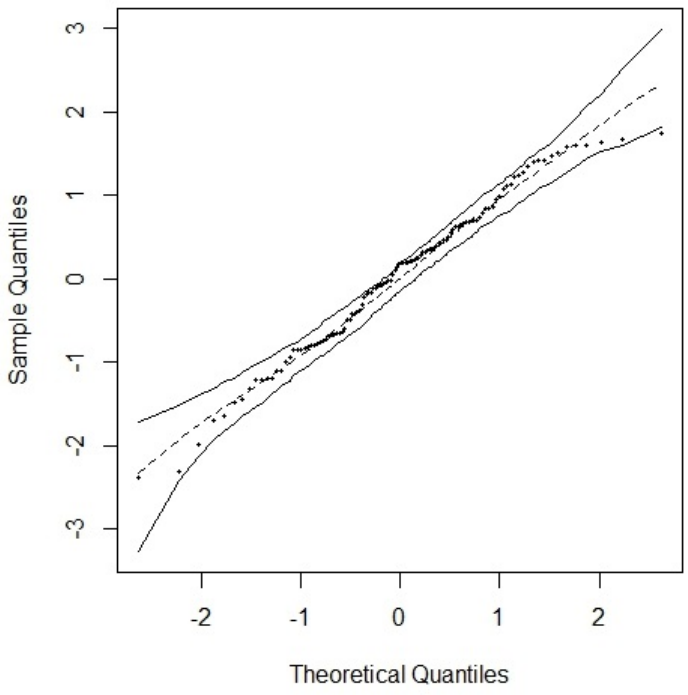
Fonte: Elaboração própria (2025).

As curvas estimadas das funções não paramétricas encontram-se ilustradas na Figura 40, acompanhadas das respectivas bandas de confiança. Observa-se que, para os indivíduos classificados no grupo *Overt*, o pico da resistência à insulina (*SSPG*) ocorre em torno do valor 150 da variável *Instest*, indicando uma resposta aguda em níveis moderados de insulina. Já para o grupo *Chemical*, o pico acontece próximo a 400, sugerindo uma resistência mais pronunciada em níveis elevados de insulina. Além disso, os coeficientes associados à variável *Glutest* são positivos para todos os grupos, indicando que um aumento na intolerância à glicose está relacionado a uma maior severidade da resistência à insulina, resultado consistente com o conhecimento fisiológico sobre diabetes. No conjunto de treinamento, o grupo Overt é composto por 28 observações (24.14%). No entanto, a estimativa para sua proporção, $p_1 = 0.060$ (6%), representa uma subestimação significativa,

possivelmente refletindo sobreposição entre subgrupos, conforme sugerido na Figura 39. Essa discrepância pode indicar dificuldade do modelo em distinguir corretamente indivíduos com características semelhantes entre os grupos minoritários. Para o grupo Chemical, que compreende 29 observações (25%) na amostra, a estimativa $p_2 = 0.444$ (44.4%) configura uma superestimação expressiva da sua prevalência. Esse resultado sugere que o modelo pode estar realocando observações de outros grupos, especialmente do grupo Overt, para o grupo Chemical. Já para o grupo Normal, com 59 observações (50.86%), a estimativa de proporção $p_3 = 0.496$ está bastante próxima da frequência observada. Isso sugere que o modelo é capaz de capturar adequadamente a estrutura do grupo majoritário, refletindo maior estabilidade na alocação de indivíduos a esse componente.

A qualidade do ajuste do modelo foi avaliada via envelope simulado baseado em resíduos quantílicos randomizados (Figura 41). Considerando um nível de significância $\alpha = 5\%$ para a construção das bandas de confiança, aproximadamente 4.3% dos pontos observados ficaram fora dessas bandas, o que indica um bom ajuste do modelo, sem violações sistemáticas das suas suposições.

Figura 41 – Envelope simulado com alfa de 5%, baseado em 500 réplicas.



Fonte: Elaboração própria (2025).

A Tabela 10 apresenta a matriz de confusão que compara as classificações clínicas, consideradas como a verdade fundamental, com as classificações obtidas com base nas probabilidades a posteriori (Seção **3.4.3**) calculadas pelo modelo de misturas finitas de modelos parcialmente lineares com P-splines. Como indicado na Tabela 10, o modelo classificou corretamente 20 das 29 observações, resultando em uma acurácia global de 68.97%.

Tabela 10 – Comparação entre classificações clínicas e classificações do modelo via algoritmo EM.

	Classificação Clínica		
Classificação EM	Overt	Chemical	Normal
Overt	1	0	0
Chemical	3	5	3
Normal	1	2	14

Fonte: Elaboração própria (2025).

Além da acurácia global, que corresponde à proporção total de classificações corretas, a acurácia balanceada é uma métrica importante especialmente em contextos com classes desbalanceadas. Ela é definida como a média da acurácia obtida separadamente em cada classe, ou seja, calcula-se a proporção de acertos dentro de cada grupo verdadeiro, e em seguida faz-se a média dessas proporções. No caso analisado, a acurácia balanceada foi de 57.93%, refletindo que, embora a acurácia global tenha sido razoável (68.97%), o desempenho do modelo variou entre os subgrupos. Esse tipo de medida é particularmente relevante em contextos clínicos, como o diagnóstico de tipos de diabetes, onde o erro em classes minoritárias pode ter implicações mais sérias do que em classes majoritárias. Para uma interpretação mais detalhada do desempenho do modelo no conjunto de teste, foram calculadas métricas por classe, sensibilidade (ou recall), precisão e F1-score (Izbicki, 2020) com base na distribuição observada das classes, conforme a classificação clínica de referência. Os resultados revelam padrões distintos de desempenho entre os subgrupos. Para o grupo Overt, o modelo apresentou desempenho bastante limitado. A sensibilidade foi de apenas 20%, indicando que a maioria dos indivíduos clinicamente classificados como Overt não foi corretamente identificada. Em contraste, a precisão foi de 100%, ou seja, todas as observações classificadas como Overt pelo modelo pertenciam de fato a esse grupo, o que se deve ao número extremamente reduzido de classificações atribuídas a essa classe. O F1-score, que sintetiza essas duas métricas por meio de uma média harmônica, foi de aproximadamente 33.3%, refletindo a grande dificuldade do modelo em recuperar corretamente os indivíduos desse subgrupo. Esse resultado está alinhado com a subestimação da proporção estimada para esse grupo, $\hat{p}_1(\text{Overt}) = 0.060$. Para o grupo Chemical, o modelo teve desempenho moderado. A sensibilidade foi de 71.4%,

indicando que a maior parte dos indivíduos com esse perfil foi corretamente identificada. No entanto, a precisão foi de apenas 45.4%, o que revela que muitas das observações classificadas como Chemical pertenciam, na verdade, a outros grupos. Como resultado, o F1-score foi de cerca de 55.6%, sugerindo uma capacidade intermediária do modelo para identificar corretamente esse subgrupo em meio à sobreposição com os demais. Por outro lado, para o grupo Normal, o desempenho do modelo foi elevado e equilibrado. Tanto a sensibilidade quanto a precisão foram de aproximadamente 82.3%, resultando em um F1-score também de 82.3%. Isso demonstra que o modelo foi eficaz em identificar os indivíduos normais com consistência, o que é compatível com sua maior representatividade no conjunto de dados e com a boa estimativa da proporção associada a esse subgrupo (Tabela 9). Em resumo, os resultados indicam que o modelo foi eficaz em capturar a estrutura do grupo majoritário (Normal), teve desempenho intermediário para o grupo Chemical e desempenho insatisfatório para o grupo Overt, especialmente devido à sua baixa sensibilidade. Esses achados reforçam a importância da avaliação por métricas específicas por classe, sobretudo em contextos desbalanceados e sensíveis, como o clínico.

5.1.1 Sugestões para Melhorias

Na configuração atual do modelo proposto, observamos que o Critério de Informação Bayesiano (BIC) seleciona uma solução com apenas uma componente (g = 1). Apesar disso, a escolha de três componentes é clinicamente justificável com base na classificação prévia dos pacientes. Entretanto, como visto nos resultados apresentados anteriormente, essa especificação resultou em estimativas imprecisas.

Algumas estratégias podem ser consideradas para aprimorar o desempenho do modelo, especialmente no que se refere à capacidade de classificação entre os subgrupos:

- *Covariáveis adicionais :* A inclusão de variáveis como idade, índice de massa corporal (IMC) ou outros marcadores metabólicos, se disponíveis, pode melhorar a separação entre os grupos Overt e Chemical. Tais variáveis podem capturar dimensões adicionais da heterogeneidade entre os indivíduos, auxiliando na discriminação dos subgrupos.

- *Balanceamento das classes:* Caso o foco seja exclusivamente na tarefa de classificação (sem interesse na interpretação dos coeficientes estimados), técnicas de reamostragem, como *oversampling* da classe minoritária ou *undersampling* da majoritária, podem ser aplicadas para reduzir o impacto do desbalanceamento entre os grupos. Essa abordagem pode melhorar métricas como sensibilidade e F1-score para as classes menos representadas.

- *Validação cruzada:* A adoção de uma abordagem de validação cruzada do tipo *k-fold* (por exemplo, com $k = 5$) permitiria uma avaliação mais robusta da capacidade preditiva do modelo, mitigando a dependência de uma única divisão entre conjunto

de treino e teste. Essa estratégia também aumenta o tamanho efetivo do conjunto de teste ao longo das iterações, tornando as métricas avaliadas mais estáveis e confiáveis.

As sugestões apresentadas, incluindo o balanceamento das classes e a adoção da validação cruzada k-fold, serão avaliadas em trabalhos futuros. Essa análise aprofundada visa aprimorar a capacidade preditiva do modelo, melhorando a discriminação entre os subgrupos clínicos, especialmente em contextos com dados desbalanceados. Uma proposta, tanto do ponto de vista inferencial quanto preditivo, seria incorporar covariáveis na modelagem das proporções, estendendo o modelo proposto neste trabalho e seguindo a linha de pesquisa das misturas de especialistas em modelos de regressão, conforme sugerido em Hunter e Young (2012), e Hwang et al. (2025), por exemplo.

5.2 ESTUDO SOBRE PRESTÍGIO OCUPACIONAL

Nesta seção, analisamos o conjunto de dados *Prestige*, disponível no pacote `car` do software R (R CORE TEAM, 2024). Esse conjunto já foi explorado anteriormente no contexto de misturas de modelos de regressão por Hwang, Seo e Oh (2025). Ele contém 102 observações, cada uma representando uma ocupação, e inclui as seguintes variáveis:

- *Prestige*: escore de prestígio ocupacional segundo a escala de Pineo-Porter, adotada como variável resposta neste estudo;
- *Edu*: número médio de anos de escolaridade dos trabalhadores em 1971, modelada como efeito linear neste estudo;
- *Income*: renda média padronizada dos trabalhadores em 1971, tratada como componente não paramétrico, dada a possível relação não linear com a variável resposta, neste estudo;
- Tipo de ocupação: classificada como Profissional (Prof), Colarinho Branco (WC) ou Colarinho Azul (BC).

Cabe destacar que quatro observações não possuem informação sobre o tipo ocupacional. Entre as demais, 31 são classificadas como pertencentes ao grupo Profissional, 23 ao grupo Colarinho Branco e 44 ao grupo Colarinho Azul.

Revisitamos os modelos de mistura de regressão ao conjunto de dados *Prestige*, no contexto de modelos parcialmente lineares via P-splines, com o objetivo de identificar subgrupos latentes entre as ocupações, adotando uma abordagem não supervisionada. Apesar do tipo de ocupação (Profissional, Colarinho Branco ou Colarinho Azul) estar disponível para a maioria das observações, essa informação não foi utilizada no processo de ajuste do modelo. Essa estratégia permite que a segmentação ocorra de forma puramente exploratória, avaliando a capacidade do modelo em recuperar padrões estruturais presentes nos dados sem depender de conhecimento prévio. Além disso, buscou-se compreender como as covariáveis, particularmente educação e renda, se associam ao prestígio dentro de cada grupo identificado. A avaliação da qualidade da classificação foi realizada a posteriori, por meio da comparação entre as alocações inferidas e os rótulos de ocupação disponíveis.

O histograma de densidade da variável *prestige* apresentado na Figura 42 não revela sinais de multimodalidade ou assimetria, algo que pudessem sugerir a existência dos três grupos. Na Figura 43, apresentamos os diagramas de dispersão tridimensionais entre as variáveis *prestige*, *income* e *edu*. A análise exploratória sugere que as relações entre *prestige* e as covariáveis variam de acordo com os subgrupos ocupacionais. Observa-se uma associação não linear entre *prestige* e *income*, enquanto a relação entre *prestige* e *edu* aparenta ser aproximadamente linear. Há indícios de que, especificamente para o grupo

Colarinho Azul, a relação entre *prestige* e *income* pode ser mais próxima de linearidade, ao contrário do observado nos demais grupos.

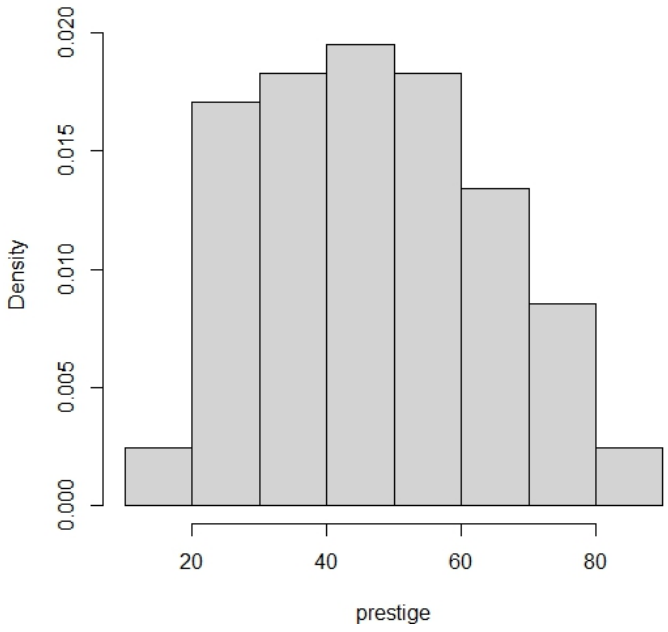
Inicialmente, seguindo a proposta de Hwang, Seo e Oh (2025), consideramos uma mistura com três componentes de modelos parcialmente lineares. Condicionalmente a $Z_{ij} = 1$, o modelo assumido é dado por:

$$\text{Prestige}_i = \beta_{01} + \beta_{1j} \cdot \text{Edu}_i + g_j(\text{Income}_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2),$$

em que $i = 1, \ldots, 102$ e $j = 1, 2, 3$. Nesse modelo, β_j representa o efeito linear da escolaridade média (*Edu*) no prestígio ocupacional, enquanto $g_j(\cdot)$ representa o efeito não paramétrico da renda média (*Income*) para cada componente da mistura. Para a construção das matrizes de base N e de penalização K, foram utilizados $k = 8$ nós, visando garantir flexibilidade adequada no ajuste dos efeitos não lineares.

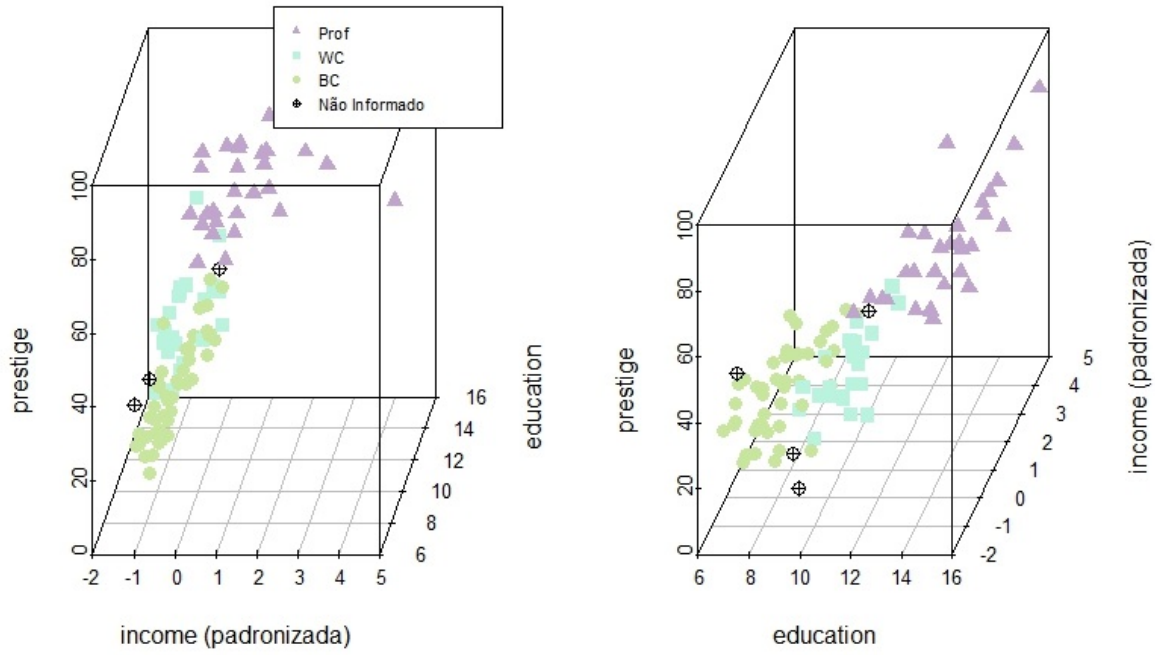
O conjunto de dados foi dividido em dois subconjuntos: treinamento e teste, sendo alocadas 82 observações para o ajuste do modelo (25 em Prof, 17 em WC e 36 em BC) e 20 observações para a avaliação da classificação. Entretanto, durante o processo de seleção aleatória, verificou-se que uma das observações destinadas ao conjunto de teste não possuía o rótulo de classe original. Dessa forma, conforme apresentado na Tabela 12, a avaliação final foi realizada com 19 observações válidas.

Figura 42 – Histograma de densidade da variável *prestige*.



Fonte: Elaboração própria (2025).

Figura 43 – Diagrama de dispersão 3D do Prestígio das ocupações pelas demais variáveis.



Fonte: Elaboração própria (2025).

As curvas estimadas das funções não paramétricas encontram-se representadas na Figura 44, acompanhadas de suas respectivas bandas de confiança. Essas curvas ilustram a relação entre a variável explicativa *income* (renda padronizada) e a variável de interesse *prestige*, permitindo avaliar o comportamento específico dessa associação em cada subgrupo identificado pelo modelo. No grupo Profissional (painel à esquerda), observa-se uma relação não linear marcada por uma forma aproximadamente parabólica. Para valores negativos de renda padronizada ($\text{income}_{\text{padronizado}} < 0$), ou seja, rendas abaixo da média, o prestígio decresce à medida que a renda se aproxima da média, atingindo um valor mínimo em torno de zero. A partir desse ponto, para rendas positivas ($\text{income}_{\text{padronizado}} > 0$), o prestígio passa a crescer com a renda, até atingir um platô próximo de $\text{income}_{\text{padronizado}} \approx 2{,}5$. Esse comportamento sugere um efeito de saturação, no qual aumentos adicionais na renda, em profissões de elevado prestígio, deixam de produzir aumentos proporcionais no reconhecimento social associado à ocupação. No grupo Colarinho Branco (painel central), o padrão estimado para a função não paramétrica apresenta uma trajetória diferente. Inicialmente, a curva cresce quase linearmente com a renda padronizada, mas a partir de aproximadamente $\text{income}_{\text{padronizado}} \approx 1{,}5$, observa-se uma queda no prestígio.

Tabela 11 – Estimativas dos parâmetros do modelo, erros padrão obtidos via bootstrap (EP.boot) e intervalos de confiança *Bootstrap* de 95%.

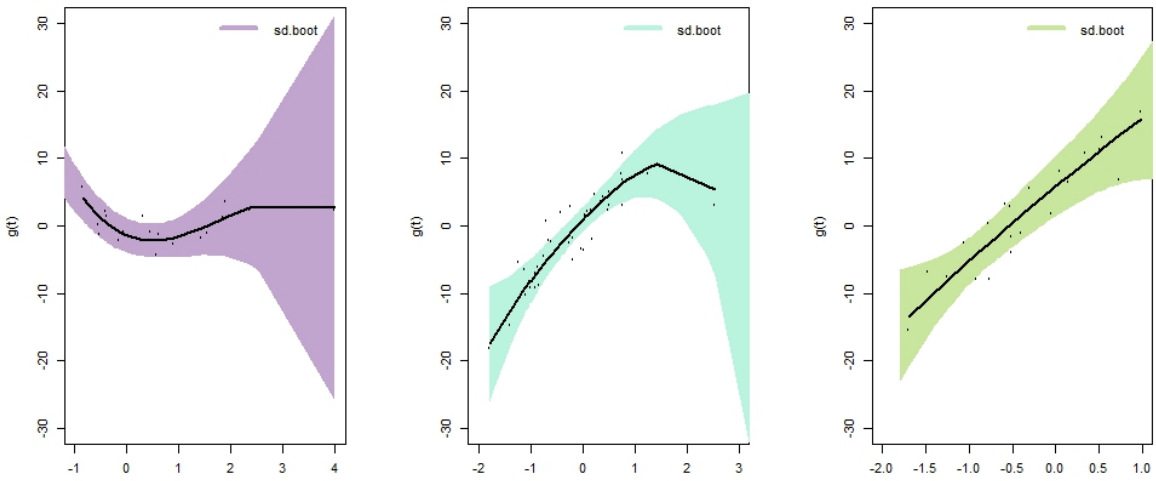
	Estimativa	Ep.boot	LI	LS
p_1 (Prof)	0.228	0.063	0.131	0.380
p_2 (WC)	0.481	0.092	0.273	0.630
β_{01} (Prof)	6.289	4.134	-0.346	15.094
β_{11} (Prof)	4.677	0.333	3.942	5.188
β_{02} (WC)	15.374	4.360	6.960	24.031
β_{12} (WC)	3.033	0.395	2.189	3.714
β_{03} (BC)	11.091	5.461	-1.459	20.142
β_{13} (BC)	1.995	0.498	1.039	3.003
σ_1^2 (Prof)	2.733	1.220	0.149	5.182
σ_2^2 (WC)	8.630	3.296	1.882	14.524
σ_3^2 (BC)	11.077	4.454	1.313	19.009
α(Prof)	1.240	-	-	-
α(WC)	0.600	-	-	-
α(BC)	19.840	-	-	-

Essa inflexão pode refletir fenômenos sociais nos quais rendas muito elevadas não são necessariamente percebidas como indicativo de maior prestígio. Em alguns contextos, profissões altamente remuneradas podem carregar conotações negativas, por exemplo, ocupações vistas como excessivamente comerciais ou distantes de funções tradicionalmente valorizadas. Alternativamente, esse padrão também pode decorrer de limitações no ajuste do modelo em regiões com menor densidade de observações, o que reduz a precisão da estimativa. Por fim, no grupo Colarinho Azul (painel à direita), a função estimada é praticamente linear, sugerindo que a relação entre renda padronizada e prestígio é constante e positiva ao longo de toda a faixa observada de renda. Neste caso, a utilização de P-splines, embora válida, pode ser considerada desnecessária, uma vez que uma especificação puramente linear seria suficiente para capturar o efeito observado. Essa simplificação reduziria a complexidade do modelo e minimizaria o risco de sobreajuste. Em síntese, os resultados mostram que o modelo de mistura com funções parcialmente lineares é capaz de captar relações diferenciadas entre *income* e *prestige* para cada subgrupo latente, revelando estruturas que não seriam identificadas por modelos homogêneos.

A qualidade do ajuste do modelo foi avaliada por meio de um envelope simulado baseado em resíduos quantílicos randomizados (Figura 45). Com um nível de significância de $\alpha = 5\%$ para a construção das bandas de confiança, observou-se que todos os pontos ficaram dentro dessas bandas, indicando um bom ajuste do modelo.

A Tabela 12 apresenta a matriz de confusão que compara as classificações originais de tipo ocupacional, consideradas como referência, com aquelas inferidas a partir das

Figura 44 – Curvas estimadas sobre os pontos referentes aos resíduos não paramétricos de cada grupo (painel à esquerda: Prof, painel central: WC e painel à direita: BC).



Fonte: Elaboração própria (2025).

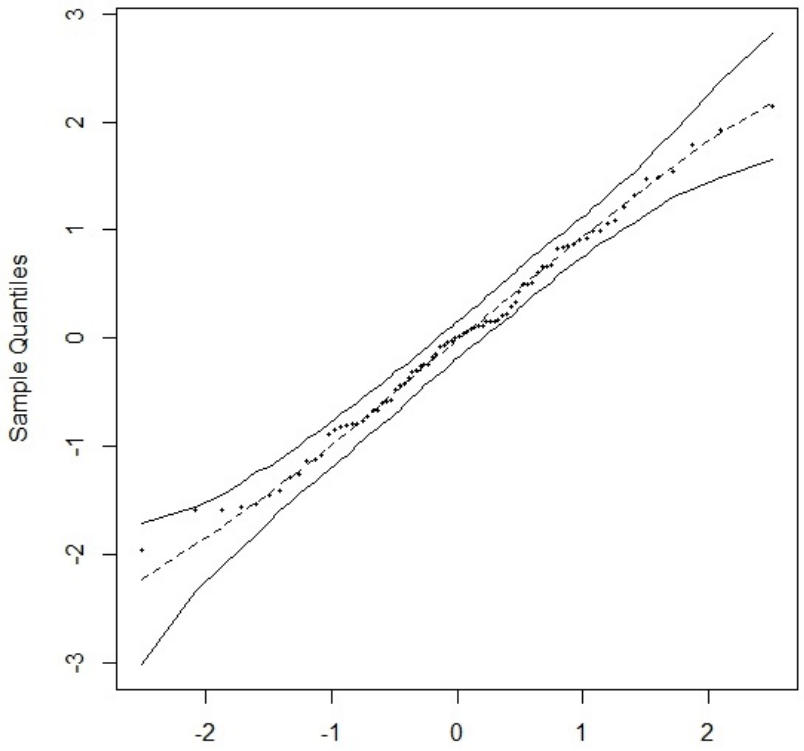
Tabela 12 – Comparação entre classificações clínicas e classificações do modelo via algoritmo EM.

	Classificação Verdadeira		
Classificação EM	Prof	WC	BC
Prof	3	0	0
WC	2	5	8
BC	1	0	0

Fonte: Elaboração própria (2025).

probabilidades a posteriori estimadas pelo modelo de mistura de regressões parcialmente lineares com P-splines. Essa matriz contempla 19 observações do conjunto de teste, dado que uma das 20 originalmente sorteadas não possuía rótulo conhecido. Com base nessa matriz, observa-se que o modelo classificou corretamente 8 das 19 observações, resultando em uma acurácia global de aproximadamente 42,1%. Embora esse valor indique desempenho modesto na tarefa de classificação, ele deve ser interpretado à luz do caráter não supervisionado do ajuste, no qual os rótulos verdadeiros não foram utilizados durante o treinamento do modelo. Além da acurácia global, foi calculada a acurácia balanceada, uma métrica apropriada para contextos com distribuição desigual entre as classes. A acurácia balanceada foi de aproximadamente 50%. Esse desempenho reflete a heterogeneidade entre os subgrupos: o modelo teve acerto perfeito para o grupo Colarinho Branco, acurácia intermediária para o grupo Profissional (50%), e não conseguiu identificar corretamente

Figura 45 – Envelope simulado com alfa de 5%, baseado em 500 réplicas.



Fonte: Elaboração própria (2025).

nenhuma observação do grupo Colarinho Azul.

A análise detalhada por classe evidencia essa heterogeneidade. Para o grupo Profissional, a sensibilidade foi de 50% (3 acertos em 6 observações reais), enquanto a precisão atingiu 100% (3 acertos em 3 classificações atribuídas), resultando em um F1-score de aproximadamente 66,7%. No grupo Colarinho Branco, o modelo obteve sensibilidade perfeita de 100% (5 em 5), mas a precisão foi baixa, cerca de 33,3% (5 acertos em 15 classificações atribuídas), levando a um F1-score próximo a 50%. Isso indica que, embora tenha identificado corretamente todos os indivíduos desse grupo, houve uma superestimação significativa, com muitas observações de outros grupos sendo classificadas como Colarinho Branco. Já para o grupo Colarinho Azul, o desempenho foi nulo, com sensibilidade, precisão e F1-score iguais a zero.

Esses resultados sugerem que o modelo, ao basear-se exclusivamente nas covariáveis educação e renda para captar padrões latentes, teve dificuldade em distinguir perfis ocupacionais que apresentam sobreposição de características, especialmente entre os grupos Colarinho Branco e os demais. A tendência em superestimar o grupo Colarinho Branco

indica que observações com perfis menos definidos acabam sendo alocadas incorretamente nesse grupo, comprometendo a precisão da segmentação.

Por fim, apesar das limitações na classificação, destaca-se o ganho inferencial proporcionado pelo modelo, que permite interpretar os coeficientes e explorar as relações não lineares específicas de cada grupo por meio das curvas estimadas, conforme discutido anteriormente.

As mesmas estratégias propostas para aprimorar o desempenho do modelo na Aplicação 1 podem ser consideradas também neste contexto, especialmente no que diz respeito à capacidade de classificação entre os subgrupos. Essas abordagens podem contribuir para uma melhor distinção dos perfis latentes e, consequentemente, para uma segmentação mais precisa dos tipos ocupacionais.

No trabalho de Hwang, Seo e Oh (2025), o ajuste e a avaliação da classificação foram realizados considerando a totalidade das observações disponíveis. Para mensurar o desempenho da clusterização do algoritmo proposto, foram utilizadas as métricas *Adjusted Rand Index* (ARI) e *Adjusted Mutual Information* (AMI) Vinh, Epps e Bailey (2009). Em nosso estudo, baseado apenas na amostra de teste, os valores obtidos para ARI e AMI foram 0.1813 e 0.1937, respectivamente, calculados com as funções `adjustedRandIndex` (pacote `mclust`) e `AMI` (pacote `aricode`) no R (R CORE TEAM, 2024). Por sua vez, Hwang, Seo e Oh (2025) reportaram resultados superiores para o modelo *Mixture of Partially Linear Experts*, com ARI de 0.4779 e AMI de 0.4506, considerando o conjunto completo de dados. Esses valores indicam que o modelo é capaz de capturar grande parte da estrutura latente presente nos dados, ainda que haja margem para melhorias. O desempenho mais elevado está relacionado ao uso do contexto de especialistas, que modelam as proporções via covariáveis, contribuindo para uma melhor capacidade classificatória. Além disso, os autores apresentaram resultados para outros modelos comparativos: *Mixture of Experts* (ARI 0.5096; AMI 0.4012) e *Finite Mixture of Partially Linear Regressions* (ARI 0.0597; AMI 0.0725). Em trabalhos futuros, pretendemos explorar extensões do modelo proposto no contexto de especialistas via P-splines, com o objetivo de aprimorar ainda mais a performance da classificação.

5.3 ESTUDO SOBRE PREÇO DE IMÓVEIS (BOSTON HOUSING)

O conjunto de dados *Boston Housing*, possui informações de 506 setores censitários da cidade de Boston, incluindo variáveis socioeconômicas, demográficas e ambientais que influenciam o valor médio das residências. Entre as principais variáveis estão o número médio de cômodos por residência, a taxa de imposto predial, a distância até centros empregatícios. Ele está disponível no pacote `mlbench` do R (R CORE TEAM, 2024) e já foi analisado em outras pesquisas, como Lopes (2025) e Ibacache-Pulgar e Paula (2013), nas quais consideram como resposta o logaritmo do valor médio de casas ocupadas pelos proprietários, em milhares de dólares (US$ 1000), e assumem que as observações provêm de uma única população.

Vamos introduzir os modelos de mistura de regressão ao conjunto de dados *Boston Housing*, no contexto de modelos parcialmente lineares via P-splines, com o objetivo de identificar subgrupos latentes entre os setores censitários, adotando uma abordagem não supervisionada. Essa estratégia permite que a segmentação ocorra de forma puramente exploratória, avaliando a capacidade do modelo em recuperar padrões estruturais presentes nos dados, sem depender de conhecimento prévio. Além disso, buscou-se compreender como as covariáveis se associam ao valor médio de imóveis dentro de cada grupo identificado. Condicionalmente a $Z_{ij} = 1$, o modelo assumido para G componentes é dado por:

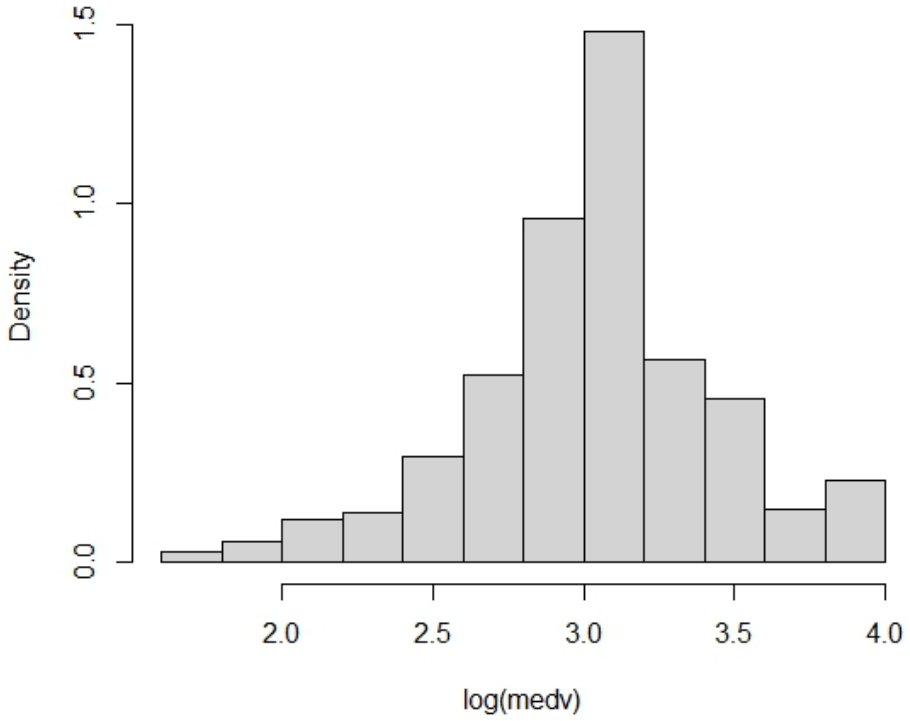
$$log(medv_i) = \beta_{0j} + \beta_{1j} rm_i + \beta_{2j} tax_i + g_j(dis_i) + \varepsilon_{ij}, \tag{5.1}$$

em que $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$; $j = 1, ..., G$; $i = 1, \cdots, 506$;

- **medv**: valor médio de casas ocupadas pelos proprietários em US$ 1.000;
- **rm**: número médio de cômodos por residência;
- **tax**: taxa de imposto predial total por US$ 10.000.
- **dis**: distâncias ponderadas até cinco centros empregadores de Boston.

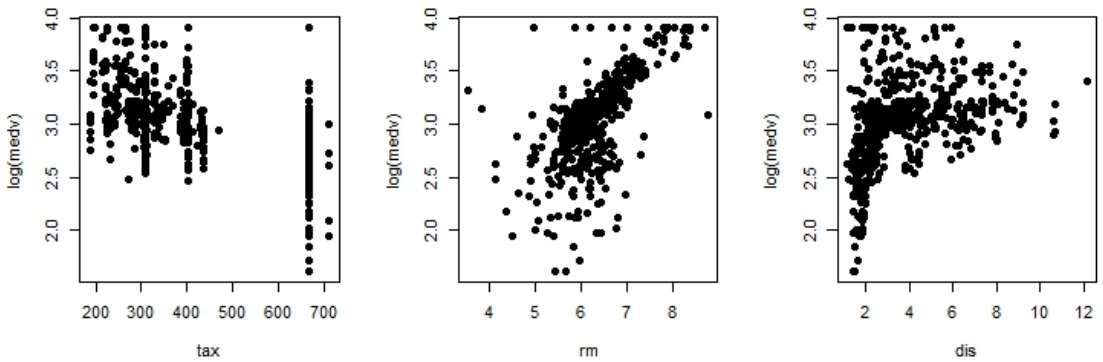
O histograma de densidade da variável *log(medv)* está apresentado na Figura 46, ele exibe uma leve assimetria à esquerda. Na Figura 47, apresentamos os diagramas de dispersão da resposta em função das covariáveis consideradas, como parte da análise exploratória. As variáveis *tax* e *rm* parecem estar associadas com a reposta de forma linear e a variável *dis* aparentemente está associada com a resposta de forma não linear. O modelo 5.4 foi ajustado considerando $G = \{1, 2, 3 \text{ e } 4\}$, para selecionar o que melhor se adequa aos dados utilizamos o Critério de Informação Bayesiano (BIC). As matrizes $\boldsymbol{N}$ e $\boldsymbol{K}$ foram construídas pela função *smoothCon* considerando k=10.

Figura 46 – Histograma de densidade da variável *log(medv)*.



Fonte: Elaboração própria (2025).

Figura 47 – Gráficos de dispersão da resposta pelas covariáveis: tax(painel à esquerda), rm(painel central) e dis(painel à direita).



Fonte: Elaboração própria (2025).

Inicialmente, com base nos resultados da Tabela 13, consideramos o modelo dado pela expressão 5.4 com três componentes. Para este conjunto de dados, a matriz de informação empírica revelou-se não invertível. Como alternativa realizamos réplicas *Bootstrap*, os intervalos de confiança foram construídos com a normalidade assintótica dos Estimadores de Máxima Verossimilhança.

Tabela 13 – Seleção do número de grupos via Critério de Informação Bayesiano (BIC).

Número de grupos (G)	BIC	Parâmetros de suavização
1	144.466	$(\alpha_1 = 0.128)$
2	38.318	$(\alpha_1 = 2.020,\ \alpha_2 = 0.113226)$
3	**8.253**	$(\boldsymbol{\alpha_1 = 1.999,\ \alpha_2 = 1.529,\ \alpha_3 = 0.263})$
4	35.883	$(\alpha_1 = 2.783,\ \alpha_2 = 4.039,\ \alpha_3 = 7.008,\ \alpha_4 = 9.898\ )$

Fonte: Elaboração própria (2025).

As estimativas com essa configuração se encontram na Tabela 14, bem como os erros padrão baseados em 500 réplicas *Bootstrap*, os intervalos de confiança assintóticos de 95%, o p-valor pela estatística Wald e a significância estatística ($\cdot = 0.1$; $* = 0.05$; $** = 0.01$; $*** = 0.001$). Os parâmetros β_{21} e β_{22} associados a variável *tax* para os grupos G1 e G2 não são estatísticamente significativos, indicando que o imposto predial não é importante para explicar o valor dos imóveis nesses setores.

Então, para garantir que apenas as covariáveis significativas ao nível de 5% estejam presentes na estrura de cada grupo, realizamos uma seleção de variáveis baseada no p-valor. Note que, os parâmetros β_{21} e β_{22} são os que apresentaram maior p-valor dentro dos grupos G1 e G2, respectivamente. Pensando nisso, o novo ajuste vai retirar essa covariável da estrutura desses grupos. De forma que, o modelo assumido para cada componente é:

Condicionalmente a $Z_{i1} = 1$

$$log(medv_i) = \beta_{01} + \beta_{11} rm_i + g_1(dis_i) + \varepsilon_{i1}, \tag{5.2}$$

onde $\varepsilon_{i1} \sim \mathcal{N}(0, \sigma_1^2);\ i = 1, \cdots, 506;$

Condicionalmente a $Z_{i2} = 1$:

$$log(medv_i) = \beta_{02} + \beta_{12} rm_i + g_2(dis_i) + \varepsilon_{i2}, \tag{5.3}$$

onde $\varepsilon_{i2} \sim \mathcal{N}(0, \sigma_2^2);\ i = 1, \cdots, 506;$

Condicionalmente a $Z_{i3} = 1$:

$$log(medv_i) = \beta_{03} + \beta_{13} rm_i + \beta_{23} tax_i + g_3(dis_i) + \varepsilon_{i3}, \tag{5.4}$$

onde $\varepsilon_{i3} \sim \mathcal{N}(0, \sigma_3^2);\ i = 1, \cdots, 506;$

Tabela 14 – Estimativas dos parâmetros do modelo, erros padrão obtidos via bootstrap (EP.boot), intervalos de confiança assintóticos de 95%, p-valor pela estatística Wald e significância.

	Estimativa	EP.boot	LI	LS	p.valor	sig.
p_1 (G1)	0.175	0.092	0.000	0.355	0.028	*
p_2 (G2)	0.523	0.209	0.113	0.933	0.006	**
β_{01} (G1)	1.190	9.623	-17.671	20.050	**0.451**	
β_{11} (G1)	0.315	0.494	-0.652	1.283	**0.261**	
β_{21} (G1)	0.000	0.027	-0.053	0.053	**0.498**	
β_{02} (G2)	0.969	0.709	-0.420	2.357	0.086	.
β_{12} (G2)	0.370	0.056	0.259	0.480	2.58×10^{-11}	***
β_{22} (G2)	-0.001	0.001	-0.002	4.88×10^{-4}	**0.156**	
β_{03} (G3)	1.833	0.368	1.111	2.555	3.26×10^{-7}	***
β_{13} (G3)	0.291	0.038	0.216	0.365	7.58×10^{-15}	***
β_{23} (G3)	-0.002	0.001	-0.003	-0.001	1.30×10^{-4}	***
σ_1^2 (G1)	0.035	0.012	0.012	0.058	0.001	**
σ_2^2 (G2)	0.010	0.019	0.000	0.047	**0.304**	
σ_3^2 (G3)	0.029	0.012	0.005	0.052	0.008	**
α_1 (G1)	1.999	-	-	-	-	-
α_2 (G2)	1.529	-	-	-	-	-
α_3 (G3)	0.263	-	-	-	-	-

Os grupos G1 e G2 incluem o intercepto e as variáveis *rm* e *dis*. Já o grupo G3 permanece com intercepto, *rm*, *tax* e *dis*. Os resultados desse ajuste estão indicados na Tabela 15. Observe que, com a nova configuração, todos os parâmetros estimados são estatisticamente significativos ao nivel de 5%, nenhum dos intervalos de confiança incluem o zero. Além disso, o valor do BIC para esse modelo foi de -3132.885. Indicando que esse modelo é mais adequado para explicar os dados do que o anterior.

Em todos os grupos formados com base nas características dos setores, observa-se uma associação positiva entre o número de cômodos (rm) e o valor médio dos imóveis. Isso sugere que, independentemente do perfil do setor, imóveis com mais cômodos tendem a apresentar maior valorização.

No entanto, para os setores pertencentes ao Grupo G3, identifica-se uma associação negativa entre a taxa de imposto predial e o valor dos imóveis. Em outras palavras, nesses setores, imóveis com impostos mais elevados tendem a ser menos valorizados

A Figura 48 mostra que a relação entre a variável dis (distância aos centros empregatícios) e o valor médio dos imóveis varia de acordo com o subgrupo:

- **G1**: Tendência decrescente. Considerando as demais variáveis constantes, o valor dos imóveis diminui, em média, com o aumento da distância;

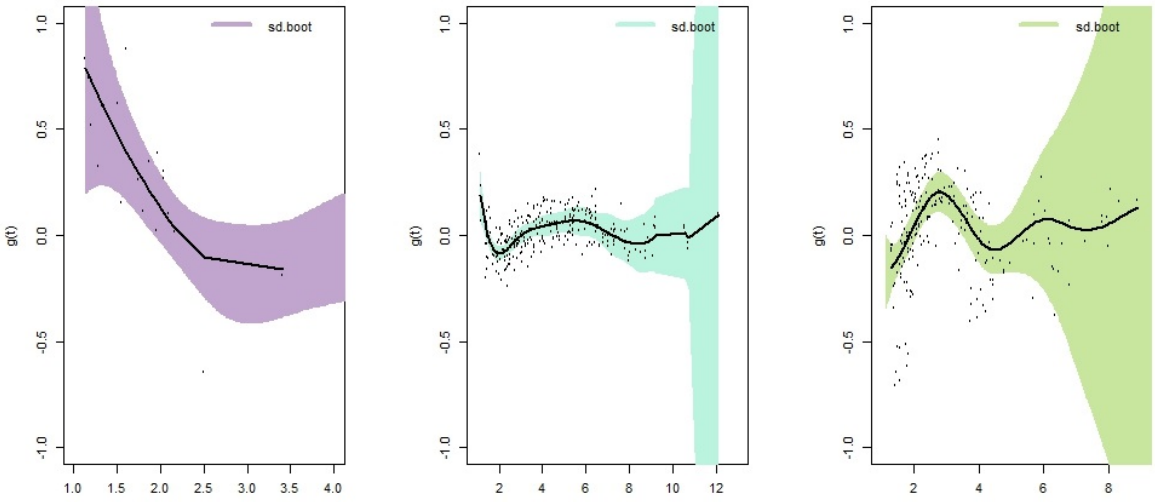
- **G2**: Comparado com os outros grupos, a curva que modela o relacionamento apresenta valores de menor magnitude. Pode indicar que para alguns setores censitários a distância não impacta muito na valorização ou desvalorização do imóvel;
- **G3**: Tendência crescente até aproximadamente 2 unidades da variável distância. Considerando as demais variáveis constantes, o valor dos imóveis aumenta, em média, com o aumento da distância.

Tabela 15 – Estimativas dos parâmetros do novo modelo, erros padrão obtidos via bootstrap (EP.boot) e intervalos de confiança assintóticos de 95%.

	Estimativa	EP.boot	LI	LS
p_1 (G1)	0.102	0.033	0.038	0.166
p_2 (G2)	0.403	0.047	0.312	0.495
β_{01} (G1)	1.449	0.572	0.328	2.571
β_{11} (G1)	0.277	0.090	0.101	0.453
β_{02} (G2)	0.548	0.093	0.366	0.730
β_{12} (G2)	0.410	0.015	0.381	0.438
β_{03} (G3)	1.763	0.177	1.416	2.109
β_{13} (G3)	0.285	0.025	0.236	0.333
β_{23} (G3)	-0.002	1.40×10^{-4}	-0.002	-0.001
σ_1^2 (G1)	0.041	0.014	0.014	0.068
σ_2^2 (G2)	0.008	0.002	0.004	0.011
σ_3^2 (G3)	0.035	0.005	0.026	0.044
α_1 (G1)	1.668	-	-	-
α_2 (G2)	0.676	-	-	-
α_3 (G3)	1.362	-	-	-

Fonte: Elaboração própria (2025).

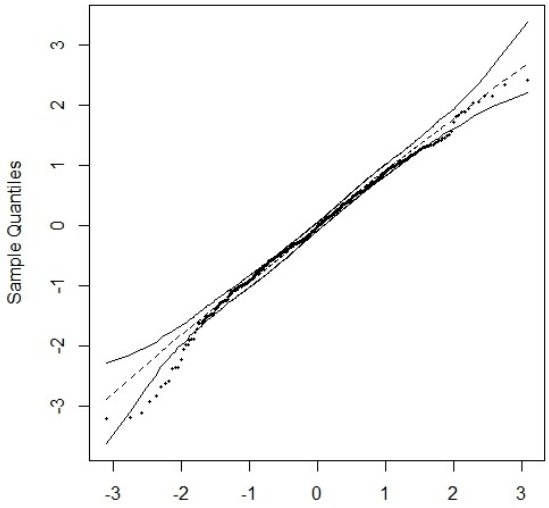
Figura 48 – Curvas estimadas para o relacionamento entre a distância e o valor do imóvel, pontos se referem aos resíduos não paramétricos de cada grupo (G1: painel à esquerda, G2: painel do centro, G3: painel à direita).



Fonte: Elaboração própria (2025).

A qualidade do ajuste do modelo foi avaliada por meio de um envelope simulado baseado em resíduos quantílicos randomizados (Figura 49). Com um nível de significância de $\alpha = 5\%$ para a construção das bandas de confiança. Cerca de 6.7% dos pontos ficaram fora, isso indica que pode haver alguma característica nos dados que o modelo não conseguiu capturar. Em trabalhos futuros, é possível explorar extensões do modelo proposto considerando distribuições pertencentes a família Skew Scale Mixtures of Normals (SSMN), elas são úteis para modelar dados com caudas pesadas e/ou assimétricos, propondo uma análise mais robusta.

Figura 49 – Envelope simulado com alfa de 5%, baseado em 500 réplicas.



Fonte: Elaboração própria (2025).

5.4 ASPECTOS COMPUTACIONAIS

Durante o procedimento de *Bootstrap*, tanto para a obtenção dos erros padrão quanto para a construção dos envelopes simulados, algumas reamostragens apresentaram falhas de convergência, sendo descartadas por não atenderem ao critério de parada do algoritmo. O número de exclusões foi monitorado e reportado.

Na primeira aplicação, realizamos 753 reamostragens com apenas 425 bem-sucedidas para o cálculo dos erros padrão e construção dos intervalos de confiança (253 apresentaram falhas de convergência e 75 apresentaram estimativas discrepantes para algum parâmetro de escala). Consideramos discrepantes estimavas classificadas como *outliers* pela função `boxplot` no R (R CORE TEAM, 2024) e as estimativas com valores menores que 1^{-10} . Na construção do envelope simulado (Figura 41), consideramos as 500 reamostras que não apresentaram falha de convergência.

Já na segunda aplicação, realizamos 516 réplicas com 390 bem-sucedidas para o cálculo dos erros padrão (16 apresentaram falhas de convergência e 110 apresentaram estimativas discrepantes para algum parâmetro de escala). Na construção do envelope simulado (Figura 45), consideramos as 500 reamostras que não apresentaram falha de convergência.

E na terceira aplicação, todas as reamostragens foram bem-sucedidas.

6 CONCLUSÃO

Neste trabalho, exploramos misturas finitas de modelos parcialmente lineares, utilizando P-splines para a estimação das componentes não paramétricas. Essa abordagem permite incorporar heterogeneidade em dados provenientes de subpopulações latentes, possibilitando a estimação de parâmetros específicos de cada componente, bem como das probabilidades a posteriori para fins de classificação.

Os estimadores de máxima verossimilhança penalizada foram obtidos por meio do algoritmo EM, com erros padrão calculados a partir da matriz de informação empírica. A seleção dos parâmetros de suavização e do número de grupos foi baseada no critério BIC.

Estudos de simulação, abrangendo cinco cenários com distintos graus de separação entre grupos e diferentes estruturas semiparamétricas, demonstraram a consistência assintótica dos estimadores. Observou-se que, com o aumento do tamanho amostral, as estimativas se aproximam dos valores verdadeiros, com redução do viés e da variabilidade, mesmo em configurações complexas, como grupos pouco separados ou com covariáveis lineares e não lineares distintas. A recuperação das curvas não paramétricas foi satisfatória, conforme indicado pelos ASEs decrescentes, e o desempenho em classificação mostrou-se robusto, sobretudo em cenários com moderada separação entre os grupos.

Nas aplicações a dados reais, incluindo conjuntos sobre diabetes (REAVEN; MILLER, 1979), prestígio ocupacional (FOX, 2022) e preços de imóveis em Boston (HARRISON; RUBINFELD, 1978), o modelo proposto apresentou bom ajuste, verificado por envelopes simulados baseados em resíduos quantílicos. As classificações obtidas refletiram padrões coerentes com as características conhecidas dos dados, evidenciando a utilidade prática da metodologia em contextos heterogêneos.

Apesar dos avanços, algumas limitações ainda merecem destaque, como a sensibilidade a valores iniciais no algoritmo EM, embora, neste trabalho, tenhamos proposto uma estratégia para sua definição e avaliado seu desempenho por meio de estudos de simulação, com resultados satisfatórios. Outras limitações incluem o custo computacional em grandes amostras e as dificuldades na inversão da matriz de informação empírica, especialmente no contexto de dados reais. Para trabalhos futuros, sugerem-se extensões para misturas de especialistas, modelando as proporções dos grupos em função de covariáveis; a incorporação de assimetria ou caudas pesadas na distribuição dos erros; bem como a integração com técnicas de aprendizado de máquina voltadas à seleção automática de variáveis. Adicionalmente, a adoção de uma abordagem bayesiana pode ser interessante, permitindo a comparação entre diferentes paradigmas inferenciais.

7 REFERÊNCIAS

1 BYRD, R. H.; LU, P.; NOCEDAL, J.; ZHU, C. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing, Philadelphia, v. 16, n. 5, p. 1190-1208, 1995.

2 DE BOOR, Carl. A practical guide to splines. Revised edition. New York: Springer, 2001. 346 p. (Applied mathematical sciences; v. 27).

3 DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), v. 39, n. 1, p. 1-38, 1977.

4 DUNN, P. K.; SMYTH, G. K. Randomized Quantile Residuals. Journal of Computational and Graphical Statistics, v. 5, n. 3, p. 236-244, set. 1996.

5 EFRON, Bradley; TIBSHIRANI, Robert J. An introduction to the bootstrap. New York: Chapman & Hall/CRC, 1994. 456 p. (Monographs on Statistics and Applied Probability)

6 EILERS, P. H. C.; MARX, B. D. Flexible smoothing with B-splines and penalties. Statistical Science, v. 11, n. 2, p. 89–121, maio 1996.

7 ENGLE, R. F., GRANGER, C. W., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. Journal of the American statistical Association, 81(394):310–320

8 FERREIRA, C. DA S.; MONTORIL, M. H.; PAULA, G. A. Partially linear models with p-order autoregressive skew-normal errors. Brazilian Journal of Probability and Statistics, v. 36, n. 4, p. 792–806, dez. 2022.

9 FRIEDMAN, H. P.; RUBIN, J. On some invariant criteria for grouping data. Journal of the American Statistical Association, v. 62, n. 320, p. 1159-1178, 1967.

10 FOX, J.; WEISBERG, S. carData: Companion to Applied Regression Data Sets. v. 3.0-5 [software], 2022. Dispon. em: https://CRAN.R-project.org/package=carData. Acesso em: 17 ago. 2025.

11 GREEN, P. J.; SILVERMAN, B. W. Nonparametric regression and generalized linear models: a roughness penalty approach. 1st ed. London: Chapman & Hall, 1994. 192 p. (Monographs on statistics and applied probability; 58).

12 HARTIGAN, J. A.; WONG, M. A. Algorithm AS 136: A k-means clustering algorithm. Applied Statistics, v. 28, n. 1, p. 100-108, 1979.

13 HARRISON, D.; RUBINFELD, D. L. Hedonic prices and the demand for clean air. Journal of Environmental Economics and Management, v. 5, p. 81-102, 1978.

14 HUBERT, L.; ARABIE, P. Comparing partitions. Journal of Classification, New York, v. 2, n. 1, p. 193–218, 1985.

15 HUNTER, D. R.; YOUNG, D. S. Semiparametric mixtures of regressions. Journal of Nonparametric Statistics, v. 24, n. 1, p. 19-38, 2012.

16 HWANG, Y.; SEO, B.; OH, S. Mixture of Partially Linear Experts. Stat, v. 14, n. 2, p. e70062, 2025.

17 IBACACHE-PULGAR, G.; PAULA, G. A.; CYSNEIROS, F. J. A. Modelos aditivos semiparamétricos sob distribuições simétricas. TEST, v. 22, p. 103-121, 2013.

18 IZBICKI, R.; DOS SANTOS, T.M. Aprendizado de máquina: uma abordagem estatística. 1 ed. São Carlos: Câmara Brasileira do Livro, 2020.

19 JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An introduction to statistical learning: with applications in R. 2. ed. New York: Springer, 2021. 607 p. (Springer Texts in Statistics).

20 LIN, T. I. Robust mixture modeling using multivariate skew t distributions. Statistics and Computing, v. 20, p. 343–356, 2010.

21 LOPES, J. S. (2025). Seleção de Modelos Aditivos Parcialmente Lineares (Trabalho de Conclusão de Curso). Universidade Federal de Juiz de Fora.

22 MCLACHLAN, G.; PEEL, D. Finite Mixture Models. New York: Wiley, 2000.

23 MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, G. Geoffrey. Introduction to linear regression analysis. 5th ed. Hoboken: Wiley, 2012.

24 R CORE TEAM. R: A Language and Environment for Statistical Computing. Vienna, Áustria: R Foundation for Statistical Computing, 2024. Disponível em: <https://www.R-project.org/>. Acesso em: 15 ago. 2025.

25 REAVEN, G. M.; MILLER, R. G. An attempt to define the nature of chemical diabetes using a multidimensional analysis. Diabetologia, Berlin, v. 16, n. 1, p. 17-24, 1979.

26 REAVEN, G. M.; MILLER, R. G. Glucose Tolerance and Diabetes Development [Conjunto de dados]. Kaggle, s.d. Disponível em: `https://www.kaggle.com/datasets/utkarshx27/diabetes-dataset`. Acesso em: 14 ago. 2025.

27 SKHOSANA, S. B.; MILLARD, S. M.; KANFER, F. H. J. A novel EM-type algorithm to estimate semi-parametric mixtures of partially linear models. Mathematics, v. 11, n. 5, p. 1087, 2023.

28 SCHWARZ, Gideon. Estimating the dimension of a model. The Annals of Statistics, [S.l.], v. 6, n. 2, p. 461–464, 1978.

29 VINH, N. X.; EPPS, J.; BAILEY, J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. Journal of Machine Learning Research, v. 11, p. 2837–2854, 2010.

30 WOOD, S. N. Generalized additive models: an introduction with R. 2. ed. Boca Raton: CRC Press, 2017. 496 p.

31 YAO, W.; XIANG, S. Mixture models: parametric, semiparametric, and new directions. Boca Raton: Chapman & Hall/CRC, 2024.

32 ZHANG, Yi; PAN, Weiquan. Estimation and inference for mixture of partially linear additive models. Communications in Statistics - Theory and Methods, v. 51, n. 8, p. 2519-2533, 2020.

33 ZELLER, C. B.; CABRAL, C. R. B.; LACHOS, V. H. Robust mixture regression modeling based on scale mixtures of skew-normal distributions. TEST, v. 25, p. 375-396, 2015.

34 ZELLER, C. B.; CABRAL, C. R. B.; LACHOS, V. H.; et al. Finite mixture of regression models for censored data based on scale mixtures of normal distributions. Advances in Data Analysis and Classification, v. 13, p. 89-116, 2018.