UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Métodos de estimação de modelos de mistura para dados com Distribuição Poisson

Claudio Henrique Leão de Almeida

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

ICMC USP SÃO CARLOS

Claudio Henrique Leão de Almeida

Métodos de estimação de modelos de mistura para dados com Distribuição Poisson

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

Área de Concentração: Estatística

Orientador: Prof. Dr. Luis Aparecido Milan

USP – São Carlos
Setembro de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

A447m

Almeida, Claudio Henrique Leão de
Métodos de estimação de modelos de mistura para dados com Distribuição Poisson / Claudio Henrique Leão de Almeida; orientador Luis Aparecido Milan. -- São Carlos, 2024.
70 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Modelo de Mistura. 2. Metropolis Hasting. 3. Algoritmo EM. I. Milan, Luis Aparecido, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Claudio Henrique Leão de Almeida

Dissertation submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Master in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Luis Aparecido Milan

USP – São Carlos
September 2024

Dedico este trabalho aos meus pais,
Maria Aparecida Leão de Almeida e José de Almeida.

AGRADECIMENTOS

Agradeço primeiramente à Deus, porque a fé com certeza sempre foi e é um conforto em todos os momentos. E também por Ele me ter feito alcançar coisas que com sozinho não conseguiria.

Agradeço aos meus pais, pois sempre foram os meus maiores admiradores. Sempre fizeram dos meus sonhos, os deles. Me mostraram o que é um amor incondicional. Obrigado por me transformarem em quem sou hoje. Essa vitória foi muito mais fácil com vocês ao meu lado. Nós conseguimos!

Agradeço aos meus irmãos, pela união e amor recíproco que temos. Também sou grato à toda a minha família, por sempre torcerem pelo meu sucesso.

Agradeço ao meu orientador Prof. Luís Aparecido Milan pela excelente parceira, por todo ensinamento e paciência.

Agradeço a todos os professores que passaram pela minha vida desde o Ensino Fundamental. Muitos enxergaram esse capítulo na minha história, e me incentivaram a segui-lo.

Agradeço aos meus amigos, em especial, a Edvaldo Coelho, Maria Luiza Matos e Renan Vinícius Rodrigues, por todo apoio e companheirismo durante esse período.

Ser mestre, vai muito além do que um dia imaginei que alcançaria. Todos os "eus"que já existiram dentro de mim estão muito felizes pelo Mestre Claudio.

“Quando não há inimigos internos,
os inimigos externos não podem nos ferir”.
(Provérbio Africano)

RESUMO

ALMEIDA, C. H. L. **Métodos de estimação de modelos de mistura para dados com Distribuição Poisson**. 2024.70 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Os modelos de mistura são utilizados quando os dados da população podem ser particionados em subpopulações. Essa metodologia permite-nos utilizar múltiplas distribuições de probabilidade, em que cada uma descreve o comportamento de cada subpopulação. Neste trabalho estudamos métodos de estimação de modelos de mistura para dados de contagem, com enfoque na abordagem bayesiana. Nele são apresentados dois métodos: EM (*expectation maximization algorithm*), MH (*Metropolis-Hasting*). O primeiro citado é baseado em máxima verossimilhança, já o MH inferência bayesiana. Foram realizadas aplicações utilizando os métodos EM e MH, em bancos de dados simulados sem e com a inclusão de variáveis. As metodologias também foram aplicadas em um banco de dados reis. À partir dos resultados, foi possível ter indícios de que os métodos já performam bem quando os parâmetros são próximos. E as estimativas são ainda melhores, para parâmetros distantes. Também verificou-se que a medida que o tamanho da amostra aumenta, essas estimações melhoram, o que era esperado.

Palavras-chave: Modelos de Mistura, Infêrencia Bayesiana, *expectation maximization algorithm*, *Metropolis-Hastings*.

ABSTRACT

ALMEIDA, C. H. L. . 2024.70 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Mixture models are only used when population data can be partitioned into subpopulations. This methodology allows the use of multiple probability distributions, so that each one determines the behavior of each subpopulation. In this work we study mixture model estimation methods for contagion data, focusing on the Bayesian approach. Two methods are presented here: EM (*expectation-maximization algorithm*), MH (*Metropolis-Hasting*). The first mentioned is based on maximum likelihood, there is no Bayesian inference. Applications were made using the EM and MH methods, in simulated databases with even variables. The methodologies are also applied to a real database. From two results, there are possible indications that the methods will perform well when the parameters are close. These estimates are even better for distant parameters. I also verified that as the sample size increases, these estimates are smaller, or what was expected.

Keywords: Mixture Models, Bayesian Inference, expectation maximization algorithm, Metropolis-Hasting.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A aplicação de modelos de mistura de distribuições tem estado presente em diversas áreas, desde estatística e aprendizado de máquinas, até biologia. Estes modelos oferecem uma abordagem flexível para modelar a complexidade inerente a conjuntos de dados multidimensionais, nos quais diferentes subpopulações podem coexistir. Ao incorporar uma combinação ponderada de diversas distribuições probabilísticas, os modelos de mistura de distribuições proporcionam uma representação mais realista da heterogeneidade subjacente aos dados, permitindo uma análise mais precisa e uma tomada de decisão mais informada.

Fazendo um revisão bibliográfica, podemos ver que Xiong, Liu e Tan (2006) utilizaram o modelo de mistura na área de bioinformática, para predizer estruturas secundárias de proteínas. Neste caso, a metodologia é aplicada para combinar as previsões de múltiplos algoritmos. Essa metodologia também foi utilizada em engenharia de materiais por Jin e Gao (2006), para segmentar imagens de microestruturas de materiais, permitindo a identificação de diferentes fases ou componentes em materiais complexos.

A aplicação desses modelos se estende à área da saúde, sendo empregados em psicologia para identificar subpopulações com comportamentos ou características psicológicas distintas, conforme evidenciado por McLachlan e Peel (2000), e em medicina, para análises de agrupamento de pacientes com base em múltiplas variáveis médicas, contribuindo para a identificação de subgrupos de pacientes com características clínicas semelhantes, abordado por Celeux e Govaert (1995).

Nessas aplicações, modelos de mistura finita sustentam uma variedade de técnicas nas principais áreas da estatística, incluindo análises de clusters, análise discriminante, análise de imagem e análise de sobrevivência, além de seu papel mais direto na análise de dados e inferência de fornecer modelos descritivos para distribuições onde uma única distribuição é aparentemente inadequada (MCLACHLAN; PEEL, 2000).

Com isso, surge o questionamento de como estimar os parâmetros de um modelo de

mistura de distribuições. Um dos métodos mais conhecidos para esta estimação é o algorítimo EM - *expectation maximization algorithm*, que é aplicado no contexto da estimação pelo método de máxima verossimilhança.

Além do algoritmo EM, também existem técnicas com enfoque bayesiano. Nesta classe, temos dois principais métodos. O primeiro é o MH (*Metropolis-Hastings*), que é utilizado para quando o número de subpopulações é conhecido (CHIB; GREENBERG, 1995). Já o segundo é chamado de *Reversible Jump MCMC* e é utilizado para os casos em que o número de subpopulações é desconhecido (GREEN, 1995).

Nesta dissertação, exploramos métodos de estimação de parâmetros em modelos de mistura, dando enfoque para os dados de contagem. Na bioestatística, utilizamos o modelo de mistura de *Poisson* para análise de sobrevivência e modelagem de dados censurados/truncados em estudos epidemiológicos e de saúde (HELD; PAWITAN, 2011). Na ciência da computação, Collins, Dasgupta e Schapire (2002) propõem uma extensão do PCA baseada no modelo de mistura de *Poisson* para redução dinâmica de dimensionalidade em análises de dados computacionais.

No âmbito econômico, destaca-se a aplicação do modelo de mistura de *Poisson* na modelagem da contagem de aves em séries temporais, especialmente em questões relacionadas à biodiversidade (MELARD; PASTEELS, 2000). Na engenharia, propõe-se a aplicação dinâmica do modelo de mistura de Poisson para resolver desafios em visão computacional, com ênfase na estimação de parâmetros em imagens de baixo nível (LIN; RADKE, 2003). Na ecologia, destaca-se a inovação do modelo de mistura de *Poisson* zero-inflado em estudos analíticos de colisões entre animais e veículos, visando entender dinamicamente padrões e fatores associados (GONÇALVES; MAIA; CARVALHO, 2017). Na medicina, apresenta-se o artigo de Yang e Zhou (2008), que propõe o uso dinâmico de modelos de regressão de mistura de Poisson para analisar a duração da estadia hospitalar, fornecendo suporte à tomada de decisões e ao planejamento de recursos médicos.

O objetivo dessa dissertação é apresentar métodos de estimação de parâmetros de modelos de mistura focado em dados de contagem.

No Capítulo **2**, apresentamos a estrutura e os componentes de um modelo de mistura, com enfoque principal para dados de contagem, descrevendo o que são variáveis não observáveis, o processo de estimação dos parâmetros via máxima verossimilhança mais utilizadas e por fim apresentando o algoritmo EM (*Expectatiob maximization algorithm*).

No Capítulo **3**, abordamos a estimação de parâmetros de modelos de mistura, seguindo o paradigma bayesiano, apresentando a definição, das distribuições a *priori* e *posteriori*, assim como algoritmos bayesianos para realização desta estimação.

Em seguida, no Capítulo **4** é apresentado a implementação das duas metodologias estudadas, o algorítmo EM e o método MH, com aplicação em diferentes conjuntos de dados simulados. Esta implementação foi feita pelo *software R* (R Core Team, 2020). No Capítulo **5** é

feita uma aplicação em um banco de dados reais. Por fim, no Capítulo **6** apresentamos conclusões iniciais assim como próximos passos da pesquisa.

CAPÍTULO

2

MODELOS DE MISTURA

Suponha uma população particionada em subpopulações. Para modelar esses casos, podemos utilizar os chamados modelos de mistura. A mistura de K distribuições tem maior flexibilidade do que apenas uma distribuição de probabilidade, como discutido por Mclachlan e Peel (2000).

Dizemos que a densidade $f(x|w,\lambda)$ é uma mistura de distribuições se ela for uma combinação linear de K densidades, isto é,

$$f(x|w,\lambda) = \sum_{k=1}^{K} w_k f_k(x_k|w_k,\lambda_k), \tag{2.1}$$

em que $f_k(x_k|w_k,\lambda_k)$ é a função de densidade do k-ésimo componente da mistura, $w_k > 0$, $\sum_{k=1}^{K} w_k = 1$, $\lambda = (\lambda_1,\lambda_2,...,\lambda_K)$ e $w = (w_1,w_2,...,w_K)$. w_k é o peso da subpopulação k na densidade geral e λ_k representa o conjunto de parâmetros associado a k-ésima componente.

A Figura 1 mostra um exemplo de modelo de mistura. Neste exemplo, apresentamos um modelo de mistura de Poisson. As distribuições de Poisson são utilizadas para modelar contagens de eventos raros em uma unidade de tempo ou espaço fixo. No entanto, em muitos casos reais, os dados podem ser compostos por uma mistura de diferentes geradores de contagens. Para ilustrar, geramos dados de exemplo a partir de uma mistura de duas distribuições de Poisson com diferentes parâmetros. A diferença nos parâmetros das distribuições de Poisson resulta em diferentes padrões de contagem nos dados observados. Ao plotar o histograma dos dados, podemos visualizar a composição da mistura e como ela contribui para a distribuição geral dos dados. Este exemplo demonstra como o modelo de mistura de Poisson pode ser aplicado para analisar e modelar dados que possuem uma estrutura complexa e heterogênea.

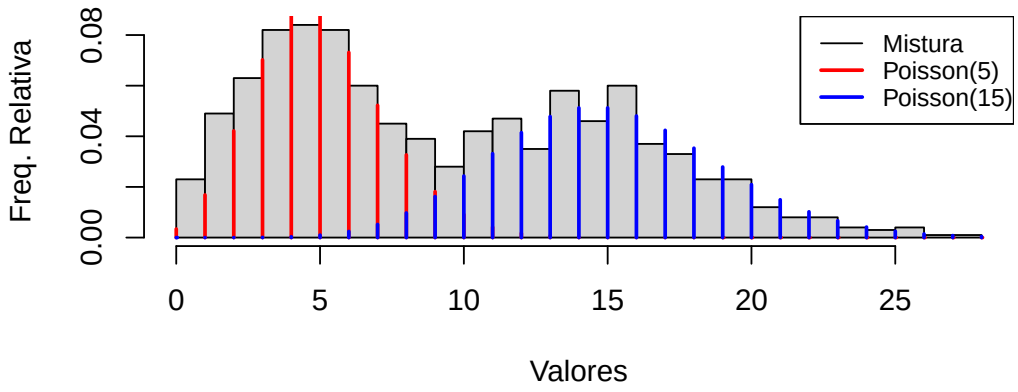


Figura 1 – Exemplo de uma mistura de duas distribuições Poisson

É importante destacar que as subpopulações utilizadas no modelo de mistura podem pertencer a distribuições diferentes.

2.0.1 Variáveis Não Observáveis

Uma das primeiras ocorrências de modelagem com mistura de distribuições ocorreu em 1887, para modelar o peso de militares franceses. Isto porque esses homens são provenientes de duas regiões diferentes: planícies e montanhas. A estrutura de mistura surge devido à indisponibilidade da origem de cada observação, ou seja, não se sabe a origem da região de cada homem. Cada peso x_i observado é proveniente da densidade f_1 (modela os pesos dos homens das planícies) ou f_2 (modela os pesos dos homens das montanhas) com probabilidades $p_1 = w_1$ e $p_2 = (1 - w_1)$, respectivamente. Esta estrutura “oculta"pode ser explorada para facilitar o procedimento de estimação dos parâmetros utilizando o fato que para toda variável aleatória, X_i, proveniente de um modelo com mistura de distribuições com k componentes, é possível associar uma variável latente $S_i = (S_{i1}, ..., S_{ik})$, de dimensão k, que indica a componente da qual a observação x_i é proveniente (SARAIVA, 2009).

Meira (2014) define X como uma sequência de valores observáveis $x = (x_1, ..., x_N)$ e S como a variável não observável na qual s_i assume valores $s_1, ..., s_N$, sendo N a quantidade total de observações. Considere também que cada s_i é independente de s_j para i diferente de j, em que $i, j \in \{1, ..., N\}$. O vetor de parâmetros para a distribuição de probabilidade ou função de densidade X é dado por $\lambda = (\lambda_1, ..., \lambda_K)$ e para a distribuição de probabilidade de S por $(1, p)$ em que $p = (p_1, ..., p_K)$.

Se considerarmos o caso em que a variável aleatória X é discreta, temos que

$$P(X_i = x_i|\lambda) = \sum_{k=1}^{K} P(S_i = k|p)P(X_i = x_i|S_i = k, \lambda, p). \tag{2.2}$$

A variável S com distribuição Multinomial e vetor de parâmetros $(1, p)$ em que $p = (p_1, ..., p_k)$ e $\sum_{k=1}^{K} p_k = 1$. Assim, é escrita a distribuição conjunta de X e S para o modelo de mistura como

$$P(X_i = x_i, S_i = s_i|\lambda, p) = P(S_i = s_i|p)P(X_i = x_i|S_i = s_i, \lambda, p). \tag{2.3}$$

Utilizando suas respectivas funções indicadoras obtém-se

$$P(X_i = x_i, S_i = s_i|\lambda, p) = \prod_{k=1}^{K} p_k^{1_{s_i}(k)} P(X_i = x_i|S_i = s_i, \lambda, p)^{1_{s_i}(k)}, \tag{2.4}$$

em que a função indicadora $1_{s_i}(k)$ assume o valor 1 se $s_i = k$ for verdadeiro e 0 caso contrário. Vale ressaltar que para cada observação i temos um valor para s_i de S e um valor para x_i de X, assim define-se os vetores $X = (X_1, ..., X_N)$ e $S = (s_1, ..., s_N)$ e calculamos a distribuição conjunta de X e S partindo da equação

$$P(X = x, S = s|\lambda, p) = \prod_{i=1}^{N} \prod_{k=1}^{K} p_k^{1_{s_i}(k)} P(X_i = x_i|S_i = s_i, \lambda_k)^{1_{s_i}(k)}. \tag{2.5}$$

2.0.2 Estimador via máxima verossimilhança

Nesta seção, foram calculadas as estimativas dos parâmetros de interesse para os modelos de mistura, λ e p, aplicando a metodologia de máxima verossimilhança.

Na Equação (2.5) os produtórios foram manipulados de forma que uma das funções indicadoras, $1_{s_i}(k)$, expresse a quantidade de observações. n_k, presente em cada componente do modelo de mistura. Observando que $n_k = \sum_{i=1}^{N} 1_{s_i}(k)$ é o número de elementos do k-ésimo componente da mistura e então reescrevemos a (2.5) da seguinte forma

$$\begin{aligned} P(X = x, S = s|\lambda, p) &= \prod_{k=1}^{K} \prod_{n=1}^{N} p_k^{1_{s_i}(k)} P(X_i = x_i|S_i = s_i, \lambda_k)^{1_{s_i}(k)} \\ &= \prod_{k=1}^{K} p_k^{\sum_{i=1}^{N} 1_{s_i}(k)} \prod_{i=1}^{N} P(X_i = x_i|S_i = s_i, \lambda_k)^{1_{s_i}(k)} \\ &= \prod_{k=1}^{K} p_k^{n_k} \prod_{i=1}^{N} P(X_i = x_i|S_i = s_i, \lambda_k)^{1_{s_i}(k)}. \end{aligned} \tag{2.6}$$

Através da função de verossimilhança, (2.6), calcula-se o ponto de máximo para cada um dos parâmetros de interesse, resultando em estimativas para λ_1,...,λ_K e $p_1, ...p_K$. Portanto, a equação (2.6) pode ser reescrita como,

$$L(\lambda, p|x,s) = \prod_{k=1}^{K} p_k^{n_k} \prod_{i=1}^{N} P(X_i = x_i|S_i = s_i, \lambda_k)^{1_{s_i}(k)}. \tag{2.7}$$

O ponto que maximiza a função de verossimilhança também maximiza o logaritmo da função de verossimilhança que é dado por

$$l(\lambda, p|x,s) = ln\Big(\prod_{k=1}^{K} p_k^{n_k} \prod_{i=1}^{N} P(X_i = x_i|S_i = s_i, \lambda_k)^{1_{s_i}(k)}\Big). \tag{2.8}$$

Devido a restrição de que $\sum_{k=1}^{K} p_k$=1, Meira (2014) utiliza multiplicadores de Lagrange para a solução da (2.8). Após, obtém-se,

$$l(\lambda, p|x,s) = \sum_{k=1}^{K} n_k ln(p_k) + \sum_{k=1}^{K}\sum_{i=1}^{N} 1_{s_i}(k) ln(P(X_i = x_i|S_i = s_i, \lambda_k)) \tag{2.9}$$
$$+\zeta(1 - \sum_{k=1}^{K} p_k),$$

em que $\zeta = \sum_{k=1}^{K} n_k$=$N$.

A derivada de (2.9) com relação a λ_k é igual a zero. Já com relação ao parâmetro p_k, encontraremos,

$$\frac{\partial l(\lambda, p|x,s)}{\partial \hat{p}_k} = \frac{n_k}{\hat{p}_k} - \zeta$$

, e,

$$\frac{n_k}{\hat{p}_k} - \zeta = 0.$$

Portanto, a estimativa de p_k é dada por,

$$\hat{p}_k = \frac{n_k}{N}.$$

Repare que as estimativas de p e λ dependem da variável não observável S e para a solução deste problema recorre-se ao algorítimo EM, o qual é apresentado a seguir.

2.1 Algoritmo EM

O algoritmo EM (*Expectation-Maximization*) é uma abordagem iterativa utilizada para estimar parâmetros em modelos estatísticos, especialmente em situações onde os dados estão incompletos ou envolvem variáveis não observáveis. Ele recebe esse nome devido aos seus dois passos fundamentais: "*expectation*"(expectativa) e "*maximization*"(maximização). No passo de

"*expectation*", o algoritmo calcula as estimativas das variáveis latentes baseadas nos valores atuais dos parâmetros. Já no passo de "*maximization*", ele atualiza os parâmetros do modelo usando as estimativas das variáveis latentes obtidas no passo anterior, juntamente com os dados observados. Esse processo é repetido iterativamente até que a convergência seja alcançada.

Além disso, o algoritmo EM possui várias extensões e remodelagens para lidar com diferentes cenários e tipos de dados que podem ser consultados em Mclachlan e Krishnan (2007).. Por exemplo, pode ser adaptado para lidar com dados censurados, truncados, ou para ajustar modelo de mistura gaussianas em problemas de clustering. Existem também variantes mais avançadas, como o EM Hierárquico e o EM-Variacional, que incorporam estruturas hierárquicas nos parâmetros do modelo ou utilizam técnicas de otimização variacional, respectivamente. Essas diversas abordagens tornam o algoritmo EM uma ferramenta poderosa e versátil em uma variedade de contextos estatísticos e de modelagem de dados.

O algoritmo EM se constrói à partir de três passos principais, que apresentaremos abaixo:

Algoritmo 1 – Algoritmo EM com variáveis não observadas

1: Inicialize $\lambda^{(0)}$
2: **para** j = 1, 2, 3, ... **faça**
3: **enquanto** $|\lambda^{(j+1)} - \lambda^{(j)}| > \tau_\lambda$ ou $|Q(\lambda^{(j+1)}, \lambda^{(j)}) - Q(\lambda^{(j)}, \lambda^{(j)})| > \tau_\lambda$ **faça**
4: Estime $Q(\lambda, \lambda^{(j)}) = E[l(\lambda)]$ com λ_i substituído por $\lambda_i^{(j)}$ ▷ Passo E
5: Escolha $\lambda^{(j+1)}$ que maximiza $Q(\lambda, \lambda^{(j)}) = E[l(\lambda)]$ em relação a λ ▷ Passo M
6: **fim enquanto**
7: **fim para**

A função densidade de probabilidade da população é desconhecida, portanto é indicado que cada subpopulação seja modelada por uma densidade pertencente à alguma família de distribuições paramétricas conhecida, facilitando assim o tratamento dos dados. É importante destacar que as subpopulações utilizadas no modelo de mistura podem pertencer a famílias de distribuições diferentes.

Dentro da metodologia de modelos de mistura, é necessário adaptar o algoritmo EM para acomodar a presença da variável não observada S, cujos valores precisam ser estimados. Para realizar a atualização de s_i, empregamos o teorema de Bayes, o teorema das probabilidades totais e a distribuição conjunta de X e S, conforme expresso na Equação 2.6, resultando na probabilidade condicional de S dado X, discutida por Meira (2014). Inicialmente, manipulamos a distribuição marginal de X, conduzindo o seguinte cálculo:

$$\begin{aligned}
&\Pr[X_i = x_i|\lambda] \\
&= \sum_{j=1}^{K} \Pr[X_i = x_i, S_i = j|\lambda, p] \\
&= \sum_{j=1}^{K} \Pr[S_i = j|p]\Pr[X_i = x_i|S_i = j, \lambda, p] \\
&= \sum_{j=1}^{K} p_j \Pr[X_i = x_i|S_i = j, \lambda, p].
\end{aligned} \tag{2.10}$$

Utilizando o teorema de Bayes e a Equação 2.3, obtemos a distribuição condicional de X dado S em:

$$\begin{aligned}
\Pr[S_i = k|X_i = x_i, \lambda, p] &= \frac{\Pr[S_i = k, X_i = x_i|\lambda, p]}{\sum_{j=1}^{K} \Pr[S_i = j, X_i = x_i|\lambda, p]} \\
&= \frac{p_k \Pr[X_i = x_i|S_i = k, \lambda, p]}{\sum_{j=1}^{K} p_j \Pr[X_i = x_i|S_i = j, \lambda, p]}.
\end{aligned} \tag{2.11}$$

O algoritmo EM padrão, Algoritmo 1, maximiza a quantidade $Q(\lambda, \lambda^{(j)})$, enquanto o algoritmo EM modificado utiliza a função de máximo argumento para determinar as probabilidades condicionais de S dado X, conforme a Equação 2.11. Em outras palavras, para obter $\hat{s}_i$, calculamos a probabilidade de o i-ésimo elemento pertencer a cada componente da mistura e selecionamos $\hat{s}_i$ como a componente com a maior probabilidade entre as K componentes. Assim, $s_i^{(j)}$ é atualizado para $\hat{s}_i^{(j)}$.

Na $(j+1)$-ésima iteração, utilizamos o valor estimado $\hat{s}_i^{(j)}$ para calcular as estimativas de máxima verossimilhança, $\hat{p}^{(j)}$ e $\hat{\lambda}^{(j)}$, dos parâmetros do modelo de mistura.

Os passos mencionados são repetidos até que ocorra convergência. Para isso, determinamos um valor para τ_p e τ_λ com base na variação máxima das estimativas dos parâmetros p e λ entre a j-ésima e a $(j+1)$-ésima iteração. Fixamos τ_p e τ_λ em 10^{-4} como tolerância máxima para essa variação.

Portanto, apresentamos o algoritmo EM modificado para o modelo de mistura, levando em consideração a probabilidade condicional de S dado X na atualização da variável não observada.

Algoritmo 2 – Algoritmo EM para Modelos de Mistura

1: Atribua um vetor de valores iniciais para $s^{(0)}$ tal que $s_i^{(0)} \in \{1,...,K\}$ para $i = 1,...,N$.
2: Inicialize $\hat{\lambda}_k^{(0)}$ e $\hat{p}_k^{(0)}$.
3: **para** j = 1, 2, 3, ... **faça**
4: **enquanto** $\max_k |\hat{p}_k^{(j)} - \hat{p}_k^{(j+1)}| > \tau_p$ ou $\max_k |\hat{\lambda}_k^{(j)} - \hat{\lambda}_k^{(j+1)}| > \tau_\lambda$ **faça** ▷ c. de parada
5: Estime $\hat{\lambda}_k^{(j)}$, Equação 2.12, e $\hat{p}_k^{(j)}$, Equação 2.13.
6: Calcule $\hat{s}_i^{(j)} = \arg\max_k \Pr[S_i = k | X_i = x_i, \hat{\lambda}, \hat{p}]$, Equação 2.15.
7: **fim enquanto**
8: **fim para**

CAPÍTULO

3

ABORDAGEM BAYESIANA

3.1 Distribuição a *priori*

A distribuição a *priori* é muito importante na inferência bayesiana. Essa distribuição deve representar, em probabilidade, o conhecimento que se tem sobre λ antes da realização do processo de inferência. Existem diversos métodos de construção/escolha de distribuição a *priori*. Para este trabalho, utilizaremos o método de distribuição a *priori* conjugada.

3.1.1 Distribuição a priori conjugada

Dada a informação que se tem sobre λ, é possível definir uma família paramétrica de densidades. Neste caso, a distribuição a *priori* é representada por uma forma funcional, na qual os parâmetros devem ser especificados mediante a este conhecimento. Estes parâmetros, indicadores da família de distribuições a *priori* são chamados de hiperparâmetros para distinguí-los dos parâmetros de interesse λ.

Em geral, esta abordagem facilita a análise e o caso mais importante é o de distribuição a *priori* conjugada. Para este, a essência é que as distribuições a *priori* e a *posteriori* pertençam a mesma classe de distribuições e assim a atualização do conhecimento que se tem de λ envolve apenas uma mudança nos hiperparâmetros. Neste caso, o aspecto sequencial do método bayesiano pode ser explorado definindo-se apenas a regra de atualização dos hiperparâmetros já que as distribuições são as mesmas. Abaixo definiremos esses conceitos matematicamente, apresentando um exemplo de como funciona a distribuição a *priori* conjugada para a distribuição Poisson, que será a principal abordada neste trabalho.

Seja F = $\{p(x|\lambda), \lambda \in \lambda\}$ uma classe de distribuições amostrais, então uma classe de distribuições P é conjugada a F se

$$\forall\ p(X|\lambda) \in F \ \text{ e } \ p(\lambda|x) \in P.$$

Exemplo 1 (Distribuição de Poisson)**.** Seja $X_1,...,X_n$ uma amostra aleatória da distribuição de Poisson com parâmetro λ. Sua função de probabilidade conjunta é dada por

$$p(x|\lambda) = \frac{\varepsilon^{-n\lambda}\lambda^t}{\prod x_i!} \propto \varepsilon^{-n\lambda}\lambda^t, t = \sum_{i=1}^n x_i.$$

O núcleo da verossimilhança é da forma $\lambda^a\varepsilon^{-b\lambda}$ que caracteriza a família de distribuições Gama. Assim, a distribuição a *priori* conjugada natural de λ é Gama com parâmetros positivos α e β. Assim sendo

$$p(\lambda) \propto \lambda^{\alpha-1}\varepsilon^{-b\lambda}, \{\alpha,\beta,\lambda\} > 0.$$

A densidade da distribuição a *posteriori* fica

$$p(\lambda|x) \propto \lambda^{\alpha+t-1}\varepsilon^{-(b+n)\lambda},$$

que corresponde à densidade Gama($\alpha+t$,$\beta+n$).

3.2 Abordagem bayesiana para quando o número de componentes é conhecido

O algoritmo de *Metropolis-Hastings* utilizam a ideia de gerar um valor de uma distribuição auxiliar e aceitá-lo com uma dada probabilidade. Este mecanismo de correção garante a convergência da cadeia para a distribuição de equilíbrio, que neste caso, é a distribuição a *posteriori* (CHIB; GREENBERG, 1995).

Suponha que a cadeia esteja no estado λ e um valor λ_0 é gerado de uma distribuição proposta q($\cdot$|λ). Note que a distribuição proposta pode depender do estado atual da cadeia, por exemplo q($\cdot$|λ) poderia ser uma distribuição normal centrada em λ. O novo valor λ_0 é aceito com probabilidade

$$\alpha(\lambda,\lambda') = \left(1, \frac{\pi(\lambda')q(\lambda|\lambda')}{\pi(\lambda)q(\lambda'|\lambda)}\right),$$

onde π é a distribuição de interesse.

O algoritmo de *Metropolis-Hastings* pode ser especificado pelos seguintes passos,

1. Inicialize o contador de iterações t=0 e especifique um valor inicial $\lambda^{(0)}$;
2. Gere um novo valor λ' da distribuição $q(\cdot|\lambda)$;
3. Calcule a probabilidade de aceitação $\alpha(\lambda,\lambda')$ e gere $\mu \sim U(0,1)$;

4. Se $\mu \geq \alpha$ então aceito o novo valor e faça $\lambda^{(t+1)} = \lambda^{'}$, caso contrário rejeito e faço $\lambda^{(t+1)} = \lambda$;

5. Incremente o contador de t para t+1 e volte ao passo 2.

Uma característica útil do algoritmo é que a distribuição alvo não precisa ser conhecida totalmente, ela só precisa ser conhecida até uma constante de proporcionalidade, uma vez que apenas a razão alvo $\frac{\pi(\lambda)}{\pi(\lambda}$ é usada na probabilidade de aceitação. Ressalta-se também que a cadeia pode permanecer no mesmo estado por muitas iterações e, na prática, um monitoramento simples e inicial que pode ser feito para evidenciar qual o modelo é melhor entre os comparados é dado pela porcentagem média de iterações para as quais os movimentos são aceitos. (EHLERS, 2007)

O algoritmo de *Metropolis-Hastings* é utilizado na estimação de parâmetros em modelos de mistura. Em um contexto de modelos de mistura, assume-se que os dados são gerados a partir de diferentes subpopulações, cada uma associada a uma distribuição de probabilidade distinta. A tarefa é estimar os parâmetros dessas distribuições, bem como as proporções das subpopulações.

Suponha que o modelo de mistura tenha K componentes, e os parâmetros a serem estimados sejam representados por $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_K)$, onde cada λ_k denota os parâmetros da k-ésima componente. Além disso, seja S um vetor latente indicando a qual componente cada observação pertence.

Os passos específicos do algoritmo Metropolis-Hastings para estimar modelos de mistura são os seguintes:

1. **Inicialização:**
   - Inicialize $t = 0$ e especifique valores iniciais para os parâmetros do modelo $\lambda^{(0)}$ e para o vetor latente $S^{(0)}$.

2. **Geração de Novo Valor:**
   - Gere novos valores para os parâmetros do modelo $\lambda^{'}$ e o vetor latente $S^{'}$ a partir de distribuições propostas adequadas.

3. **Cálculo da Função de Aceitação:**
   - Calcule a função de aceitação $\alpha(\lambda, \lambda')$ usando a verossimilhança do modelo de mistura e a função de densidade a *priori*, se aplicável.

4. **Aceitação/Rejeição:**
   - Gere uma variável aleatória $\mu \sim U(0,1)$.
   - Se $\mu \geq \alpha$, rejeite os novos valores e faça $\lambda^{(t+1)} = \lambda^{(t)}$ e $S^{(t+1)} = S^{(t)}$.

- Caso contrário, aceite os novos valores e faça $\lambda^{(t+1)} = \lambda'$ e $S^{(t+1)} = S'$.

5. **Atualização da Variável Latente:**

   - Para cada observação i no conjunto de dados:
     - Calcule a probabilidade condicional $P(s_i^{(t+1)} = k|x_i, \lambda^{(t+1)})$ para cada componente do modelo de mistura, onde x_i é a observação e $\lambda^{(t+1)}$ são os parâmetros do modelo na iteração $t+1$.
     - Atribua a $S_i^{(t+1)}$ o valor da componente correspondente amostrada de acordo com as probabilidades condicionais calculadas.
   - Repita esse processo para todas as observações no conjunto de dados.

6. **Iteração:**

   - Incremente t para $t+1$.
   - Volte ao passo 2 para continuar gerando novos valores dos parâmetros do modelo e do vetor latente.

Este processo iterativo permite explorar o espaço de parâmetros do modelo de mistura, ajustando-se gradualmente aos dados observados e convergindo para uma estimativa dos parâmetros que descrevem adequadamente a estrutura de mistura nos dados. Essa abordagem é especialmente valiosa em situações onde as subpopulações não são diretamente observáveis.

3.2.1 Algoritmo Metrópolis Hastings utilizando distribuição Poisson

Os passos específicos do algoritmo Metropolis-Hastings para estimar modelos de mistura com distribuição Poisson são os seguintes:

1. **Inicialização:**

   - Inicialize $t = 0$ e especifique valores iniciais para os parâmetros do modelo $\lambda^{(0)}$ e para o vetor latente $\mathbf{S}^{(0)}$.

2. **Geração de Novo Valor:**

   - Gere novos valores para os parâmetros do modelo $\lambda^{'}$ e o vetor latente $\mathbf{S}^{'}$ a partir de distribuições propostas adequadas para uma distribuição Poisson. Por exemplo:

$$\lambda_j^{'} \sim \text{Normal}(\lambda_j, \sigma)$$

   sendo σ fixado pelo modelador. Neste caso, os valores assumidos pela normal foram truncados em somente positivos.

$$\Pr[S_i = k \,|\, X_i = x_i, \lambda, p] = \frac{p_k \Pr[X_i = x_i \,|\, S_i = k, \lambda, p]}{\sum_{j=1}^{K} p_j \Pr[X_i = x_i \,|\, S_i = j, \lambda, p]} \tag{3.1}$$

3. **Cálculo da Função de Aceitação:**

   - Calcule a função de verossimilhança $L(\mathbf{y}|\lambda)$ para os dados $\mathbf{y}$, assumindo uma distribuição Poisson:
   $$L(\mathbf{y}|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$
   onde y_i é a i-ésima observação, λ_i é o valor médio associado a y_i, e n é o número de observações.
   - Calcule a função de densidade a *priori* $\pi(\lambda)$ para os parâmetros.
   - Calcule a função de aceitação $\alpha(\lambda, \lambda')$ como:
   $$\alpha(\lambda, \lambda') = \min\left(1, \frac{L(\mathbf{y}|\lambda')\pi(\lambda')q(\lambda|\lambda')}{L(\mathbf{y}|\lambda)\pi(\lambda)q(\lambda'|\lambda)}\right)$$
   onde $q(\lambda'|\lambda)$ é a distribuição proposta para os parâmetros.

4. **Aceitação/Rejeição:**

   - Gere uma variável aleatória $\mu \sim U(0,1)$.
   - Se $\mu \geq \alpha$, rejeite os novos valores e faça $\lambda^{(t+1)} = \lambda^{(t)}$ e $\mathbf{S}^{(t+1)} = \mathbf{S}^{(t)}$.
   - Caso contrário, aceite os novos valores e faça $\lambda^{(t+1)} = \lambda'$ e $\mathbf{S}^{(t+1)} = \mathbf{S}'$.

5. **Iteração:**

   - Incremente t para $t+1$.
   - Volte ao passo 2 para continuar gerando novos valores dos parâmetros do modelo e do vetor latente.

3.2.2 Algoritmo Metrópolis Hastings utilizando distribuição Poisson com inclusão de covariáveis

Os passos específicos do algoritmo Metropolis-Hastings para estimar modelos de mistura com distribuição Poisson com a inclusão de convariáveis são os seguintes:

1. **Inicialização:**

   - Inicialize $t = 0$ e especifique valores iniciais para os parâmetros do modelo $\lambda^{(0)}$, para o vetor latente $\mathbf{S}^{(0)}$, e para os parâmetros das covariáveis $\beta^{(0)}$.

2. **Geração de Novo Valor:**

   - Gere novos valores para os parâmetros do modelo λ', o vetor latente $\mathbf{S}'$, e os parâmetros das covariáveis β' a partir de distribuições propostas adequadas. Por exemplo:
   $$\lambda_j^{'} \sim \text{Normal}(\lambda_j, \sigma), \quad \beta_k^{'} \sim \text{Normal}(\beta_k, \sigma_\beta)$$

neste caso, os valores assumidos pela normal foram truncados em somente positivos.

3. **Cálculo da Função de Aceitação:**

   - Calcule a função de verossimilhança $L(y|\lambda, \mathbf{X}, \beta)$ para os dados y, levando em consideração as covariáveis:

   $$L(y|\lambda, \mathbf{X}, \beta) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, \quad \text{em que} \quad \lambda_i = e^{\mathbf{X}_i^T \beta + \Sigma_{j=1}^{K} z_{ij} \lambda_j}$$

   em que y_i é a i-ésima observação, λ_i é o valor médio associado a y_i, e n é o número de observações.

   - Calcule a função de densidade a *priori* $\pi(\lambda, \beta)$ para os parâmetros e as covariáveis.
   - Calcule a função de aceitação:

   $$\alpha(\lambda, \lambda', \beta, \beta') = \min\left(1, \frac{L(y|\lambda', \mathbf{X}, \beta')\pi(\lambda', \beta')q(\lambda, \beta|\lambda', \beta')}{L(y|\lambda, \mathbf{X}, \beta)\pi(\lambda, \beta)q(\lambda', \beta'|\lambda, \beta)}\right)$$

   onde $q(\lambda, \beta|\lambda', \beta')$ é a distribuição proposta para os parâmetros e as covariáveis.

4. **Aceitação/Rejeição:**

   - Gere uma variável aleatória $\mu \sim U(0, 1)$.
   - Se $\mu \geq \alpha$, rejeite os novos valores e faça $\lambda^{(t+1)} = \lambda^{(t)}$, $\mathbf{S}^{(t+1)} = \mathbf{S}^{(t)}$, e $\beta^{(t+1)} = \beta^{(t)}$.
   - Caso contrário, aceite os novos valores e faça $\lambda^{(t+1)} = \lambda'$, $\mathbf{S}^{(t+1)} = \mathbf{S}'$, e $\beta^{(t+1)} = \beta'$.

5. **Iteração:**

   - Incremente t para $t + 1$.
   - Volte ao passo 2 para continuar gerando novos valores dos parâmetros do modelo, do vetor latente e dos parâmetros das covariáveis.

Este processo iterativo permite explorar o espaço de parâmetros do modelo de mistura com distribuição Poisson, ajustando-se gradualmente aos dados observados e convergindo para uma estimativa dos parâmetros que descrevem adequadamente a estrutura de mistura nos dados. Essa abordagem é especialmente valiosa em situações onde as subpopulações não são diretamente observáveis.

CAPÍTULO

4

RESULTADOS

4.1 Estudos de Simulação

4.1.1 Modelo sem a inclusão de covariáveis

Utilizamos o *software* R para desenvolvimento dos estudos de simulação. Em especial utilizamos a função *rpoisson* para geração de valores provenientes da distribuição Poisson. Para representar um modelo de mistura de distribuições Poisson trucadas no zero, geramos K sub amostras de uma distribuição Poisson com parâmetro λ_K. As subamostras simuladas são de tamanhos iguais. A chance de cada observação pertencer a subamostra K é equiprovável, ou seja, $\frac{1}{K}$.

Foram geradas amostras com os seguintes valores de N: 100(a), 500(b) e 1000(c). Cada uma dessas amostras foram geradas fixando os valor de K em 2.

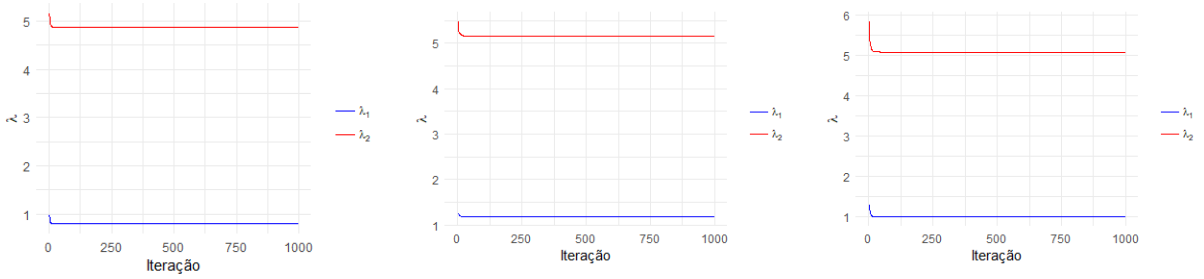
No cenário *a* foram estabelecidos valores de K próximos, com o intuito de saber como o modelo se comporta com mistura de distribuições com parâmetros próximos. Já no cenário *b* utilizamos valores de K mais distantes. Os cenários são apresentados na Tabela 1.

Vale ressaltar que para esse estudo serão consideradas 1000 iterações para o EM e 5000 para o MH.

Tabela 1 – Cenários Simulados

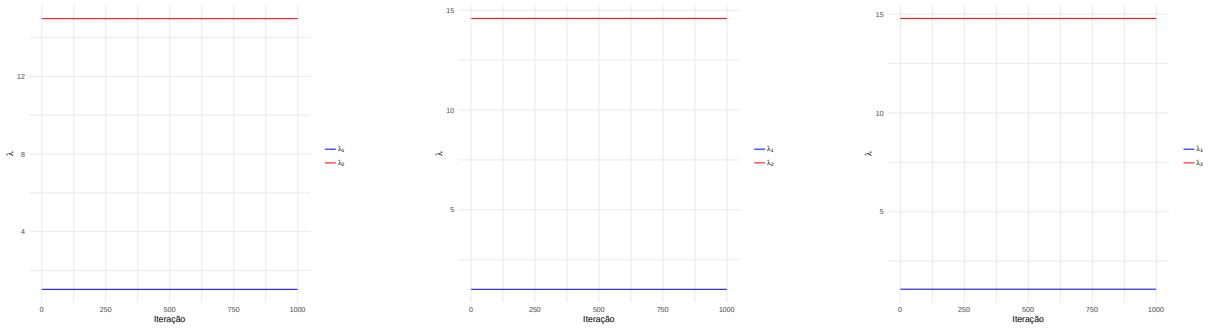
		$\lambda = (\lambda_1, ..., \lambda_K)$	
K	p_k	***a***	***b***
2	1/2	(1; 5)	(1; 15)

Abaixo, apresentamos os resultados encontrados para as estimações dos parâmetros dos modelos simulados de mistura de *Poisson*.



(a) (b) (c)

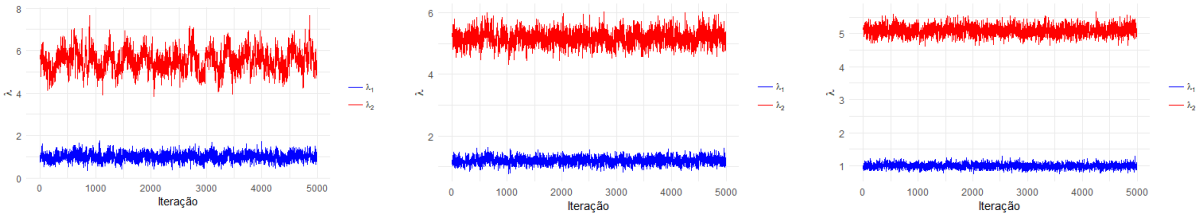
Figura 2 – Resultados do EM, para os dados simulados com $\lambda = 1$ e 5, considerando 1000 iterações e uma amostra de tamanhos 100(a), 500(b) e 1000(c) respectivamente.



(a) (b) (c)

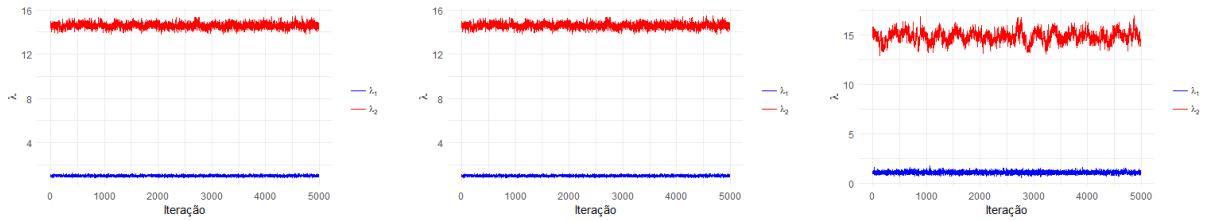
Figura 3 – Resultados do EM, para os dados simulados com $\lambda = 1$ e 15 considerando 1000 iterações e uma amostra de tamanhos 100(a), 500(b) e 1000(c)

Analisando conjuntamente as Figuras 2 e 3 temos que, o algoritmo EM já performa bem quando os parâmetros são próximos. No entanto, para parâmetros distantes, as estimações são muito próximas dos reais valores. Vale ressaltar que a medida que o tamanho da amostra aumenta, essas estimações ficam ainda melhores.



(a) (b) (c)

Figura 4 – Resultados do MH, para os dados simulados com $\lambda = 1$ e 5, considerando 1000 iterações e uma amostra de tamanhos 100(a), 500(b) e 1000(c)



(a) (b) (c)

Figura 5 – Resultados do MH, para os dados simulados com $\lambda = 1$ e 15 considerando 5000 iterações e uma amostra de tamanhos 100(a), 500(b) e 1000(c)

Quando avaliamos as Figuras 4, 5 constatamos que os resultados apresentados pela métrica MH são bem próximos aos encontrados utilizando o algoritmo EM.

Abaixo, na Tabela 2 e Tabela 3, também apresentamos os resultados médios das iterações, para as estimações dos parâmetros dos modelos simulados de mistura de *Poisson*.

Para o algoritmo MH, podemos fazer uma análise mais detalhada. A Tabela 3 apresenta os resultados de simulações para diferentes cenários, destacando a precisão das estimativas $\hat{\lambda}_k$ em relação aos valores verdadeiros λ_k. Observa-se que, com o aumento do tamanho da amostra (100, 500, 1000), a precisão das estimativas melhora significativamente, evidenciada pela redução do desvio padrão (s) e pela menor variabilidade nas estimativas mínimas e máximas. Além disso, as estimativas $\hat{\lambda}_k$ são consistentes e próximas aos valores verdadeiros λ_k, especialmente em amostras maiores. A distribuição das estimativas é simétrica, com medianas próximas às estimativas e percentis 5% e 95% equidistantes da mediana. Em resumo, a análise demonstra que maiores tamanhos de amostra resultam em estimativas mais precisas e menos variáveis, destacando a importância de amostras grandes em simulações para previsões mais confiáveis.

Tabela 2 – Resultados dos cenários simulados - EM

$\lambda = (\lambda_1, ..., \lambda_K)$		
	K=2	
	$a = (1; 5)$	$b = (1; 15)$
100	(1.21; 4.81)	(1.08; 14.79)
500	(1.18; 5.16)	(1.02; 14.59)
1000	(0.99; 5.08)	(1.03; 14.97)

Tabela 3 – Resultados dos cenários simulados - MH

K	Amostra	λ_k	$\hat{\lambda_k}$	s	Mín	Máx	Mediana	P5%	P95%
2	100	**1**	1.04	0.20	0.25	1.84	1.04	0.69	1.32
	100	**5**	5.75	0.52	3.58	7.92	5.75	4.70	6.42
	500	**1**	1.20	0.11	0.71	1.69	1.20	1.01	1.39
	500	**5**	5.19	0.24	4.22	6.15	5.19	4.80	5.60
	1000	**1**	1.02	0.07	0.72	1.33	1.02	0.88	1.12
	1000	**5**	5.14	0.14	4.54	5.73	5.14	4.86	5.33
	100	**1**	1.15	0.14	0.48	1.83	1.15	0.88	1.33
	100	**15**	14.88	0.59	12.57	17.19	14.88	13.83	15.77
	500	**1**	1.04	0.06	0.80	1.28	1.04	0.92	1.13
	500	**15**	14.63	0.25	13.67	15.59	14.63	14.19	15.01
	1000	**1**	1.07	0.05	0.86	1.27	1.07	0.96	1.11
	1000	**15**	14.94	0.17	14.28	15.59	14.94	14.71	15.26

Corroborando os resultados anteriores, os algoritmos EM (*expectation maximization algorithm*) e MH (*Metropolis-Hasting*) performam melhor para parâmetros distantes, ou seja, apresentam valores mais próximos dos reais valores.

Além disso, também foram estimadas o peso (proporção) que cada subpopulação tem no modelo de mistura. Como podemos observar na Tabela 4 e Tabela 5, para as situações em que os parâmetros são próximos, o MH consegue dividir melhor as observações. Já para quando os parâmetros são distantes, ambos os parâmetros apresentam resultados praticamente iguais.

A Tabela 5 apresenta os pesos das subpopulações em diferentes cenários simulados, mostrando como as estimativas $\hat{\rho}$ se comparam aos pesos verdadeiros ρ_k. Observa-se que, conforme o tamanho da amostra aumenta (100, 500, 1000), a precisão das estimativas melhora, evidenciada pela redução do desvio padrão (σ) e pela menor variabilidade entre os valores mínimos e máximos. A mediana e os percentis 5% e 95% indicam que as distribuições das estimativas são simétricas e centradas em torno do valor verdadeiro. Em resumo, a análise demonstra que maiores tamanhos de amostra resultam em estimativas de peso mais precisas e menos variáveis.

Tabela 4 – Pesos das subpopulações dos cenários simulados - EM

$\lambda = (\lambda_1, ..., \lambda_K)$		
	K=2	
	$a = (1; 5)$	$b = (1; 15)$
100	(0.43; 0.57)	(0.50; 0.50)
500	(0.59; 0.41)	(0.50; 0.50)
1000	(0.50; 0.50)	(0.50; 0.50)

Tabela 5 – Pesos das subpopulações dos cenários simulados - MH

K	n	ρ_k	$\hat{\rho}$	σ	Mín	Máx	Mediana	p5%	P95%
2	100	**0.50**	0.55	0.07	0.28	0.82	0.55	0.44	0.67
	100	**0.50**	0.45	0.07	0.18	0.72	0.45	0.33	0.56
	500	**0.50**	0.54	0.04	0.39	0.70	0.54	0.49	0.62
	500	**0.50**	0.46	0.04	0.30	0.61	0.46	0.38	0.51
	1000	**0.50**	0.51	0.02	0.41	0.61	0.51	0.46	0.50
	1000	**0.50**	0.49	0.02	0.39	0.59	0.49	0.45	0.54
	100	**0.50**	0.50	0.05	0.30	0.70	0.50	0.42	0.58
	100	**0.50**	0.50	0.05	0.30	0.70	0.50	0.42	0.58
	500	**0.50**	0.50	0.02	0.41	0.58	0.50	0.46	0.53
	500	**0.50**	0.50	0.02	0.41	0.58	0.50	0.46	0.53
	1000	**0.50**	0.50	0.02	0.44	0.56	0.50	0.48	0.53
	1000	**0.50**	0.50	0.02	0.44	0.56	0.50	0.47	0.52

4.1.2 Modelo com a inclusão de covariáveis

A simulação dos dados do modelo de mistura de distribuições de Poisson truncada no zero foi realizada em duas etapas principais:

1. **Geração de Covariáveis e Cálculo de λ**: Inicialmente, foram geradas duas covariáveis x_1 e x_2 a partir de uma distribuição normal com média 0 e variância 1. Essas covariáveis foram utilizadas para calcular os parâmetros λ_1 e λ_2 das distribuições Poisson, usando as seguintes fórmulas:

$$\lambda_1 = \exp(\beta_{0,1} + \beta_{1,1}x_1 + \beta_{2,1}x_2)$$

$$\lambda_2 = \exp(\beta_{0,2} + \beta_{1,2}x_1 + \beta_{2,2}x_2)$$

2. **Geração das Classes e Dados da Mistura de Poisson**: As classes foram geradas aleatoriamente com probabilidade $\pi_1 = 0.5$ e $\pi_2 = 0.5$. Com base nas classes, os dados foram gerados a partir de distribuições Poisson com os parâmetros λ_1 e λ_2. Se a observação pertencesse à classe 0, o dado foi gerado de uma Poisson com parâmetro λ_1; caso contrário, de uma Poisson com parâmetro λ_2.

Foi gerada uma amostra com tamanho $N = 1000$. Para este estudo, foram considerados dois componentes ($K = 2$) com λ = (10, 40).

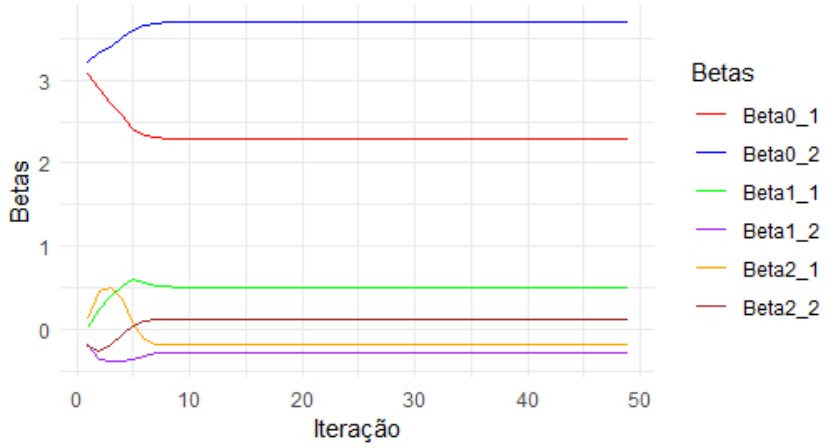
Para este estudo, foram realizadas 10000 iterações do algoritmo *Metropolis-Hastings* e a quantidade suficiente até convergência do algoritmo *Expectation-Maximization*, incorporando as covariáveis nas simulações. As covariáveis foram geradas a partir de uma distribuição normal com média 0 e variância 1, e incluídas como fatores explicativos nos modelos de mistura de Poisson.

Os parâmetros específicos utilizados para a geração dos dados são:

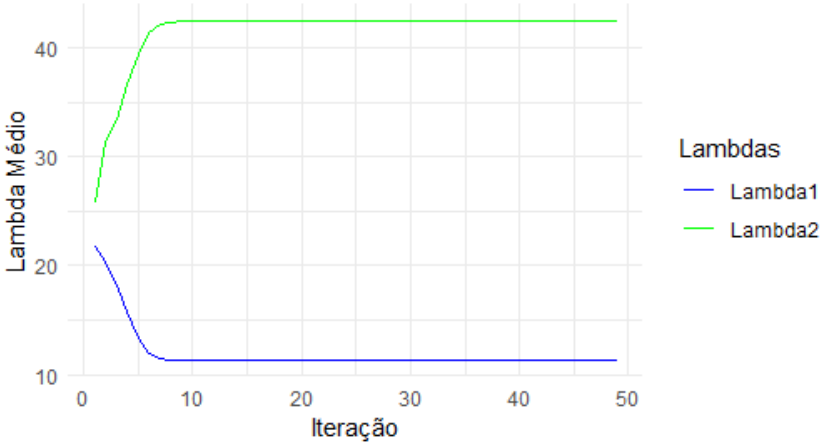
- $\rho_1 = 0.52, \rho_2 = 0.48$;
- $\beta_{0,1} = 2.3, \beta_{0,2} = 3.7$;
- $\beta_{1,1} = 0.5, \beta_{1,2} = -0.3$;
- $\beta_{2,1} = -0.2, \beta_{2,2} = 0.1$.

Com esses dados, é possível analisar como a inclusão das covariáveis impacta a modelagem da mistura de distribuições de *Poisson*.

Abaixo, apresentamos os resultados encontrados para as estimações dos parâmetros dos modelos simulados de mistura de *Poisson*. Para garantir a qualidade dos resultados obtidos pelo método *Metropolis Hastings*, desconsideramos as primeiras 1000 iterações utilizando o método de *burn-in* e consideramos um salto de 15 nas iterações subsequentes.



(a)



(b)

Figura 6 – Resultados do EM, para os dados simulados

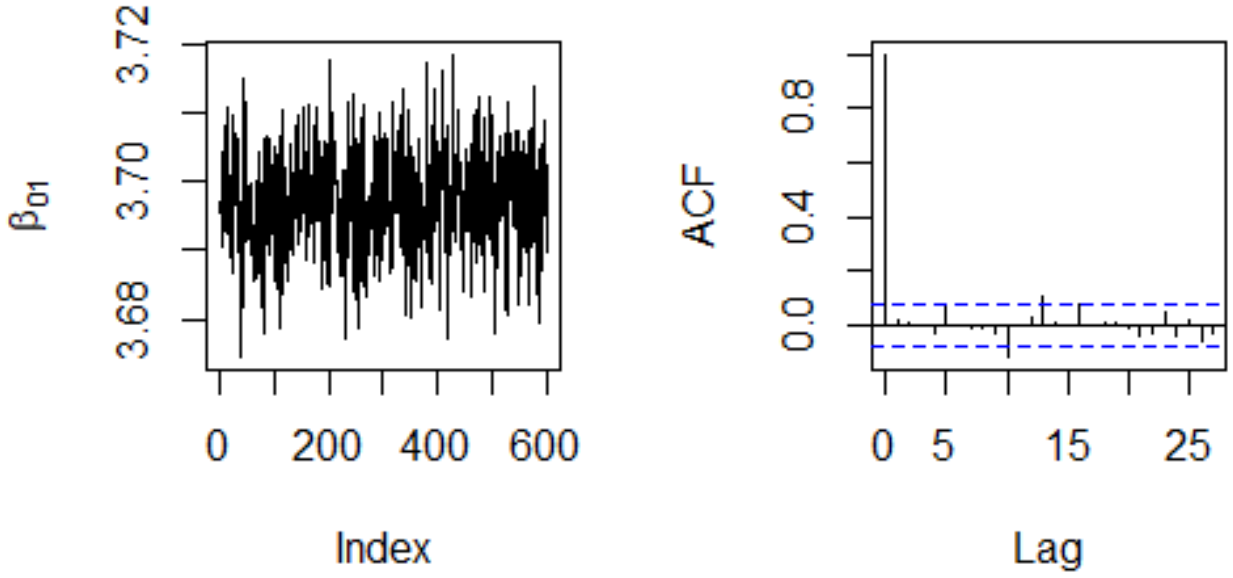


Figura 7 – Resultados do MH, considerando os dados simulados para o parâmetro β_{01}

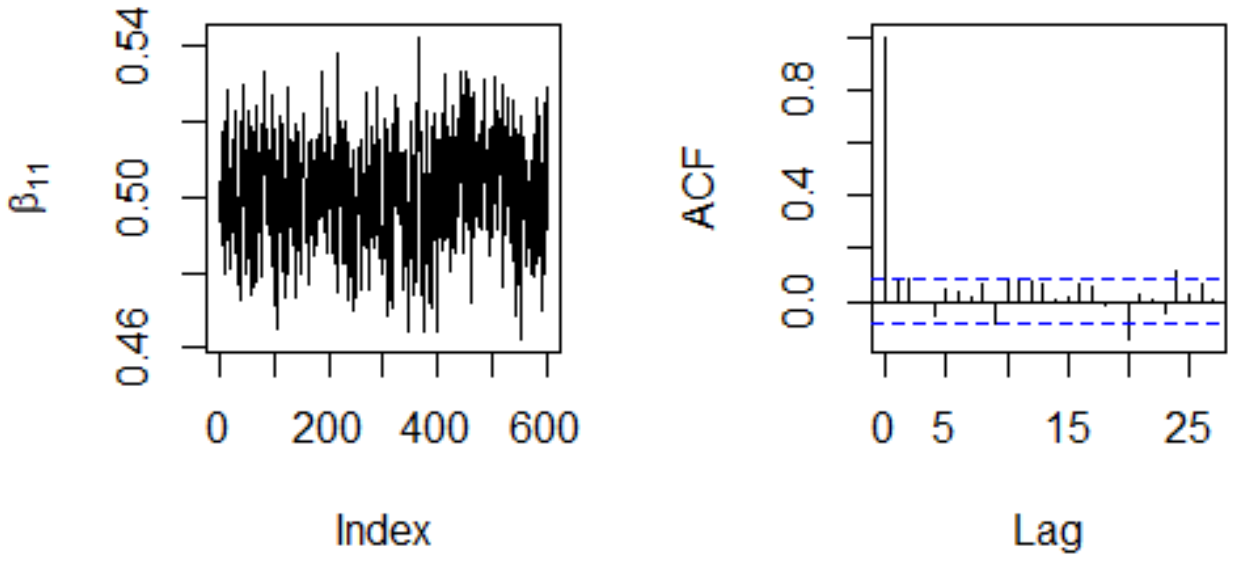


Figura 8 – Resultados do MH, considerando os dados simulados para o parâmetro β_{11}

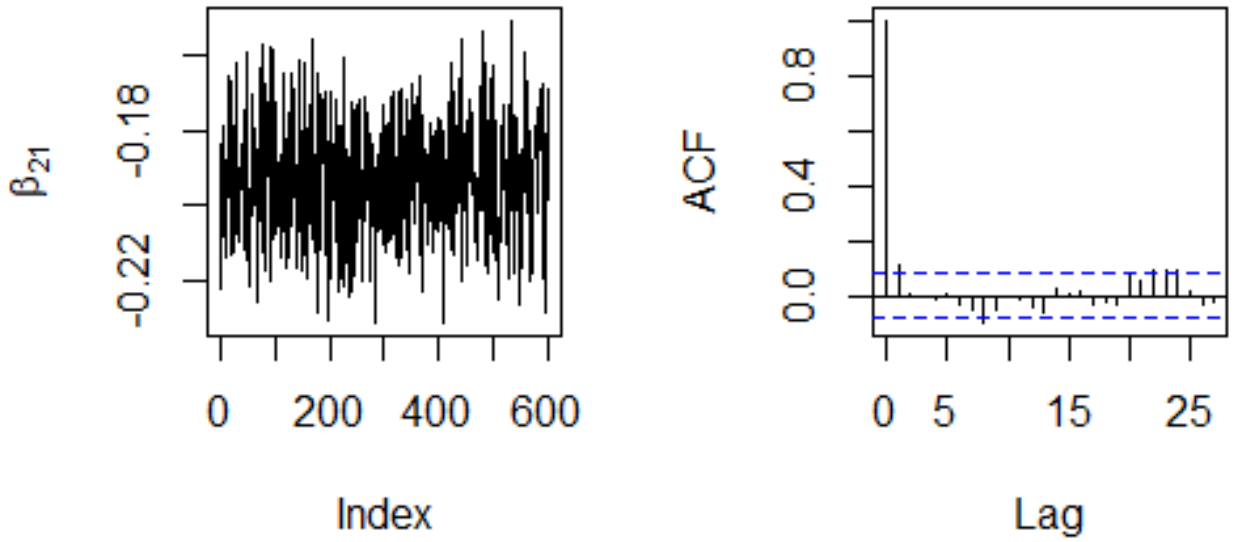


Figura 9 – Resultados do MH, considerando os dados simulados para o parâmetro β_{21}

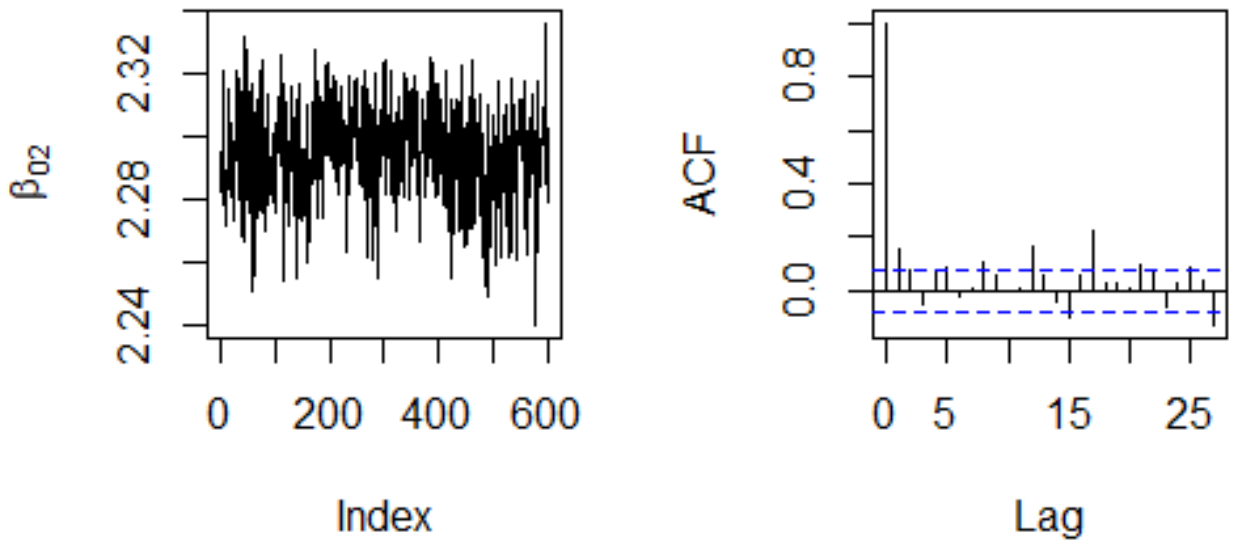


Figura 10 – Resultados do MH, considerando os dados simulados para o parâmetro β_{02}

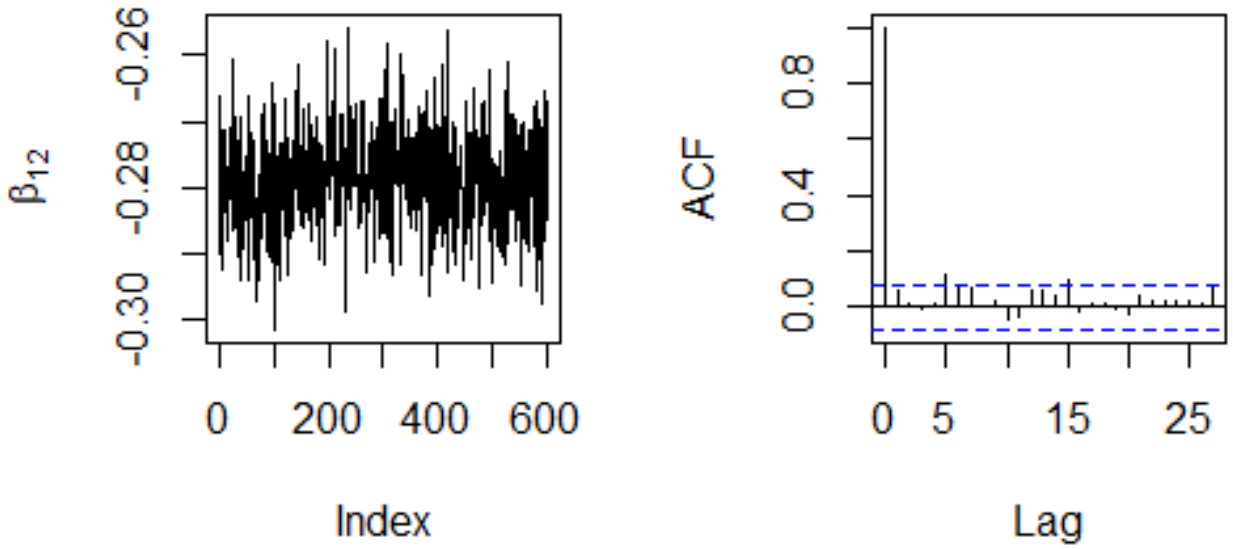


Figura 11 – Resultados do MH, considerando os dados simulados para o parâmetro β_{12}

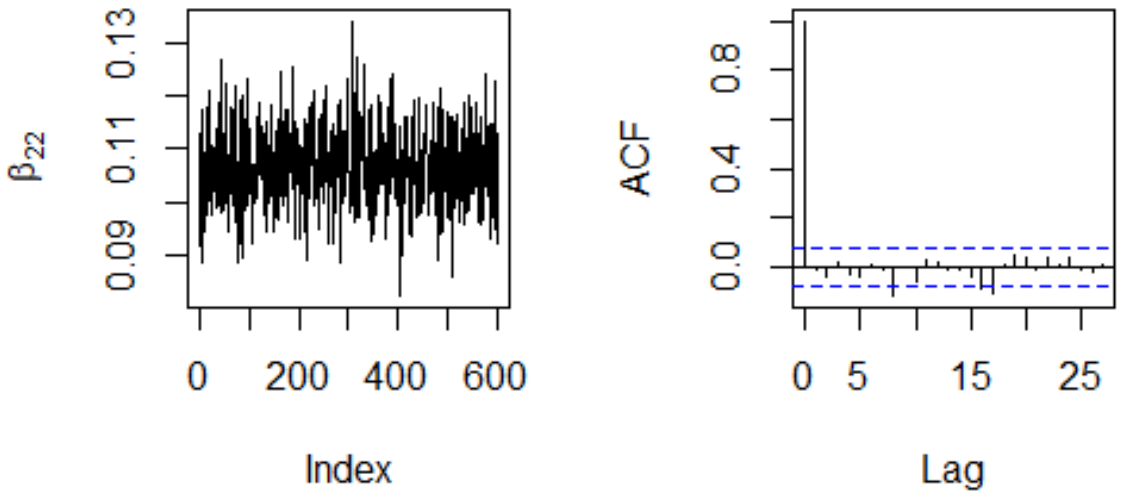


Figura 12 – Resultados do MH, considerando os dados simulados para o parâmetro β_{22}

Tabela 6 – Tabela Descritiva das estimativas dos parâmetros β_s, através do MH

	β	$\hat{\beta}$	s_k	Min	Max	Mediana	p5%	p95%
β_{01}	3.70	3.70	0.01	3.67	3.72	3.70	3.68	3.71
β_{11}	0.50	0.50	0.01	0.45	0.55	0.50	0.48	0.52
β_{21}	-0.19	-0.19	0.02	-0.24	-0.14	-0.19	-0.22	-0.17
β_{02}	2.29	2.29	0.02	2.23	2.35	2.29	2.27	2.32
β_{12}	-0.28	-0.28	0.01	-0.31	-0.25	-0.28	-0.29	-0.27
β_{22}	0.11	0.11	0.01	0.08	0.14	0.11	0.09	0.12

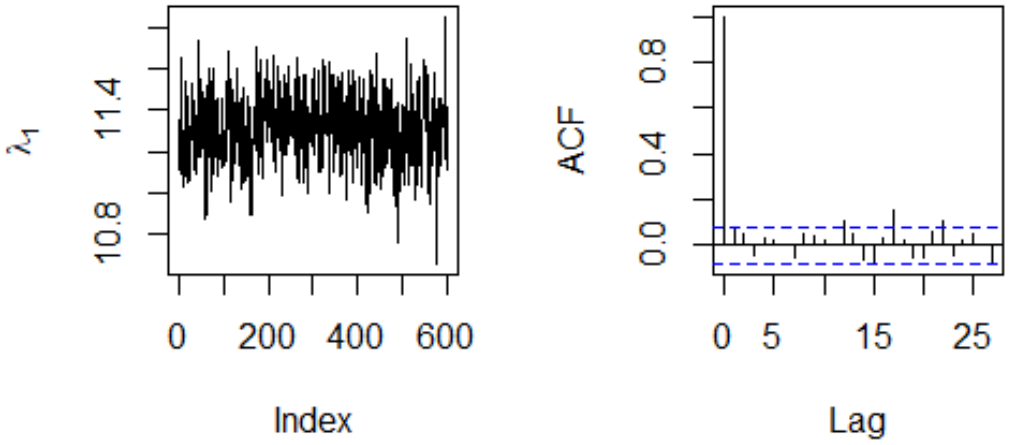


Figura 13 – Resultados do MH, considerando os dados simulados para o parâmetro λ_1

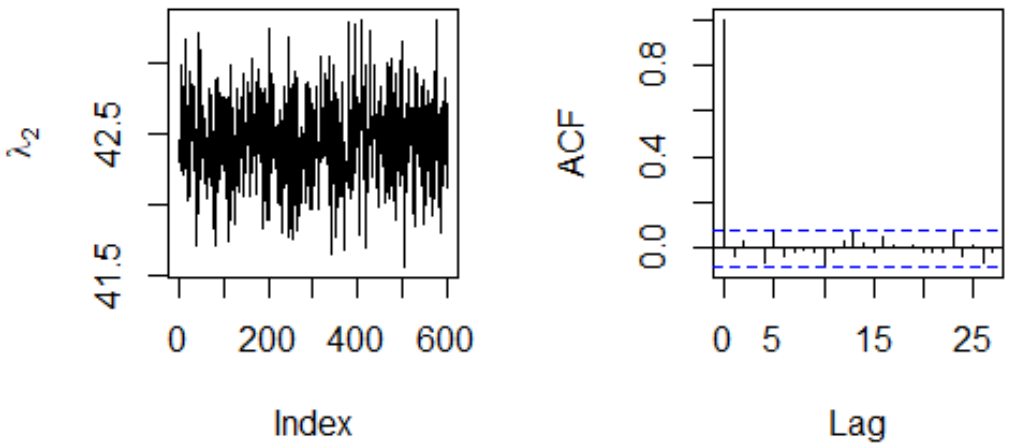


Figura 14 – Resultados do MH, considerando os dados simulados para o parâmetro λ_2

Tabela 7 – Tabela Descritiva das estimativas dos parâmetros λ_s, através do MH

	λ	$\hat{\lambda}$	SD	Min	Max	Mediana	p5%	p95%
λ_1	10	11.25	0.17	10.54	11.97	11.25	11.04	11.58
λ_2	40	42.43	0.30	41.35	43.52	42.43	41.96	42.93

Tabela 8 – Tabela Descritiva das estimativas dos parâmetros ρ_s, através do MH

	ρ	$\hat{\rho}$	SD	Min	Max	Mediana	p5%	p95%
ρ_1	0.52	0.52	0.02	0.46	0.59	0.52	0.50	0.55
ρ_2	0.48	0.48	0.02	0.41	0.54	0.48	0.45	0.50

A análise dos gráficos traceplot e ACF confirma que as cadeias de Markov atingiram uma mistura adequada e convergência após o período de *burn-in*. A rápida queda da autocorrelação nas amostras sugere baixa dependência serial, resultando em estimativas estáveis e confiáveis dos parâmetros β e λ. Esses resultados reforçam a validade das estimativas obtidas e a eficiência do algoritmo de MCMC utilizado no modelo de mistura de distribuições de Poisson truncada no zero com inclusão de covariáveis.

Os gráficos gerados para o algoritmo EM mostram uma rápida convergência dos parâmetros estimados para os valores simulados. Observamos que os β_s e os λ_s convergem de forma suave e eficiente. Por exemplo, $\beta_{0,1}$ e $\beta_{0,2}$ convergem rapidamente para aproximadamente 2.3 e 3.7, respectivamente. Da mesma forma, λ_1 e λ_2 convergem para 10 e 40, respectivamente. Essa rápida convergência e a estabilidade das estimativas indicam que o algoritmo EM é altamente eficaz para a identificação da estrutura subjacente dos dados.

Por outro lado, os gráficos para o algoritmo Metropolis-Hastings apresentam uma convergência mais lenta e com maiores flutuações iniciais. Embora os parâmetros eventualmente se aproximem dos valores esperados, o processo é mais demorado e menos estável. Por exemplo, os betas $\beta_{0,1}$ e $\beta_{0,2}$ mostram uma variação significativa antes de estabilizarem próximos aos valores simulados. Os lambdas também exibem flutuações antes de convergirem para os valores esperados. Essas características refletem a natureza estocástica do algoritmo MH, que tende a exigir mais iterações para alcançar a convergência, que nesse caso é uma convergência em distribuição.

A principal diferença entre os métodos está na natureza dos mesmos, enquanto o EM consiste na busca do máximo da função de verossimilhança, apresentando portanto uma informação pontual, no MH temos uma convergência em distribuição onde o objeto da busca é a distribuição do parâmetro que pretendemos estimar e cujos valores simulados permitem estimar, além da média, a mediana, medidas de variabilidade como variância, desvio padrão, entre outras, a densidade, medidas de assimetria entre outras.

Dadas essas diferenças, podemos concluir que trata-se de ferramentas que estão em patamares diferentes e que a comparação não se aplica.

CAPÍTULO

5

APLICAÇÃO

5.1 Descrição dos Dados

Os dados utilizado neste trabalho foi disponibilizado pelo fbref.com (2021), um site que oferece um extenso histórico de estatísticas do futebol, incluindo dados sobre jogadores, times e ligas de diversos países. Mangerona (2022) já utilizou esse mesmo conjunto de dados para analisar os principais fatores que influenciam o número de gols feitos pelos jogadores no campeonato inglês.

O conjunto de dados original incluía informações de todos os 532 jogadores que participaram do campeonato inglês na temporada 2020-2021. Foram selecionados apenas 162 jogadores das posições de ataque e meio-campo, excluindo goleiros e zagueiros, que não têm como objetivo principal marcar gols e também os jogadores que não marcaram nenhum gol.

As variáveis mais pessoais dos jogadores, como nome, nacionalidade, nome da equipe, posição em campo, ano de nascimento e minutos jogados divididos por 90, foram excluídas, pois não agregariam ao objetivo do trabalho. Além disso, foram removidas 4 variáveis derivadas do xG: gols esperados normais, gols esperados normais por chute, diferença entre gols marcados e xG, e diferença entre gols normais marcados e xG, uma vez que xG já estava sendo utilizado na análise.

Assim, o conjunto de dados final para o estudo incluiu as seguintes 14 variáveis:

1. **Idade:** Idade do jogador;
2. **Gols:** Quantidade de gols marcados por jogador na temporada;
3. **TC:** Total de chutes por jogador, não incluindo cobranças de pênaltis;
4. **CaG:** Total de chutes por jogador com direção ao gol (excluindo chutes para fora do gol);

5. **SoT:** Porcentagem de chutes por jogador com direção ao gol (CaG/TC);
6. **Sh:** Total de chutes por 90 minutos;
7. **SooT:** Total de chutes com direção ao gol por 90 minutos;
8. **Gols/TC:** Quantidade de gols marcados por total de chutes;
9. **Gols/CaG:** Quantidade de gols marcados por total de chutes com direção ao gol;
10. **Dist:** Distância média do gol, em jardas, de todas as finalizações;
11. **FK:** Chutes de falta por jogador;
12. **PB:** Pênaltis convertidos por jogador;
13. **PT:** Pênaltis batidos por jogador; e
14. **xG:** Gols previstos por jogador (incluem pênaltis). Essa informação é disponibilizada na fonte da base de dados.

5.2 Análise Descritiva

Com o banco de dados, primeiramente, foi realizada uma análise descritiva. Utilizando o *software* estatístico R, os seguintes resultados foram obtidos.

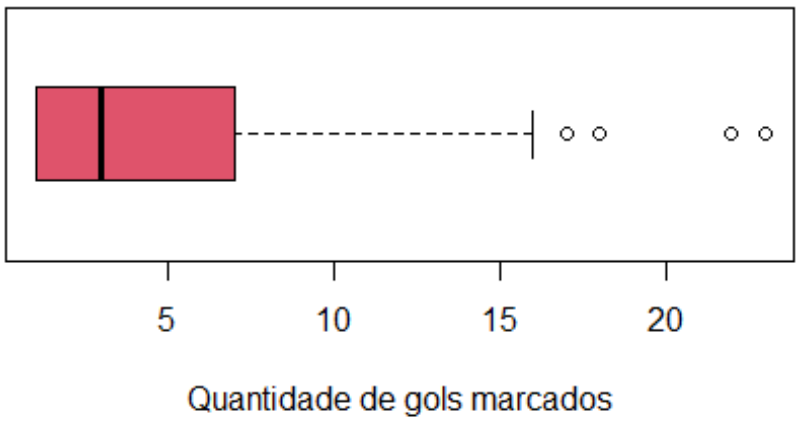


Figura 15 – Boxplot da quantidade de gols marcados, considerando em todas as observações da base de dados.

Tabela 9 – Descritiva da quantidade de gols marcados pelos jogadores na liga inglesa.

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
1	1	3	4,767	7	23

A partir da Figura 15 e da Tabela 9, observa-se que os jogadores de ataque e meio-campo da Premier League na temporada 2020-2021 marcaram, em média, poucos gols. Isto porque 75% dos jogadores analisados marcaram até 7 gols ao longo do campeonato.

Além disso, mais de 25% dos jogadores não marcaram nenhum gol. Considerando isso, a aplicação de um modelo com inflação de zeros poderia ser interessante. Contudo, como os dados serão modelados por um modelo de regressão, cada jogador terá sua própria média estimada a partir das variáveis do estudo, o que permitirá identificar os jogadores que marcaram zero gols.

Por outro lado, alguns jogadores, como Harry Kane do Tottenham, destacam-se como artilheiros do campeonato, com 23 gols marcados em 38 jogos.

Após essa análise inicial da variável resposta, o próximo passo é verificar graficamente quais variáveis preditoras parecem ter uma maior correlação com Y, fornecendo uma ideia inicial dos fatores que influenciam o número de gols marcados pelos jogadores.

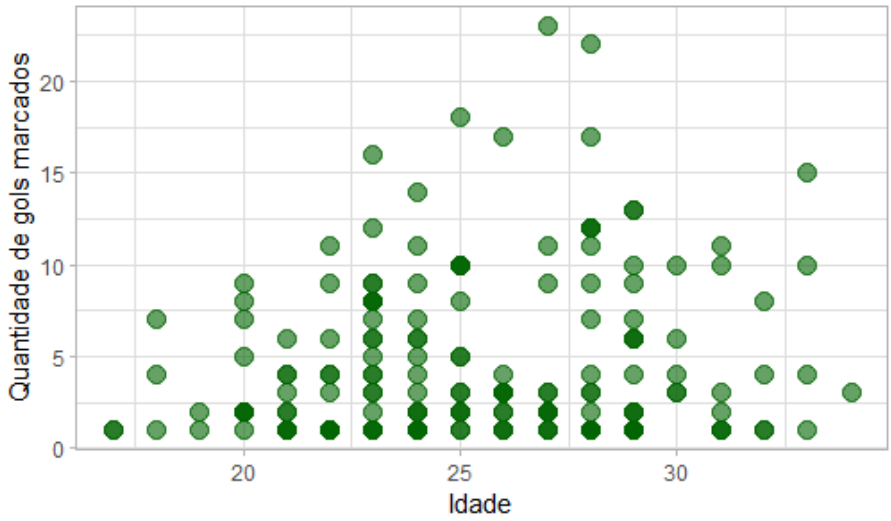


Figura 16 – Diagrama de dispersão da idade por gols marcados.

A partir do diagrama de dispersão acima, não se observa uma relação aparente entre a idade do jogador e o número de gols marcados. Isso sugere que, independentemente da idade, os jogadores não têm uma vantagem clara em termos de número de gols marcados. Nesse tipo de gráfico, as cores indicam a frequência em que o valor aparece na base. Quanto mais escuro, maior a frequência.

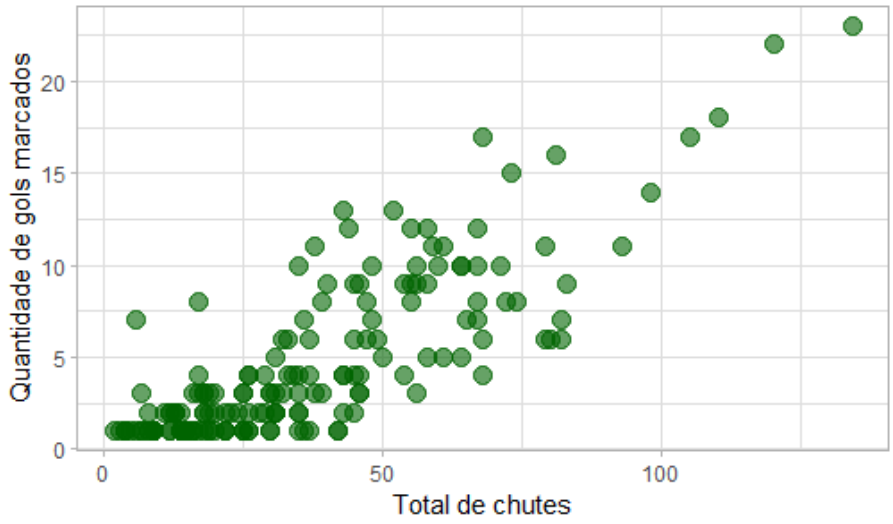


Figura 17 – Diagrama de dispersão de chutes por gols marcados.

Ao contrário do caso anterior, a Figura 17 mostra uma forte relação linear positiva entre o número total de chutes e a quantidade de gols marcados. Isso faz sentido, pois quanto mais o jogador chuta, maiores são as chances de marcar gols.

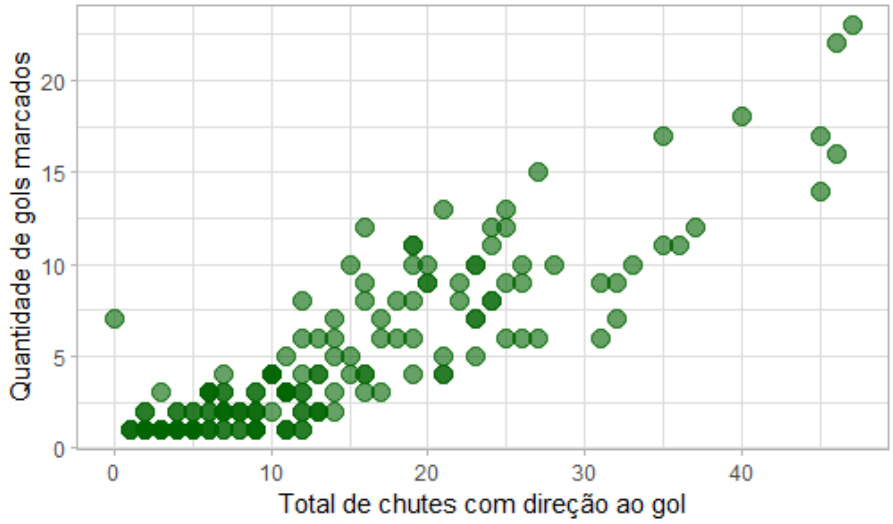


Figura 18 – Diagrama de dispersão de chutes com direção ao gol por gols marcados.

De maneira similar ao total de chutes, a variável de chutes com direção ao gol também apresenta uma forte correlação linear positiva com o número de gols marcados. Isso indica que a precisão no chute é um fator importante para marcar gols.

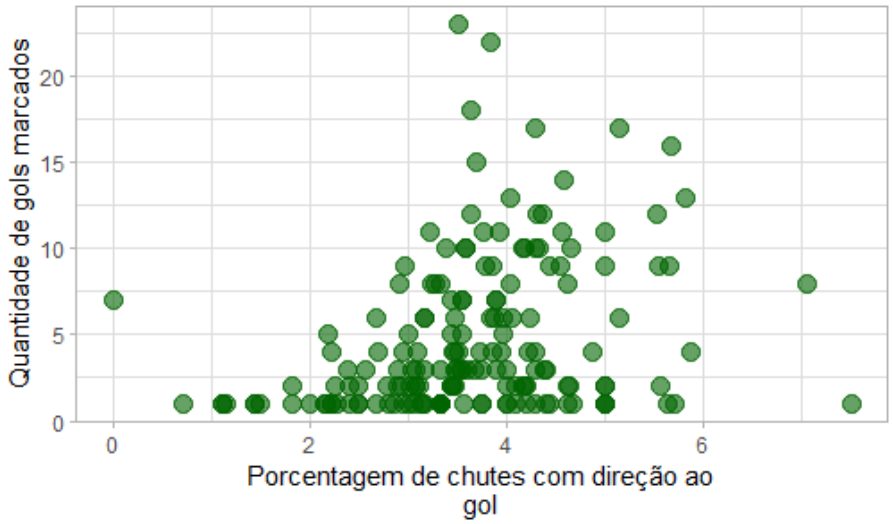


Figura 19 – Diagrama de dispersão da porcentagem de chutes com direção ao gol por gols marcados.

A Figura 19 mostra que a porcentagem de chutes com direção ao gol não parece ter uma relação clara com o número de gols marcados.

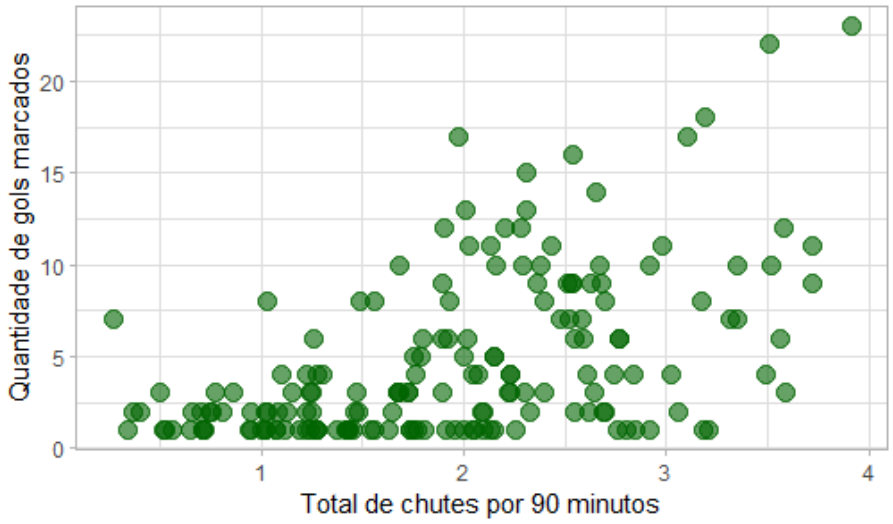


Figura 20 – Diagrama de dispersão de chutes em 90 minutos por gols marcados.

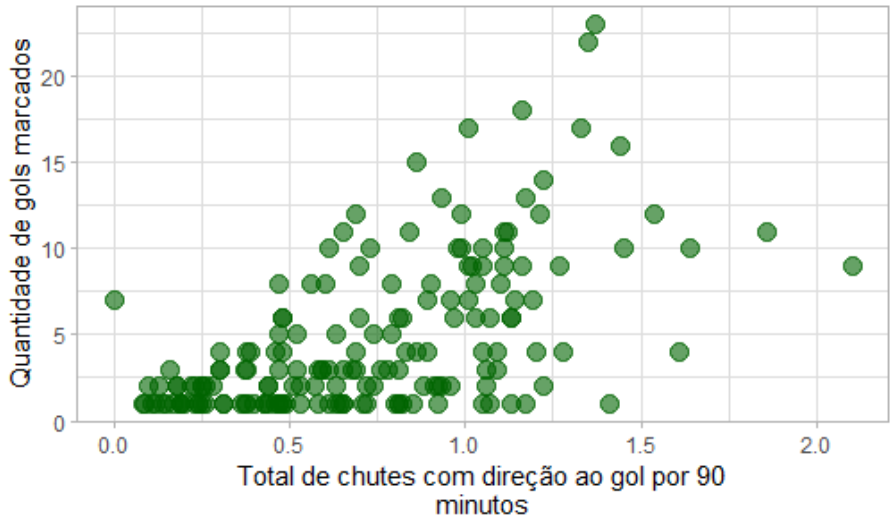


Figura 21 – Diagrama de dispersão de chutes com direção ao gol em 90 minutos por gols marcados.

As Figuras 20 e 21 sugerem uma relação linear positiva entre o número total de chutes (com direção ou não ao gol) por 90 minutos e a quantidade de gols marcados. Observa-se que, conforme o total de chutes aumenta, a quantidade de gols também tende a aumentar.

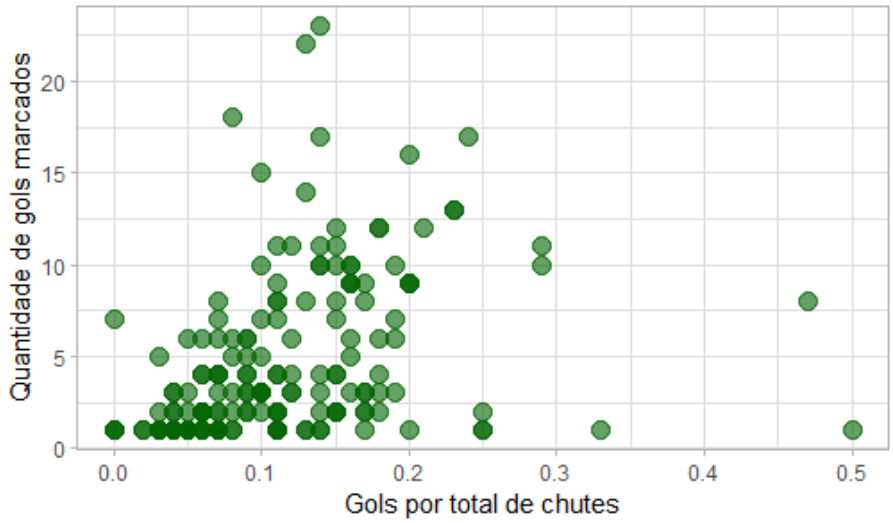


Figura 22 – Diagrama de dispersão de gols por chutes por gols marcados.

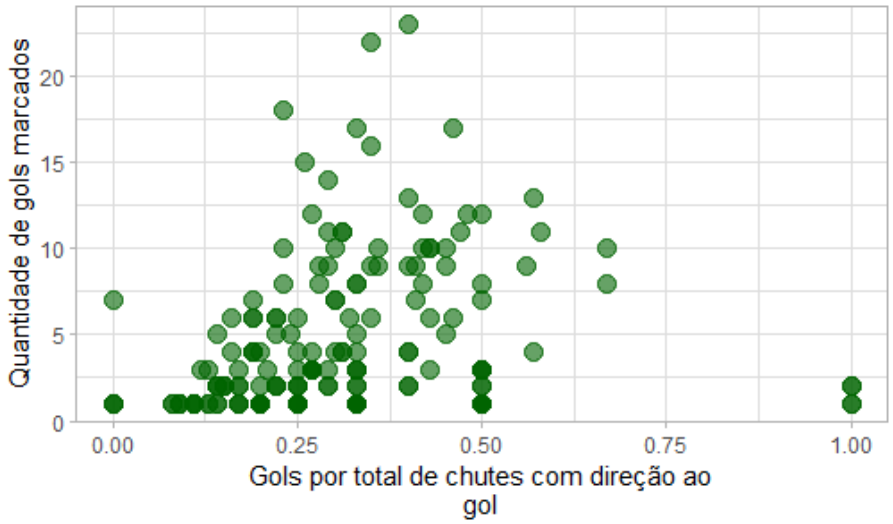


Figura 23 – Diagrama de dispersão de gols por chutes com direção ao gol por gols marcados.

As Figuras 22 e 23 não mostram uma relação clara entre gols por chutes (em direção ou não ao gol) e a quantidade de gols marcados. Isso pode parecer contraditório, mas é importante notar que essas variáveis são funções da variável resposta e, portanto, podem ser descartadas na modelagem.

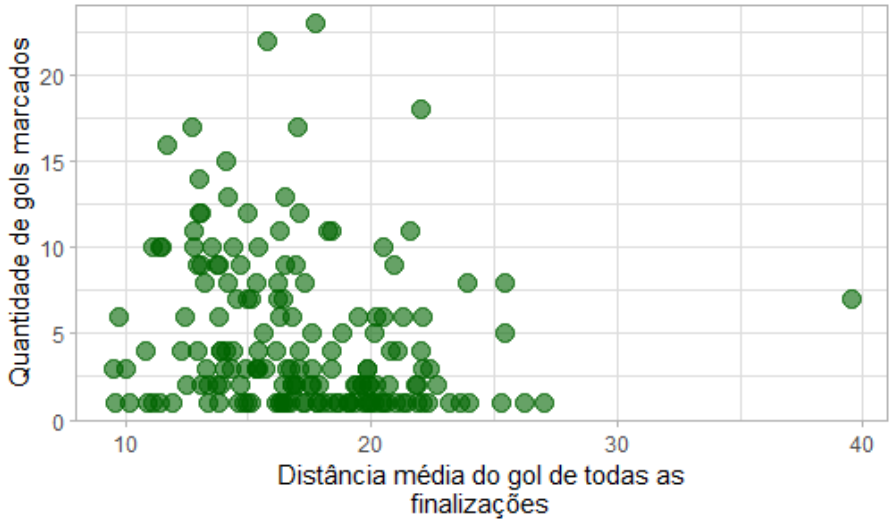


Figura 24 – Diagrama de dispersão de distância por gols marcados.

Por fim, a Figura 24 mostra que a distância média do gol de todas as finalizações tem uma correlação negativa com o número de gols marcados. Isso significa que jogadores que chutam mais perto do gol têm uma maior probabilidade de marcar gols.

Outro aspecto importante que pode atrapalhar o ajuste do modelo é a presença de variáveis preditoras altamente correlacionadas entre si. Nesse caso, é essencial selecionar aquelas que não apresentem correlações tão fortes para a continuação do estudo.

Nesse sentido, a Figura 25 apresenta a matriz de correlação entre as variáveis disponíveis.

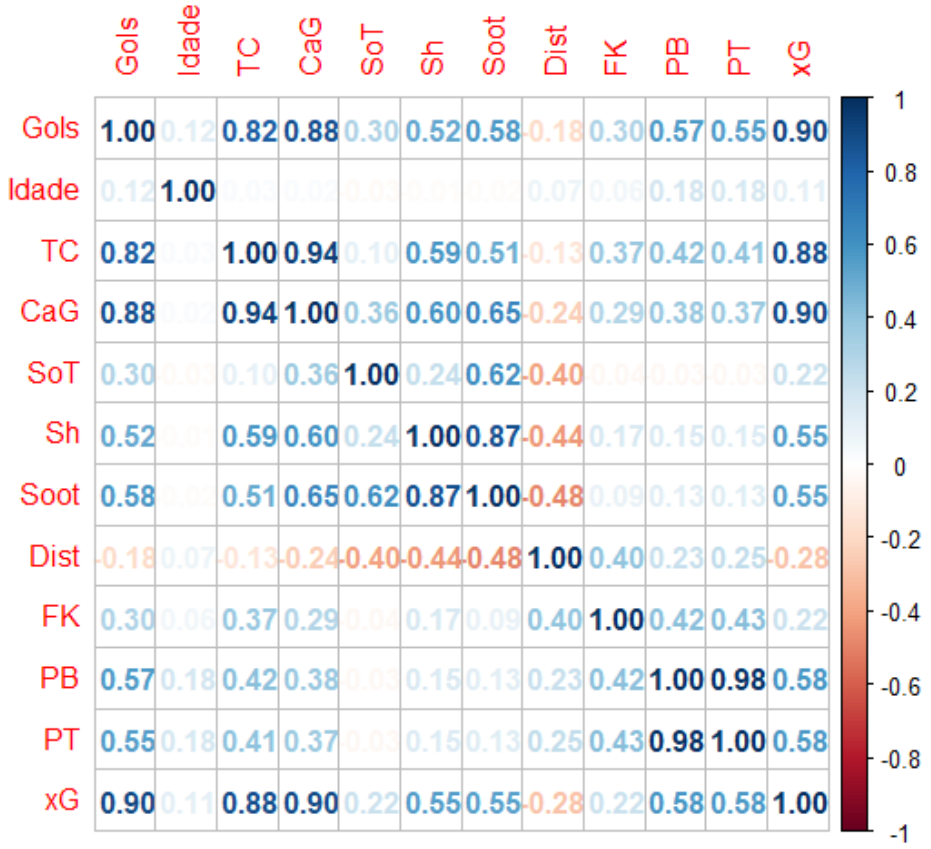


Figura 25 – Matriz de correlação.

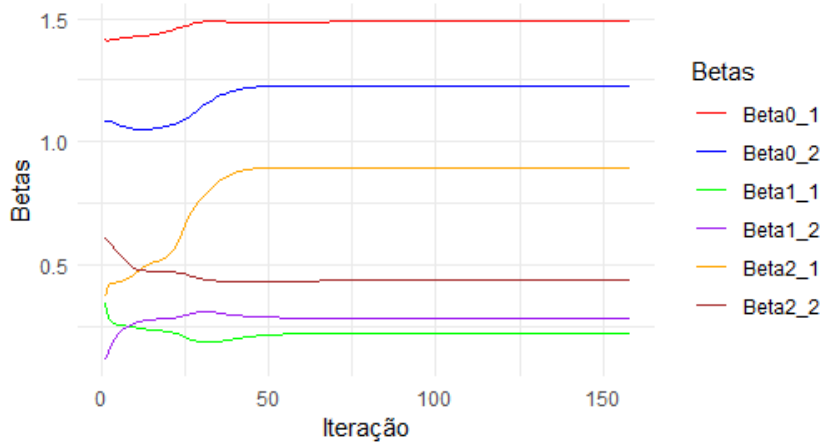
A partir da Figura 25, notamos a forte relação entre gols previstos por jogador (xG) com total de chutes por jogador (TC) e total de chutes por jogador com direção ao gol (CaG), ambas correlações maiores que 88%. Possivelmente, o xG de cada jogador é construído com forte influência da quantidade de vezes que o jogador chuta a bola ao gol.

Dessa forma, o TC será mantido no estudo, enquanto as outras citadas serão foram removidas da análise, dado que as informações que estas apresentam são praticamente as mesmas do TC. Ademais, vemos uma forte relação entre pênaltis batidos por jogador (PT) com pênaltis convertidos por jogador (PB), com 98% de correlação. Apesar de uma se referir à quantidade de vezes que o jogador bateu pênalti e a outra se referir à quantidade de vezes que a bola de fato entrou no gol, ambas parecem dar a mesma informação. Nesse sentido, a variável pênaltis batidos por jogador (PT) permanecerá no estudo.

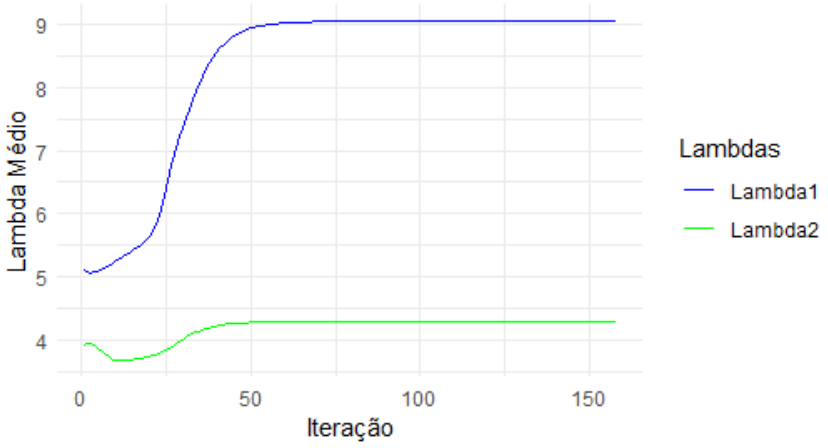
Por fim, outra relação que chama a atenção é entre porcentagem de chutes por jogador com direção ao gol (SoT) e total de chutes com direção ao gol por 90 minutos (SooT), com 62% de correlação. De fato, as duas variáveis representam situações bem parecidas no contexto de um jogo de futebol e, para o presente trabalho, será mantida a variável SooT.

Portanto, permanecem na análise as variáveis Idade, Sh, SooT, Dist, FK, PT e TC. Nesta dissertação, estamos nos restringindo a mistura de até 2 componentes e também com a presença de até 2 variáveis preditoras. Dessa forma, seguiremos utilizando as variáves Soot e TC que são variáveis fortemente relacionadas com a variável resposta e com pouca correlação entre si. Abaixo, apresentamos os resultados encontrados para as estimações dos parâmetros dos modelos simulados de mistura de Poisson. Para garantir a qualidade os resultado obtidos pelo método *Metropolis Hastings*, desconsideramos as primeiras 1000 iterações utilizando o método de burn-in e consideramos um salto de 15 nas iterações subsequentes.

Para as análises, considere β_{mk}, sendo k o componente e m a covariável associada. Por fim, m = 0 se refere ao intercepto, $m = 1$ a variável Soot e $m = 2$ a variável TC.



(a)



(b)

Figura 26 – Resultados das estimativas dos parâmetros β_s e λ_s, através do EM

Tabela 10 – Tabela Descritiva das estimativas dos parâmetros β_s e λ_s, através do EM

β_{01}	β_{11}	β_{21}	β_{02}	β_{12}	β_{22}	λ_1	λ_2
1.49	0.22	0.89	1.23	0.28	0.43	9.05	4.27

A Figura 26 e Tabela 10 apresenta as estimativas dos parâmetros β_s e λ_s, obtidas através do algoritmo EM. Os valores dos parâmetros β_{01}, β_{11} e β_{21} são 1.49, 0.22 e 0.89, respectivamente. Esses parâmetros representam as estimativas para o primeiro componente do modelo, indicando que β_{01} tem o maior impacto inicial. Para o segundo componente do modelo, as estimativas dos parâmetros β_{02}, β_{12} e β_{22} são 1.23, 0.28 e 0.43, respectivamente, sugerindo que β_{02} também tem um impacto inicial significativo, mas levemente inferior ao β_{01}. As estimativas dos parâmetros λ_1 e λ_2, 9.05 e 4.27, respectivamente.

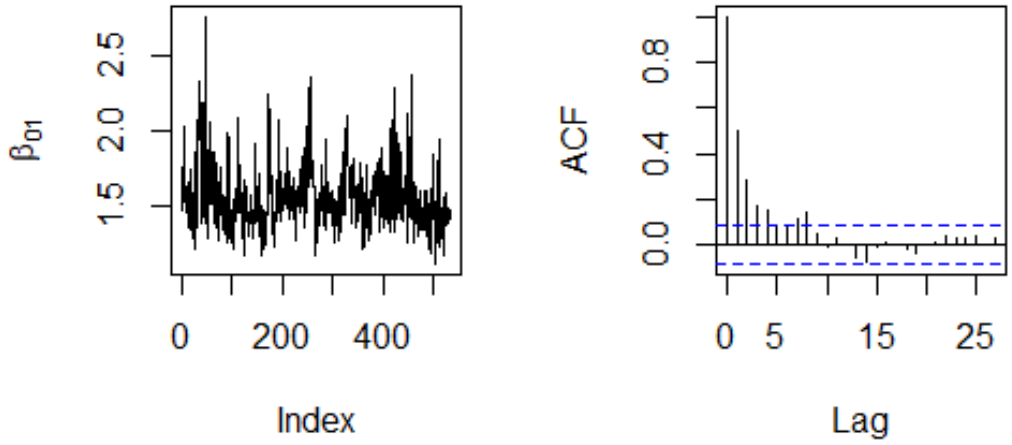


Figura 27 – Resultados do MH, para o parâmetro β_{01}

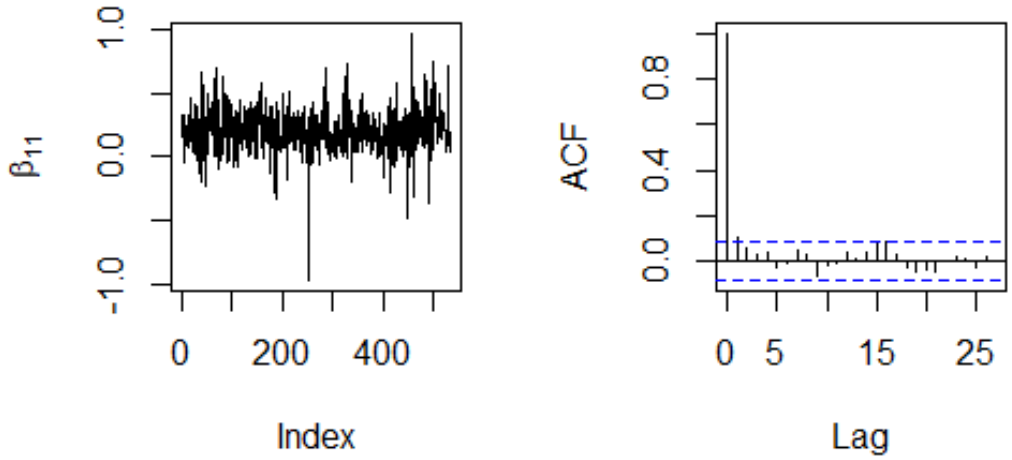


Figura 28 – Resultados do MH, para o parâmetro β_{11}

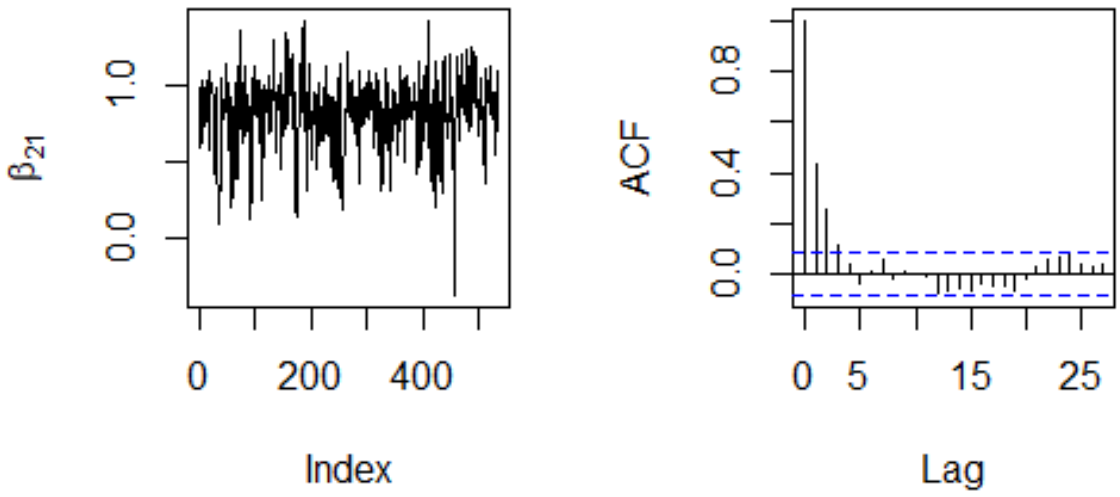


Figura 29 – Resultados do MH, para o parâmetro β_{21}

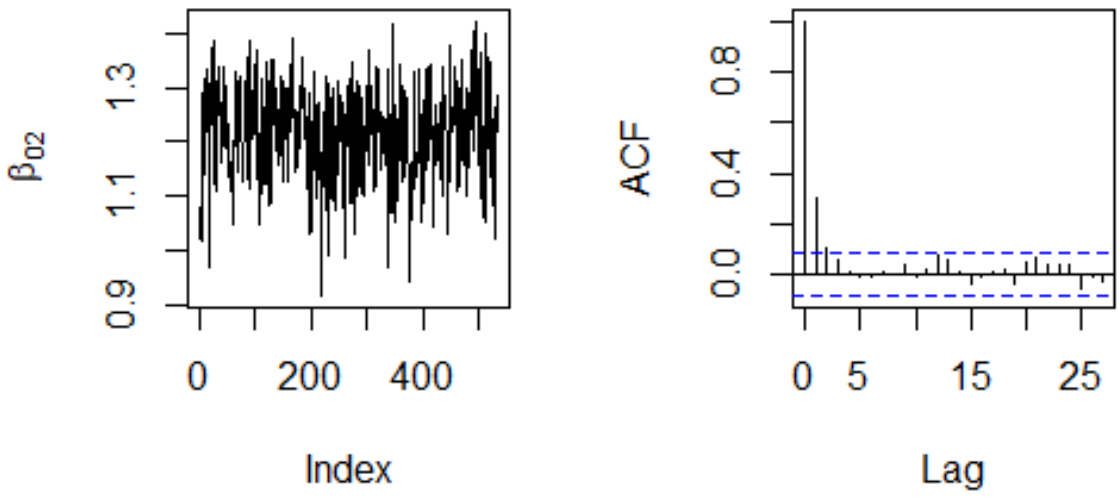


Figura 30 – Resultados do MH, para o parâmetro β_{02}

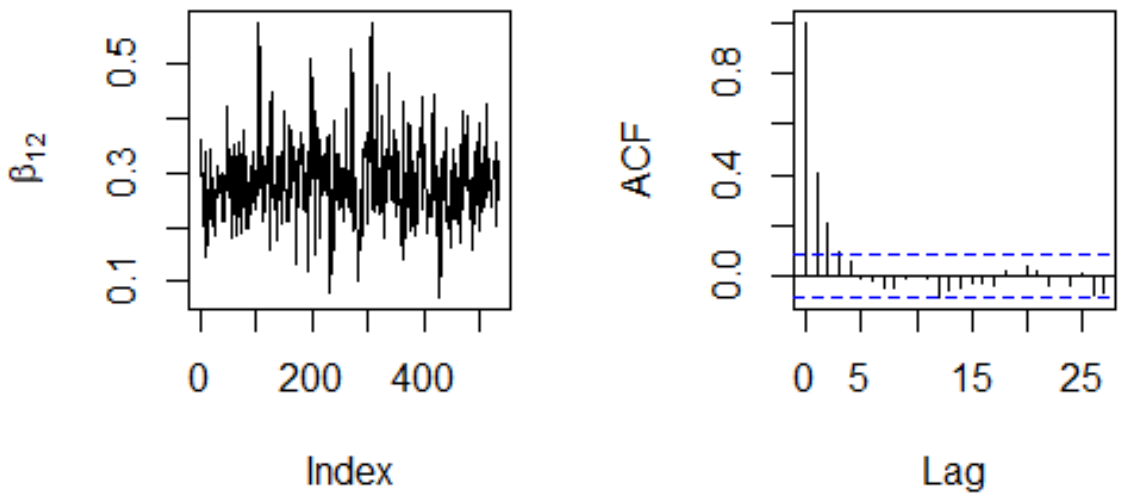


Figura 31 – Resultados do MH, para o parâmetro β_{12}

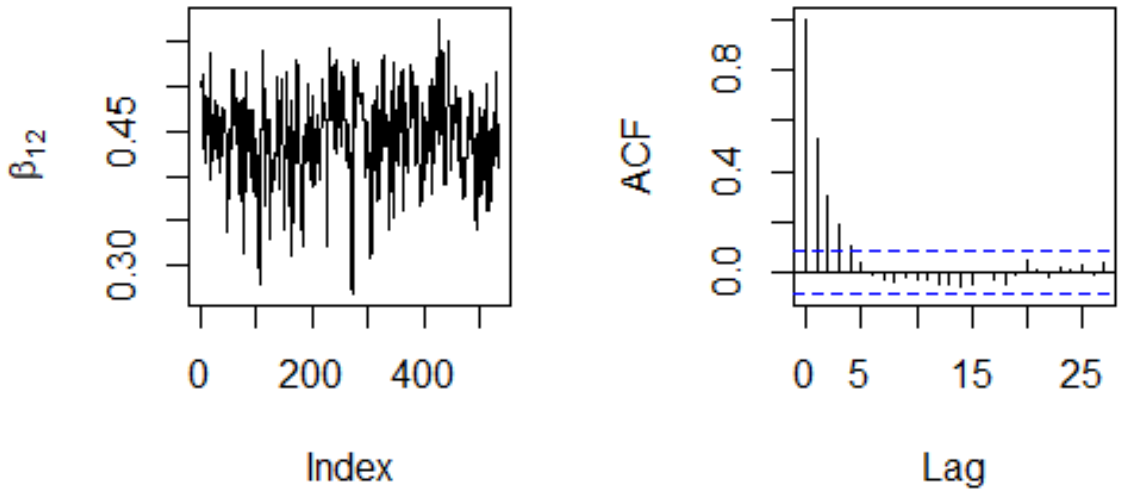


Figura 32 – Resultados do MH, para o parâmetro β_{22}

Tabela 11 – Tabela Descritiva das estimativas dos parâmetros β_s, através do MH

β	$\hat{\beta}$	s_k	Min	Max	Mediana	p5%	p95%
β_{01}	1.56	0.22	1.11	2.75	1.56	1.27	2.03
β_{11}	0.20	0.18	-0.97	0.97	0.20	-0.05	0.50
β_{21}	0.81	0.23	-0.37	1.42	0.81	0.33	1.13
β_{02}	1.22	0.08	0.92	1.42	1.22	1.07	1.34
β_{12}	0.29	0.07	0.07	0.58	0.29	0.19	0.41
β_{22}	0.44	0.05	0.27	0.58	0.44	0.36	0.52

A análise dos gráficos traceplot e ACF indica que as cadeias de Markov atingiram uma mistura adequada e convergência após o período de burn-in para todas as estimativas. A rápida diminuição da autocorrelação nas amostras sugere uma baixa correlação serial, resultando em estimativas estáveis e confiáveis dos parâmetros β e λ. Esses resultados reforçam a validade das estimativas obtidas e a eficiência do algoritmo de MCMC utilizado no modelo de mistura de distribuições de Poisson truncada no zero com inclusão de covariáveis. Vale ressaltar, que estimativas via MH estão muito próximas as obtidas pelo algoritmo EM.

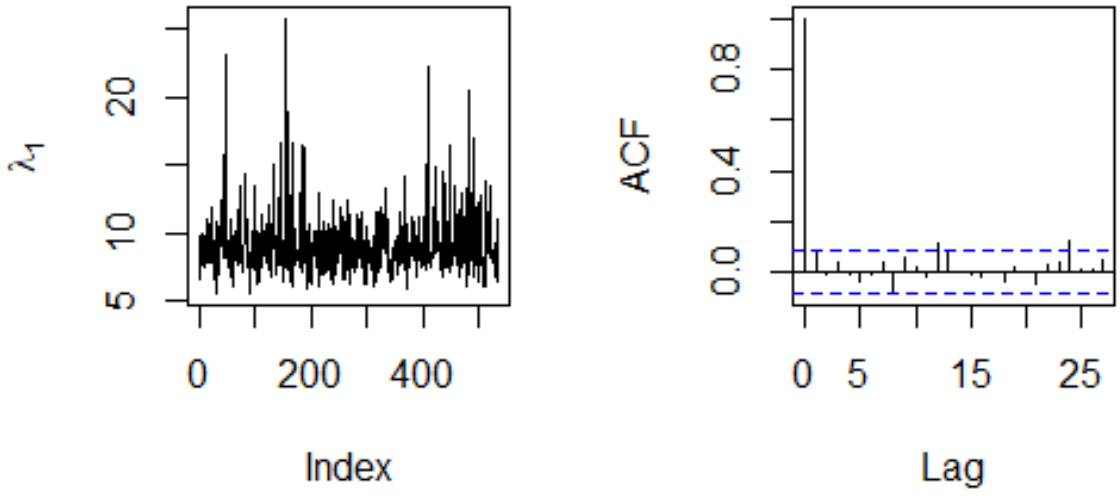


Figura 33 – Resultados do MH, para o parâmetro λ_1

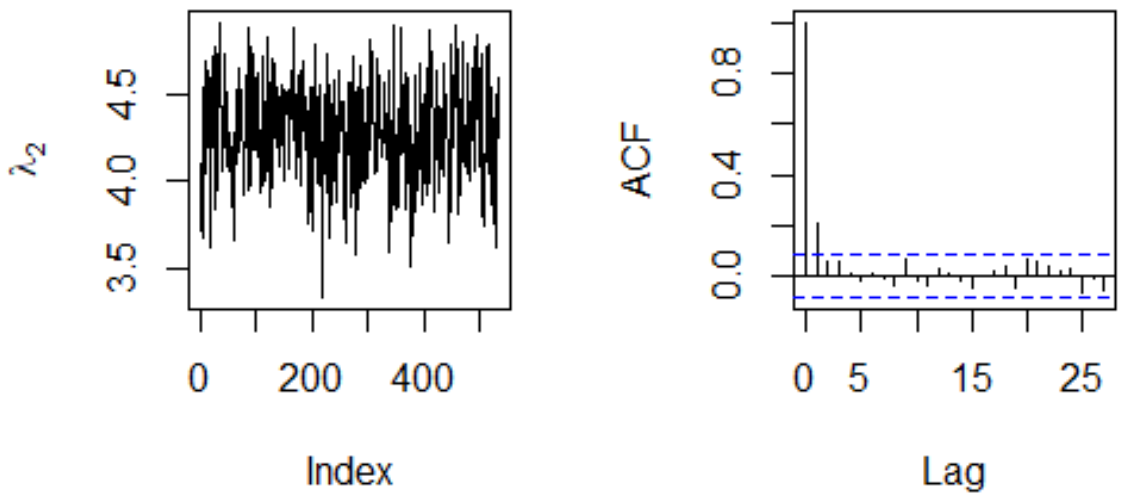


Figura 34 – Resultados do MH, para o parâmetro λ_2

Os gráficos apresentados nas Figura 33 e Figura 34, mostram as trajetórias das estimativas dos parâmetros λ_1 e λ_2 ao longo das iterações e suas respectivas funções de autocorrelação (ACF). Nos gráficos de λ_1, as estimativas se estabilizam após várias iterações, apesar de ainda haver alguma variabilidade, e a maioria dos valores de ACF está dentro dos limites de significância, sugerindo independência após algumas defasagens. Similarmente, para λ_2, as estimativas são mais estáveis e a ACF indica autocorrelação não significativa após algumas defasagens. Esses resultados são indicativos de um bom comportamento das simulações, onde as estimativas dos parâmetros se tornam aproximadamente independentes ao longo das iterações.

Tabela 12 – Tabela Descritiva das estimativas dos parâmetros λ_s, através do MH

λ_k	$\hat{\lambda}_k$	s_k	Min	Max	Median	p.5%	p.95%
λ_1	9.20	2.22	5.55	25.74	9.20	6.53	12.93
λ_2	4.30	0.28	3.33	4.92	4.30	3.82	4.73

Tabela 13 – Tabela Descritiva das estimativas dos parâmetros ρ_s, através do MH

k	$\hat{\rho}_k$	s_k	Min	Max	Mediana	p5%	p95%
1	0.28	0.12	0.00	0.65	0.28	0.08	0.47
2	0.72	0.12	0.35	1.00	0.72	0.53	0.92

A Tabela 12 apresenta as estimativas dos parâmetros λ_k. Para λ_1, a estimativa pontual é de 9.623, com um desvio padrão de 5.22, indicando uma variabilidade considerável, e um intervalo de confiança de 6.63 a 13.08. Em contraste, λ_2 possui uma estimativa de 4.27 e um

desvio padrão menor, de 0.28, refletindo uma maior precisão, com um intervalo de confiança de 3.85 a 4.77.

A Tabela 13 apresenta as estimativas dos parâmetros ρ_s. Para k=1, a estimativa pontual é de 0.33, com um desvio padrão de 0.12, e um intervalo de confiança de 0.07 a 0.49, indicando uma variabilidade baixa nas estimativas. Para o componente 2, temos uma estimativa de 0.67, também com um desvio padrão de 0.12, com um intervalo de confiança de 0.51 a 0.93.Portanto, O MH indica que a componente 2 tem um peso maior do que a 1 no modelo de mistura.

CAPÍTULO

6

CONCLUSÕES

Na pesquisa desenvolvida para essa dissertação foram estudadas duas metodologias para estimar os parâmetros do modelo de mistura de *Poisson* com truncamento no zero. Para a mistura, empregou-se a variável não observável S para definir em qual componente cada elemento pertence.

A primeira metodologia é o algorítimo EM - *expectation maximization algorithm*, que é aplicado no contexto da estimação pelo método de máxima verossimilhança. A segunda é o MH (*Metropolis-Hastings*),uma abordagem bayesiana, que é utilizada para quando o número de componentes é conhecido. Descrevemos e implementamos essa medidas, e realizamos estudos de simulação e também aplicação em um banco de dados reais.

A principal diferença entre os métodos está na natureza dos mesmos, enquanto o EM consiste na busca do máximo da função de verossimilhança, apresentando portanto uma informação pontual, no MH temos uma convergência em distribuição onde o objeto da busca é a distribuição do parâmetro que pretendemos estimar e cujos valores simulados permitem estimar, além da média, a mediana, medidas de variabilidade como variância, desvio padrão, entre outras, a densidade, medidas de assimetria entre outras.

Tanto o algoritmo EM quanto o MH performaram bem nos estudos de simulação, apresentando estimativas bem próximas dos reais parâmetros. Para a base de dados reais, algumas parâmetros estimados através do MH, apresentaram uma leve autocorrelação, mas os resultados estão bem próximos dos apresentados pelo algoritmo EM.

Neste trabalho, tanto nos estudos de simulação quanto nas aplicações, está sendo considerado um número pequeno de variáveis e componentes da mistura. Portanto, ainda não existe evidências de que essas conclusões obtidas são válidas para situações onde existe alta dimensionalidade. Um outro estudo pode ser desenvolvido para fazer tal verificação.

Além do estudo mencionado no parágrafo anterior, outros trabalhos futuros podem ser

desenvolvidos. Pode-se considerar por exemplo outras medidas para estimar os parâmetros do modelo de mistura para dados de contagem. Também poder ser considerada outra distribuição para os dados, por exemplo, Binomial Negativa.

REFERÊNCIAS

CELEUX, G.; GOVAERT, G. Gaussian parsimonious clustering models. **Pattern Recognition**, v. 28, n. 5, p. 781–793, 1995. Citado na página 21.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hastings algorithm. **The American Statistician**, [American Statistical Association, Taylor Francis, Ltd.], v. 49, n. 4, p. 327–335, 1995. ISSN 00031305. Disponível em: <http://www.jstor.org/stable/2684568>. Citado nas páginas 22 e 34.

COLLINS, A.; DASGUPTA, S.; SCHAPIRE, R. E. A generalization of principal component analysis to the exponential family. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2002. p. 617–624. Citado na página 22.

EHLERS, R. S. **Inferência Bayesiana**. [S.l.]: IME-USP São Paulo, 2007. Citado na página 35.

FBREF.COM. **2020-2021 Premier League Estatísticas de Chutes**. 2021. Disponível em: <https://fbref.com/pt/comps/9/10728/shooting/2020-2021-Premier-League-estatisticas>. Citado na página 51.

GONÇALVES, F.; MAIA, M. A.; CARVALHO, F. Multivariate zero-inflated poisson models for wildlife-vehicle collision data. **Journal of Environmental Management**, v. 193, p. 537–545, 2017. Citado na página 22.

GREEN, P. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. **Biometrika**, v. 82, 09 1995. Citado na página 22.

HELD, L.; PAWITAN, Y. **Statistical Methods for Censored and Truncated Data**. [S.l.]: Springer Science Business Media, 2011. ISBN 978-1-4419-0851-5. Citado na página 22.

JIN, Y.; GAO, X. Finite mixture model-based image segmentation. **IEEE Transactions on Image Processing**, v. 15, n. 10, p. 3066–3072, 2006. Citado na página 21.

LIN, H.; RADKE, R. J. Parameter estimation for the poisson mixture model and its application to low-level vision. In: **Proceedings Ninth IEEE International Conference on Computer Vision**. [S.l.: s.n.], 2003. v. 1, p. 182–189. Citado na página 22.

MANGERONA, R. A. M. **A estatística no futebol: uma análise dos principais fatores que influenciam o número de gols feitos pelos jogadores no campeonato inglês**. Tese (Trabalho de Conclusão de Curso) — Depertamento de Estatística da Universidade Federal de São Carlos, 2022. Citado na página 51.

MCLACHLAN, G.; KRISHNAN, T. The em algorithm and extensions. **John Wiley Sons**, v. 382, p. Citado nas páginas 22, 28 e 30, 2007. Citado na página 29.

MCLACHLAN, G.; PEEL, D. Finite mixture model. **John Wiley Sons**, v. 44, 01 2000. Citado nas páginas 21 e 25.

MCLACHLAN, G. J.; PEEL, D. **Finite Mixture Models**. [S.l.]: John Wiley & Sons, 2000. Citado na página 21.

MEIRA, S. A. **Modelo de Mistura com Dependência Markoviana de Primeira Ordem**. Tese (Tese de Doutorado) — Departamento de Estatística da Universidade Federal de São Carlos., 2014. Citado nas páginas 26, 28 e 29.

MELARD, G.; PASTEELS, J. M. Forecasting count series of animals: Exponential smoothing or dynamic regression? an application to breeding bird survey data. **International Journal of Forecasting**, v. 16, n. 4, p. 463–476, 2000. Citado na página 22.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>. Citado na página 22.

SARAIVA, E. F. **Modelo de Mistura com Número de Componentes Desconhecido: Estimação via Método Split-Merge**. Tese (Tese de Doutorado) — Departamento de Estatística da Universidade Federal de São Carlos., 2009. Citado na página 26.

XIONG, H.; LIU, H.; TAN, A. C. Protein secondary structure prediction using model-averaged consensus. **BMC Bioinformatics**, v. 7, n. 1, p. 1–15, 2006. Citado na página 21.

YANG, M.-H.; ZHOU, X.-H. Poisson mixture regression models for the analysis of the length of hospital stay. **Statistics in Medicine**, v. 27, n. 11, p. 1979–1997, 2008. Citado na página 22.


ICMC
USP
SÃO CARLOS