Aplicação do MCMC na Modelagem dos Gols dos Campeonatos da CBF de 2013 a 2020

Igor Patrício Michels

Escola de Matemática Aplicada, EMAp/FGV
Rio de Janeiro/RJ, Brasil

6 de Junho de 2021

Resumo: Esse trabalho tem como objetivo descrever a modelagem de gols por jogo nos campeonatos realizados pela CBF no período de 2013 a 2020 por meio da utilização de Cadeias de Markov Monte Carlo para a estimação de parâmetros. Dois modelos são propostos, um com uma Distribuição de Poisson e outro com uma Distribuição Gamma adaptada.

Palavras chave: MCMC, Método Iterativo, Modelagens de Gols, Distribuição de Poisson.

1 Introdução

Em grandes competições esportivas sempre ocorrem especulações sobre o resultado de uma prova, jogo ou da própria competição. Essas especulações se intensificam em épocas de Copa do Mundo, principalmente com a mídia trazendo animais "oráculos", os quais prevem os resultados de algumas partidas, como o polvo Paul, na Copa do Mundo de 2010, e o gato Aquiles, da Copa do Mundo de 2018. Matematicamente, a modelagem esportiva tem aplicações práticas para jogos de azar, melhoria nas técnicas de coaching para motivar equipes ou jogadores, no jornalismo e nos próprios clubes com o intuito de melhorar o desempenho dos seus jogadores e de prevenir lesões, além de estudar o desempenho de seus rivais, podendo elaborar uma estratégia ótima para enfrentar os mesmos.

Nesse sentido, estudos relacionados aos resultados esportivos são cada vez mais frequentes, uma vez que buscam explicar o poder ofensivo e defensivo dos clubes, sendo a modelagem global dos dados, isto é, modelar o esporte como um todo ao invés de modelar uma competição ou os clubes dentro da mesma, um passo inicial para esse estudo.

Outro fator relevante é o modelo e a técnica utilizada para realizar a modelagem. Nesse trabalho, será proposto um modelo se valendo da Distribuição de Poisson, a qual é defendida na literatura para modelar a quantidade de gols em uma partida em virtude da ideia de que os gols são eventos independentes, bem como o resultado de duas partidas distintas. Uma abordagem secundária se dará pela Distribuição Gamma com certas adaptações. Já técnica utilizada nesse trabalho será a estimação por meio de Cadeias de Markov Monte Carlo (MCMC).

2 Metodologia

Conforme citado anteriormente, o objetivo desse estudo é modelar a quantidade de gols em uma partida das competições da CBF. Para isso, os dados utilizados foram coletados diretamente do site da CBF por meio das súmulas das partidas em um trabalho anterior [4], com tratamento realizado via Python. As simulações foram realizadas por meio da mesma linguagem com a implementação do MCMC de maneira similar a vista em [3].

A modelagem foi realizada tanto globalmente quanto individualmente a nível de competição, isto é, foram analisados tanto os dados dos jogos como um todo, juntando todos os jogos de todas as competições, quanto os jogos de uma competição específica, como apenas os jogos da Série A, por exemplo. Entretanto, nesse trabalho será discutido apenas os resultados da questão

global, os gráficos desenvolvidos pela análise de uma competição específica podem ser vistos em [5].

A escolha pela linguagem Python se deve pela facilidade na implementação e execução do MCMC, bem como para a geração dos gráficos, ambas tarefas realizadas com utilização das ferramentas oriundas dos pacotes `Numpy`, `Matplotlib`, `Statistics` e `Scipy`.

3 Markov Chain Monte Carlo

Nessa seção iremos abordar a modelagem das Cadeias de Markov Monte Carlo no contexto das distribuições Poisson e Gamma de modo a encontrar uma distribuição a posteriori que busque explicar os dados observados. A escolha da distribuição de Poisson se dá pelo fato da mesma ser discreta e não negativa, assim como o fenômeno que estamos modelando. Já a Gamma foi escolhida como distribuição alternativa, uma vez que a mesma é contínua e estritamente positiva, pela qual é possível comparar o desempenho do MCMC com relação a outra distribuição de probabilidade.

3.1 Modelagem com a Poisson

Podemos iniciar a modelagem vendo os dados que dispomos. Dessa forma, analisando a quantidade de gols por partida nas competições da CBF de 2013 a 2020 temos o dado observado na Figura 1.

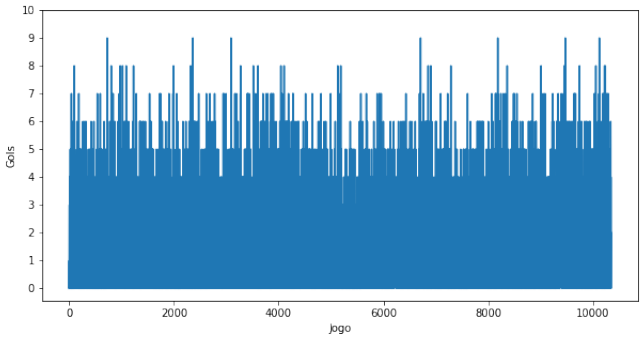

Figura 1: Gols por jogo no período de 2013 a 2020 nas competições da CBF.

Da forma como os dados estão expostos é difícil fazer qualquer análise sobre sua distribuição. Dessa forma, ao trazer os mesmos dados, mas em forma de histograma, pode-se ter uma análise mais eficiente de como os mesmos se comportam. Assim, a Figura 2 traz a informação de quantos jogos possuem cada quantidade de gols nesse mesmo período.

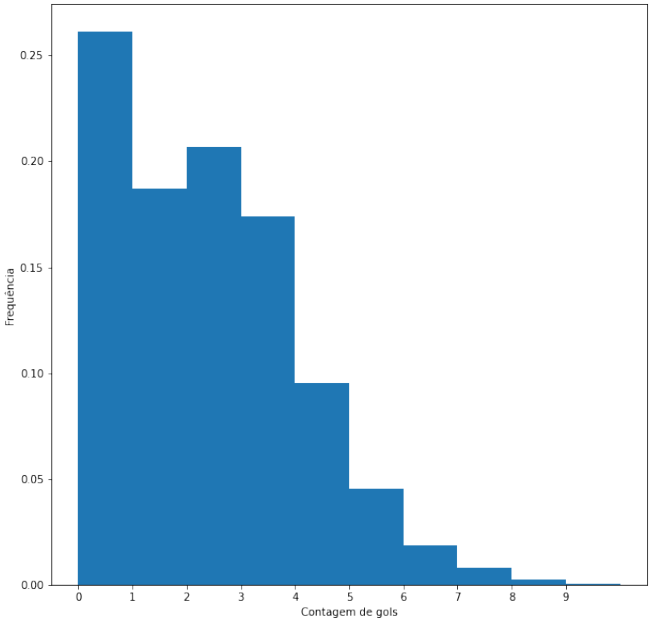

Figura 2: Histograma de jogos por quantidade de gols no período de 2013 a 2020 nas competições da CBF.

Pela Figura 2 pode-se observar uma certa similaridade dos dados com a Distribuição de Poisson para uma média λ. Esta similaridade não é uma mera coincidência, dado que podemos considerar independência entre os gols dentro de uma mesma partida e também independência entre as partidas, isto é, o resultado de um jogo não afeta o resultado de outro jogo. Dessa forma, a soma dos gols em uma partida também pode ser modelada como um processo de Poisson. Na literatura, pode-se citar [1] e [2] como dois dos trabalhos que se valeram dessa abordagem para a modelagem de gols.

Tendo visto que faz sentido modelar a distribuição do gols por meio de uma Distribuição

de Poisson, um estimador natural que surge para λ se dá pelo Estimador de Máxima Verossimilhança, o qual muito utilizado na estatística frequentista. Dessa forma, pode-se estimar

$$\lambda_{EMV} = \mu, \tag{1}$$

sendo μ a média dos dados. Utilizando λ_{EMV} como parâmetro da Poisson obtém-se a distribuição ilustrada na Figura 3.

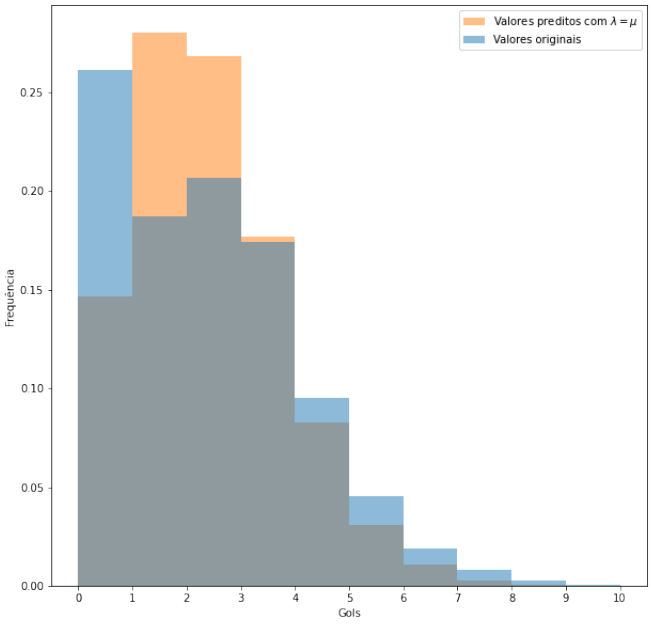

Figura 3: Distribuição obtida com λ_{EMV}.

Já a estimação se valendo do MCMC se dará por meio de um algoritmo iterativo. Para tanto, definem-se as seguintes funções:

- uma função de transição de estados, isto é, uma função que recebe uma proposta atual x para λ e retorna uma proposta x_n aleatória que, nesse estudo, se dá por $x_n \sim N(x, 0.5)$;
- uma função para calcular a log verossimilhança dos dados sob uma Distribuição de Poisso com parâmetro λ, a qual será representada por $L(\lambda)$;
- uma função de aceitação que calculará uma probabilidade p que, nesse estudo, será dada por

$$p = \begin{cases} 1, & \text{se } L(x_n) > L(x) \\ \exp\left(L(x_n) - L(x)\right), & \text{c.c.} \end{cases}. \tag{2}$$

  Feito isso, aceita-se a proposta x_n com probabilidade p.

Feitas tais definições, pode-se definir valor inicial para x e, em cada iteração do algoritmo, o mesmo realizará o seguinte procedimento:

- por meio da função de transição obtém-se x_n, uma nova proposta para λ;
- calcula-se $L(x)$ e $L\left(x_n\right)$;
- a função de aceitação verifica se aceita ou rejeita x_n;
- o valor de x_n é armazenado numa lista de aceitos/rejeitados, conforme função de aceitação, bem como numa lista de propostas.

Como resultado, o algoritmo retornará as listas citadas na última parte da iteração, possibilitando analisar o comportamento do algoritmo durante suas iterações.

3.2 Modelagem com a Gamma

Primeiramente, conforme citado no início desta seção, temos que a Distribuição Gamma é contínua e estritamente positiva, indo contra os dados disponíveis, que são discretos (a quantidade de gols sempre é um número inteiro) e que podem ser nulos (pode ocorrer empate de zero a zero).

Dessa forma, uma maneira de utilizar a Gamma para modelar essa quantidade de gols se dá com alguns ajustes como, por exemplo, incrementar a quantidade de gols em 0.1, fazendo com que os dados sejam estritamente positivos. Para resolver o problema da continuidade, pode-se analisar os valores à posteriori sob um histograma onde cada bin é o intervalo entre dois inteiros consecutivos ou, de forma equivalente, operar com o piso da distribuição.

Com esses ajustes, pode-se encontrar, de modo análogo a abordagem anterior, um estimador natural para a quantidade de gols. Isto é, por meio do Método dos Momentos, pode-se ver que

$$\alpha = \frac{\mu^2}{\sigma^2} \text{ e } \beta = \frac{\mu}{\sigma^2}, \tag{3}$$

sendo μ e σ^2 a média amostral e a variância amostral, respectivamente, apresentam uma boa estimativa para os parâmetros a serem procurados. A modelagem de uma Gamma com os parâmetros da Equação 3 resulta no modelo da Figura 4.

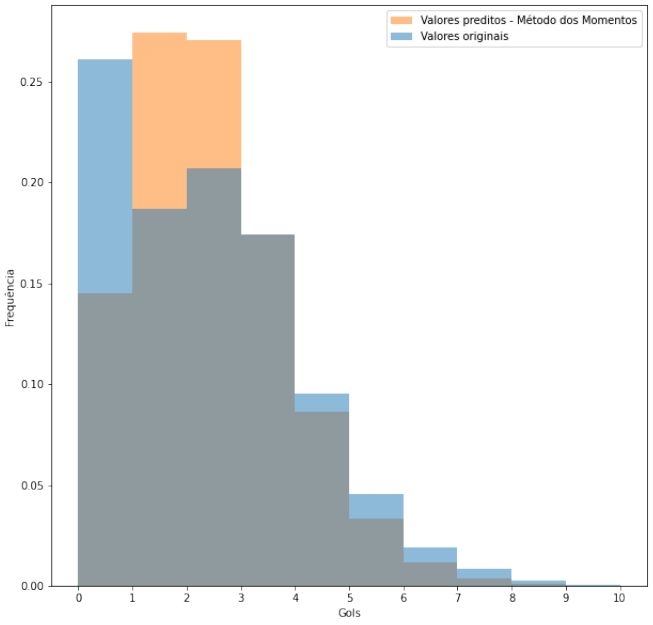

Figura 4: Distribuição obtida estimando α e β pelo Método dos Momentos.

Já a modelagem pelo MCMC poderá ser realizada de modo similar ao realizado com a Distribuição de Poisson, isto é, por meio de um processo iterativo similar.

Dentre as funções definidas anteriormente, ocorre modificações na função de transição de estados, que, ao contrário da abordagem anterior onde retornava uma proposta escalar, passará a retornar um vetor $x_n = (\alpha_n, \beta_n)$ de modo que, nesse trabalho, $\alpha_n \sim N(\alpha, 0.05)$ e $\beta_n \sim N(\beta, 5)$. E a função de log verossimilhança é modificada de modo que calcula-se a log verossimilhança de uma Distribuição Gamma de parâmetros $x_p = (\alpha_p, \beta_p)$, retornando $L(x_p)$. A função de aceitação permanece a mesma, isto é, a probabilidade de aceitar uma proposta continua sendo expressa pela Equação 2.

Definidas tais funções, pode-se definir os passos do algoritmo de modo análogo a modelagem com a Distribuição de Poisson.

4 Resultados

Nessa seção discutiremos alguns dos resultados observados durante a modelagem do fenômeno observado. Vale ressaltar que os dados observados contemplam todos os jogos de todas as competições da CBF, isto é, desde os jogos da Série A até os jogos da Série D e da Copa do Brasil. A quantidade de gols nos jogos observados irá variar conforme a competição. Por simplicidade, será aplicado o algoritmo sobre a distribuição ilustrada na Figura 2, entretanto as mesmas ideias se aplicam para as demais competições. Os gráficos dispostos nesse documento, bem como os gráficos da modelagem das demais competições, podem ser encontrados em [5].

4.1 Distribuição de Poisson

Aplicando a modelagem descrita na Seção 3.1 obtém-se, com $\lambda_0 = 10$ e 50.000 iterações, uma sequência de propostas similares a ilustrada na imagem superior da Figura 5. Já a imagem inferior destaca o caminho percorrido pelas propostas aceitas.

Já a Figura 6 traz as mesmas informações que a Figura 5, mas com um "burn-in" de 20% das iterações com a intenção de retirar da visualização as discrepâncias geradas por uma priori não informativa.

Tendo tais dados, pode-se analisar a região de aceitação para λ, isto é, a parte do espaço em que se concentram os valores para os quais a proposta λ é aceita. Com essa ideia, pode-se observar na Figura 7 a região de aceitação de λ, bem como sua distribuição. Novamente, considera-se os dados com o "burn-in".

Dessa forma, pode-se modelar o fenômeno observado tomando como λ o valor médio das pro-

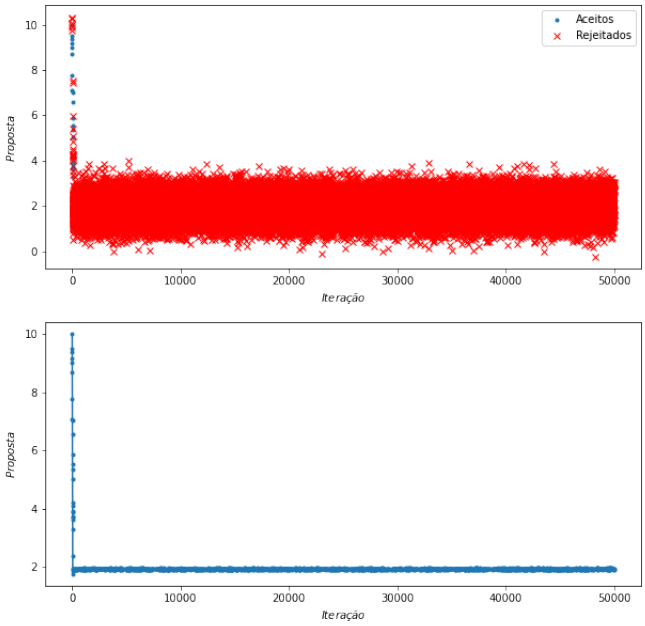

Figura 5: Na parte superior observam-se todas as propostas do algoritmo e na inferior o caminho percorrido pelas propostas aceitas.

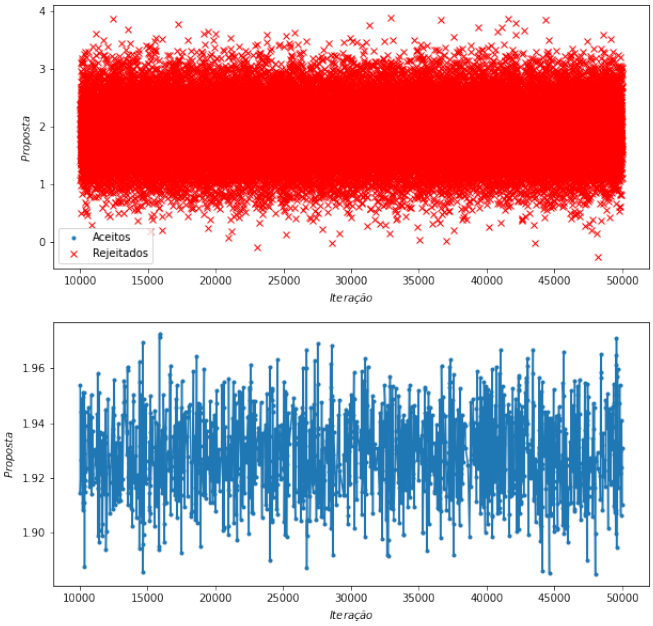

Figura 6: Na parte superior observam-se as últimas 40.000 propostas do algoritmo e na inferior o caminho percorrido pelas propostas aceitas dentre elas.

postas aceitas. Com isso, obtém-se a distribuição à posteriori dos dados, conforme ilustrado na Figura 8. A ideia de utilizar o valor médio se dá com o intuito de minimizar os efeitos gerados pela probabilidade de aceitar uma proposta caso o valor da log verossimilhança diminua.

Comparando a posteriori com a estimativa obtida com λ_{EMV}, isto é, as distribuições das Figuras 8 e 3, nota-se ambas parecem similares. Analisando o algoritmo utilizado pode-se perceber que isso é algo natural, uma vez que o algoritmo aceita uma sugestão λ_n se a log verossimilhança aumenta ou, caso contrário, com uma probabilidade igual a razão entre a verossimilhança em λ_n e λ.

4.2 Distribuição Gamma

Já a aplicação da modelagem com a Distribuição Gamma descrita na Seção 3.2 requer um pouco mais de atenção, uma vez que a mesma é mais complexa por requerer dois parâmetros. Dessa forma, usando como parâmetros iniciais o vetor $x_0 = (1, 1)$, isto é, $\alpha_0 = \beta_0 = 1$, bem como 100.000 iterações e o incremento citado de 0.1 na quantidade de gols por jogo, pode-se aplicar o algoritmo, o qual irá fornecer dados com os quais geram-se observações similares as realizadas com a modelagem pela Distribuição de Poisson. Dessa forma, a Figura 9 ilustra os passos do algoritmo.

De maneira análoga a modelagem anterior, o "burn-in" dos 20% primeiros passos poderá possibilitar uma análise mais concreta das iterações. Assim, a Figura 10 apresenta os últimos 80.000 passos do algoritmo.

A região de aceitação para uma proposta é um pouco mais complexa que no caso anterior, uma vez que $x \in \mathbb{R}^2$. Entretanto, por meio de um mapa de calor é possível visualizar a região de aceitação conjunta para α e β, conforme ilustrado na Figura 11.

Por fim, a distribuição à posteriori gerada pelo algoritmo aplicado a Distribuição Gamma pode

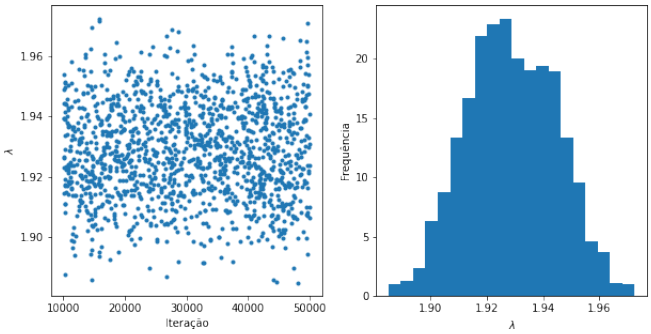

Figura 7: À esquerda observa-se a região de aceitação de λ, enquanto à direita observa-se a distribuição do mesmo.

ser dada por meio de uma Distribuição Gamma com parâmetros $\alpha = \mu_\alpha$ e $\beta = \mu_\beta$, onde μ_α e μ_β representam, respectivamente, a média dos α e dos β aceitos após aplicar o "burn-in". Novamente, a aplicação da média está relacionada com o intuito de minimizar os erros cometidos pelo fator aleatório na aceitação das propostas. A distribuição gerada por tais parâmetros pode ser encontrada na Figura 12.

5 Conclusão

O modelo dado pela Distribuição de Poisson modela a maior parte dos jogos possuindo 1, 2 ou 3 gols, acertando grande parte dessas predições, mas errando uma parte considerável da mesmas. Além disso, o modelo também acaba errando, para baixo, a quantidade de de partidas que terminam com mais de 3 gols ou as em que não há gols. Boa parte desses erros é irrelevante, dado que a taxa de acerto é alta, entretanto, quando se trata de um placar com mais de 6 gols o modelo já não é totalmente capaz de prever.

Em contrapartida, a modelagem dada pela Distribuição Gamma já resulta em um modelo que contempla mais facilmente placares com muitos gols, inclusive superestimando as probabilidades desses placares. Além dos placares com muitos gols, essa modelagem superestima placares magros, como zero a zero ou um a zero, com o modelo dando uma probabilidade muito acima da frequência observada nos dados ao primeiro destes placares. Já os placares com quantidade de

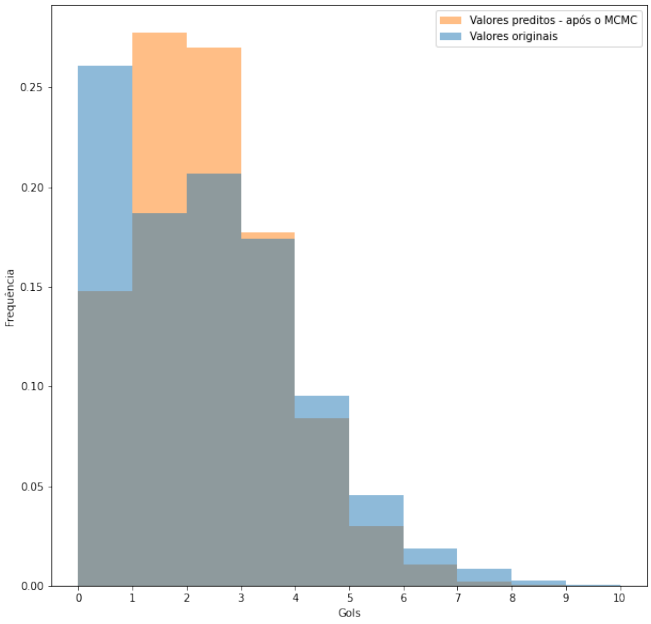

Figura 8: Distribuição a posteriori com distribuição de Poisson.

gols entre 2 e 5 são subestimados pelo modelo.

Comparativamente, ambos os modelos aparentam ter eficiência similar, se distinguindo apenas pelos pontos onde aparentam superestimar e subestimar os placares. Já levando em conta a eficiência do algoritmo, percebe-se que a modelagem pela Distribuição de Poisson é mais eficiente, uma vez que, por possuir apenas um parâmetro, apresenta bons resultados de convergência com menos iterações, enquanto o modelo com a Distribuição Gamma necessita de mais iterações para ter um resultado similar.

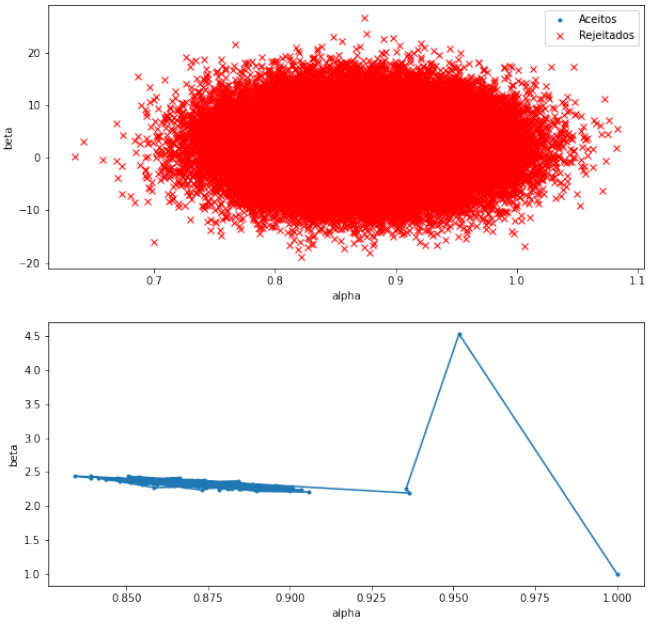

Figura 9: Na imagem superior observam-se todas as propostas do algoritmo e na inferior o caminho percorrido pelas propostas aceitas.

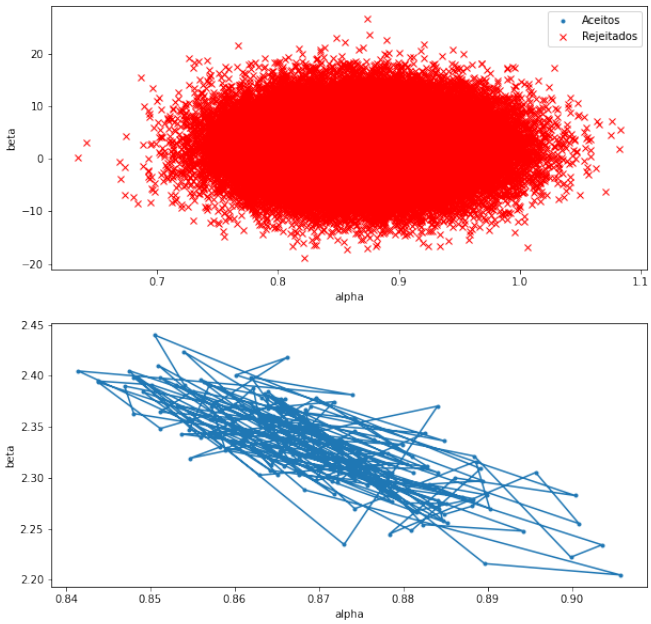

Figura 10: Na imagem superior observam-se as últimas 80.000 propostas do algoritmo e na inferior o caminho percorrido pelas propostas aceitas dentre elas.

Referências

[1] Inan, T. (2020). Using poisson model for goal prediction in European football. *Journal of Human Sport and Exercise, in press*. doi: `https://doi.org/10.14198/jhse.2021.164.16`.

[2] Maher, M. J. (1982). Modelling Association Football Scores.

[3] Moukarzel, J. (2019). Markov Chain Monte Carlo. *MCMC*. `https://github.com/Joseph94m/MCMC`.

[4] Michels, I. (2021). Brazilian Soccer. *Brazilian Soccer*. `https://github.com/IgorMichels/Brazilian_Soccer`.

[5] Michels, I. (2021). Statistical Modeling. *Statistical Modeling*. `https://github.com/IgorMichels/Statistical_Modeling`.

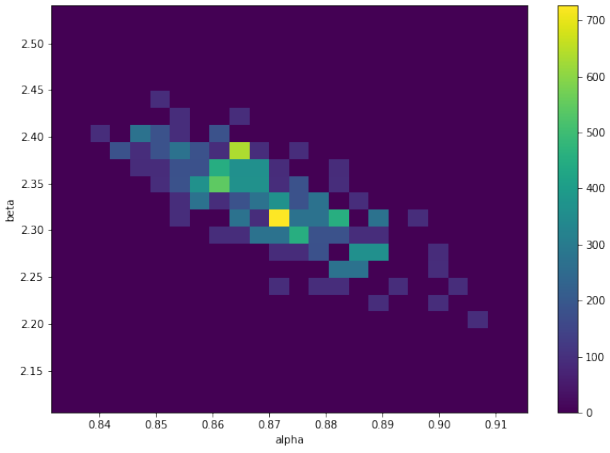

Figura 11: Região de aceitação conjunta de α e β.

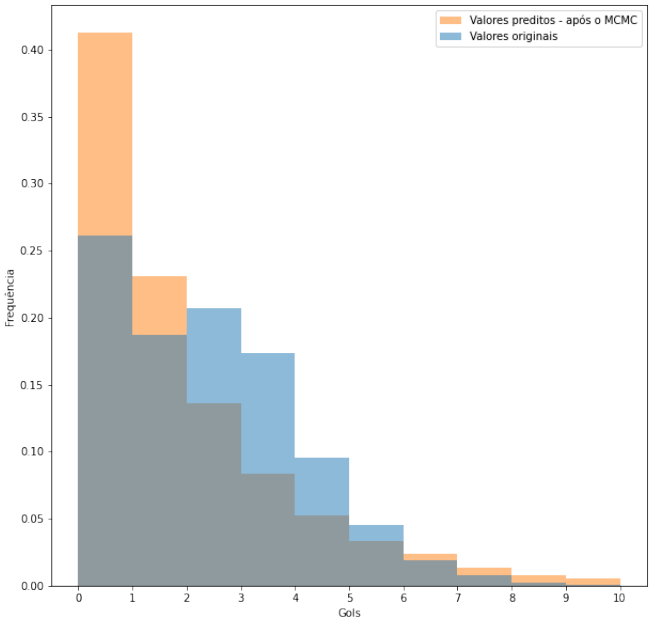

Figura 12: Distribuição à posteriori $Gamma(\mu_\alpha, \mu_\beta)$.