Trabalho 2ª VAE

de

Inferência Bayesiana

Escola Nacional de Ciências Estatísticas.

29 de julho de 2024

Tema: Técnicas de MCMC: Análise e Aplicabilidade em Séries Temporais de Streams de Música no Spotify

Alunos: Antonia Xavier

João Vitor Prisco da Silva.

# Introdução

Atualmente, a quantidade de streams de uma música é um parâmetro fundamental para determinar seu sucesso. No contexto das plataformas digitais, um stream (equivalente a unidade de ouvintes da música) é contabilizado cada vez que uma mídia é reproduzida. Entre as principais plataformas de streaming musical, destacam-se o Spotify, Youtube Music, Apple Music e Deezer, sendo o Spotify a maior no quesito número de usuários

Diversos indicadores do mundo do entretenimento utilizam o número de streams no Spotify como parâmetro para qualidade/sucesso de músicas por todo o mundo. Por exemplo, a famosa revista estadunidense Billboard, que lança semanalmente sua lista Billboard Hot 100, considerada a mais relevante para determinar quais músicas fizeram mais sucesso nos Estados Unidos a cada semana.

Dado o contexto é crucial conhecer a quantidade de streams de uma música em determinado período, pois isso permite quantificar seu sucesso. Tal informação é valiosa para profissionais da indústria musical e do entretenimento, e interessante para a sociedade em geral. De acordo com (Kortaba, 2017), “O estudo do consumo de música é importante para a sociologia, pois contém e reflete a linguagem, valores, sentimentos, preocupações e objetivos de uma pessoa”.

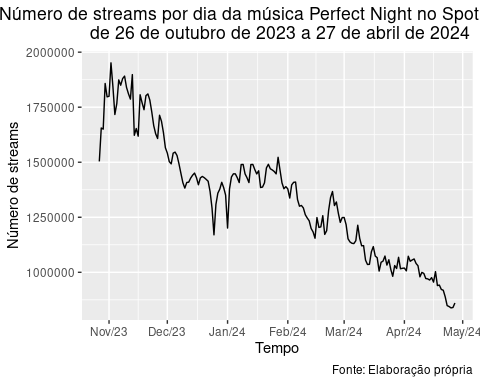
Com isso, este relatório tem como objetivo ajustar um modelo autorregressivo (AR (1)) à série temporal de streams diários do Spotify da música “Perfect Night”, do grupo sul-coreano Le Sserafim. A escolha da música se deu pelo gosto pessoal dos autores e pela fácil disponibilidade dos dados. No entanto, ao testar outras séries musicais, observou-se que o modelo AR (1) também se ajusta bem a elas, demostrando que o método pode ser aplicado a outras músicas de forma análoga.

# Dados

Cada dado da nossa série representa a soma de todas as reproduções da música ao longo de 24 horas. É relevante esclarecer que esse número não reflete o total de pessoas que ouviram a música na plataforma, pois um único usuário pode escutar a mesma música várias vezes. O Spotify, em sua documentação, também especifica que a música precisa ser ouvida por pelo menos 30 segundos para que o stream seja contabilizado.

Os dados utilizados neste estudo são provenientes da conta da rede social X, Le Sserafim on Spotify, de usuário @LSSRFMonSpotify, que compilou diariamente o número de streams de diversas músicas do grupo. Esta conta registrou os streams da música escolhida desde o seu lançamento em 26 de outubro de 2023 até o presente momento. No entanto, para facilitar a análise e a compilação manual dos dados a partir do X para o software R, consideramos os dados até 27 de abril de 2024, abrangendo um intervalo total de 184 dias (aproximadamente seis meses). O gráfico abaixo ilustra o número de streams diários ao longo do tempo.

*Gráfico 1: Série temporal do número de streams da música no Spotify por dia período de 26 de outubro de 2023 a 27 de abril de 2024.*



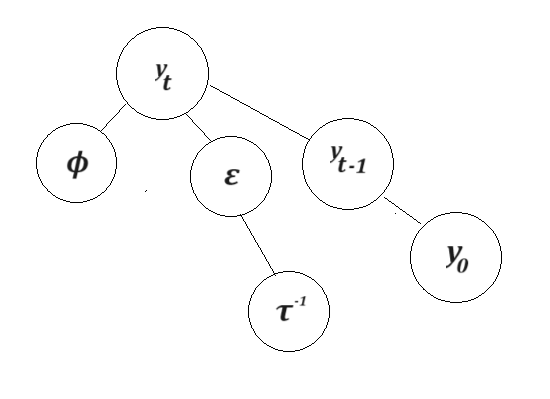
# Métodos

Queremos ajustar um modelo que seja capaz de prever o número de streams de um dia com base na informação do dia anterior. Como os nossos dados apresentam valores muito grandes, na casa dos milhões, dividimos toda a nossa série por mil. Consideramos essa transformação como nossa variável aleatória.

Onde t varia no conjunto discreto de 1 a 184. Queremos que nosso modelo se ajuste como um AR (1), ou seja:

Assim, podemos montar a seguinte hierarquia para o nosso modelo: depende , e que por sua vez depende de enquanto depende de y\_0.

*Figura 1: Relações hierárquicas dos hiperparâmetros do modelo.*



Iremos considerar independência a priori e definir como nosso vetor paramétrico onde . O núcleo da distribuição a posteriori é dado por:

Iremos gerar uma amostra da nossa posteriori. Consideramos as seguintes distribuições a priori pouco informativas:

Onde **,** é uma distribuição uniforme em -1 e 1 porque para o modelo autorregressivo ser estacionário temos essa condição, além disso, julgamos que pode assumir qualquer valor entre -1 e 1, sem qualquer tendência. No caso, **,** pode ser interpretado como o número de streams um dia antes do lançamento da música. Como é uma normal, faz sentido ser uma normal também. Em relação a sua média e precisão usamos a média e precisão de outra música do mesmo artista na semana de lançamento, no caso a música Unforgiven. Para tínhamos a restrição que a variância precisa ser positiva e fomos testando valores. Faz sentido nossa variância ser alta pois a escala dos nossos dados é bem grande, então a média é grande, assim, faz sentido a variância ser na mesma grandeza, portanto será muito pequeno.

Dessa forma, como  possuí distribuição normal, é fácil ver que ***yt***terá uma distribuição função de verossimilhança para o nosso AR (1) é da forma:

Dessa forma, como as condicionais completas são completamente conhecidas podemos usar o amostrador de Gibbs para se obter uma amostra da distribuição a posteriori de . O algoritmo de Gibbs para obter uma amostra da distribuição conjunta de é:

1. Determinar um valor inicial arbitrário para . Chamemos esses valores de
2. Iniciar o contador no i = 1
3. Obter um novo valor através de gerações sucessivas de distribuições
4. Altera-se o contador i para i + 1
5. Repete-se o item 3 e 4 até obter convergência

# Resultados

Utilizamos o pacote JAGS no R para aplicar o algoritmo de Gibbs aos nossos dados, empregando a função de verossimilhança e as distribuições a priori informadas anteriormente. A documentação do pacote recomenda o uso de pelo menos três cadeias. Iniciamos nossa simulação com quatro cadeias em paralelo, com a intenção de aumentar o número de cadeias caso encontrássemos problemas de convergência. No entanto, conseguimos bons resultados com quatro cadeias.

Inicialmente, realizamos 3000 iterações e obtivemos convergência. Contudo, o número final de cadeias não foi suficiente para aplicar o critério de Raftery e Lewis (1992) de análise de convergência. Portanto, aumentamos para 9000 iterações, utilizando 20% desse valor como burn-in (1800 iterações de burn-in). Assim, terminamos com 7200 iterações válidas, das quais 3600 foram salvas.

Os gráficos abaixo mostram as amostras geradas. Através da inspeção visual, podemos facilmente observar que todas as cadeias convergiram. A figura ilustra as amostras geradas para e . A linha vermelha representa a média, que pode ser usada como estimador pontual para esses parâmetros

Utilizamos o pacote JAGS no R para aplicar Gibbs aos nossos dados utilizando a função de verossimilhança e as distribuições a priori informadas acima. A documentação do pacote utilizado recomenda que se utilize pelo menos 3 cadeias. Iniciamos nossa simulação com 4 cadeias em paralelo com a ideia de aumentar o número de cadeias caso tenhamos algum problema de convergência, entretanto conseguimos bons resultados com 4 cadeias. Da mesma forma, começamos com 3000 iterações e conseguimos convergir, mas o número final de cadeias não foi o suficiente para usar o critério de Raftery e Lewis (1992) de análise de convergência. Utilizamos então 9000 iterações e 20% desse valor como burn-in (1800 de burnin). Terminamos então com 3600 iterações salvas.

Abaixo vemos os gráficos das amostras geradas. Fica fácil ver pelo critério visual que todas convergiram. A figura abaixo mostra a amostra gerada para e e a linha vermelha representa a média que pode ser usada como estimador pontual para eles.

*Figura 2: Amostra da posteriori de por iteração, com a média destacada em vermelho.*

Uma imagem contendo Gráfico

Descrição gerada automaticamente

Além da análise visual, podemos usar os critérios de Raftery e Lewis e Geweke para a análise de convergência de nossas cadeias. A tabela abaixo mostra os resultados utilizando o critério de Raftery por meio da função ‘raftery.diag’ do R.

***Tabela 1****: Critério de Raftery*

| **Variável** | **Burn-in (M)** | **Total (N)** | **Lower bound (NMin)** | **Dependence factor (I)** |
| --- | --- | --- | --- | --- |
|  | 2 | 3787 | 3746 | 0.988 |
|  | 2 | 3787 | 3746 | 0.977 |
|  | 2 | 4029 | 3746 | 0.997 |

Utilizando o critério de Raftery e Lewis (1992), que estima quantas iterações são necessárias para que o amostrador de Gibbs apresente convergência à distribuição estacionária indicou boa convergência das cadeias MCMC para os parâmetros ***ϕ, τ-1*** e ***y0***​. O burn-in (aquecimento) é pequeno (2-3 iterações) e o total de iterações necessárias é razoável (3787-4029). O fator de dependência é baixo (1.01-1.08), sugerindo que a autocorrelação seja baixa. Abaixo vemos o gráfico de autocorrelação dos nossos parâmetros. Que sugere só haver autocorrelação significativa no Lag 1, que é o que se espera de um AR (1).

*Gráfico 2: Autocorrelações dos parâmetros estimados.*

Gráfico, Histograma

Descrição gerada automaticamente

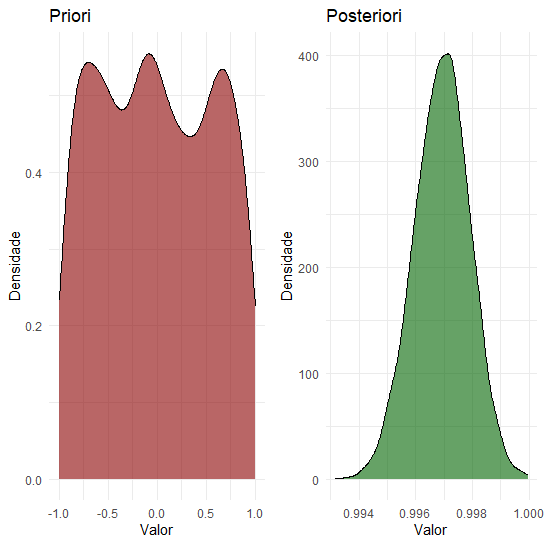
Analisamos a convergência também pelo critério de Geweke (1992) cujo diagnóstico de convergência para cadeias de Markov é baseado em um teste de igualdade das médias das primeiras 10% cadeias e da última metade das cadeias. O diagnóstico pressupõe que se as amostras são extraídas da distribuição estacionária da cadeia, as duas médias são iguais. Os valores abaixo indicam os valores encontrados no critério de Geweke para as nossas cadeias usando a função ‘Geweke.diag’ no R.

*Tabela 2: Saídas critério de Geweke.*

|  | **Diagnóstico de Geweke** |
| --- | --- |
|  | 0.5161 |
|  | 0.1582 |
|  | 0.7992 |

Considerando um nível de significância de 5%, não existem evidências contra a convergência dos parâmetros  *.*

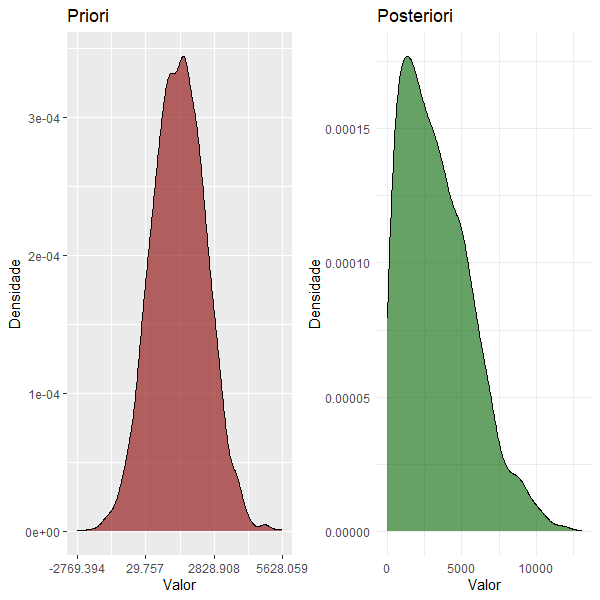
*Gráfico 3: Comparação visual entre a Distribuição a Priori e a Posteriori, do parâmetro ϕ.*



Observando as duas distribuições, levamos a crer que no caso da distribuição a priori, é indicado uma incerteza ou falta de informação prévia, uma vez que não observamos os dados previamente, tratando assim valores entre –1 e 1 com probabilidades similares.

No caso da distribuição a posteriori, a distribuição Normal resultante, representada por um formato de sino, centrada em valores perto de 1, nos mostra que a maior densidade de probabilidade está concentrada ao redor deste valor, ilustrando claramente como a evidência leva a uma redução drástica na incerteza, transformando a crença inicial ampla e uniforme, em uma crença posterior muito mais precisa e concentrada.

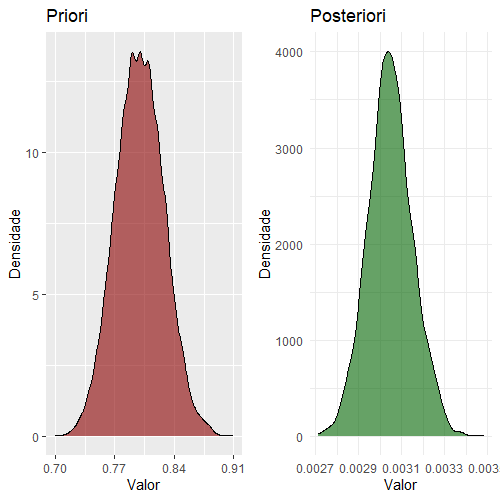
*Gráfico 4: Comparação visual entre a Distribuição a Priori e a Posteriori, do parâmetro Y0.*



A mudança de uma densidade a priori extremamente concentrada para uma densidade a posteriori com um intervalo HPD muito mais amplo indica que, apesar dos dados sustentarem a estimativa central original, eles também introduzem uma considerável incerteza. Isso pode ocorrer em situações em que os dados observados são muito variáveis ou não são fortemente informativos, obrigando a revisão da certeza inicial expressa pela distribuição a priori.

Esta atualização reflete a natureza adaptativa da inferência bayesiana, em que as crenças iniciais são ajustadas para incorporar a evidência empírica, resultando em uma visão mais abrangente e, neste caso, mais incerta sobre o parâmetro em questão.

*Gráfico 5: Comparação visual entre a Distribuição a Priori e a Posteriori, do parâmetro τ -1.*



A mudança substancial dos parâmetros sugere que os dados observados foram altamente informativos e indicaram que o valor verdadeiro do parâmetro é muito menor do que inicialmente acreditávamos com base na priori.

Sob perda quadrática, podemos dizer que a média da nossa posteriori é o nosso estimador para o vetor . Utilizando esses valores nosso modelo ajustado para o AR (1) ficou da forma:

A tabela abaixo mostra a mediana e o intervalo de alta densidade a posteriori (intervalo HPD) com 95% de credibilidade encontrado para e aproximado para quatro casas decimais.

*Tabela 3: Exibição da mediana e intervalos HPD dos parâmetros.*

|  | **Mediana** | **Intervalo HPD** |
| --- | --- | --- |
|  | 0.9969 | [0.9949, 0.9987] |
|  | 1388.866 | [-829.1377,3460.345] |
|  | 0.0030 | [0.0028,0.0032] |

Podemos interpretar esses intervalos da seguinte forma: há 95% de probabilidade de que o verdadeiro valor do parâmetro esteja entre 0.9949 e 0.9987, de que o valor verdadeiro do parâmetro y0 esteja entre -829.1377 e 3460.345. há 95% de probabilidade de que o verdadeiro valor de esteja entre .0028 e 0.0032 O gráfico abaixo mostra o histograma dos parâmetros com o intervalo HPD e a média em destaque em verde.

*Gráfico 6: Histogramas dos parâmetros com seus respectivos intervalos HPD.*

Gráfico, Histograma

Descrição gerada automaticamente

# Previsão

Usando o modelo ajustado, podemos prever valores futuros para a nossa série. Retirando a última observação e usando a equação do modelo prevemos como , enquanto o valor real é 860.646. Uma previsão bem satisfatória. Podemos de forma análoga gerar nossa distribuição preditiva para o , sendo ela representada pelo gráfico a seguir:

*Gráfico 7: Histograma da distribuição preditiva para* ***.***

Gráfico, Histograma

Descrição gerada automaticamente

**Conclusão**

Em resumo, este estudo proporcionou uma análise robusta do comportamento dos streams diários da música "Perfect Night" no Spotify, utilizando um modelo AR (1) eficazmente ajustado e validado. A aplicação do método de Gibbs foi relativamente simples e eficiente na obtenção dos parâmetros do modelo, com um número reduzido de iterações. Além disso, observamos que o modelo AR (1) demonstrou ser adequado para diversas outras músicas testadas, evidenciando sua aplicabilidade generalizada.

# 

# Referências

**KOTARBA**, Joseph A. **Understanding society through popular music**. Routledge, 2013.

***Geweke.diag: Geweke’s convergence diagnostic****, rdrr.io, 1992, disponível em: https://rdrr.io/cran/coda/man/geweke.diag.html*

**GAZZINELLI**, Rafael Moraes. Inferência Bayesiana na análise de dados de experimentos planejados. 2013.

Apostila de Aula.