UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
Departamento de Métodos Estatísticos - IM
Estatística computacional

TÍTULO
Projeto final

Aluno: Silvaneo Vieira dos Santos Junior
Professor: Carlos Tadeu Pagani Zanini

29 de dezembro de 2022

ÍNDICE

1 Introdução

2 Metodologia

Neste trabalho estaremos interessados em avaliar o ajuste de Modelos Lineares Generalizados (GLM) usando uma variação da metodologia proposta em Marotta, Alves, and Migon (2022). Nesta sessão apresentaremos a base da metodologia utilizada, começando por uma breve introdução aos GLM, em seguida apresentaremos uma versão simplificada do Teorema da Projeção (Amari 2016) e a forma como aplicaremos esse teorema para o problema que desejamos resolver. Finalizamos essa sessão com algumas considerações sobre algumas formas de se calcular as integrais necessárias para a aplicação da metodologia.

2.1 Modelos Lineares Generalizados

A classe dos Modelos Lineares Generalizados (McCulloch and Searle 2001; Dobson and Barnett 2018) é uma classe de modelos muito ampla e que generaliza os Modelos Lineares com resposta Normal. De modo geral, vamos assumir que temos um conjunto de observações Y_i, tais que:

$$\begin{aligned} Y_i|\theta &\sim F(\lambda_i), \\ g(\lambda_i) &= \eta = x_i'\theta, \end{aligned} \tag{1}$$

onde g, chamada função de ligação, é uma função contínua e monótona, η é o preditor linear, X é a matriz de planejamento, x_i é a i-ésima coluna de X, θ (possivelmente um vetor) são os parâmetros latentes que representam o efeito das variáveis explicativas e F representa uma distribuição pertencente a família exponencial e indexada pelo parâmetro λ_i (possivelmente um vetor). Nesse trabalho sempre vamos considerar que g e X são conhecidos.

Sendo F pertencente a família exponencial, então temos que, por definição, podemos escrever a densidade de $Y_i|\theta$ como:

$$f(y_i|\theta) = \exp\left\{\lambda_i \cdot H(y_i) - A(\lambda_i) + B(y_i)\right\}, \tag{2}$$

onde as funções H, A e B são conhecidas. Neste trabalho o vetor $H(y_i)$ será chamado de vetor de estatísticas suficientes.

Adiante apresentamos um importante resultado associado as distribuições pertencentes à família exponencial (Migon, Gamerman, and Louzada 2014):

Teorema 1: *Seja Y_i com distribuição pertencente à família exponencial conforme (2), então vale que:*

$$\mathbb{E}[H_j(y_i)] = \frac{\partial A(\lambda_i)}{\partial \lambda_{ij}}. \tag{3}$$

Alguns exemplos de distribuições pertencentes à família exponencial são: Normal com média e variância desconhecidas, Gamma com parâmetro de forma e escala desconhecidos, Beta, Multinomial com parâmetro de tentativas conhecido, Binomial Negativa com número de fracassos conhecido, Poisson, Geométrica, Rayleigh, Pareto com locação conhecida e Pareto Assimétrica com locação conhecida. Neste trabalho estaremos especialmente interessados nos casos Normal com média e variância desconhecidas, Rayleigh e Pareto Assimétrica com locação conhecida.

Abordaremos neste trabalho como realizar a análise Bayesiana para dados provenientes de um modelo observacional conforme especificado em (1). Para isso, vamos especificar uma priori π_0 para θ e, para realizar qualquer que seja a análise, devemos encontrar a distribuição a posteriori de θ (π_n). Como pode ser visto em Migon, Gamerman, and Louzada (2014), podemos escrever π_n como:

$$\pi_n(\theta|y_1, ..., y_n) = \frac{\pi_0(\theta)\prod_{i=1}^n f(y_i|\theta)}{\int \pi_0(\theta)\prod_{i=1}^n f(y_i|\theta)d\theta}. \tag{4}$$

Infelizmente, a menos do caso onde F representa a distribuição Normal com varância conhecida, não é possível obter uma solução analítica para $\int \pi_0(\theta) \prod_{i=1}^n f(y_i|\theta) d\theta$, sendo necessário recorrer a métodos de integração numérica para obter $\pi_n(\theta|y_1, ..., y_n)$ ou estimativas de θ. Tendo em mente que métodos de integração numérica podem ter um custo computacional muito elevado, especialmente quando a dimensão de θ é grande, Marotta, Alves, and Migon (2022) propõe o uso de aproximações para π_n de modo que possamos obter uma forma analítica aproximada para a posteriori de θ. No trabalho original, a metodologia é apresentada para Modelos Dinâmicos Lineares Generalizados, porém os GLM são um caso particular desta classe, de modo que a metodologia proposta também é válida para os modelos em que estamos interessados neste trabalho, contudo, usaremos uma versão modificada desta metodologia, de modo que não discutiremos o trabalho original de Marotta, Alves, and Migon (2022), mas apresentaremos diretamente a versão modificada (que é mais simples).

2.2 Inferência aproximada via Teorema da Projeção

Como visto na sessão anterior, para realizar o processo de inferência, precisamos obter $\pi_n(\theta|y_1, ..., y_n) = \frac{\pi_0(\theta) \prod_{i=1}^n f(y_i|\theta)}{\int \pi_0(\theta) \prod_{i=1}^n f(y_i|\theta) d\theta}$, sendo que nem sempre podemos obter uma forma analítica fechada para π_n. Seguindo a abordagem proposta em Marotta, Alves, and Migon (2022), para solucionar este problema, iremos aproximar π_n por uma $\widehat{\pi}_n$ que seja próxima da posteriori verdadeira, mas que tenha propriedades úteis, por exemplo, tal que $\int \pi_0(\theta) \prod_{i=1}^n f(y_i|\theta) d\theta$ seja tratável. Para a escolha de $\widehat{\pi}_n$, vamos escolher, dentro de uma família de distribuições, aquela que minimiza a divergência de Kullback-Leibler (KL), definida como:

$$KL(p||q) = \mathbb{E}_p[\ln(f_q(X)) - \ln(f_p(X))], \tag{5}$$

onde p e q são distribuições de probabilidade, f_p e f_q são funções de densidade ou de massa de probabilidade associadas, respectivamente, a p e q e $\mathbb{E}_p$ representa o valor esperado calculado considerando que X tem distribuição p.

Para os fins deste trabalho, é natural escolher $\widehat{\pi}_n$ tal que $KL(\pi_n, \widehat{\pi}_n)$ é minimal (ver capítulo 5 de MacKay (2002) para uma interpretação intuitiva da divergência KL). Por conveniência, vamos escolher $\widehat{\pi}_n$ como pertencente à família Normal, de modo que, dado π_n, basta encontrar os parâmetros de média e variância para $\widehat{\pi}_n$ tais que $KL(\pi_n, \widehat{\pi}_n)$ é otimal. Para auxiliar no processo de encontrar os parâmetros ótimos de $\widehat{\pi}_n$, podemos usar o seguinte teorema (ver Amari (2016)):

Teorema 2 (da Projeção): \emph{ Seja p e q duas distribuições de variáveis aleatórias contínuas [1] pertencentes à família exponencial, conforme (2), isto é, existem densidades de probabilidade f_p e f_q associadas, respectivamente, a p e q tais que:

$$f_p(x) = \exp\{\lambda_p \cdot H_p(x) - A_p(\lambda_p) + B_p(x)\} \, f_q(x) = \exp\{\lambda_q \cdot H_q(x) - A_q(\lambda_q) + B_q(x)\}. \tag{6}$$

Então, fixados os parâmetros λ_p associados à distribuição p, os parâmetros λ_q da distribuição q que minimizam a divergência KL são únicos (quando existem) e satisfazem o seguinte sistema:

$$\mathbb{E}_p[H_q] = \mathbb{E}_q[H_q]. \tag{7}$$

Vale ainda que, se existe λ_q que satisfaz o sistema acima, então existe um mínimo para a divergência KL com relação a λ_q e a solução para o sistema 7 é única. }

Apresentaremos a prova parcial do Teorema 2, pois a enunciação do Teorema da Projeção apresentada em Amari (2016) é muito mais geral do que o teorema que usaremos. Vamos nos limitar a provar que os parâmetros λ_q que minimizam a divergência KL satisfazem 7, sendo que os argumentos para a existência e unicidade do mínimo e da unicidade da solução do sistema podem ser encontrardo em Amari (2016).

Para provar o Teorema 2, primeiro observe que a divergência KL de p com relação a q pode ser escrita como:

[1] O Teorema ainda vale sem essa restrição

$$\begin{aligned} KL(p||q) &= \mathbb{E}_p[\lambda_q \cdot H_q(X) - A_q(\lambda_q) + B_q(X) - (\lambda_p \cdot H_p(X) - A_p(\lambda_p) + B_p(X))]) \\ &= \lambda_q \cdot \mathbb{E}_p[H_q(X)] - A_q(\lambda_q) + \mathbb{E}_p[B_q(X)] - \lambda_p \cdot \mathbb{E}_p[H_p(x)] + A_p(\lambda_p) - \mathbb{E}_p[B_p(X))]. \end{aligned} \tag{8}$$

Veja que, se λ_q minimiza $KL(p||q)$, usando propriedades da divergência KL na família exponencial (i.e., que ela é contínua e duas vezes diferenciável em relação à λ_q), temos que:

$$\frac{\partial}{\partial \lambda_{qi}} KL(p||q) = 0, \forall i. \tag{9}$$

Mas veja que:

$$\frac{\partial}{\partial \lambda_{qi}} KL(p||q) = \mathbb{E}_p[H_{qi}(X)] - \frac{\partial}{\partial lambda_{qi}} A_q(\lambda_q), \tag{10}$$

pois os valores esperados com relação a p não dependem de λ_q.

Daí obtemos que, se λ_q minimiza $KL(p||q)$, então:

$$\mathbb{E}_p[H_{qi}(X)] = \frac{\partial}{\partial \lambda_{qi}} A_q(\lambda_q), \forall i. \tag{11}$$

Usando a equação (3) do Teorema 1 e lembrando que a equação acima deve valer para todas as coordenadas de λ_q, obtemos que:

$$\mathbb{E}_p[H_q(X)] = \mathbb{E}_q[H_q(X)].$$

E isso conclui a prova parcial do Teorema 2. □

No caso específico em que estamos trabalhando, vamos tomar q como pertencente à família Normal (em alguns casos, multivariada), de modo que:

$$H_q(X) = (X, XX')'. \tag{12}$$

Na prática, isso significa que a q que melhor aproxima p é aquela que tem o mesmo vetor de médias e a mesma matriz de covariância.

A princípio, a metodologia descrita até o momento parece bastante simples: Basta aproximar π_n por $\widehat{\pi}_n$ e realizar toda a inferência com base nesta última distribuição que, pertencendo à família Normal, é bem fácil de se trabalhar. Porém, talvez o leitor tenha observado que há uma certa incosistência na metodologia: Não podemos calcular π_n ,pois não temos forma analítica fechada para $\int \pi_0(\theta) \prod_{i=1}^n f(y_i|\theta) d\theta$, então iremos aproximar a posteriori verdadeira por $\widehat{\pi}_n$, sendo que, para isto, devemos calcular $\mathbb{E}_p[H_q(X)]$. Ora, se não conseguimos calcular $\int \pi_0(\theta) \prod_{i=1}^n f(y_i|\theta) d\theta$, não é razoável assumir que conseguimos calcular $\mathbb{E}_p[H_q(X)]$! De certa forma, estaríamos trocando um problema por outro de igual complexidade (se não maior!).

Felizmente, trabalhar com $\mathbb{E}_p[H_q(X)]$ não é tão problemático quanto trabalhar diretamente com π_n, pois precisamos apenas do valor de $\mathbb{E}_p[H_q(X)]$, e não de uma expressão analítica para ele, Desta forma, podemos calcular $\mathbb{E}_p[H_q(X)]$ usando métodos de integração numérica, especificamente, vamos trabalhar com 4 métodos para calcular os valores esperados desejados: Quadratura Gaussiana, Aproximação de Laplace, Monte Carlo e pelo método proposto em Tierney and Kadane (1986) e refinado em TierneyKadane2. Usaremos Quadratura Gaussiana nos Casos Normal com variância desconhecida e Rayleight, pois nestes casos temos de lidar com integrais univariadas, de modo que o custo computacional de se usar Quadratura Gaussiana é despresível. Para o caso Laplace Assimétrica, usaremos os outros métodos afim de conseguir um ajuste satisfatório.

3 Qualidade da aproximação

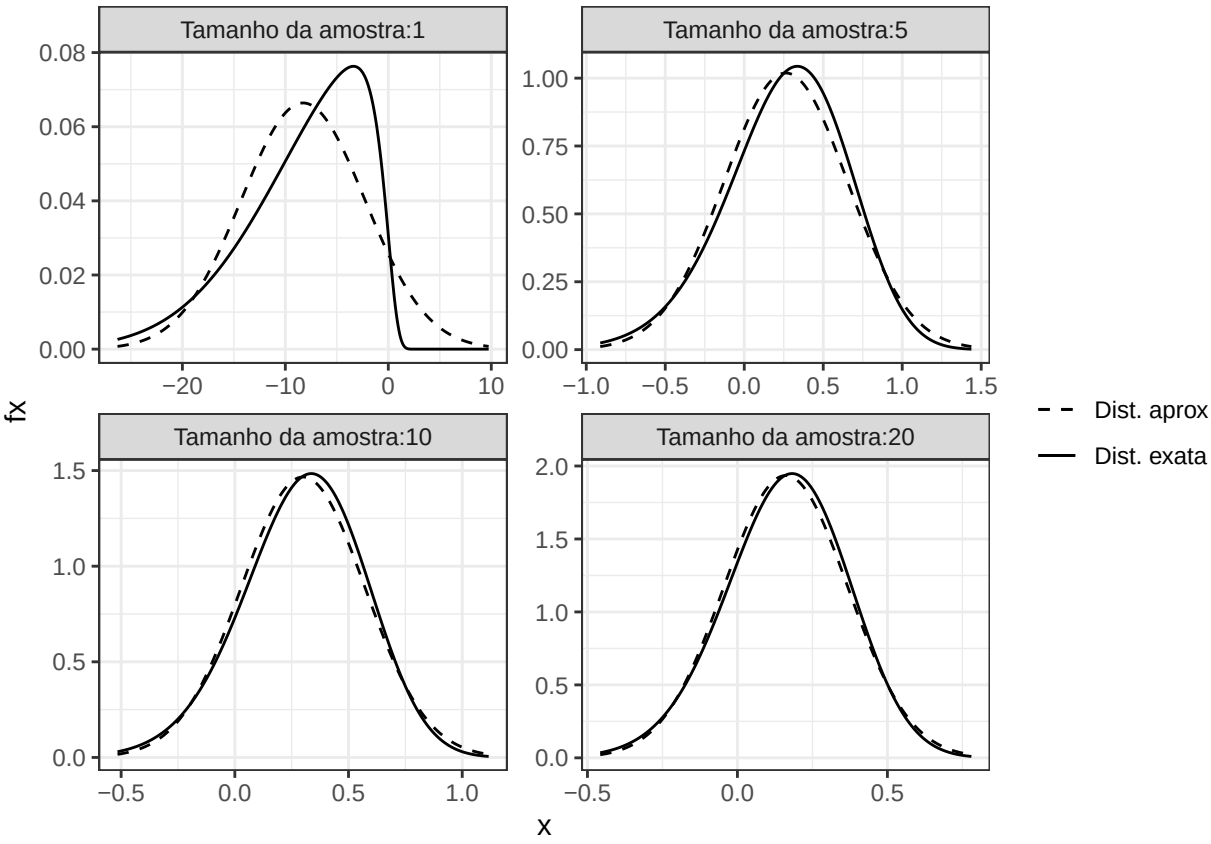
Como visto na sessão anterior, o Teorema 2 no fornece uma forma simples de aproximar uma distribuição de probabilidade por outra, desde que ambas pertençam à família exponencial, especificamente, esse teorema nos diz qual é **a melhor** distribuição que aproxima nossa posteriori. Naturalmente, **a melhor** distribuição pode não ser boa, por isso, é necessário alguma investigação sobre o assunto. Felizmente, graças as propriedades da família exponencial (ver Amari (2016) e, para o caso mais geral, Tierney and Kadane (1986) e Migon, Gamerman, and Louzada (2014)), temos que a aproximação será muito boa desde que o tamanho da amostra seja suficientemente grande.

Para exemplificar o comportamento descrito acima, vamos exibir adiante uma comparação entre as posterioris aproximadas para vários tamanhos de amostra. Neste caso vamos supor um modelo muito simples:

$$Y_i|\theta \sim Poisson(\lambda), \ln(\lambda) = \theta, \theta \sim \mathcal{N}(0, 100), \tag{13}$$

onde tomaremos $\lambda = 1$ para gerar os dados.

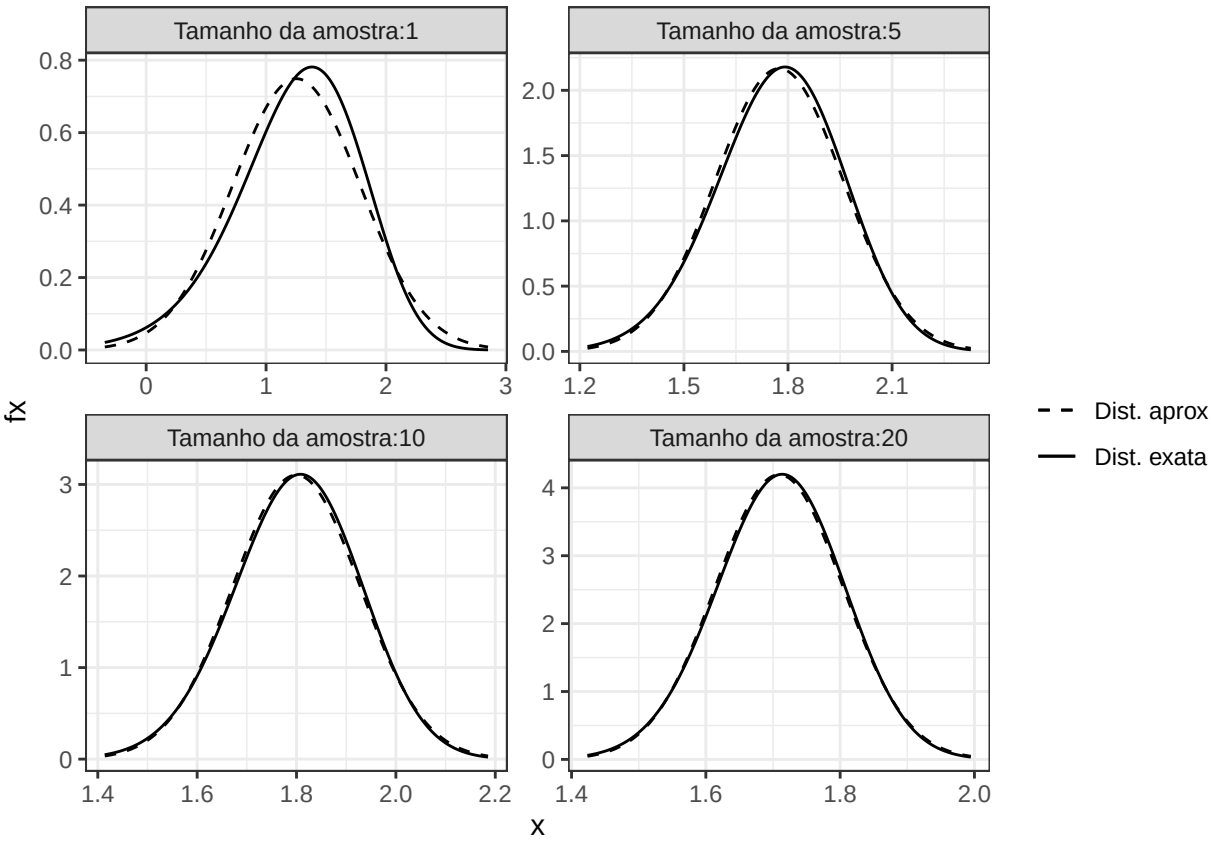
```
set.seed(13031998)

data_plot=data.frame()

for(n in c(1,5,10,20)){
y=rpois(n,1)
y_stat=sum(y)

f=function(x){exp(y_stat*x-n*exp(x)-(x**2)/(2*100))}
c=integrate(f,-Inf,Inf)$value
x_mean=integrate(function(x){x*f(x)},-Inf,Inf)$value/c
x2_mean=integrate(function(x){(x**2)*f(x)},-Inf,Inf)$value/c
s2=x2_mean-x_mean**2

x=seq(x_mean-3*sqrt(s2),x_mean+3*sqrt(s2),l=1000)

data_plot=rbind(data_plot,data.frame(x=x,fx=f(x)/c,gx=dnorm(x,x_mean,sqrt(s2)),n=n))
}

data_plot$n=factor(paste0('Tamanho da amostra:',data_plot$n),levels=paste0('Tamanho da amostra:',c(1,5,

ggplot(data_plot)+
  geom_line(aes(x=x,y=fx,linetype='Dist. exata'))+
  geom_line(aes(x=x,y=gx,linetype='Dist. aprox'))+
  scale_linetype_manual('',values=c('dashed','solid'))+
  theme_bw()+
  facet_wrap(.~n,scales = 'free')
```



Veja que temos uma aproximação com qualidade muito boa, mesmo para amostras relativamente pequenas. Vale observar que, no caso da distribuição Poisson, a qualidade da aproximação dependende da magnitude dos dados observados. De modo geral, a qualidade vai depender da quantidade de informação na amostra, sendo que é fácil ver (via, por exemplo, a informação de Fisher) observações de valores maiores tem muito mais informação do que obserções de valores menores (especialmente o 0). O caso que mostramos anteriormente seria uma caso **ruim**, pois a taxa verdadeira da Poisson (isto é, a taxa usada para gerar os dados) foi igual a 1. Adiante, vamos mostrar o mesmo exemplo, mas agora gerando dados de uma Poisson com taxa 5 (que ainda é um valor relativamente baixo):

```
set.seed(13031998)

data_plot=data.frame()

for(n in c(1,5,10,20)){
y=rpois(n,5)
y_stat=sum(y)

f=function(x){exp(y_stat*x-n*exp(x)-(x**2)/(2*100))}
c=integrate(f,-Inf,Inf)$value
x_mean=integrate(function(x){x*f(x)},-Inf,Inf)$value/c
x2_mean=integrate(function(x){(x**2)*f(x)},-Inf,Inf)$value/c
s2=x2_mean-x_mean**2

x=seq(x_mean-3*sqrt(s2),x_mean+3*sqrt(s2),l=1000)

data_plot=rbind(data_plot,data.frame(x=x,fx=f(x)/c,gx=dnorm(x,x_mean,sqrt(s2)),n=n))
}
```

```
data_plot$n=factor(paste0('Tamanho da amostra:',data_plot$n),levels=paste0('Tamanho da amostra:',c(1,5,

ggplot(data_plot)+
  geom_line(aes(x=x,y=fx,linetype='Dist. exata'))+
  geom_line(aes(x=x,y=gx,linetype='Dist. aprox'))+
  scale_linetype_manual('',values=c('dashed','solid'))+
  theme_bw()+
  facet_wrap(.~n,scales = 'free')
```



Observe que, agora, a aproximação é razoável mesmo para uma amostra com apenas 1 elemento, sendo que ela é praticamente idêntica à distribuição exata para amostras de tamanho maior que 10.

O modelo descrito em (13) é útil para exemplificar propriedades gerais da aproximação, porém ele não representa bem o tipo de modelo que gostaríamos de ajustar em problemas reais, uma vez que, nesta especificação, as observações y_i são i.i.d.. De modo geral, estaremos interessados em um modelo da forma descrita em (1), onde teremos um conjunto de covariáveis das quais queremos estimar o efeito. No exemplo apresentado tivemos de lidar apenas com uma variável, de modo que as integrais a serem calculadas eram univariadas, o que permitiu o uso de quadratura Gaussiana com um custo computacional desprezível. Ao lidar com um conjunto de variáveis as integrais com as devemos trabalhar passam a ser multivariadas, o que torna o uso de métodos de integração determinísticos inviável (para um conjunto grande de parâmetros no modelo).

Felizmente, há uma solução para o problema mencionado acima. Primeiro, para apresentar essa proposta, suponha que há apenas uma observação, de modo que nossa posteriori é propocional à $f(y_1|\theta)\pi_0(\theta)$. Veja que, no nosso modelo, θ depende de y_1 apenas através do preditor linear η_1, que por sua vez é univariado. Apartir da priori Normal de θ, temos uma priori Normal para η_1 e podemos obter a posteriori aproximada para η_1 usando a metodologia descrita anteriormente (como η_1 é sempre univariado, temos que as integrais podem ser resolvidas facilmente com métodos numéricos determinísticos). Uma vez obtida a posteriori Normal para η_1,

é fácil obter a posteriori para θ (mesma fórmula usada em Modelos Dinâmicos Lineares, ver Petris, Petrone, and Campagnoli (2009), West and Harrison (1997) ou Kalman (1960)):

$$\begin{aligned} \eta_1|y_1 \sim \mathcal{N}(\mu, \sigma^2) \quad \theta \sim \mathcal{N}(\vec{m}_0, V_0) \to \theta|y_1 \sim \mathcal{N}(\vec{m}_1, V_1), \\ \vec{m}_1 = \vec{m}_0 + V_0 x_1 (x_1' V_0 x_1)^{-1} (\mu - x_1' \vec{m}_0), \\ \vec{V}_1 = \vec{V}_0 + V_0 x_1 (x_1' V_0 x_1)^{-1} (\sigma^2 - x_1' V_0 x_1)(x_1' V_0 x_1)^{-1} x_1' V_0. \end{aligned} \tag{14}$$

4 Aplicações

4.1 Caso Normal com variância desconhecida

4.2 Caso Rayleight

4.3 Caso Laplace Assimétrica

5 Conclussões

Referências

Amari, Shun-ichi. 2016. *Information Geometry and Its Applications*. 1st ed. Springer Publishing Company, Incorporated.

Dobson, A. J., and A. G. Barnett. 2018. *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press. https://books.google.com.br/books?id=YOFstgEACAAJ.

Kalman, Rudolph Emil. 1960. "A New Approach to Linear Filtering and Prediction Problems." *Transactions of the ASME–Journal of Basic Engineering* 82 (Series D): 35–45.

MacKay, David J. C. 2002. *Information Theory, Inference and Learning Algorithms*. USA: Cambridge University Press.

Marotta, Raíra, Mariane Branco Alves, and Helio S. Migon. 2022. "K-Parametric Dynamic Generalized Linear Models: A Sequential Approach via Information Geometry." arXiv. https://doi.org/10.48550/ARXIV.2201.05387.

McCulloch, Charles E, and Shayle R. Searle. 2001. *Generalized, Linear and Mixed Models*. Wiley, New York.

Migon, H. S., D. Gamerman, and F. Louzada. 2014. *Statistical Inference: An Integrated Approach, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press. https://books.google.com.br/books?id=2VfNBQAAQBAJ.

Petris, Giovanni, Sonia Petrone, and Patrizia Campagnoli. 2009. *Dynamic Linear Models with r*. useR! Springer-Verlag, New York.

Tierney, Luke, and Joseph B. Kadane. 1986. "Accurate Approximations for Posterior Moments and Marginal Densities." *Journal of the American Statistical Association* 81 (393): 82–86. https://doi.org/10.1080/01621459.1986.10478240.

West, Mike, and Jeff Harrison. 1997. *Bayesian Forecasting and Dynamic Models (Springer Series in Statistics)*. Hardcover; Springer-Verlag.