UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Departamento de Métodos Estatísticos - IM

Estatística computacional

Modelos Lineares Generalizados: Uma abordagem sequencial via Geometria da Informação

Projeto final

Aluno: Silvaneo Vieira dos Santos Junior

Professor: Carlos Tadeu Pagani Zanini

12 de janeiro de 2023

ÍNDICE

1 Introdução

Este trabalho tem como proposta aplicar a metodologia desenvolvida em Marotta, Alves, and Migon (2022) para o ajuste de Modelos Dinâmicos Lineares Generalizados (GDLM) via Geometria da Informação (Amari 2016) no caso particular dos"modelos não-dinâmicos", isto é, dos Modelos Lineares Generalizados (GLM). A ideia deste projeto é apresentar uma versão modificado do trabalho original, dando ênfase nas particularidades associadas aos GLMs.

Começaremos o trabalho apresentando a metodologia usada, descrevendo de forma geral a classe dos GLMs e apresentando os resultados relevantes associados à Geometria da Informação.

Em seguida, discutiremos a qualidade do ajuste obtido pela metodologia proposta, apresentando também uma alternativa com menor custo computacional.

Por último, finalizaremos o trabalho apresentando um exemplo com dados simulados, discutindo a qualidade do ajuste como um todo e as consequências de certas simplificações.

Os códigos utilizados para os ajustes (incluindo os que foram omitidos neste documento) e criação deste documento podem ser encontrado neste repositório.

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.

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:

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

onde g, chamada função de ligação, é uma função contínua e monótona, η_i é o preditor linear, X é a matriz de planejamento, x_i é a i-ésima coluna de X, β (possivelmente um vetor) são os parâmetros latentes que representam o efeito das variáveis explicativas e F representa uma distribuição pertencente à 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 à família exponencial, então temos que, por definição, podemos escrever a densidade de $Y_i|\beta$ como:

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

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 2.1. *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 no caso Normal com média e variância desconhecidas.

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 β e, para realizar qualquer que seja a análise, devemos encontrar a distribuição a posteriori de β (π_n). Como pode ser visto em Migon, Gamerman, and Louzada (2014), podemos escrever π_n como:

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

Infelizmente, a menos do caso onde F representa a distribuição Normal com variância conhecida, não é possível obter uma solução analítica para $\int \pi_0(\beta)\prod_{i=1}^n f(y_i|\beta)d\beta$, sendo necessário recorrer a métodos de integração numérica para obter $\pi_n(\beta|y_1,...,y_n)$ ou estimativas de β.

Tendo em mente que métodos de integração numérica podem ter um custo computacional muito elevado, especialmente quando a dimensão de β 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 β. No trabalho original, a metodologia é apresentada para Modelos Dinâmicos Lineares Generalizados, porém os GLM são um caso particular desta classe de modelos, de modo que a metodologia proposta também é válida para os modelos em que estamos interessados. Dito isso, usaremos na verdade 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 anteriormente, para realizar o processo de inferência, precisamos obter $\pi_n(\beta|y_1,...,y_n) = \pi_0(\beta)\prod_{i=1}^n f(y_i|\beta)\left(\int \pi_0(\beta)\prod_{i=1}^n f(y_i|\beta)d\beta\right)^{-1}$, 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 $\int \pi_0(\beta)\prod_{i=1}^n f(y_i|\beta)d\beta$ 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 e que explica o porque a escolha feita é "natural"). 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 2.2 (da Projeção)**.** *Sejam 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:*

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

[1] O Teorema ainda vale sem essa restrição.

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 encontrado em Amari (2016).

Demonstração. Para provar o teorema 2.2, primeiro observe que a divergência KL de p com relação a q pode ser escrita como:

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

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 2.1 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 2.2. □ □

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 inconsistência na metodologia: Não podemos calcular π_n ,pois não temos forma analítica fechada para $\int \pi_0(\beta) \prod_{i=1}^n f(y_i|\beta) d\beta$, 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 $\int \pi_0(\beta) \prod_{i=1}^n f(y_i|\beta) d\beta$, 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 3 métodos para calcular os valores esperados desejados: Quadratura Gaussiana, Monte Carlo e pelo método proposto em Tierney and Kadane (1986) e refinado em Tierney, Kass, and Kadane (1989). Usaremos Quadratura Gaussiana nos Casos Normal com variância desconhecida e Rayleigh [2], pois nestes casos temos de lidar com integrais univariadas, de modo que o custo computacional de se usar Quadratura Gaussiana é desprezível. Para o caso Laplace Assimétrica [2], usaremos os outros métodos afim de conseguir um ajuste satisfatório.

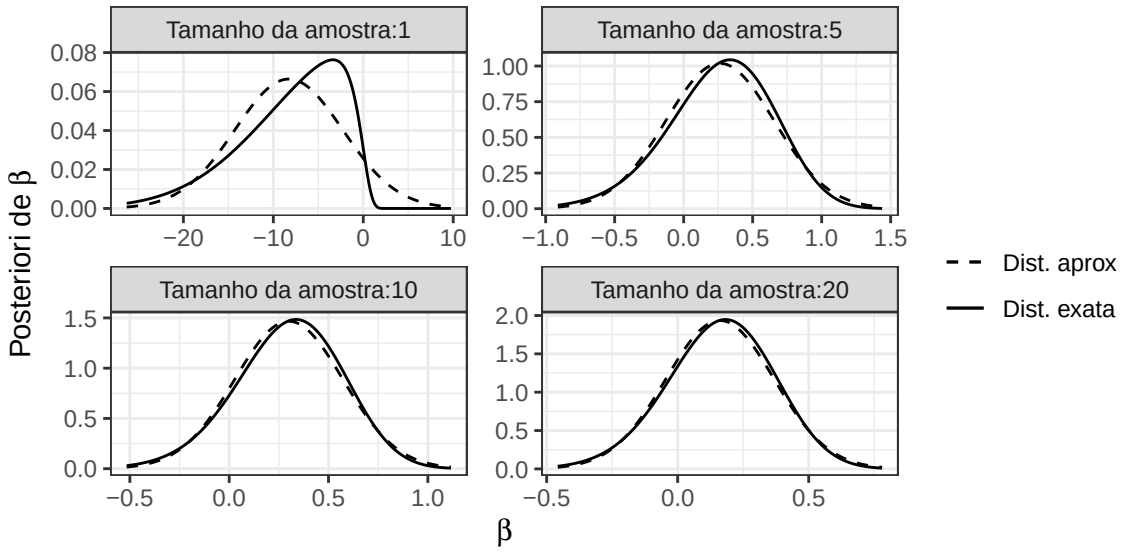
3 Qualidade da aproximação

Como visto na sessão anterior, o teorema 2.2 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:

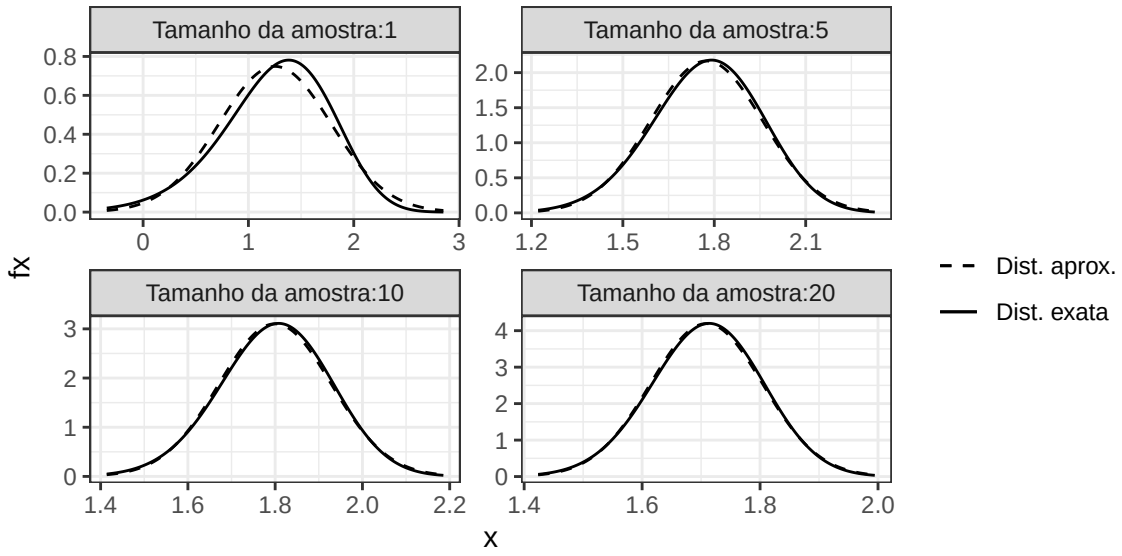
$$\begin{aligned} Y_i|\beta &\sim Poisson(\lambda), \\ \ln(\lambda) &= \beta, \\ \beta &\sim \mathcal{N}(0, 100), \end{aligned} \tag{13}$$

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



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 dependente da magnitude dos dados observados. De modo geral, a qualidade vai depender da quantidade de informação na amostra, sendo que é fácil ver que (especificamente no caso Poisson) observações de valores maiores tem muito mais informação do que observações de valores menores (especialmente o 0). O caso que mostramos anteriormente seria um "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):

[2]Esse caso foi omitido do trabalho final para não ultrapassar o limite de páginas.



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 modelo onde β é um vetor, as integrais com as quais devemos trabalhar passam a ser multivariadas, o que torna o uso de métodos de integração determinísticos inviável (para um modelo com muitos parâmetros).

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 $f(y_1|\beta)\pi_0(\beta)$. Veja que, no nosso modelo, β depende de y_1 apenas através do preditor linear η_1, que por sua vez é univariado. A partir da priori Normal de β, 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 β (mesma fórmula usada em Modelos Dinâmicos Lineares, ver Petris, Petrone, and Campagnoli (2009), West and Harrison (1997) ou Kalman (1960)):

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

Com as equações acima podemos atualizar a distribuição de β de forma computacionalmente eficiente, independente da dimensão de β. Infelizmente, isso só resolve o nosso problema para amostras de tamanho unitário, de fato, para uma amostra de tamanho n qualquer, devemos obter a posteriori conjunta dos n preditores lineares para podermos usar a fórmula (14), para isso deveríamos resolver algumas integrais n variadas, ou seja, o problema inicial ainda persiste.

Resta então apresentar um último resultado teórico que finalmente vai nos permitir fugir da "maldição da dimensionalidade". Veja que, em geral, vale que:

$$f(\beta|y_1, ..., y_n) \propto f(y_n|\beta, y_1, ..., y_{n-1}) f(y_{n-1}|\beta, y_1, ..., y_{n-2}) \cdots f(y_2|\beta, y_1) f(y_1|\beta) f(\beta). \tag{15}$$

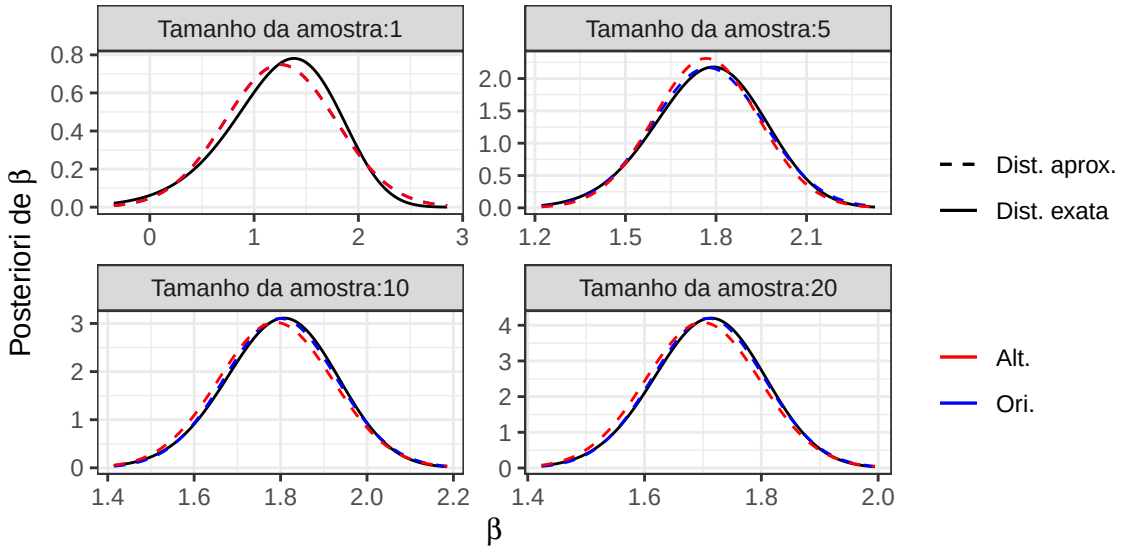
Usando que, por hipótese, os y_i's são independentes dado β, podemos simplificar a equação acima da seguinte forma:

$$\begin{aligned} f(\beta|y_1,...,y_n) &\propto f(y_n|\beta)f(y_{n-1}|\beta)\cdots f(y_2|\beta)f(y_1|\beta)f(\beta), \\ &= f(\beta)\prod_{i=1}^{n} f(y_i|\beta). \end{aligned} \tag{16}$$

Observe que $f(y_1|\beta)f(\beta) \propto f(\beta|y_1)$, de modo que podemos obter uma forma aproximada para $f(y_1|\beta)f(\beta)$ usando a metodologia proposta. Com isto, podemos escrever $f(y_2|\beta)f(y_1|\beta)f(\beta) \propto f(y_2|\beta, y_1)f(\beta|y_1)$ e, interpretando $f(\beta|y_1)$ como uma nova priori para β, podemos usar a metodologia proposta para obter uma forma aproximada para $f(\beta|y_1, y_2)$. Repetindo esse processo sequencialmente na amostra podemos obter uma forma aproximada para $f(\beta|y_1, ..., y_n)$ usando apenas integrais univariadas. Para exemplificar que esta abordagem tem um custo computacional menor do que a original, digamos que, ao usar Quadratura Gaussiana para resolver uma integral, devemos avaliar a função $f(y|\beta)$ k vezes para cada dimensão, de modo que, se fôssemos calcular as integrais diretamente, deveríamos avaliar a função k^n (teríamos uma malha n dimensional e repartiríamos cada dimensão em k segmentos) no caso da integração pelos preditores lineares, ou k^r no caso da integração diretamente nos parâmetros latentes β. Em contrapartida, usando a aproximação sequencial da amostra, devemos avaliar a função $f(y_1|\beta)$ apenas $k \times n$ vezes!

Claramente essa proposta reduz drasticamente a quantidade de vezes que devemos computar a função $f(y|\beta)$, porém há um preço a se pagar por essa simplificação: Ao usar a distribuição aproximada diversas vezes, há um acumulo do erro de aproximação. Sendo assim, devemos avaliar se, usando esse método simplificado, não estamos tendo perdas significativas na qualidade da aproximação.

Adiante apresentamos uma comparação entre as posterioris aproximadas pelo método original e alternativo para os dados do exemplo Poisson com taxa 5:



Veja que o método alternativo introduz, de fato, um erro adicional significativo, mas que é, contudo, tolerável, uma vez que a redução do custo computacional é muito grande e, como veremos na próxima sessão, em termos práticos, a diferença entre a posteriori exata e a posteriori aproximada pelo método alternativo é desprezível (pelo menos em alguns casos).

4 Aplicações

Nesta sessão apresentaremos uma das aplicação do método proposto em dados simulados. No exemplo escolhido faremos comparações entre a posteriori obtida pela abordagem descrita na sessão anterior (doravante chamado de método KL) e a posteriori obtida usando métodos de Monte Carlo via Cadeias de Markov (MCMC). Para o obtenção da posteriori por MCMC usaremos o algorítmos de Gibbs (Gamerman and Lopes 2006), sendo que para amostrar das marginais completas de cada parâmetro usaremos Metropolis-Hastings com propostas independentes (Gamerman and Lopes 2006), especificamente, podemos usar a abordagem KL para obter aproximações para as marginais completas e então usar a distribuição aproximada como proposta, desta forma a cadeia gerada deve convergir rapidamente e ser pouco auto correlacionada (supondo, claro, que a aproximação seja boa). Uma vantagem dessa abordagem para a criação da cadeia é que a aproximação KL só precisa ser feita uma vez (podemos aproveitar o ajuste que já foi feito com a metodologia KL), uma vez que, se $\vec{\beta} \sim \mathcal{N}(\vec{\mu}, \Sigma)$, então:

$$\vec{\beta}_1 | \vec{\beta_2} \sim \mathcal{N}(\vec{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\vec{\beta_2} - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}), \tag{17}$$

onde $\vec{\beta}_1$ e $\vec{\beta_2}$ formam uma partição de $\vec{\beta}$ tal que:

$$\begin{bmatrix} \vec{\beta}_1 \\ \vec{\beta}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \vec{\mu}_1 \\ \vec{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right).$$

Inicialmente, o intuito era apresentar 3 exemplos nesta sessão, porém, devido à restrição sob o tamanho do documento, as demais aplicações (com dados com distribuição Rayleigh e Laplace Assimétrica) serão omitidas, uma vez que o resultado obtido no caso Normal com variância desconhecida representa bem a qualidade do ajuste nos demais casos.

4.1 Caso Normal com variância desconhecida

A primeira etapa para esse exemplo é a criação de um conjunto de dados simulados. Ao longo desta sessão testaremos ajustes com diversos tamanhos de amostra[3], em todos os casos o processo de ajuste e criação dos dados será o mesmo.

Primeiro, vamos criar um conjunto de variáveis auxiliares X_{1i}, X_{2i}, X_{4i}[4] tais que:

$$\begin{aligned} X_1 &\sim \mathcal{G}(5, 2.5), \\ X_2 &\sim \mathcal{N}(3, 4), \\ X_4 &\sim Poisson(20). \end{aligned} \tag{18}$$

Vamos então gerar uma amostra Y_i tal que:

$$\begin{aligned} Y_i &\sim \mathcal{N}(\mu_i, \tau_i^{-1}), \\ \mu_i &= \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2, \\ \ln(\tau_i) &= \beta_3 + X_{4i}\beta_4. \end{aligned} \tag{19}$$

Sendo que, neste exemplo, usamos $\vec{\beta} = (1, 0.5, 0.25, 1, -0.04)'$ como o valor verdadeiro e tomamos como priori para $\vec{\beta}$ a distribuição Normal Multivariada com vetor de médias nulo e matriz de covariância igual a identidade.

Um detalhe importante a se notar neste exemplo é que estamos trabalhando com uma distribuição que possui dois parâmetros (μ e τ), desta forma, para cada elemento da amostra, temos dois preditores lineares, ou seja,

[3] Os ajustes foram feitos e analisado, poréo não exibiremos todos os gráficos.

[4] Não criamos uma variável X_{3i} para que a notação fique mais intuitiva , uma vez que β_3 será o intercepto associado à precisão.

mesmo com a abordagem alternativa, ainda precisaríamos resolver integrais duplas para obter a aproximação KL. Porém, de forma muito conveniente, é possível obter todos os valores esperados desejados a partir de integrais univariadas, para isto vamos usar que:

$$\mathbb{E}[f(\vec{\eta})|y] = \mathbb{E}[\mathbb{E}[f(\vec{\eta})|\eta_2, y]|y],$$

daí, se $\mathbb{E}[f(\vec{\eta})|\eta_2, y]$ tem forma analítica fechada e conhecemos algo proporcional à distribuição marginal de $\eta_2|y$, então $\mathbb{E}[f(\vec{\eta})|y]$ pode ser escrita como uma integral que depende apenas de η_2.

De fato, temos que, se τ fosse conhecido, a distribuição a priori de η_1 conjugaria com a distribuição de y, de modo que, se $\eta_1|\eta_2 \sim \mathcal{N}(\mu_1, \tau_1^{-1})$[5], então:

$$\begin{aligned} \eta_1|\tau, y &\sim \mathcal{N}(\mu_1^*, \tau_1^{*-1}), \\ \mu_1^* &= \frac{\tau_1\mu_1 + \tau y}{\tau + \tau_1}, \\ \tau_1^* &= \tau + \tau_1. \end{aligned} \tag{20}$$

Daí:

$$\begin{aligned} \mathbb{E}[\eta_1|\eta_2, y] &= \mu_1^*, \\ \mathbb{E}[\eta_1^2|\eta_2, y] &= \tau_1^{*-1} + \mu_1^{*2}, \\ \mathbb{E}[\eta_1\eta_2|\eta_2, y] &= \mu_1^*\eta_2. \end{aligned} \tag{21}$$

Com o resultado acima, se encontrarmos algo proporcional à distribuição marginal de $\eta_2|y$, é possível obter a média e a matriz de covariância de $\vec{\eta}$ usando Quadratura Gaussiana univariada.

Para obter algo proporcional à distribuição marginal de $\eta_2|y$, basta observar que, se $Y|\eta_1, \eta_2 \sim \mathcal{N}(\eta_1, \exp\{-\eta_2\})$ e $\eta_1|\eta_2 \sim \mathcal{N}(\mu_1, \tau_1^{-1})$, então:

$$Y|\eta_2 \sim \mathcal{N}\left(\mu_1, \exp\{-\eta_2\} + \tau_1^{-1}\right)$$

Ademais, vale que:

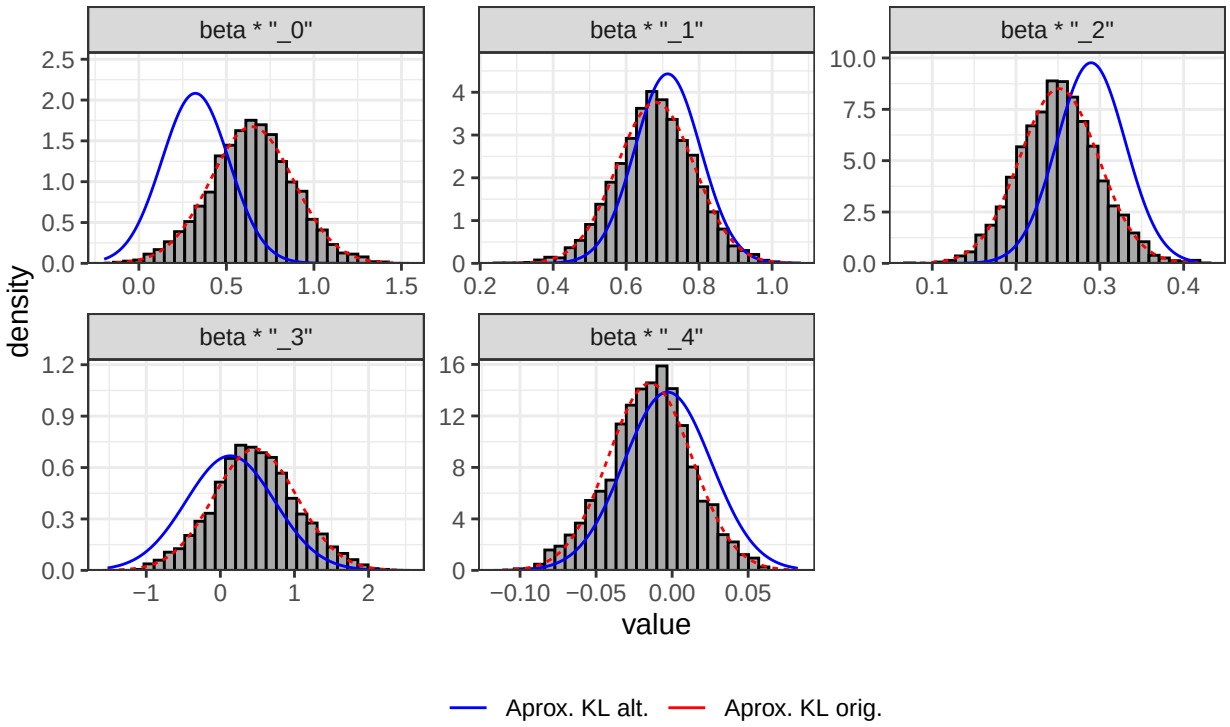
$$f(\eta_2|y) \propto f(y)f(\eta_2).$$

Com os resultados descritos acima foi possível ajustar os modelos desejados com um custo computacional negligenciável.

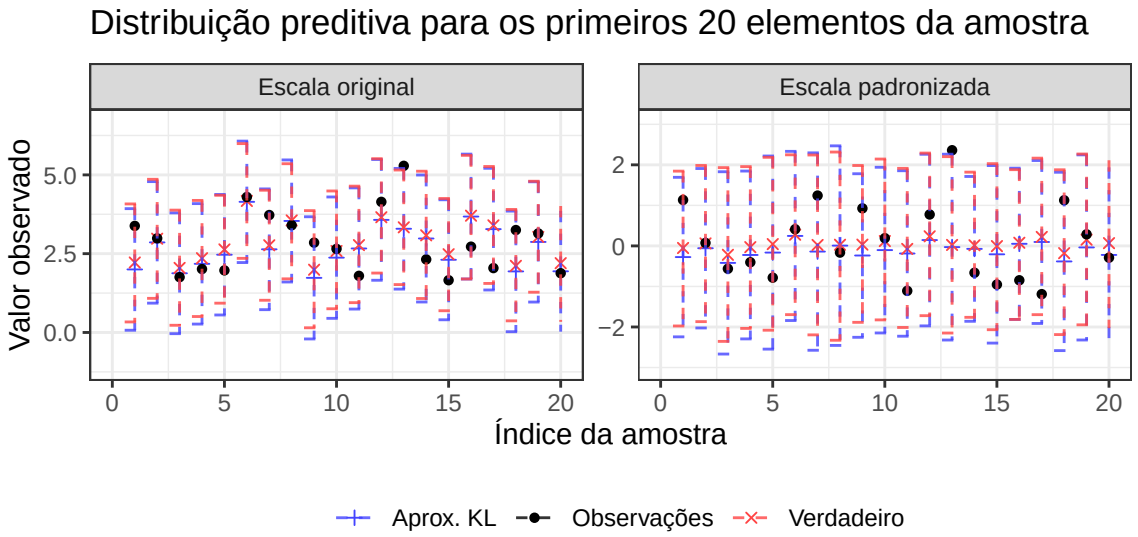
Adiante vamos exibir algumas comparações entre a distribuição posteriori de $\vec{\beta}$ aproximada pelo método KL (original e alternativo)[6] e o histograma de uma amostra da distribuição posteriori verdadeira de $\vec{\beta}$:

[5] Como a distribuição conjunta de η_1, η_2 é Normal bivariada, é possível obter μ_1 e τ_1^{-1} facilmente, basta observar que se X, Y tem distribuição Normal Multivariada, de modo que X tem média μ_X e variância σ_X^2, Y tem média μ_Y e variância σ_Y^2 e a correlação entre X e Y é igual à ρ, então $X|Y \sim \mathcal{N}\left(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y), \left(1 - \rho^2\right)\sigma_X^2\right)$

[6] Vale lembrar que a abordagem original é aproximar a posteriori completa de $\vec{\beta}$ por uma Normal com mesmo vetor de médias e mesma matriz de covariância (utilizamos uma amostra obtida por MCMC da posteriori de β para encontrar o vetor de médias e a matriz de convariância "exatas"). Em contrapartida, a abordagem alternativa faz a mesma aproximação, porém de forma sequencial, elemento a elemento da amostra. A abordagem original é substancialmente mais cara, pois exige que calculemos integrais multivariadas.



Podemos observar pelo gráfico acima que a abordagem original produz uma aproximação que é essencialmente igual à distribuição verdadeira (representa bem ela), em contrapartida, a abordagem alternativa produz uma distribuição que é significativamente diferente da distribuição correta, especialmente para o parâmetro β_0, que representa o intercepto da média. Apesar da discrepância significativa entre a distribuição aproximada pelo método alternativo e a distribuição correta, podemos observar no gráfico a seguir que a distribuição preditiva para as observações não é tão diferente:



Veja que as estimativas pontuais e intervalares (intervalos simétricos centrados na média com 95% de credibilidade) para as observações não são muito discrepantes, de modo geral. Esse comportamento acontece devido a uma certa "compensação" do viés da aproximação KL alternativa: Apesar do intercepto estar subestimado, o efeito das covariáveis está superestimado. Ademais, apesar de visualmente a posteriori

aproximada pelo método KL alternativo estar muito discrepante, a magnitude do viés não é de fato tão significativa, de modo que o impacto na distribuição preditiva é pequeno, mesmo se não houvesse compensação no viés.

É relevante observar também a forma como o tamanho da amostra afeta a qualidade das aproximações. Infelizmente, teremos de omitir os gráficos associados a essa análise para não ultrapassar o limite estipulado de páginas, contudo, de modo geral, observamos que a qualidade da aproximação aumenta conforme o tamanho da amostra aumente e, de forma análoga, a aproximação fica pior conforme o tamanho da amostra diminui. Dito isso, para amostras de tamanho superior a 20 a qualidade do ajuste ficou aceitável, sendo que, como veremos adiante, há uma redução massiva no custo computacional ao se utilizar esta abordagem.

A última comparação que precisamos fazer antes de finalizar esta aplicação é com respeito ao tempo computacional, especificamente, gostaríamos de comparar o tempo de execução do método KL alternativo com o tempo de amostragem da abordagem por MCMC. Naturalmente, a comparação entre o tempo de execução dos métodos não é direta, afinal, um dos métodos exige amostragem, sendo que seu tempo de execução pode ser reduzido ao simplesmente tomar uma amostra menor (diminuindo a precisão das estimativas). Além disso, a comparação entre a abordagem por MCMC não é muito justa, uma vez que estamos usando a aproximação KL como proposta, o que faz com que a cadeia gerada convirja muito mais rápido do que o usual (como referência, a taxa média de aceitação do passo de Metrolis foi de cerca de 96%, isto é, a proposta utilizada é bem próxima da distribuição verdadeira, o que agiliza bastante a convergência do algoritmo) e tenha uma auto correlação também muito baixa.

Com todas essas observações em mente, concluímos que a forma mais "justa" de comparar as duas abordagens seria através do tempo necessário para gerar amostras de tamanhos variados da posteriori usando cada um dos métodos. Na tabela a seguir vamos exibir os tempos registrados:

	Tamanho da amostra					
	500		5,000		50,000	
Nº obs.	MCMC	KL	MCMC	KL	MCMC	KL
20	0.83	0.48	5.80	0.47	53.09	0.50
100	0.81	0.52	5.90	0.52	53.52	0.55
500	0.95	0.81	7.01	0.82	65.93	0.82
1,000	1.16	1.16	8.78	1.20	80.91	1.21

Tabela 1: Comparação entre o tempo de execução (em segundos) de cada um dos métodos. Vale destacar que este tempo inclui o tempo de ajuste do modelo e o tempo para amostrar da posteriori (no caso do método MCMC os dois processos são inseparáveis), ademais, para a amostra obtida pelo método MCMC, não foi feito *burn-in* e nem nenhum tipo de checagem quanto a convergência da cadeia.

A tabela acima mostra claramente que a abordagem KL alternativa é muito superior em termos de custo computacional, sendo o ajuste e amostragem praticamente instantâneos, ademais, essa abordagem também escala muito melhor com o tamanho da amostra desejada, quase não havendo aumento no tempo de execução para se gerar amostras maiores.

5 Conclusões

Levando em conta todas as análises feitas para este projeto (incluindo algumas que não entraram neste documento, para manter a brevidade), podemos concluir que a abordagem proposta oferece um custo computacional baixíssimo em troca de uma pequena (mas significativa) perda na precisão, sendo que a perda de precisão vem quase completamente da parte sequencial da metodologia (i.e., a aproximação da posteriori por uma gaussiana via Teorema da Projeção introduz pouco erro nas estimativas). Tendo isso em mente, uma progressão natural para este trabalho seria buscar formas de mitigar o erro de aproximação devido à parte sequencial (talvez fazer a atualização da posteriori por subconjuntos não unitários da amostra, por exemplo). Caso seja possível obter um ajuste que mantenha o custo computacional, mas apresente uma precisão maior na aproximação da posteriori (o que é possível, pois vimos que a aproximação pela abordagem KL original é

muito boa), então teríamos um método muito geral (toda a metodologia descrita vale para qualquer dado cuja distribuição pertença à família exponencial) e bem acessível.

De modo geral, acredito que podemos considerar que o projeto foi bem sucedido, uma vez que os resultados obtidos tanto para o caso apresentado como para os casos omitidos (Rayleigh e Laplace Assimétrico) foram satisfatórios. Ademais, acredito que o trabalho tenha permitido o uso de diversos conceitos e métodos vistos ao longo da disciplina de Estatística Computacional, de modo que o desenvolvimento deste projeto esteve bem alinhado com o conteúdo trabalhado em aula.

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.

Gamerman, D., and H. F. Lopes. 2006. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. https://books.google.com.br/books?id=yPvECi/_L3bwC.

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.

Tierney, Luke, Robert E. Kass, and Joseph B. Kadane. 1989. "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions." *Journal of the American Statistical Association* 84 (407): 710–16. https://doi.org/10.1080/01621459.1989.10478824.

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