# Inferência variacional bayesiana como alternativa ao MCMC

**Larissa de Carvalho Alves**
Escola Nacional de Ciências Estatísticas - ENCE

V Seminário Internacional de Estatística com R

Junho 2021

## Sumário

- O que é a inferência variacional?
- MCMC x VB
- Formulação do VB
- Aplicações
  - Exercício 1: MCMC x VB
  - Exercício 2: Covid-19
- Conclusão

# O que é a inferência variacional?

- A inferência variacional (VI) ou variacional bayesiano (VB) é um método de aprendizado de máquinas que usa otimização para **aproximar** densidades.
- Um dos algoritmos mais usados para resolver o problema de otimização foi introduzido por Bishop (2006) e denominado por CAVI - inferência variacional de ascensão por coordenadas.
- Esse método é especialmente útil na inferência Bayesiana pois ela tem como principal meta obter a distribuição a posteriori que na maioria dos problemas é desconhecida.
- Blei et al. (2018) revisa as ideias por trás da inferência variacional e apresenta exemplos sob o ponto de vista bayesiano.

# MCMC x VB

- Método tradicional: MCMC
  - Computacionalmente intensivo;
  - Gera uma amostra da distribuição a posteriori (exato);
  - Bem difundido e com muitos pacotes no software R.
- Método alternativo: VB
  - Aproximado;
  - Mais rápido que o MCMC;
  - Contas semelhantes ao Amostrador de Gibbs (MCMC);
  - Poucos pacotes no software R.

# Formulação VB (Blei et al. [2018])

- Ideia: escolher uma família de densidades e encontrar o membro desta família que esteja próximo da distribuição alvo.
- A proximidade é medida pela divergência de Kullback-Leibler.
- Seja $q(\theta)$ a família de densidades variacionais, então tem-se como objetivo obter $q^*(\theta)$ tal que

$$q^*(\theta) = argmin_{q(\theta)} KL(q(\theta)||p(\theta|y))$$

onde
$KL(q(\theta)||p(\theta|y)) = E[\log q(\theta)] - E[\log p(y,\theta)] + \log p(y)$ e $p(y)$ é a evidência do modelo.

# Formulação VB (Blei et al. [2018])

- Não é possível calcular KL, então otimiza-se a função ELBO.

$$ELBO(q) = E[\log p(y, \theta)] - E[\log q(\theta)].$$

- **Maximizar o ELBO** é equivalente a minimizar a divergência de KL.
- O ELBO pode ser usado como um **critério de seleção de modelos** devido a sua relação com $p(y)$.
- Para maximizar o ELBO vamos particionar o vetor $\theta$ e usar como densidades variacionais $q_l(\theta_l)$ tal que

$$q(\theta) = \prod_{l=1}^{m} q_l(\theta_l).$$

- O valor ótimo para cada $q_l(\theta_l)$ é obtido por

$$q_l^*(\theta_l) \propto \exp\{E_{-l}[\log p(\theta_l|\theta_{-l}, y)]\}.$$

# Algoritmo VB

(1) Inicializar: $q_l(\theta_l)$

(2) Calcular $q_l(\theta_l) \propto \exp\{E_{-l}[\log p(\theta_l|\theta_{-l}, y)]\}$ para $l = 1, \ldots, m$

(3) Calcular ELBO(q)

(4) Repetir passos (2) e (3) até convergir.

# Exercício 1: MCMC x VB

- Objetivo: Comparar os métodos MCMC e VB por meio de dados simulados de um modelo de regressão linear, com relação ao tempo computacional e a estimação dos parâmetros.

Considere o seguinte modelo (Drugowitsch, 2019):

$$
\begin{aligned}
y|X,\beta,\phi &\sim N(X\beta, \phi^{-1}I_n) \\
\beta|\phi,\tau &\sim N(0, (\phi D_\tau)^{-1}) \\
\tau_j &\sim Ga(c_0, d_0), \quad j = 1, \ldots, p \\
\phi &\sim Ga(a_0, b_0)
\end{aligned}
$$

onde $y$ tem dimensão $n$, $\beta = (\beta_1, \ldots, \beta_p)^T$, $X$ é a matriz de regressores de dimensão $n \times p$ e $D_\tau = diag(\tau_1, \ldots, \tau_p)$.

# Exercício 1: MCMC x VB

- Distribuição a priori: $\phi \sim Ga(0.1, 0.1)$, $\tau_j \sim Ga(1, 0.1)$.
- Simulação dos dados: $n = 10000$, $p = 10$, $\phi = 0.4$, $\tau_j = 1 \; \forall j$ e cada coluna de $X$ foi gerada de uma distribuição $N(0, I_n)$.
- Os coeficientes de regressão e as observações foram gerados a partir da estrutura hierárquica do modelo.
- Para o MCMC foi utilizado o pacote "R2jags"e o critério de Raftery e Lewis (1992) para verificar convergência.
- Foram geradas 5 mil iterações, descartadas mil como aquecimento da cadeia e considerou-se espaçamento de 4, resultando em uma amostra de tamanho 1000 da distribuição a posteriori.

# Exercício 1: MCMC x VB

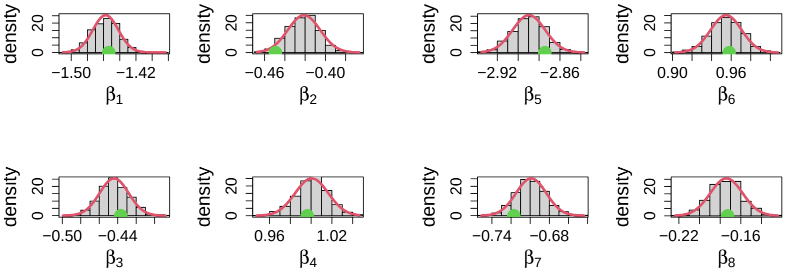


Figura: *Comparação entre MCMC (histograma) e VB (curva em vermelho).*

# Exercício 1: MCMC x VB

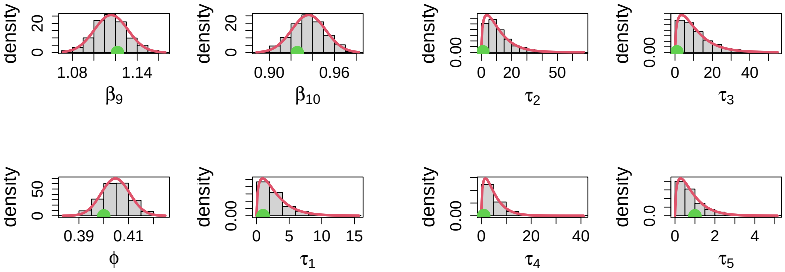


Figura: *Comparação entre MCMC (histograma) e VB (curva em vermelho).*

# Exercício 1: MCMC x VB

- VB apresenta resultados semelhantes ao MCMC.
- Principal diferença: tempo computacional.

Tabela: Tempo computacional

| | |
|---|---|
| MCMC | 39.58 seg |
| VB | 0.55 seg |

- O tempo do MCMC é aproximadamente 72 vezes maior nesta simulação.

# Exercício 2: Covid-19

- Objetivo: capturar a tendência do número de casos diários por Covid-19 no Brasil de 10 de março de 2020 até 18 de março de 2021.
- Fonte dos dados: CovidLP (UFMG)

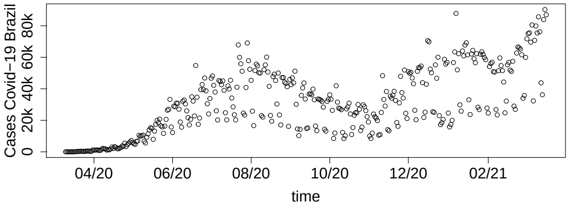


Figura: Casos diários de Covid-19 no Brasil.

- Ajustou-se aos dados (na escala logarítmica) um modelo de regressão spline com penalidade Lasso para selecionar o número de nós significativos.

# Exercício 2: Covid-19

- Considere o seguinte modelo de regressão não paramétrica

$$y_i = f(x_i) + \epsilon_i,$$

  onde
  - $(y_i, x_i)$ para $i = 1, \ldots, n$ é a coleção de observações,
  - $f(x_i) = E[y_i|x_i]$ são os valores obtidos por uma função suave $f$,
  - $\epsilon_i$ é uma sequência de variáveis aleatórias não correlacionadas com média $0$ e precisão desconhecida $\phi$.
- Uma possível abordagem para estimar $f$ é assumir que a curva pode ser bem aproximada por uma função spline.

# Exercício 2: Covid-19

- Dado uma sequência de nós $\kappa = (\kappa_1, \ldots, \kappa_K)$ tal que $\kappa_1 < \ldots < \kappa_K$, um modelo de regressão spline pode ser escrito como

$$f(x, \beta) = \beta_0 + \sum_{j=1}^{p} \beta_j x^j + \sum_{k=1}^{K} \beta_{p+k} (x - \kappa_k)_+^p,$$

  onde $p$ é o grau do polinômio e $\beta$ é o vetor de coeficientes de dimensão $K + p + 1$.

- $(u)_+^p = max(0, u^p)$ formam a base de funções de potências truncadas.

# Exercício 2: Covid-19

- Atribuímos uma distribuição a priori Laplace aos coeficientes da base de potências truncadas.

- O modelo de regressão spline para seleção de nós pode ser escrito como:

$$
\begin{aligned}
y|X,\beta,\phi &\sim N(X\beta,\phi^{-1}I_n) \\
\beta^{(1)} &\sim N(m_0, C_0) \\
\beta^{(2)}|\phi,\tau &\sim N(0,\phi^{-1}D_\tau) \\
\tau_j|\lambda &\sim Exp(\lambda), \quad j=1,\ldots,K \\
\phi &\sim Ga(a_0,b_0) \\
\lambda &\sim Ga(g_0,h_0)
\end{aligned}
$$

# Exercício 2: Covid-19

- Comparamos modelos com $p = 3$ e $K = 10, 20$ e $30$.
- O ELBO foi usado como critério para indicar o melhor modelo.

Tabela: ELBO - Covid-19 - Brasil.

| | K = 10 | K = 20 | K = 30 |
|---|---|---|---|
| Brasil | **-251,39** | -260,12 | -260,22 |

# Exercício 2: Covid-19

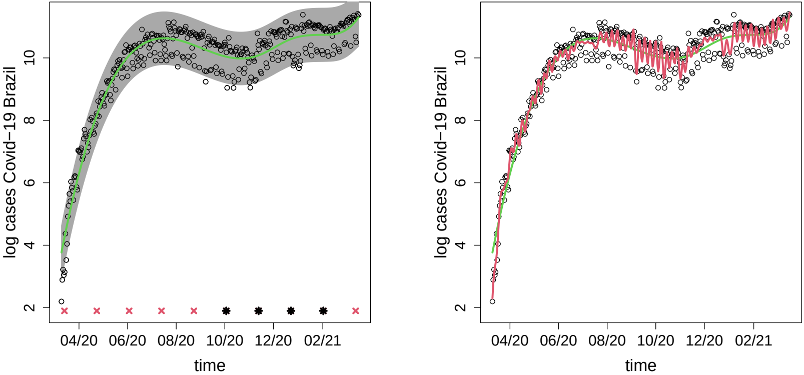


Figura: Esquerda: Ajuste do modelo de regressão spline penalizada ($p = 3$ e 4 nós significativos de um total de $K = 10$) aos dados do Brasil na escala log. Direita: Ajuste do modelo proposto (verde) e de um modelo usando a função "smooth.spline" do software R (vermelho).

# Conclusão

- Na simulação apresentada o VB tem desempenho semelhante ao MCMC e tempo computacional menor.
- Foram apresentados pontos negativos e positivos associados ao VB.
- Alguns pacotes do VB no software R: 'varbvs', 'VariationalBayes' e 'sparsevb'.

# Referências


C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer Science & Business Media, 2006

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: a review for statisticians. *J. Amer. Statist. Assoc.*, 112(518):859-877, 2018.

Center of Astrostatistics. Shapley galaxy dataset. Disponível em: https://astrostatistics.psu.edu/datasets/Shapley_galaxy.html. Acesso em: 12 de maio de 2021.

Departamento de Esttística UFMG. CovidLP. Disponível em: http://est.ufmg.br/covidlp/home/pt/. Acesso em: 20 de março de 2021.

J. Drugowitsch. Variational Bayesian inference for linear and logistic regression. *arXiv preprint arXiv:1310.5438*.

R Development Core Team. 2021:*R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, https://www.rproject.org/.

# Obrigada!

larissa.alves@ibge.gov.br

# VB aplicado a um modelo de regressão linear

Considere o seguinte modelo de regressão:

$$
\begin{aligned}
y|X,\beta,\phi &\sim N(X\beta, \phi^{-1} I_n) \\
\beta|\phi,\tau &\sim N(0, (\phi D_\tau)^{-1}) \\
\tau_j &\sim Ga(c_0, d_0) \;\; j = 1, \ldots, p \\
\phi &\sim Ga(a_0, b_0)
\end{aligned}
$$

onde $y$ é o vetor de variável resposta $n \times 1$, $\beta = (\beta_1, \ldots, \beta_p)^T$ são os parâmetros da regressão, $X$ é a matriz de regressores de dimensão $n \times p$ e $D_\tau = diag(\tau_1, \ldots, \tau_p)$ .

- Drugowitsch (2019) denomina essa representação hierárquica de determinação automática de relevância (ARD).

# VB aplicado a um modelo de regressão linear

- Seja $\theta$ o vetor de parâmetros e variáveis latentes, então, a distribuição a posteriori a seguir não é conhecida:

$$p(\theta|y) \propto p[y|X, \beta, \phi] \; p[\beta|\phi, \tau] \; p[\tau] \; p[\phi].$$

- Neste contexto as densidades variacionais consideradas foram:

$$\log(q(\theta)) = \log(q_1(\beta, \phi)) + \log(q_2(\tau)).$$

- Para cada parcela da soma obtivemos:

$$\log q_1(\beta, \phi) = \log N(\beta|m_\beta, \phi^{-1}C_\beta) \times Ga(\phi|a_\phi, b_\phi)$$

$$\log q_2(\tau) = \log \prod_{j=1}^{p} Ga(\tau_j|c_\tau, d_{\tau_j})$$

- As quantidades $m_\beta$, $C_\beta$, $b_\phi$, $d_{\tau_j}$ dependem de valores esperados que são atualizados a cada iteração do algoritmo.

# Formulação MCMC

- Podemos usar o Método de Monte Carlo via cadeias de Markov (MCMC), mais especificamente o Amostrador de Gibbs, para obter uma amostra da distribuição a posteriori, por meio das distribuições condicionais completas, $p(\theta_l|\theta_{-l}, y)$.

$$(\beta|\cdot) \sim N\left((X^TX + D_\tau)^{-1}X^Ty, \frac{1}{\phi}(X^TX + D_\tau)^{-1}\right)$$

$$(\phi|\cdot) \sim Ga\left(\frac{n}{2} + \frac{p}{2} + a_0, b_0 + \frac{1}{2}[(y - X\beta)^T(y - X\beta) + \beta^TD_\tau\beta]\right)$$

$$(\tau_j|\cdot) \sim Ga\left(\frac{1}{2} + c_0, d_0 + \frac{1}{2}\beta_j^2\phi\right)$$