Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

A Melhor Pior Priori: Minimizando a Variância de Amostragem por Importância para Troca de Prioris

Amanda Perez[1], Luiz Max Carvalho[2], Diego Mesquita[3]
FGV EMAp, Rio de Janeiro, RJ

O paradigma bayesiano de inferência estatística apresenta a grande conveniência de permitir a incorporação de conhecimento *a priori* ao processo de modelagem. Em especial, em muitos contextos, é de grande interesse estimar $\mathbb{E}_p[h(\theta)]$, onde h é uma função (mensurável) de teste e θ é o parâmetro de interesse do modelo, cuja distribuição *a posteriori* é $p(\theta \mid X) \propto \mathcal{L}(X \mid \theta)\pi(\theta)$. Entretanto, nem sempre é possível obter essa posteriori, seja por limitações de poder computacional ou mesmo por má convergência das cadeias de Markov ao realizar MCMC.

Nesses casos, uma alternativa é a troca de prioris: utiliza-se uma priori mais simples que a de interesse ("priori substituta") para a obtenção de uma distribuição *a posteriori*. Após isso, podemos utilizar amostragem por importância (IS) para amostrar da *a posteriori* de interesse. No entanto, existe a possibilidade de a variância do estimador ser muito alta ou mesmo infinita, a depender da escolha da priori substituta e da função de teste h [2]. No presente trabalho, utilizamos métodos de descida do gradiente para encontrar uma priori substituta que minimize essa variância.

É conhecido que, no caso geral, essa variância pode ser minimizada por uma distribuição conhecida com forma fechada [3]. Contudo, se o interesse é estimar $\mathbb{E}_p[h(\theta)]$, a variância é minimizada quando escolhe-se como priori substituta a distribuição cuja posteriori q seja dada por:

$$q(\theta \mid X) \propto |h(\theta)|\, p(\theta \mid X). \tag{1}$$

Essa escolha depende da posteriori de interesse $p(\theta|X)$, que não nos é acessível, impossibilitando usar diretamente o resultado em (1). Nossa proposta, então, é escolher uma priori $\pi_s(\theta)$ para substituir a priori de interesse $\pi(\theta)$ tal que sua posteriori $p_s(\theta|X)$ minimize a divergência KL entre p_s e q. Usando $D_{\mathrm{KL}}(p_s\|q)$ como função de perda, pode-se simplificar o problema de otimização a:

$$\pi_s^\star(\theta) = \underset{\pi_s \in \Pi_s}{\arg\min} \quad \log c_s + \mathbb{E}_{p_s}\left[\frac{\pi_s(\theta)}{|h(\theta)|\ \pi(\theta)}\right], \tag{2}$$

onde c_s é a constante que normaliza a distribuição p_s, isto é, se $\mathcal{L}(X \mid \theta)$ é a verossimilhança dos dados, então $p_s(\theta \mid X) = c_s\ \pi_s(\theta)\ \mathcal{L}(X \mid \theta)$. Caso nos limitemos ao uso de prioris substitutas conjugadas, teremos c_s e p_s conhecidas, o que nos permite avaliar (aproximadamente) a função de perda descrita em (2), usando Monte Carlo. Assim, fixando uma forma paramétrica para π_s, podemos utilizar algoritmos de descida do gradiente para obter os parâmetros ótimos da distribuição.

Além disso, foi possível generalizar o resultado para o caso em que há múltiplas funções de teste a serem consideradas (digamos $h_1, \cdots, h_m$). Para esse caso, buscou-se minimizar a média das variâncias do estimador de IS para cada uma das funções, obtendo como distribuição *a posteriori* substituta ótima: $g(\theta \mid X) \propto p(\theta \mid X)\sqrt{\sum_{i=1}^m h_i(\theta)^2}$. Esta expressão é muito semelhante à apresentada em (1), tendo como diferença apenas que a função h agora seria uma função de $h_1, \cdots, h_m$.

[1]perez.amanda.de.m@gmail.com
[2]luiz.fagundes@fgv.br
[3]diego.mesquita@fgv.br

Este detalhe auxilia na implementação para diferentes funções de teste simultaneamente, pois permite minimizar $D_{\mathrm{KL}}(p_s\|g)$ de maneira análoga, bastando tomar $|h(\theta)| = \sqrt{\sum_{i=1}^{m} h_i(\theta)^2}$.

Para ambos os casos, realizamos experimentos envolvendo um cenário simples e computacionalmente tratável. Assumimos uma sequência de observações $X_1, \cdots, X_n \sim \mathcal{N}(0, \sigma^2)$, com média conhecida e variância $\theta := \sigma^2$ desconhecida. Os dados utilizados foram gerados amostrando de uma normal de média zero e variância pequena, o que sugeriria modelar a priori de $\sigma = \sqrt{\theta}$ como uma distribuição com suporte em $\mathbb{R}^{>0}$ e alta densidade na vizinhança de zero, e.g., $\mathcal{N}(0, 1)$ truncada em zero. Mas, supondo a impossibilidade de usar essa distribuição para inferir σ, poderíamos tomar $\theta \sim \text{GamaInversa}(\alpha, \beta)$ como priori (conjugada) e aplicar IS para obter a posteriori de interesse.

Para essa implementação, foi utilizada a linguagem Python, com as bibliotecas PyTorch, PyStan e Numpy. A otimização dos parâmetros foi feita utilizando o algoritmo Adam [1].

Aplicando o método proposto para diferentes funções de teste, tanto individual quanto simultaneamente, obtivemos resultados promissores, que apontam que, ao utilizar a Gama Inversa com os parâmetros ótimos encontrados, o estimador de IS converge mais rapidamente e para resultados mais próximos dos obtidos por MCMC para a posteriori de interesse (vide a Figura 1). Esses resultados indicam o potencial do nosso método, e trabalhos futuros poderiam ampliar a variedade de experimentos e generalizar nossa abordagem para casos com prioris substitutas não-conjugadas.

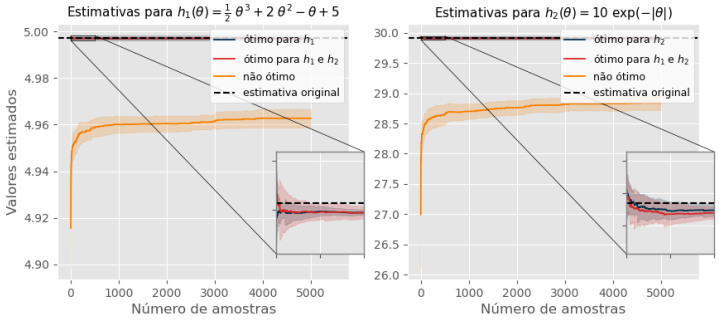

Figura 1: Estimativas obtidas por IS em 30 realizações do experimento para diferentes prioris substitutas e duas funções de teste ao variar o número de amostras de importância. Fonte: dos autores.

Agradecimentos

Este trabalho recebeu o apoio da Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro FAPERJ (SEI-260003/000709/2023), da Fundação de Amparo à Pesquisa do Estado de São Paulo FAPESP (2023/00815-6) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico CNPq (404336/2023-0).

Referências

[1] D. Kingma e J. Ba. "Adam: A Method for Stochastic Optimization". Em: **International Conference on Learning Representations (ICLR)**. San Diego, CA, USA, 2015.

[2] W. Neiswanger e E. Xing. "Post-Inference Prior Swapping". Em: **Proceedings of the 34th International Conference on Machine Learning**. Vol. 70. 2017, pp. 2594–2602. DOI: 10.48550/ARXIV.1606.00787.

[3] A. B. Owen. **Monte Carlo theory, methods and examples**. Acessado em 14/03/2024, https://artowen.su.domains/mc/Ch-var-is.pdf. 2013. Cap. 9.