Trabalho apresentado no XXXIX CNMAC, Uberlândia - MG, 2019.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Um estudo comparativo do uso de métodos de calibração bayesiana aproximada em modelos estocásticos

Heber L. Rocha [1]
Regina C. Almeida [2]
Renato S. Silva [3]
Laboratório Nacional de Computação Científica (LNCC), Petrópolis/RJ, Brasil
Ernesto A. B. F. Lima [4]
Institute for Computational Engineering and Sciences (ICES), Austin/TX, USA

Resumo. Neste trabalho realizamos um estudo comparativo do uso de três métodos de inferência Bayesiana em modelos matemáticos estocásticos. Consideramos que não dispomos de função de verossimilhança, de modo que a calibração é realizada tendo como base a abordagem bayesiana aproximada. Utilizamos dois métodos clássicos de computação bayesiana aproximada (ABC), o ABC Rejeição e o ABC Monte Carlo Sequencial, e um terceiro método que propõe o uso conjunto de um metamodelo com o ABC, requerendo um número limitado de simulações, denominado AABC. Os métodos foram utilizados para a resolução de dois problemas inversos simples, um deles tendo solução exata. Apesar de sensíveis à escolha dos parâmetros, todos os três métodos se mostraram adequados, com o AABC apresentando potencial destaque na redução do custo computacional.

Palavras-chave. Problema Inverso, Modelo Matemático Estocástico, Computação Bayesiana Aproximada.

1 Introdução

O principal objetivo de modelos matemáticos e computacionais é prever o valor de uma (ou mais) quantidade de interesse, que eventualmente será utilizada para uma determinada tomada de decisão. Neste contexto, a presença de incertezas em vários níveis na modelagem e simulação coloca em questão a confiabilidade da predição. É primordial que a calibração dos parâmetros do modelo leve em consideração tanto incertezas na modelagem quanto nos dados. Estes ingredientes são implicitamente considerados no processo de calibração via inferência bayesiana, foco deste trabalho. Especificamente, na inferência Bayesiana desejamos estimar os valores dos parâmetros $\theta \in \Theta$ associados ao modelo $\mathcal{M}$,

[1]heberlr@lncc.br
[2]rcca@lncc.br
[3]rssr@lncc.br
[4]lima@ices.utexas.edu

DOI: 10.5540/03.2020.007.01.0448

tal que as saídas/respostas (QoI) do modelo $y \in \mathcal{Y}$ melhor se ajustem aos valores experimentais $y_o \in \mathcal{Y}$. Como $\mathcal{M}$ é estocástico, dado um θ', sua execução repetida fornece uma gama de possíveis QoIs. Para denotar que y é gerado simulando o modelo para um dado θ', escrevemos $y \overset{\text{sim}}{\sim} \pi(y|\theta')$, em que $\pi(y|\theta')$ representa o simulador. Assumindo que o conhecimento inicial sobre os parâmetros é expresso pela distribuição *a priori* $\pi(\theta)$ e que a função de verossimilhança $p(y_o|\theta)$ é conhecida, a atualização do conhecimento sobre os parâmetros é feita através do teorema de Bayes $\pi(\theta|y_o) \propto p(y_o|\theta)\pi(\theta)$. A ideia básica deste teorema é que, após observar y_o, o conhecimento sobre θ aumenta, e é expresso pela distribuição de probabilidade *a posteriori* $\pi(\theta|y_o)$. A determinação analítica de $\pi(\theta|y_o)$ é restrita a casos de limitado interesse prático. A resolução numérica pode ser feita facilmente usando métodos de amostragem do tipo Monte Carlo. Além de serem computacionalmente intensivos, estes métodos requerem o conhecimento explícito da função de verossimilhança ($p(y_o|\theta)$), o que nem sempre é possível, seja por não se conhecer de fato a relação entre y_o e θ ou por ser intratável na prática por relacionar número grande de amostras e/ou parâmetros ou por complexidade funcional [4]. Nestas situações, a computação bayesiana aproximada (ABC - *Approximate Bayesian Computation*) tem se tornado uma alternativa atrativa por ser um método "livre de função de verossimilhança". Esta abordagem aplicada a modelos de natureza estocástica é o foco deste trabalho, conforme detalhado a seguir.

2 Computação Bayesiana Aproximada

O ABC é uma técnica de inferência Bayesiana sem a construção explícita da função de verossimilhança [1]. Esta abordagem tem se tornado bastante popular, e vem sendo aplicada em problemas inversos de diversas áreas do conhecimento [6]. A ideia básica do ABC consiste em considerar uma versão aumentada da distribuição *a posteriori*: ao invés de $\pi(\theta|y_o) \propto p(y_o|\theta)\pi(\theta)$, usamos simulações do modelo para obter a forma:

$$\pi(\theta, y|y_o) \propto p(y_o|y, \theta)\pi(y|\theta)\pi(\theta). \tag{1}$$

A função peso $p(y_o|y, \theta)$ é uma medida da similaridade entre o valor simulado y e o observado y_o. Quando $y = y_o$, $p(y_o|y_o, \theta)$ é uma constante positiva de modo que $\pi(\theta, y_o|y_o) \propto p(y_o|\theta)\pi(\theta)$. Uma aproximação para a distribuição de interesse é obtida calculando a marginal integrando em $\mathcal{Y}$, isto é,

$$\pi_{ABC}(\theta|y_o) \propto \pi(\theta) \int_{\mathcal{Y}} p(y_o|y, \theta)\pi(y|\theta)dx. \tag{2}$$

Assim, $\pi_{ABC}(\theta|y_o)$ representa a distribuição *a posteriori* $\pi(\theta|y_o)$ quando $p(y_o|y, \theta)$ for um ponto de massa em $y = y_o$ e se anular para qualquer outro valor em $\mathcal{Y}$. A probabilidade dos dados simulados serem semelhantes aos observados é, em geral, muito pequena, resultando em taxas de aceitação proibitivamente baixas. Tal condição pode ser relaxada definindo-se uma apropriada função distância (por exemplo, a Euclidiana) entre o simulado e o observado, $d(y_o, y)$, a qual deve ser menor que um certo limiar $\epsilon > 0$ pré-definido, de modo a aceitar amostras tais que $y \approx y_o$. Neste caso, passamos a obter uma aproximação para $\pi_{ABC}(\theta|y_o)$ que está condicionada ao evento $d(y_o, y) \leq \epsilon$. Uma outra estratégia é utilizar a comparação com os dados observados usando estatísticas resumidas. Neste caso, define-se

DOI: 10.5540/03.2020.007.01.0448

um vetor $S(\cdot)$ de estatísticas resumidas de dimensão menor que a dos valores observados y_o. Se $S(\cdot)$ for uma estatística altamente informativa ou suficiente, então $\pi(\theta|s_o) \approx \pi(\theta|y_o)$ ou $\pi(\theta|s_o) = \pi(\theta|y_o)$, respectivamente, em que $s_o = S(y_o)$. Usando uma função kernel de suavização K, é comum utilizar duas aproximações para $p(y_o|y,\theta)$, a saber:

$$p_\epsilon(y_o|y,\theta) = \frac{1}{\epsilon}K\left(\frac{d(y_o,y)}{\epsilon}\right) \quad (3) \quad \text{e} \quad p_\epsilon(y_o|y,\theta) = \frac{1}{\epsilon}K\left(\frac{d(s_o,s)}{\epsilon}\right). \quad (4)$$

Observe que $\lim_{\epsilon\to 0} p_\epsilon(y_o|y,\theta)$ é um ponto de massa em y_o. K e ϵ devem ser escolhidos de modo que (3) e (4) ponderem mais as regiões de $\mathcal{Y}$ em que os dados gerados pelo modelo são mais similares aos observados ($y \approx y_o$ ou $s = S(y) \approx S(y_o)$), respectivamente. Por outro lado, usar $S(\cdot)$ é vantajoso quando a dimensão dos dados é grande ou quando os dados estão incompletos. Além disso, os eventos $S(y) \approx S(y_o)$ são significativamente mais prováveis que os $y \approx y_o$, melhorando a eficiência computacional do método (aumenta a taxa de aceitação). Substituindo as definições (3) e (4) em (1) obtemos, respectivamente,

$$\pi(\theta,y|y_o) \propto K\left(\frac{d(y_o,y)}{\epsilon}\right)\pi(y|\theta)\pi(\theta) \quad (5) \quad \text{e} \quad \pi(\theta,y|y_o) \propto K\left(\frac{d(s_o,s)}{\epsilon}\right)\pi(y|\theta)\pi(\theta). \quad (6)$$

Dentre as funções kernel mais comuns destacamos a uniforme, a triangular, a gaussiana e a de Epanechnikov, dentre outras [6]. Diferenças significativas aparecem apenas quando ϵ é grande. Quando este limiar é suficientemente pequeno, as aproximações convergem para a distribuição alvo. O kernel de densidade uniforme é usado mais frequentemente e é o adotado neste trabalho. Nele, considera-se que y (ou $S(y)$) está uniformemente distribuído em uma esfera de raio ϵ centrada em y_o (ou $S(y_o)$), como expresso por:

$$p_\epsilon(y_o|y,\theta) = \begin{cases} 1, & d(y_o,y) \le \epsilon\ ; \\ 0, & \text{caso contrário.} \end{cases} \quad (7)$$

A base conceitual dos métodos ABC [8] é sumarizada pelo método ABC-Rejeição apresentado no Algoritmo 1, no qual $d(y_o,y)$ foi utilizada como medida de distância. A taxa de aceitação usando o algoritmo ABC-Rejeição é em geral muito baixa. A escolha do limiar de aceitação ϵ influencia diretamente a precisão e a eficiência do método. Se $\epsilon = 0$, a taxa de aceitação é computacionalmente proibitiva. Aumentando este limiar, a taxa de aceitação aumenta, resultando em distribuições *a posteriori* mais "largas" que a real, podendo ocorrer distorções inaceitáveis. Note também que a escolha da(s) estatística(s) resumida(s) também é uma tarefa delicada pois ela deve capturar ou sintetizar apropriadamente as informações sobre os parâmetros.

Algoritmo 1: ABC-Rejeição

```
for i = 1 to M do
    repeat ;
        gere amostras θ* a partir de π(θ) ;
        gere amostras y* a partir do simulador π(y|θ*) ;
    until d(y_o, y*) ≤ ε ;
    θ^(i) ← θ* ;
end
```

Em linhas gerais, muitos trabalhos têm sido desenvolvidos objetivando o desenvolvimento de algoritmos de amostragem mais eficientes para o cálculo de $\pi_{ABC}(\theta|y_o)$ e que ao

mesmo tempo reduzam sua dependência com ϵ. A seguir descrevemos um método ABC bastante popular, o ABC Monte Carlo Sequencial [7]. Descrevemos também o método desenvolvido em [2], que é uma aproximação para a computação Bayesiana aproximada.

2.1 ABC Monte Carlo sequencial (ABC-SMC)

O método SMC, também denominado filtro de partículas, é baseado no conceito geral de amostragem por importância e reamostragem [5]. Partindo de uma população inicial para os parâmetros, esta população é propagada, são obtidas distribuições intermediárias até que se obtenha uma que represente a distribuição *a posteriori* desejada. Isto é realizado atendendo à limiares de aceitação de valores decrescentes $\epsilon_1 > \epsilon_2 > \ldots > \epsilon_T > 0$. A propagação das amostras é feita basicamente através de três processos: (i) mutação, usando um kernel de transição; (ii) correção, avaliando pesos associados à amostragem por importância; e (iii) seleção, considerando reamostragem para reduzir a variabilidade dos pesos por importância.

No primeiro passo do ABC-SMC, considera-se uma população inicial com amostras $\{\theta_1^{(1)}, \theta_1^{(2)}, \ldots, \theta_1^{(N)}\}$ geradas da distribuição *a priori* $\pi_1(\theta)$ usando o algoritmo ABC-Rejeição usual, de modo que, com $y \sim \pi(y|\theta_1^{(i)})$, obtém-se $d(y_o, y) \leq \epsilon_1$. Nos t passos seguintes, $t = 2, \ldots, T$, são geradas novas populações usando funções de transição K_t tais que $\theta_t^{(i)} \sim K_t(\theta|\theta^\star)$, com $\theta^\star$ escolhido aleatoriamente dentre os $\theta_{t-1}^{(i)}$ com probabilidade ω_{t-1}^i [3]. Deste modo, em cada passo são obtidas distribuições $\pi_t(\theta)$ com $d(y_o, y) \leq \epsilon_t$, para $y \sim \pi(y|\theta_i^{(t)})$, até atingir a precisão desejada no passo T (a distribuição alvo π_T).

2.2 Aproximação da Computação Bayesiana Aproximada (AABC)

O método AABC foi desenvolvido em [2] para problemas em que uma rodada do simulador computacional de interesse implica em um custo computacional relativamente grande. A ideia básica é obter e utilizar apenas um certo número de simulações *a priori*. Em um típico algoritmo ABC, a distribuição *a posteriori* π_{ABC} que deseja-se estimar é obtida após realizar um número grande de iterações M do método, de modo que desse número são selecionadas apenas as amostras para as quais a distância entre o simulado e observado é menor ou igual à uma tolerância ϵ. O método AABC permite obter uma estimativa π_{AABC} para π_{ABC} utilizando m iterações, com $m << M$. O método AABC pode ser resumido basicamente em três etapas:

1. Amostragem de um número limitado de realizações (m) da distribuição *a priori* dos parâmetros e as respectivas amostras dos dados via o simulador.
2. Amostragem de um novo valor do parâmetro $\theta^\star$ da distribuição a priori e cálculo dos pesos associados usando o kernel de Epanechnikov; seleção de um conjunto com k elementos que tenham peso não nulos. Geração de uma amostra da distribuição de Dirichlet parametrizada pelos pesos.
3. Comparação das estatísticas resumidas dos dados simulados com as dos dados observados, aceitando-se ou rejeitando o valor do parâmetro $\theta^\star$.

DOI: 10.5540/03.2020.007.01.0448

3 Resultados

Nos exemplos 1 e 2 a seguir as populações finais apresentam 1000 e 3000 amostras, respectivamente. No ABC-SMC utilizamos 3 populações intermediárias e kernel de transição gaussiano com desvio σ_{smc}.

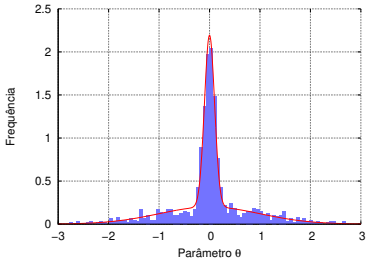
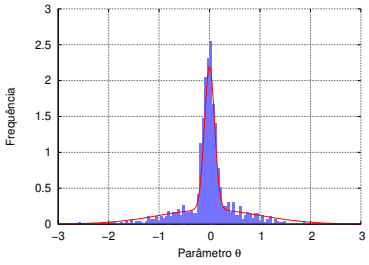
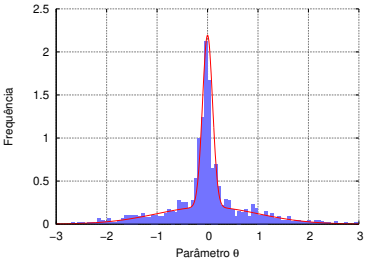
3.1 Exemplo 1: distribuição *a posteriori* conhecida

O exemplo proposto em [7] consiste em determinar a distribuição *a posteriori* da média associada à observação $y_0 = 0$ partindo de uma distribuição *a priori* $\pi(\theta) = U(-10, 10)$ com

$$d(y, y_0) = \begin{cases} \left|\frac{1}{n}\sum_{i=1}^{n} x_i\right|, & \text{com probabilidade 0.5;} \\ |x_1|, & \text{com probabilidade 0.5;} \end{cases} \qquad \text{com} \quad x_i \sim \mathcal{N}(\theta, 1). \tag{8}$$

A distribuição *a posteriori* exata é dada por $\pi(\theta|y_o) = 0.5\mathcal{N}(0, 0.01)+0.5\mathcal{N}(0, 1)$. A Tabela 1 apresenta as distribuições *a posteriori* obtidas para cada método usando $\epsilon = 0,025$, $\epsilon_{smc} = \{2,0; 0,5; 0,025\}$, $\epsilon_{aabc} = 0,025$, $\sigma_{smc} = 1,0$ e $k = 20$. Todos eles foram capazes de obter boas aproximações para a distribuição alvo, com os respectivos valores mais prováveis bastante próximos. Entretanto, o custo de solução com o ABC-Rejeição é quase 400 vezes maior que o com o AABC, enquanto que com o ABC-SMC quase 80 vezes superior (ver números das simulações entre parênteses).

Tabela 1: Comparação das distribuições *a posteriori* obtidas para cada método com a exata (em vermelho). As quantidades de simulações executadas e os valores mais prováveis estão indicados em parênteses e abaixo dos gráficos, respectivamente.

ABC-Rejeição (396173)	ABC-SMC (77142)	AABC (1000)
		
$-1,2176 \times 10^{-4}$	$5,2056 \times 10^{-4}$	$7,02 \times 10^{-4}$

3.2 Exemplo 2: Modelo de Verhulst estocástico

A dinâmica representada na Figura 1, referente ao crescimento *in vitro* de células SUM-149PT (câncer de mama triplo-negativo), pode ser modelada por:

$$\begin{cases} \dfrac{dC}{dt} = rC\left(1 - \dfrac{C}{\kappa}\right) + \mathcal{N}\left(0, \dfrac{\bar{C}}{100}\right) \\ C(t_0) = \bar{C} \end{cases} \tag{9}$$

Com C representando a densidade de células cancerosas, a condição inicial $\bar{C}$, a taxa de crescimento r e a capacidade suporte κ são calibradas usando os dados ilustrados na Figura 1 e as seguintes distribuições *a priori*: $\bar{C} \sim \mathcal{U}(2500, 5000)$, $r \sim \mathcal{U}(3 \times 10^{-2}, 5 \times 10^{-2})$, e $\kappa \sim \mathcal{U}(3, 3 \times 10^4, 3, 65 \times 10^4)$ (em unidades apropriadas). Utilizamos 10 réplicas do modelo para um dado conjunto de parâmetro e usamos a média da resposta do modelo como estatística resumida. Desse modo, é possível comparar os dados simulados com os experimentais através da distância $d(\mathcal{S}(C_i), \mathcal{S}(C_o)), i = 1, \ldots, 10$. Utilizamos $\epsilon = 1587$, $\epsilon_{smc} = \{5000; 3000; 1587\}$, $\epsilon_{aabc} = 5000$, $\sigma_{smc} = \{250; 0,002; 350\}$ e $k = 20$.

Na Tabela 2 apresentamos as distribuições *a posteriori* marginais obtidas com cada método. As distribuições são qualitativamente similares e a ordem de grandeza dos valores mais prováveis dos parâmetros é a mesma. Com 3000 amostras aceitas em todas as populações finais, o custo associado ao AABC é significativamente menor que o dos demais métodos.

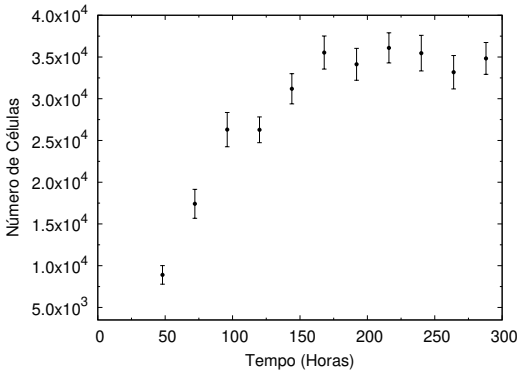


Figura 1: Crescimento *in vitro* da linhagem celular SUM-149PT.

4 Conclusões

A ABC permite realizar a inferência bayesiana sem o conhecimento explícito da função de verossimilhança. Neste trabalho realizamos um estudo utilizando dois métodos clássicos (ABC-Rejeição e ABC-SMC) e um metamodelo com o ABC (AABC). Para os dois exemplos apresentados aqui, todos os métodos conduziram a resultados satisfatórios. O método AABC obteve a distribuição alvo com um número significativamente menor de execuções do que os outros métodos. Esta é uma propriedade intrínseca deste método, construído a partir de um metamodelo definido por um número limitado de simulações do modelo original. Seu uso pode ser potencializado através do aprimoramento do metamodelo, foco da pesquisa em andamento.

Agradecimentos: HLR agradece o apoio da CAPES (projeto 1576953/2016-3). Agradecemos ao Center for Computational Oncology, UTexas at Austin, pelos dados.

Referências

[1] M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[2] E. O. Buzbas and N. A. Rosenberg. AABC: Approximate approximate bayesian computation for inference in population-genetic models. *Theoretical Population Biology*, 99:31–42, 2015.

[3] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *J. R. Statist. Soc. B*, 68(3):411–436, 2006.

[4] G. Karabatsos and F. Leisen. An approximate likelihood perspective on ABC methods. *ArXiv e-prints*, 2018.

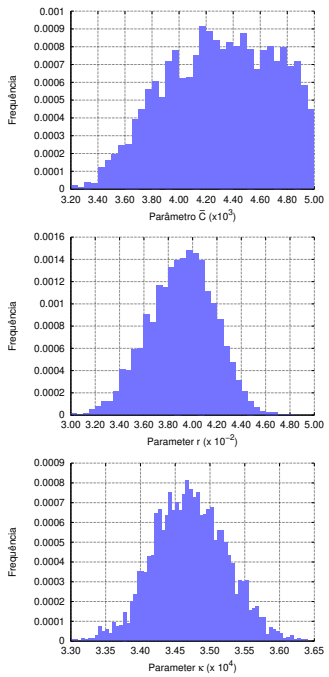
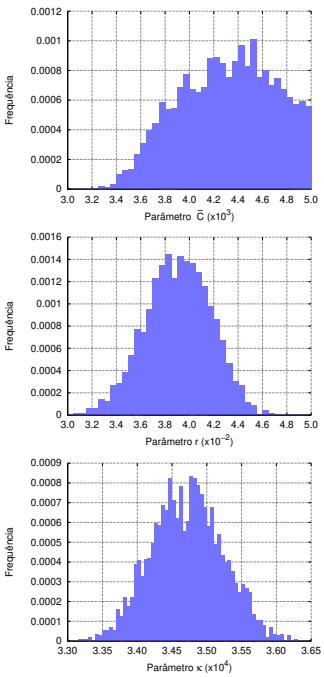
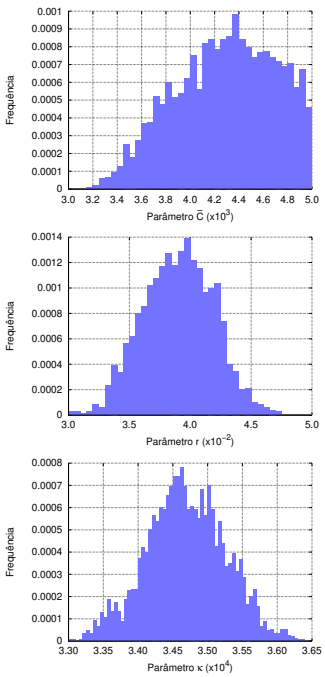
[5] S. A. Sisson and Y. Fan. Likelihood-free Markov chain Monte Carlo. In S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, chapter 12, pages 313–335. Chapman and Hall/CRC Press, 2011.

[6] S. A. Sisson, Y. Fan, and M. A. Beaumont. Overview of approximate bayesian computation. *ArXiv e-prints*, 2018.

[7] S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential Monte Carlo without likelihoods. *PNAS*, 104(6):1760–1765, 2007.

[8] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.

Tabela 2: Distribuições *a posteriori* marginais de cada parâmetro obtidas para cada método. As quantidades de simulações executadas estão indicadas em parênteses e os valores mais prováveis estão indicados em colchetes na ordem $\bar{C}, r$ e κ (em unidades apropriadas).

ABC-Rejeição (77996)	ABC-SMC (84247)	AABC (10000)
		
$[4092; 0,039; 34625]$	$[4153; 0,041; 35129]$	$[3776; 0,0369; 33716]$

