Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Um Estudo de Cadeias Markovianas: Algoritmos Dinâmicos de Monte Carlo

S. J. Ferreira

Relatório Técnico RTE-02/2000

Relatório Técnico
Série Ensino

Um Estudo de Cadeias Markovianas: Algoritmos Dinâmicos de Monte Carlo

Sabino J. Ferreira

Departamento de Estatística
Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627
31.270-010 - Belo Horizonte - MG

Resumo

Algoritmos de Monte Carlo nos permitem simular em computadores um tipo de processo estocástico que são as Cadeias de Markov. Frequentemente, encontramos em trabalhos envolvendo simulações de Monte Carlo, frases do tipo:

> "Fizemos um total de 10^6 iterações do algoritmo, descartamos as primeiras 5×10^5 iterações para equilibrar o processo e nos cálculos das médias temporais usamos uma configuração gerada à cada 100 iterações do algoritmo"

Exceto nos casos dos trabalhos que informem também o valor dos tempos de autocorrelação exponencial e integrado, afirmações como acima não tem qualquer valor. Sem conhecimento dos tempos de autocorrelação, não temos condições de saber se 5×10^5 iterações são suficientes ou não para equilibrar o processo e nem quais são os erros estatísticos das médias temporais calculadas. Para explicar todos os conceitos acima, apresentamos inicialmente as definições e os resultados básicos da teoria de Cadeias de Markov. Em seguida mostramos porque o conhecimento dos tempos de autocorrelação exponencial e integrado é essencial na análise estatística dos dados obtidos em simulações de Monte Carlo e o que fazer na prática para estimar estes tempos de autocorrelação.

Palavras Chaves: algoritmos de Monte Carlo, cadeias de Markov, tempo de autocorrelação

1 Introdução

Nestas notas, farei um estudo de um exemplo de processo estocástico do tipo cadeias de Markov, os chamados *algoritmos dinâmicos de Monte Carlo*, que recebem a denominação geral MCMC (Markov Chain Monte Carlo). Todo trabalho envolvendo o método de Monte Carlo tem a mesma estrutura geral: dada uma distribuição de probabilidade π sobre algum espaço de configurações ou estados $S \equiv \{C\}$, gostaríamos de formular um algoritmo que nos permitisse gerar uma sequência de configurações C_n de S com probabilidade $\pi(C_n)$. Ou seja, dada uma função real f definida em S, gostaríamos que a sequência de configurações C_n geradas pelo nosso algoritmo fosse tal que:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(C_n) \;=\; \sum_{C\in S} f(C)\pi(C) \;=\; E_\pi[f] \tag{1}$$

Na prática, partindo de uma configuração inicial *arbitrária*, utilizamos este algoritmo para simular o processo estocástico correspondente em computadores. Considerando como uma unidade de tempo uma iteração básica do algoritmo, calculamos médias temporais do tipo:

$$\overline{f} \;=\; \frac{1}{N}\sum_{n=1}^{N} f(C_n) \tag{2}$$

que tomamos como estimativas da média estatística ou esperança $E_\pi[f]$. Como, em geral, os processos estocásticos simulados são do tipo cadeia de Markov, na segunda seção farei uma pequena revisão dos resultados básicos da teoria de cadeias de Markov. Também discutirei em quais condições as médias temporais $\overline{f}$ convergem para a média estatística $E_\pi[f]$, bem como a taxa desta convergência (tempo de autocorrelação) e os erros estatísticos nas estimativas de $E_\pi[f]$.

Na terceira seção, serão discutidas questões práticas relacionadas com a análise estatística dos dados obtidos a partir de algoritmos dinâmicos de Monte Carlo. Na verdada,e esta é parte mais importante destas notas: pressupondo que o leitor tenha um conhecimento básico da teoria de Cadeias de Markov no nível dos cursos de processos estocásticos oferecidos nos cursos de graduação em Estatística, o objetivo é explicar porque, ao se fazer uma simulação de MCMC, devemos descartar inicialmente um determinado número de configurações geradas e depois, ao calcular as médias temporais (2), somente incluir configurações geradas que sejam separadas entre si por um determinado número de passos do algoritmo.

2 Cadeias de Markov e Métodos Dinâmicos de Monte Carlo: Teoria Geral

Como afirmamos anteriormente, a estrutura geral de qualquer algoritmo de MCMC é: dada uma medida de probabilidade π sobre determinado espaço de configurações

S, gostaríamos de um algoritmo que gerasse uma sequência de configurações $C_n \in S$ com probabilidade $\pi(C_n)$.

Podemos classificar os métodos de MCMC em duas categorias: *estáticos* e *dinâmicos*. Os métodos estáticos são aqueles em que as sucessivas configurações C_n da cadeia são *estatisticamente independentes*. Estas técnicas são usadas principalmente em integrações numéricas em espaços com um número pequeno de dimensões [1]. Entretanto, estas técnicas são completamente inadequadas para tratar problemas onde temos um grande número de graus de liberdade não independentes.

No caso do métodos de MCMC *dinâmicos*, a idéia é construir um processo estocástico que tenha um espaço de estados S e π como única distribuição de equilíbrio. Ou seja, simulamos este processo estocástico em um computador partindo de uma configuração inicial arbitrária. Uma vez que o sistema tenha atingido o equilíbrio, utilizamos as configurações geradas para calcular a média temporal do tipo (2). Sob determinadas condições, esperamos que no limite onde o número de iterações tende ao infinito, a média temporal convirja para a média $E_\pi[.]$. Dizemos, então, que estamos gerando uma *evolução temporal estocástica* para um dado sistema.

Em geral, os processos estocásticos simulados são *processos de Markov*. No que segue, vamos apresentar alguns resultados básicos da teoria geral de processos de Markov. Por simplicidade, vamos considerar que nosso espaço de estados S seja discreto (finito ou infinito porém contável). Ver em [2, 3, 6] um tratamento para cadeias de Markov em espaços de estados *finitos* e em [4] no caso em que o espaço de estados S é contável. No caso mais geral, o espaço de estados S é um subconjunto de $\mathbf{R}^d$. Neste caso, muitos resultados da teoria do caso discreto podem ser estendidos para o caso geral com as respectivas substituições de somas por integrais, matrizes por núcleos de operadores porém as demonstrações são bem mais difíceis (ver em [5]).

De uma maneira abreviada, podemos afirmar que uma cadeia de Markov (que é um processo de Markov com tempo discreto) com espaço de estados S é uma sequência de variáveis aleatórias $X_0, X_1, X_2, \ldots$ que assumem valores em S tal que as sucessivas transições $X_t \to X_{t+1}$ são *estatisticamente independentes*. Mais precisamente, dizemos que uma cadeia de Markov é especificada por dois ingredientes:

- Uma *distribuição inicial* α. Aqui, α é uma distribuição de probabilidade em S e o processo é definido de tal forma que $\mathbb{P}(X_0 = x) = \alpha_x$.

- Uma *matriz probabilidade de transição* $P = \{p_{xy}\}_{x,y\in S} = \{p(x \to y)\}_{x,y\in S}$. Aqui a matriz P é tal que $p_{xy} \geq 0$, para todo par x, y e a soma $\sum_y p_{xy} = 1$ para todo x. O processo é definido de tal forma que $\mathbb{P}(X_{t+1} = y \,|\, X_t = x) = p_{xy}$.

Ou seja, a cadeia de Markov é completamente especificada pelas probabilidades conjuntas

$$\mathbb{P}(X_0 = x_0,\, X_1 = x_1,\, X_2 = x_2,\, \ldots,\, X_n = x_n) \;\equiv\; \alpha_{x_0}\, p_{x_0 x_1}\, p_{x_1 x_2} \cdots p_{x_{n-1} x_n}\,. \quad (3)$$

onde o produto das probabilidades expressa justamente o fato das transições sucessivas serem independentes.

Definimos também a probabilidade de transição após *n-passos* da cadeia como:

$$p_{xy}^{(n)} \;=\; \mathbb{P}(X_{t+n} = y \,|\, X_t = x)\ . \tag{4}$$

onde $p_{xy}^{(0)} = I(x = y)$, ou seja, $I = 1$ quando $x = y$, de outro modo, $I = 0$ e $p_{xy}^{(1)} = p_{xy}$, sendo que em geral, $\{p_{xy}^{(n)}\}$ são elementos de matriz de P^n.

Uma cadeia de Markov é dita ser *irredutível* se a partir de um estado qualquer é possível se atingir qualquer outro estado de S, i.e., se para cada par $x, y \in S$, existir um inteiro $n \geq 0$ tal que $p_{xy}^{(n)} > 0$. Em geral, na literatura sobre cadeias de Markov com espaço de estados *finito*, o termo "ergódico" é usado como sinônimo para "irredutível", porém isto não é válido no caso de espaços de estados genéricos. Aqui, estaremos sempre tratando de cadeias de Markov irredutíveis.

Para cada estado $x \in S$, definimos o *período de* x (denotado por d_x) como o maior divisor comum dos números $n > 0$ tal que $p_{xx}^{(n)} > 0$. Se $d_x = 1$, o estado x é chamado *aperiódico*. Pode-se mostrar que, para uma cadeia irredutível, todos os estados tem o mesmo período. Neste caso podemos nos referir a uma cadeia de período d. Além disto, o espaço de estados pode ser particionado em subconjuntos $S_1, S_2, \ldots, S_d$, ao longo dos quais a cadeia move-se ciclicamente, i.e., $p_{xy}^{(n)} = 0$ sempre que $x \in S_i$, $y \in S_j$ e $j - i \neq n \pmod d$. Também pode-se mostrar que uma cadeia é *irredutível e aperiódica* se e somente se, para todo par x, y, existir N_{xy} tal que $p_{xy}^{(n)} > 0$ para *todo* $n \geq N_{xy}$.

Agora vamos tratar de um tópico fundamental na teoria de cadeia de Markov: o problema da convergência para o equilíbrio. A distribuição de probabilidade $\pi = \{\pi_x\}_{x \in S}$ é denominada uma *distribuição de probabilidade estacionária* (ou *distribuição de probabilidade de equilíbrio*) de uma cadeia de Markov P no caso em que:

$$\sum_x \pi_x\, p_{xy} \;=\; \pi_y \qquad \text{para todo } y \in S. \tag{5}$$

No caso da distribuição de equilíbrio existir, temos o seguinte resultado:

Teorema Seja P a matriz probabilidade de transição de uma cadeia de Markov *irredutível* de período d. Se existir uma distribuição de equilíbrio π, então ela é única e tal que $\pi_x > 0$ para todo x. Além disto,

$$\lim_{n\to\infty} p_{xy}^{(nd+r)} \;=\; \begin{cases} d\pi_y & \text{se } x \in S_i,\ y \in S_j \text{ com } j - i = r \pmod d \\ 0 & \text{se } x \in S_i,\ y \in S_j \text{ com } j - i \neq r \pmod d \end{cases} \tag{6}$$

para todo x, y. Em particular, se P é aperiódica, então

$$\lim_{n\to\infty} p_{xy}^{(n)} \;=\; \pi_y\ . \tag{7}$$

Os limites acima significam que:

$$\lim_{n\to\infty} \|p_{x.}^{(n)} - \pi_{(.)}\| \;=\; 0\ . \tag{8}$$

onde a distância entre duas distribuições π^1 e π^2 dada por:

$$\|\pi^1 - \pi^2\| \;=\; \sup_{x \in S} |\pi_x^1 - \pi_x^2|\ . \tag{9}$$

Este teorema nos mostra que, independente da distribuição inicial α, a cadeia de Markov converge para a distribuição de equilíbrio π quando $t \to \infty$. Considerando funções reais f, definidas no espaço de estados S, que sejam de quadrado integrável com respeito a π $(\sum_x \pi_x |f(x)|^2)^{1/2} < \infty$) e sob a mesmas condições deste teorema, podemos provar que:

1. Médias temporais destas funções f convergem para médias estatísticas com relação a π com probabilidade 1 (lei dos grande números).

2. Esta convergência se faz com flutuações da ordem de $\sim n^{-1/2}$ (teorema central do limite).

Detalhes dos enunciados e provas destes teoremas podem ser vistos em [4].
Temos agora uma receita geral de um método dinâmico de Monte Carlo para gerar configurações de acordo com determinada distribuição de probabilidade π. Para isto, temos que construir uma matriz probabilidade de transição $P = \{p_{xy}\} = \{p(x \to y)\}$ que satisfaça as seguintes condições :

(A) *Irredutibilidade* - Para cada par $x, y \in S$, deve existir um inteiro $n \geq 0$ tal que $p_{xy}^{(n)} > 0$.

(B) *Estacionariedade de* π - Para todo $y \in S$,

$$\sum_x \pi_x \, p_{xy} = \pi_y. \tag{10}$$

Notamos que uma maneira conveniente de satisfazer a condição (B) é satisfazer uma condição mais forte:

(B′) *Balanço Detalhado* - Para cada par $x, y \in S$,

$$\pi_x p_{xy} = \pi_y p_{yx}. \tag{11}$$

onde basta somar (B′) sobre x para obter novamente (B). A condição (B′) é chamada *condição do balanço detalhado*. Uma cadeia de Markov satisfazendo a (B′) é denominada *reversível* [2, 3].
O teorema e suas consequências acima mostram que a simulação de uma cadeia de Markov P constitui um legítimo método de Monte Carlo para se estimar médias com respeito à medida de probabilidade π. Porém, existe um detalhe a qual devemos ficar atentos: como a cadeia de Markov é inicializada em uma distribuição α, que em geral, é diferente da distribuição de equilíbrio π, as configurações inicialmente geradas na cadeia constituem uma espécie de "transiente inicial" e não refletem a distribuição de equilíbrio desejada π. Este transiente introduz nas médias temporais um *erro sistemático* ou *bias* ou *tendência*. Na prática, para eliminar este bias, temos que descartar um certo número de configurações geradas no "transiente inicial" e só então passamos a utilizar as configurações geradas na cadeia para calcular as médias temporais. Este processo de descarte das configurações geradas no transiente inicial

se denomina "burn-in" ou aquecimento da cadeia. Entretanto, assintoticamente o bias é da ordem de n^{-1} e portanto muito menor do que as flutuações estatísticas que são da ordem de $n^{-1/2}$.

Entretanto, apesar de um algoritmo de MCMC ser correto, pode acontecer que ele não seja *eficiente*. O problema é que os sucessivos estados X_0, X_1, $\ldots$ da cadeia de Markov são *correlacionados*, algumas vezes fortemente, fazendo com que a variância das médias temporais obtidas pela simulação dinâmica de Monte Carlo seja muito maior que no caso dos algoritmos estáticos, onde os sucessivos estados são independentes. Para precisar melhor a natureza deste problema vamos considerar uma função real f que seja de quadrado integravel em relação a π. Consideramos também uma cadeia de Markov *estacionária*, ou seja, começamos com o sistema já na distribuição de equilíbrio π ou então "equilibramos" o sistema antes de usar as configurações geradas. Assim, seja $\{f_t\} \equiv \{f(X_t)\}$ um processo estocástico estacionário com média:

$$\mu_f \equiv E_\pi[f] = \sum_{x \in S} \pi_x \, f(x) \tag{12}$$

e *função autocorrelação não normalizada**

$$\begin{aligned} C_{ff}(t) &\equiv E_\pi[f_s f_{s+t}] - \mu_f^2 \\ &= \sum_{x,y} f(x) \, [\pi_x p_{xy}^{(|t|)} - \pi_x \pi_y] \, f(y). \end{aligned} \tag{13}$$

A *função autocorrelação normalizada* será:

$$\rho_{ff}(t) \equiv C_{ff}(t)/C_{ff}(0). \tag{14}$$

Típicamente, $\rho_{ff}(t)$ decai exponencialmente ($\sim e^{-|t|/\tau}$) quando t é grande. Definimos, então, o chamado *tempo de autocorrelação exponencial* como:

$$\tau_{exp,f} = \limsup_{t \to \infty} \frac{t}{-\log |\rho_{ff}(t)|} \tag{15}$$

e

$$\tau_{exp} = \sup_f \, \tau_{exp,f}. \tag{16}$$

Para mostrar que o tempo τ_{exp} controla a taxa de convergência para o equilíbrio de uma cadeia representada por uma matriz probabilidade de transição P, vamos considerar que cadeia é inicializada em uma distribuição α diferente da distribuição de equilíbrio π. Após t passos da cadeia, a distribuição de probabilidade é $\alpha P_x^t = \sum_{y \in S} \alpha_y p_{yx}^{(t)}$. Pode-se mostrar† que a diferença entre as distibuições no sentido de (9) é tal que:

$$\|\alpha P^t - \pi\| \;\leq\; \exp(-t/\tau_{exp}) \; \|\alpha - \pi\|. \tag{17}$$

Dada uma função real f no espaço de estados S, definimos o chamado *tempo de autocorrelação integrado* como:

*Também denominada função autocovariância.

†Ver os detalhes em [7]

$$
\begin{aligned}
\tau_{int,f} &= \frac{1}{2} \sum_{t=-\infty}^{\infty} \rho_{ff}(t) \qquad (18) \\
&= \frac{1}{2} + \sum_{t=1}^{\infty} \rho_{ff}(t)
\end{aligned}
$$

O fator $\frac{1}{2}$ é inserido apenas para que $\tau_{int,f} \approx \tau_{exp,f}$ se $\rho_{ff}(t) \sim e^{-|t|/\tau}$ quando $\tau \gg 1$. O tempo de autocorrelação integrado controla o erro estatístico nas estimativas de Monte Carlo para a média $E_\pi[f]$. Mais precisamente, considerando a média temporal

$$
\bar{f} \equiv \frac{1}{n} \sum_{t=1}^{n} f_t \qquad (19)
$$

sua variância é dada por:

$$
\begin{aligned}
\mathrm{var}(\bar{f}) &= \frac{1}{n^2} \sum_{r,s=1}^{n} C_{ff}(r-s) \qquad (20) \\
&= \frac{1}{n} \sum_{t=-(n-1)}^{n-1} \left(1 - \frac{|t|}{n}\right) C_{ff}(t) \qquad (21) \\
&\approx \frac{1}{n}(2\tau_{int,f})\, C_{ff}(0) \quad \text{for } n \gg \tau \qquad (22)
\end{aligned}
$$

A conclusão é que a variância de $\bar{f}$ é um fator de $2\tau_{int,f}$ vezes maior do que seria obtido no caso dos sucessivos valores $\{f_t\}$ serem estatisticamente independentes. Ou seja, em uma média temporal com n configurações, o número de "configurações efetivamente independentes" é aproximadamente $n/2\tau_{int,f}$.

Em resumo, os tempos de autocorrelação τ_{exp} e $\tau_{int,f}$ desempenham papéis diferentes nas simulações de Monte Carlo. τ_{exp} representa uma cota superior do número de iterações n_{desc} que devemos descartar no início, antes que o sistema esteja equilibrado. Em geral, $n_{desc} \approx 20\tau_{exp}$ é um valor adequado. Este é o processo de "burn-in" a que nos referimos anteriormente. Por outro lado, $\tau_{int,f}$ controla os erros estatísticos cometidos nas estimativas de $E_\pi[f]$ uma vez que o sistema esteja equilibrado. Em geral, podemos considerar na prática que τ_{exp} e $\tau_{int,f}$ são da mesma ordem de grandeza, pelo menos no caso de funções f que sejam de quadrado integrável.

3 Análise Estatística de Dados obtidos por Métodos Dinâmicos de Monte Carlo

Em muitos trabalhos publicados com simulações de Monte Carlo, frequentemente encontramos afirmações do tipo:

> "Foram feitas um total de 100000 iterações, descartamos as primeiras 50000 iterações para equilibrar o sistema e no cálculo das média temporais usamos uma configuração a cada 100 iterações do algoritmo."

Exceto nos casos dos trabalhos que informem também o valor dos tempos de autocorrelação exponencial e integrado, afirmações como acima não tem qualquer valor. Sem conhecimento dos tempos de autocorrelação, não temos condições de saber se 50000 iterações são suficientes ou não para equilibrar o sistema e quais são os erros estatísticos das médias temporais calculadas.
Nesta seção, vamos discutir dois pontos essenciais em qualquer simulação pelo método de Monte Carlo:

1 - *Bias de Inicialização* - Se a cadeia de Markov começa de acordo com uma distribuição de probabilidade α diferente da distribuição de equilíbrio π, então haverá um "transiente inicial" durante o qual as configurações geradas não refletem a distribuição de equilíbrio desejada π. Isto gera um erro sistemático (bias) nas médias temporais.

2 - *Autocorrelação no Equilíbrio* - Uma vez equilibrada, as configurações geradas pela cadeia de Markov apresentam correlação estatística entre si. Esta correlação faz com que o erro estatístico (variância) das médias temporais seja um fator $2\tau_{int,f}$ maior do que seria no caso em que as sucessivas configurações são estatisticamente independentes.

1 – Frequentemente, a cadeia de Markov é inicializada em alguma configuração específica $x \in S$, neste caso, $\alpha_y = I(y = x)$. Ou, então, poderíamos fazer com que a cadeia de Markov fosse inicializada com uma configuração aleatória escolhida de acordo com uma distribuição de probabilidade α distinta da distribuição de equilíbrio π. Neste caso, dizemos que o sistema está inicialmente "fora do equilíbrio". O teorema na seção anterior garante que o sistema converge para o equilíbrio quando $t \to \infty$, mas nada afirma sobre a *taxa* desta convergência.
Conhecendo o tempo de autocorrelação exponencial τ_{exp}, podemos estabelecer uma cota superior para o tempo (número de iterações do algoritmo de MCMC) que teríamos que considerar antes que pudéssemos afirmar que, para propósitos práticos, o sistema está equilibrado. Por exemplo, se esperarmos um tempo de $20\tau_{exp}$, então a distância da distribuição de equilíbrio no sentido de (9) seria de, no máximo, e^{-20} ($\approx 2 \times 10^{-9}$) vezes o desvio inicial. Infelizmente, não sabemos em geral como calcular τ_{exp} ou mesmo obter uma cota superior rigorosa para ele. Neste caso, tentamos estimar τ_{exp} empiricamente. O que se faz é medir a função de autocorrelação normalizada $\rho_{ff}(t)$ para um conjunto adequadamente escolhido de funções f, já que tipicamente esperamos que $\rho_{ff}(t)$ decaia exponencialmente como ($\sim e^{-|t|/\tau}$). Porém, se ao realizar este estudo empírico, deixarmos de incluir neste conjunto alguma função que tenha uma superposição com algum modo lento do processo, estaremos subestimando τ_{exp}.
Pode também acontecer da taxa real de convergência para o equilíbrio, a partir de uma distribuição inicial α, ser muito maior que a estimativa fornecida por τ_{exp}.

Neste caso, o que se faz empiricamente para se determinar quando o equilíbrio é alcançado é a identificação visual de quando o transiente inicial termina nos gráficos das médias $\overline{f}$ de um conjunto selecionado de observaveis como funções do "tempo", que representa o número de iterações N da cadeia.
O perigo maior para todos métodos empíricos para determinação de quando o equilíbrio é alcançado é a possibilidade da ocorrência de *metaestabilidade*. Apesar de tudo indicar que o equilíbrio foi alcançado, pode acontecer do sistema ficar confinado a uma região do espaço de configurações bem distante das configurações de equilíbrio. É claro que a obtenção de uma cota superior rigorosa para τ_{exp} seria uma garantia da inexistência de metaestabilidade. Também poderíamos tentar heuristicamente obter argumentos teóricos que indicassem ser bastante improvável a presença de metaestabilidade. Entretanto, se não temos como justificar teoricamente a ausência de metaestabilidade, sempre é útil ter uma idéia das possíveis regiões de metaestabilidade no espaço de configurações. Neste caso, o aconselhavel seria fazer diversas simulações com diferentes condições iniciais que correspondessem a configurações típicas destas possíveis regiões de metaestabilidade e depois testar a consistência da médias temporais obtidas. É claro que a consistência dos resultados não é uma garantia da ausência de metaestabilidade, mas, com certeza representa um aumento no grau de confiança destes.
Agora, podemos considerar que, de um modo ou outro, sabemos quanto tempo devemos esperar para que o equilíbrio seja alcançado. Isto significa que, ao fazer simulações de MCMC, devemos descartar todas as configurações geradas dentro do transiente inicial caracterizado por um certo tempo n_{desc} e só incluir configurações subsequentes em nossas médias. Como já dissemos, isto não é necessário assintoticamente, já que os erros sistemáticos gerados pelo transiente inicial são da ordem de τ/n, enquanto os erros estatísticos são da ordem de $(\tau/n)^{1/2}$. Na prática, pode acontecer do coeficiente de τ/n ser razoavelmente grande no caso da configuração inicial estar muito distante da configuração de equilíbrio.
2 – Como explicamos na seção anterior, a variância das médias temporais $\bar{f}$ obtidas através dos métodos dinâmicos de MCMC é um fator de $2\tau_{int,f}$ maior que a variância das médias calculadas a partir de configurações descorrelacionadas. Ou seja, uma simulação com um total de n iterações contém na verdade $n/2\tau_{int,f}$ "configurações efetivamente independentes". Isto significa que a *eficiência computacional* de um algoritmo é determinada essencialmente pelo tempo de autocorrelação.
O conhecimento de $\tau_{int,f}$ também é essencial para se determinar o número total n de iterações de uma simulação, bem como as barras de erros na estimativas de $E_\pi[f]$. Como as barras de erros são da ordem de $(\tau/n)^{1/2}$, se quisermos uma precisão de 1% nas estimativas, precisamos de um número de iterações $n \approx 10000\tau$.
Além disto, como um teste de auto-consistência para os resultados obtidos em uma simulação, o número total de iterações n deve ser muito maior que as estimativas de τ produzidas pela mesma simulação. De outro modo estes resultados não seriam confiáveis.
Para ter uma idéia da importância relativa do bias inicial e da autocorrelação ao longo da cadeia nos resultados das simulações de MCMC, vamos assumir que o tempo de autocorrelação exponencial τ_{exp} e o tempo de autocorrelação integrado

$\tau_{int,f}$, pelo menos para funções f de interesse, sejam comparáveis (o que, em geral, acontece). Neste caso, o bias inicial é um problema relativamente simples se comparado ao da autocorrelação no equilíbrio. Para se eliminar o bias inicial, para propósitos práticos, basta descartarmos da ordem de $\approx 20\tau$ configurações iniciais, porém, para que as nossas estimativas tenham precisão de pelo menos 1%, o número total de configurações geradas tem que ser $\approx 10000\tau$ ou maior.

No restante desta seção, vamos considerar que os dados foram obtidos a partir de simulações de MCMC onde as configurações geradas já estão devidamente "equilibradas". Mostramos como estimar o tempo de autocorrelação $\tau_{int,f}$ e como calcular barras de erro (nossas referências básicas são os livros de análise de séries temporais [9, 10]).

Seja $\{f_t\}$ um processo estocástico real com média com respeito a distribuição de probabilidade π:

$$\mu \;\equiv\; E_\pi[f_t] \;=\; \sum_{x\in S} f(x)\pi_x \;, \tag{23}$$

função correlação não-normalizada

$$\begin{aligned} C(t) \;&\equiv\; E_\pi[f_s f_{s+t}] - \mu_f^2 \\ &=\; \sum_{x,y} f(x)\,[\pi_x p_{xy}^{(|t|)} - \pi_x\pi_y]\, f(y). \end{aligned} \tag{24}$$

função correlação normalizada

$$\rho(t) \;\equiv\; C(t)/C(0) \;, \tag{25}$$

e tempo de autocorrelação integrado

$$\tau_{int} \;=\; \frac{1}{2}\sum_{t=-\infty}^{\infty} \rho(t) \;. \tag{26}$$

O nosso objetivo é obter estimativas de μ, $C(t)$, $\rho(t)$ e τ_{int} a partir de um número finito (porém grande) de valores f_1, ..., f_n gerados por este processo estocástico.

O estimador natural de μ é a média temporal

$$\overline{f} \;\equiv\; \frac{1}{n}\sum_{i=1}^{n} f_i \;. \tag{27}$$

É um estimador sem bias, já que $E_\pi[f] = \mu$ e tem variância

$$\mathrm{var}(\overline{f}) \;=\; \frac{1}{n}\sum_{t=-(n-1)}^{n-1} (1 - \frac{|t|}{n})\; C(t) \tag{28}$$

$$\approx\; \frac{1}{n}\,(2\tau_{int})\; C(0) \qquad \text{for } n \gg \tau \tag{29}$$

Mesmo que nosso único interesse seja calcular a quantidade estática μ, é necessário que tenhamos uma estimativa da grandeza dinâmica τ_{int} para calcular as barras de erro corretas para μ.

O estimador natural para $C(t)$ é

$$\widehat{C}(t) \ \equiv \ \frac{1}{n \ - \ |t|} \sum_{i=1}^{n-|t|} (f_i \ - \ \mu)(f_{i+|t|} \ - \ \mu) \tag{30}$$

quando a média μ é conhecida, e

$$\widehat{\widehat{C}}(t) \ \equiv \ \frac{1}{n \ - \ |t|} \sum_{i=1}^{n-|t|} (f_i \ - \ \overline{f})(f_{i+|t|} \ - \ \overline{f}) \tag{31}$$

no caso de não conhecermos μ. É importante observar que, para cada t, enquanto a função autocorrelação $C(t)$ é um *número real*, os estimadores $\widehat{C}(t)$ e $\widehat{\widehat{C}}(t)$ são *variáveis aleatórias*. Adiante veremos que esta distinção é importante também na prática. Como estimadores de $C(t)$, $\widehat{C}(t)$ não possui bias, enquanto $\widehat{\widehat{C}}(t)$ tem um bias da ordem de $1/n$ [10]. A variância e covariância de $\widehat{C}(t)$ são [9, 10]:

$$\begin{aligned}
\mathrm{var}(\widehat{C}(t)) \ &= \ \frac{1}{n} \sum_{m=-\infty}^{\infty} \Big[C(m)^2 + C(m+t)C(m-t) + \kappa(t,\, m,\, m+t)\Big] \\
&\quad + o\left(\frac{1}{n}\right) \qquad (32) \\
\mathrm{cov}(\widehat{C}(t), \widehat{C}(u)) \ &= \ \frac{1}{n} \sum_{m=-\infty}^{\infty} [C(m)C(m+u-t) + C(m+u)C(m-t) \\
&\quad + \kappa(t,\, m,\, m+u)] + o\left(\frac{1}{n}\right) \qquad (33)
\end{aligned}$$

onde $t, u \geq 0$ e κ é a função autocorrelação de 4-pontos

$$\begin{aligned}
\kappa(r,\, s,\, t) \ \equiv \ & E_\pi[(f_i - \mu)(f_{i+r} - \mu)(f_{i+s} - \mu)(f_{i+t} - \mu)]\rangle \\
& - C(r)C(t-s) - C(s)C(t-r) - C(t)C(s-r)\,.
\end{aligned} \tag{34}$$

É importante observarmos que o comportamento $\widehat{C}(t)$ é idêntico ao de $\widehat{\widehat{C}}(t)$ considerando termos até a ordem de $1/n$.
Como estimador natural de $\rho(t)$, vamos usar:

$$\widehat{\rho}(t) \ \equiv \ \widehat{C}(t)/\widehat{C}(0) \tag{35}$$

no caso da média μ ser conhecida, e

$$\widehat{\widehat{\rho}}(t) \ \equiv \ \widehat{\widehat{C}}(t)/\widehat{\widehat{C}}(0) \tag{36}$$

quando não conhecemos μ. Para n suficientemente grande, a variância e covariância de $\widehat{\rho}(t)$ também podem ser calculadas a partir de (33) (ver em [9, 10]).
Neste caso o chamado estimador "natural" para τ_{int} seria

$$\widehat{\tau}_{int} \ \stackrel{?}{\equiv} \ \frac{1}{2} \sum_{t=-(n-1)}^{n-1} \widehat{\rho}(t) \tag{37}$$

ou analogamente com $\widehat{\widehat{\rho}}$. Porém esta escolha *não é correta*. O que ocorre é que o estimador definido em (37) tem uma variância que não tende a zero quando n tende a infinito [9]. De uma maneira não rigorosa, podemos dizer que isto ocorre porque as autocorrelações $\widehat{\rho}(t)$ para $|t| \gg \tau$ contém muito "ruído" e pouco "sinal". Porém, o número destas contribuições é da ordem de n e estas, ao serem somadas, resultam em uma variância da ordem de 1 [9]. A solução é a introdução de uma janela $\lambda(t)$ tal que $\lambda \approx 1$, para $|t| \lesssim \tau$ e $\lambda \approx 0$ para $|t| \gg \tau$, que funcione como um "corte" para a soma em (37). Neste caso, o estimador para τ_{int} é:

$$\widehat{\tau}_{int} \;\equiv\; \frac{1}{2} \sum_{t=-(n-1)}^{n-1} \lambda(t)\, \widehat{\rho}(t) \,. \tag{38}$$

Dada a sua simplicidade, uma boa escolha é a janela retangular

$$\lambda(t) = \begin{cases} 1 & \text{se } |t| \leq M \\ 0 & \text{se } |t| > M \end{cases} \tag{39}$$

onde M é um convenientemente escolhido *fator de corte*. O preço a pagar é a introdução de um bias:

$$\text{bias}(\widehat{\tau}_{int}) \;=\; -\frac{1}{2} \sum_{|t|>M} \rho(t) + o\left(\frac{1}{n}\right) \,. \tag{40}$$

Quanto à variância de $\widehat{\tau}_{int}$, a partir de (33), podemos obter

$$\text{var}(\widehat{\tau}_{int}) \;\approx\; \frac{2(2M+1)}{n}\, \tau_{int}^2 \,, \tag{41}$$

com a condição que $\tau \ll M \ll n$. Vemos então que a escolha de M é um compromisso entre o bias e a variância: podemos fazer o bias pequeno se tomarmos M grande o suficiente tal que $\rho(t)$ seja desprezível para $|t| > M$, enquanto mantemos a variância pequena, controlando para que M não fique maior que o necessário para controlar o tamanho do bias. Em geral, na prática, basta tomar M poucas vezes maior que τ. Porém, uma pergunta permanece: quantas vezes maior? Uma possível resposta é dada por uma técnica semi-empírica denominada "janelamento automático" [8] onde M é escolhido como o *menor* inteiro tal que

$$M \;\geq\; c\, \widehat{\tau}_{int}(M) \tag{42}$$

Quando a função de autocorrelação normalizada tem decaimento exponencial ($\rho(t) \sim e^{-|t|/\tau}$ para t grande) é suficiente considerar $c \approx 4$ (já que $e^{-4} < 2\%$). Porém quando $\rho(t)$ apresenta um decaimento assintótico menor que exponencial, tomamos para c valores entre 6 e 10. Na prática, quando o número de iterações $n \gtrsim 1000\tau$, este procedimento funciona bem. Mas, em geral, a escolha de c é feita de uma maneira empírica. Seria interessante que o próprio algoritmo de "janelamento automático" fornecesse um valor otimizado para c sempre como um compromisso entre o bias e a variância, dependendo, é claro, de n.

Referências

[1] J.M. Hammersley and D.C. Handscomb, *Monte Carlo Methods* (Methuen, London, 1964).

[2] J. G. Kemeny and J. L. Snell, *Finite Markov Chains* (Springer, New York, 1976).

[3] M. Iosifescu, *Finite Markov Processes and Their Applications* (Wiley, Chichester, 1980).

[4] K. L. Chung, *Markov Chains with Stationary Transition Probabilities*, 2*nd* ed. (Springer, New York, 1967).

[5] E. Nummelin, *General Irreducible Markov Chains and Non-Negative Operators*, (Cambridge University Press, Cambridge, 1984).

[6] S. Ross, *Stochastic Processes*, 2*nd* ed. (Wiley, New York, 1996).

[7] A. D. Sokal e L. E. Thomas, J. Stat. Phys. **54**, 797 (1989).

[8] N. Madras e A.D. Sokal, J. Stat. Phys. **50**, 109 (1988).

[9] M.B. Priestley, *Spectral Analysis and Time Series*, 2 vols. (Academic, London, 1981).

[10] T.W. Anderson, *The Statistical Analysis of Time Series* (Wiley, New York, 1971).