Análise de Dados e Simulação

Márcia Branco

Universidade de Sao Paulo
Instituto de Matematica e Estatistica
http:www.ime.usp.br/m̃branco

Amostragem por Importância

Amostragem por Importância

- Muitas vezes é usando o termo em inglês *Importance Sampling* e a sigla *IS*.
- Pode ser considerada como uma técnica de redução de variância, mas também pode ser usada simplesmente por facilidade de simulação.
- Considera uma densidade proposta da qual se saiba simular, obtem-se as amostras de MC a partir desta proposta. O estimador de MC-IS é obtido fazendo uma média ponderada dos valores simulados, em que os pesos dependem da razão entre as densidades de interesse e proposta $\frac{f(x)}{g(x)}$.

Amostragem por Importância

Interesse:

$$E[h(X)] = \int h(x)f(x)dx.$$

Em que X é uma v.a. com f.d.p. $f(x)$.
Seja $g(x)$ uma f.d.p. com mesmo suporte de X, então

$$E[h(X)] = \int h(x)\frac{f(x)}{g(x)}g(x)dx.$$

O estimador de Monte Carlo por importância (MC-IS) é dado por:

$$T_{IS}(X) = \frac{1}{R}\sum_{i=1}^{R}\frac{f(x_i)}{g(x_i)}h(x_i).$$

Em que $x_1, x_2, \ldots, x_R$ são simulados de densidade proposta g.

Teorema: A variância do estimador $T_{IS}(X)$ é minima quando a densidade proposta é dada por

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(z)|f(z)dz}.$$

Consequências:

- O uso da simulação direta de f só é a solução ótima no caso de h constante. Nos demais casos, sempre é preferível o uso de amostras simuladas da proposta g^*.
- Na prática é quase impossível obter g^*. Note que, se $h > 0$ então g^* depende da integral de interesse (desconhecida).
- Na busca de uma proposta próxima da ótima podemos considerar g tal que $\frac{|h|f}{g}$ é quase constante.

Amostragem por Importância

- O estimador MC-IS não se comporta bem quando $Var[w(X)] = \infty$, em que $w(X) = \frac{f(X)}{g(X)}$ é o peso.
- Não é necessário conhecer completamente as f.d.p, basta saber seu núcleo:

$$T_{IS}(X) = \frac{\sum_{i=1}^{R} h(x_i)w(x_i)}{\sum_{i=1}^{m} w(x_i)}$$

em que os pesos são $w(x_i) = \frac{f^*(x_i)}{g^*(x_i)}$. f^* e g^* são os núcleos das densidades f e g, respectivamente.

Observações sobre o comportamento dos pesos $w(x) = \frac{f(x)}{g(x)}$

Como x_i são simulados da proposta g para a maioria dos valores simulados $f(x_i) < g(x_i)$ e portanto $w(x_i) < 1$. No entanto,

$$E_g[W(X)] = \int \frac{f(x)}{g(x)} g(x)dx = \int f(x)dx = 1.$$

Logo, alguns valores de x_i devem produzir pesos muito altos (maiores que 1). Para manter a estabilidade do produto $h(x)w(x)$ devemos ter valores pequenos de $h(x)$ nesses pontos.

O método de amostragem por importância tem um bom comportamento quando deseja-se estimar pequenos valores de probabilidades.

Exemplo 1: Estimando a área da cauda da distribuição $N(0,1)$.

$$\theta = P(Z > a) = \int_a^{\infty} \phi(x)dx$$

em que $\phi(x)$ é a f.d.p. da $N(0,1)$.

Estimador ingênuo: $I = 1$ se $Z > a$ e $I = 0$ caso contrário.
Simular $z_1, \ldots, z_R$ da $N(0,1)$ e obter a proporção de $I = 1$.

Se a for um valor alto isso pode ser pouco eficiente pois a probabilidade de se observar $Z > a$ será muito baixa.

Amostragem por Importância

Estimador *IS* : Considera como proposta $g(x) = e^{-(x-a)} \quad x > a$.

Neste caso os pesos são $w(x) = \phi(x)e^{(x-a)}$

Simular $y_1, \ldots, y_R$ da $Exp(1)$ e fazer $x_j = y_j + a$.
O estimativa é obtida por:

$$\frac{1}{R}\sum_{j=1}^{R} w(x_j)$$

Exemplo 2: X tem distribuição t-Student com ν graus de liberdades. A integral de interesse é

$$E\left[\sqrt{\left|\frac{X}{1-X}\right|}\right] = \int_{-\infty}^{\infty} \sqrt{\left|\frac{x}{1-x}\right|} \frac{1}{K}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2} dx.$$

Vamos considerar aqui o uso do *IS*, com duas alternativas de propostas:

(a) Cauchy. Neste caso os pesos, sem considerar as constantes de padronização, são dados por:

$$w(x) = \frac{\left(1+x^2\right)}{\left(1+\frac{x^2}{\nu}\right)^{(\nu+1)/2}}$$

(b) $N(0, \nu/(\nu-2))$. Neste caso os pesos, sem considerar as constantes de padronização, são dados por:

$$w(x) = \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} e^{\frac{(\nu-2)}{2\nu}x^2}.$$

- A distribuição de Cauchy tem caudas mais pesadas que a t-Student e a Normal tem caudas mais leve.
- Podemos mostrar que $Var[W(X)]$ é infinita no caso da proposta normal e finita no caso da proposta Cauchy.
- Espera-se um comportamento melhor do uso do *MC-IS* no caso (a).

Tarefa: Pesquisar sobre densidade *tilted* e com isso pode ser usada com *MC-IS* para estimar

$$\theta = P(S \geq a)$$

em que $S = \sum_{i=1}^{n} X_i$ com X_i v.a. independentes com f.d.p. f_i .