Núcleo Estimador da Função de Distribuição com Incorporação de Restrições de Suporte

Gregorio Saravia Atuncar
Lupércio França Bessegato

Departamento de Estatística
Universidade Federal de Minas Gerais
30123-970 Belo Horizonte - MG
e-mails: gregorio@est.ufmg.br,
lupercio@est.ufmg.br

Resumo

Seja F_n^* o núcleo estimador usual da função de distribuição de uma variável aleatória, baseado em uma amostra $X_1, \cdots, X_n$. Em muitas aplicações práticas, sabe-se que $X \geq c$ e/ou $X \leq d$, para constantes c e d dadas ou a serem estimadas. Este trabalho estuda modificações de F_n^* que denominaremos "Método da Reflexão Invertida", que incorpora estas informações adicionais no estimador F_n, o qual possui suporte $[c, \infty]$. Este estimador é interpretado de uma maneira que permite o uso das conhecidas propriedades de convergência do núcleo estimador no estudo do comportamento de F_n. Apresentamos também resultados de simulações de grande porte, avaliando o desempenho do estimador, além de uma comparação com a estimação de função de distribuição através da integral do núcleo estimador da função densidade. Este trabalho é baseado em Schuster (1985).

Palavras-chave: Núcleo-estimador, Método da Reflexão Invertida, Problema de Fronteira, Função Característica.

1 Introdução

Dada uma amostra aleatória $X_1, \cdots, X_n$, de uma variável aleatória contínua X, com função de distribuição F, define-se o estimador de F, avaliado no ponto x, por:

$$F_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \tag{1}$$

onde K é uma função de distribuição, sendo denominado núcleo e h_n é chamado de parâmetro de suavidade.

Assumiremos que a função de densidade $k = K'$ é limitada, simétrica, continuamente diferenciável, tem suporte compacto e $0 < \int t^2 k(t)dt = k_2 < \infty$. Assumiremos também que $h_n \to 0$ e $nh_n \to \infty$, quando $n \to \infty$. A taxa de

convergência e a suavidade do núcleo estimador dependem da escolha de uma largura de janela. A partir daqui, para simplificação, escreveremos h em lugar de h_n e, quando não houver indicação dos limites de integração, assume-se que a integral é sobre toda a reta.

Em muitas aplicações práticas, sabe-se que $X \geq c$ e/ou $X \leq d$, dadas as constantes c e d. Adicionalmente, pode-se conhecer os valores de $f(c)$ e $f(d)$. Nestes casos é importante modificar o estimador $F_n^*(x)$ de $F(x)$ para incorporar esta informação adicional. Neste trabalho, estudamos modifcações que denominaremos "imagem invertida" de F_n^* que incorpora a informação adicional em nosso estimador F_n. Assim, podemos interpretar F_n de maneira a podermos usar as mais conhecidas propriedades de convergência dos núcleos estimadores.

2 Restrições no Suporte de F

O problema básico com o núcleo estimador usual é que dada a definição de F_n^* em (1), ela pode assumir valores positivos, mesmo para $x < c$, embora $F(x) = 0$ para $x < c$. Se X_i é próximo a c, então parte da contribuição de X_i para $F_n^*(x)$, dada por $K_i(x) = \frac{1}{n}K\left(\frac{x - X_i}{h}\right)$, avança sobre o intervalo $(-\infty, c)$ (fig. 1). O estimador proposto F_n incorpora a massa de probabilidade alocada em $(-\infty, c)$ de volta para $[c, \infty)$, subtraindo o da "imagem refletida",dado por $\frac{1}{n}K\left(\frac{2c - x - X_i}{h}\right)$ para $K_i(x)$. Para $x \geq c$ o efeito é subtrair o termo "imagem refletida" $F_n^*(2c - x)$ de $F_n^*(x)$ para produzir o estimador $F_n(x)$.

Motivados pela abordagem dada por Schuster (1985) [5] no tratamento de problemas de fronteira na estimação da função de densidade, estamos propondo o núcleo estimador corrigido, dado pela expressão abaixo, que aloca a probabilidade distribuída para os valores de $x < c$ de volta para o intervalo $[c, \infty)$ que é o suporte de F e que supomos conhecido. Assumiremos que $X \geq c$, onde c é conhecida.

$$\begin{aligned} F_n(x) &= F_n^*(x - c) - F_n^*(2c - x), se\ x \geq c \\ &= 0, se\ x < c \end{aligned} \tag{2}$$

Sem perda de generalidade, verifica-se para $c = 0$ que o estimador proposto é funcção de distribuição. Assim, dado que $F_n(x) = F_n^*(x) + F_n^*(-x)$, temos que:

1. $\lim_{x \to 0} F_n(x) = 0$
2. $\lim_{x \to \infty} F_n(x) = \lim_{x \to \infty} F_n^*(x) - \lim_{x \to \infty} F_n^*(-x) = 1$
3. Se $x < y$, então:

$$\begin{aligned} F_n(x) &= F_n^*(x) - F_n^*(-x) \\ &\leq F_n^*(y) - F_n^*(-x) \\ &\leq F_n^*(y) - F_n^*(-y) \end{aligned}$$

atendendo portanto aos requisitos de uma função de distribuição com suporte em$[0, \infty)$.

A partir das evidências empíricas alentadoras oferecidas pelas simulações efetuadas, estamos estudando as características assintóticas deste estimador, além de seu desempenho em situações em que possamos nos defrontar com o problema de fronteira na estimação da função de distribuição.

3 Escolha da Janela Ótima

A escolha do núcleo K não é muito crucial, mas a escolha do parâmetro de suavidade é um sério problema que tem sido tratado exaustivamente na literatura. Em geral, h é escolhido de maneira que $\hat{F}_n(x)$ seja um ótimo estimador de F, de acordo com alguma medida de desempenho, sendo comum o uso do erro quadrático médio integrado, MISE (Mean Integrated Squared Error) que é definido como:

$$MISE(h) = E \int \left\{ \hat{F}_n(x) - F(x) \right\}^2 \tag{3}$$

Está disponível há algum tempo uma expressão para a janela que minimiza o $MISE(h)$, verificando-se que este valor ótimo, h_{opt} infelizmente depende da função desconhecida F. Precisamos então estimar h_{opt} a partir dos dados observados. De Bowman et al (1998) [3], obtemos a expressão da janela ótima:

$$h_{opt} = \left\{ \frac{\int W(x)[1 - W(x)]\, dx}{\left[\int z^2 dW(z)\right]^2 \int \left[F''(x)\right]^2 \, dx} \right\}^{1/3} n^{-1/3} \tag{4}$$

A literatura aborda de várias maneiras a escolha da janela ótima h_{opt}, havendo uma grande demanda por procedimentos automáticos para seleção da janela.

Uma abordagem possível na escolha da janela ótima é através da utilização de método "plug-in", que estima o valor da única quantidade desconhecida na expressão do erro quadrático médio integrado assintótico, ou seja, a parcela dependente da função que se quer estimar ($\int [F'']^2$, no caso da estimação da função de distribuição). Salienta-se que o método "plug-in" tem a aparente vantagem de, em seu cálculo, não necessitar de uma rotina de otimização.

Chiu [4] prestou importante colaboração ao propor estimadores "plug-in" ajustados, baseados em funções características. Utilizamos o estimador da função de distribuição análogo ao estimador proposto por Chiu baseado na estimação de $H = \int \left[F''(x)\right]^2 dx$, que é a única quantidade desconhecida na expressão de h_{opt} (4). Esse valor é estimando usando a função característica empírica da amostra. Em Atuncar, Bessegato e Duczmal [1], verifica -se que H pode ser aproximada por:

$$\hat{H} = \frac{1}{\pi} \int_0^{\Lambda} \lambda^2 \left[|\hat{\varphi}(\lambda)|^2 - \frac{1}{n} \right] d\lambda \tag{5}$$

onde $\Lambda = min \left\{ \lambda : |\hat{\varphi}(\lambda)|^2 \leq \frac{C}{n} \right\}$, para algum $C > 1$.

Substituímos $\hat{H}$ em lugar de H na expressão (4) e obtemos um estimador $\hat{h}_{opt}$, para h_{opt}.

4 Simulação

Foram realizadas simulações de grande porte. Os resultados mostram que a correção melhora significativamente o estimador. Resultados típicos dessas simulações são mostrados nas figuras (1), (2). A figura (1) mostra os gráficos de F, F_n e F_n^* com tamanho amostral $n = 100$. Pode-se observar a qualidade dos estimadores. A figura (2) mostra os gráfico com $n = 50$. Pode-se observar nesse gráfico que F_n tem um desempenho muito fraco e a correção melhora significativamente o desempenho do estimador.

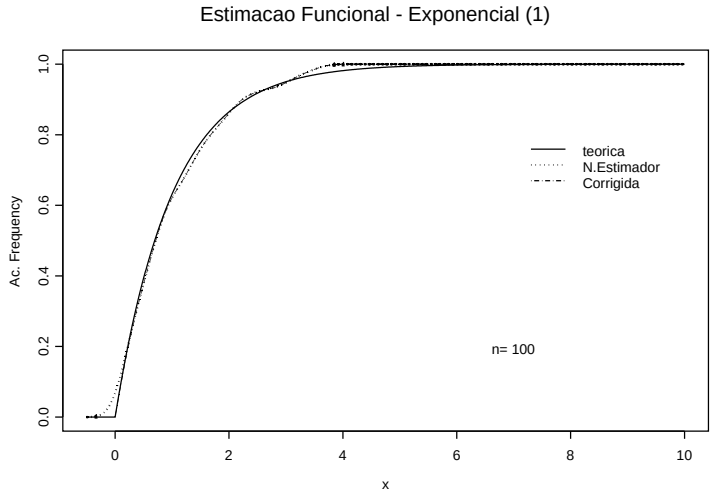

Figura 1: Exponencial $\lambda = 1$, amostra n=100.

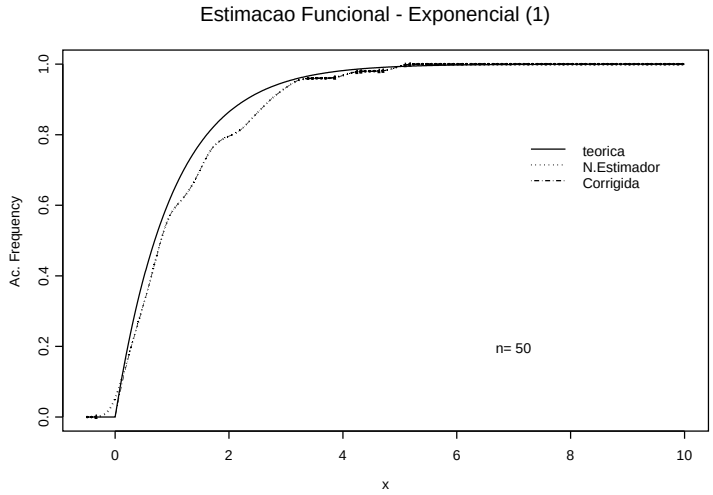

Figura 2: Exponencial $\lambda = 1$, amostra n=50.

Referências

[1] ATUNCAR, G. S.; BESSEGATO, L. F.; DUCZMAL, L. H. *A consistent estimator for the optimal bandwidth: the distribution function case.* Artigo submetido.

[2] ATUNCAR, G. S.; TRAVASSOS, A. P. *Boundary problems of the kernel estimator and their approach.* Em preparação.

[3] BOWMAN, A. W.; HALL, P.; PRVAN, T. *Bandwidth selection for the smoothing of distribution functions.* Biometrika, 85, pag. 799-808, 1998.

[4] CHIU, S. T. *Bandwidth selection for kernel density estimation.* The Annals of Statistics, 33, pag. 1883-1905, 1991.

[5] SCHUSTER, E. F. *Incorporating support constraints into nonparametric estimators of densities.* Commun. Statist.-Theor.Meth., 14(5), pag. 1123-1136, 1985.