16º SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA

AVALIAÇÃO DO MÉTODO DE VALIDAÇÃO CRUZADA PARA ESTIMAR A JANELA ÓTIMA – DADOS COM CENSURA

Gregório S. Atuncar
Lupércio F. Bessegato
Rudger M. Chagas

Departamento de Estatística - UFMG

RESUMO

Quando se tem dados completos (não censurados) o método de validação cruzada estabilizada é eficiente na escolha da janela ótima para a estimação da função de densidade usando núcleos estimadores. Porém na pratica muitas vezes temos dados censurados, pois problemas de tempo e recursos impossibilitam a utilização de dados não censurados. Marron e Padgett desenvolveram o método de validação cruzada para escolha da janela ótima para dados com censura. No presente trabalho, foram desenvolvidos métodos computacionais para avaliação do método de Marron e Padgett.

INTRODUÇÃO

Núcleo estimadores de uma função de densidade desconhecida de dados censurados foram estudados recentemente por muitos autores. Esses estimadores são mostradas na próxima seção.

Da mesma forma que em amostras completas, a escolha da janela ótima é crucial para que haja um bom desempenho do núcleo estimador. Quando a janela ótima é muito pequena existe muita variância no ajuste de um conjunto de dados e quando ela é muito grande há um grande vício.

No caso de amostras completas foi mostrado que o erro médio quadrado integrado (MISE) pode ser assintóticamente decomposto em um termo de variância, um de vicio quadrático e alguns outros termos sem importância para escolha da janela ótima. Esta decomposição também pode ser usada para o caso de amostras censuradas. Uma aproximação é dada para o núcleo estimador normal.

Porém esta representação de MISE não é muito utilizada para encontrar a janela ótima porque contém termos que são mais difíceis de estimar do que a própria função de densidade. Como esse problema também ocorre no caso de

amostras não censuradas, foi considerada uma outra maneira de selecionar a janela ótima: O método de validação-cruzada estabilizado.

Para avaliar o método de validação-cruzada, foi implementado um programa em linguagem C, que calcula qual é a janela ótima que minimiza a função de validação cruzada.

OS ESTIMADORES E A SELEÇÃO DA JANELA ÓTIMA

No caso de dados censurados é muito utilizado o estimador da função de distribuição definido a seguir:

Sejam $X_1^0, X_2^0, ..., X_n^0$ tempos de sobrevivência (independentes e identicamente distribuídos). Essas observações são censurados por variáveis aleatórias i.i.d $U_1,...,U_n$ que são independentes dos X^0_i's. Denote a distribuição dos X^0_i's por F^0 e dos U_i's por H. Seja $H^*=1-H$. Assume-se que F^0 é absolutamente continua com densidade f^0 e que H é contínua.

Os valores censurados observados são denotados por pares (X_i, Δ_i), $i= 1,...,n,$ onde

$X_i=\min\{X^0_i, U_i\}$ e $\Delta_i = 1_{[X0i \leq Ui]}$.

Baseado em (X_i, Δ_i), $i= 1,...,n,$ um estimador para a função de sobrevivência $1-F_0(t)$ é o estimador do produto-limite (PL), proposto por Kaplan e Meier. Seja (Z_i, Λ_i), $i= 1,...,n,$ os X^0_i's ordenados com seus correspondentes Δ_i's. O estimador PL de $1-F_0(t)$ é definido por

$$P_n(t) = \begin{cases} 1, se 0 \leq t \leq Z_1 \\ \prod_{i=1}^{k-1} \left(\frac{n-i}{n-i+1} \right)^{\Lambda i}, se Z_{k-1} \leq t \leq Z_k, k = 2,...,n \\ 0, se t > Z_n \end{cases}$$

e seja s_j o "salto" da função de sobrevivência em Z_j, que é

$$s_j = \begin{cases} 1 - Pn(Z_2), \ j=1 \\ Pn(Z_J) - Pn(Z_{j+1}), j = 2, ..., n-1 \\ Pn(Z_n), j = n. \end{cases}$$

Então para j<n, s_j=0 se e somente se Λ_i, isto é, Zj é uma observação censurada.

O estimador da função de densidade F_n, considerando o núcleo gaussiano, é definido por

$$f_n(x) = h^{-1} \sum_{j=1}^{n} s_j \left\{ \frac{e^{-\frac{1}{2}\left(\frac{x-Z_j}{n}\right)^2}}{\sqrt{2\pi}} \right\}$$

A função de distribuição das variáveis aleatórias de censura (U_i's) é definida por

$$H_n^*(t) \begin{cases} 1, 0 \le t \le Z_1 \\ \prod_{i=1}^{k-1} \left(\frac{n-i+1}{n-i+2} \right)^{1-\Lambda_i}, Z_{k-1} \le t \le Z_k, k = 2,...,n \\ \prod_{i=1}^{n} \left(\frac{n-i+1}{n-i+2} \right)^{1-\Lambda_i}, Z_n < t \end{cases}$$

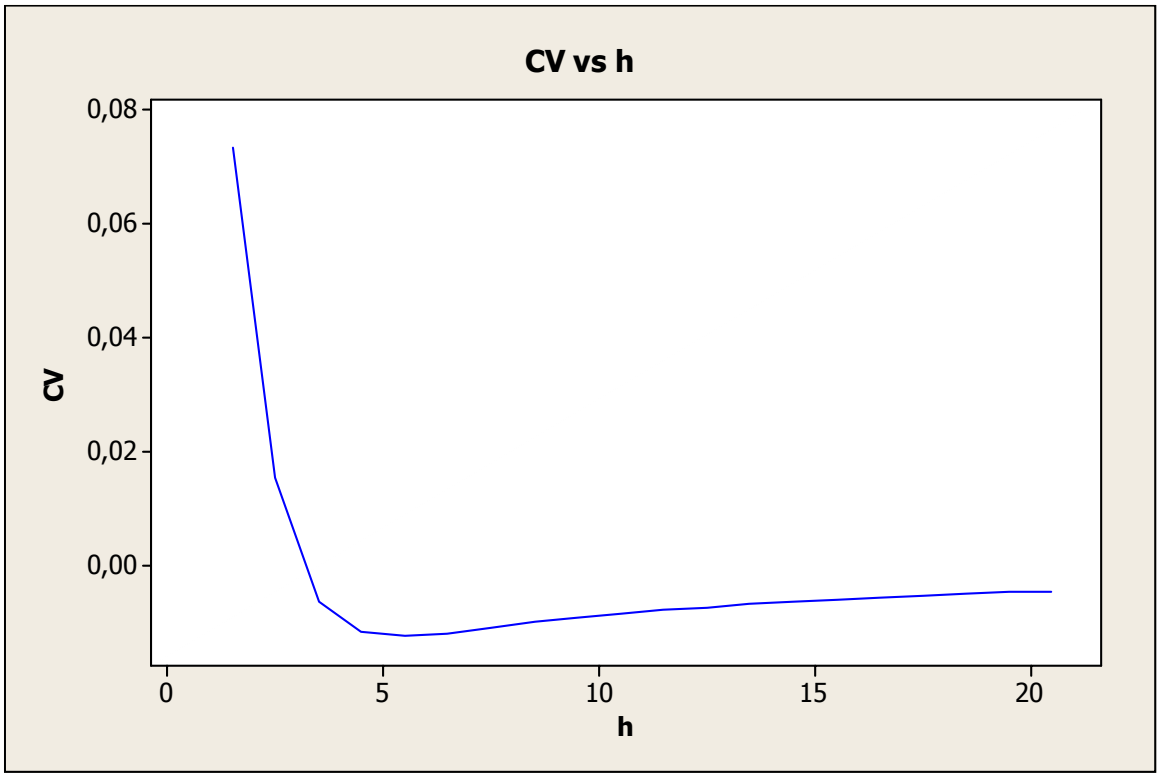
O critério de validação cruzada para encontrar a janela ótima consiste em minimizar a seguinte função:

$$CV(h) = \int [f(x)]^2 w(x) dx - 2n^{-1} \sum_{i=1}^{n} f_i(X_i) \frac{w(X_i)}{H_n^*(X_i)} 1_{[\Delta_i=1]},$$

onde

$$f_{n,i}^*(x) = \sum_{j \neq i} \frac{1}{(n-1) H_n^*(x) h} \left\{ \frac{e^{-\frac{1}{2}\left(\frac{x-Z_j}{n}\right)^2}}{\sqrt{2\pi}} \right\} 1_{[\Delta_i=1]}.$$

Simulações foram realizadas para avaliar o comportamento da função *CV(h)*. Um resultado típico dessas simulações é mostrado a seguir:



REFERÊNCIA BIBLIOGRÁFICA

Maron, J.S.; Padgett, W.J. Asymptotically optimal Bandwidth Selection For Kernel Density Estimator From Randomly Right-Censored Samples.