UMA IMPLEMENTAÇÃO DO MÉTODO ESTABILIZADO DE VALIDAÇÃO CRUZADA PARA A ESCOLHA DA JANELA ÓTIMA EM ESTIMAÇÃO FUNCIONAL

Gregorio Saravia Atuncar
Prof. Adjunto do Departamento de Estatística da UFMG
E-mail: gregorio@est.ufmg.br

Paula Jacqueline de Oliveira
Mestranda em Estatística pelo Departamento de Estatística da UFMG

Evander C. Damasceno
Mestrando em Estatística pelo Departamento de Estatística da UFMG

Frederico Rodrigues Borges da Cruz
Prof. Adjunto do Departamento de Estatística da UFMG

Resumo

A densidade invariante de processos estocásticos é um conceito fundamental em estatística. Freqüentemente, porém, não se dispõe de um modelo paramétrico adequado; recorrendo-se, por isso, aos estimadores funcionais não paramétricos. Estes, no entanto, são de implementação complexa, visto que utilizam algoritmos computacionalmente. Torna-se portanto essencial a utilização de uma linguagem poderosa de programação com a flexibilidade do C++. Este artigo disponibiliza uma biblioteca orientada por objetos para a implementação do método estabilizado de validação cruzada direcionado à escolha automática da janela ótima para a estimação da densidade invariante de processos estocásticos. O programa será utilizado como alternativa mais eficiente ao já desenvolvido anteriormente pelos autores em forma de macro no software estatístico MINITAB.

Palavras Chave: Programação Orientada por Objetos, Processos Estocásticos, Densidade Invariante.

Abstract

The invariant density of an stochastic process is very important in statistic. We can infer about it. The nonparametric estimator of the invariant density are a kind of complex to be implemented. In this paper we present the computational aspects of the stababilizes cross-validation method (SVC), proposed in the literature, to estimate the optimal bandwidth to be used in the definition of the kernel density estimator of the invariant density. We work using C++. The SVC method choose automatically na estimato of the optimal bandwidth. In this work we have the intentio of offer na alternative to a software, inside MINITAB, developed by the authors. That work is part of a major project in developement.

Key Words: Oriented for Objects Programing, Stochasthics Processes, Invariant Density.

1. Introdução

Os métodos não paramétricos aplicados a processos estocásticos tem assumido progressiva importância entre as metodologias estatísticas. Resultados assintoticamente ótimos tem sido demonstrados e novos estimadores e preditores propostos. Trata-se de uma alternativa, mais flexível, à abordagem clássica freqüentemente embasada em rígidas suposições pouco razoáveis em algumas situações práticas.

A utilização de algoritmos computacionalmente intensivos de implementação complexa, no entanto, afastou as primeiras iniciativas de incorporação dos métodos em um software estatístico de uso geral. Somente em 1990 com os sofwares STATXACT e DISFREE lançados respectivamente pela Cytel Software Corporation e Biosoft Cambridge alguns princípios básicos da metodologia foram incluídos a pacotes estatísticos para PC's.

Atualmente, com a disponibilização de linguagens poderosas e flexíveis de programação, como o C++, tornou-se possível uma avaliação empírica dos métodos até então dificultada pela escassez de recursos computacionais adequados. Além disso, o paradigma de programação orientada por objetos, definido por alguns como programação de tipos abstratos de dados (TAD) e suas relações, tem possibilitado a otimização do desempenho dos programas e sua conseqüente difusão.

Este artigo disponibiliza uma biblioteca orientada por objetos para a implementação do método estabilizado de validação cruzada direcionado à escolha automática do valor ótimo da janela em estimação funcional. Deve-se ressaltar que o mesmo é parte integrante de um estudo de maior amplitude, no qual os autores abordam o comportamento dos métodos de escolha da janela ótima, disponíveis na literatura, para processos auto-regressivos e variáveis aleatórias independentes e identicamente distribuídas.

2. Estimadores funcionais não-paramétricos

A função densidade de probabilidade (fdp) é um conceito fundamental em estatística. Seu conhecimento permite inferir acerca de inúmeras características da variável aleatória de interesse. Freqüentemente, porém, não se dispõe de um modelo paramétrico adequado. Recorre-se assim, aos estimadores funcionais não-paramétricos.

A partir destes estimadores pode-se obter a fdp de variáveis aleatórias independentes e a densidade invariante de processos estocásticos, ou seja, a distribuição de probabilidade, no limite, da variável X indexada pelo parâmetro de tempo t.

Existem, na literatura, inúmeros estimadores funcionais não paramétricos, muitos dos quais baseiam-se no conceito frequentista de probabilidade. Dentre eles destaca-se o estimador pelo método do núcleo definido a seguir.

2.1 Estimador funcional pelo método do núcleo

Suponha uma amostra aleatória X_1, X_2,..., X_n. Define-se o estimador pelo método do núcleo como:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right) \qquad (1)$$

onde *h* é denominado ***"amplitude da janela"*** ou ***"parâmetro de amortização"*** e *k(x)* é uma função de densidade, chamada núcleo.

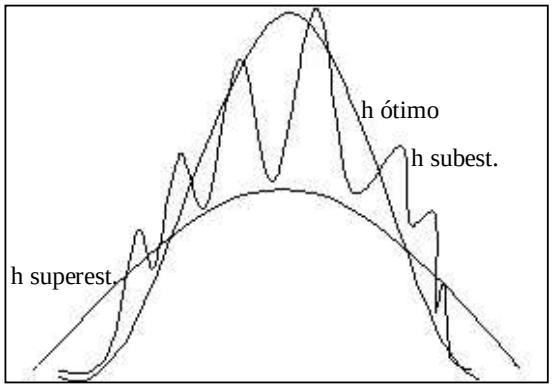
Nota: *Assume-se, neste trabalho, que k é simétrica e que:*

$$0 < \int_{-\infty}^{\infty} t^2 k(t)dt = k_2 < \infty \qquad (2)$$

Pode-se observar que a estimativa de *f(x)* dependerá tanto da função *núcleo, k(x)*, quanto da *amplitude da janela, h,* que determinam respectivamente a forma e a suavidade da curva estimada. A escolha apropriada do núcleo encontra-se bem estudada na literatura que aponta o normal e o de Epanechnikov como boas alternativas.

A escolha do valor ótimo para a amplitude da janela, tema deste artigo, é um ponto crucial para a obtenção de uma boa estimativa da fdp de interesse. Sua subestimação implica no aumento da velocidade de oscilação da curva estimada, enquanto que a superestimação pode obscurecer uma possível multimodalidade nos dados (figura 1).

Figura 1: Estimativas de densidade pelo método do núcleo



O critério para a escolha do valor ideal da janela está relacionado com a finalidade da estimativa da função densidade. Em algumas situações tem-se apenas o propósito exploratório dos dados no sentido de estabelecer hipóteses e suspeitas. Nestes casos, pode-se escolhe-lo intuitivamente através de gráficos para a estimativa de *f(x)*, obtidos por meio de vários valores de *h*. Este critério, porém, nem sempre é satisfatório, por ser extremamente subjetivo. Além disso, tem-se grande demanda por métodos automáticos de busca do valor de *h* que otimiza a estimativa da função densidade de probabilidade de interesse.

Dentre os métodos automáticos de escolha de *h*, propostos na literatura, destacam-se o plug-in originalmente proposto por Woodroofe (1970), o de validação cruzada proposto por Rudemo (1982) e Bowman (1984) e as respectivas modificações propostas por Chiu (1991). Em projetos desenvolvidos anteriormente pelos autores concluiu-se que os métodos modificados por Chiu são similares quando comparados entre si e mais eficazes que os originais, no que se refere à variância das estimativas obtidas. Sendo assim, optou-se pela implementação do método estabilizado de

validação cruzada que, em etapas posteriores, a partir da utilização dos recursos de herança e polimorfismos disponíveis no C++ suportará a implementação do método plug-in modificado.

2.2 Método Estabilizado de Validação Cruzada

Suponha um estimador $\hat{f}$ da função densidade de probabilidade *f* de uma variável aleatória. O erro quadrático integrado (EQI) é definido por:

$$EQI = \int (\hat{f} - f)^2 = \int \hat{f}^2 - 2\int \hat{f}f + \int f^2 \qquad (1)$$

Pode-se perceber que o último termo de (2) não depende de $\hat{f}$, então a idéia é se minimizar, sob h, a quantidade R($\hat{f}$) definida por:

$$R(\hat{f}) = \int \hat{f}^2 - 2\int \hat{f}f \qquad (2)$$

o que é o mesmo que minimizar o próprio *EQI*.

No entanto, pode-se perceber que R($\hat{f}$) ainda depende da fdp desconhecida *f(x)*. Sugere-se assim, que esta quantidade seja estimada pelos próprios dados.

Rudemo (1982) e Bowman (1984) propuseram o método de validação cruzada (MVC) para a estimação do valor ótimo de *h*. Segundo estes autores estima-se R($\hat{f}$) por:

$$M_o(h) = \int \hat{f} - \frac{2}{n}\sum_i \hat{f}_{-i}(X_i) \qquad (3)$$

Onde: $\hat{f}_{-i}(X_i) = \frac{1}{(n-1)h}\sum_{j\neq i} k\left(\frac{X_i - X_j}{h}\right)$

A quantidade $M_0(h)$, como se pode verificar, depende apenas dos dados, e além disso $E(M_0(h)) = E(R(\hat{f}))$ e essa é a idéia que suporta o método.

De modo a facilitar a utilização de recursos computacionais, segundo Silverman (1986), define-se $K^{(2)}$ como sendo a convolução da função núcleo com ela mesma. Assumindo que K é simétrica e realizando-se a mudança de variável u = h-1x tem-se os seguintes resultados:

$$\int \hat{f}(x)^2 dx = \frac{1}{n^{-2}h}\sum_i\sum_j K^{(2)}\left(\frac{X_i - X_j}{h}\right) \qquad (4)$$

$$\frac{1}{n}\sum \hat{f}_{-i}(X_i) = \frac{1}{n(n-1)}\sum_i\sum_j \frac{1}{h}K\left(\frac{X_i - X_j}{h}\right) - \frac{1}{(n-1)h}K(0) \qquad (5)$$

Substituindo-se (4) e (5) em (3) e substituindo-se $(n-1)^{-1}$ por n^{-1} obtém-se uma boa aproximação de $M_0(h)$, dada por:

$$M_1(h) = \frac{1}{n^2 h} \sum_i \sum_j K^* \left(\frac{X_i - X_j}{h} \right) + \frac{2}{nh} K(0) \qquad (6)$$

onde: $K^*(t) = K^{(2)}(t) - 2K(t)$

Sendo assim, o valor ótimo da janela é estimado pelo valor de *h* que minimiza $M_1(h)$.

As estimativas de h-ótimo obtidos pelo método original de validação cruzada, porém, apresentam uma grande variabilidade, como observado por Chiu (1991). No intuito de reduzir esta variabilidade o mesmo autor propõe o método estabilizado de validação cruzada. Este sugere a substituição de $M_1(h)$ pela quantidade Sn(h) definida por:

$$S_n(h) = \frac{\pi}{nh} \int_0^{\Lambda} K^2(x) + \int_0^{\Lambda} \left(|\tilde{\phi}(\lambda)|^2 - \frac{1}{n} \right) \left(W^2(\lambda h) - 2W(\lambda h) \right) d\lambda \qquad (7)$$

onde: $\tilde{\phi}(\lambda)$ é a função característica empírica definida por:

$$\tilde{\phi}(\lambda) = \frac{1}{n} \sum_{j=1}^{n} e^{i\lambda X_j} \qquad W(t) = \int e^{itX} K(x) dx \qquad (8)$$

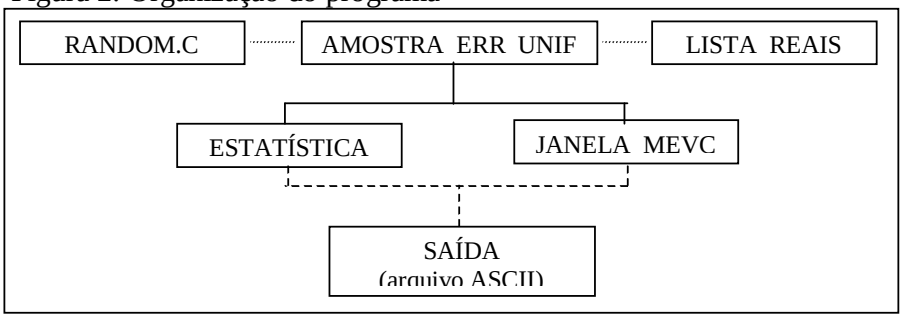
e Λ é o primeiro valor de λ para o qual $|\tilde{\phi}(\lambda)|^2 \leq \frac{c}{n}$ para $c \geq 1$

3. Estrutura do programa

O programa foi implementado de forma abrangente e interativa com o usuário, que somente necessitará de um conhecimento razoável da notação e do método abordado. O principal objetivo foi a disponibilização de recursos que permitissem a realização, em larga escala, de simulações do método estabilizado de validação cruzada com enfoque em processos auto-regressivos de primeira ordem cujos erros são uniformemente distribuídos.

Neste sentido e de forma a otimizar o desempenho do programa foi utilizada a técnica de programação por objetos na linguagem C++. Desta forma o programa resultante conta com 3 objetos que se comunicam por meio de recursos como herança e compartilhamento de membros privados (FRIEND) e um programa de execução (figura 2).

Figura 2: Organização do programa



3.1 Gerador de processos auto-regressivos

Um processo auto-regressivo de primeira ordem pode ser definido como um tipo de processo estocástico em que a variável aleatória X indexada pelo parâmetro de tempo t estabelece uma relação de dependência apenas com a observação imediatamente anterior indexada por (t-1). Assim, sua modelagem possui a seguinte forma:

$$X_t = \rho X_{t-1} + \varepsilon_t$$

onde: ρ é o valor da auto-correlação de primeira ordem.
ε_t é série de resíduos.

Para simular amostras deste tipo de processo estocástico foi utilizado o tipo abstrato de dados LISTA para inserção e armazenamento das observações calculadas em um vetor de tamanho pré-determinado pelo usuário. Para geração da série de resíduos foi utilizado um gerador de números aleatórios com distribuição uniforme implementado em C.

Ressalta-se que a cada execução, através de funções intrínsecas do C++, o objeto referente a amostra é substituído utilizando-se o recurso de alocação dinâmica de memória.

Figura 2: Objeto AmostraErrUnif – protótipo do gerador de amostras auto-regressivas

```
class AmostraErrUnif {
 public:
  AmostraErrUnif() {};
  ~AmostraErrUnif(void) {};
  ListaR GeraAmostra(int *semente, int tam, float rho, float Obs1);
};
```

3.2 Objeto Estatística

Este objeto foi desenvolvido para disponibilizar informações exploratórias dos dados, pertinentes ao bom desempenho do método abordado. Neste sentido, foram implementados estimadores de parâmetros de alocação, simetria e variância os quais serão compartilhados como os demais objetos.

Figura 3: Objeto Estatística – protótipo

```
class Estatistica {
 public:
  Estatistica () {};
  ~Estatistica (void) {};
  float Mean (ListaR *amostra);
  float Median (ListaR *amostra);
  float StDev (ListaR *amostra);
};
```

3.3 Objeto JanelaMEVC

A classe JanelaMEVC implementa o método estabilizado de validação cruzada para estimação da janela ótima. Esta implementação foi realizada de forma a possibilitar, via recursos de polimorfismo, a incorporação de outros métodos a serem posteriormente avaliados.

Foram utilizadas em seu desenvolvimento o tipo abstrato de dados LISTAS e funções virtuais, além de construtores e destrutores, para compor e manusear vetores com alocação dinâmica de memória.

Figura 4: Objeto JanelaMEVC - protótipo

```
class JanelaMEVC : public Estatistica {
 private:
  float lambdaMax;
  int tamAmostra;
  ListaR VetorLambda;
  ListaR VetorPhi2;
 public:
  JanelaMEVC() {};
  ~JanelaMEVC() {};
  void InicializaTamanho(ListaR *amostra);
  float CalculaPhi2(ListaR *amostra, float lambda);
  float CalculaSn(float h, float norma2);
  void CalculaLambdaMax(ListaR *amostra, float norma1);
  void MonteParLambdaPhi(ListaR *amostra, float norma2);
  float EstimaHOtimo(ListaR *amostra, float norma2);
};
```

As principais funções da classe Janela_MEVC são a CalculaLambdaMax e EstimaHOtimo. Ambas são descritas a seguir.

Função CalculaLambdaMax

No cálculo das expressões $S_n(h)$ é necessário encontrar o limite superior de integração (Λ). Para tal, fundamentado no decrescimento exponencial da curva, precisa-se buscar o primeiro valor de λ tal que $|\tilde{\phi}(\lambda)|^2 \leq c/n$ onde c > 1.

A rigor, os valores de λ podem variar de zero a infinito, porém, no sentido de viabilizar os cálculos, foi assumido que, como apontado por Chiu (1991), pag 1900, para k1>5 e $\delta > 0$:

$$P(\Lambda \leq n^{1/k_1+\delta}) \underset{n \to \infty}{\to} 1$$

A partir deste resultado pôde-se dizer que, com alta probabilidade, Λ pertencerá ao intervalo $(0, n^{1/k_1+\delta}]$.

Função EstimaHOtimo

Esta função implementa o algoritmo de *busca recursiva* para a minimização da expressão Sn(h). Ressalta-se que o referido algoritmo é menos eficiente para o caso geral, porém, em se tratando de funções convexas com apenas um ponto de mínimo possui desempenho satisfatório.

3.4 Programa principal

Figura 5: Arquivo de execução – membros principais

```
#include <stdio.h>
#include "listar.cpp"
#include "hotimo.cpp"

int main(void){
 ...

 Amostra = Ar.GeraAmostra(&semente, n, 0.6, 1.0);
 H.InicializaTamanho(&Amostra);

 relogio.Start();
 H.CalculaLambdaMax(&Amostra, norma1);
 H.MonteParLambdaPhi(&Amostra, norma2);
 hOtimo = H.EstimaHOtimo(&Amostra, norma2);
 relogio.PrintElapsedTime();
....

 return (0);
}
```

3.5 Saída

A saída é apresentada em um arquivo ASCII no qual as linhas representam as amostras geradas e as colunas descrevem informações pertinentes à avaliação do método estabilizado de validação cruzada, no que se refere ao vício e variância do estimador proposto. Além disso, pela incorporação da função relógio no programa, encontra-se disponível, na saída, o tempo de processamento referente à aplicação do método em cada uma das amostras geradas.

Exemplo: Considere um exemplo no qual o usuário requisitasse a simulação de 10 amostras, cada qual com 200 observações de um processo auto-regressivo de primeira ordem com valor de auto-correlação 0.8. Então obteria um arquivo de saída com a seguinte estrutura:

TAMANHO	ORDEM	LAMBDAMAX	HOTIMO
150	1	1.37208	0.580
150	2	1.27415	0.680
150	3	1.47002	0.610
150	4	1.17622	0.510
150	5	1.47002	0.575
150	6	1.27415	0.645
150	7	1.56795	0.650
150	8	1.37208	0.550
150	9	1.27415	0.525
150	10	1.56795	0.535

4. Uma aplicação dos estimadores funcionais não paramétricos

Uma empresa de prestação de serviços, utiliza uma medida padrão denominada "Unidade de Trabalho (UT)" para o pagamento dos honorários de seus sócios. O valor da UT varia mensalmente, de acordo com a receita e despesas mensais da empresa, da seguinte forma:

$$UT = \frac{\text{Receita mesal - Despesas mensais}}{\#UT''s \text{ processadas no mes}}$$

Como se pode notar, o valor da UT é representado em moeda vigente e estabelece relação crescente e decrescente respectivamente com a receita e despesas mensais da empresa. No entanto, esta oscilação deve ser controlada no sentido de não comprometer o bom funcionamento da empresa. A meta traçada é se manter o valor de R$0,27 por UT o que, em alguns momentos, pode ser inviável no mercado.

O objetivo aqui almejado avaliar a probabilidade, a partir das 124 observações disponíveis, de que, no limite, o valor da UT estivesse incluído no intervalo (0,20; 0,27). Este intervalo foi estabelecido de modo a satisfazer mutuamente a empresa e a conveniência do mercado.

Após uma análise inicial da série de UT's, pôde-se ajusta-la a um modelo auto-regressivo de primeira ordem (ρ = 0,6582), obtendo como resíduos uma série de ruídos brancos. Sendo assim, espera-se o bom funcionamento dos métodos de estimação funcional, para o ajuste da densidade e consequentemente para a avaliação da referenciada probabilidade.

H-ótimo

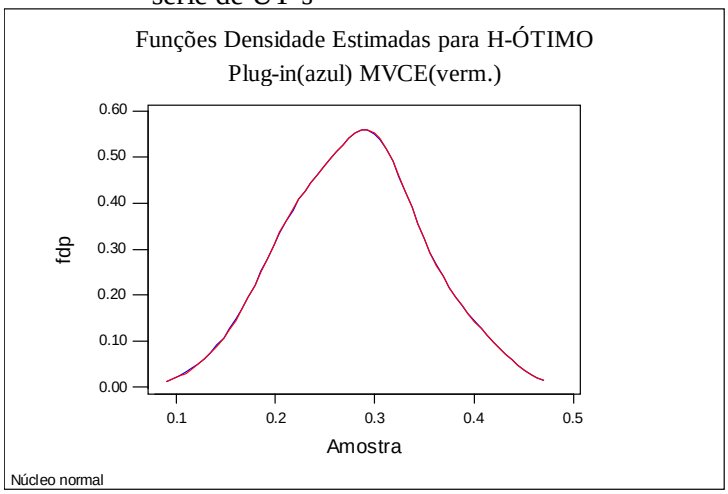
A partir dos métodos abordados e após uma transformação linear nos dados (Y=12X), foram obtidos as seguintes estimativas para o valor ótimo do parâmetro de suavização:

Tabela 5: Valores estimados para a amplitude da janela

Método de Estimação	H estimado
Plug-in Modificado	0,365177
Estabilizado de Valicação Cruzada	0,360000

Pode-se observar que os valores estimados a partir dos dois métodos estão bastante próximos, com os mesmos foi estimada a função de densidade para a série de UT's apresentada abaixo.

Figura 4 : Função densidade invariante estimada para a série de UT's



Como os métodos forneceram estimativas bastante próximas para o valor ótimo de h, as curvas estimadas tornaram-se sobrepostas.

Probabilidades de interesse

A partir da estimação do valor de h, foi possível, através de uma aproximação do caso contínuo pelo discreto obter a probabilidade de que, no limite, o valor da Unidade de trabalho esteja no intervalo de (0,20; 0,27).

Tabela 4: Probabilidades estimadas

Método de Estimação	**P(0.2 < UT < 0.27)**
Plug-in	0.6572
Estabilizado de Validação Cruzada	0.6426

Serão apresentados a seguir alguns dos resultados das simulações obtidos, inclusive a formatação, da saída da macro. Deve-se ainda frisar que, apesar de terem sido tomados vários valores de auto-correlação serão mostrados apenas as saídas para $\rho = 0$ e $\rho = 0.6$ para o caso em que $\varepsilon \sim N(0,1)$.

6. Considerações finais

Este artigo é parte integrante do projeto *"Escolha da Janela Ótima em Estimação Funcional: Caso Markoviano"*, de mesma autoria, cujo objetivo principal é avaliar empiricamente o comportamento de alguns dos métodos automáticos de escolha do valor ótimo de *h* na estimação da densidade invariante de processos markovianos.

Neste são contemplados processos auto-regressivos de primeira ordem, cujos erros seguem uma distribuição normal, ou uniforme, ou exponencial dupla.

Além disso, deve-se salientar que o programa resultante deste artigo disponibiliza uma biblioteca orientada por objetos para a realização de simulações de processos auto-regressivos de primeira ordem e não estando previstos outros tipos de processos que somente serão incorporados em etapas futuras.

7. Referências

Atuncar, G.S., Oliveira, P.J. (1998) Escolha da janela ótima em estimação funcional: caso markoviano. Relatório técnico, Departamento de Estatística UFMG.

Atuncar, G.S., Damasceno, E.C. E Mendonça, P.A. (1997) *Escolha da janela ótima em estimação funcional.* Relatório técnico, Departamento de Estatística UFMG.

Bosq,D. (1998) Nonparametric Statistics for Stochastic Processes, Springer - Verlag, New York.

Chiu, S.T. (1991) *Bandwidth selection for Kernel density estimation*. The Annals of Statistics, **33**, 1883-1905.

Hart, J. D. (1984) *Efficiency of a Kernel Density Estimator Under an Autoregressive Dependente Model*. Journal of the American Statistical Associatior, 79, 110-117.

Hart, J.D. E Vieu, P. (1990) *Data-driven Bandwidth Choice for Density Estimation Based on Dependence Data*. The Annals of Statistics, 18, 873-890.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analisys***.** Chapman adn Hall London.