IX Encontro de Modelagem Computacional
Belo Horizonte,MG
15 a 17 de novembro de 2006
CEFET-MG - IPRJ/UERJ
ISBN: 978-85-99836-02-6

ROTINAS EM R PARA NÚCLEO ESTIMADORES

Lupércio França Bessegato- lupercio@est.ufmg.br
Gregorio Saravia Atuncar- gregorio@est.ufmg.br
Luiz Henrique Duczmal- duczmal@est.ufmg.br
Universidade Federal de Minas Gerais, Departamento de Estaística
av. Antonio Carlos, 6627. - 31270-901 - Belo Horizonte, MG, Brasil

Resumo. *O método de suavização por núcleo estimadores tem sido amplamente utilizado em estimações funcionais em suas várias aplicações. A eficiência da metodologia é bastante sensível ao valor adotado para o parâmetro de suavização. Dentre as vária metodologias disponíveis para sua escolha, destacam-se aquelas que utilizam o método 'plug-in' baseado na função característica empírica. Neste trabalho detalhamos uma biblioteca de funções com a metodologia citada, desenvolvida para uso no pacote estatístico livre R. A título de exemplo, apresentamos uma aplicação da metodologia na análise de uma situação real.*

Keywords: *Núcleo estimador, escolha da janela ótima, método 'plug-in', função caraccterística, processo de Poisson*

1. INTRODUCÃO

Ao fazermos inferência de um modelo específico, é possível obter um ganho muito grande em eficiência, mas somente se o modelo assumido for pelo menos aproximadamente verdadeiro. Se o modelo assumido não estiver correto, as inferências podem ser piores e inúteis, levando a enganos grosseiros na interpretação dos dados. Assim, infelizmente, a força da modelagem paramétrica é também sua fraqueza.

Os métodos de suavização oferecem uma ponte entre não estabelecer nenhuma hipótese na estrutura formal (abordagem puramente não-paramétrica) e estabelecer hipóteses muito fortes (abordagem paramétrica). A adoção de uma hipótese relativamente fraca de que a verdadeira forma funcional é suave, permite a extração de mais informações dos dados do que seria possível por uma abordagem puramente não paramétrica, assim como o abandono de hipóteses paramétricas rígidas, fornecendo, em conseqüência análises ao mesmo tempo flexíveis e robustas. Assim, os métodos de suavização oferecem com eficiência uma maneira de ressaltar importantes estruturas subjacentes aos dados.

A suavização por núcleo estimador é um método bastante difundido na estimação da função densidade de probabilidade, da função de distribuição de probabilidade subajacente a um conjunto de dados observados, além de ser uma metodologia aplicada na estimação da função de intensidade de um processo de Poisson não-homogêneo e de uma função de regressão.

Uma questão crucial na aplicação desta metodologia é a determinação da janela h, que controla o grau de suavização dos dados. Se h é muito pequeno, admite-se demasiado ruído amostral e se h é muito grande, perdem-se características da curva devido à supersuavização. A taxa de convergência e a suavidade do estimador dependem da escolha da largura desta janela, tornando-se de extrema relevância estudar estimadores de janela ótima para a obtenção da estimação mais apropriada da função que exprima a probabilidade de ocorrência dos dados.

A literatura aborda de várias maneiras a escolha da janela ótima h_{op}. Embora, na prática, seja possível escolher o parâmetro de suavização de maneira subjetiva, há uma grande demanda por procedimentos automáticos para seleção da janela. O seletor automático mais estudado é o da função escore de validação cruzada de mínimos quadrados. Embora o minimizador da função escore de validação cruzada seja uma uma estimativa consistente da janela ótima e possua normalidade assintótica, verifica-se que as estimativas da janela proveniente dos procedimentos de validação cruzada apresentam uma grande variabilidade, impactando assim a estimativa funcional desejada. Estudos de simulação indicaram que o seletor tende a escolher valores de janela menores, com mais freqüência que o predito pelos teoremas assintóticos.

Outra abordagem possível na escolha da janela ótima é através do método 'plug-in' que estima o valor da única quantidade desconhecida na expressão que define o valor ótimo de h, ou seja, a parcela dependente da função que se quer estimar ($\int [f'']^2$, no caso da estimação da função de densidade ou $\int [F'']^2$ no caso da estimação da função de distribuição). Salienta-se que o método 'plug-in', quando aplicável, tem a vantagem de, em seu cálculo, não necessitar de uma rotina de otimização. Os estimadores 'plug-in' utilizados no presente trabalho baseiam-se em funções características, cujo comportamento foi estudado por Damasceno (2000), no caso da estimação da função de densidade e por Bessegato (2001), no caso da estimação da função de distribuição.

O objetivo principal deste trabalho é apresentar e detalhar um conjunto de funções desenvolvidas em R que são úteis na estimação funcional de um conjunto isolado de dados observados. Apresenta-se uma aplicação do método.

2. ESTIMAÇÃO FUNCIONAL

Inicialmente, nosso problema de interesse será estimar a função de densidade de probabilidade f ou a função de distribuição F de uma variável aleatória contínua, a partir de uma amostra aleatória $X_1, \cdots, X_n$.

O estimador da função de densidade mais utilizado é o histograma. Nesse gráfico a área da barra é equivalente à proporção de observações no intervalo ao qual pertencem. A amplitude desses intervalos basicamente controla a suavidade do procedimento. Este estimador tem duas importantes limitações: a dependência do comprimento do intervalo e o fato de não constituir uma curva contínua.

Essa última limitação do histograma incentivou a procura de estimadores contínuos. Assim, utilizando a idéia de probabilidade freqüentista, estabeleceu-se o estimador $\hat{f}$ da densidade f de uma variável aleatória contínua X, cuja expressão é dada por:

$$\hat{f}(x) = \frac{1}{2nh} \left[\# \; de \; X_i's \; \in (x-h; x+h) \right] \tag{1}$$

escolhendo-se um número h pequeno. Esta função é chamada de estimador natural.

Para superar algumas das limitações deste estimador, em particular o fato de $\hat{f}$ não ser contínua, generaliza-se sua expressão, tomando uma função núcleo k satisfazendo a condição de que $\int_{-\infty}^{\infty} k(x)dx = 1$. Esta função tem a finalidade de ponderar a distância entre os dados observados e o ponto em que se deseja estimar a densidade. Em geral, assume-se que k é uma função de densidade simétrica. Assim, o núcleo estimador com núcleo k é dado por:

$$\hat{f}(x) = \frac{1}{nh} \sum_{j=1}^{n} k\left(\frac{x - X_j}{h} \right) \tag{2}$$

Pode-se verificar que quando o parâmetro de suavidade h for muito pequeno, o resultado da estimativa da função de densidade tende a produzir estruturas falsas que apresentam cur-

vas muito irregulares. Já quando h é escolhido grande o resultado da estimativa da função de densidade tende a super-suavizar f.

Da própria definição do núcleo-estimador, seguem algumas propriedades elementares, dentres as quais a mais importante talvez seja a de que $\hat{f}$ herdará todas as propriedades de continuidade e diferenciabilidade do núcleo k. Assim, se k for uma função de densidade normal, então $\hat{f}$ será uma curva suave tendo derivadas de todas as ordens.

O método do núcleo estimador também é utilizado na estimação da função de distribuição. De maneira similar ao caso da função de densidade, define-se o núcleo estimador de F, avaliado no ponto x, por:

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{3}$$

onde o núcleo K é uma função de distribuição de probabilidade. De maneira análoga ao caso da estimação da função de densidade, a taxa de convergência e a suavidade do núcleo estimador dependem da escolha de uma largura de janela. A partir daqui, para simplificação,quando não houver indicação dos limites de integração, assume-se que a integral é sobre toda a reta.

No caso em que os pontos $X_1, \cdots, X_n$ são uma realização no intervalo $[0, T]$ de um processo de Poisson não homogêneo, o núcleo estimador de sua função de intensidade λ avaliado no ponto t é dado por:

$$\hat{\lambda}(t) = \frac{1}{h}\sum_{i=1}^{n} k\left(\frac{t - X_i}{h}\right) \tag{4}$$

A única diferença em relação ao estimador definido em Eq. (2) é a ausência do fator n^{-1}, já que essa função pode ser interpretada como o número de observações em um determinado intervalo, enquanto a densidade é uma proporção de eventos.

Quando o suporte da função a ser estimada é limitado, o núcleo estimador tem um desempenho insatisfatório nas fronteiras do intervalo. Embora se saiba que o núcleo estimador seja viciado, nas proximidades das fronteiras do intervalo o vício do estimador é ainda maior que em seu interior, com tendência à subestimação.

Para correção deste problema já foram propostas várias soluções. Devido a sua grande simplicidade de implementação e a sua natureza intuitiva, o método da reflexão se destaca dos demais e baseia-se na idéia de transformar os dados de modo que, na escala transformada, a função a ser estimada tenha suporte ilimitado. Isto é feito refletindo-se os pontos em torno do ponto de fronteira.

Implementamos a rotina para o caso em $X \geq 0$. Neste caso, o núcleo estimador corrigido da função de densidade é dado por:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left[k\left(\frac{x - X_i}{h}\right) + k\left(\frac{-x - X_i}{h}\right)\right]. \tag{5}$$

A janela h é selecionada através do método 'plug-in' modificado, já adotado nas demais estimações funcionais. Em todos as situações de estimação relacionadas anteriormente, assume-se que a função de densidade $k = K'$ é limitada, simétrica, continuamente diferenciável, com suporte compacto e $0 < \int t^2 k(t)dt = k_2 < \infty$. Assume-se também que $h_n \to 0$ e $nh_n \to \infty, \text{quando } n \to \infty$.

3. ESCOLHA DA JANELA ÓTIMA

A escolha do núcleo não é muito crucial, mas a escolha da janela é um sério problema que tem sido tratado exaustivamente na literatura. Em geral, h é escolhido de maneira que o núcleo estimador seja um estimador ótimo da função de interesse, de acordo a alguma medida de desempenho. No caso da função de densidade é comum avaliar o desempenho do núcleo estimador $\hat{f}$ através do Erro Quadrático Integrado (ISE-Integrated Squared Error), dado por:

$$ISE(h) = \int \left\{ \hat{f}_h(x) - f(x) \right\}^2 dx \tag{6}$$

e seu valor esperado, o Erro Quadrático Médio Integrado (MISE-Mean Squared Error), dado por:

$$MISE(h) = E\left[\int \left\{ \hat{f}_h(x) - f(x) \right\}^2 dx \right] = \int Var\hat{f}(x)dx + \int vicio^2\hat{f}(x)dx \tag{7}$$

sendo que, a igualdade final pode ser obtida através de propriedades elementares de média e variância.

A partir de uma análise assintótica do $MISE(h)$ pode-se verificar que pequenos valores de h aumentam a variância assintótica e, desta maneira, a estimativa $\hat{f}(x)$ resultante terá uma aparência muito irregular. Por outro lado, como o vício na estimação de $f(x)$ depende diretamente da amplitude da janela h, valores grandes de h reduzem a variância assintótica de $\hat{f}(x)$ mas aumenta o vício assintótico. Salienta-se que o vício na estimação de $f(x)$ não depende diretamente do tamanho da amostra, mas da amplitude da janela h.

A partir da expressão do 'MISE' assintótico, o valor ótimo de h é dado por:

$$h_{op} = k_2^{-2/5} \left[\int k(t)^2 dt \right]^{1/5} \left[\int f''(x)^2 dx \right]^{-1/5} n^{-1/5} \tag{8}$$

onde k_2 é uma constante que depende do núcleo utilizado. Da Eq. (8), percebe-se que a janela não é disponível na prática, pois, ela depende da função de densidade desconhecida f.

Há vários métodos propostos para estimar a janela ótima a partir de uma amostra aleatória $X_1, \cdots, X_n$.

Uma aproximação comum na seleção automática da janela é obter uma estimativa do $MISE(h)$ e através de sua minimização estima-se h_{op} . O Método da Validação Cruzada por Mínimos Quadrados, proposto por Rudemo (1982) e Bowman (1984) é o mais antigo e o mais estudado dos métodos de escolha da janela. Entretanto, verifica-se que este método sofre muito com a variação amostral, isto é, para diferentes amostras da mesma distribuição, as janelas estimadas apresentam uma grande variabilidade. Uma outra desvantagem do método é que a função a ser minimizada apresenta freqüentemente diversos mínimos, com alguns deles espúrios, situados na região de sub-suavização. Por outro lado, algumas vezes ocorre o problema de não existir mínimo.

Uma outra 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 na Eq. (7), ou seja, $\int (\hat{f}'')^2$, parcela dependente da função de densidade que se quer estimar. Esta estimativa será utilizada na Eq. (8) para se obter h_{op}. Salienta-se que o método 'plug-in' tem a aparente vantagem de, em seu cálculo, não necessitar de uma rotina de otimização.

Este trabalho foca a estimação da janela ótima de núcleo estimadores de diversos tipos de funções utilizando-se de funções caracteristicas amostrais. Chiu (1991) prestou importante colaboração ao propor inicialmente estimadores 'plug-in' ajustados, para a função de densidade, baseados em funções características para encontrar uma expressão equivalente à quantidade desconhecida $G = \int (f''(x))^2$ na Eq. (8).

Pela Fórmula da Inversão e Identidade de Parseval, prova-se que:

$$G = \int [f''(x)]^2 \, dx = \frac{1}{2\pi} \int_0^\infty \lambda^4 \, |\varphi(\lambda)|^2 \, d\lambda \tag{9}$$

sendo que $\varphi(\lambda)$ é a função característica da densidade, definida por:

$$\varphi(\lambda) = \int e^{i\lambda x} f(x) dx \tag{10}$$

A função característica amostral é definida por:

$$\hat{\varphi}(\lambda) = \frac{1}{n} \sum_{j=1}^{n} e^{i\lambda X_j} \tag{11}$$

A partir desta definição, temos que:

$$|\hat{\varphi}(\lambda)|^2 = \left[\frac{\sum_j cos\,(\lambda X_j)}{n} \right]^2 + \left[\frac{\sum_j sen\,(\lambda X_j)}{n} \right]^2 \tag{12}$$

Chiu (1991), buscando uma melhor regra de seleção da janela, propôs modificar a função característica amostral abaixo de alguma freqüência de corte λ. Seu estimador de G é dado por:

$$\hat{G} = \frac{1}{\pi} \int_0^\Lambda \lambda^4 \left[|\hat{\varphi}(\lambda)|^2 - \frac{1}{n} \right] \, d\lambda. \tag{13}$$

Primeiramente, encontramos Λ que é o primeiro valor de λ tal que $|\hat{\varphi}(\lambda)|^2 < c/n$ para $c > 1$. Essa constante c impacta no resultado mesmo quando f é suficientemente suave, verificando-se que para $c = 3$ a variância do estimador é a menor. Prova-se ainda que $\hat{G}$ é um estimador com boas propriedades assintóticas.

Usando estes resultados, a partir da Eq. (8), estabelece- se um estimador para a janela ótima dado por:

$$\hat{h}_{op} = k_2^{-2/5} \left[\int k(t)^2 dt \right]^{1/5} \left[\hat{G} \right]^{-1/5} n^{-1/5}. \tag{14}$$

O método oferece uma seleção completamente automática, com baixo esforço computacional em comparação ao método de validação cruzada, já que não necessita uma rotina de otimização.

No caso do núcleo estimador da função de distribuição, dado pela Eq. (3), h é escolhido de maneira que $\hat{F}(x)$ seja um ótimo estimador de F de acordo com alguma medida de desempenho, sendo comum o uso do MISE definido como:

$$MISE(h) = E \int \left\{ \hat{F}_n(x) - F(x) \right\}^2 \tag{15}$$

Está disponível há algum tempo uma expressão para a janela que minimiza o $MISE(h)$, verificando-se que este valor ótimo, h_{op}, como no caso da densidade, depende infelizmente da função desconhecida F. Precisamos então estimar h_{op} a partir dos dados observados. De Bowman et al. (1998), obtemos a expressão da janela ótima dada por:

$$h_{op} = \left\{ \frac{\int K(x)[1-K(x)]\,dx}{\left[\int z^2 dK(z)\right]^2 \int \left[F''(x)\right]^2\,dx} \right\}^{1/3} n^{-1/3} \tag{16}$$

Utilizando um estimador análogo àquele da densidade, Bessegato (2001), propôs um estimador para $H = \int \left[F''(x)\right]^2 dx$, que é a única quantidade desconhecida na expressão de h_{op} em Eq. (16). Este estimador, é dado por:

$$\hat{H} = \frac{1}{\pi} \int_0^{\Lambda} \lambda^2 \left[\left|\hat{\varphi}(\lambda)\right|^2 - \frac{1}{n} \right] d\lambda \tag{17}$$

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

Substituindo $\hat{H}$ em lugar de H na Eq. (16) e obtem-se um estimador $\hat{h}_{op}$, para h_{op}, ou seja:

$$\hat{h}_{op} = \left\{ \frac{\int K(x)[1-K(x)]\,dx}{\left[\int z^2 dK(z)\right]^2 \hat{H}} \right\}^{1/3} n^{-1/3} \tag{18}$$

A consistência forte de $\hat{H}$ e de $\hat{h}_{op}$ encontram-se provadas em Bessegato (2001).

Dada a equivalência entre os métodos estabilizado de validação cruzada e 'plug-in'modificado, métodos muito difundidos e com implementação computacional conhecida, Travassos (2003) verificou seu bom desempenho no contexto da estimação da função de intensidade.

No caso do estimador de densidade com correção de fronteira (Eq. 5), a janela h é selecionada através do método 'plug-in' modificado, da mesma maneira que nas demais estimações funcionais já mencionadas neste artigo.

4. FUNÇÕES EM R

O objetivo do conjunto de funções apresentado neste trabalho é facilitar a estimação funcional de interesse referente a um único conjunto de dados. A biblioteca utiliza o pacote estatístico R (R Development Core Team, 2006) e pode ser aplicada na estimação da função de densidade, função de distribuição e função intensidade de processos, além da correção de fronteira naqueles casos apontados anteriormente neste trabalho.

Em todas estas situações, a escolha da janela ótima dá-se através do método 'plug-in', utilizando a função característica empírica. Além da estimação funcional propriamente dita pode-se obter os resultados de cada etapa do procedimento, ou seja, a determinação do limite de integração Λ, a estimativa da parcela desconhecida (G, no caso da densidade e H, no caso da distribuição), o cálculo de h_{op} e o gráfico da estimativa funcional. Dependendo das condições dos dados observados, há uma rotina que transforma os dados, facilitando a implementação da metodologia. A Tabela 1 apresenta um resumo das funções disponíveis.

O método de Simpson foi utilizado nas integrações numéricas que estão presentes na maioria das etapas do processo de integração. Simulações intensivas mostraram-no adequado ao procedimento. Dedicamos cuidados especiais com o truncamento da integral, já que em grande

Tabela 1: Resumo das Funções das Principais da Biblioteca

Comando	Argumentos	Descrição
alg.lambda	(vetor.dados, cota.sup=10, cota.inf=0)	Calcula Λ
alg.G	(vetor.dados, lambda)	Calcula $\hat{G}$
alg.H	(vetor.dados, lambda)	Calcula $\hat{H}$
alg.hop.f	(vetor.dados,$\hat{G}$,flag)	Calcula h_{op} densidade
alg.hop	(vetor.dados,$\hat{H}$,flag)	Calcula h_{op} distribuição
plug.smooth	(vetor.dados,rotulo=NULL)	Estimação da densidade
plug.smooth.F	(vetor.dados,rotulo=NULL)	Estimação da distribuição
plug.trans	(vetor.dados,rotulo=NULL)	Estimação da densidade com transformação
plug.trans.F	(vetor.dados,rotulo=NULL	Estimação da distribuição com transformação
plug.reflex	(vetor.dados,rotulo=NULL)	Estimação da densidade com correção efeito de fronteira
plug.intensidade	(vetor.dados,rotulo=NULL)	Estimação intensidade do processo

parte dos cálculos, pelo menos um dos limites de integração tende ao infinito de maneira que as rotinas utilizem intervalos de integração que assegurem a precisão dos resultados.

Há algumas bibliotecas relacionadas com núcleo estimador e disponíveis em R. Destacamos a biblioteca sm, de autoria de Bowman & Azzalini (2005) desenvolvida para GNU R e a biblioteca KernSmooth, de Wand (2005) Ambos os pacotes enfatizam a estimação da densidade, dando-se a escolha da janela através de variações do método de validação cruzada.

A biblioteca e as instruções detalhadas para sua utilização estão disponíveis no endereço ftp://ftp.est.ufmg.br/pub/nucleo ou diretamente com o primeiro autor [1].

5. APLICAÇÃO

A metodologia foi aplicada na estimação da função de intensidade de acidentes de trabalho em uma empresa industrial, no período de janeiro de 1998 a dezembro de 2001. A empresa possui vários setores de negócios em uma mesma unidade fabril. Como os processos de trabalho são distintos, buscou-se explorar e analisar suas características também através de técnicas de suavização. Os dados encontram-se na página mencionada acima.

A unidade em questão possui uma fábrica de material elétrico de grande porte com até três turnos de trabalho diários, dependendo do nível de produção necessário ao atendimento das encomendas. Na mesma unidade encontram-se outros setores de sua divisão de prestação de serviços.

Consideramos os acidentes de trabalho como um processo pontual, em que $N(t)$ é uma variável aleatória de contagem que denota a quantidade de acidentes de trabalho no intervalo de tempo $[0, t]$. A função média deste processo é dada por $\Lambda(t) = E[N(t)]$, sendo nosso interesse estudá-la através de sua função de intensidade, definida como:

$$\lambda(t) = \frac{d}{dt}\Lambda(t). \tag{19}$$

[1]Autor correspondente: Lupercio F. Bessegato. E-mail: lupercio@est.ufmg.br

Nossa análise supôs a existência de um processo de Poisson não- homogêneo, com 222 ocorrências registradas. Na escolha do parâmetro de suavização, utilizamos a metodologia que efetua a correção dos efeitos de fronteira e para garantir a eficiência da análise, efetuamos uma transformação dos dados do tipo $Y/\sigma(X)$.

Além do processo global dos acidentes, analisamos e comparamos vários dos processos relacionados, sendo que, por suas características, mostrou-se mais significativa a estratificação por tipo de afastamento, em que acidente com afastamento significa que houve dias de trabalho perdidos após a ocorrência do acidente de trabalho. Outros tipos de estratificação mostraram-se pouco informativos.

O processo global de acidentes foi observado entre 1/1/1998 e 31/12/2001. Da função de intensidade estimada (Fig. 1), percebemos um período de relativa estabilidade até setembro de 2000, com uma taxa de acidentes crescentes a partir de então, indicando uma deterioração do processo de acidentes. Ao aprofundarmos em nossa análise, pudemos perceber a estabilidade do processo de acidentes de trabalho com afastamento e a deterioração, a partir de setembro de 2000, do processo de acidentes sem afastamento. Estas situações podem ser verificadas pelas respectivas estimativas da função de intensidade, superpostas na Fig. 2.

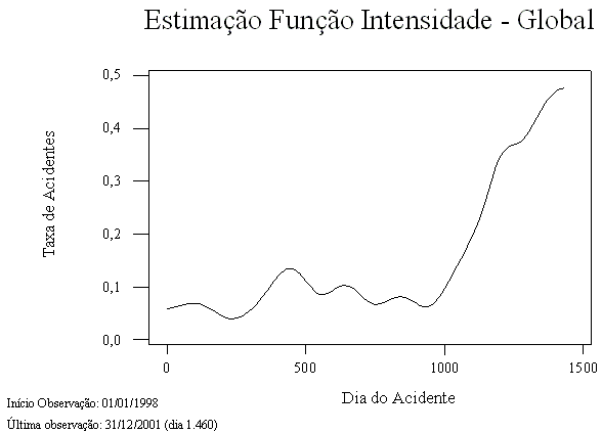

Figura 1: Estimação da intensidade do processo global de acidentes.

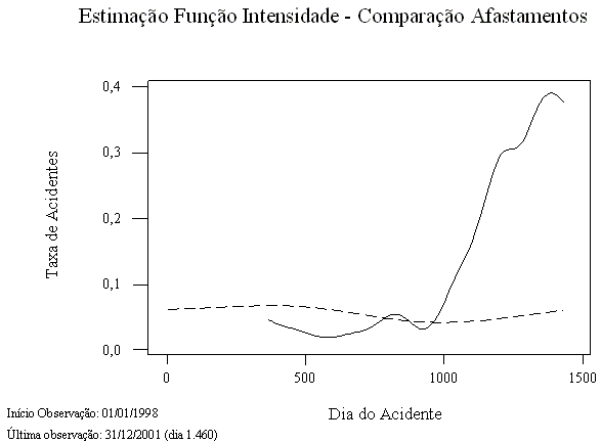

Figura 2: Comparação da intensidade dos processos por afastamento.

Fica claro nesta aplicação a importância da suavização na análise exploratória, proporcionando uma visualização do comportamento da taxa de acidentes ao longo do tempo que permitiu a compreensão do histórico do processo, indicando os eventos que interferiram em sua evolução. Além disso auxiliou grandemente na verificação da consistência dos dados e na indicação dos próximos passos na modelagem deste processo. Aparentemente houve sub-notificação dos acidentes sem afastamento em parte do período estudado com a importante evidência empírica de que não há relação causal entre os dois processos.

Agradecimento

O primeiro autor agradece o apoio financeiro parcial da Fapemig.

REFERÊNCIAS

Atuncar, G. S., Bessegato, L. F., & Duczmal, L. H., 2003. A consistent estimator for the optimal bandwidth: the distribution function case. Relatório Técnico RTP-06/2003, UFMG/Dep. de Estatística.

Bessegato, L. F., 2001. *Escolha do parâmetro de suavidade na estimativa da função de distribuição*. Dissertação de Mestrado, UFMG/Departamento de Estatística, Belo Horizonte, Minas Gerais, Brasil.

Bowman, A. W., 1984. An alternative method of cross validation for the smoothing of density estimates. *Biometrika*, vol. , n. 71, pp. 353–360.

Bowman, A. W. & Azzalini, A., 2005. *sm: Smoothing methods for nonparametric regression and density estimation*. R package version 2.1-0.

Bowman, A. W., Hall, P., & Prvan, T., 1998. Bandwidth selection for the smoothing of distribution functions. *Biometrika*, vol. , n. 85, pp. 799–808.

Chiu, S. T., 1991. Bandwidth selection for kernel density estimation. *Annals of Statistics*, vol. , n. 33, pp. 1883–1905.

Damasceno, E. C., 2000. *Escolha do parâmetro de suavidade em estimação funcional*. Dissertação de Mestrado, UFMG/Departamento de Estatística, Belo Horizonte, Minas Gerais, Brasil.

R Development Core Team, 2006. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rudemo, M., 1982. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, vol. , n. 9, pp. 65–78.

Travassos, A. P. A., 2003. *Problemas de fronteira dos núcleos estimadores e suas abordagens*. Dissertação de Mestrado, UFMG/Departamento de Estatística, Belo Horizonte, Minas Gerais, Brasil.

Wand, M., 2005. *KernSmooth: Functions for kernel smoothing for Wand & Jones (1995)*. R package version 2.22-15.

AN R LIBRARY FOR KERNEL ESTIMATION

Abstract

The kernel method for functional estimatin has been largely used to estimate functions such as the density function, distribution function and intensity function of a process. The efficiency of the method is very sensitive to the value adopted for the bandwidth parameter. Among the best methods availablefor the bandwidth choice, we have the plug-in method based on the empirical characteristic function. In this work we present a library of functions with the aforementioned method, developed to be used in the free R statistical package. For illustration purposes, we present an application of the method in the analysis of a real situation.

Keywords

Kernel estimator, bandwidth choice; plug-in method; characteristic function; Poisson process.