

AVALIANDO O DESEMPENHO DO NÚCLEO ESTIMADOR NA ESTIMAÇÃO DA FUNÇÃO DE DENSIDADE NORMAL BIVARIADA

Cléia do Nascimento Cavalcante

Sueli Aparecida Mingoti

Gregório Saravia Atuncar

Frederico R. B. Cruz

Departamento de Estatística - ICEx - UFMG

31270-901 - Belo Horizonte - MG

{cnc,sueli,gregorio,fcruz}@est.ufmg.br

Resumo

Neste artigo apresenta-se o método não-paramétrico de núcleo estimador para a estimação de funções de densidade multivariadas. A abordagem utilizada é a da transformação inversa proposta por Fukunaga (1972). Este procedimento tem a vantagem de fazer com que os parâmetros de suavização (ou janelas ótimas) possam ser estimados para cada variável individualmente quando a transformação é escolhida de modo que as novas variáveis sejam não correlacionados. Assim, os métodos automáticos de seleção de janelas amplamente discutidos na literatura podem ser utilizados para o caso multivariado. Neste artigo, apresentamos um estudo de simulação com a distribuição normal bivariada mostrando que o estimador de núcleo pode ser uma boa alternativa. O método torna-se mais importante ainda em situações em que não temos clareza de qual é a forma matemática da função densidade que está gerando os dados observados. Temos realizado simulações em algumas situações não normais e os resultados são surpreendentes. O método empregado para a seleção da janela ótima é o plug-in modificado proposto por Chiu (1991).

Palavras Chaves: Função de densidade multivariada; Estimação não-paramétrica; Núcleo estimadores multivariados, Método plug-in.

Abstract

In this paper an evaluation of the kernel method to estimate a multivariate density is presented. The considered approach is the inverse transformation proposed by Fukunaga (1972). Under this approach univariate techniques to find the smoothing parameter h for each variable for the univariate case can be used because the original data is transformed in such a way that the transformed data is non correlated. So, the density of the transformed random vector can be estimated as a product of univariate density estimators. For estimating the univariate densities the plug-in modified method proposed by Chiu (1991) is used. The estimator of the density of original random variables is obtained by back transformation. Some simulation results by using the bivariate normal distribution are presented. The simulations show that the kernel estimator had a good performance even for moderate sample sizes. The method become more important in situations on which there is not clear information about the mathematical form of the joint density. Data were simulated for some non normal situations and the results were promising.

Keywords: Multivariate density function; Non-parametric estimates; Multivariate kernel estimators; Plug-in method.

1. Introdução

A necessidade de estimação da função de densidade aparece em várias áreas de aplicação. Como exemplo, podemos citar seu uso em análise discriminante (Johnson & Wichern, 2002; Cooley & Maceachern, 1998). Este artigo tem como foco apresentar o método não-paramétrico de núcleo estimadores (Silverman, 1986; Hand 1982) para a estimação de funções de densidades multivariadas.

Uma comparação deste método com os métodos de estimação paramétricos dos momentos e de máxima verossimilhança será efetuada para a distribuição normal bivariada buscando conhecer as vantagens e desvantagens do método de estimação não-paramétrico. A abordagem paramétrica tem como característica a admissão de que as observações provêm de um vetor aleatório, cuja função densidade é conhecida, exceto os parâmetros envolvidos. Para situações nas quais não se tem conhecimento sobre a distribuição do vetor aleatório de interesse, são necessários métodos de estimação que não dependam da forma particular da função de densidade de probabilidade. Nestes casos, utiliza-se a abordagem não-paramétrica. Em ambos os campos quando se têm fenômenos a serem estudados a partir de dados contendo muitas variáveis, os métodos estatísticos delineados para obter informações a partir destes conjuntos de dados, são denominados métodos de estatística multivariada.

Dois procedimentos são citados na literatura para a estimação de densidades multivariadas via núcleos estimadores: o produto de núcleos normais independentes (Scott, 1992), e a estimação da densidade através de transformação inversa (Fukunaga, 1972). Em ambos os procedimentos, a estimação de função densidade multivariada depende da escolha do parâmetro de suavidade, ou seja, da busca da janela ótima h_i para cada variável e esta busca é feita dentro da abordagem já existente na literatura no caso univariado, uma vez que a finalidade da transformação é tornar as novas variáveis não-correlacionadas, de modo que a estimação da janela h_i possa ser feita para cada variável individualmente. Desta forma, a proposta para este artigo é utilizar a estimação da função de densidade multivariada via transformação inversa.

Para a estimação do parâmetro de suavidade ou janela h_i são abordados na literatura vários métodos de escolha automática. Dentre eles, o método plug-in modificado (Chiu, 1991) e o método estabilizado de validação cruzada proposto por Chiu (1991) como forma alternativa ao método de validação cruzada proposto por Rudemo (1982) e Bowman (1984). Neste artigo será abordado somente o método plug-in modificado. A função núcleo escolhida é o núcleo normal. Neste artigo será feita também uma avaliação da possibilidade de se trabalhar com janelas do mesmo valor h para todas as variáveis ao invés de janelas com valores diferentes.

O artigo está organizado da seguinte forma. Na seção 2, apresentamos a estimação da função de densidade multivariada pelo método do núcleo. Na seção 3, discutimos o problema da escolha da janela ótima para estimação. Resultados são apresentados na seção 4. Conclusões e observações finais encerram o artigo.

2. Estimação da Função de Densidade Multivariada pelo Método do Núcleo

A estimação da função de densidade multivariada via núcleos estimadores é uma generalização da estimação da densidade univariada pelo método do núcleo. Há dois procedimentos de estimação multivariada citados na literatura: (i) produtos de núcleos normais independentes (Scott, 1992) e (ii) estimação da densidade por transformação inversa apresentada por Fukunaga (1972) e descrita em Silverman (1986).

Neste estudo optamos em apresentar o segundo procedimento proposto por Fukunaga (1972), onde a estimação de função de densidade multivariada é feita por transformação inversa com a finalidade de utilizar abordagem já existente na literatura relacionado a estimação de parâmetro de suavidade para o caso univariado, visto que a janela h pode ser estimada individualmente para cada variável transformada quando estas são não-correlacionadas.

Para estimar a função de densidade multivariada através da transformação inversa, considere X um vetor aleatório definido por, $X_i^T = [X_1, X_2, \dots, X_p]$ de uma distribuição multivariada, com vetor de médias $\mu^T = [\mu_1, \mu_2, \dots, \mu_p]$ e matriz de covariância $\Sigma_{p \times p}$, com função de densidade $f = (x_1, x_2, \dots, x_p)$. O objetivo é estimar esta função $f(\cdot)$, através de transformação inversa. A estratégia utilizada é descrita a seguir. Primeiramente faz-se uma transformação linear do vetor X de modo que as novas variáveis sejam não-correlacionadas, visando assim encontrar as janelas dentro do contexto de estimação univariada. Seja a transformação $Y = C_{p \times p}^{-1} X_{p \times p}$, onde C é uma matriz não-singular. Uma escolha para a matriz C é, $C = \Sigma^{-1/2}$. Sua obtenção pode ser feita a partir da decomposição espectral da matriz Σ (Johnson e Wichern, 2002). A seguir, usa-se os dados transformados Y para estimar os valores das

janelas h_i que seriam apropriados para a estimação da função de densidade de Y . Desta forma, o valor da função de densidade de Y , no ponto $y=(y_1, y_2, \dots, y_n)$, estimado usando o método do núcleo normal, considerando $h_i=h$ é dado por:

$$\hat{g}(y) = \frac{1}{(2\pi)^{p/2} nh^p} \sum_{i=1}^n k \left\{ -\frac{1}{2nh^2} (y - Y_i)^T (y - Y_i) \right\},$$

onde $k(\cdot)$ é a função núcleo normal. Após encontrar as estimativas de h_i , estima-se a função de densidade de X , através da transformação inversa, isto é:

$$\hat{f}(x) = g(CX) |C|,$$

onde g é a função de densidade de Y e $|C|$ denota o determinante da matriz C . Usando este procedimento, e estimando a matriz de covariâncias Σ_{pxp} pela matriz de covariâncias amostral S_{pxp} , o núcleo estimador da função de densidade multivariada do vetor X avaliada em $x=(x_1, x_2, \dots, x_p)$, considerando $h_i = h$, $i=1, 2, \dots, p$, é dado por:

$$\hat{f}(x) = \frac{|S|^{-1/2}}{(2\pi)^{p/2} nh^p} \sum_{i=1}^n \exp \left\{ -\frac{1}{2nh^2} (x - X_i)^T S^{-1} (x - X_i) \right\}.$$

Quando supusermos os parâmetros de suavidade h_i diferentes para cada variável i , teremos:

$$\hat{f}(x) = \frac{|S|^{-1/2}}{(2\pi)^{p/2} n \prod_{i=1}^n h_i} \sum_{i=1}^n \exp \left\{ -\frac{1}{2} \left[\left(\frac{x - X_i}{h_i} \right)^T S^{-1} \left(\frac{x - X_i}{h_i} \right) \right] \right\}.$$

3. Escolha da Janela Ótima

Para se estimar a função de densidade multivariada pelo método do núcleo estimador, necessita-se encontrar os valores de h_i , $i=1, 2, \dots, p$, que otimizem a estimativa da função de densidade. Com isso, tem-se uma grande procura pela implementação de métodos automáticos para a escolha de h_i . A busca da janela ótima h_i para a estimação da função densidade multivariada pelo método do núcleo é feita pelo método univariado existente na literatura, uma vez que o propósito da transformação de Y é tornar as novas variáveis não-correlacionadas, de modo que a otimização da janela possa ser feita para cada variável individualmente.

Dentre os vários métodos de seleção de h abordados na literatura, encontram-se o método plug-in modificado e o método estabilizado de validação cruzada para variáveis aleatórias univariadas (Chiu, 1991). Neste artigo utilizamos apenas o método plug-in modificado o qual será brevemente descrito na seção a seguir.

3.1. Medidas de Discrepância

Estudos têm sido feitos relacionados à escolha da janela ótima para apenas uma variável X com função de densidade $f(\cdot)$, tendo como o objetivo encontrar uma medida de discrepância mínima e que mantenha um equilíbrio entre a variância e o vício. Uma medida global de discrepância muito usada é o erro quadrático médio integrado (EQMI), e pode ser rescrita em função do vício quadrático integrado e da variância integrada definida por:

$$EQMI = \frac{1}{4} h^4 k_2^2 \int (f''(x))^2 dx + \frac{1}{nh} \int (k(t))^2 dt + o \left(h^4 + \frac{1}{nh} \right)$$

A partir dessa relação, o valor de h que minimiza o EQMI é dado por:

$$h_{\text{ótimo}} = k_2^{-2/5} \left(\int k(t) dt \right)^{1/5} \left(\int (f''(x))^2 dx \right)^{-1/5} n^{-1/5},$$

onde $k_2 = \int t^2 k(t) dt$. A idéia do método plug-in modificado é aproximar o único termo desconhecido $G = \int (f''(x))^2$ por:

$$\hat{G} = \frac{1}{\pi} \int_0^\Lambda \lambda^4 \left(|\tilde{\phi}(\lambda)|^2 - \frac{1}{n} \right) d\lambda,$$

onde $\tilde{\phi}(\cdot)$ é a função característica de X . Desta forma o estimador pelo método plug-in modificado da janela ótima é definido como:

$$h_p = \left[\hat{\theta}_p + \frac{R'_n(\hat{\theta}_p)}{\tilde{A}_n(\hat{\theta}_p)} \right] n^{-1/5}$$

onde:

$$\hat{\theta}_p = \left\{ \int k^2(x) dx \right\}^{1/5} \left\{ \int x^2 k(x) dx \right\}^{-2/5} \hat{G}^{-1/5},$$

$$\tilde{R}_n(h) = (24\pi)^{-1} n^{-2/5} h^6 \int_0^\Lambda \lambda^6 \left\{ |\tilde{\phi}(\lambda)|^2 - \frac{1}{n} \right\} d\lambda \int x^2 k(x) dx \int x^4 k(x) dx \text{ e}$$

$$\tilde{A}_n(h) = (4)^{-1} h^4 \hat{G} \left\{ \int x^2 k(x) dx \right\}^2 + h^{-1} \int k^2(x) dx - \tilde{R}_n(h).$$

Maiores detalhes sobre esse método podem ser visto em Damasceno (2000).

4. Resultados

Com o objetivo de comparar o desempenho do núcleo estimador com os estimadores multivariados paramétricos para densidades multivariadas conhecidos na literatura, um pequeno estudo simulado foi feito para a distribuição normal bivariada. Amostras de tamanhos iguais a $n=40$, $n=100$ e $n=1000$ foram considerados para a implementação da metodologia do núcleo estimador e para a estimação da janela h_i de cada variável, $i=1,2$. Conjuntos de teste de tamanhos $m=50$ e $m=100$ foram usados para testar a metodologia, isto é, estes conjuntos são gerados de acordo com a distribuição normal bivariada e os valores reais da função de densidade são comparados com a estimativa obtida via núcleo estimador.

Os estimadores foram comparados em relação ao erro médio, ao erro médio absoluto, erro quadrático médio e ao erro médio relativo. Para a implementação computacional foi utilizado o *software* S-Plus 2000® tanto para gerar as amostras da normal bivariada quanto para a estimar os valores dos parâmetros de suavidade h_i . Foram geradas amostras aleatórias de tamanhos: $n=40$, $n=100$, e $n=1000$ de uma distribuição normal bivariada com vetor de médias $\mu = (1.0; 3.0)$ e matriz de covariâncias $\Sigma = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$. Estas amostras foram usadas para a estimação dos valores de h_1 -ótimo e h_2 -

ótimo. A determinação destas janelas foi feita usando-se as amostras transformadas pelo método descrito na seção 2. Na avaliação da qualidade de ajuste do método do núcleo usou-se dois conjuntos de testes com tamanhos $m=50$ e $m=100$ também gerados de acordo com a distribuição normal bivariada especificada.

A Tabela 1 e a Tabela 2 apresentam os resultados obtidos para o caso em que $m=50$ e $m=100$ respectivamente. Nestes quadros FN denota a função de densidade estimada pelo método do núcleo, FM a função de densidade estimada pelo método dos momentos e FMV a função densidade estimada por máxima verossimilhança. A Figura 1 apresenta os gráficos gerados com os valores da função de

densidade teórica das $m=50$ observações de teste bem como aqueles gerados com as estimativas da função densidade em cada método de estimação discutido neste artigo. A Figura 2 apresenta os gráficos gerados com $m=100$. De acordo com estes resultados constatou-se que para pequenas amostras os estimadores paramétricos foram melhores que o núcleo estimador. Para amostras maiores ($n \geq 100$) o núcleo estimador teve um comportamento melhor que os estimadores paramétricos. Observou-se também que à medida que o tamanho da amostra aumenta as estimativas das janelas ótimas h_i de cada componente tendem a ser similares. É importante salientar que como os dados gerados são normais já se esperava que os estimadores paramétricos apresentassem bons resultados. O ponto relevante é que o núcleo estimador apresentou resultados comparáveis aos paramétricos e qualidade superior para amostras maiores mostrando que pode ser um competidor para aquelas situações nas quais não se tem clareza de qual a forma matemática da distribuição de probabilidades que está gerando os dados amostrais. Finalmente, é importante observar o efeito do tamanho de amostra na construção do gráfico da densidade da distribuição normal bivariada. Neste artigo o gráfico da distribuição teórica foi feito com apenas 50 ou 100 observações o que não caracteriza um tamanho amostral adequado para se ver a superfície da distribuição normal original adequadamente uma vez que os gráficos gerados são bimodais. Para se ver a distribuição normal com mais propriedade é necessário em torno de 400 a 500 observações.

Tabela 1: Medidas de Desempenho dos Estimadores de Densidade - $m=50$.

Medidas de Desempenho	Tamanho da Amostra	FN	FM	FMV
Erro Médio	40	0,016052	-0,001643	0,001322
	100	0,016052	-0,017718	-0,016308
	1000	-0,000131	0,002096	0,003096
Erro Médio Absoluto	40	0,022814	0,017515	0,017416
	100	0,016879	0,018735	0,017631
	1000	0,001281	0,007194	0,007344
Erro Quadrático Médio	40	0,000820	0,000466	0,000449
	100	0,000438	0,000494	0,000436
	1000	0,000002	0,000008	0,000008
Erro Médio Relativo	40	0,318589	0,300363	0,297589
	100	0,304902	0,345070	0,323634
	1000	0,023365	0,102619	0,091026

Tabela 2: Medidas de Desempenho dos Estimadores de Densidade - $m=100$.

Medidas de Desempenho	Tamanho da Amostra	FN	FM	FMV
Erro Médio	40	0,014286	0,002094	0,004831
	100	0,003713	0,003952	0,005019
	1000	0,001049	0,001241	0,001126
Erro Médio Absoluto	40	0,017476	0,007486	0,008072
	100	0,010354	0,011102	0,010501
	1000	0,001014	0,001517	0,001439
Erro Quadrático Médio	40	0,000628	0,000082	0,000102
	100	0,000131	0,000141	0,000151
	1000	0,000001	0,000003	0,000003
Erro Médio Relativo	40	0,417187	0,148554	0,149276
	100	0,232212	0,241907	0,244091
	1000	0,028041	0,032042	0,030513

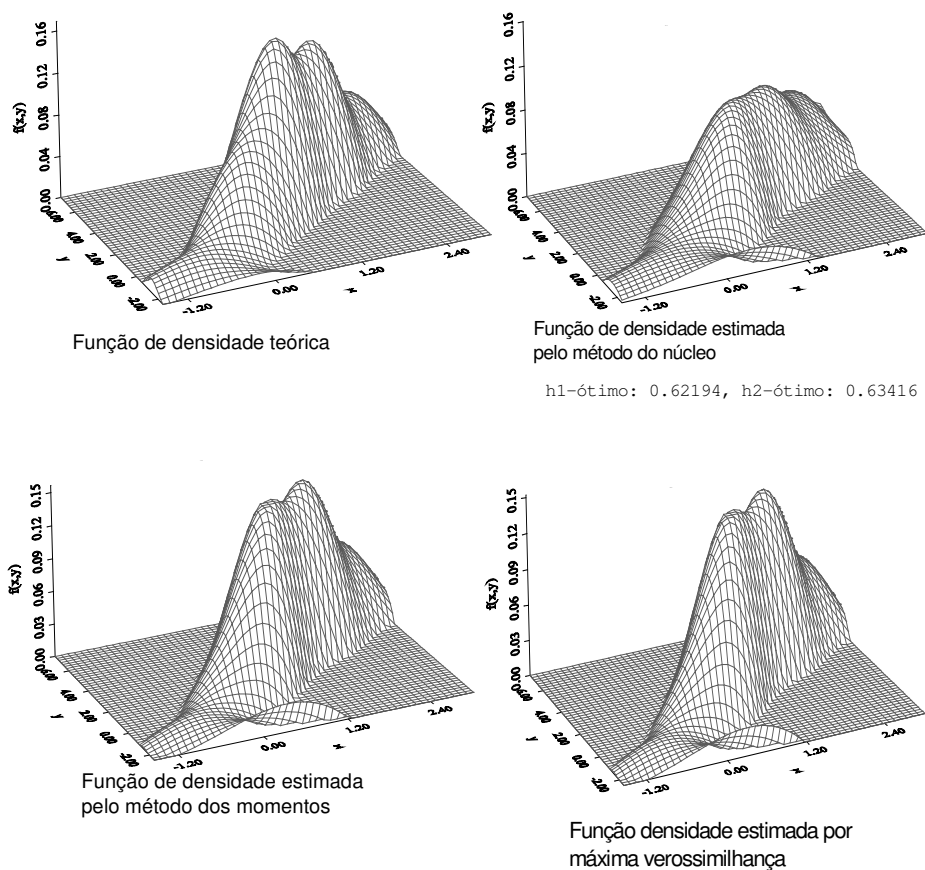


Figura 1: Gráficos de superfície para as funções de densidade multivariadas, com a estimação da janela h com $n=40$ observações e o conjunto de testes $m=50$.

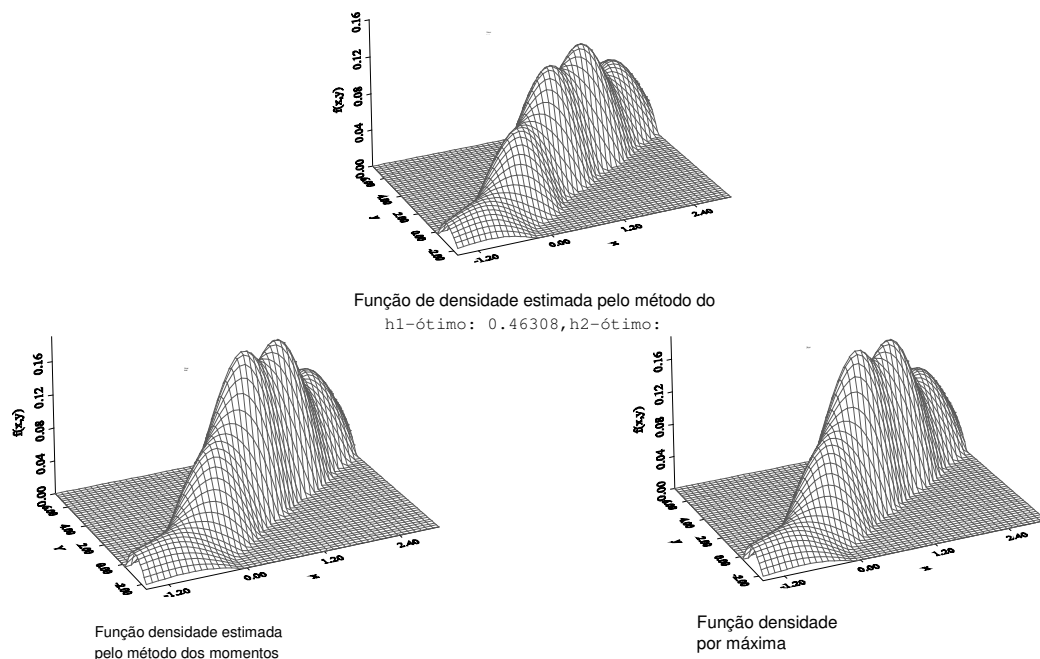


Figura 2: Gráficos de superfície para as funções de densidade multivariadas, com a estimação da janela h com $n=100$ observações e o conjunto de testes $m=50$.

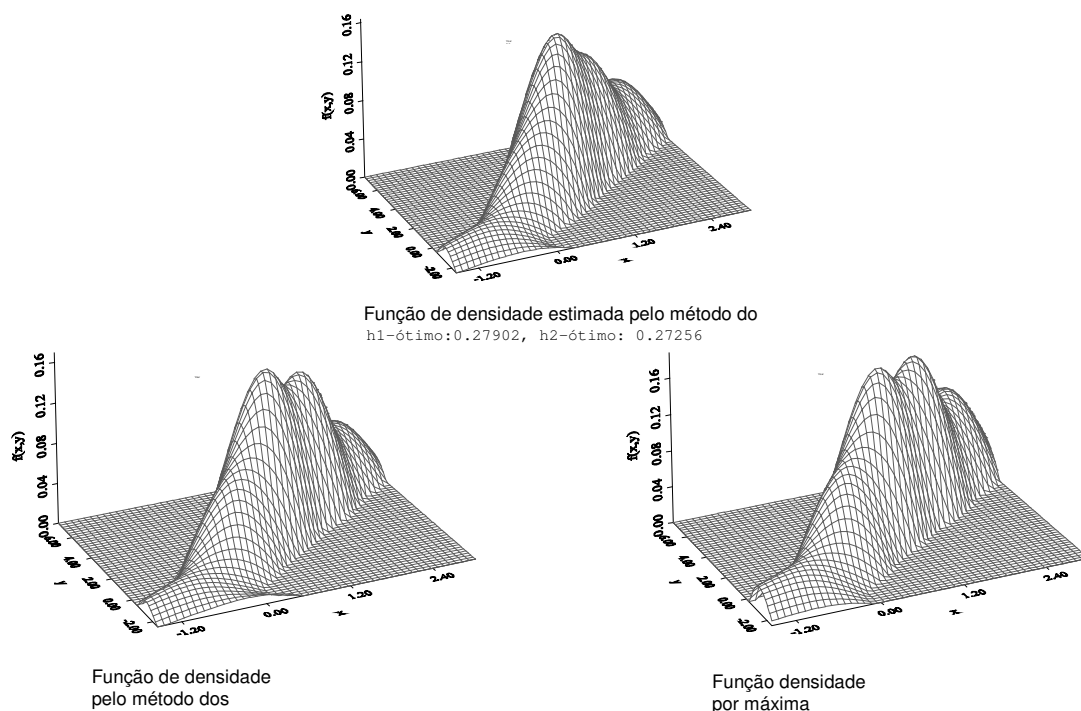
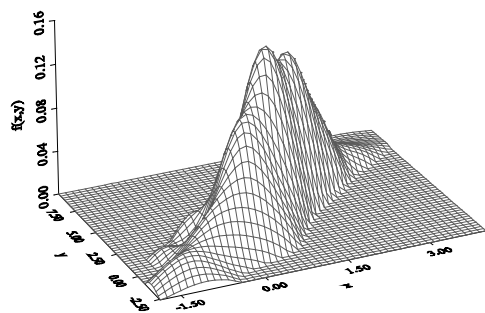
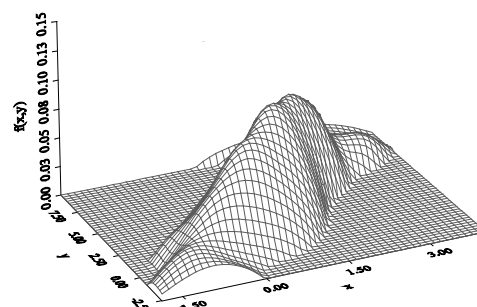


Figura 3: Gráficos de superfície para as funções de densidade multivariadas, com a estimação da janela h com $n=1000$ observações e o conjunto de testes $m=50$.

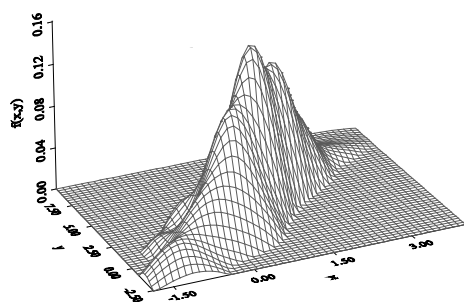


Função de densidade teórica

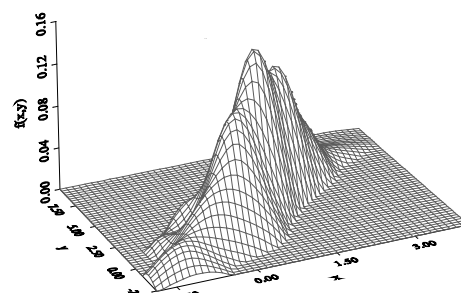


Função densidade estimada pelo método do núcleo

h1-ótimo: 0.60992 h2-ótimo: 0.63391

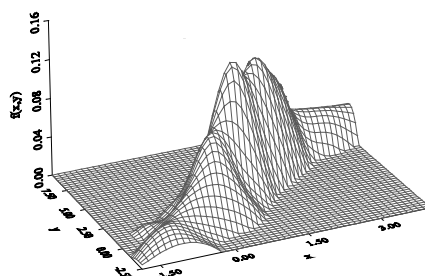


Função de densidade estimada
pelo método dos momentos



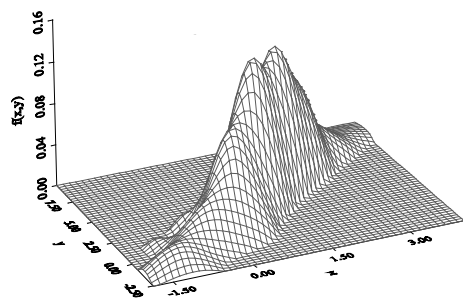
Função densidade estimada
por máxima verossimilhança

Figura 4: Gráficos de superfície para as funções de densidade multivariadas, com a estimação da janela h com $n=40$ observações e o conjunto de testes $m=100$.

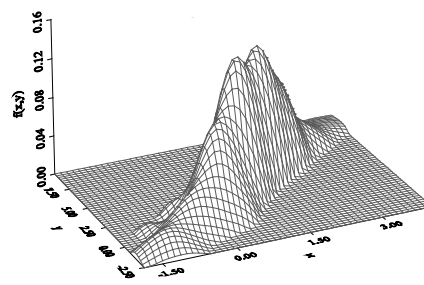


Função densidade estimada pelo método do núcleo

h1-ótimo: 0.48195 h2-ótimo: 0.47873



Função de densidade estimada
pelo método dos momentos



Função densidade estimada
por máxima verossimilhança

Figura 5: Gráficos de superfície para as funções de densidade multivariadas, com a estimação da janela h com $n=100$ observações e o conjunto de testes $m=100$.

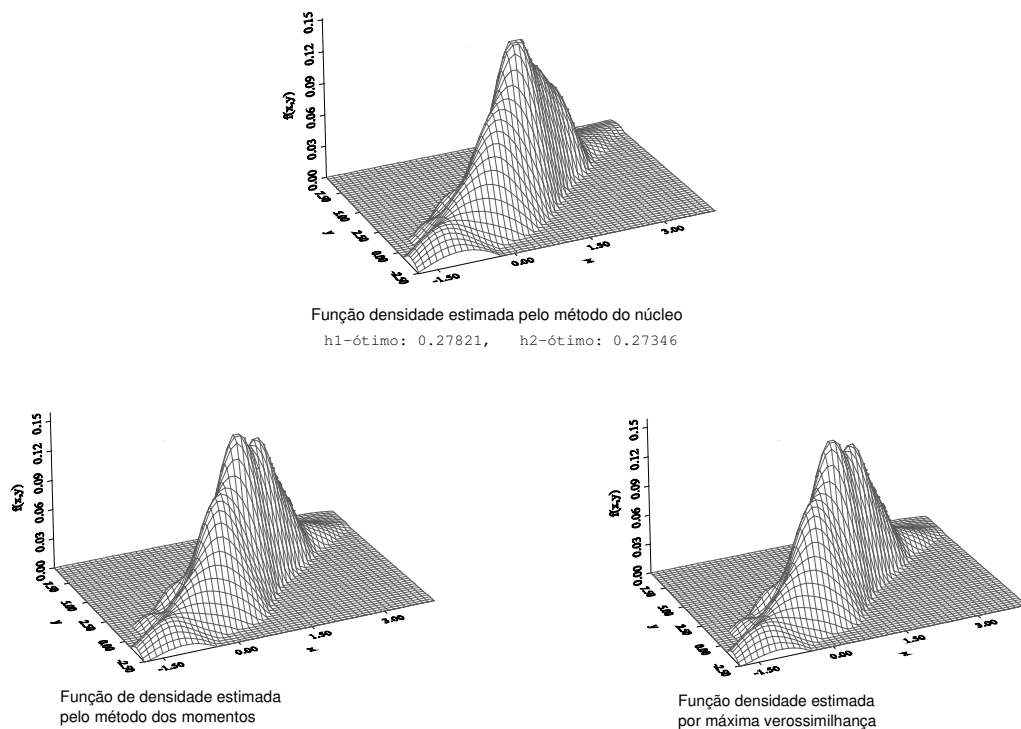


Figura 6: Gráficos de superfície para as funções de densidade multivariadas, com a estimação da janela h com $n=1000$ observações e o conjunto de testes $m=100$.

5. Considerações Finais

Este estudo mostra que o núcleo estimador pode ser facilmente utilizado para a estimação de funções de densidade multivariadas quando se utiliza o método de transformação inversa proposto por Fukunaga (1972). A simulação nos deu resultados promissores no sentido de que esta pode ser uma metodologia com potencial para resolver problemas nos quais não se tem certeza sobre a forma matemática da função de densidade geradora dos dados amostrais. O estudo mostra ainda que os valores da janela ótima são aproximadamente iguais com a qualidade da aproximação melhorando à medida que o tamanho da amostra aumenta. Este é um resultado importante, pois pode diminuir o tempo computacional exigido para a implementação do método de estimação de h_i nos casos em que o número de variáveis é mais elevado. Neste caso, o problema seria simplificado escolhendo-se um valor de h comum para todas as variáveis.

Este estudo, no entanto, é apenas preliminar uma vez que foi simulado um único modelo normal bivariado. Futuramente, pretende-se, através de simulações de grande porte, avaliar de forma mais completa o desempenho do método de núcleo na estimação de funções de densidades multivariadas, em situações nas quais se tem dados normais e em situações mais gerais de dados multivariados não-normais.

Agradecimentos

Agradecemos à CAPES – Brasil e ao CNPq, pelo apoio financeiro que possibilitou a execução deste trabalho.

Referências Bibliográficas

- (1) Atuncar, G.S. e Oliveira, P.J. (1999) *Escolha da Janela Ótima em Estimação Funcional: Caso Markoviano*, Relatório Técnico, Departamento de Estatística da UFMG.

- (2) Atuncar, G.S. e Travassos, A.P.A. (1998) *Implementação de Dois Métodos de Escolha da Janela Ótima em Estimção Funcional*, Relatório de Iniciação Científica, Departamento de Estatística da UFMG.
- (3) Anderson, T.W. (1984) *An introduction to Multivariate Statistical Analysis*, 2nd ed. New York:Wiley.
- (4) Bowman, A. (1984) An alternative method of Cross-Validation for the Smoothing of Density Estimates, *Biometrics*, (71) 353-360.
- (5) Cooley, A. C, Maceachern, S.N. (1998) Classification via kernel product estimators, *Biometrika*, 85, 4, 823-833.
- (6) Chiu, S.T. (1991) Bandwidth Selection for Kernel Density Estimation. *The Annals of Statistics*, Vol. 33, 1883-1905.
- (7) Damasceno, E.C. (2000) *Escolha do Parâmetro de Suavidade em Estimção Funcional*, Dissertação de Mestrado, Departamento de Estatística da UFMG.
- (8) Fukunaga, K. (1972) *Introduction to Statistical Pattern Recognition*, New York: Academic Press.
- (9) Hand, D. J. (1982) *Kernel Discriminant Analysis*. New York: Wiley.
- (10) Johnson, R. A., Wichern, D.W. (2002) *Applied Multivariate Statistical Analysis*, 5th ed. New Jersey.
- (11) Rudemo, M. (1982) Empirical choice of histograms and kernel density estimators, *Scandinavian Journal of Statistics*, 9, 65-78.
- (12) Scott, D.W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley.
- (13) Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.
- (14) Simonoff, J.S. (1996) *Smoothing Methods in Statistics*. Springer Series in Statistics.
- (15) Travassos, A.P.A. (2003) *Problemas de Fronteiras dos Núcleos Estimadores e suas abordagens*, Dissertação de Mestrado, Departamento de Estatística da UFMG.
- (16) Wand, M.P. & Jones. M.C. (1995) *Kernel Smoothing Monographs on Statistics and Applied Probability* 60, Chapman and Hall.
- (17) Woodroffe, M. (1970) On choosing a delta sequence, *Annals of Mathematical Statistics*, 41, 1665-