UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

Uma Abordagem Bayesiana para Seleção de Janela Ótima em Estimação de Densidades Multivariadas

Max Sousa de Lima, Gregório Saravia Atuncar

Relatório Técnico
RTP 01/2008
Série Pesquisa

Uma Abordagem Bayesiana para Seleção da Matriz de Suavização em Estimação de Densidades Multivariadas.

Max Sousa de Lima [1]
Gregório Saravia Atuncar[2]

Resumo

A estimação de densidades multivariadas através do método Núcleo Estimador tem muitos campos de aplicação como, por exemplo, análise de regressão não-paramétrica, análise de discriminante, etc. No entanto, ela tem recebido pouca atenção na literatura devido a grande dificuldade para selecionar a matriz de suavização H em um espaço p-dimensional. O fato é que as dificuldades computacionais crescem com aumento de p. Para superar estas dificuldades propomos neste trabalho uma abordagem Bayesiana para seleção da matriz H. Nesta abordagem, a escolha de H é realizada através da minimização de uma função de perda para H. Estudos empiricos mostraram resultados bastante satisfatórios para estimação de misturas de densidades multivariadas com componentes correlacionadas. Observou-se ainda a influência das especificações a priori para H nas estimativas das densidades.

palavra chave: núcleo-estimador, matrizes aleatórias, estimação bayesiana

[1]Prof. do Dep. de Estatística da UFAM/maxlima@ufam.edu.br
[2]Prof. do Dep. de Estatística da UFMG/Atuncar@est.ufmg.br

1 Introdução

O método Núcleo Estimador é uma ferramenta muito útil para estimação de densidades devido a uma boa visualização dos dados que ele nos fornece (Scot, 1992). Em um espaço p-dimensional (com $p \geq 2$) o método depende da especificação (ou estimação) de uma matriz de suavização, denotado por $\mathbf{H}$. Para grandes valores de p, a escolha de $\mathbf{H}$ é muito difícil, pois requer procedimentos computacionais intensivos e avaliação (ou estimação) de derivas de altas ordens (Duong and Hazelton, 2003). Devido a essas dificuldades este método tem recebido pouca atenção na literatura (Zhang at all, 2006). Por isso, apresentamos um procedimento bayesiano muito simples para seleção de $\mathbf{H}$. Este procedimento pode ser aplicado com facilidade para escolha de $\mathbf{H}$ em qualquer dimensão.

Seja $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p)'$ um vetor aleatório com função de densidade $f(\mathbf{y}) = f(y_1, ..., y_p)$. Dada uma amostra $\mathbf{x}$ de tamanho n da distribuição de $\mathbf{X}$ e, uma matriz $\mathbf{H}$, de dimensão $p \times p$, definida positiva, temos que um estimador não-paramétrico para $f(\mathbf{y})$ é dado por,

$$\tilde{f}(\mathbf{y}|\mathbf{H}) = \int_{\mathbb{R}^p} K_{\mathbf{H}}(\mathbf{y} - \mathbf{x}) d\hat{F}_n(\mathbf{x}) \tag{1}$$

onde, $(\mathbf{y} - \mathbf{x}) = (y_1 - x_{i1}, ..., y_p - x_{ip})', i = 1, 2, ..., n$, $\mathbf{H}$ é uma matriz de suavização, $K_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1/2} K\{\mathbf{H}^{-1/2}(\mathbf{u})\}$ é uma função núcleo $p-$variada, em geral simétrica, tal que $\int_{R^p} K_{\mathbf{H}}(\mathbf{u}) du = 1$ e $\hat{F}_n(\mathbf{x})$ é um estimador para função de distribuição $F(\mathbf{x})$. Se $\hat{F}_n(\mathbf{x})$ é função de distribuição empírica, então dizemos que a densidade f possui núcleo-estimador (Wand and Jones, 1995),

$$\hat{f}(\mathbf{y}|\mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} |\mathbf{H}|^{-1/2} K\{\mathbf{H}^{-1/2}(\mathbf{y} - \mathbf{x}_i)\} \tag{2}$$

Em altas dimensões, existem várias opções para a escolha de $\mathbf{H}$ (Wand and Jones, 1993). por exemplo, a matrix $\mathbf{H}$ pode ser restrita a classe das matrizes diagonais, essa classe pode ser adequada se o vetor $\mathbf{X}$ é composto por componentes não correlacionados. No entanto, se os componentes são correlacionados, o uso de elementos não nulos fora da diagonal de $\mathbf{H}$ pode ser usado para corrigir esse efeito de correlação entre os componentes de $\mathbf{X}$. Isto porque uma matriz de suavização completa permite uma orientação arbitrária a função núcleo. Quando os componentes de $\mathbf{X}$ são correlacionados, duas abordagens são bastante comuns na literatura: 1) os dados são reescalados tal que as variâncias amostrais sejam iguais em cada dimensão e então, a classe usada para $\mathbf{H}$ é $\mathcal{H}_1 = \{h\mathbf{D} : h > 0\}$, onde $\mathbf{D} = diag(\mathbf{C})$ e $\mathbf{C}$ é a matriz de covariância amostral; 2) toma-se uma transformação linear nos dados tal

que a matriz de covariância amostral é a matriz identidade e $\mathcal{H}_2 = \{h\mathbf{C} : h > 0\}$. Wand and Jones (1994) demostram que essas abordagens podem ser inadequadas. No entanto, o método proposto nesse trabalho não requer a reescala ou transformação dos dados.

É conhecido na literatura que o desempenho do núcleo estimador depende primariamente da escolha de $\mathbf{H}$ e não da função núcleo (Scott, 1992). Por isso, ao longo dos anos estudos têm sido realizados em busca de metodologias que estimem de forma automática o valor ótimo de $\mathbf{H}$. estes estudos foram iniciados por Cacoullos (1966), que utilizou uma matriz diagonal de janelas $\mathbf{H} = hI$ para a estimação, em que h é o parâmetro de suavidade univariado, constante para todas as variáveis e I é uma matriz identidade $p \times p$. Epanechnikov (1969) investigou a utilização de uma matriz diagonal com diferentes janelas para cada variável, isto é, $\mathbf{H} = diag(h_1, ..., h_p)$. Fukunaga (1972) propôs utilizar transformações nos dados, de modo que não seja necessário trabalhar com a matriz completa de janelas; a mesma idéia foi abordada por Silverman (1986). Deheuvels (1977) foi o primeiro a discutir o caso em que a matriz de janelas é completa. Wand e Jones (1994) desenvolvem o método plug-in para selcionar uma matriz $\mathbf{H}$ completa em dados multivariados. Duong and Hazelton (2003) apresentam alguns casos onde o método proposto por Wand e Jones (1994) falha. Em resposta a esse problema, esses autores propõem um Plug-in alternativo, este tem a vantagem de sempre produzir uma $\mathbf{H}$ finita. Zhang et al.(2006) propuseram o primeiro procedimento bayesiano para a estimação de $\mathbf{H}$ completa. Eles combinam o método de validação cruzada por verossimilhança (Habbema et al,1974) com métodos de Monte Carlo via Cadeias de Markov-MCMC.

Em geral, a escolha de $\mathbf{H}$ visa otimizar alguma medida de distância entre a densidade estimada e a densidade verdadeira. Usualmente, adota-se o erro médio quadrático integrado. Entretanto, nós propomos uma abordagem bayesiana para estimação de $\mathbf{H}$ baseada em uma função de perda para $\mathbf{H}$. Nossa abordagem é uma extensão das idéias de Gangopadhyay e Cheung (2002) que propuseram uma abordagem bayesiana para a estimação do parâmetro h, no caso univariado.

2 Procedimento Bayesiano para Estimação da matriz H

2.1 Densidade Filtrada

Dado uma matriz $\mathbf{H}$, defina

$$f(\mathbf{y}|\mathbf{H}) = \int_{\mathbb{R}^p} K_{\mathbf{H}}(\mathbf{y} - \mathbf{x}) dF(\mathbf{x}) \tag{3}$$

Neste caso, $f(\mathbf{y}|\mathbf{H})$ representa uma densidade filtrada. O valor da função $f(\mathbf{y}|\mathbf{H})$ é definido pela média de todos os pontos pertencentes a uma matriz $\mathbf{H}$ de janelas h_{kl} cujo o centro é o ponto $\mathbf{y}$. Observa-se que (1) é um estimador natural para $f(\mathbf{y}|\mathbf{H})$ e, se $\hat{F}_n(\mathbf{x})$ é a função de distribuição empírica, temos que (2) é um estimador para $f(\mathbf{y})$ com uma janela fixa $\mathbf{H}$ para cada ponto $\mathbf{y}$. No entanto, a escolha de uma janela fixa pode ser razoável somente em regiões onde a densidade é homogênea. De modo que o ideal é que a densidade possa ser estimada com janelas de formato e tamanho diversos. Esta forma de estimação é descrita na próxima seção.

2.2 Estimação Bayesiana de H

Considere que $K_{\mathbf{H}}(\mathbf{u})$ é núcleo gaussiano, dado por

$$K_{\mathbf{H}}(\mathbf{u}) = (2\pi)^{-p/2} e^{-\frac{1}{2}\mathbf{u}'\mathbf{u}}$$

de modo que $K_{\mathbf{H}}(\mathbf{y}-\mathbf{x})$, representa a densidade de uma distribuição $\mathcal{N}(\mathbf{x}, \mathbf{H})$ dada por:

$$(2\pi)^{-p/2}|\mathbf{H}|^{-1/2} exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{x})'\mathbf{H}^{-1}(\mathbf{y}-\mathbf{x})\right\} \tag{4}$$

Então, para cada n e $\mathbf{H}$, $\tilde{f}$ definida em (1) é uma densidade de probabilidade e $\mathbf{H}$ é parâmetro de escala para o núcleo gaussiano.

Do ponto de vista Bayesiano, a incerteza sobre o parâmetro $\mathbf{H}$ é modelada por uma distribuição de probabilidade, que é chamada de disrtribuição a priori. Desse modo, seja $\pi(\mathbf{H})$ uma distribuicao a priori sobre todos os possiveis valores de $\mathbf{H}$, temos que a distribuicao a posteriori de $\mathbf{H}$ dado o ponto $\mathbf{y}$, é

$$\pi(\mathbf{H}|\mathbf{y}) = \frac{f(\mathbf{y}|\mathbf{H})\pi(\mathbf{H})}{\int f(\mathbf{y}|\mathbf{H})\pi(\mathbf{H})d\mathbf{H}}. \tag{5}$$

Seja $\hat{\mathbf{H}}$ um estimador arbitrário de $\mathbf{H}$, e $L = L(\hat{\mathbf{H}}, \mathbf{H})$ uma função de perda. Então, a perda esperada a posteriori, associada a $\hat{\mathbf{H}}$, é dada por:

$$\mathbb{E}[L(\hat{\mathbf{H}}, \mathbf{H})|\mathbf{y}] = \int L(\hat{\mathbf{H}}, \mathbf{H})\pi(\mathbf{H}|\mathbf{y})d\mathbf{H}$$

de forma que a regra de Bayes consiste em escolher o estimador que minimiza esta perda esperada. Então, o estimador de Bayes para $\mathbf{H}$ no ponto $\mathbf{y}$ é dado por

$$\hat{\mathbf{H}}(\mathbf{y}) = \underset{\hat{\mathbf{H}} \in \mathcal{H}}{Min} \left[\mathbb{E}[L(\hat{\mathbf{H}}, \mathbf{H})|\mathbf{y}] \right].$$

Onde $\mathcal{H}$ é o espaço das matrizes positivas definidas. Duas funções de perda comumente usadas são:

$$L_1 = tr(\hat{\mathbf{H}}\mathbf{H}^{-1}) - log|\hat{\mathbf{H}}\mathbf{H}^{-1}| - p.$$

e,

$$L_2 = tr(\mathbf{H} - \hat{\mathbf{H}})^2.$$

onde tr(M) representa o traço da matriz M.

A primeira função é chamada perda de entropia (Stein, 1956) e a segunda é uma função de perda quadrática, para mais funções de perda ver (Haff, 1977). Se a função de perda L_1 é utilizada, temos que o estimador de Bayes para $\mathbf{H}$ é,

$$\hat{\mathbf{H}}(\mathbf{y}) = \left[\mathbb{E}_\pi(\mathbf{H}^{-1}) \right]^{-1}, \quad \text{(Yang and Berger, 1994)}$$

e sob L_2 temos,

$$\tilde{\mathbf{H}}(\mathbf{y}) = \left[\mathbb{E}_\pi(\mathbf{H}) \right], \text{(Press, 1982)}.$$

Em que $\mathbb{E}_\pi$ é o valor esperado com respeito a distribuição a posteriori de $\mathbf{H}$. Observa-se que para as funções de perda descritas acima, o estimador de Bayes é o valor esperado de alguma função $g(\mathbf{H})$ com respeito a distribuição a posteriori.

2.3 Priori e Estimadores de Bayes para H

2.3.1 Escolha da *Priori*

A distribuição a *priori* para $\mathbf{H}$ é a Wishart invertida, denotada por $\mathbf{H} \sim \mathcal{WI}_p(d, V)$ com densidade,

$$\pi(\mathbf{H}) = \frac{|\mathbf{V}|^{d/2}|\mathbf{H}|^{-(d+p+1)/2} exp\{-tr(\mathbf{V}\mathbf{H}^{-1})/2\}}{2^{dp/2}\Gamma_p(d/2)}. \tag{6}$$

Em que $d \geq p$ representa os graus de liberdade ou parâmetro de precisão, V é uma matrix $p \times p$ definida positiva, comumente chamada de matrix escala ou parâmetro de escala e $\Gamma_p(.)$ é a função gama generalizada. A família de distribuições $Wishart$ é muito flexível e inclui uma variedades de formas, pode ser facilmente manipulada para incorporar as classes $\mathcal{H}_1$, $\mathcal{H}_2$ dentre outras. Para uma discussão completa e propriedades veja Press (1982, cap V). Alguns resultados que serão usados neste trabalho, são:

- $\Phi = \mathbf{H}^{-1}$ possui distribuição Wishart com parâmetros (d,V) e $\mathbb{E}(\Phi) = dV^{-1}$ tal que V/d é uma estimativa para $\mathbf{H}$.
- Se V possui elementos v_{kl} então para $d > p+1$, a $\mathbb{E}[h_{kl}] = v_{kl}/(d-p-1)$ e para $d > p+3$,

$$Var[h_{kl}] = \frac{(d-p+1)v_{kl}^2 + (d-p-1)v_{kk}v_{ll}}{(d-p)(d-p-1)^2(d-p-3)}$$

e,

$$Var[h_{kk}] = \frac{2v_{kk}^2}{(d-p-1)^2(d-p-3)}, \quad Cv[h_{kk}] = \left(\frac{2}{d-p-3}\right)^{1/2}.$$

onde Cv é o coeficiente de variação de h_{kk}.

- Quando $d \to (p+3)$ a $Var[h_{kl}] \to \infty$, se $d \to (p+5)$ então $\mathbf{H}$ está centrada em $V/4$ com $Cv[h_{kk}] = 1$. Por fim, se $d \to \infty$ a distribuição de $\mathbf{H}$ está concentrada próxima de uma matriz nula $\mathbf{0}$.

2.3.2 Calculo dos Estimadores

Uma dificuldade técnica para obtenção de $\pi(\mathbf{H}|\mathbf{y})$ em (5) é que $f(\mathbf{y}|\mathbf{H})$ é desconhecida, ou seja, a distribuição a *posteriori* de $\mathbf{H}$ não pode ser obtida diretamente. No entanto se assumirmos, para cada n e amostra $\mathbf{x}$, que $\hat{F}_n(\mathbf{x})$ é a função de distribuição empírica, então temos que uma estimativa da distribuição a posteriori é dada por

$$\hat{\pi}(\mathbf{H}|\mathbf{y}) = \frac{\hat{f}(\mathbf{y}|\mathbf{H})\pi(\mathbf{H})}{\int \hat{f}(\mathbf{y}|\mathbf{H})\pi(\mathbf{H})d\mathbf{H}}. \tag{7}$$

O produto no numerador em (7) é dado por,

$$\frac{1}{|\mathbf{V}|^{-d/2}2^{dp/2}\Gamma_p(d/2)} \int_{R^p} \left[|\mathbf{H}|^{-(d+p+2)/2} exp\{-tr(\mathbf{\Delta}_i\mathbf{H^{-1}})/2\}\right] d\hat{F}_n(\mathbf{x}) \tag{8}$$

com $\mathbf{\Delta}_i = [(\mathbf{y}-\mathbf{x}_i)(\mathbf{y}-\mathbf{x}_i)' + \mathbf{V}]$. Como (6) é uma densidade de probabilidade, ou seja, integra 1. temos que o denominador em (7) é,

$$\frac{2^{(d+1)p/2}\Gamma_p((d+1)/2)\sum_{i=1}^n |\mathbf{\Delta}_i|^{-(d+1)/2}}{n|\mathbf{V}|^{-d/2}(2\pi)^{p/2}2^{dp/2}\Gamma_p(d/2)}. \tag{9}$$

combinando (8) e (9) obtemos,

$$\begin{aligned}\hat{\pi}(\mathbf{H}|\mathbf{y}) &= \frac{1}{2^{(d+1)p/2}\Gamma_p((d+1)/2)} \times \frac{\sum_{i=1}^n |\mathbf{H}|^{-(d+p+2)/2} exp\{-tr(\mathbf{\Delta}_i\mathbf{H^{-1}})/2\}}{\sum_{i=1}^n |\mathbf{\Delta}_i|^{-(d+1)/2}} \\ &= \sum_{i=1}^n \omega_i \mathcal{WI}_p(d+1, \Delta_i)\end{aligned}$$

onde

$$\omega_i = \frac{|\mathbf{\Delta}_i|^{-(d+1)/2}}{\sum_{i=1}^{n} |\mathbf{\Delta}_i|^{-(d+1)/2}}.$$

Então, usando os resultados da seção (2.3.1), temos que sob a função de perda L_1 o estimador de Bayes para $\mathbf{H}$ no ponto é,

$$\hat{\mathbf{H}}(\mathbf{y}) = \frac{\left[\sum_{i=1}^{n} \omega_i \Delta_i^{-1}\right]^{-1}}{d+1}$$

e sob L_2,

$$\tilde{\mathbf{H}}(\mathbf{y}) = \frac{\sum_{i=1}^{n} \omega_i \Delta_i}{d-p}.$$

Observa-se que

$$\mathbf{\Delta}_i = [(\mathbf{y} - \mathbf{x}_i)(\mathbf{y} - \mathbf{x}_i)' + \mathbf{V}] = [(\mathbf{y} - \mathbf{x}_i)(\mathbf{y} - \mathbf{x}_i)' + (d-p-1)\mathbb{E}(\mathbf{H})]$$

De modo que $\hat{\mathbf{H}}(\mathbf{y})$ e $\tilde{\mathbf{H}}(\mathbf{y})$, podem ser representados por uma combinação entre as distâncias espaciais, ponderadas, do ponto $\mathbf{y}$ ao vetor de dados amostrais e o valor esperado a *priori* para $\mathbf{H}$. É possível mostrar que o ponderador ω_i decresce a medida que as distâncias espaciais do ponto $\mathbf{y}$ ao vetor de dados amostrais crescem. Ou seja, pontos amostrais distantes de $\mathbf{y}$ contribuem muito pouco para a estimação de $\mathbf{H}$. Portanto, esses estimadores obtidos para $\mathbf{H}$, são consistentes no sentido de que, se o interesse é estimar f pontualmente em $\mathbf{y}$, então somente a vizinha de $\mathbf{y}$ deve ser de vital importância.

2.3.3 Escolha dos Parâmetros da *Priori*

Para finalizarmos devemos especificar os hiperparâmetros $(d, \mathbf{V})$. Esta especificação pode ser realizada da seguinte forma:

1. A Escolha de d: pode ser feita fixando um valor para o coeficiente de variação Cv descrito na seção (2.3.1). Por exemplo, se temos muita incerteza sobre o valor esperado a *priori* para $\mathbf{H}$, podemos escolher $3 < d - p \leq 5$. Tipicamente estudos empíricos tem nos mostrado que $d - p = 5$ é a melhor opção.

2. Escolha de $\mathbf{V}$: na prática é muito dificíl a escolha de V, pois certamente não teremos, a *priori*, informação suficiente sobre $\mathbf{H}$ que nos leve a uma escolha adequada de V. No entanto, estudos empíricos realizados e apresentados por vários autores, como Silverman (1986), Scott(1992) e

Wand(1993) têm mostrado que uma boa regra de referência é $\mathbf{H}_0 \propto \Sigma_0$, onde Σ_0 é a variãncia dos dados e $\mathbf{H}_0$ é chamada de estimativa piloto para $\mathbf{H}$. Em implementações práticas, se Σ_0 é desconhecido, ele pode ser subtituído por um estimador para a variância.

3. A Escolha do par $(d, \mathbf{V})$: pode ser feita diretamente usando a regra de referência generalizada de Scott. Neste caso $(d, V) = (n^{2/(p+4)} + p, \Sigma_0)$. Observa-se que sob essa escolha, a *priori* $\mathbf{H}$ está concentrada pròximo de uma matriz nula e é aproximadamente $n^{-2/(p+4)}\Sigma_0$. O fato da escolha de d depender do tamanho da amostra n pode ser importante para garantir a consistência da densidade estimada.

3 Estudos Numérico

nesta seção apresentamos a perfomance do método proposto para seleção da matriz de suavizção através de um conjunto de dados simulados. Este conjunto de dados é gerado de densidades conhecidas em um espaço p-dimensional, com p=2, 3 e 4. A medida usada para avaliar a perfomance do método foi erro quadrático integrado (ISE). Para cada densidade foram utilizados 50 conjunto de dados com amostras de tamanho 100, 200 e 500. Como o estimador é pontual, contruímos as seguintes malhas:

- malha de 80 pontos em cada dimensão, para p=2.
- malha de 40 pontos em cada dimensão, para p=3.
- malha de 20 pontos em cada dimensão, para p=4.

O gráfico de contorno para as verdadeiras densidades bivariadas é apresentado na figura (1) e a superfície de contorno para as densidades trivariadas é mostrada na figura (2). As densidades avaliadas foram:

- Densidade A: Densidade assimétrica, é uma mistura do produto de uma distribuição Normal condicionada no valor observado de uma distribuição Weibull.

$$f_A(\mathbf{y}|\alpha, \beta) = \frac{1}{2}\mathcal{N}(y_1/2, 1)Y_1(\alpha_1, \beta_1) + \mathcal{N}(y_2/4, 1)Y_2(\alpha_2, \beta_2).$$

  onde, $Y_i(\alpha_i, \beta_i)$ é a distribuição Weibull com parãmetros de escala $\alpha_1 = \alpha_2 = 10$, parâmetros de forma $\beta_1 = 2$ e $\beta_2 = 4$.

- Densidade B: Distribuição Bimodal I, é uma mistura de duas densidades t Student Bivariadas,

$$f_B(\mathbf{y}|\mu,\Sigma,\nu) = \frac{1}{2}t_p(\mathbf{y}|\mu_1,\Sigma_1,\nu_1) + \frac{1}{2}t_p(\mathbf{y}|\mu_2,\Sigma_2,\nu_2)$$

onde $t_p(\mathbf{y}|\mu,\Sigma,\nu)$ denota a distribuição t p-dimensional com parâmetro de locação μ, matriz de dispersão Σ e ν graus de liberdade e p=2. Os parãmetros são:

$$\mu_1 = (3,3)' \quad \text{e} \quad \mu_2 = (-3,-3)'$$

$$\mathbf{\Sigma_1} = \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix} \quad \text{e} \quad \mathbf{\Sigma_2} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

e $\nu_1 = \nu_2 = 3$.

- Densidade C: Densidade bimodal II, é uma mistura de duas normais com correlação moderada, porém em sentidos opostos.

$$f_C(\mathbf{y}|\mu,\Sigma,\nu) = \frac{1}{2}\phi(\mathbf{y}|\mu_1,\Sigma_1) + \frac{1}{2}\phi(\mathbf{y}|\mu_2,\Sigma_2)$$

onde $\phi(\mathbf{y}|\mu,\Sigma)$ denota a densidade da distribuição normal multivariada com média μ, matriz de variância-covariância Σ, e

$$\mu_1 = (1,1)' \quad \text{e} \quad \mu_2 = (-1,-1)'$$

$$\mathbf{\Sigma_1} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \quad \text{e} \quad \mathbf{\Sigma_2} = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

- Densidade D: Densidade trimodal , é uma mistura de três normais.

$$f_D(\mathbf{y}|\mu,\Sigma,\nu) = \frac{9}{20}\phi(\mathbf{y}|\mu_1,\Sigma_1) + \frac{9}{20}\phi(\mathbf{y}|\mu_2,\Sigma_2) + \frac{2}{20}\phi(\mathbf{y}|\mu_3,\Sigma_3)$$

com,

$$\mu_1 = (-6/5,6/5)' \quad \text{e} \quad \mu_2 = (6/5,-6/5)' \quad \text{e} \quad \mu_3 = (0,0)$$

$$\mathbf{\Sigma_1} = \begin{bmatrix} 9/25 & 27/250 \\ 27/250 & 9/25 \end{bmatrix}, \quad \mathbf{\Sigma_2} = \begin{bmatrix} 9/25 & -27/125 \\ -27/125 & 9/25 \end{bmatrix}$$

e,

$$\mathbf{\Sigma_3} = \begin{bmatrix} 1/4 & 1/80 \\ 1/80 & 1/4 \end{bmatrix}$$

- Densidade E: é uma mistura de dois Modelos auto-regressivos de primeira ordem em série temporal com $f_E = \frac{1}{2}\phi(\mathbf{y}|\mu_1, \Sigma_1) + \frac{1}{2}\phi(\mathbf{y}|\mu_2, \Sigma_2)$, onde

$$\mathbf{\Sigma_1} = \frac{1}{1-\rho_1^2}\begin{bmatrix} 1 & \rho_1 & \rho_1^2 \\ \rho_1 & 1 & \rho_1 \\ \rho_1^2 & \rho_1 & 1 \end{bmatrix}, \quad \mu_1 = (4,4,4)' \quad \text{e} \quad \rho_1 = 0.9$$

e,

$$\mathbf{\Sigma_2} = \frac{1}{1-\rho_2^2}\begin{bmatrix} 1 & \rho_2 & \rho_2^2 \\ \rho_2 & 1 & \rho_2 \\ \rho_2^2 & \rho_2 & 1 \end{bmatrix}, \quad \mu_2 = (1,1,1)' \quad \text{e} \quad \rho_2 = 0.7$$

- Densidade F: é uma mistura de duas densidades t trivariada,

$$f_F(\mathbf{y}|\mu, \Sigma, \nu) = \frac{1}{2}t_p(\mathbf{y}|\mu_1, \Sigma, \nu) + \frac{1}{2}t_p(\mathbf{y}|\mu_2, \Sigma, \nu)$$

com $\nu = 3$, $\mu_1 = (1.5, 1.5, 1.5)'$, $\mu_2 = (-1.5, -1.5, -1.5)'$ e Σ é a matrix identidade.

- Densidade G: é uma mistura de normais com densidade $f_G = \frac{1}{2}\phi(\mathbf{y}|\mu_1, \Sigma_1) + \frac{1}{2}\phi(\mathbf{y}|\mu_2, \Sigma_2)$ onde,

$$\mathbf{\Sigma_1} = \frac{1}{0.19}\begin{bmatrix} 1 & 0.9 & 0.81 & 0.729 \\ 0.9 & 1 & 0.9 & 0.81 \\ 0.81 & 0.9 & 1 & 0.9 \\ 0.729 & 0.81 & 0.9 & 1 \end{bmatrix}, \quad \mu_1 = (4,4,4,4)'$$

e,

$$\mathbf{\Sigma_2} = \frac{1}{51}\begin{bmatrix} 1 & 0.7 & 0.49 & 0.343 \\ 0.7 & 1 & 0.7 & 0.49 \\ 0.49 & 0.7 & 1 & 0.7 \\ 0.343 & 0.49 & 0.7 & 1 \end{bmatrix}, \quad \mu_1 = (1,1,1,1)'$$

- Densidade H: é uma mistura de duas normais com estrutura de correlação circular com densidade $f_H = \frac{1}{2}\phi(\mathbf{y}|\mu_1, \Sigma_1) + \frac{1}{2}\phi(\mathbf{y}|\mu_2, \Sigma_2)$, onde

$$\mathbf{\Sigma_1} = \sigma_1^2\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_1 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_1 & \rho_2 & \rho_1 & 1 \end{bmatrix}, \quad \mu_1 = (1,1,1,1)', \rho_1 = \rho_2 = 0.6, \sigma_1^2 = 1.$$

e,

$$\mathbf{\Sigma_2} = \sigma_2^2\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_1 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_1 & \rho_2 & \rho_1 & 1 \end{bmatrix}, \quad \mu_2 = (-1,-1,-1,-1)', \rho_1 = 0.5, \rho_2 = 0.7, \sigma_2^2 = 1.$$

Na tabela 1 é apresentado o valor médio e o desvio padrão do erro quadrático integrado para método Bayesiano proposto com $(d, V) = (n^{2/(p+4)} + p, \Sigma_0)$ o qual é denotado por MB1, também usamos $(d, V) = (n^{2/(p+4)} + p, \hat{\Sigma})$ que é denotado por MB2. Para efeito de comparação, apresentamos os resultados para o método plug-in, denotado por PI, proposto por Duong e Hazelton (2003) com transformação esférica nos dados(sphering transformation of data). Esse método tem sido considerado, na literatura, como um dos melhores métodos. Ressaltamos que análise bayesiana é restrita aos resultados obtidos com a função perda de entropia. Ao analisar-mos os resultados, observamos que:

- Nas densidade A, C e E, os métodos PI, MB1 e MB2 são praticamente equivalentes.
- Nas densidades B, D e F o método PI é melhor que MB1 e MB2 em todos os tamanhos de amostra.
- Nas densidades G e H, MB1 e MB2, apresentaram melhores resultados que o PI.

Uma melhor comparação do desempenho de cada método pode ser feita através da visualização do box plot do log(ISE), o qual é apresentado na figura (3). Observa-se ainda que não há muita perda de eficiência no método baysiano ao subtiuirmos Σ_0 por $\hat{\Sigma}$.

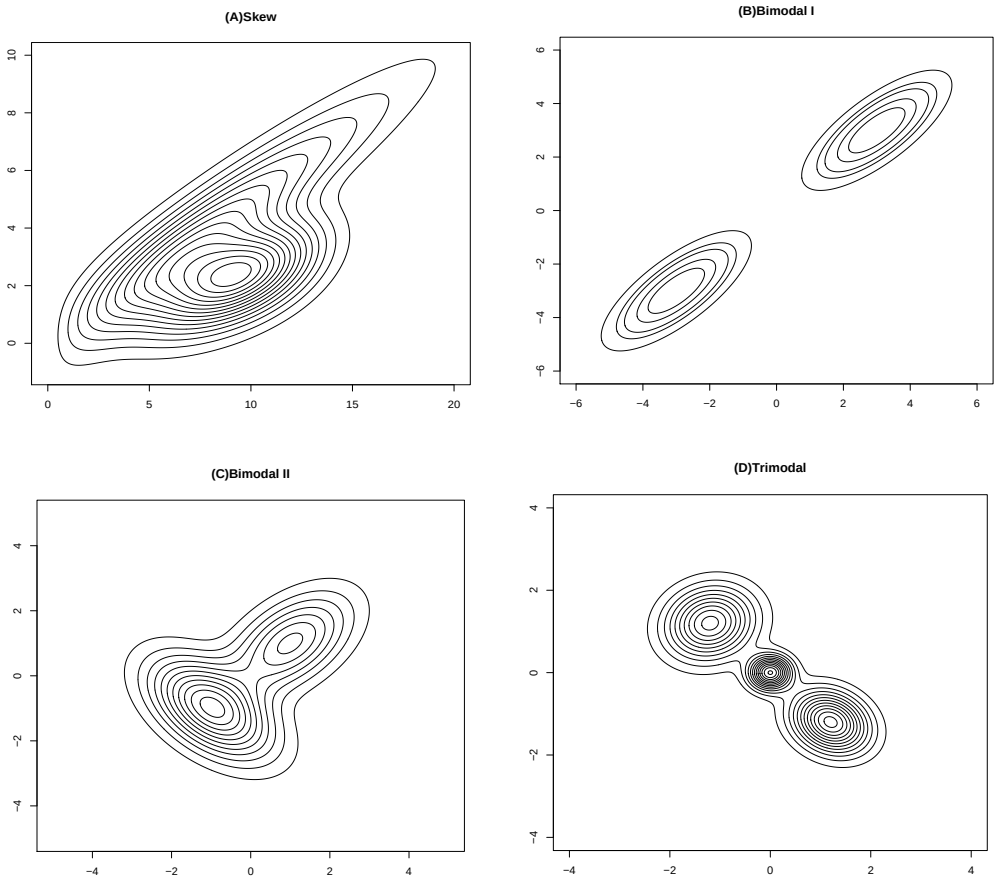

Figura 1: Contour Plot for proposed bivariate densities

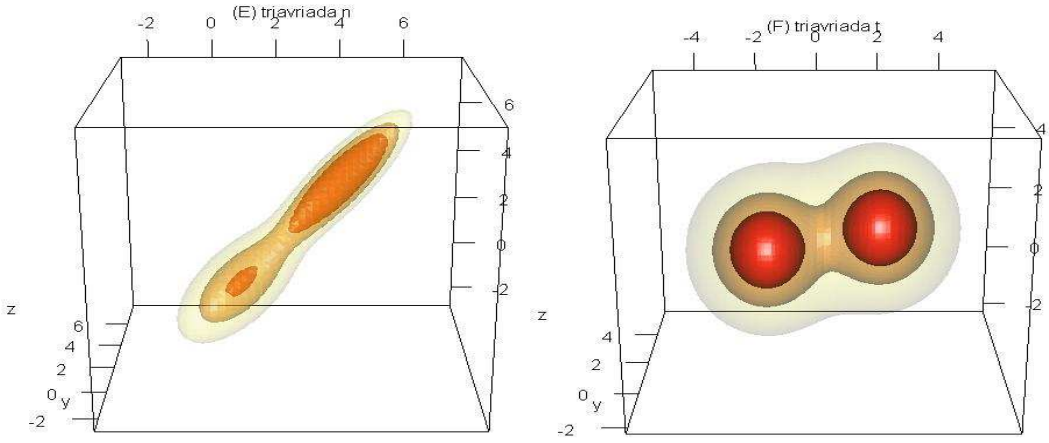

Figura 2: Contour Surface for proposed trivariate densities

Tabela 1: *mean and standard deviation of ISE for proposed densities*

		Mean			Sd		
Density	n	MB1	MB2	PI	MB1	MB2	PI
f_A	100	0.00497	0.00510	0.00479	0.00094	0.00112	0.00114
	200	0.00461	0.00465	0.00449	0.00070	0.00072	0.00074
	500	0.00452	0.00454	0.00445	0.00049	0.00053	0.00055
f_B	100	0.01204	0.01200	0.00876	0.00156	0.00178	0.00215
	200	0.01049	0.01029	0.00681	0.00107	0.00132	0.00145
	500	0.00881	0.00880	0.00550	0.00055	0.00079	0.00064
f_C	100	0.00517	0.00524	0.00518	0.00150	0.00153	0.00151
	200	0.00372	0.00375	0.00371	0.00097	0.00097	0.00097
	500	0.00249	0.00251	0.00243	0.00054	0.00056	0.00057
f_D	100	0.01586	0.01598	0.01475	0.00346	0.00350	0.00379
	200	0.01279	0.01276	0.01088	0.00207	0.00208	0.00228
	500	0.00941	0.00938	0.00726	0.00147	0.00148	0.00157
f_E	100	0.00308	0.00310	0.00311	0.00032	0.00042	0.00049
	200	0.00284	0.00286	0.00284	0.00025	0.00029	0.00034
	500	0.00260	0.00261	0.00259	0.00019	0.00022	0.00026
f_F	100	0.00380	0.00371	0.00289	0.00029	0.00042	0.00048
	200	0.00341	0.00334	0.00237	0.00025	0.00039	0.00043
	500	0.00290	0.00287	0.00183	0.00017	0.00022	0.00021
f_G	100	0.00105	0.00106	0.00111	0.00010	0.00011	0.00012
	200	0.00095	0.00095	0.00098	0.00008	0.00009	0.00010
	500	0.00085	0.00085	0.00086	0.00007	0.00007	0.00008
f_H	100	0.00279	0.00280	0.00303	0.00039	0.00039	0.00043
	200	0.00246	0.00247	0.00261	0.00027	0.00028	0.00028
	500	0.00235	0.00235	0.00258	0.00022	0.00023	0.00022

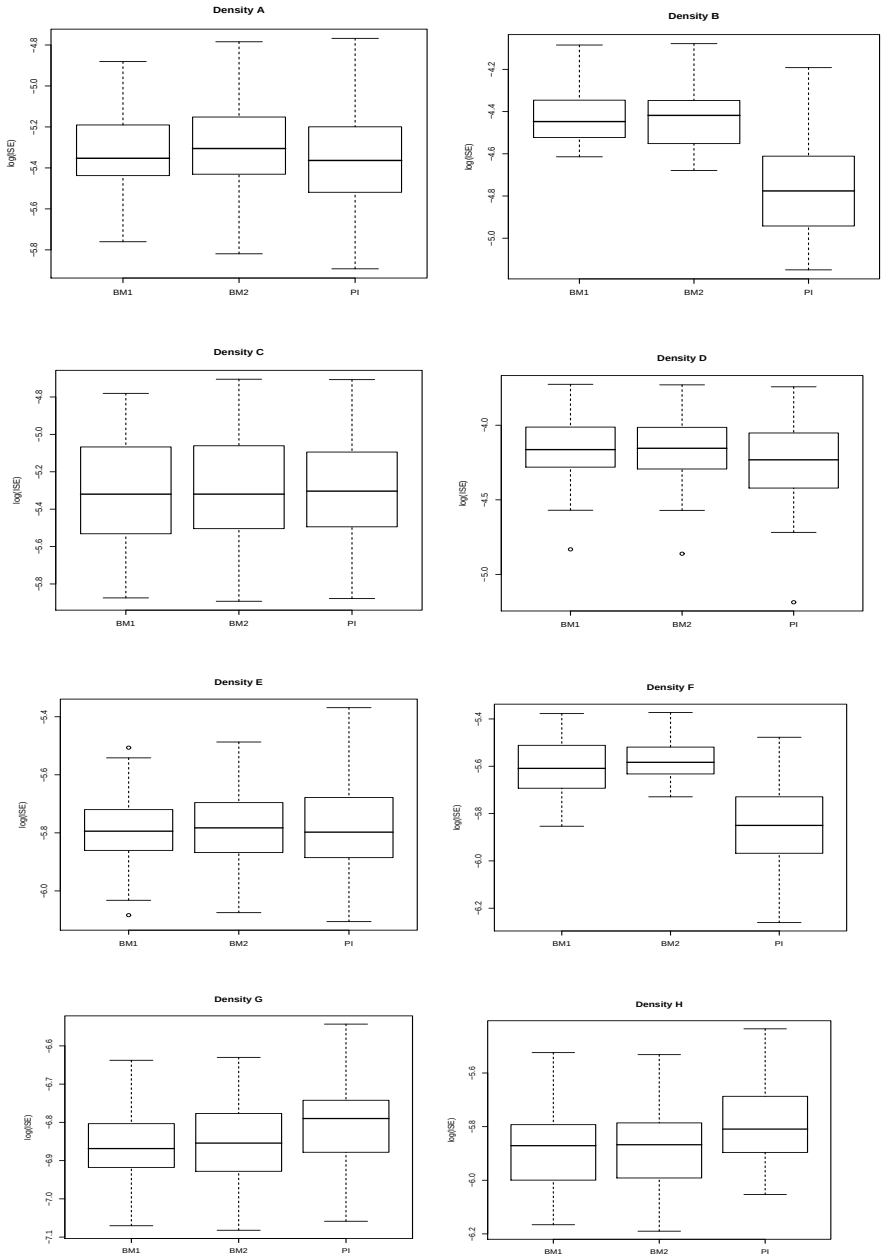

Figura 3: Box plots of log(ISE) for samples of size n=100 from proposed densities

4 Aplicação a Dados Reais: Análise de Clusters

Nesta seção a metodologia proposta é aplicada a um conjunto de dados de 4 variáveis para 3 espécie de Iris (espécie de flor) discutido em Scott(1992). Para 50 flores de cada uma das espécies (Iris Setosa, Versicolour e Virginica) foram obtidas observações (em cm) das variáveis: Largura da pétala, Comprimento da pétala, Largura da Sépala Comprimento da Sépala. É conhecido na literatura que existem três grupos neste conjunto de dados. O Scaterplot apresentado na figura (4) mostra o comportamento destas variáveis observadas.

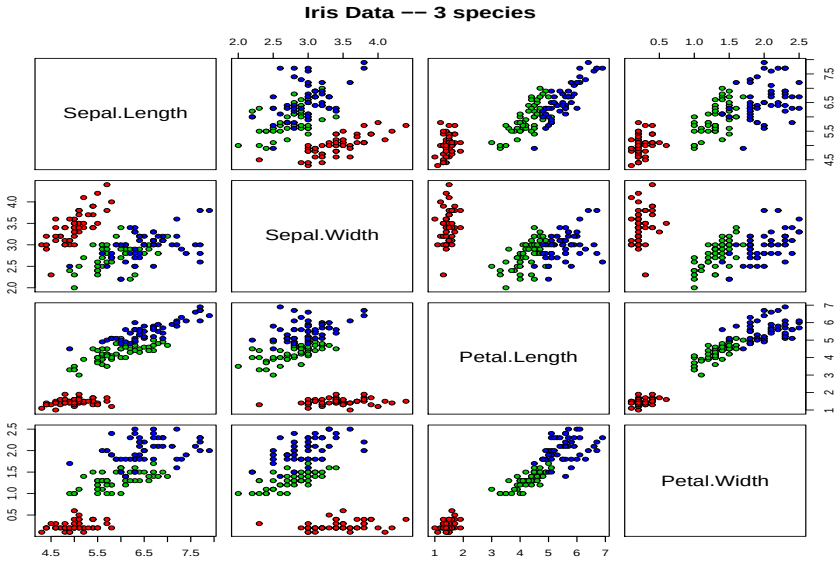

Figura 4: Escaterplot para o "iris data"

Usamos o método proposto com $(d, V) = (n^{2/(p+4)} + p, \hat{\Sigma})$. Na presença de pouca (ou nhenhuma) informação usamos $V = \hat{\Sigma}$. Onde $\hat{\Sigma}$ representa a matriz de Variância-Covariância amostral. Observe que esta *priori* depende dos dados e gera um conflito com a idéia de *priori*. No entanto os resultados obtidos com dados simulados na seção 3, mostraram que não há muita perda de consistência nos resultados quando obtemos V com um valor estimado dos dados. Observe, agora, na figura (5) que a através densidade estimada pelo método proposto, identifica claramente os 3 clusters existentes nesse conjunto de dados.

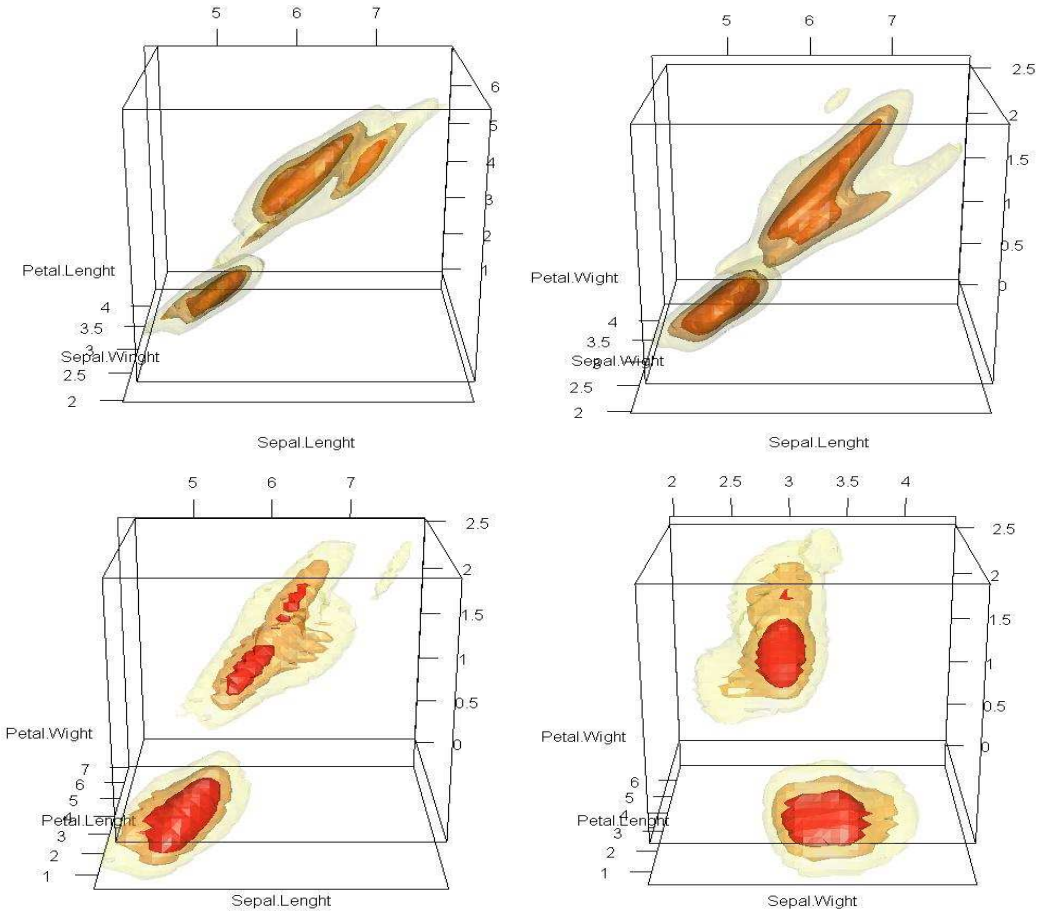

Figura 5: Densidade Estimada para o "iris data"

5 Considerações Finais

Neste trabalho foi desenvolvido um método Bayesiano para Estimacão da matriz de suavizacão usada em estimacão de densidades multivariadas através do núcleo-estimador. As grande vantagem deste método é que ele não necessita que as variáveis aleatória sejam independentes, e também, não requer o uso de métodos computacionais intensivos e isto, facilita sua aplicabilidade a qualquer conjunto de dados de qualquer dimensão. Resultados de estudos empirícos mostraram que dimensões $p > 3$, ele é mais eficiente que o método Plug-in alternativo proposto por Duong e Hazelton (2003). Sua aplicabilidade a análise de cluster também mostrou que o método possui uma boa perfomance para identificar cluster em altas dimensões.

Agradecimentos

O primeiro Autor agradece o suporte financeiro concedido pela Fundação de Amparo à Pesquisa do Estado do Amazonas-**Fapeam**.

Referências

[1] COMANICIU,D. and MEER. P. (1999). *Distribution Free Decomposition of Multivariate Data*. Pattern Analysis & Applications, N° 2: 22 − 30.

[2] CACOULLOS, T. (1966) *Estimation of a multivariate density*. Annals of the Institute of Statistical Mathematics, v. 18, p. 179 − 189.

[3] DEHEUVELS, P. (1977) *Estimation non paramétrique de la densité par histogrammes généralises*. Publications del Institute Statistique de Université Paris, v. 22, p. 1 − 23.

[4] Duong, T. Hazelton, M.L. (2003) *Plug-in bandwidth selectors for bivariate kernel density estimation*. Journal Nonparametric Statist, v. 15, p. 17−30.

[5] EPANECHNIKOV, V. (1969). *Non parametric estimation of a multivariate probability density*. Theory of Probability and Its Applications, v. 14, p. 153 − 158.

[6] FUKUNAGA, K. (1972) *Introduction to statistical pattern recognition*. [S.l.]: New York: Academic Press.

[7] GANGOPADHYAY, A.; CHEUNG, K. (2002). *Bayesian approach to the choice of smoothing parameter in kernel density estimation*. Nonparametric Statistics, v. 14(6), p. 655 − 664.

[8] HÄARDLE, W., MÄULLER, M., (2000). *Multivariate and semiparametric kernel regression*. In Schimek, M.G. (Eds.), Smoothing and Regression: Approaches, Computation, and Application. John Wiley & Sons, New York, 357 − 392.

[9] HABBEMA, J., HERMANS, J., BROEK, K. van den. (1974) *A stepwise discriminant analysis program using density estimation*. In COMPSTAT, ed. G. Bruckmann, Pyshica Verlag,Viena, p. 101 − 110.

[10] PRESS,S.J. (1982) *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. 2nd ed. New York: Krieger.

[11] SCOTT, D. (1992) *Multivariate Density Estimation: Theory, Practice and Visualization*. [S.l.]: New York: John Wiley.

[12] SILVERMAN, B. (1986) *Density estimation for statistics and data analysis*. [S.l.]: London: Chapman & Hall.

[13] WAND, M., JONES, M. (1994)*Multivariate plug-in bandwidth selection.* Computational Statistics, v.9, p. 97 − 116.

[14] ZHANG, X.; KING, M.; HYNDMAN, R. (2006) *A bayesian approach to bandwidth selection for multivariate kernel density estimation*. Computational Statistics Data Analysis,v. 50, p. 3009 − 3031.