Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

Estudo comparativo entre os métodos de Rosenblatt-Parzen e Grenander na estimação de densidades

DISSERTAÇÃO DE MESTRADO

Autor: Fernando Lucambio Pérez
Orientador: Dr. Mauro S. de F. Marques

1998

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP

Lucambio Pérez, Fernando

L962e Estudo comparativo entre os métodos de Rosenblatt-Parzen e Grenander na estimação de densidades / Fernando Lucambio Pérez -- Campinas, [S.P. :s.n.], 1998.

Orientador : Mauro S. de F. Marques

Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Teoria da estimação. 2. Estatística não paramétrica. 3. Probabilidades. 4. Amostragem (Estatística). I. Marques, Mauro Sergio de Freitas. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Estudo comparativo entre os métodos de Rosenblatt-Parzen e Grenander na estimação de densidades

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por FERNANDO LUCAMBIO PÉREZ e aprovada pela comissão julgadora.

Campinas, 1 de Junho de 1998

Prof. Dr. Mauro S. de Freitas Marques

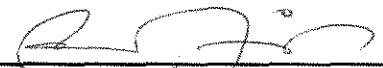
Orientador

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica, UNICAMP, como requisito parcial para a obtenção do Título de MESTRE em ESTATÍSTICA.

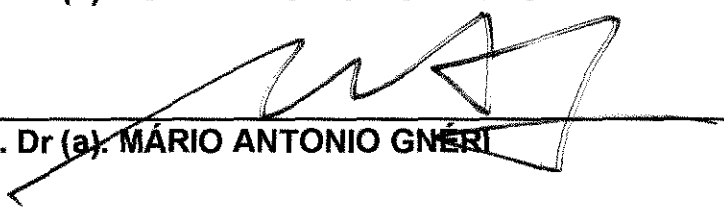
Dissertação de Mestrado defendida e aprovada em 07 de maio de 1998

pela Banca Examinadora composta pelos Profs. Drs.

Prof (a). Dr (a). MAURO SÉRGIO DE FREITAS MARQUES



Prof (a). Dr (a). JOSÉ ANTONIO CORDEIRO



Prof (a). Dr (a). MÁRIO ANTONIO GNÉRI

Índice

Resumo

Desde 1890 diferentes formas de estimar uma função densidade de probabilidade têm sido propostas. Uma destas é devida a Pearson entre 1890 e 1900, e é obtida como solução de uma equação diferencial (Johnson, N. & Kotz, S., 1988).

A partir de 1956 os métodos de estimação de funções de densidade de probabilidade não paramétricos têm-se consolidado como uma alternativa sofisticada ao tratamento tradicional de estudar conjuntos de dados. Esta alternativa baseia-se na possibilidade de analisar os dados sem assumir um comportamento distribucional específico.

Sobre o problema da estimação de funções de densidade de probabilidade trata o Capítulo I desta dissertação. Descrevemos também de maneira resumida algumas das propostas para obter estes estimadores e definimos propriedades estatísticas que serão estudadas nas diferentes situações consideradas.

O Capítulo II dedica-se ao estudo de duas propostas de estimadores da função de densidade. O primeiro estimador estudado é o de Rosenblatt-Parzen. As primeiras idéias deste estimador devem-se a Rosenblatt(1956), idéias posteriormente generalizadas por Parzen(1962), obtendo-se o atualmente conhecido como estimador de Rosenblatt-Parzen ou *"kernel"*. A seguir estuda-se um caso particular do estimador proposto por Grenander(1981). O estimador obtido segundo esta metodologia é conhecido como estimador de Grenander ou *"sieves"* de convolução.

Geman & Hwang(1982) mostraram a forma do estimador de Grenander quando a densidade gaussiana é utilizada, na convolução, como a função núcleo. No Capítulo III estudamos como obter a forma do estimador *"sieves"* de convolução em situações mais gerais de duas maneiras diferentes. Uma destas maneiras é uma generalização das idéias de Geman & Hwang(1982) e a outra utiliza o modelo de dados incompletos. Estes resultados constituem a proposta teórica mais importante.

Como a forma dos estimadores de Grenander obtidos através de convoluções pode ser vista como um modelo de mistura finita de densidades, realizamos no Capítulo IV um estudo do algoritmo EM para o caso particular destes modelos. Nele, apresentamos a teoria geral dos modelos de mistura de densidade e provamos que é possível utilizar o algoritmo EM na estimação de densidades segundo a proposta de Grenander no caso de convoluções, exemplificando este algoritmo quando é utilizada na convolução a densidade gaussiana.

Finalmente, comparamos a performance do estimador de Grenander em relação ao estimador de Rosenblatt-Parzen, através de dados simulados para diferentes funções de densidade. Este estudo constitui o Capítulo V.

Capítulo I

A estimação de funções de densidade de probabilidade

A função de densidade de probabilidade da variável aleatória X é um conceito fundamental na Estatística. Esta é uma função real mensurável e não negativa satisfazendo

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

A necessidade de utilizar estimadores da função f aparece freqüentemente em situações tais como:

i) Análise exploratória, onde se descrevem aspectos como a multimodalidade, o comportamento nas caudas e a simetria, onde a amostra pode ser obtida experimentalmente ou gerada através de simulações, pois o gráfico resume convenientemente a informação relativa à forma da distribuição da amostra.

ii) Análise confirmatória, para tomada de decisões através de diferentes métodos, tais como: análise discriminante não paramétrica, análise de clusters, testes para a moda, dentre outros.

Alguns pesquisadores como Silverman(1986), Tapia & Thompson(1990), Wand & Jones(1995) exemplificam diferentes formas de estimar a função de densidade baseados em conjuntos de dados reais. No entanto, no nosso trabalho utilizaremos dados simulados para mostrar a performance dos estimadores a serem considerados.

Para exemplificar o estimador mais simples de uma função de densidade de probabilidade, chamado de histograma, utilizaremos uma amostra de tamanho 50 da densidade χ^2 com 6 graus de liberdade.

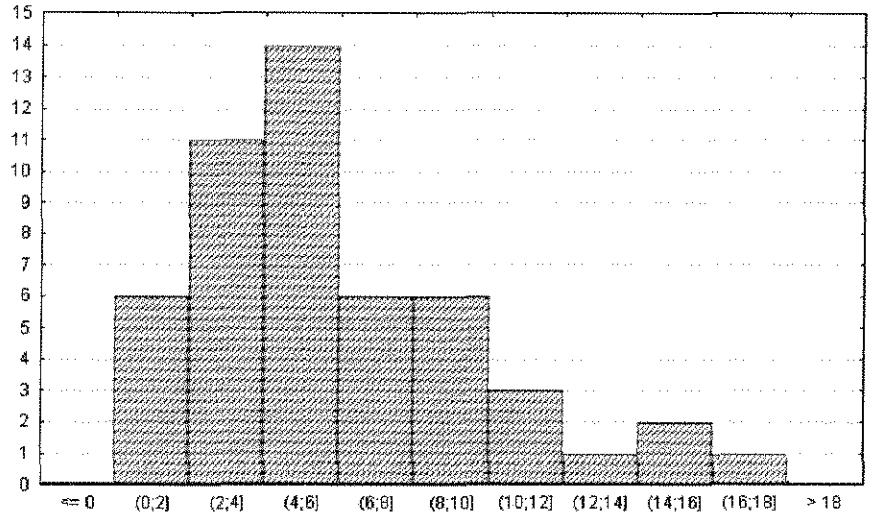
Robertson(1967) provou que, dados os intervalos $I_1,...,I_k$, o histograma $\hat{f}$ é um estimador de máxima verossimilhança dentre os estimadores expressados como funções simples e semi-contínuas superiormente, isto se o fecho de cada intervalo contém duas ou mais observações, veja também Wegman(1975).

Este estimador é calculado segundo a expressão

$$\hat{f}(x) = \frac{n_i}{n|I_i|},$$

para todo $x \in I_i$ e para cada $i=1,...,k$. Na expressão anterior, $|I_i|$ representa o comprimento do intervalo i e n_i o número de elementos na amostra em I_i.

O seguinte gráfico mostra a forma do histograma, segundo a proposta de Robertson(1967).



Pode-se observar que este estimador tem duas limitações importantes: a dependência do comprimento do intervalo e o fato de o histograma não constituir uma função contínua. A primeira destas limitações foi amplamente estudada por Wegman(1975). Ele provou que os pontos extremos de cada intervalo I_k devem ser coincidentes com observações e que, se o número mínimo de observações em cada intervalo m aumenta, conforme aumenta o tamanho da amostra, o estimador $\hat{f}$ é consistente.

A segunda limitação importante do histograma, isto é, o fato de ele não constituir uma função contínua, incentivou diversos estudos na procura de estimadores contínuos, como, por exemplo, Parzen(1962), Grenander(1981) dentre outros. Neste trabalho estudaremos fundamentalmente estimadores contínuos da função de densidade de probabilidade.

1.1 Alguns métodos de estimação

Um método amplamente utilizado em Estatística para obter estimadores é o método de máxima verossimilhança. Na situação específica da estimação de funções de densidade de probabilidade este método é descrito da seguinte forma.

Suponhamos que $X_1, \ldots, X_n$ seja uma amostra aleatória da variável X com função de

densidade contínua f_0, desconhecida. Definimos a verossimilhança de uma densidade qualquer f como

$$L(f)=\prod_{i=1}^{n} f(X_i). \qquad (1.2)$$

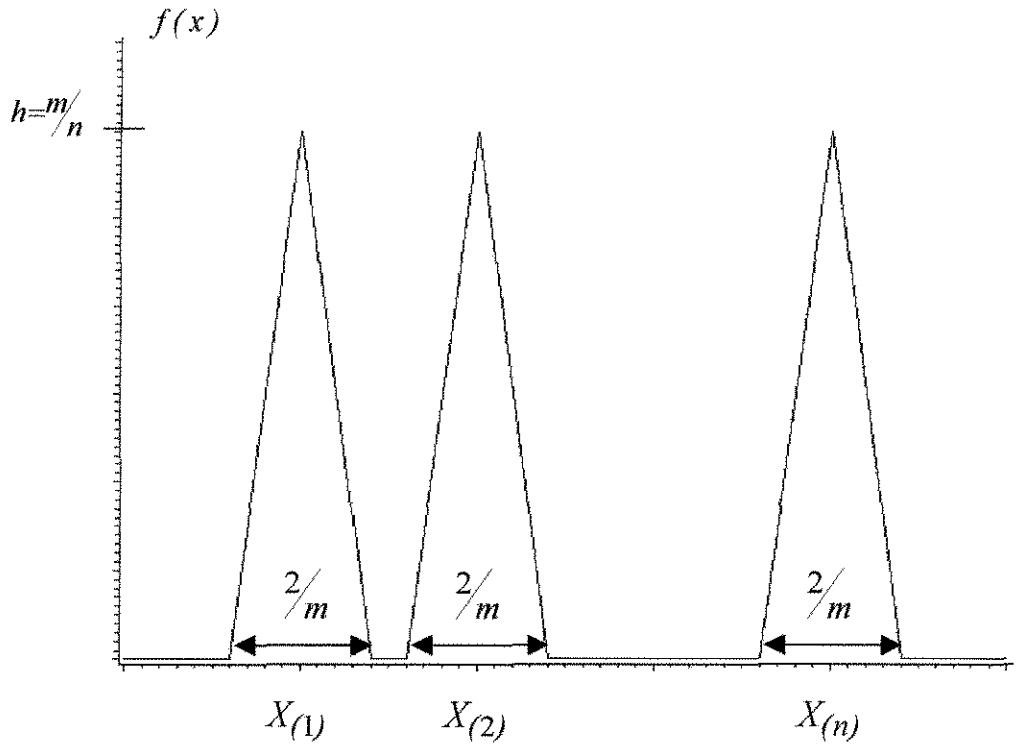
O problema de otimização que se tenta solucionar é

$$\text{maximizar } L(f),$$

onde

$$\int_R f(t)dt = 1 \text{ e } f \geq 0, \text{ real e mensurável.}$$

Este problema não tem solução. Por exemplo, numa amostra $X_1, \ldots, X_n$, seja a função de densidade f da forma mostrada no seguinte gráfico.



Observemos que, a medida em que o comprimento $2/m$ de cada intervalo tende a zero (quando $m \to \infty$) a altura $h = m/n$, tende ao ∞. Desta forma a verossimilhança $L(f) \to \infty$.

Uma forma de conseguir que a função de vesossimilhança $L(f)$ tenha máximo é penalizá-la. Para isto define-se um funcional integrável $\Phi: H \to R$ (H aqui é um espaço de Hilbert) e dada uma amostra $X_1,\ldots,X_n$, a verossimilhança penalizada por Φ de $f \in H$ é definida como

$$\hat{L}(f) = \prod_{i=1}^{n} f(X_i) exp(-\Phi(f)). \qquad (1.4)$$

O estimador de máxima verossimilhança penalizado é aquele obtido como solução do problema:

$$\text{maximizar } \hat{L}(f),$$

$$\text{restrito a } f \in H, \int_a^b f(t)dt = 1 \quad \text{e} \quad f(t) \geq 0 \ \forall t \in (a,b).$$

Dependendo da escolha do espaço H e da função Φ, pode ser gerada uma grande variedade destes estimadores. Por exemplo, De Mostricher *et al*(1975) propuseram maximizar (1.4) escolhendo H como um espaço de Sobolev[1] restrito. Nesta situação, a solução do problema (1.4) existe e é única, veja Tapia & Thompson(1990). Também foi proposto um algoritmo numérico para obter este estimador.

Good & Gaskins(1971) propuseram duas alternativas para escolher a função Φ, a primeira destas alternativas é chamada de primeiro estimador de Good & Gaskins, definido como

$$\Phi(f) = \alpha \int_{-\infty}^{\infty} \frac{f'(t)^2}{f(t)} dt, \ \alpha > 0.$$

Para obter o chamado segundo estimador de Good & Gaskins, eles definiram

$$\Phi(f) = \alpha \int_{-\infty}^{\infty} f'(t)^2 dt + \beta \int_{-\infty}^{\infty} f''(t)^2 dt, \ \alpha \geq 0, \beta > 0.$$

A função f deve ser contínua, de quadrado integrável e satisfazer $\int f(\log f)^2 dx < \infty$. Pode-se ver em Tapia & Thompson(1990) um estudo detalhado da existência e unicidade da solução do problema (1.4) para cada um dos estimadores de Good & Gaskins.

---

[1] Este espaço é definido como $H_0^s(a,b) = \left\{ f : f^{(j)} \in L^2(a,b)\ j = 1,\cdots,s \text{ e } f^{(j)}(a) = f^{(j)}(b)\ j = 0,\cdots,s-1 \right\}$ e produto escalar $< f,g >= \int_a^b f^{(s)}(t) g^{(s)}(t) dt$. Aqui $f^{(j)}$ é a j-ésima derivada da função f.

Outra forma de maximizar $L(f)$ foi proposta por Grenander(1981). A idéia nesta nova situação é restringir o espaço H, o espaço no qual se faz a maximização, da seguinte forma: para cada $m > 0$, escolhe-se o subespaço $S_m \subset H$ onde o estimador de máxima verossimilhança da densidade f exista. Fazendo estes subespaços S_m crescerem apropriadamente com o tamanho da amostra, consegue-se que o estimador achado em cada S_m convirja à função de densidade que originou a amostra. Os estimadores assim achados se chamam de *"sieves"*.

1.2 Propriedades assintóticas

Fryer(1976), por exemplo, estuda os estimadores da função de densidade mantendo finito o tamanho da amostra em situações específicas. No entanto, na maioria dos trabalhos as propriedades estatísticas destes estimadores são provadas assintoticamente no tamanho de amostra.

De maneira natural define-se o estimador $\hat{f}$ como não-tendencioso see $E_f[\hat{f}(x)] = f(x)$, para todo x real. No entanto, esta definição não tem utilidade na estimação de funções de densidade de probabilidade pois não existe estimador não-tendencioso $\hat{f}$ para toda função de densidade contínua (Rosenblatt, 1956). Esta propriedade importante de todo estimador se define então da seguinte forma.

Definição 1.1: A seqüência de estimadores $\{\hat{f}_n\}$, da função de densidade f, se diz assintóticamente não-tendenciosa see

$$\lim_{n\to\infty} E_f[\hat{f}_n(x)] = f(x),$$

para todo x real.

Em geral, juntamente com a propriedade de não-tendenciosidade do estimador $\hat{f}$, estuda-se também a sua consistência. Dependendo da forma de convergência escolhida, observam-se diferentes formas de consistência de $\hat{f}$. Se $\hat{f}(x) \to f(x)$ em probabilidade para todo x real, afirma-se que $\hat{f}$ é fracamente consistente. Se a convergência acontece quase certamente, $\hat{f}$ é fortemente consistente. Outros tipos de convergência dependem também do que se entenda por critério de erro.

Por exemplo, se a função f é suposta de quadrado integrável o ajuste obtido por $\hat{f}$ é medido segundo o erro quadrático médio (*EQM*). Se $EQM(x) \to 0$ para todo x real, quando $n \to \infty$, diz-se que $\hat{f}$ é estimador consistente fracamente de f em média quadrática.

Outros critérios importantes medem como a função $\hat{f}$, como um todo, estima f. Um deles é o erro quadrático médio integral(*EQMI*), cuja definição é dada abaixo:

$$EQMI = \int_{-\infty}^{\infty} EQM(\hat{f}(x))dx. \qquad (1.9)$$

Um problema com estas formas de medir o ajuste obtido com $\hat{f}$ é que o estudo nas caudas das densidades têm pouca importância. Estas e outras objeções ao *EQM* , *EQMI* e outras medidas de ajuste encontram-se em Donoho & Johstone(1989).

Numa tentativa de avaliar globalmente o ajuste do estimador $\hat{f}$ diminuindo a limitação das medidas de ajuste mencionadas, Devroye & Gyorfi(1985) e Devroye(1987) propuseram o erro absoluto integral(*EAI*), definido como

$$EAI = \int_{-\infty}^{\infty} |\hat{f}(x) - f(x)|dx. \qquad (1.10)$$

Esta medida possui como propriedades a invariância a transformações monótonas, ser bem definida em espaços normados e limitada em $0 \leq EAI \leq 2$, além de ser útil na definição da consistência do estimador.

Definição 1.2: O estimador $\hat{f}$, da função de densidade f se diz consistente see

$$\int_{-\infty}^{\infty} |\hat{f}_n(x) - f(x)|dx \to 0, \qquad (1.11)$$

em probabilidade quando $n \to \infty$.

Se a convergência a zero do *EAI* for quase certa, $\hat{f}$ é dita fortemente consistente. Esta forma de medir a consistência do estimador $\hat{f}$ está relacionada com a distância de Kullback-Leibler, definida em Kullback & Leibler(1951). Tem-se observado que a obtenção de resultados de consistência segundo o *EAI* é muito mais complicada que aqueles esforços necessários para obter resultados semelhantes segundo o *EQM* e suas modificações.

Capítulo II

Estimadores Rosenblatt-Parzen e Grenander da função de densidade

Neste Capítulo serão tratados resumidamente dois estimadores para a função de densidade. O primeiro deles foi proposto originalmente por Rosenblatt(1956) e generalizado por Parzen(1962) e é conhecido como estimador de Rosenblatt-Parzen, também chamado estimador núcleo, este estimador será denotado por $\tilde{f}$. O segundo estimador a ser considerado, proposto por Grenander(1981), é conhecido por estimador de *"sieves"* e será denotado por $\hat{f}$.

Na primeira parte do Capítulo, na Seção 2.1, estudaremos o estimador da função de densidade proposto por Rosenblatt. Posteriormente, na Seção 2.2 justificaremos e definiremos a extensão feita por Parzen. Propriedades serão consideradas nas seções seguintes e, a partir dessas, trataremos do problema da estimação do parâmetro de alisamento. A definição e exemplos de estimadores de *"sieves"* serão apresentados nas Seções 2.8 e 2.9, e suas propriedades estudadas na Seção 2.10.

2.1 Introdução

A partir dos trabalhos de Rosenblatt(1956) e Parzen(1962), o estimador núcleo tem sido bastante estudado. Veja, por exemplo, Silverman(1986) e Wand & Jones(1995). Fatores que contribuíram para esta ampla utilização são a simplicidade e as boas propriedades de $\tilde{f}$.

Seja $X_1,...,X_n$ uma amostra aleatória de uma variável aleatória absolutamente contínua com função de distribuição de probabilidade F e densidade f. Um possível estimador do valor $F(x)$ é dado pela função de distribuição empírica, definida como

$$\hat{F}_n(x) = \frac{1}{n}\{\text{ número de observações na amostra} \le x\} = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \le x). \quad (2.1)$$

Vejamos algumas propriedades importantes deste estimador . Note que $n\hat{F}(x) \sim B(n,F(x))$. Com isto

$$i)\ E(n\hat{F}_n(x)) = nF(x) \quad \Rightarrow \quad E(\hat{F}_n(x)) = F(x),$$

ii) $Var(n\hat{F}_n(x)) = nF(x)(1-F(x)) \quad \Rightarrow Var(\hat{F}_n(x)) = \frac{1}{n} F(x)(1-F(x))$.

Portanto, $\hat{F}_n(x)$ é um estimador não-tendencioso de $F(x)$ e, utilizando a desigualdade de Tchebychev, temos que

$$P\left(|\hat{F}_n(x) - F(x)| > \varepsilon\right) \leq \frac{Var(\hat{F}_n(x))}{\varepsilon^2} = \frac{F(x)(1-F(x))}{n\varepsilon^2} \underset{n\to\infty}{\to} 0,$$

implicando na consistência fraca de $\hat{F}_n(x)$.

Além das propriedades mencionadas, podemos observar que o estimador definido em (2.1) é o estimador de máxima verossimilhança da função de distribuição de probabilidade F no ponto x. Segundo o Teorema de Glivenko-Cantelli, veja, por exemplo, Rao(1973) o estimador $\hat{F}_n(x)$ converge uniformemente em probabilidade a $F(x)$, isto é, para todo $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left\{ \sup_{-\infty<x<\infty} |\hat{F}_n(x) - F(x)| > \varepsilon \right\} = 0.$$

$\hat{F}_n(\cdot)$ é, portanto, um "bom" estimador de $F(\cdot)$. A derivada de $\hat{F}_n(\cdot)$ seria um estimador lógico de $f(\cdot)$. No entanto, $\hat{F}_n$ é uma função não diferenciável, pois sempre que x assume um dos valores X_k, $k=1,...,n$, a função $\hat{F}_n$ tem salto igual a $\frac{1}{n}$. Rosenblatt(1956) propôs utilizar como estimador de $f(\cdot)$ uma aproximação limite que define um estimador da função de densidade em termos de $\hat{F}_n$, isto é, dado que

$$f(x) = \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h},$$

Rosenblatt propôs como estimador de f a função

$$\tilde{f}_n(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h}, \qquad (2.2)$$

para h um número real positivo pequeno, que será chamado de parâmetro de alisamento, e n grande. O estimador , definido em (2.2) é uma função de densidade para todo $h>0$.

Vejamos a forma do estimador proposto por Rosenblatt numa amostra de tamnaho 50 simulada de uma mistura de funções de densidade gaussianas, nestes gráficos assumiram-se três valores diferentes para o parâmetro de alisamento h:

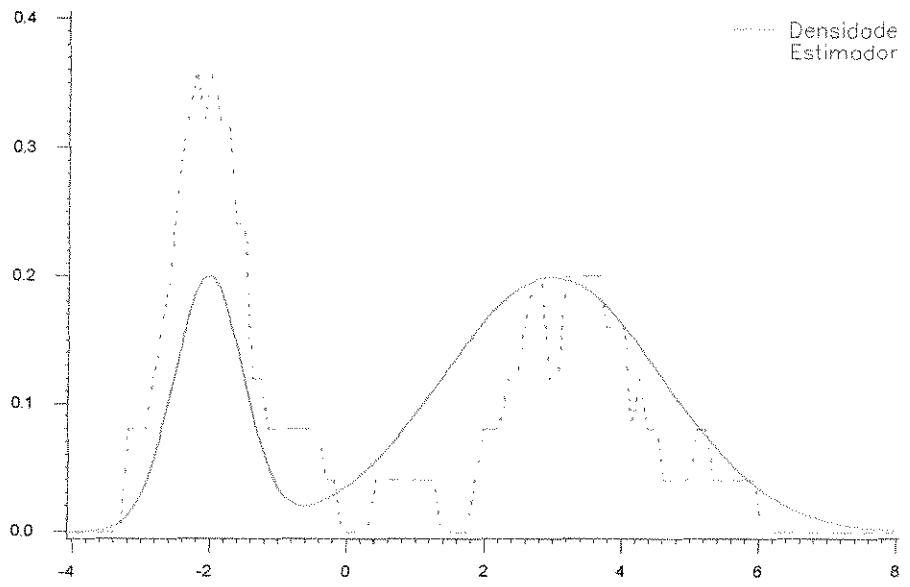


Gráfico 2.1 Estimador de Rosenblatt para a mistura $\frac{1}{2}N(-2, 0.25)+\frac{1}{2}N(3, 2.56)$, parâmetro de alisamento h=0.5.

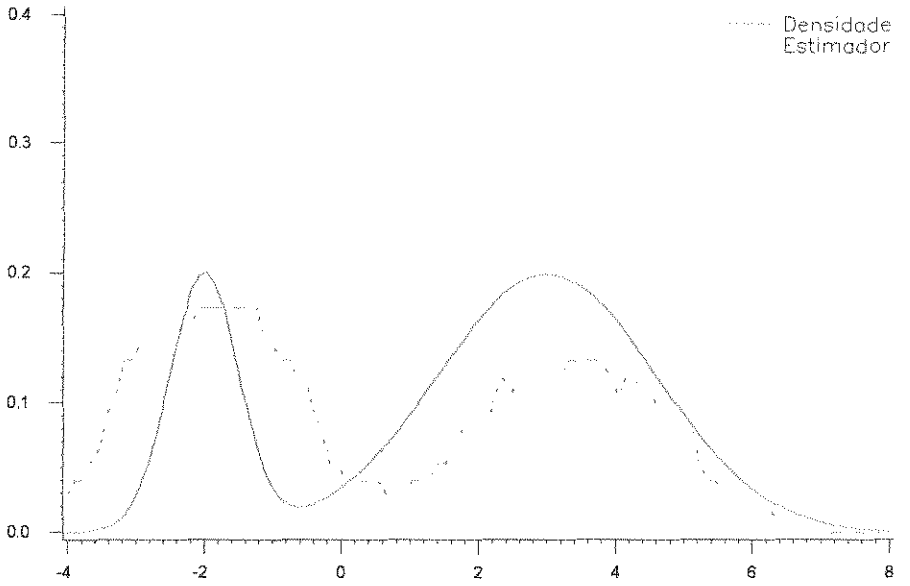


Gráfico 2.2 Estimador de Rosenblatt para a mistura $\frac{1}{2}N(-2, 0.25)+\frac{1}{2}N(3, 2.56)$, parâmetro de alisamento h=1.5.

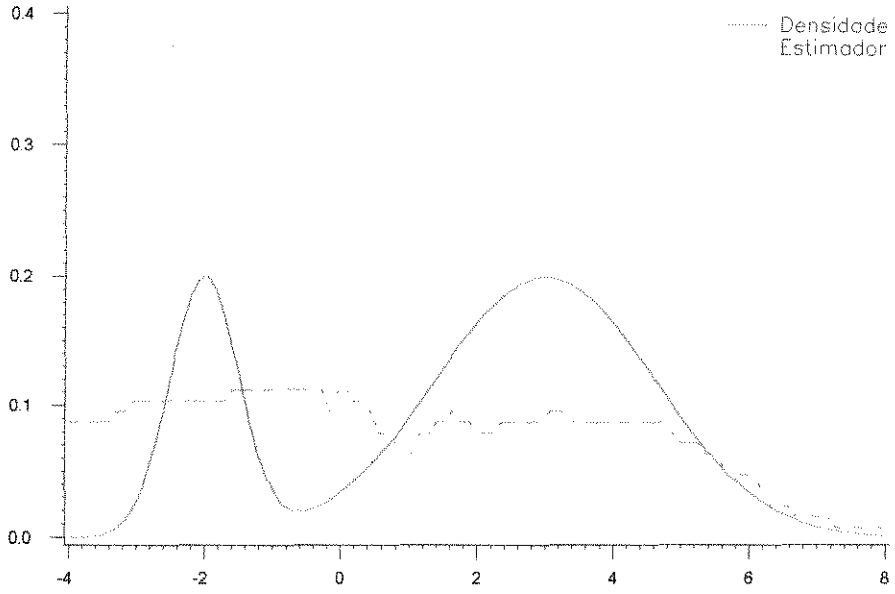


Gráfico 2.3 Estimador de Rosenblatt para a mistura $\frac{1}{2}N(-2, 0.25)+\frac{1}{2}N(3, 2.56)$, parâmetro de alisamento h=2.5.

Na expressão (2.2), o parâmetro h é desconhecido. Nos gráficos anteriores foram escolhidos arbitrariamente os valores de h iguais a 0.5, 1.5 e 2.5. Esses valores mostraram a dependência do estimador de Rosenblatt do parâmetro de alisamento.

Como foi apontado por Rosenblatt, é facil ver que

$$\left(n\hat{F}_n(x-h), n(\hat{F}_n(x+h)-\hat{F}_n(x-h)), n(1-\hat{F}_n(x+h)\right)$$

é um vetor aleatório trinomial com parâmetros

$$\left(n, F(x-h), F(x+h)-F(x-h), 1-F(x+h)\right).$$

A partir deste resultado temos que

$$E\left(\tilde{f}_n(x)\right)=\frac{1}{2nh}E\left(n(\hat{F}_n(x+h)-\hat{F}_n(x-h))\right)=\frac{n(F(x+h)-F(x-h))}{2nh}=$$
$$=\frac{(F(x+h)-F(x-h))}{2h},$$

$$Var\left(\tilde{f}_n(x)\right) = \frac{(F(x+h) - F(x-h) - (F(x+h) - F(x-h))^2)}{4h^2 n}.$$

Estas expressões permitirão o cálculo do erro quadrático médio*(EQM)* de $\tilde{f}_n(x)$, que é dado por

$$EQM(\tilde{f}_n(x)) = E\left((\tilde{f}_n(x) - f(x))^2\right) = Var(\tilde{f}_n(x)) + Vicio^2(\tilde{f}_n(x)).$$

Temos então que

$$EQM(\tilde{f}_n(x)) = \frac{(F(x+h) - F(x-h) - (F(x+h) - F(x-h))^2)}{4h^2 n} +$$

$$+\left(\frac{F(x+h) - F(x-h)}{2h} - f(x)\right)^2. \qquad (2.3)$$

Assumindo que *f*, a densidade desconhecida, seja ao menos três vezes diferenciável, pode ser mostrado que

$$F(x+h) - F(x-h) = 2hf(x) + \frac{h^3}{3} f''(x) + O(h^4)^1.$$

De fato, a expansão em série de Taylor no ponto *x* da função $f(x-ht)$ é

$$f(x-ht) = f(x) - htf'(x) + \frac{h^2 t^2}{2} f''(x) - \frac{h^3 t^3}{6} f'''(x) + O(h^4). \qquad (2.4)$$

Por outro lado, temos que

$$F(x+h) - F(x-h) = \int_{x-h}^{x+h} f(y)dy,$$

e, fazendo *y=x-ht*, obtemos

$$F(x+h) - F(x-h) = \int_{x-h}^{x+h} f(y)dy = \int_{-1}^{1} f(x-ht)dt =$$

[1] Diremos que $f(x) = O(g(x))$ quando $x \to x_0$ se $|f(x)/g(x)|$ é limitado numa vizinhança de x_0.

$$= \int_{-1}^{1} \left(f(x) - htf'(x) + \frac{h^2 t^2}{2} f''(x) - \frac{h^3 t^3}{6} f'''(x) + O(h^4) \right) h dt$$

$$= 2hf(x) + \frac{1}{3} h^3 f''(x) + O(h^4).$$

Podemos escrever agora as expressões dos momentos do estimador $\tilde{f}_n(x)$ em função do parâmetro $f(x)$, ou seja,

$$E\left(\tilde{f}_n(x)\right) = \frac{2hf(x) + \frac{1}{3} h^3 f''(x) + O(h^4)}{2h} = f(x) + \frac{1}{6} h^2 f''(x) + O(h^3),$$

$$Var\left(\tilde{f}_n(x)\right) = \frac{1}{4nh} \left(2f(x) + \frac{1}{3} h^2 f''(x) + O(h^3) - h\left(2f(x) + \frac{1}{3} h^2 f''(x) + O(h^3) \right)^2 \right)^2.$$

Se $h \to 0$, o estimador de Rosenblatt é assintóticamente não-tendencioso. No entanto, esta condição não basta para ser consistente. Podemos observar isto na expressão da variância. Se $h \to 0$, o numerador tende a $f(x)$ noentanto, a variância cresce ao infinito.

Para obter a consistência do estimador teremos que exigir também que $nh \to \infty$ quando $n \to \infty$. No caso do erro quadrático médio, substituindo a diferença $F(x+h) - F(x-h)$ por sua aproximação $2hf(x) + \frac{1}{3} h^3 f''(x) + O(h^4)$, obtida em (2.3), temos que

$$EQM(\tilde{f}_n(x)) = \frac{1}{4h^2 n} \left[2hf(x) + \frac{h^3}{3} f''(x) + O(h^4) - (2hf(x) + \frac{h^3}{3} f''(x) + O(h^4))^2 \right]$$

$$+ \frac{1}{4h^2} \left[2hf(x) + \frac{h^3}{3} f''(x) + O(h^4) - 2hf(x) \right]^2 =$$

$$= \frac{f(x)}{2hn} + \frac{h^4 (f''(x))^2}{36} + \frac{1}{nh} \left\{ \frac{h^2 f''(x)}{12} - \frac{h^2 f''(x) O(h^4)}{6} - \right.$$

$$\left. - \frac{h^3 f(x) f''(x)}{3} - \frac{h^5 (f''(x))^2}{36} - hf(x)^2 - f(x) O(h^4) \right\} +$$

$$O(h^4)\left\{\frac{1}{4h^2n}-\frac{O(h^4)}{4h^2n}+\frac{O(h^4)}{4h^2}+\frac{hf''(x)}{6}\right\}=$$

$$=\frac{f(x)}{2hn}+\frac{h^4}{36}(f''(x))^2+o\left(\frac{1}{hn}+h^4\right)^2. \qquad (2.5)$$

A igualdade anterior é obtida tendo em vista que

$$\frac{1}{nh}\left\{\frac{h^2f''(x)}{12}-\frac{h^2f''(x)O(h^4)}{6}-\frac{h^3f(x)f''(x)}{3}-\frac{h^5(f''(x))^2}{36}-hf(x)^2-f(x)O(h^4)\right\}$$

é uma função de ordem menor que $\frac{1}{nh}$, ou seja, $o\left(\frac{1}{nh}\right)$ e a função

$$O(h^4)\left\{\frac{1}{4h^2n}-\frac{O(h^4)}{4h^2n}+\frac{O(h^4)}{4h^2}+\frac{hf''(x)}{6}\right\}$$

é da mesma ordem que $O(h^4)$.

A expressão (2.5) mostra que $EQM(\tilde{f}_n(x))\to 0$ quando $n\to\infty$. Além disso, fixando f, x e n podemos minimizar (2.5) em relação a h, obtendo

$$h_{opt}=\left(\frac{9}{2}\frac{f(x)}{(f''(x))^2}\right)^{1/5}n^{-1/5} \qquad (2.6)$$

e o correspondente erro quadrático médio mínimo será

$$EQM(\tilde{f}_n(x))=\frac{5}{4}9^{-1/5}2^{-4/5}(f(x))^{4/5}(f''(x))^{2/5}n^{-4/5}. \qquad (2.7)$$

O valor ótimo h_{opt}, apresentado em (2.6), satisfaz as condições exigidas no parâmetro de alisamento h para o estimador de Rosenblatt ser consistente e assintoticamente não tendencioso. Observamos também que (2.6) não é de utilidade prática, já que depende da função de densidade desconhecida f e da sua segunda derivada.

A utilidade de (2.6) é que obtemos o grau de dependência do parâmetro h do tamanho da amostra n, assim como do ponto x no qual se esteja fazendo a estimação. É importante notar

[2] Diremos que $f(x)=o(g(x))$ quando $x\to x_0$ se $|f(x)/g(x)|\to 0$ quando $x\to x_0$.

que além da continuidade da função de densidade f, deve-se exigir continuidade na segunda derivada e, pelo fato de esta estar no denominador, deve-se ter cuidado nos possíveis pontos de inflexão, ou seja, nos pontos onde f'' for zero. Nestes, h_{opt} será qualquer função de n que satisfaça $nh \to \infty$.

2.2 Estimador de Rosenblatt-Parzen

A partir do trabalho de Rosenblatt(1956), foram direcionados os esforços para a obtenção de estimadores da função de densidade contínuos. Observemos que o estimador (2.2) pode ser expresso na forma

$$\tilde{f}_n(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} = \frac{1}{2nh}\sum_{i=1}^{n} 1_{(x-h,x+h]}(X_i). \qquad (2.8)$$

Se definimos

$$K(x) = \tfrac{1}{2} 1_{[-1,1]}(x),$$

o estimador $\tilde{f}_n$ em (2.8) constitui uma integral de Rieman-Stieltjes com respeito a $\hat{F}_n$, isto é,

$$\tilde{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-y}{h}\right) d\hat{F}_n(y). \qquad (2.9)$$

A proposta de Parzen(1962) foi substituir as funções indicadoras na soma em (2.8) por determinadas funções K, chamadas de funções núcleo. Mudando a função K, obtemos da expressão anterior uma grande variedade de estimadores.

O estimador proposto por Parzen será exemplificado nos Gráficos 2.4, 2.5 e 2.6, a seguir, para distintos valores de h. Estes gráficos foram construídos assumindo a função K como

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

e utilizando os mesmos valores de h escolhidos nos Gráficos 2.1, 2.2 e 2.3.

Note que a dependência do estimador $\tilde{f}_n$ do parâmetro de alisamento h permanece.

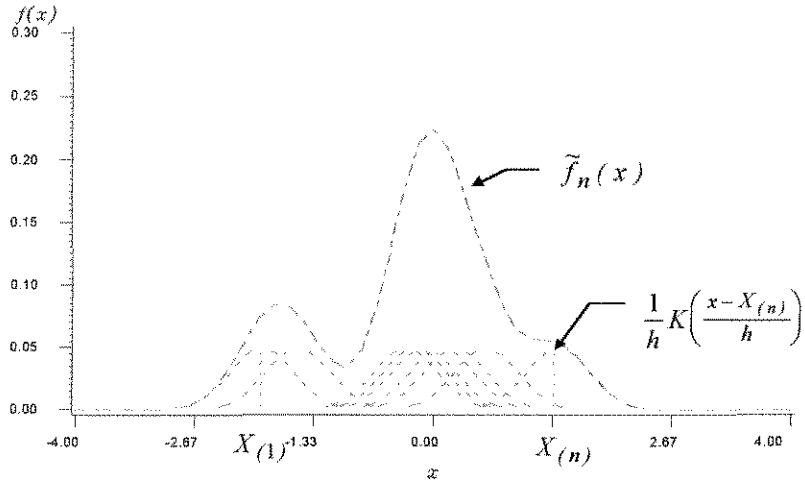


Gráfico 2.4 Estimador de Rosenblatt-Parzen mostrando as funções núcleo individuais, h=0.5.

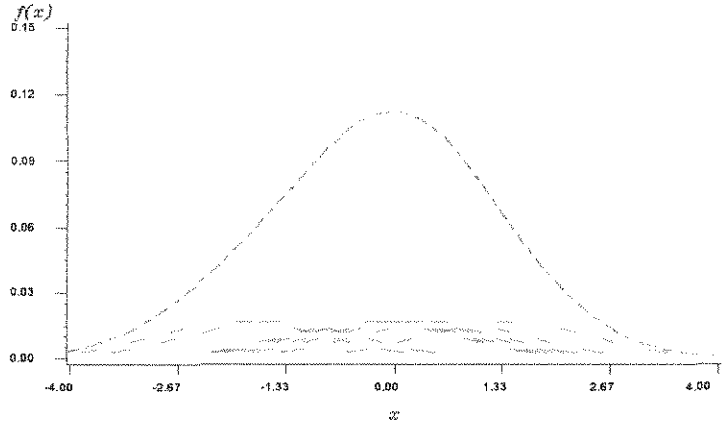


Gráfico 2.5 Estimador de Rosenblatt-Parzen mostrando as funções núcleo individuais, h=1.5.

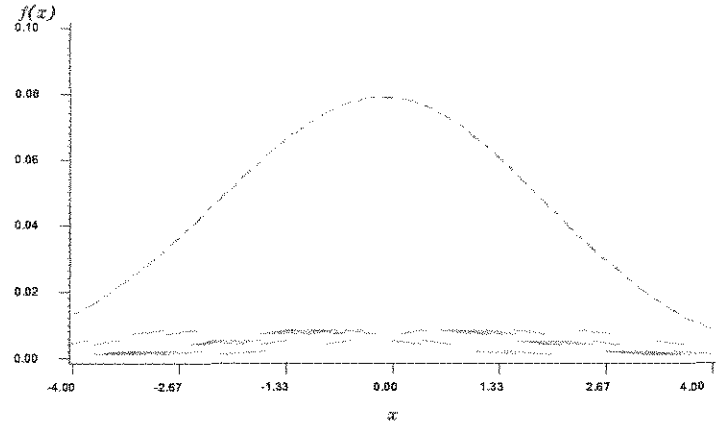


Gráfico 2.6 Estimador de Rosenblatt-Parzen mostrando as funções núcleo individuais, h=2.5.

Definição 2.1: Seja $X_1,...,X_n$ uma amostra aleatória da função de densidade f. Diremos que $\tilde{f}_n$ é o estimador de Rosenblatt-Parzen de f se

$$\tilde{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) d\hat{F}_n(y) \qquad (2.10)$$

onde $h > 0$ e K é uma função real positiva satisfazendo $\int_{-\infty}^{\infty} K(x)dx = 1$.

O estimador de Rosenblatt-Parzen será denotado da mesma forma que o estimador de Rosenblatt. Podemos observar das expressões (2.8), (2.9) e (2.10) que este último constitui uma generalização do estimador de Rosenblatt.

Da definição anterior, obtemos que

$$\int_{-\infty}^{\infty} \tilde{f}_n(x)dx = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(\frac{x - X_i}{h}\right)\frac{1}{h}dx = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K(x)dx = \frac{1}{n}\sum_{i=1}^{n} 1 = 1.$$

Portanto $\tilde{f}_n$ é uma função de densidade de probabilidade. Além disso, se K for uma função contínua e derivável, então $\tilde{f}_n$ também o será.

Direcionaremos nossa atenção somente nas funções núcleo que sejam contínuas e deriváveis e, dada a restrição na definição do estimador de Rosenblatt-Parzen de terem integral um, consideraremos somente funções núcleo que sejam funções de densidade com derivadas contínuas. Além destas, outras condições sobre a função núcleo K e sobre h permitirão obter propriedades estatísticas do $\tilde{f}_n$.

2.3 Propriedades

Como é natural no estudo de estimadores, estamos interessados nos momentos, com este objetivo, observemos que

$$\tilde{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} V_{ni}(x),$$

onde

$$V_{ni}(x) = \frac{1}{h} K\left(\frac{x - X_i}{h}\right),$$

são variáveis independentes identicamente distribuídas. Escrito desta forma, temos que

$$E(\tilde{f}_n(x)) = \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{1}{h} K\left(\frac{x - X_i}{h}\right)\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(V_{ni}) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy, \qquad (2.11)$$

e

$$Var(\tilde{f}_n(x)) = \frac{1}{n^2}\sum_{i=1}^{n} Var\left(\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(V_{ni}) = \frac{1}{n} Var(V_n),$$

onde

$$V_n = \frac{1}{n}\sum_{i=1}^{n} V_{ni}.$$

Então,

$$Var(\tilde{f}_n(x)) = E(V_n^2) - E^2(V_n) =$$

$$= \int_{-\infty}^{\infty} \frac{1}{h^2} K^2\left(\frac{x-y}{h}\right) f(y)dy - \left(\int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy\right)^2. \qquad (2.12)$$

2.3.1 Propriedades de $\hat{f}_n$ para n finito

Dos resultados (2.11) e (2.12), chegamos à conclusão de que os momentos do estimador de Rosenblatt-Parzen se reduzem aos momentos da função núcleo utilizada e independem de n. Isto é importante conceitualmente pois mostra que, tomando uma amostra grande, não será possível reduzir o vício. Torna-se necessário então procurar condições para obter estimadores assintoticamente não-tendenciosos.

No entanto, em situações muito específicas, podem ser obtidas expressões tratáveis e com significado intuitivo dos momentos do estimador $\tilde{f}_n$. Um amplo estudo destas situações encontra-se em Deheuvels(1977). Também Fryer(1976) obteve expressões explícitas para o vício, *EQM* e *EQMI* ao estimar a densidade normal utilizando o estimador de Rosenblatt-Parzen com função núcleo gaussiana mantendo o tamanho de amostra finito.

Segundo a definição do estimador $\tilde{f}_n$ observamos que depois de escolhida a função núcleo K, para se ter completamente especificada a forma do estimador, temos que achar uma expressão para o parâmetro de alisamento h. Neste sentido Deheuvels(1977) utilizou o *EQM*, e a partir dele obteve o valor ótimo de h, achado em (2.6), minimizando o *EQM*.

Voltamos a ressaltar que este valor além de depender da função de densidade desconhecida f, depende também do ponto x no qual se esteja fazendo a estimação, ou seja, o h ótimo em (2.6) constitui uma expressão local. Para contornar esta dificuldade propõe-se utilizar como medida de ajuste o erro quadrado médio integral *EQMI*, definido em (1.9), e que nesta situação será

$$EQMI(\tilde{f}_n) = E\int_{-\infty}^{\infty}(\tilde{f}_n(x) - f(x))^2\,dx .$$

No estudo de Deheuvels(1977), obteveram-se expressões exatas do *EQMI* quando a função núcleo K é simétrica e a função a estimar gaussiana padrão. Nesta situação o *EQMI* será escrito como:

$$EQMI(\tilde{f}_n) = \int_{-\infty}^{\infty}(E(\tilde{f}_n(x)) - f(x))^2\,dx + \int_{-\infty}^{\infty} Var(\tilde{f}_n(x))dx = B_1 + B_2 ,$$

e, segundo Deheuvels(1977), Proposição (2,7) página 19, temos que

$$B_1 = \sum_{r=2}^{\infty} h^{2r}\left\{\frac{(-1)^r}{r!\sqrt{\pi}\,2^{2r+1}}\right\}\left\{\sum_{p=1}^{r-1}\binom{2p}{2r}\int_{-\infty}^{\infty}x^{2p}K(x)dx\int_{-\infty}^{\infty}x^{2r-2p}K(x)dx\right\},$$

e

$$B_2 = \frac{1}{nh}\left[\int_{-\infty}^{\infty}K^2(x)dx - h\sum_{r=0}^{\infty}h^{2r}\left\{\frac{(-1)^r}{r!\sqrt{\pi}\,2^{2r+1}}\right\}\left\{\sum_{p=0}^{r}\binom{2p}{2r}\int_{-\infty}^{\infty}x^{2p}K(x)dx\int_{-\infty}^{\infty}x^{2r-2p}K(x)dx\right\}\right].$$

Vejamos nos exemplos a seguir a forma particular de B_1 e B_2. Estes e outros exemplos acham-se em Deheuvels(1977).

Exemplo 2.1: Seja $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, x real e função núcleo $K(x) = 1$ se $|x| < \frac{1}{2}$. Então:

$$B_1 = \sum_{r=2}^{\infty}h^{2r}\left\{\frac{(-1)^r}{r!(2r+1)\sqrt{\pi}}\right\}\left\{\frac{1}{r+1} - \frac{1}{2^{2r-1}}\right\},\quad B_2 = \frac{1}{nh} - \frac{1}{n\sqrt{\pi}}\sum_{r=0}^{\infty}h^{2r}\frac{(-1)^r}{r!(2r+1)(2r+2)2^{2r}}.$$

Exemplo 2.2: Seja $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, x real e função núcleo $K(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, x real.

Então:

$$B_1 = \frac{1}{2\sqrt{\pi}}\left[\frac{1}{\sqrt{1+h^2}} - \frac{2}{\sqrt{1+\frac{1}{2}h^2}} + 1\right], \quad B_2 = \frac{1}{2nh\sqrt{\pi}}\left[1 - \frac{h}{\sqrt{1+h^2}}\right].$$

Se a função núcleo K for normal de média zero e variância h^2+1, o *EQMI* correspondente ao estimador $\tilde{f}_n$ seria a soma de B_1 e B_2.

Estes exemplos servem para mostrar o difícil trabalho com as expressões exatas do *EQMI* do estimador $\tilde{f}_n$. Além disso, os valores de B_1 e B_2 achados nestas situações particulares podem ser utilizados para minimizar o *EQMI* em relação a h e compará-lo com o valor h obtido minimizando-se expressões aproximadas do *EQMI*. Este estudo será realizado na Seção 2.5.2.

2.3.2 Propriedades assintóticas

Da expressão (2.11) observamos que o estimador de Rosenblatt-Parzen é tendencioso e, além disso, a variância em (2.12) não permite decidir sobre a sua consistência. Aproximações dos momentos em determinadas situações auxiliam no estudo do estimador.

Seguindo os estudos de Parzen(1962), Silverman(1986) e Wand & Jones(1995), vejamos primeiramente condições para que o estimador de Rosenblatt-Parzen seja assintoticamente não tendencioso. Considerando o parâmetro de alisamento h como dependente do tamanho da amostra, isto é $h = h_n$ e

$$\lim_{n\to\infty} h_n = 0,$$

o seguinte resultado sobre o limite da esperança de $\tilde{f}_n$ foi provado em Parzen(1962).

Teorema 2.1: Suponhamos a função núcleo K satisfazendo $\lim_{y\to\infty} yK(y) = 0$ e $\{h_n\}$ uma seqüência positiva convergindo a zero. Então, a esperança do estimador $\tilde{f}_n$ pode ser escrita como

$$E\left(\tilde{f}_n(x)\right) = \frac{1}{h_n}\int_{-\infty}^{\infty} K\left(\frac{y}{h_n}\right) f(x-y)dy, \qquad (2.13)$$

e, em cada ponto de continuidade de f,

$$\lim_{n\to\infty} E\left(\tilde{f}_n(x)\right) = f(x)\int_{-\infty}^{\infty} K(y)dy = f(x). \qquad (2.14)$$

Da expressão (2.14) temos que $\tilde{f}_n$ é assintoticamente não-tendencioso.

Outra propriedade importante geralmente exigida de todo estimador é a consistência. No seguite teorema, Parzen(1962) obtém o limite da variância do estimador de Rosenblatt-Parzen.

Teorema 2.2: Seja a função núcleo K limitada e de quadrado integrável. Neste caso o estimador da função de densidade $\tilde{f}_n$ tem variância satisfazendo

$$lim_{n\to\infty}\, nh_n Var(\tilde{f}_n(x)) = f(x)\int_{-\infty}^{\infty} K^2(y)dy \qquad (2.15)$$

em todo ponto x de continuidade de f, se $h_n \to 0$ quando $n \to \infty$.

Do Teorema 2.2 obtemos que $Var(\tilde{f}_n(x))$ é da ordem $O(1/nh)$. Logo, para que o estimador de Rosenblatt-Parzen seja consistente exigiremos que $nh_n \to \infty$ quando $n \to \infty$. Nessa situação

$$lim_{n\to\infty} Var(\tilde{f}_n(x)) = 0.$$

Sabemos que o EQM de $\tilde{f}_n(x)$ pode ser decomposto em $Var(\tilde{f}_n(x))+\{\text{vício}(\tilde{f}_n(x))\}^2$. Pelo Teorema 2.1, temos que

$$\{\text{vício}(\tilde{f}_n(x))\}^2 \to 0$$

em cada ponto x de continuidade da função de densidade f e, como conseqüência do Teorema 2.2,

$$Var(\tilde{f}_n(x)) \to 0$$

quando $n \to \infty$.

Expressões exatas dos momentos do estimador $\tilde{f}_n$ não são de muita utilidade. Com o objetivo de obter aproximações, Parzen(1962) obtém expressões aproximadas da esperança e variância de V_n. Vejamos isto.

Pela definição de esperança, temos

$$E(V_n^{2+\delta}) = \int_{-\infty}^{\infty} \left(\frac{1}{h_n} K\left(\frac{x-y}{h_n} \right) \right)^{2+\delta} f(y)dy,$$

para algum $\delta>0$. Utilizando a mudança de variáveis $y = x - h_n t$ e a expansão em série de Taylor em torno do ponto x da função $f(x - ht)$ dada em (2.4), obtemos

$$E(V_n^{2+\delta}) = \int_{-\infty}^{\infty} \frac{1}{h_n^{2+\delta}} K^{2+\delta}(t) f(x - h_n t) h_n dt$$

$$= \int_{-\infty}^{\infty} \frac{1}{h_n^{2+\delta}} K^{2+\delta}(t) \left(f(x) - h_n t f'(x) + \frac{h_n^2 t^2}{2} f''(x) - \frac{h_n^3 t^3}{6} f'''(x) + O(h^4) \right) h_n dt$$

.

Desconsiderando os termos nos quais aparecem as derivadas de ordem superior da função f, chegamos a

$$E(V_n^{2+\delta}) \approx \frac{1}{h_n^{1+\delta}} f(x) \int_{-\infty}^{\infty} K^{2+\delta}(t)dt. \qquad (2.16)$$

De maneira similar chegamos à seguinte expressão aproximada para a variância

$$Var(V_n) \approx \frac{1}{h_n} f(x) \int_{-\infty}^{\infty} K^2(y)dy. \qquad (2.17)$$

Com o objetivo de utilizar o Teorema Central do Limite de Lindeberg, verifiquemos se a condição de Lyapunov é satisfeita, isto é vejamos se para algum $\delta>0$

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} E|V_{ni} - E(V_{ni})|^{2+\delta} = 0, \qquad (2.18)$$

onde $s_n^2 = Var(\sum_{i=1}^{n} V_{ni})$.

Neste caso, dadas as condições das variáveis aleatórias V_{ni}, obtemos que $s_n^2 = nVar(V_n)$ e $\sum_{i=1}^{n} E|V_{ni} - E(V_{ni})|^{2+\delta} = nE|V_n - E(V_n)|^{2+\delta}$.

Utilizaremos as aproximações dos momentos em (2.14) para provar a validade do limite. Observemos que

$$E|V_n - E(V_n)|^{2+\delta} = \int_{-\infty}^{\infty} \left| \frac{1}{h_n} K\left(\frac{x-y}{h_n}\right) - \frac{1}{h_n}\int_{-\infty}^{\infty} K\left(\frac{x-y}{h_n}\right) f(y)dy \right|^{2+\delta} f(y)dy .$$

Da expressão da decomposição em série de Taylor em torno do ponto x da função $f(x-ht)$ dada em (2.4), obtemos

$$E|V_n - E(V_n)|^{2+\delta} \approx \frac{f(x)}{h_n^{1+\delta}} \int_{-\infty}^{\infty} \left| K(t) - \int_{-\infty}^{\infty} K\left(\frac{x-y}{h_n}\right) f(y)dy \right|^{2+\delta} dt ,$$

e, multiplicando e dividindo $h_n^{1+\delta}$ em (2.18), chegamos ao seguinte resultado

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} E|V_{ni} - E(V_{ni})|^{2+\delta} = \lim_{n\to\infty} \frac{nE|V_n - E(V_n)|^{2+\delta}}{n^{1+\delta/2} Var^{1+\delta/2}(V_n)}$$

$$= \lim_{n\to\infty} \frac{h_n^{1+\delta} E|V_n - E(V_n)|^{2+\delta}}{(nh_n)^{\delta/2} h_n^{1+\delta/2} Var^{1+\delta/2}(V_n)} .$$

Substituindo as expressões de $E|V_n - E(V_n)|^{2+\delta}$ e $Var(V_n)$, temos

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} E|V_{ni} - E(V_{ni})|^{2+\delta} =$$

$$= \lim_{n\to\infty} \frac{h_n^{1+\delta} \frac{f(x)}{h_n^{1+\delta}} \int_{-\infty}^{\infty} \left| K(t) - \int_{-\infty}^{\infty} K\left(\frac{x-y}{h_n}\right) f(y)dy \right|^{2+\delta} dt}{(nh_n)^{\delta/2} h_n^{1+\delta/2} \left(\frac{1}{h_n} f(x) \int_{-\infty}^{\infty} K^2(y)dy \right)^{1+\delta/2}} .$$

Este último limite pode ser escrito como

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} E|V_{ni} - E(V_{ni})|^{2+\delta} = \lim_{n\to\infty} \frac{C_1}{(nh_n)^{\delta/2} f^{\delta/2}(x) C_2} = 0 ,$$

já que, sendo C_1 e C_2 constantes positivas, para todo $\delta>0$, $\int_{-\infty}^{\infty} K^{2+\delta}(y)dy < \infty$.

Assim, do Teorema Central do Limite de Lindeberg, obtemos que

$$\lim_{n\to\infty} P\left(\frac{\tilde{f}_n(x)-E(\tilde{f}_n(x))}{\sqrt{Var(\tilde{f}_n(x))}} \le c\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} e^{-\frac{y^2}{2}} dy = \Phi(c)$$

para todo real c.

O grau de aproximação para a distribuição normal do estimador $\tilde{f}_n(x)$ pode ser obtido utilizando o resultado do artigo de Esseen(1945), Teorema 2, Capítulo IV:

$$F(x) = \Phi(x) + \frac{\mu_3}{6\sigma^3\sqrt{2\pi n}}(1-x^2)e^{-\frac{x^2}{2}} + o\left(\frac{1}{\sqrt{n}}\right), \qquad (2.19)$$

onde μ_3 representa o terceiro momento central, σ o desvio padrão e Φ a função de distribuição gaussiana.

Utilizaremos (2.19) para escrever a função de distribuição de probabilidade $\tilde{F}$ da variável aleatória $(V_n - E(V_n))/\sqrt{Var(V_n)}$ em termos de Φ e admitiremos válidas as condições para a obtenção desta aproximação diferente de Parzen(1962).

Sob estas condições

$$\tilde{F}(x) \cong \Phi(x) + \frac{E(V_n^3)}{6Var^{3/2}(V_n)\sqrt{2\pi n}}(1-x^2)e^{-\frac{x^2}{2}} + o\left(\frac{1}{\sqrt{n}}\right).$$

Substituindo os momentos pelas suas aproximações, isto é,

$$E(V_n^3) \approx \frac{1}{h_n^2 f(x)} \int_{-\infty}^{\infty} K^3(t)dt$$

e

$$Var^{3/2}(V_n) \approx \left[\frac{1}{h_n} f(x) \int_{-\infty}^{\infty} K^2(t)dt \right]^{3/2},$$

temos

$$\tilde{F}(x) \cong \Phi(x) + \frac{1}{6\sqrt{nh_n f(x)}\sqrt{2\pi n}} \left[\left\{ \int_{-\infty}^{\infty} K^3(y)dy \right\} \Big/ \left\{ \int_{-\infty}^{\infty} K^2(y)dy \right\}^{\frac{3}{2}} \right] (1-x^2)e^{-\frac{x^2}{2}} + o\left(\frac{1}{\sqrt{n}}\right).$$

Parzen(1962), Teorema 3A, demonstrou que o estimador $\tilde{f}_n$ também é uniformemente consistente para a função de densidade f, ou seja, para todo $\varepsilon > 0$

$$P\left[\sup_{-\infty<x<\infty} |\tilde{f}_n(x) - f(x)| < \varepsilon \right] \to 1$$

quando $n \to \infty$.

Como conseqüência dos Teoremas 2.1 e 2.2, se o parâmetro de alisamento h_n for uma função do tamanho de amostra satisfazendo $h_n \to 0$ e $nh_n \to \infty$, quando $n \to \infty$, as propriedades consideradas não restringem a função núcleo K.

2.4 Estimadores de núcleo admissíveis

O conceito de estimadores de núcleo admissíveis foi introduzido por Cline(1988) e serve como um critério na escolha da função núcleo.

Definição 2.2: Diremos que a função núcleo K é admissível, se a única função núcleo K_1, satisfazendo

$$E\int_{-\infty}^{\infty} \left(\frac{1}{nh} \sum_{i=1}^{n} K_1\left(\frac{x-X_i}{h}\right) - f(y) \right)^2 dy \le E\int_{-\infty}^{\infty} \left(\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) - f(y) \right)^2 dy, \quad (2.20)$$

para toda f, é $K_1 = K$. O estimador $\tilde{f}_n$ é admissível se for definido através de funções núcleo admissíveis.

O trabalho de Cline(1988) caracteriza as funções núcleo admissíveis.

Teorema 2.3: A função núcleo K é admissível se e somente se, sua transformada de Fourier[3] ψ é real e satisfaz $0 \leq \psi \leq 1$.

Deste teorema obtemos que para a função núcleo K ser admissível, ela tem que ser simétrica. Outra propriedade destas funções é serem unimodais e que de qualquer rescalamento se obtém novamente uma função núcleo admissível. Isto mostra que a escolha de tais funções é possível independentemente da escolha do parâmetro de alisamento. Varias são as funções de densidade que podem servir como núcleos admissíveis. Entre estas estão a gaussiana, a logística, a exponencial dupla e outras.

2.5 Parâmetro de alisamento *h* e a função núcleo ótimos

Estudamos na Seção anterior um critério para a escolha da função núcleo. Vejamos aqui outro desses critérios induzidos pelo erro quadrático médio integral (*EQMI*). O objetivo de usar o *EQMI* é obter medidas globais da qualidade do ajuste do estimador da função de densidade.

Suponhamos que escolhemos uma função núcleo K admissível. Isto não traz resultados importantes quanto ao parâmetro de alisamento, por esta razão é que se trabalha com o *EQMI*, esta medida de ajuste permite relacionar a escolha da função K com o parâmetro h. Ainda assim, os resultados obtidos relacionando estas funções são de utilidade teórica.

Tentaremos obter um valor ótimo do parâmetro de alisamento h, no sentido de minimizar o *EQMI*, similarmente ao que foi feito no estudo do estimador de Rosenblatt minimizando o *EQM*.

Observamos que o integrando na definição do *EQMI*, segundo foi definido em (1.9), é não negativo, logo será possível inverter a ordem de integração, obtendo-se

$$EQMI(\tilde{f}_n) = \int_{-\infty}^{\infty} EQM(\tilde{f}_n(x))dx = \int_{-\infty}^{\infty} \{E(\tilde{f}_n(x)) - f(x)\}^2 dx + \int_{-\infty}^{\infty} Var(\tilde{f}_n(x))dx,$$

isto é, o *EQMI* é a soma da integral do vício ao quadrado e a integral da variância.

Para a variância de $\tilde{f}_n$ temos, de (2.12) e (2.17), a seguinte expressão aproximada

---

[3] A transformada de Fourier ψ da função núcleo K é definida como $\psi(t) = \int_{-\infty}^{\infty} e^{-itx} K(x)dx$.

$$Var(\tilde{f}_n(x)) \approx \frac{1}{nh_n} f(x) \int_{-\infty}^{\infty} K^2(y)dy .$$

Falta-nos ainda uma expressão aproximada para a integral do vício ao quadrado. Para isto seguiremos o estudo de Parzen(1962), Silverman(1986) e Tapia & Thompson(1990).

Consideremos a Transformada de Fourier ψ da função núcleo K e definamos a função característica empírica de $\hat{F}_n$, isto é,

$$\varphi_n(t) = \int_{-\infty}^{\infty} e^{itx} d\hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} e^{itX_k} ,$$

onde $X_1,\ldots,X_n$ é uma amostra aleatória da densidade f. Temos então, pelo Teorema de Inversão, que

$$\tilde{f}_n(x) = \frac{1}{nh_n} \sum_{k=1}^{n} K\left(\frac{x - X_k}{h_n}\right) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \psi(h_n t)\varphi_n(t)dt .$$

Além disso, se

$$\varphi(t) = E(\varphi_n(t)) = \frac{1}{n} \sum_{k=1}^{n} E(e^{itX_k}),$$

então

$$E(\tilde{f}_n(x)) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \psi(h_n t)\varphi(t)dt .$$

Se $\int_{-\infty}^{\infty} |\varphi(t)|dt < +\infty$, o Teorema de Inversão é válido, com isto, a função de densidade f existe e pode ser escrita como

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-itx} \varphi(t)dt ,$$

obtendo-se

$$vício(\tilde{f}_n(x)) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}[\psi(h_n t) - 1]\varphi(t)dt .$$

Se existir um número positivo r tal que

$$k_r = \lim_{t\to 0}\left[\frac{1-\psi(t)}{|t|^r}\right] \tag{2.21}$$

seja diferente de zero e finito, este será chamado de *expoente característico* da Transformação ψ e k_r de *coeficiente característico*. Desenvolvendo em série de Taylor a função e^{itx}, obtemos

$$k_r = \lim_{t\to 0}\left[\frac{1-\int_{-\infty}^{\infty} e^{itx}K(x)dx}{|t|^r}\right] =$$

$$= \lim_{t\to 0}\frac{1}{|t|^r}\left[1-\int_{-\infty}^{\infty} K(x)dx - \cdots - \frac{(it)^r}{r!}\int_{-\infty}^{\infty} x^r K(x)dx + O(t^{r+1})\right],$$

e, a partir deste resultado,

$$\frac{vício(\tilde{f}_n(x))}{h_n^r} = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\frac{\psi(h_n t)-1}{|h_n t|^r}|t|^r\varphi(t)dt \to k_r f^{(r)}(x), \tag{2.22}$$

onde

$$f^{(r)}(x) = \left.\frac{d^r f}{dx^r}\right|_x = -\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}|t|^r\varphi(t)dt .$$

Temos assim expressões aproximadas para os termos que definem o *EQMI*. Assim,

$$EQMI(\tilde{f}_n(x)) \approx h_n^{2r}k_r^2\int_{-\infty}^{\infty} f^{(r)}(x)^2dx + \frac{1}{nh_n}\int_{-\infty}^{\infty} K^2(x)dx .$$

Dos possíveis valores assumidos por r, o de maior interesse é 2. Nesta situação, devido às suposições sobre a função núcleo K, obtemos de (2.22) que

$$vicio(\tilde{f}_n(x)) \to h_n^2 k_2 f''(x),$$

sendo

$$k_2 = \frac{1}{2}\int_{-\infty}^{\infty} x^2 K(x)dx .$$

Logo, a expressão do *EQMI* será aproximadamente

$$EQMI(\tilde{f}_n(x)) \approx \frac{h_n^4}{4}\left(\int_{-\infty}^{\infty} x^2 K(x)dx\right)^2 \int_{-\infty}^{\infty} f''(x)^2 dx + \frac{1}{nh_n}\int_{-\infty}^{\infty} K^2(x)dx . \qquad (2.23)$$

Observemos na expressão acima que se escolhermos h_n muito pequeno, diminuiremos o vício, mas teremos aumentado o valor da variância. O objetivo então é achar h_{opt}, o h_n ótimo no sentido de ser aquele que reduz ao mínimo o *EQMI*. Para isto Parzen(1962, Lema 4a) obteve que se A, B, α e β são números positivos dados, então o valor x para o qual $Ax^{\alpha} + Bx^{-\beta}$ assume o mínimo é $x = (\beta B / \alpha A)^{1/(\alpha+\beta)}$.

Fazendo uma analogia entre a função minimizada pelo Lema 4a em Parzen(1962) e a expressão em (2.23), obtemos

$$h_{opt} = \frac{1}{n^{1/5}}\left(\int_{-\infty}^{\infty} K^2(x)dx\right)^{1/5}\left(\int_{-\infty}^{\infty} x^2 K(x)dx\right)^{-2/5}\left(\int_{-\infty}^{\infty} f''(x)^2 dx\right)^{-1/5}. \qquad (2.24)$$

Substituindo o h_{opt} em (2.23) obteremos o *EQMI* ótimo. Este será

$$EQMI_{opt}(\tilde{f}_n) \approx \left(\int_{-\infty}^{\infty} K^2(x)dx\right)^{4/5}\left(\int_{-\infty}^{\infty} f''(x)^2 dx\right)^{1/5}\left(\int_{-\infty}^{\infty} x^2 K(x)dx\right)^{2/5} n^{-4/5}.$$

A expressão em (2.24) não tem utilidade prática pelo fato de depender da função de densidade f desconhecida. No entanto, algumas conclusões úteis podem ser obtidas.

Observemos primeiramente que h_{opt} vai convergir a zero quando o tamanho da amostra cresce, na razão de $n^{-1/5}$. Dado que $\int_{-\infty}^{\infty} f''(x)^2 dx$ mede a rapidez das flutuações da densidade f, devemos esperar que valores pequenos de h_n sejam apropriados para as densidades com rápidas flutuações.

Outro fato interessante relacionado com a expressão (2.24) é que a aproximação obtida é boa inclusive para pequenos tamanhos de amostra (n=10). Na tabela a seguir, obtida por Deheuvels(1977), se compara o valor de h que minimiza o *EQMI* exato e o *EQMI* aproximado para as funções núcleo consideradas nos Exemplos 2.1 e 2.2.

n	h_{opt} exato	h_{opt} aproximado
10	2.475	2.326
20	2.126	2.025
50	1.747	1.686
100	1.510	1.468

Tabela 2.1: Valores exatos e aproximados de h_{opt} para as funções núcleo unitária $K(x)=1$ se $|x|<\frac{1}{2}$

n	h_{opt} exato	h_{opt} aproximado
10	0.758	0.668
20	0.642	0.582
50	0.520	0.484
100	0.445	0.422

Tabela 2.2: Valores exatos e aproximados de h_{opt} para as funções núcleo gaussianas.

Uma forma de utilizar a expressão (2.24) na prática é calcular $\int_{-\infty}^{\infty} f''(x)^2 dx$ supondo f a densidade gaussiana com variância σ^2. Neste caso, obtém-se $\int_{-\infty}^{\infty} f''(x)^2 dx = \frac{3}{8}\pi^{-1/2}\sigma^{-5}$ e, se também for utilizada a função núcleo gaussiana, obtemos $h_{opt} = \left(\frac{4}{3}\right)^{1/5} \sigma n^{-1/5}$.

Definido o $EQMI_{opt}$ ótimo em relação a h_n, pensemos agora na otimalidade desta medida de ajuste em relação à função núcleo K.

Observemos que, se $\int_{-\infty}^{\infty} x^2 K(x)dx$ não é idual a 1, dado que estamos supondo K como admissível, poderemos fazer uma transformação de variável adequada de maneira a obter o valor desejado desta integral, então, minimizar o $EQMI_{opt}$ em relação à K, implica minimizar $\int_{-\infty}^{\infty} K^2(x)dx$ restrito a $\int_{-\infty}^{\infty} K(x)dx$ e $\int_{-\infty}^{\infty} x^2 K(x)dx$ serem iguais a 1. Num contexto diferente, Hodges & Lehmann(1956) mostraram que a solução deste problema é a função

$$K_E(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right) & -\sqrt{5} \le t \le \sqrt{5} \\ 0 & \text{caso contra'rio} \end{cases}, \qquad (2.25)$$

conhecida como função núcleo de Epanechnikov(1969).

Esta função servirá então como padrão de eficiência do núcleo do estimador de Rosenblatt-Parzen.

Definição 2.3: A *eficiência(eff)* para uma função núcleo K é proporcional à razão dos valores de $EQMI$ ótimos obtidos segundo as funções núcleo de Epanechnikov e K, isto é

$$eff(K) = \left\{EQMI_{opt}(K_E) / EQMI_{opt}(K)\right\}^{5/4}. \qquad (2.26)$$

Observemos que, para a utilização prática desta medida, a expressão em (2.26) será escrita como

$$eff(K) = \frac{3}{5\sqrt{5}}\left\{\int_{-\infty}^{\infty} x^2 K(x)dx\right\}^{-1/2}\left\{\int_{-\infty}^{\infty} K^2(x)dx\right\}^{-1}.$$

Silverman(1986) e Wand & Jones(1995) incluem tabelas apresentando a eficiência de diferentes funções núcleo. Destas obtém-se como conclusão que dados os valores muito próximos a um para todas elas, este não constitui um fator determinante na escolha da função núcleo a ser utilizada. Por exemplo, a eficiência da função núcleo uniforme no intervalo $(-1/2, 1/2)$ é 0.9295 e a eficiência da função núcleo gaussiana é 0.9512. Logo a utilização da função gaussiana como núcleo está bem justificada.

2.6 Estimação do parâmetro de alisamento *h*

Muito é conhecido a cerca da relação entre a razão de convergência de $\hat{f}_n$ para f e o grau de dependência assintótica do parâmetro h como função do tamanho de amostra n. No entanto, para n fixo, o estimador $\hat{f}_n$ é sensível à escolha de h e não existe uma metodologia geral para a determinação deste parâmetro.

Estudando esta problemática, vários pesquisadores referiram-se à escolha prática de h. Chow *et al*(1983) observaram que "a relação geralmente bem compreendida entre o parâmetro de alisamento assintótico e a razão de convergência do estimador não resulta em um guia prático para a implementação do estimador em dados reais". No mesmo contexto, Silverman(1978) opinou que "observamos uma considerável necessidade de métodos objetivos

de determinação do parâmetro de alisamento apropriado para uma dada amostra". De forma mais genérica, Wahba(1981) afirmou: "o maior problema na estimação de densidades é a escolha do parâmetro de alisamento, que é parte de todos os estimadores da função de densidade ...".

Para exemplificar a situação descrita no parágrafo anterior, geramos uma amostra de tamanho 50 da mistura de densidades gaussianas $\frac{1}{2}N(-2,0.25)+\frac{1}{2}N(3,2.56)$.

Utilizaremos o estimador de Rosenblatt-Parzen com nucleo gaussiano, isto é

$$\tilde{f}_n(x)=\frac{1}{\sqrt{2\pi}}\frac{1}{50h}\sum_{i=1}^{50}e^{-\frac{(x-X_i)^2}{2h^2}}.$$

O valor do parâmetro de alisamento h é arbitrariamente escolhido como h=0.2, 0.5 e 1.5. Nos Gráficos 2.8, 2.9 e 2.10 a seguir apresentam-se a amostra gerada de tamanho 50, a função de densidade e a forma do estimador para os diferentes valores de h.

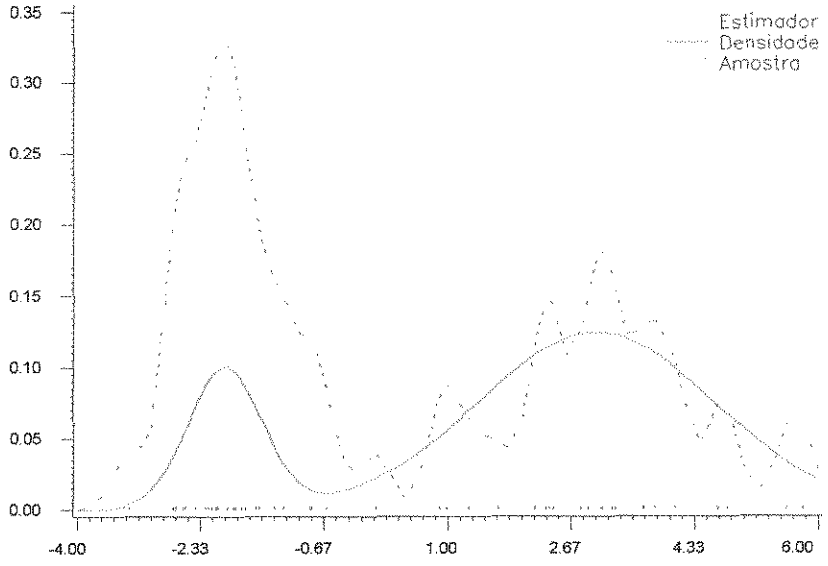


Gráfico 2.8 Estimador de Rosenblatt-Parzen, parâmetro de alisamento h=0.2.

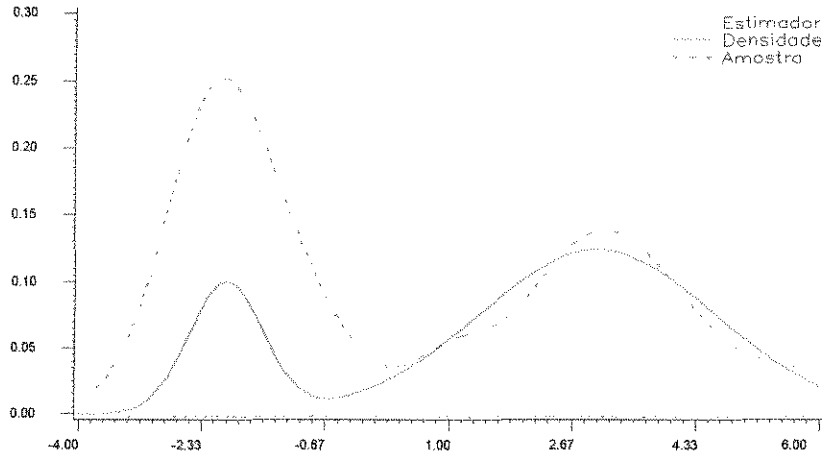


Gráfico 2.9 Estimador de Rosenblatt-Parzen, parâmetro de alisamento h=0.5.

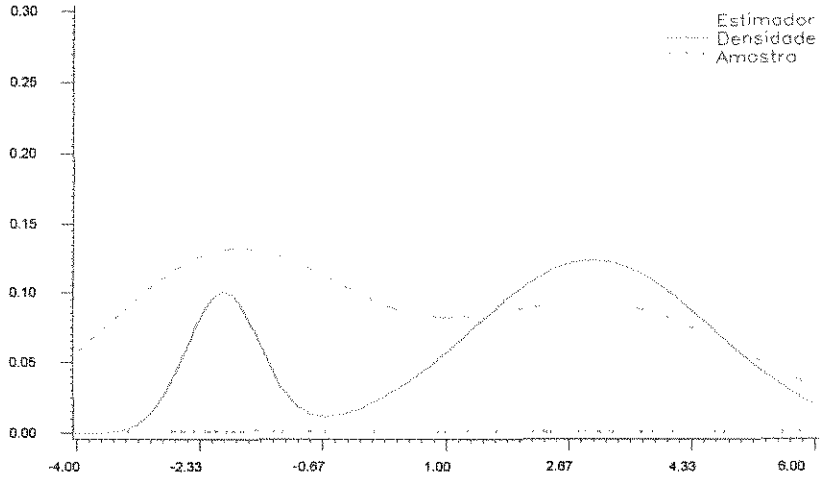
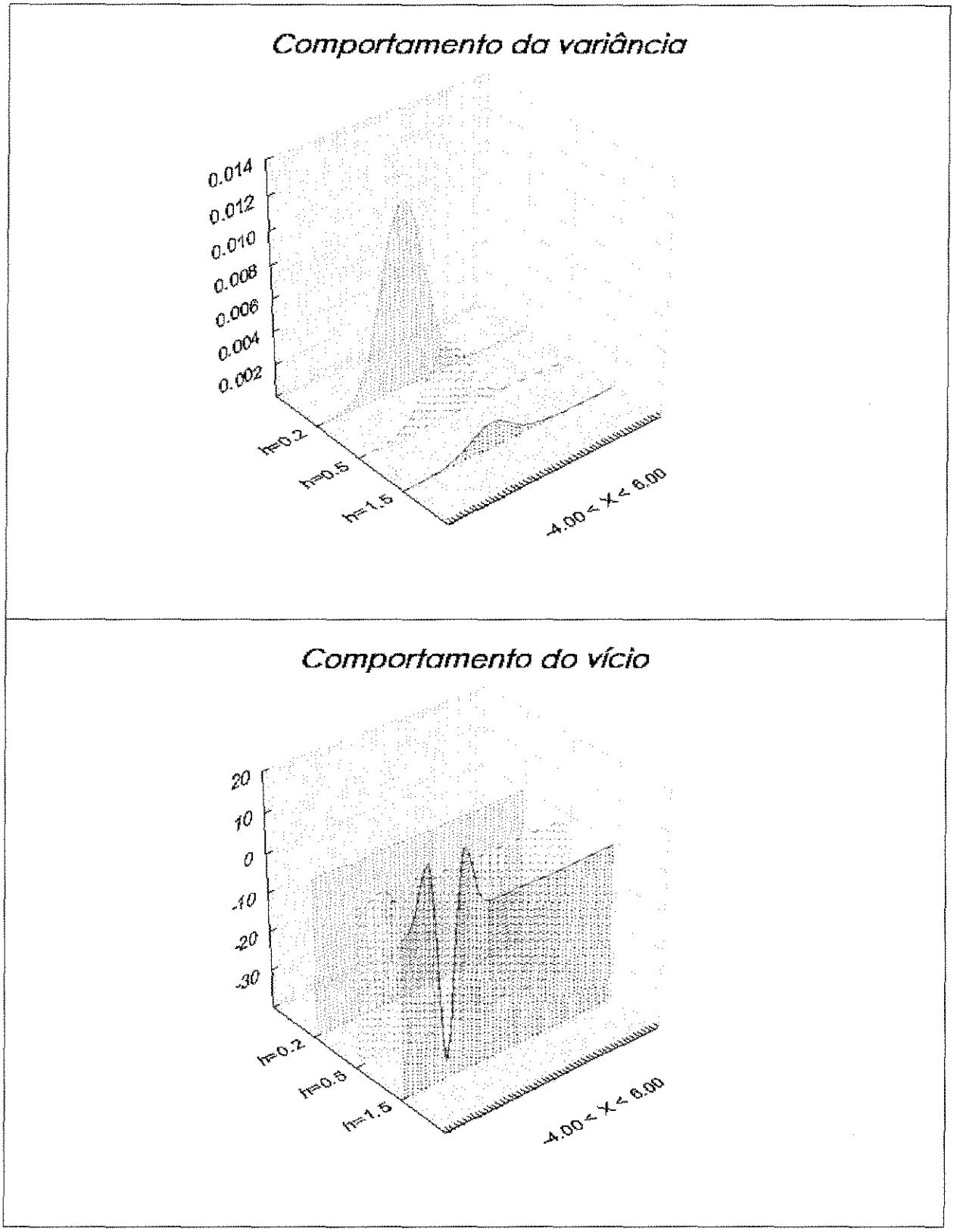


Gráfico 2.10 Estimador de Rosenblatt-Parzen, parâmetro de alisamento h=1.5.

Observemos nos seguintes gráficos o comportamento do vício e da variância para os diferentes valores de h.


Comportamento da variância
0.014
0.012
0.010
0.008
0.006
0.004
0.002
h=0.2
h=0.5
h=1.5
-4.00 < X < 6.00
Comportamento do vício
20
10
0
-10
-20
-30
h=0.2
h=0.5
h=1.5
-4.00 < X < 6.00

Pode-se observar que, quando h=1.5 obtêm-se os menores valores de variância, no entanto o vício é comprometido, situação contrária quando h=0.2. Daí a necessidade de estimar um valor que pondere estas duas formas de medir o ajuste do estimador de Rosenblatt-Parzen.

Para isto devemos considerar o parâmetro de alisamento como função do tamanho de amostra. Das variadas formas propostas na literatura, provavelmente a de maior versatilidade e mais amplamente estudada é o método de validação cruzada.

2.6.1 Estimação de *h* via validação cruzada

A idéia geral deste procedimento de estimação é medir, como função do parâmetro de alisamento, a habilidade do estimador para interpretar ou ajustar os dados observados. O parâmetro de alisamento h será escolhido como aquele que maximiza esta medida. A validação cruzada é obtida eliminando-se uma observação e calculando este estimador na observação eliminada.

Denotemos por $\hat{f}_{n-1}^{i}$ o estimador de Rosenblatt-Parzen calculado depois de eliminar a i-ésima observação, isto é,

$$\hat{f}_{n-1}^{i}(x)=\frac{1}{n-1}\sum_{j\neq i}\frac{1}{h}K\left(\frac{x-X_j}{h}\right). \qquad (2.27)$$

Este estimador não depende de X_i e $\hat{f}_{n-1}^{i}(X_i)$ pode ser escolhido como medida apropriada da contribuição de X_i ao valor de h: se $\hat{f}_{n-1}^{i}(X_i)$ for grande, então pode ser dito que $\hat{f}_{n-1}^{i}(X_i)$ "antecipa" a observação X_i e que h é apropriado. Valores pequenos de $\hat{f}_{n-1}^{i}(X_i)$ sugerem que a observação X_i pode ser esquecida e interpretá-lo como evidência de que o valor de h é impróprio.

Variando i ao longo da amostra, obtemos n de tais medidas de ajuste que então podem ser combinadas na expressão de verossimilhança

$$L(h)=\prod_{i=1}^{n}\hat{f}_{n-1}^{i}(X_i), \qquad (2.28)$$

ou utilizar as $\hat{f}_{n-1}^{i}(X_i)$ numa expressão aproximada ao erro quadrático integral (EQI), dada por

$$EQI_h \approx \int_{-\infty}^{\infty} \tilde{f}_n^2(x)dx - 2\frac{1}{n}\sum_{i=1}^{n} \tilde{f}_{n-1}^{i}(X_i) + \int_{-\infty}^{\infty} f^2(x)dx. \qquad (2.29)$$

Escolhendo $\hat{h}_n$ como aquele que maximiza (2.28) ou minimiza (2.29), obtemos o estimador de $\hat{h}$.

O primeiro destes procedimentos foi proposto por Habbema *et al*(1974) e, separadamente, por Duin(1976). O segundo procedimento foi sugerido por Rudemo(1982) e Bowman(1984).

2.6.2 Validação cruzada mínimos quadrados

Observamos que, minimizar o erro quadrático integral,

$$EQI_h = \int_{-\infty}^{\infty} (\tilde{f}_n(x) - f(x))^2 dx = \int_{-\infty}^{\infty} \tilde{f}_n^2(x)dx - 2\int_{-\infty}^{\infty} \tilde{f}_n(x)f(x)dx + \int_{-\infty}^{\infty} f^2(x)dx,$$

é equivalente a minimizar

$$R_h = \int_{-\infty}^{\infty} \tilde{f}_n^2(x)dx - 2\int_{-\infty}^{\infty} \tilde{f}_n(x)f(x)dx,$$

dado que a última integral em (2.29) não depende de *h*. O princípio básico da validação cruzada mínimos quadrados é construir um estimador de R_h e minimizá-lo em relação a *h* para obter o estimador do parâmetro de alisamento. Vejamos como construir $\tilde{R}_h$, um estimador de R_h.

Definamos

$$\tilde{R}_h = \int_{-\infty}^{\infty} \tilde{f}_n^2(x)dx - 2\frac{1}{n}\sum_{i=1}^{n} \tilde{f}_{n-1}^{i}(X_i). \qquad (2.30)$$

Como

$$E\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{f}_{n-1}^{i}(X_i)\right) = E(\tilde{f}_{n-1}^{n}(X_n)) = E\int_{-\infty}^{\infty} \tilde{f}_{n-1}^{n}(x)f(x)dx = E\int_{-\infty}^{\infty} \tilde{f}_n(x)f(x)dx,$$

temos que $E(\widetilde{R}_h) = E(R_h)$, e com isto, provamos que $\widetilde{R}_h + \int_{-\infty}^{\infty} f^2(x)dx$ é um estimador não tendencioso do erro quadrático médio integral, $MEQI_h = E(EQI_h)$.

Com isto, minimizar $\widetilde{R}_h$ em relação a h seria uma boa alternativa na estimação do parâmetro de alisamento.

Certos estudos, veja por exemplo Park & Marron(1990), mostram que esta forma de estimar o parâmetro de alisamento pode não ser adequada e apresenta severas dificuldades computacionais. Wand & Jones(1995) mostram que o $\hat{h}_n$ achado minimizando (2.30) é altamente variável. Scott & Terrell(1987) propõem uma versão de validação cruzada mínimos quadrados a qual é tendenciosa porém ganha em estabilidade, o $\hat{h}_n$ obtido assim, tem sempre uma variância menor em relação ao resultado obtido quando é minimizado (2.30) diretamente. Outras formas de estimação de h relacionadas com o EQI podem ser achadas em Wand & Jones(1995).

Apesar destas dificuldades, boas propriedades foram provadas para o $\hat{h}_n$. Stone(1984) provou que, assintoticamente, este estimador é a melhor escolha de parâmetro de alisamento no sentido de minimizar o EQI. Bowman(1984) mostrou, via simulações, que $\widetilde{f}_{n,\hat{h}_n}$ apresenta resultados satisfatórios em densidades com caudas compridas. Outras boas propriedades de $\hat{h}_n$ foram provadas em Hall & Marron(1987a, 1987b) e Hall(1982, 1983).

2.6.3 Validação cruzada pseudo - máxima verossimilhança

Maximizando a função de verossimilhança $L(h)$ em (2.28), obtemos o estimador de h por validação cruzada pseudo-máxima verossimilhança. No obstante, esta maximização pode ser drasticamente afetada por valores amostrais quando f está próxima de zero. Exemplos desta problemática foram analisados em Marron(1985).

Marron(1985) considerou varias formas de solução desta dificuldade, que é a maior limitação deste procedimento, conduzindo ao refinamento que será adotado como método de estimação do parâmetro h.

O algoritmo proposto por Marron(1985) para a estimação de h consiste no seguinte:

I. achar um intervalo $[a,b]$ no qual seja conhecido que f é limitada e positiva,

II. definir a verossimilhança estimada $\hat{L}_h = \prod_{i=1}^{n} \tilde{f}^{i}_{n-1,h}(X_i)^{1_{[a,b]}(X_i)} e^{-\rho(X_i)}$,

onde $\rho(x) = \int_a^b \frac{1}{h} K\left(\frac{y-x}{h}\right) dy$,

III. assumir $\hat{h}_n$ como aquele que maximiza $\hat{L}_h$.

Note que a validação cruzada é realizada só naquelas observações que pertençam ao intervalo $[a,b]$. Pela vantagem computacional que oferece este método, será o adotado nas simulações referidas ao estimador de Rosenblatt-Parzen incluídas no Capítulo V.

2.7 Propriedades assintóticas do estimador $\hat{f}_{n,h}$

As propriedades assintóticas do estimador de Rosenblatt-Parzen foram estudadas supondo o parâmetro de alisamento conhecido. Nesta Seção estudaremos propriedades quando o parâmetro de alisamento h for estimado por validação cruzada, em especial, quando o método de estimação do parâmetro h é pseudo - máxima verossimilhança.

Para medir o desempenho do estimador $\hat{f}_{n,h}$ em relação a f, no caso do estimador ser obtido via pseudo - máxima verossimilhança, utilizaremos a seguinte versão de erro quadrático médio, definido como

$$EQM'(h) = \frac{1}{n}\sum_{i=1}^{n} (\hat{f}^{i}_{n-1,h}(X_i) - f(X_i))^2 / f^2(X_i) 1_{[a,b]}(X_i).$$

A justificativa desta proposta de critério de ajuste pode ser vista em Marron(1983). A seguir incluem-se suposições e restrições para o teorema que demonstra o bom desempenho do estimador $\hat{f}_{n,h}$.

Para algum $\delta > 0$, definamos as seqüências $\{\underline{h}_n\}$ e $\{\overline{h}_n\}$ como sendo

$$\underline{h}_n = n^{-1+\delta} \text{ e } \overline{h}_n = n^{-\delta}.$$

O espaço das funções de densidade f é restrito àquelas que sejam limitadas e positivas no intervalo $[a,b]$, suporemos que existam constantes M, M' e $\gamma, \xi > 0$, tais que para todo par x, y

i. $|f(x)-f(y)|\leq M|x-y|^{\gamma}$

ii. $|K(x)-K(y)|\leq M'|x-y|^{\xi}$

Suponhamos também que, ou a variável aleatória X tem algum momento, ou a função núcleo K tem suporte compacto. O seguinte teorema, provado em Marron(1985), é então válido.

Teorema 2.4: Dadas as suposições anteriores e se $\{\hat{h}_n\}$ for uma sequência de pontos de máximos de

$$\hat{L}(h)=\prod_{i=1}^{n}\tilde{f}^{i}_{n-1,h}(X_i)^{1_{[a,b]}(X_i)}e^{-\rho(X_i)},$$

sujeitos à restrição $\hat{h}_n \in [\underline{h}_n, \overline{h}_n]$, então

$$\frac{EQM'(\hat{h}_n)}{\inf\limits_{h\in[\underline{h}_n,\overline{h}_n]} EQM(h)} \to 1 \text{ quase certamente.}$$

Foi mostrado em Marron(1985), via simulações que a restrição $h\in[\underline{h}_n,\overline{h}_n]$ não constitui um problema nas aplicações práticas. Além disso, foi mostrado também que se $\{\hat{h}_n\}$ for uma seqüência qualquer de pontos de máximos de $\hat{L}(h)$, então $\hat{h}_n \to 0$ quase certamente.

Resultados similares ao Teorema 2.4, no caso de estimar h através de mínimos quadrados, foram estabelecidos por Stone(1984), ele utilizou o *EQM*.

2.8 Estimador de Grenander

Nesta Seção nos dedicaremos ao estudo de uma outra proposta de estimador da função de densidade desenvolvido por Grenander(1981) e conhecido como estimador de Grenander. Outra referência importante é Geman & Hwang(1982). A definição deste estimador será apresentada nesta Seção, exemplos na Seção 2.9 e resultados gerais sobre existência e consistência serão estudados na Seção 2.10.

O método de Grenander, também conhecido como *"sieves"*, é uma técnica mediante a qual resultados da estimação paramétrica podem ser aplicados a problemas não paramétricos. A idéia é utilizar, por exemplo, o método de máxima verossimilhança na estimação de densidade, só que, como foi dito no Capítulo I, o espaço paramétrico, nesta situação, o conjunto de todas as funções de densidades de probabilidade, denominado Θ, é muito grande para a existência do

estimador de máxima verossimilhança, ou seja, não existe $g \in \Theta$ que maximize a função de verossimilança

$$L_n(g) = \prod_{i=1}^{n} g(X_i).$$

Como o método de máxima verossimilhança não fornece um estimador adequado neste contexto, foi sugerido por Grenander a maximização da verossimilhança numa seqüência de subconjuntos do espaço paramétrico. Essa seqüência foi chamada de *"sieves"*, dela se exige ser não decrescente com o tamanho da amostra e que a sua união seja densa no espaço paramétrico original.

Assumamos que (Θ, d) seja um espaço métrico de funções de densidade de probabilidade e que a função de densidade de probabilidade desconhecida seja f_0, o parâmetro a estimar. Consideramos também como espaço amostral o espaço de medida $(R, \mathcal{B}, dx)$, com dx medida de Lebesgue e $\mathcal{B}$ os borelianos. No espaço amostral definido temos a família de medidas de probabilidade ou de distribuições de probabilidade $\{P_f : f \in \Theta\}$, com a propriedade de serem identificáveis e absolutamente contínuas em relação a dx, isto é, $(dP_f / dx)(x) = f(x)$. Assim, podemos definir o experimento estatístico como sendo a tripla $(X, \mathcal{B}, \{P_f : f \in \Theta\})$.

Definição 2.4: Sejam $\{S_m\}$ uma seqüência de subconjuntos de Θ, $(X, \mathcal{B}, \{P_f : f \in \Theta\})$ o experimento estatístico e $X_1, \ldots, X_n$ uma amostra aleatória de tamanho n da função de densidade de probabilidade desconhecida f_0. Diremos que $\{S_m\}$ constitui uma seqüência *"sieves"* se satisfizer que: $S_m \subseteq S_{m+1}$, $\cup S_m$ denso em Θ e que o problema de maximização

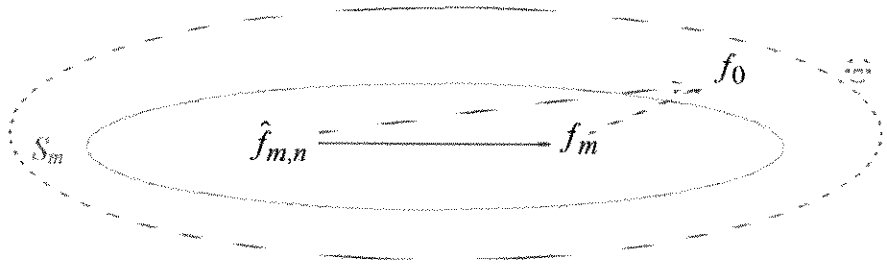
$$\hat{f}_{m,n} = \arg \max_{g \in S_m} L_n(g) = \arg \max_{g \in S_m} \prod_{i=1}^{n} g(X_i),$$

tem solução para todo m e n.

O índice m é chamado de parâmetro do *"sieves"* e deve ter uma relação estreita com o tamanho da amostra.

Desta definição obtemos uma seqüência de estimadores de Grenander $\{\hat{f}_{m,n}\}$, que se deve provar é consistente em algum sentido, isto é, $\hat{f}_{m,n} \to f_0$, em algum sentido, quando $n, m \to \infty$. Este procedimento de estimação foi descrito em Grenander(1981) e é conhecido como *"método de sieves"*.

O desenho a seguir tenta ilustrar o método. Nele, f_m é um elemento de S_m que, quando m cresce, aproxima f_0 num certo sentido, que será esclarecido no Teorema 2.6.



Temos que encontrar condições para que, à medida em que, os estimadores de Grenander $\hat{f}_{m,n}$ se aproximem de f_m, quando n cresce, os f_m se aproximem do parâmetro f_0. Isto só será possível se $m = m_n$ e $m_n \to \infty$, quando $n \to \infty$.

Alguns exemplos que ilustrarão este procedimento são apresentados na próxima Seção e podem ser encontrados em Grenander(1981). Exemplos tratando de outros problemas de estimação por *"sieves"* encontram-se em Geman(1982), Nguyen(1982), McKeague(1986), Karr(1987), Antoniadis(1988) e Moulin *et al*(1992).

2.9 Exemplos de estimadores de Grenander

Histograma

Seja $X_1, \ldots, X_n$ uma amostra aleatória da função de densidade desconhecida f_0. Definindo para cada $m > 0$ os conjuntos

$$S_m = \left\{ f \in \Theta: f \text{ constante nos intervalos } \left[\frac{k-1}{2^m}, \frac{k}{2^m}\right), \; k = 0, \pm 1, \pm 2, \cdots \right\},$$

e, fazendo m crescer, obtemos que a seqüência $\{S_m\}$ constitui um *"sieves"*, e o estimador de Grenander associado para um dado tamanho de amostra n e uma dimensão do *"sieves"* m é

$$\hat{f}_{m,n}(x) = \frac{2^m}{n} \sum_{i=1}^{n} \mathbb{1}_{\left[\frac{k-1}{2^m}, \frac{k}{2^m}\right)}(X_i) \mathbb{1}_{\left[\frac{k-1}{2^m}, \frac{k}{2^m}\right)}(x),$$

que constitui o histograma com amplitude de intervalo $\frac{1}{2^m}$.

Estimadores de máxima verossimilhança penalizada

Este método de estimação estuda o problema de estimar a função de densidade de probabilidade desconhecida f_0 no conjunto Θ que maximize

$$\sum_{i=1}^{n} \log f(X_i) + \lambda\phi(f)$$

onde ϕ é um determinado funcional de penalização. A solução do problema descrito é o estimador de *"sieves"* obtido definindo-se a seqüência

$$S_m = \{f \in \Theta : \phi(f) \leq m\}.$$

Um estudo da relação entre os estimadores de máxima verossimilhança penalizada e os de *"sieves"* acha-se em Gimenez(1993).

Estimador de Séries Ortogonais

A idéia deste método é estimar f estimando os coeficientes da sua expansão em série de Fourier. Sejam $\widetilde{\Theta} \subseteq \mathrm{H} \subseteq \Theta$, H é um espaço de Hilbert separável. Neste caso

$$f \in \widetilde{\Theta} \Leftrightarrow f = \sum_{k=0}^{\infty} \theta_k \phi_k ,$$

onde $\{\phi_k\}_{k=0,\ldots,\infty}$ é uma base ortonormal de H.

No contexto de estimação por *"sieves"*, definindo a seqüência

$$S_m = \left\{ f \in \widetilde{\Theta} : f = \sum_{k=0}^{m} \vartheta_k \phi_k \right\},$$

obtemos o estimador *"sieves"* por séries ortogonais.

O *"sieves"* de convolução

Um tipo de *"sieves"* proposto por Grenander, mas pouco estudado, é conhecido como *"sieves"* de convolução. Neste caso, para uma dada densidade K, usualmente simétrica em torno de zero,

$$S_m = \left\{ f \in \Theta : f(x) = \int mK(m(x-y))F(dy),\ F \text{ função de distribuição de probabilidade} \right\}.$$

Os m são reais não-negativos e a função K é chamada de núcleo.

Este *"sieves"* fornece um estimador que está estreitamente relacionado com o estimador de Rosenblatt-Parzen estudado nas Seções 2.2 a 2.7. Será visto que, se $m \to \infty$ quando $n \to \infty$ numa velocidade apropriada, o estimador $\hat{f}_{m,n}$ será consistente. Este tipo de *"sieves"* será estudado com detalhes no Capítulo III, enfatizando particularmente o caso onde a função núcleo K é a densidade gaussiana.

2.10 Propriedades assintóticas do estimador de Grenander

O método de *"sieves"* produz uma grande variedade de estimadores e propriedades de existência e consistência podem ser obtidas de maneira unificada.

As referências básicas para os próximos resultados são Grenander(1981) e Geman & Hwang(1982). Os conceitos a seguir nos permitirão apresentar os teoremas de existência e consistência dos estimadores de Grenander. Como anteriormente, seja $f \in \Theta$ o espaço das funções de densidade com uma métrica d e S_m um *"sieves"*.

Para $f \in S_m$, definamos a bola

$$B_m(f,\varepsilon) = \{ g \in S_m : d(f,g) < \varepsilon \}.$$

Entenderemos por "entropia formal" a função

$$H(f,g) = E_f(\ln g(X)) = \int \ln g(x) f(x) dx,$$

onde $f, g \in \Theta$.

Associada a H temos a informação de Kullback-Leibler definida por $H(f,f) - H(f,g)$.

O conjunto dos estimadores de máxima verossimilhança em S_m dado o tamanho amostral n é definido como

$$M_m^n = \left\{ \hat{f}_{m,n} \in S_m : \hat{f}_{m,n} = arg \max_{g \in S_m} \prod_{i=1}^{n} g(X_i) \right\}.$$

O conjunto de entropia máxima em S_m será

$$A_m^n = \left\{ f_{m,n} \in S_m : f_{m,n} = \arg\max_{g \in S_m} H(f_0, g) \right\}.$$

O seguinte teorema trata da existência de uma seqüência $\{m_n\}$ para a qual cada um dos conjuntos $M_{m_n}^n$ é não-vazio e qualquer seqüência de estimadores em $M_{m_n}^n$ é consistente.

Teorema 2.5: Suponhamos que uma seqüência $S_m \subseteq \Theta$ seja escolhida de forma que:

i) Para todo *m*, toda $f \in S_m$ e todo $\varepsilon > 0$, a função

$$h_\varepsilon(x) = \sup_{g \in B_m(f,\varepsilon)} g(x)$$

é mensurável em *x*.

ii) Para todo *m* e quase todo *x* (segundo a medida *dx*)

$$\lim_{\varepsilon \to 0} h_\varepsilon(x) = f(x),$$

para toda $f \in S_m$. Isto significa que $f(x)$ é semicontínua superiormente em S_m.

iii) Para todo *m* e toda $f \in S_m$, existe $\varepsilon > 0$ tal que a entropia formal

$$H(f_0, h_\varepsilon) = E_{f_0}(\ln h_\varepsilon(X)) = \int \ln h_\varepsilon(x) f_0(x) dx$$

é finita.

iv) Os conjuntos S_m são compactos para todo *m*.

Se

$$\sup_{f_{m,n} \in A_m^n} d(f_m, f_0) \to 0,$$

quando $m \to \infty$, então, para todo *n*, *m* o conjunto M_m^n é não vazio, e para toda seqüência m_n de crescimento suficientemente lento,

$$\sup_{\hat{f}_{m,n} \in M_m^n} d(\hat{f}_{m,n}, f_0) \to 0,$$

quase certamente.

Este teorema dá condições para que uma seqüência *"sieves"* seja suficientemente regular para produzir estimadores consistentes, mas não diz nada sobre a velocidade com a qual m_n deve tender ao infinito. Neste sentido, uma razão de convergência razoável foi proposta por Shen & Wing(1994), eles sugerem $m_n = O(n^{-1/2})$ sem que, no entanto, seja ótima.

O método de estimação não tem aplicação se não pudermos identificar uma seqüência $\{m_n\}$ que garanta a consistência qualquer que seja o verdadeiro parâmetro. O teorema seguinte fornece um meio para caracterizar uma tal seqüência $\{m_n\}$.

Teorema 2.6: Suponhamos que uma seqüência *"sieves"* seja escolhida de tal forma que:

i) Para todos m e n o conjunto M_m^n é não-vazio quase certamente.

ii) Se para alguma seqüência $f_m \in S_m$, $H(f_0, f_m) \to H(f_0, f_0)$ quando $m \to \infty$, então $f_m \to f_0$ quando $m \to \infty$.

iii) Existe uma sequência $f_m \in S_m$ tal que $H(f_0, f_m) \to H(f_0, f_0)$.

iv) Para cada $\delta > 0$ e cada m, seja

$$D_m = \{f \in S_m : H(f_0, f) \leq H(f_0, f_m) - \delta\},$$

onde $\{f_m\}$ é a seqüência definida em *iii)*.

iv) Dados os conjuntos $O_1, \cdots, O_l$ em S_m tais que $\sup_{g \in O_k} g$ seja mensurável para cada k, seja

$$\rho_m = \sup_k \inf_{t \geq 0} E_{f_0}\left[exp\left[t \ln\left\{ \frac{\sup_{g \in O_k} g(X)}{\sup_{g \in S_m} g(X)} \right\} \right] \right].$$

v) Consideremos a seqüência $m_n \to \infty$ e suponhamos que para $\delta > 0$ seja possível encontrar $O_1^{m_n}, \cdots, O_l^{m_n}$ tais que

$$D_{m_n} \subset \bigcup_{k=1}^{l_{m_n}} O_k^{m_n},$$

$$\sup_{g \in O_k^{m_n}} g \text{ mensurável,}$$

e

$$\sum_{n=1}^{\infty} l_{m_n} (\rho_{m_n})^n < \infty .$$

Então, $\sup_{\hat{f}_{m_n} \in M_{m_n}^n} \hat{f}_{m_n} \to f_0$ quase certamente.

A aplicação deste teorema não é fácil. Exemplos de como identificar a seqüência podem ser achados em Geman & Hwang(1982). A métrica a ser escolhida para o espaço (Θ,d), e em termos da qual é expressa a consistência dos estimadores, é geralmente sugerida pela condição *ii)*.

Estes teoremas são bastante gerais e de difícil aplicação. Assim, um estudo particular para cada tipo de *"sieves"* envolvendo formas explícitas para o estimador é desejável e, em algumas situações, estudos com dados simulados é a única alternativa.

Capítulo III

O *"sieves"* de Convolução

Neste Capítulo, estudaremos especificamente o *"sieves"* de convolução. Em particular, a obtenção de uma forma fechada deste específico estimador de Grenander de duas maneiras diferentes. As situações onde a convolução utiliza como núcleo a densidade gaussiana e a exponencial dupla servirão como exemplos.

3.1 *"Sieves"* de Convolução

Seja X uma variável aleatória absolutamente contínua, com função de densidade f_0 e $X_1, \ldots, X_n$ uma amostra aleatória de X. De acordo com o que foi definido no Capítulo II, assumiremos o experimento estatístico como sendo a tripla $(X, \mathcal{B}, \{P_f : f \in \Theta\})$, onde as distribuições P_f são absolutamente contínuas com densidade f e Θ é o conjunto de todas as funções de densidade de probabilidade. Denotaremos o conjunto das funções de distribuição de probabilidade em R como $\mathfrak{I}$. Definamos a seqüência *"sieves"* como

$$S_m = \left\{ f \in \Theta : f(x) = \int mK(m(x-y))F(dy), F \in \mathfrak{I} \right\} \qquad (3.1)$$

onde K é uma função de densidade de probabilidade e m é um real não-negativo. Esta forma de definir os elementos dos conjuntos S_m consegue transferir propriedades desejáveis às funções f a partir da função núcleo K. Por exemplo, se K for limitada e uniformemente contínua, então f tambem será.

Observamos que os conjuntos S_m definidos em (3.1) formam uma seqüência não-decrescente.

Sejam $m < m' \Rightarrow \exists\, \varepsilon > 0$ tal que $m = \varepsilon m'$, então

$$f_m(x) = \int \varepsilon m' K(\varepsilon m'(x-y))F(dy) = \int m' K(m'(x-y'))F_\varepsilon(dy') = f_{m'}(x),$$

obtendo assim $S_m \subset S_{m'}$. Este mesmo desenvolvimento serve para provar $S_{m'} \subset S_m$, logo $S_m = S_{m'}$.

Com relação à propriedade de $\cup S_m$ ser denso em Θ, lembremos que a família de funções

núcleo, definida como $\{mK(mx): m > 0\}$, é conhecida como *aproximantes da identidade*. Veja, por exemplo, Wheeden & Zygmund(1977), onde é provado que o conjunto

$$S'_m = \{f \in S_m : F \text{ é função de distribuição de probabilidade absolutamente contínua}\},$$

é denso em Θ. Dado que $S'_m \subset S_m$, S_m é denso em Θ.

3.2 Existência e consistência do estimador *"sieves"* de convolução

Não encontramos na literatura resultados que mostrem, de forma geral, a existência do estimador Grenander obtido via *"sieves"* de convolução. Geman(1981) prova a existência de estimadores *"sieves"* de convolução restrito à função núcleo gaussiana, Walter & Blum(1984) no caso de função núcleo ser a exponencial dupla. Os métodos usados por estes autores são distintos.

A Proposição 3.1 a seguir oferece uma demonstração unificada da existência de $\hat{f}$ pertencente ao conjunto dos estimadores de Grenander no caso do *"sieves"* de convolução S_m, dado um tamanho da amostra n e K com uma única moda em zero.

Proposição 3.1: Suponhamos que a função núcleo K satisfaça $max_{x \in R} K(x) = K(0)$. Então o conjunto

$$M^n_{m_n} = \left\{ \hat{f}_{m_n} \in S_{m_n} : \hat{f}_{m_n} = arg \max_{g \in S_{m_n}} \prod_{i=1}^{n} g(X_i) \right\}$$

é não-vazio.

Prova.

Sejam $X_{(1)}=min(X_1, \ldots, X_n)$ e $X_{(n)}=max(X_1, \ldots, X_n)$ a primeira e a n-ésima estatísticas de ordem respectivamente. Consideraremos duas situações:

i) Suponha $X_{(1)} = X_{(n)}$, então, todos os elementos da amostra são iguais o que implica que o conjunto $M^n_{m_n}$ seja

$$M^n_{m_n} = \left\{ \hat{f}_{m_n} \in S_{m_n} : \hat{f}_{m_n} = arg \max_{g \in S_{m_n}} \prod_{i=1}^{n} g(X_{(i)}) \right\}.$$

Com isto

$$M_{m_n}^n = \left\{ \hat{f}_{m_n} \in S_{m_n} : \hat{f}_{m_n} = arg \max_{g \in S_{m_n}} g^n(X_{(1)}) \right\},$$

pela forma das funções g de S_{m_n} e devido a termos fixado o argumento, chegamos que

$$\max_{g \in S_{m_n}} g^n(X_{(1)}) = \max_{g \in S_{m_n}} g(X_{(1)}).$$

Este último é equivalente a maximizar $g(x)$, $g \in S_{m_n}$, ou seja

$$\max_{x \in R} g(x) = \max_{x \in R} \int_R m_n K\big(m_n(x-y)\big) F(dy).$$

Dado que $\max_{x \in R} K(x) = K(0)$, pela forma com que a função K foi escolhida, temos

$$\max_{x \in R} g(x) = \int_R m_n K\big(m_n(x-y)\big) F(dy) =$$

$$= \int_R \max_{x \in R} m_n K\big(m_n(x-y)\big) F(dy) =$$

$$= \int_R m_n K\big(m_n(0)\big) F(dy) = m_n K(0),$$

já que a função integrando não depende da variável de integração e a medida segundo F é 1.

O máximo de g é assumido então quando $y = X_{(1)}$, e a forma da função na qual este máximo é atingido será

$$m_n K\big(m_n(x - X_{(1)})\big),$$

constituindo o único elemento de $M_{m_n}^n$, mostrando assim que este conjunto é não-vazio quando $X_{(1)} = X_{(n)}$.

ii) Suponha $X_{(1)} < X_{(n)}$. Seja $\hat{h}_k \subset S_{m_n}$ uma seqüência em S_{m_n} tal que

$$\lim_{k\to\infty}\prod_{i=1}^{n}\hat{h}_k(X_i)=\sup_{g\in S_{m_n}}\prod_{i=1}^{n}g(X_i).$$

A existência desta seqüência é garantida pela definição de supremo de um conjunto.

Definimos a sequência F_k como aquela cujos elementos são as funções de distribuição de probabilidade que produzem os $\hat{h}_k$, isto é,

$$\hat{h}_k(x)=\int_R m_n K\big(m_n(x-y)\big)F_k(dy).$$

A partir de F_k, definimos a medida $\mu_{\tilde{F}_k}$ como

$$\mu_{\tilde{F}_k}(B)=\mu_{F_k}(B) \text{ se } B\subset(X_{(1)},X_{(n)})$$

$$\mu_{\tilde{F}_k}(\{X_{(1)}\})=\mu_{F_k}((-\infty,X_{(n)}])$$

$$\mu_{\tilde{F}_k}(\{X_{(n)}\})=\mu_{F_k}([X_{(n)},+\infty))$$

para k=1,2, ... Desta definição, temos que

$$\mu_{\tilde{F}_k}(\{X_{(1)}\})=F_k(X_{(1)})-F_k(-\infty)=F_k(X_{(1)})$$

$$\mu_{\tilde{F}_k}(\{X_{(n)}\})=F_k(+\infty)-F_k(X_{(n)})=1-F_k(X_{(n)}).$$

Logo, da definição da medida de Lebesgue-Stieltjes $\mu_{\tilde{F}_k}$, obtém-se univocamente as $\tilde{F}_k$ como sendo funções de distribuição de probabilidade, definidas como

$$\tilde{F}_k(y)=0 \qquad \text{se } y\in(-\infty,X_{(1)})$$

$$\tilde{F}_k(y)=F_k(y) \qquad \text{se } y\in[X_{(1)},X_{(n)}]$$

$$\tilde{F}_k(y)=1 \qquad \text{se } y\in(X_{(n)},+\infty).$$

Seja $\tilde{h}_k(x)=\int_R m_n K\big(m_n(x-y)\big)\tilde{F}_k(dy)$ e, dado que $\tilde{F}_k$ é função de distribuição de probabilidade, temos que $\tilde{h}_k\in S_{m_n}$, k=1,2, ...

$$\int_R m_n K(m_n(x-y))\widetilde{F}_k(dy) = \int_{(-\infty,X_{(1)}]} m_n K(m_n(x-y))\widetilde{F}_k(dy) +$$

$$\int_{(X_{(1)},X_{(n)})} m_n K(m_n(x-y))\widetilde{F}_k(dy) + \int_{[X_{(n)},+\infty)} m_n K(m_n(x-y))\widetilde{F}_k(dy).$$

Observemos que $\widetilde{F}_k$ é a *variação total*[1] de F_k nos intervalos $(-\infty, X_{(1)})$ e $(X_{(n)},+\infty)$. Então

$$\int_{(-\infty,X_{(1)}]} m_n K(m_n(x-y))F_k(dy) \le \int_{(-\infty,X_{(1)}]} m_n K(m_n(x-y))\widetilde{F}_k(dy),$$

$$\int_{[X_{(n)},+\infty)} m_n K(m_n(x-y))F_k(dy) \le \int_{[X_{(n)},+\infty)} m_n K(m_n(x-y))\widetilde{F}_k(dy).$$

Com isto, temos limitada a verossimilhança na densidade h_k da seguinte forma,

$$\prod_{i=1}^{n} h_k(X_i) \le \prod_{i=1}^{n} \widetilde{h}_k(X_i),$$

e também

$$\lim_{k\to\infty} \prod_{i=1}^{n} \widetilde{h}_k(X_i) = \sup_{g\in S_{mn}} \prod_{i=1}^{n} g(X_i).$$

Temos então que $\{\widetilde{F}_k\}$ é uma seqüência *tight*[2] dado que o suporte destas distribuições é o

---

[1] Seja $\varphi: F \to R$ uma função σ- aditiva. As medidas $\varphi^+(A) = sup_{E\subset A}\varphi(E)$ e $\varphi^-(A) = -inf_{E\subset A}\varphi(E)$, $A \subseteq F$ são chamadas de "variações inferior e superior de φ". A medida $|\varphi(A)| = \varphi^+(A) + \varphi^-(A)$ é chamada de "variação total de φ".

[2] Uma seqüência de funções $\widetilde{F}_k$ se diz *tight* se para cada $\varepsilon > 0$, existem a e b reais, $a < b$, tais que $\widetilde{F}_k(a) < \varepsilon$ e $\widetilde{F}_k(b) > 1 - \varepsilon \ \forall k$. Para caracterizar este tipo de seqüências temos que se as $\{\widetilde{F}_k\}$ são de suporte compacto $\forall k$, então formam uma seqüência *tight*, veja Billingsley(1979).

boreliano $[X_{(1)}, X_{(n)}]\ \forall k$; que é fechado e limitado, logo compacto. O fato de $\{\widetilde{F}_k\}$ ser *tight* garante que $lim_{k\to\infty} \widetilde{F}_k = F_\infty$, onde F_∞ é função de distribuição de probabilidade.

Se

$$\widetilde{f}_\infty(x) = \int_R m_n K(m_n(x-y)) F_\infty(dy),$$

então $f_\infty \in S_{m_n}$ e $f_\infty(x) = \lim_{k\to\infty} \widetilde{h}_k(x)$ em cada x. Tem-se então

$$\prod_{i=1}^{n} \widetilde{f}_\infty(X_i) = \lim_{k\to\infty} \prod_{i=1}^{n} \widetilde{h}_k(X_i) = \sup_{g\in S_{m_n}} \prod_{i=1}^{n} g(X_i).$$

Segue que $\widetilde{f}_\infty \in M^n_{m_n}$, logo $M^n_{m_n}$ é não-vazio. ♦

Veremos a seguir que a condição *ii)* do Teorema 2.6 é satisfeita. Este resultado foi provado em Geman(1981) e nos garante as condições para a consistência destes estimadores.

As condições para a consistência dos estimadores de Grenander da função de densidade de probabilidade f_0 são apresentadas a seguir. Esta é demonstrada de modo geral, isto é, será válida qualquer que seja a seqüência *"sieves"* de convolução S_m escolhida.

Proposição 3.2: Seja f_0 uma função de densidade satisfazendo

$$\int_{-\infty}^{\infty} f_0(x) \ln f_0(x) dx < \infty.$$

Se, para cada n, T_n é um conjunto de funções de densidade f_n tal que

$$\lim_{n\to\infty} \sup_{f_n \in T_n} \int_{-\infty}^{\infty} f_0(x) \ln\left\{\frac{f_0(x)}{f_n(x)}\right\} dx = 0,$$

então tambem

$$\lim_{n\to\infty} \sup_{f_n \in T_n} \int_{-\infty}^{\infty} |f_n(x) - f_0(x)| dx = 0. \qquad (3.2)$$

Note que assim se obtém a validez da condição *ii)* do Teorema 2.6.

Prova.

Para cada $c>1$, definamos a variável x_c sendo $x_c > 1$ e tal que

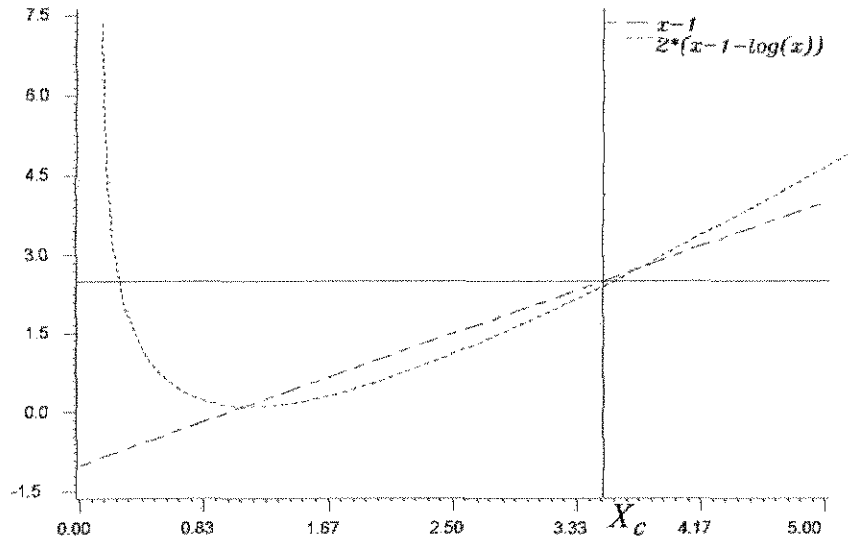
$$x_c - 1 = c(x_c - 1 - \log x_c).$$

Observemos que

$$x - 1 < c(x - 1 - \log x) \tag{3.3}$$

para toda $x > x_c$.

Estudemos a relação anterior. Por exemplo, no caso $c = 2$ temos o seguinte gráfico



O ponto de interceção das curvas diferente de $x = 1$, está ressaltado no gráfico. Esse é o ponto x_c. A partir dele verifica-se a desigualdade (3.3)

Provemos a relação em (3.3). Para isto definamos a função $f(x) = c(x - 1 - \log x)$. Com

$$f''(x) = c\left(1/x^2\right) > 0, \forall x \neq 0,$$

temos assim que a função f é convexa, isto significa que o segmento de reta ligando os pontos $(1,0)$ e $(x_c, f(x_c))$ está acima do gráfico da função f. Este segmento pertence à reta $x - 1$. Com isto temos que a relação em (3.3) é satisfeita $\forall x > x_c$.

Dado que x_c é a solução da equação $x_c - 1 = c(x_c - 1 - \log x_c)$, temos que

$$\lim_{c\to\infty} \frac{x_c - 1}{x_c - 1 - \log(x_c)} = \lim_{c\to\infty} \frac{1}{1 - 1/x_c} = \infty,$$

obtemos que $\lim_{c\to\infty} 1/x_c = 1$, e então

$$\lim_{c\to\infty} x_c = 1. \tag{3.4}$$

Além disto, temos que

$$x - 1 - \log x \geq 0 \tag{3.5}$$

para todo $x \geq 0$.

Escolhamos $c_n \to \infty$ tal que

$$\lim_{n\to\infty} c_n \sup_{f_n \in T_n} \int_{-\infty}^{\infty} f_0(x) \log \frac{f_0(x)}{f_n(x)} dx = 0.$$

$$\overline{\lim_{n\to\infty}} \sup_{f_n \in T_n} \int_{-\infty}^{\infty} |f_n(x) - f_0(x)| dx = 2 \overline{\lim_{n\to\infty}} \sup_{f_n \in T_n} \int_{f_n > f_0} f_0(x) \left(\frac{f_n(x)}{f_0(x)} - 1 \right) dx$$

$$= 2 \overline{\lim_{n\to\infty}} \sup_{f_n \in T_n} \int_{1 < \frac{f_n}{f_0} \leq X_{c_n}} f_0(x) \left(\frac{f_n(x)}{f_0(x)} - 1 \right) dx + 2 \overline{\lim_{n\to\infty}} \sup_{f_n \in T_n} \int_{\frac{f_n}{f_0} > X_{c_n}} f_0(x) \left(\frac{f_n(x)}{f_0(x)} - 1 \right) dx.$$

Utilizemos agora o resultado em (3.3). Identifiquemos $\frac{f_n}{f_0}$ como sendo o x em (3.3). Então

$$2 \overline{\lim_{n\to\infty}} \sup_{f_n \in T_n} \int_{1 < \frac{f_n}{f_0} \leq x_{c_n}} f_0(x) \left(\frac{f_n(x)}{f_0(x)} - 1 \right) dx \leq 2 \overline{\lim_{n\to\infty}} (x_{c_n} - 1) \tag{3.6}$$

e

$$2\ \overline{\lim_{n\to\infty}}\ \sup_{f_n\in T_n}\ \int_{\frac{f_n}{f_0}>x_{c_n}} f_0(x)\left(\frac{f_n(x)}{f_0(x)}-1\right)dx$$

$$\leq 2\ \overline{\lim_{n\to\infty}}\ c_n\ \sup_{f_n\in T_n}\ \int_{\frac{f_n}{f_0}>x_{c_n}} f_0(x)\left(\frac{f_n(x)}{f_0(x)}-1-log\left(\frac{f_n(x)}{f_0(x)}\right)\right)dx.$$

O resultado em (3.4) implica que o limite em (3.6) é zero, e (3.5) nos permite afirmar que

$$2\ \overline{\lim_{n\to\infty}}\ c_n\ \sup_{f_n\in T_n}\ \int_{\frac{f_n}{f_0}>x_{c_n}} f_0(x)\left(\frac{f_n(x)}{f_0(x)}-1-log\left(\frac{f_n(x)}{f_0(x)}\right)\right)dx$$

$$\leq 2\ \overline{\lim_{n\to\infty}}\ c_n\ \sup_{f_n\in T_n}\ \int_{-\infty}^{\infty} f_0(x)\left(\frac{f_n(x)}{f_0(x)}-1-log\left(\frac{f_n(x)}{f_0(x)}\right)\right)dx.$$

Dado que as integrais $\int_{-\infty}^{\infty} f_0(x)dx$ e $\int_{-\infty}^{\infty} f_n(x)dx$ são ambos iguais a 1, obtemos

$$\overline{\lim_{n\to\infty}}\ c_n\ \sup_{f_n\in T_n}\ \int_{-\infty}^{\infty} f_0(x)\left(\frac{f_n(x)}{f_0(x)}-1-log\left(\frac{f_n(x)}{f_0(x)}\right)\right)dx=$$

$$=\overline{\lim_{n\to\infty}}\ c_n\ \sup_{f_n\in T_n}\ \int_{-\infty}^{\infty}\left(f_n(x)-f_0(x)-f_0(x)log\left(\frac{f_n(x)}{f_0(x)}\right)\right)dx$$

$$=\overline{\lim_{n\to\infty}}\ c_n\ \sup_{f_n\in T_n}\ \int_{-\infty}^{\infty} f_0(x)log\left(\frac{f_0(x)}{f_n(x)}\right)dx=0,$$

implicando no resultado do teorema. ♦

Destas Proposições obtemos que o estimador de Grenander com *"sieves"* de convolução da densidade desconhecida f_0 existe e é consistente. Estaremos interessados agora na forma deste estimador. Estudos neste sentido só foram encontrados em Geman(1981), onde é derivada a forma do estimador no caso de ser a função núcleo gaussiana e em Walter & Blum(1984), onde é obtido o estimador quando a função núcleo é exponencial dupla.

A seguir estudaremos duas formas de obter $\hat{f}$. Na primeira seguimos as idéias expostas em Geman(1981) para o caso da função núcleo gaussiana. Na segunda, utilizaremos o *modelo de dados incompletos* e a caracterização do estimador de máxima verossimilhança da função de distribuição dada por Laird(1978). Chegaremos que o estimador $\hat{f}$ é uma mistura ponderada, possivelmente finita, da função núcleo. Obtemos também que, se a função núcleo for gaussiana, o número de misturas é finito e está relacionado com o tamanho da amostra.

A segunda forma de encontrar o estimador de Grenander via convoluções é mais simples, provando-se também que o estimador procurado é uma mistura ponderada da função núcleo. Conseguimos garantir que o número de misturas é finito, no entanto, quando particularizado à função núcleo gaussiana, não se consegue relacionar teoricamente o número de misturas ao tamanho da amostra.

3.3 Primeira forma de obter $\hat{f}$ via *"sieves"* de convolução

A demonstração da Proposição a seguir foi baseada nas idéias em Geman(1981) no caso da função K gaussiana.

Proposição 3.3: Assumamos que a função núcleo K seja derivável e tenha um único máximo em 0. Então, se $\hat{f} \in M_{m_n}^n$, esta será da forma

$$\hat{f}(x) = \sum_{i=1}^{q} p_i m_n K\big(m_n(x - y_i)\big), \tag{3.7}$$

onde q não é necessariamente finito.

Prova.

Seja F uma função de distribuição de probabilidade tal que

$$\hat{f}(x) = \int_R m_n K\big(m_n(x - y)\big) F(dy), \tag{3.8}$$

com $\hat{f} \in M_{m_n}^n$.

A idéia da demonstração é construir uma função auxiliar que nos permita utilizar a condição necessária de extremo através da derivada de Gateaux e cujo resultado seja equivalente a maximizar a verossimilhança restrita a S_{m_n}.

A derivada de Gateaux nos oferece uma condição necessária de extremo (Veja Teorema I.1, Apêndice I). Como assumimos que $\hat{f} \in M_{m_n}^n$, ou seja, $\hat{f}$ é uma função que maximiza o funcional

$$L_n(g) = \prod_{i=1}^{n} g(X_i), \; g \in S_{m_n},$$

temos que

$$\delta L(\hat{f}; h) = 0, \; \forall h \in S_{m_n},$$

onde $\delta L(\hat{f}; h)$ denota o diferencial Gateaux de L em $\hat{f}$ com incremento h.

Dito de outra forma, dado que $\hat{f}$ anula a derivada de Gateaux do funcional L, por ser o ponto de máximo, obtemos

$$\frac{d}{d\alpha} L(\hat{f} + \alpha h)\Big|_{\alpha=0} = 0.$$

para todo $h \in S_{m_n}$.

Construamos primeiramente a função auxiliar. Para $s \in R$ com $F(s) > 0$, $\varepsilon > 0$ e $z \in R$, seja $\mu_{G_{\varepsilon,s,z}}$ uma medida definida como

$$\mu_{G_{\varepsilon,s,z}}(B) = \mu_F([s-\varepsilon, s+\varepsilon] \cap (B - z)), \tag{3.9}$$

onde B é um Boreliano de R.

A medida definida em (3.9), corresponde à função de distribuição $G_{\varepsilon,s,z}$, isto é, $G_{\varepsilon,s,z}$ é uma função real não decrescente contínua à direita. Escrevamo-la em termos da função de distribuição de probabilidade F.

$$\mu_{G_{\varepsilon,s,z}}((-\infty, t]) = \mu_F([s-\varepsilon, s+\varepsilon] \cap ((-\infty, t-z]) = \mu_F([s-\varepsilon, \min(s+\varepsilon, t-z)]),$$

logo

$$\begin{cases} G_{\varepsilon,s,z}(t) = F(\min(s+\varepsilon, t-z)) - F((s-\varepsilon)^-) & \text{se} \quad t - z > s - \varepsilon \\ G_{\varepsilon,s,0}(t) = F(\min(s+\varepsilon, t)) - F((s-\varepsilon)^-) & \text{se} \quad t > s - \varepsilon \end{cases}$$

Definindo

$$F_{\varepsilon,s}(t) = F(t) - G_{\varepsilon,s,0}(t) = F(t) - F(min(s+\varepsilon,t)) + F((s-\varepsilon)^{-}),$$

temos que $F_{\varepsilon,s} + G_{\varepsilon,s,z}$ é função de distribuição de probabilidade de uma variável aleatória Z.

Pela definição anterior temos que

$$F_{\varepsilon,s}(t) + G_{\varepsilon,s,z}(t) = F(t) - F(min(s+\varepsilon,t)) + F(min(s+\varepsilon,t-z)) \qquad (3.10)$$

que é soma e diferença de funções contínuas à direita, logo é contínua à direita.

A função em (3.10) é não-decrescente. De fato, definamos $H = F_{\varepsilon,s} + G_{\varepsilon,s,z}$. Sejam t_1 e t_2 números reais ($t_1 < t_2$), e seja $z > 0$, tal que $z < t_2 - t_1$. Podemos ter as seguintes situações:

$$s+\varepsilon \le t_1 - z,\ t_1 - z < s+\varepsilon \le t_1,\ t_1 < s+\varepsilon \le t_2 - z,\ t_2 - z < s+\varepsilon \le t_2 \text{ ou } t_2 \le s+\varepsilon.$$

Para cada uma das possibilidades expressadas acima, $H(t_1) \le H(t_2)$. Vejamos isto em cada uma destas situações:

i) $s+\varepsilon \le t_1 - z$

$$F(min(s+\varepsilon,t)) = F(min(s+\varepsilon,t-z))$$

qualquer seja o valor de t, logo

$$H(t_1) = F(t_1) \le F(t_2) = H(t_2).$$

ii) $t_1 - z < s+\varepsilon \le t_1$

$$F(min(s+\varepsilon,t_1-z)) = F(t_1-z) \le F(min(s+\varepsilon,t_2-z)) = F(s+\varepsilon)$$

$$F(min(s+\varepsilon,t_1)) = F(s+\varepsilon) = F(min(s+\varepsilon,t_2)) = F(s+\varepsilon)$$

$$H(t_1) = F(t_1) - F(s+\varepsilon) + F(t_1-z),$$

dado que

$$F(t_1-z) \le F(s+\varepsilon)$$

temos

$$H(t_1) \le F(t_1) \le F(t_2) = H(t_2).$$

iii) $t_1 < s+\varepsilon \le t_2-z$

$$F(min(s+\varepsilon, t_1 - z)) = F(t_1 - z) \le F(min(s+\varepsilon, t_2 - z)) = F(s+\varepsilon)$$

$$F(min(s+\varepsilon, t_1)) = F(t_1) \le F(min(s+\varepsilon, t_2)) = F(s+\varepsilon)$$

$$H(t_1) = F(t_1 - z) \le F(t_2) = H(t_2).$$

iv) $t_2-z < s+\varepsilon \le t_2$

$$F(min(s+\varepsilon, t_1 - z)) = F(t_1 - z) \le F(min(s+\varepsilon, t_2 - z)) = F(t_2 - z)$$

$$F(min(s+\varepsilon, t_1)) = F(t_1) \le F(min(s+\varepsilon, t_2)) = F(s+\varepsilon),$$

dado que

$$F(t_2) \ge F(s+\varepsilon)$$

temos

$$H(t_2) \ge F(t_2 - z) \ge F(t_1 - z) = H(t_1).$$

v) $s+\varepsilon > t_2$

$$F(min(s+\varepsilon, t_1 - z)) = F(t_1 - z) \le F(min(s+\varepsilon, t_2 - z)) = F(t_2 - z)$$

$$F(min(s+\varepsilon, t_1)) = F(t_1) \le F(min(s+\varepsilon, t_2)) = F(t_2)$$

$$H(t_1) = F(t_1 - z) \le F(t_2 - z) = H(t_2).$$

O mesmo argumento serve para mostrar que $H(t_1) \le H(t_2)$ é válido nas outras situações, ou seja, a relação continua válida se $z > 0$ e $z > t_2 - t_1$, se $z < 0$ e $z < t_2 - t_1$ ou se $z < 0$ e $z > t_2 - t_1$.

Construiremos a seguir uma função diferenciável segundo Gateaux e provaremos que maximizar este funcional é equivalente, no limite, a maximizar a verossimilhança da função $\hat{f}$. Para a construção deste funcional utilizaremos a função $F_{\varepsilon,s} + G_{\varepsilon,s,z}$.

Definamos

$$f_{\varepsilon,s,z}(x) = \int_R m_n K\big(m_n(x-y)\big)\mu_{F_{\varepsilon,s}}(dy) + \int_R m_n K\big(m_n(x-y)\big)\mu_{G_{\varepsilon,s,z}}(dy).$$

A partir de (3.8), obtemos $\hat{f} = f_{\varepsilon,s,0}$, isto é devido ao fato de $F = F_{\varepsilon,s} + G_{\varepsilon,s,0}$.

Note que

$$\frac{\partial}{\partial z}\sum_{i=1}^{n} \log f_{\varepsilon,s,z}(X_i)\Big|_{z=0} \tag{3.11}$$

é coincidente com a derivada de Gateaux (veja Apêndice I) do funcional que define a verossimilhança L, isto é, a derivada em (3.11) é coincidente com a derivada do funcional

$$\log L(\hat{f}) = \sum_{i=1}^{n} \log \hat{f}(X_i).$$

Utilizando a notação da derivada de Gateaux, temos

$$\delta \log L(\hat{f}) = \sum_{i=1}^{n} \frac{1}{\hat{f}(X_i)} \frac{\partial}{\partial z} \int_{-\infty}^{\infty} m_n K\big(m_n(X_i - y)\big) \mu_{G_{\varepsilon,s,z}}(dy)\Big|_{z=0}. \tag{3.12}$$

Da definição de derivada é possível expressar (3.12) em termos de F, mostrando a coincidência com a derivada de Gateaux;

$$\frac{\partial}{\partial z}\int_{R} m_n K\big(m_n(X-y)\big)\mu_{G_{\varepsilon,s,z}}(dy) =$$

$$= \lim_{\Delta z \to 0} \frac{\int_{R} m_n K\big(m_n(X-y)\big)\mu_{G_{\varepsilon,s,z+\Delta z}}(dy) - \int_{R} m_n K\big(m_n(X-y)\big)\mu_{G_{\varepsilon,s,z}}(dy)}{\Delta z}$$

$$= \lim_{\Delta z \to 0} \frac{\int_{(s-\varepsilon,s+\varepsilon)} m_n K\big(m_n(X-y)\big)F(dy - z - \Delta z) - \int_{(s-\varepsilon,s+\varepsilon)} m_n K\big(m_n(X-y)\big)F(dy - z)}{\Delta z}$$

$$= \lim_{\Delta z \to 0} \int_{(s-\varepsilon,s+\varepsilon)} \frac{m_n\Big(K\big(m_n(X-y)\big)F(dy - z - \Delta z) - K\big(m_n(X-y)\big)F(dy - z)\Big)}{\Delta z}$$

$$= \lim_{\Delta z \to 0} \int_{(s-\varepsilon,s+\varepsilon)} \frac{m_n\Big(K\big(m_n(X-y+z+\Delta z)\big) - K\big(m_n(X-y+z)\big)\Big)}{\Delta z} F(dy). \tag{3.13}$$

Substituindo (3.13) em (3.12), obtemos

$$\sum_{i=1}^{n}\frac{1}{\hat{f}(X_i)}\frac{\partial}{\partial z}\int_{-\infty}^{\infty}m_n K\big(m_n(X_i-y)\big)\mu_{G_{\varepsilon,s,z}}(dy)\Big|_{z=0} =$$

$$\sum_{i=1}^{n}\frac{1}{\hat{f}(X_i)}\lim_{\Delta z\to 0}\int_{(s-\varepsilon,s+\varepsilon)}\frac{m_n\big(K\big(m_n(x-y+z+\Delta z)\big)-K\big(m_n(x-y+z)\big)\big)}{\Delta z}F(dy)\Big|_{z=0} \quad (3.14)$$

Com o objetivo de mudar a ordem das operações de integração e diferenciação em (3.14), vejamos se as condições do Teorema da Convergência Dominada são válidas.

Dado que

$$m_n K\big(m_n(X-y)\big)\le m_n K(0)$$

e

$$\int_{-\infty}^{\infty}m_n K\big(m_n(X-y)\big)\mu_{G_{\varepsilon,s,z}}(dy)\le m_n K(0)\int_{-\infty}^{\infty}\mu_{G_{\varepsilon,s,z}}(dy),$$

por (3.9) temos que

$$m_n\int_{-\infty}^{\infty}\mu_{G_{\varepsilon,s,z}}(dy)\le m_n K(0),$$

já que $\int_{-\infty}^{\infty}\mu_{G_{\varepsilon,s,z}}(dy)\le 1$.

Pelo Teorema da Convergência Dominada, temos

$$\sum_{i=1}^{n}\frac{1}{\hat{f}(X_i)}\frac{\partial}{\partial z}\int_{-\infty}^{\infty}m_n K\big(m_n(X_i-y)\big)\mu_{G_{\varepsilon,s,z}}(dy)\Big|_{z=0} =$$

$$=\sum_{i=1}^{n}\frac{1}{\hat{f}(X_i)}\int_{-\infty}^{\infty}\frac{\partial}{\partial z}m_n K\big(m_n(X_i-y)\big)\mu_{G_{\varepsilon,s,z}}(dy)\Big|_{z=0}. \quad (3.15)$$

A igualdade em (3.15) é válida pois, pela definição de derivada,

$$\lim_{\Delta z\to 0}\frac{m_n\left(K\left(m_n(X-y+z+\Delta z)\right)-K\left(m_n(X-y+z)\right)\right)}{\Delta z}\Bigg|_{z=0}=m_n\frac{\partial}{\partial z}K\left(m_n(X-y+z)\right)\Bigg|_{z=0}.$$

Utilizando o resultado de (3.14) em (3.15), temos

$$\sum_{i=1}^{n}\frac{1}{\hat{f}(X_i)}\int_{(s-\varepsilon,s+\varepsilon)}\lim_{\Delta z\to 0}\frac{m_n\left(K\left(m_n(X_i-y+z+\Delta z)\right)-K\left(m_n(X_i-y+z)\right)\right)}{\Delta z}F(dy)\Bigg|_{z=0}=$$

$$\sum_{i=1}^{n}\frac{1}{\hat{f}(X_i)}\int_{(s-\varepsilon,s+\varepsilon)}\frac{\partial}{\partial z}m_nK\left(m_n(X_i-y+z)\right)F(dy)\Bigg|_{z=0}.$$

Multiplicando e dividindo por $\mu_F([s-\varepsilon,s+\varepsilon])$ na expressão anterior e fazendo $\varepsilon\to 0$, segue que

$$\sum_{i=1}^{n}\frac{1}{\hat{f}(X_i)}\lim_{\varepsilon\to 0}\frac{\mu_F([s-\varepsilon,s+\varepsilon])}{\mu_F([s-\varepsilon,s+\varepsilon])}\int_{(s-\varepsilon,s+\varepsilon)}\frac{\partial}{\partial z}m_mK\left(m_n(X_i-y+z)\right)F(dy)\Bigg|_{z=0}=$$

$$=\sum_{i=1}^{n}\frac{F(s)}{\hat{f}(X_i)}\frac{\partial}{\partial z}m_nK\left(m_n(X_i-y+z)\right)\Bigg|_{z=0}, \qquad (3.16)$$

para qualquer s no suporte de F.

Segundo a definição da função $\hat{f}$, o máximo do funcional L em S_{m_n} é atingido nos $\hat{f}\in M_{m_n}^n$. Pela condição necessária de extremo segundo a derivada de Gateaux (Teorema I.1, Apêndice I), temos

$$\delta\log L(\hat{f})=\delta\sum_{i=1}^{n}\log\hat{f}(X_i)=\frac{\partial}{\partial z}\sum_{i=1}^{n}\log f_{\varepsilon,s,z}(X_i)\Bigg|_{z=0}=0.$$

Então

$$\sum_{i=1}^{n} \frac{F(s)}{\hat{f}(X_i)} \frac{\partial}{\partial z} m_n K\big(m_n(X_i - y + z)\big)\Big|_{z=0} = 0,$$

obtendo-se

$$\sum_{i=1}^{n} \frac{1}{\hat{f}(X_i)} m_n \frac{\partial}{\partial z} K\big(m_n(X_i - y + z)\big)\Big|_{z=0} = 0. \qquad (3.17)$$

Definindo q como o número de valores de y que satisfazem (3.17), podemos escrever a função $\hat{f}$ em (3.7) como

$$\hat{f}(x) = \sum_{i=1}^{q} p_i m_n K\big(m_n(x - y_i)\big). \quad \blacklozenge$$

Desta forma obtemos uma expressão fechada para o estimador de Grenander via convolução. Observemos que em (3.17) podemos ter infinitas soluções. Será visto posteriormente que em casos particulares o número q é finito e que está relacionado com o tamanho da amostra.

Na próxima Seção estudaremos outra forma de obter a expressão do estimador $\hat{f}$. Consideraremos o modelo de amostragem conhecido como *modelo de dados incompletos*. Trabalhando assim, consegue-se provar que o número de misturas na definição de $\hat{f}$ é finito, ou seja, q é finito. No entanto, quando particularizamos a uma determinada função núcleo, não conseguimos relacionar o número de misturas com o tamanho da amostra.

3.4 Segunda forma de obter o estimador $\hat{f}$, via *"sieves"* de convolução

Em problemas nos quais é necessária a estimação da função de distribuição de probabilidade F, um estimador amplamente utilizado é a função de distribuição empírica. Quando assumido o *modelo de dados incompletos* são necessários estudos específicos para caracterizar o estimador de máxima verossimilhança da função F. Um estudo neste sentido acha-se em Laird(1978).

Devido ao fato de o estimador *"sieves"* de convolução para a função de densidade de probabilidade, ser definido pela convolução da função núcleo conhecida K com a função de distribuição de probabilidade F, mostraremos que, utilizando o estimador de máxima

verossimilhança $\hat{F}$ obtido segundo o *modelo de dados incompletos*, é possível chegar à forma do estimador de Grenander da função de densidade de probabilidade.

3.4.1 Caracterização do estimador de máxima verossimilhança $\hat{F}$ da função de distribuição

Existe uma ampla bibliografia sobre o *modelo de dados incompletos*, uma das versões tendo sido introduzida no trabalho de Robbins(1964).

Seja $(X_1,Y_1),\ldots,(X_n,Y_n)$ uma amostra aleatória do vetor de variáveis aleatórias (X,Y), onde Y tem função de distribuição F. Assumiremos que o espaço amostral de Y é um intervalo I e que esta variável é não observável. Condicionalmente em $Y=y$, cada X é independentemente distribuída com função de densidade de probabilidade $K(x|Y=y)$. Assumiremos a variável X observável. Em geral, a variável X será absolutamente contínua e a variável Y, na maioria das vezes, é do tipo discreta. No entanto, chamaremos aqui de densidade a função conjunta.

A densidade marginal de cada X será

$$f(x|F)=\int K(x|y)F(dy), \tag{3.18}$$

onde assumimos que a fórmula da função núcleo K conhecida.

Foi visto na Seção 2.1 que a função de distribuição empírica $\hat{F}_n(x)$ é um bom estimador de $F(x)$, a função de distribuição de probabilidade quando assumido o modelo de dados completos. A seguir, faremos suposições e definiremos conceitos que permitirão caracterizar um estimador da função de distribuição de probabilidades quando o modelo adotado é o de dados incompletos.

Podemos observar em (3.18) que a forma de obter a função de densidade marginal de X guarda relação com a definição dos elementos dos conjuntos S_m em (3.1), ou seja, é do tipo dos elementos dos conjuntos que formam a seqüência *"sieves"* de convolução.

Nosso objetivo é caracterizar o estimador de máxima verossimilhança $\hat{F}$ da função de distribuição de probabilidade F como auto-consistente. Desta forma provaremos que $\hat{F}$ é uma função escada e, com isto, reduziremos a integral em (3.18) a uma soma ponderada.

Esta caracterização do estimador $\hat{F}$ será feita, por simplicidade de notação, quando o vetor (X,Y) for do tipo absolutamente contínuo. Na situação em que a variável Y é discreta, as demonstrações seguem as mesmas idéias.

Introduziremos a seguir conceitos que permitirão o desenvolvimento das provas do Lema 3.1 e os Teoremas 3.1 e 3.2.

Definamos $\Im_1$ como o subconjunto das funções de distribuição de probabilidade F que geram densidades f estritamente positivas através da convolução em (3.18), isto é

$$\Im_1 = \{F \in \Im : f(x|F) > 0\},$$

onde $\Im$ é o conjunto das funções de distribuição de probabilidade.

Seja

$$l(F) = \sum_{i=1}^{n} \log f(X_i|F),$$

o logaritmo da função de verossimilhança da função de densidade condicional de X. A densidade conjunta do vetor (X,Y) será

$$p(x,y|F) = K(x|y)g(y), \tag{3.19}$$

onde g é a função de densidade da variável aleatória Y. A densidade condicional de $(Y|X = x)$, definida para toda $F \in \Im_1$, é

$$h(y|x,F) = \frac{K(x|y)g(y)}{f(x|F)}, \tag{3.20}$$

com função de distribuição de probabilidade $H(\cdot|x,F)$.

Definição 3.1: Diremos que o estimador $\tilde{F}$ da função de distribuição de probabilidade F é auto-consistente se satisfaz

$$\tilde{F}(y) = \frac{1}{n}\sum_{i=1}^{n} H(y|X_i,\tilde{F}).$$

Esta forma de definir auto-consistência se deve a Turnbull(1974, 1976), que generalizou a definição dada em Efron(1967). Um dos atrativos deste conceito é a coincidência dos estimadores de máxima verossimilhança de F com os estimadores auto-consistentes.

Suponhamos a função $K(\cdot|y)$ limitada, implicando que a $f(\cdot|F)$ seja também limitada. Além disto, pela definição da densidade condicional de Y dado $X = x$ em (3.20), temos a

equivalência

$$h(y|x,F)=0 \;\Leftrightarrow\; K(x|y)g(y)=0 \;\; \forall F \in \Im_1 .$$

Seja $\hat{F}$ o estimador de máxima verossimilhança de F segundo o *modelo de dados incompletos*, ou seja, $\hat{F}$ satisfaz

$$l(\hat{F}) \geq l(F) \qquad \forall F \in \Im .$$

Consideremos o subconjunto das funções de distribuição de probabilidade F tais que, se $\hat{F}$ for positivo no ponto y, então F também o será, isto é, definamos o subconjunto $\Im_0$ como

$$\Im_0 = \left\{ F \in \Im_1 : \hat{F}(y) > 0 \Rightarrow F(y) > 0 \right\}.$$

Definamos a função $Q(F|\hat{F})$, para toda $F \in \Im_0$, como sendo

$$Q(F|\hat{F}) = \sum_{i=1}^{n} \int \log p(X_i, y|F) H(dy|X_i, F). \tag{3.21}$$

Lema 3.1: Para toda $F \in \Im_0$, $Q(\hat{F}|\hat{F}) \geq Q(F|\hat{F})$.

Prova.

Tomando o logaritmo em (3.20), obtemos

$$\log h(y|x,F) = \log K(x|y)g(y) - \log f(x|F),$$

para todo $F \in \Im_0$ e todo y tal que $h(y|x,\hat{F}) > 0$. Pela definição da função de densidade conjunta p em (3.19) e pela relação em (3.20), temos

$$\log p(x,y|F) = \log h(y|x,F) + \log f(x|F).$$

Então, Q pode ser escrita em função de $l(F)$, o logaritmo da verossimilhança, da seguinte forma

$$Q(F|\hat{F}) = \sum_{i=1}^{n} \int \log f(X_i|F) H(dy|X_i, \hat{F}) + \sum_{i=1}^{n} \int \log h(y|X_i, F) H(dy|X_i, \hat{F}).$$

Dado que a integral $\int H(dy|X_i, \hat{F}) = 1$, e que a função $f(X_i|F)$ não depende da variável de integração, chegamos a

$$Q(F|\hat{F}) = l(F) + \sum_{i=1}^{n} \int \log h(y|X_i, F) H(dy|X_i, \hat{F}). \qquad (3.22)$$

O Lema 1, em Wald(1949), será de muita utilidade.

Seja X uma variável aleatória com função de densidade f_{θ_0}, *então para qualquer* $\theta \neq \theta_0$, *temos que*

$$E \log f_{\theta}(X) < E \log f_{\theta_0}(X).$$

Observemos que o segundo termo em (3.22) pode ser visto como a esperança da função $\log h(\cdot|X_i, F)$. Utilizando o resultado anterior, temos

$$E\left(\log h(Y|X_i, F)\right) < E\left(\log h(Y|X_i, \hat{F})\right),$$

e $L(F) \leq L(\hat{F})$, sendo, portanto, válida a afirmação no Lema 3.1. ♦

No Teorema a seguir, provaremos que o estimador $\hat{F}$ é auto-consistente. Seguiremos as idéias em Laird(1978).

Teorema 3.1: O estimador $\hat{F}$ de máxima verossimilhança de F é auto-consistente.

Prova.

Do Lema 3.1 sabemos que $\hat{F}$ maximiza a função $Q(F|\hat{F})$. Pela definição de $p(x, y|F)$, a função em (3.21) pode ser escrita como

$$Q(F|\hat{F}) = \sum_{i=1}^{n} \int \log K(X_i|y) H(dy|X_i, \hat{F}) + \sum_{i=1}^{n} \int \log g(y) H(dy|X_i, \hat{F}),$$

para toda $F \in \Im_0$. O primeiro termo não depende de g e o segundo pode ser escrito na forma

$$n \int \log g(y) \frac{1}{n} \sum_{i=1}^{n} H(dy|X_i, \hat{F}).$$

Utilizando o Lema 1 em Wald(1949), obtemos que $Q(F|\hat{F}) \leq Q(\tilde{F}|\hat{F})$ para toda $F \in \Im_0$, onde

$$\widetilde{F}(y)=\frac{1}{n}\sum_{i=1}^{n}H(y|X_i,\hat{F}),$$

com igualdade, se e somente se, $F(y)=\widetilde{F}(y)$, quase certamente $\mu_{\widetilde{F}}$. Dado que $\hat{F}\in\Im_0$, temos que $Q(\hat{F}|\hat{F})<Q(\widetilde{F}|\hat{F})$, o que constitui uma contradição ao Lema 3.1, a menos que $F(y)=\widetilde{F}(y)$, quase certamente $\mu_{\widetilde{F}}$.

Vejamos agora que $F(y)=\widetilde{F}(y)$ para todo $y\in I$, ou seja, provemos que

$$\left\{y:F(y)=\widetilde{F}(y)\right\}=I.$$

Pela definição,

$$\widetilde{F}(y)=\frac{1}{n}\sum_{i=1}^{n}H(y|X_i,\hat{F})=\frac{1}{n}\hat{F}(y)\sum_{i=1}^{n}\frac{K(X_i|y)}{f(X_i|\hat{F})}, \qquad (3.23)$$

dado que $F\in\Im_0$, então

$$F(y)=0\Rightarrow\hat{F}(y)=0,$$

e pela expressão anterior temos

$$\hat{F}(y)=0\Rightarrow\widetilde{F}(y)=0.$$

Suponhamos agora que, para algum $y\in I$

$$\widetilde{F}(y)=0 \text{ e } \hat{F}(y)>0.$$

Logo, y pertence ao conjunto $\Omega_y=\left\{y:K(X_i|y)=0 \text{ para } i=1,\ldots,n\right\}$ e

$$\int_{\Omega_y}\hat{F}(dy)=\varepsilon>0.$$

Poderíamos então construir uma função de distribuição $\widetilde{F}'$, satisfazendo $\widetilde{F}'(y)=0$ para todo $y\in\Omega_y$ e $\widetilde{F}'(y)=\hat{F}(y)/(1-\varepsilon)$ se $y\notin\Omega_y$, tendo então $L(\widetilde{F}')=L(\hat{F})-n\log(1-\varepsilon)$ que é maior que $L(\hat{F})$, contradizendo a propriedade de máxima verossimilhança de $\hat{F}$. ♦

Desta forma, provamos que

$$\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} H(y|X_i, \hat{F}).$$

Este Teorema é a base para caracterizar $\hat{F}$ como função discreta e com um número finito de pontos de saltos (isto será visto nos Teoremas 3.2 e 3.3). Os resultados são obtidos a partir de combinações das seguintes Condições.

Condição 3.1: A função $K(x|y) > 0$ é analítica para todo $y \in I$.

Condição 3.2: Para algum k=1,2,... a k-ésima derivada com respeito a y de $K(x|y)$ é não negativa ou não positiva, e é distinta de zero em algum ponto $y_0 \in I$.

Condição 3.3: Existe um intervalo fechado $[a,b]$ em I, tal que para qualquer par de pontos (y,y'), $y \in [a,b]$, $y' \notin [a,b]$, satisfaz-se $K(x|y) > K(x|y')$.

As Condições 3.1 e 3.2 são suficientes para provar que $K(x|y)$ é positiva num número discreto de pontos em I (Teorema 3.2). A Condição 3.1 é uma imposição direta sob a densidade condicional $K(x|y)$ e a Condição 3.2 é fácil de verificar.

Por exemplo, se X for normal com média y e variância θ, a Condição 3.2 será válida para k=2 e $y_0 = x - C - \theta^{1/2}$, onde $C > 0$ é uma constante real. De fato

$$K(x|y) = \frac{C}{\theta} e^{-(x-y)^2/2\theta},$$

e a segunda derivada com respeito a y é

$$\frac{\partial^2}{\partial y^2} K(x|y) = \frac{C(x-y)^2}{\theta^3 e^{-(x-y)^2/2\theta}} - \frac{C}{\theta^2 e^{-(x-y)^2/2\theta}}.$$

Temos que provar, por exemplo, que existe um y_0 para o qual a derivada acima é positiva. Isto se reduz a encontrar um y_0 para o qual $\frac{(x-y)^2}{\theta} > 1$.

Observemos que, se $y_0 = x - C - \theta^{1/2}$, temos

$$\frac{(x-(x - C - \theta^{1/2}))^2}{\theta} > \frac{(\theta^{1/2})^2}{\theta} = 1.$$

A Condição 3.3 é satisfeita sempre que a função $K(x|y)$ tenha um número finito de modas finitas. Esta condição não constitui uma restrição relevante.

Os Teoremas a seguir foram demonstrados em Laird(1978) e caracterizarão o estimador $\hat{F}$.

Teorema 3.2: Sob as Condições 3.1 e 3.2, o estimador $\hat{F}$ tem salto positivo somente em pontos discretos.

Prova.

Provaremos que o estimador $\hat{F}$ não se pode incrementar positivamente de maneira contínua em nenhum intervalo. Pelo Teorema 3.1, expressão (3.23), temos que $\hat{F}$ é auto-consistente, isto é, satisfaz

$$d\hat{F}(y)\left\{n-\sum_{i=1}^{n}\frac{K(X_i|y)}{f(X_i|\hat{F})}\right\}=0$$

para todo $y\in I$.

Suponhamos que exista algum intervalo $[c,d]\subseteq I$ tal que $d\hat{F}(y)>0$ para todo $y\in[c,d]$. Então a função

$$W(y)=\sum_{i=1}^{n}\frac{K(X_i|y)}{f(X_i|\hat{F})}-n$$

teria que ser nula para todo $y\in[c,d]$.

Pela Condição 3.1 a função $W(y)$ é analítica em $[c,d]$, podemos então desenvolver $W(y)$ em série de Taylor em torno do ponto y_0 indicado na Condição 3.2. Desta forma teremos uma série de potências de $(y-y_0)$ identicamente nula, implicando que todos os coeficientes são nulos.

No entanto, dado que $0<f(X_i|F)<\infty$ para todo i, o coeficiente do k-ésimo termo do desenvolvimento em série de Taylor da função W tem que ser positivo ou negativo. Logo, pela Condição 3.2, W não pode ser identicamente zero em I. ♦

É de interesse saber quantos saltos tem a função $\hat{F}$. Em geral, mesmo em situações simples, não é possível contar o número de saltos, mas é possível estabelecer condições para que esta quantidade seja finita.

Teorema 3.3: Sob as Condições 3.1, 3.2 e 3.3, $\hat{F}$ possui um número finito de saltos.

Prova.

Observamos primeiramente que os pontos nos quais a função $\hat{F}$ tem incremento positivo se encontram unicamente no intervalo $[a,b]$ definido na Condição 3.3. Para ver isto, observemos também que o estimador $\hat{F}$ induz um estimador de máxima verossimilhança $\hat{f}$ através da convolução em (3.18), isto é, através da transformação

$$\hat{f}(x|\hat{F}) = \int K(x|y)\hat{F}(dy),$$

pela propriedade de invariância dos estimadores de máxima verossimilhança.

Definamos por L o conjunto dos pontos de I nos quais o estimador $\hat{F}$ incrementa-se positivamente, isto é, seja

$$L = \{y \in I : \hat{F}(y^+) - F(y) > 0\}.$$

Observemos que $L \subseteq [a,b]$. Para ver isto, suponhamos que $y' \in L$ é externo ao intervalo $[a,b]$. Assim, o estimador $\hat{f}$ não poderia conservar a propriedade de ser máximo da função de verossimilhança em relação a algum outro estimador que seja positivo no mesmo conjunto L e que também seja positivo em algum outro ponto $z \in L$, que não pertença ao intervalo $[a,b]$, pela Condição 3.3.

Sabemos também pelo Teorema 3.2 que os $y \in L$ são pontos isolados, então o número

$$\varepsilon = \min_{y',y'' \in L} |y' - y''|,$$

é estritamente positivo.

Desta forma temos que a quantidade de pontos em L será no máximo

$$\frac{b-a}{\varepsilon}.$$ ♦

Dado que o estimador de máxima verossimilhança $\hat{F}$ da função de distribuição de probabilidade F é auto-consistente, ele também será uma função de distribuição de probabilidade. Como consequência dos Teoremas 3.2 e 3.3, este estimador será caracterizado univocamente pelas localizações dos saltos $\hat{\lambda}_1,\ldots,\hat{\lambda}_k$, $\hat{\lambda}_j \in I$ e pelas magnitudes $\hat{\pi}_j$, $j=1,\ldots,k$

de cada salto.

Desta forma reduzimos a estimação de $\hat{F}$ à estimação dos parâmetros de uma mistura de k funções de densidades K. Este problema foi tratado, por exemplo, em Hasselblad(1966, 1969), Dempster *et al* (1977), Peter & Walker(1978), Redner & Walker(1984), entre outros.

3.4.2 O estimador $\hat{f}$ supondo o *modelo de dados incompletos*

Vejamos como o estudo do estimador $\hat{F}$, na Seção anterior, pode ser utilizado para obter a forma do estimador $\hat{f}$ da função de densidade através de *"sieves"* de convolução.

Provaremos primeiramente que, segundo o *modelo de dados incompletos*, o estimador $\hat{f}$ é obtido substituindo-se na convolução em (3.1) a função de distribuição de probabilidade F pelo seu estimador de máxima verossimilhança $\hat{F}$, isto é, obteremos que

$$\hat{f}(x) = \int mK(m(x-y)\hat{F}(dy),$$

pertence ao conjunto

$$M_{m_n}^n = \left\{ \hat{f}_{m,n} \in S_{m_n} : \hat{f}_{m,n} = arg \max_{g \in S_{m_n}} \prod_{i=1}^n g(X_i) \right\},$$

definido em (2.31).

Lema 3.2: Seja $\hat{f}(x) = \int m_n K(m_n(x-y))\hat{F}(dy)$, então $\hat{f} \in M_{m_n}^n$.

Prova.

A prova constitui uma aplicação direta da propriedade de invariância dos estimadores de máxima verossimilhança obtida por Zehna(1966).

Sejam $F \in \Im$ o conjunto de todas as funções de distribuição de probabilidade definido em (3.1), $L(F): \Im \to R$ a função de verossimilhança e $\hat{F} \in \Im$ o ponto em $\Im$ tal que $L(\hat{F}) \geq L(F)\ \forall F \in \Im$, isto é, o estimador de máxima verossimilhança de F.

Definamos a função $u: \Im \to \Lambda$, onde Λ é o conjunto das funções de densidade f da forma

$$f(x) = \int mK(m(x-y))F(dy).$$

Definamos também $\hat{f} \in \Lambda$ como $\hat{f} = u(\hat{F})$, $\hat{f}$ é único por ser u função. Para cada $f \in \Lambda$, definamos $\Im_f = \{F \in \Im : u(F) = f\}$ e $M_f = \sup_{F \in \Im_f} L(F)$, M_f é uma função real chamada de *função de verossimilhança induzida* pela transformação u.

Desta forma $M_{\hat{f}} = L(\hat{F})$ e, dado que

$$M_f = \sup_{F \in \Im_f} L(F) \leq \sup_{F \in \Im} L(F) = L(\hat{F}) = M_{\hat{f}} \quad \forall f \in \Lambda. \qquad \blacklozenge$$

Como foi visto, fixando o valor de m_n, os elementos do conjunto S_{m_n} podem ser obtidos pelo *modelo de dados incompletos* e a convolução

$$f(x|F) = \int K(x|y)F(dy).$$

O Teorema 3.3 caracteriza o estimador $\hat{F}$ e, o Lema 3.2 nos garante que se utilizamos $\hat{F}$ na convolução $\int K(x|y)F(dy)$, podemos obter de forma alternativa à Proposição 3.3 o estimador $\hat{f}$ via *"sieves"* de convolução, este resultado apresenta-se a seguir.

Proposição 3.4: Sob as Condições 3.1, 3.2 e 3.3 o estimador *"sieves"* de convolução $\hat{f}(x) = \int m_n K\big(m_n(x-y)\big)\hat{F}(dy)$ será da forma

$$\hat{f}(x) = \sum_{j=1}^{k} m_n \hat{\pi}_j K\Big(m_n(x - \hat{\lambda}_j)\Big), \tag{3.24}$$

onde k é finito.

Prova.

Válidas as Condições 3.1, 3.2 e 3.3 dos Teoremas 3.2 e 3.3, podemos escrever $\hat{F}$ como

$$\hat{F}(y) = \sum_{j=1}^{k} \hat{\pi}_j \mathbf{1}_{\{\hat{\lambda}_j\}}(y).$$

Desta forma, temos que o estimador de *"sieves"* de convolução será

$$\hat{f}(x) = \int m_n K(m_n(x-y))\hat{F}(dy) = \int m_n K(m_n(x-y)) \sum_{j=1}^{k} \hat{\pi}_j 1_{\{\hat{\lambda}_j\}}(y)dy,$$

de onde obtemos

$$\hat{f}(x) = \sum_{j=1}^{k} \hat{\pi}_j m_n \int K(m_n(x-y)) 1_{\hat{\lambda}_j}(y)dy = \sum_{j=1}^{k} \hat{\pi}_j m_n K\left(m_n(x-\hat{\lambda}_j)\right). \quad \blacklozenge$$

Passaremos agora a estudar dois casos específicos. O primeiro deles é o *"sieves"* de convolução gaussiano que pode ser obtido a partir dos raciocínios, tanto da Proposição 3.3 quanto da Proposição 3.4. A importância dele é que, utilizando os procedimentos contidos na Proposição 3.3 prova-se que q é finito, ou seja, neste caso teremos uma soma finita em (3.7) e obtemos condições sob a quantidade de somandos.

Outro caso particular é o *"sieves"* de convolução por exponencial dupla, que devido à não-diferenciabilidade da função núcleo, não pode ser estudado pelos raciocínios contidos nas Proposições 3.3 ou 3.4, mas será provado que o estimador assume uma forma similar àquele apresentado em (3.24).

3.5 O *"sieves"* de convolução gaussiano

Se na definição dos subconjuntos que formam a seqüência *"sieves"*, escolhemos como função núcleo a gaussiana, obtemos o chamado *"sieves"* de convolução gaussiano, isto é, diremos que $\hat{f}$ é um estimador *"sieves"* de convolução gaussiano se

$$\hat{f}(x) = \int \frac{m_n}{\sqrt{2\pi}} e^{-\frac{m_n^2(x-y)^2}{2}} F(dy).$$

Foi mostrado que o número de somandos q, nesta situação, é no máximo n. Este resultado é obtido a partir da Proposição 3.2 e está baseado nas idéias da Proposição 1, página A31 em Geman(1981).

Proposição 3.5: Seja $\hat{f} \in M_{m_n}^n$, com função núcleo gaussiana. Então $\hat{f}$ tem a forma

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{p_i m_n^2}{\sqrt{2\pi}} e^{-\frac{m_n^2(x-y_i)^2}{2}},$$

para alguns $y_1,\ldots,y_n$ e $p_1,\ldots,p_n$ satisfazendo $p_i \geq 0 \ \forall\, i=1,2,\ldots,n$ e $\sum_{i=1}^{n} p_i = 1$. Além disso, se

$X_{(1)} < max(y_1,..., y_n)$, então $X_{(1)} < min(y_1,..., y_n) \leq max(y_1,..., y_n) < X_{(n)}$.

Prova.

A partir da forma dos estimadores *sieves* de convolução obtida na Proposição 3.3, provaremos que a quantidade de somandos em (3.7) é finita. Para isto, estudaremos a seguir a função

$$\sum_{i=1}^{n} \frac{1}{\hat{f}(X_i)} m_n \frac{\partial}{\partial z} K\big(m_n(X_i - y + z)\big)\Big|_{z=0},$$

obtida em (3.17).

A derivada na expressão anterior assume a forma

$$\frac{\partial}{\partial z} m_n K\big(m_n(x - y + z)\big)\Big|_{z=0} = m_n^2(x - y + z) e^{\left.\frac{m_n^2(x-y+z)^2}{2}\right|_{z=0}}.$$

Avaliando em $z = 0$, temos

$$\frac{\partial}{\partial z} m_n K\big(m_n(x - y + z)\big)\Big|_{z=0} = -m_n^2(x - y) e^{\frac{m_n^2(x-y)^2}{2}}.$$

Segue que a expressão em (3.16) será

$$T(y) = \sum_{i=1}^{n} \frac{(X_i - y)}{\hat{f}(X_i)} e^{\frac{m_n^2(X_i-y)^2}{2}}.$$

O conjunto de funções

$$\left\{ \frac{(x_i - y)}{\hat{f}(x_i)} e^{\frac{m_n^2(x_i-y)^2}{2}} \right\}_{i=1}^{n},$$

forma um *sistema extendido de Tchebychev*, ou *ET-system* (veja Apêndice II). Duas conseqüências disto são:

i) $Z_1 = \{y: T(y) = 0\}$ tem no máximo $2n$-1 elementos,

ii) $Z_2 = \{y: T(y) = 0, {}^{d}/_{dy}\, T(y) \le 0\}$ tem no máximo n elementos.

Dado que o número de elementos de Z_1 define a quantidade de somandos q em (3.7), temos que estes são no máximo $2n$-1. Provaremos a seguir que q é no máximo n; para isto temos que mostrar

$$ {}^{d}/_{dy}\, T(y)\big|_{y=s} \le 0, \tag{3.25} $$

para qualquer s no suporte de $\hat{F}$.

Como $\hat{f} \in M^{n}_{m_n}$, pela Proposição 3.3, podemos escrever

$$ \hat{f}(x) = \sum_{i=1}^{q} p_i m_n K\big(m_n(x - s_i)\big), $$

onde $\{s_1, \ldots, s_q\}$ é o suporte de $\hat{F}$, $q \le 2n$-1, $p_i > 0$, i=1,2, ... ,q e $\sum_{i=1}^{q} p_i = 1$.

Fixando $l \in \{1,2, \ldots, q\}$, definimos f_ε para cada ε>0 como

$$ f_\varepsilon(x) = \sum_{\substack{i=0 \\ i \ne l}}^{q+1} p_i m_n K\big(m_n(x - s_i)\big) + \frac{p_l m_n}{2} K\big(m_n(x - s_l - \varepsilon)\big) + \frac{p_l m_n}{2} K\big(m_n(x - s_l + \varepsilon)\big), $$

onde $s_0 = -\infty$ e $s_{q+1} = +\infty$.

Para provar que $f_\varepsilon \in S_m$, temos que achar uma função de distribuição de probabilidade F_ε tal que $f_\varepsilon(x) = \int m_n K(m_n(x - y)) F_\varepsilon(dy)$.

Para isto definamos,

$$ F_\varepsilon(y) = \sum_{\substack{k=1 \\ k \ne l}}^{q} \left(\sum_{j=1}^{k} p_j \right) \mathbf{1}_{[s_{k-1}, s_k)}(y) + \left(\sum_{j=1}^{k} p_j + \frac{p_l}{2} \right) \mathbf{1}_{[s_{l-1}, s_l - \varepsilon)}(y) + \left(\sum_{j=1}^{k} p_j + \frac{p_l}{2} \right) \mathbf{1}_{[s_{l-1}, s_l + \varepsilon)}(y). $$

Observemos que se $\varepsilon \to 0$, $f_\varepsilon \to \hat{f} \in M_{m_n}^n$ e, com isto,

$$\frac{\partial^2}{\partial \varepsilon^2} \sum_{i=1}^{n} \log f_\varepsilon(X_i) \Big|_{\varepsilon=0} \leq 0. \tag{3.26}$$

Pensemos aqui que, o estimador $\hat{f}$ maximiza a função de verossimilhança em S_{m_n}, com isto, a segunda derivada tem que ser negativa.

Calculando a derivada em (3.26),

$$\begin{aligned} \frac{\partial^2}{\partial \varepsilon^2} \sum_{i=1}^{n} \log f_\varepsilon(X_i) \Big|_{\varepsilon=0} &= \frac{\partial}{\partial \varepsilon} \sum_{i=1}^{n} \frac{\partial}{\partial \varepsilon} \log f_\varepsilon(X_i) \Big|_{\varepsilon=0} = \\ &= \frac{\partial}{\partial \varepsilon} \sum_{i=1}^{n} \frac{\partial / \partial \varepsilon f_\varepsilon(X_i)}{f_\varepsilon(X_i)} \Big|_{\varepsilon=0}. \end{aligned} \tag{3.27}$$

Pela definição de f_ε, temos que, a derivada no numerador em (3.27) é dada por

$$\partial / \partial \varepsilon f_\varepsilon(x) = \frac{p_l m_n^3}{2} \left((x - s - \varepsilon) e^{\frac{m_n^2 (x - s_l - \varepsilon)^2}{2}} - (x - s + \varepsilon) e^{\frac{m_n^2 (x - s_l + \varepsilon)^2}{2}} \right).$$

Consequentemente

$$\frac{\partial}{\partial \varepsilon} \sum_{i=1}^{n} \frac{\partial / \partial \varepsilon f_\varepsilon(X_i)}{f_\varepsilon(X_i)} \Big|_{\varepsilon=0} =$$

$$= \frac{\partial}{\partial \varepsilon} \sum_{i=1}^{n} \frac{\frac{p_l m_n^3}{2} \left((X_i - s_l - \varepsilon) e^{\frac{m_n^2 (X_i - s_l - \varepsilon)^2}{2}} - (X_i - s_l + \varepsilon) e^{\frac{m_n^2 (X_i - s_l + \varepsilon)^2}{2}} \right)}{f_\varepsilon(X_i)} \Bigg|_{\varepsilon=0}.$$

Derivando em (3.27), obtemos

$$\frac{\partial^2}{\partial\varepsilon^2}\sum_{i=1}^{n}\log f_\varepsilon(X_i)\Big|_{\varepsilon=0}=\sum_{i=1}^{n}\frac{\left(\frac{\partial^2}{\partial\varepsilon^2}f_\varepsilon(X_i)\right)f_\varepsilon(X_i)-\left(\frac{\partial}{\partial\varepsilon}f_\varepsilon(X_i)\right)^2}{f_\varepsilon^2(X_i)}\Bigg|_{\varepsilon=0}.$$

Dado que $\frac{\partial}{\partial\varepsilon}f_\varepsilon(x)\Big|_{\varepsilon=0}=0$ e

$$\frac{\partial^2}{\partial\varepsilon^2}f_\varepsilon(x)\Big|_{\varepsilon=0}=\frac{p_l m_n^3}{2}\left\{\left(m_n^2(x-s_l-\varepsilon)^2-1\right)e^{\frac{m_n^2(x-s_l-\varepsilon)^2}{2}}-\left(m_n^2(x-s_l+\varepsilon)^2-1\right)e^{\frac{m_n^2(x-s_l-\varepsilon)^2}{2}}\right\}\Bigg|_{\varepsilon=0},$$

obtemos

$$\frac{\partial^2}{\partial\varepsilon^2}\sum_{i=1}^{n}\log f_\varepsilon(X_i)\Big|_{\varepsilon=0}=p_l m_n^2\sum_{i=1}^{n}\frac{\left(m_n^2(X_i-s_l)^2-1\right)e^{\frac{m_n^2(X_i-s_l)^2}{2}}}{\sum_{j=1}^{q}p_j e^{\frac{m_n^2(X_i-s_j)^2}{2}}}.$$

Segue que a derivada em (3.25) é

$$\frac{d}{dy}T(y)\Big|_{y=s_l}=\frac{d}{dy}\sum_{i=1}^{n}\frac{(X_i-y)e^{\frac{m_n^2(X_i-y)^2}{2}}}{\sum_{j=1}^{q}p_j m_n e^{\frac{m_n^2(X_i-s_j)^2}{2}}}\Bigg|_{y=s_l}=$$

$$=\sum_{i=1}^{n}\frac{\left(m_n^2(X_i-s_l)^2-1\right)e^{\frac{m_n^2(X_i-s_l)^2}{2}}}{\sum_{j=1}^{q}p_j e^{\frac{m_n^2(X_i-s_j)^2}{2}}}.$$

Com isto

$$\frac{\partial^2}{\partial \varepsilon^2}\sum_{i=1}^{n} \log f_{\varepsilon}(X_i)\Big|_{\varepsilon=0} = p_l m_n^2 \frac{d}{dy}T(y)\Big|_{y=s_l} \leq 0.$$

Logo $\frac{d}{dy}T(y) \leq 0$, para qualquer s no suporte de $\hat{F}$. Isto prova que $q \leq n$, como conseqüência de *ii)*.

A última afirmação nesta Proposição, refere-se ao suporte do estimador $\hat{F}$. Se algum s no suporte de $\hat{F}$ é tal que $s \leq X_{(1)}$, então para todo $i = 1,2,\ldots,n$ $\hat{f}(X_i)$ é estritamente crescente em função de s (lembremos que o máximo da função núcleo K é $K(0)$), implicando que $\prod_{i=1}^{n}\hat{f}(X_i)$ seja crescente, isto contradiz o fato de $\hat{f} \in M_{m_n}^{n}$. Se $s \geq X_{(n)}$ obteríamos, de maneira similar, que $\prod_{i=1}^{n}\hat{f}(X_i)$ seria uma função decrescente em s, contradizendo que $\hat{f} \in M_{m_n}^{n}$. ♦

3.6 O *"sieves"* de convolução via exponencial dupla

Este tipo de estimador é tratado em Walter & Blum(1984). Devido à não diferenciabilidade da função núcleo, não é possível considerar este estimador como caso particular dos propostos nas seções anteriores. Seguindo as idéias de Walter & Blum(1984), chegaremos a um resultado próximo àqueles apresentados nas Proposições 3.3 e 3.4.

A importância do trabalho com esta função é que, a pesar de não ser diferenciável, possui propriedades desejáveis nas funções núcleo, como a de ser simétrica e contínua. Desta forma consegue-se ampliar o conjunto das funções a serem utilizadas como núcleo na convolução com a qual se obtêm os elementos dos conjuntos das seqüência *"sieves"*.

A seqüência *"sieves"*, utilizando a função núcleo exponencial dupla, é da forma,

$$S_{m_n}''' = \left\{ f \in \Theta : f(x) = \int \frac{m_n}{2} e^{-m_n|x-y|} F(dy) \right\}. \qquad (3.28)$$

Associado à S_{m_n}''' temos o conjunto,

$$M_{m_n}''' = \left\{ \hat{f} \in S_{m_n}''' : \hat{f}(x) = arg \max_{g \in S_{m_n}'''} \prod_{i=1}^{n} g(X_i) \right\}.$$

Como sempre, a problemática é identificar a função de distribuição F que gera os elementos de M_{m_n}'''. Isto é feito na seguinte Proposição.

Proposição 3.6: Se $\hat{f} \in M_{m_n}'''$, então é da forma

$$\hat{f}(x) = \frac{m_n}{2} \sum_{i=1}^{q} p_i e^{-m_n|x-y_i|},$$

onde q não é necessariamente finito.

Prova.

Suponhamos $\hat{f}$ uma função contínua. Foi provado em Laird(1978) que o estimador de máxima verossimilhança é auto-consistente, isto é, satisfaz a relação

$$\frac{1}{n}\sum_{i=1}^{n} \frac{e^{-m_n|X_i-y|}\hat{f}(y)}{\int e^{-m_n|X_i-z|}\hat{f}(z)dz} = \hat{f}(y). \tag{3.29}$$

Notemos que $e^{-m_n|x-y|}$ satisfaz a equação diferencial

$$(D^2 - m_n^2)e^{-m_n|x-y|} = -2m_n 1_{\{0\}}(x-y). \tag{3.30}$$

Seja (a,b) um intervalo no qual $\hat{f} > 0$ e, em (a,b), definamos $h_{m_n}(x) = \int e^{-m_n|x-y|}\hat{f}(z)dz$. A idéia desta demonstração é provar que o estimador $\hat{f}$ não pode ser contínuo. Para isto utilizando a equação diferencial (3.30) chega-se a uma relação que contradiz a propriedade de auto-consistência de $\hat{f}$.

Da equação diferencial em (3.30) temos que

$$D^2 e^{-m_n|x-y|} - m_n^2 e^{-m_n|x-y|} = -2m_n 1_{\{0\}}(x-y).$$

Dividindo em ambos os lados da equação anterior por $h_{m_n}(x)$, temos

$$\frac{D^2 e^{-m_n|x-y|}}{h_{m_n}(x)} - \frac{m_n^2 e^{-m_n|x-y|}}{h_{m_n}(x)} = -\frac{2m_n 1_{\{0\}}(x-y)}{h_{m_n}(x)}.$$

Avaliando em $x = X_i$, somando em i e multiplicando por $1/n$, obtemos

$$\frac{1}{n}\sum_{i=1}^{n} \frac{D^2 e^{-m_n|X_i-y|}}{h_{m_n}(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{m_n^2 e^{-m_n|X_i-y|}}{h_{m_n}(X_i)} = -\frac{1}{n}\sum_{i=1}^{n} \frac{2m_n 1_{\{0\}}(X_i-y)}{h_{m_n}(X_i)}.$$

Assim,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{D^2 e^{-m_n|X_i-y|}}{\int e^{-m_n|X_i-y|}\hat{f}(z)dz} = \frac{1}{n}\sum_{i=1}^{n} \frac{m_n^2 e^{-m_n|X_i-y|}}{\int e^{-m_n|X_i-y|}\hat{f}(z)dz} - \frac{1}{n}\sum_{i=1}^{n} \frac{2m_n 1_{\{0\}}(X_i-y)}{\int e^{-m_n|X_i-y|}\hat{f}(z)dz},$$

ou seja,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{D^2 e^{-m_n|X_i-y|}}{\int e^{-m_n|X_i-y|}\hat{f}(z)dz} = \frac{1}{n}\sum_{i=1}^{n}\frac{m_n^2 e^{-m_n|X_i-y|}}{\int e^{-m_n|X_i-y|}\hat{f}(z)dz} - \frac{1}{n}\sum_{i=1}^{n}\frac{2m_n 1_{\{0\}}(X_i-y)}{\int e^{-m_n|X_i-y|}\hat{f}(z)dz}.$$

Dividindo por $\hat{f}(y)$ ambos os lados em (3.30), temos que

$$\frac{1}{n}\sum_{i=1}^{n}\frac{e^{-m_n|X_i-y|}}{\int e^{-m_n|X_i-z|}\hat{f}(z)dz} = 1.$$

Derivando duas vezes, ou seja, aplicando o operador D^2 e utilizando a relação (3.30), obtemos

$$D^2\frac{1}{n}\sum_{i=1}^{n}\frac{e^{-m_n|X_i-y|}}{\int e^{-m_n|X_i-y|}\hat{f}(z)dz} = \frac{1}{n}\sum_{i=1}^{n}\frac{m_n^2 e^{-m_n|X_i-y|}}{\int e^{-m_n|X_i-y|}\hat{f}(z)dz} - \frac{1}{n}\sum_{i=1}^{n}\frac{2m_n 1_0(X_i-y)}{\int e^{-m_n|X_i-y|}\hat{f}(z)dz} = 0.$$

A igualdade anterior somente é verificada quando $y = X_i$, $i = 1,2,\ldots,n$. Isto contradiz o fato de $\hat{f} > 0$ no intervalo *(a,b)*, concluindo-se então que $\hat{f}$ é discreta.

Capítulo IV:

Algoritmo EM. Aplicação aos modelos de misturas finitas de densidades

4.1 Introdução

O objetivo do algoritmo EM é estimar, por máxima verossimilhança, os parâmetros de uma função de densidade de probabilidade $\alpha(x|\theta)$, $\theta = (\theta_1, \ldots, \theta_s) \in \mathbb{R}^s$, da qual se tem uma amostra aleatória $X_1, \ldots, X_n$. Este algoritmo será útil nas situações onde não é possível obter de maneira explícita a expressão dos estimadores como solução da equação de verossimilhança. Por exemplo, Chapman(1956) chegou a este algoritmo tentando estimar os parâmetros da densidade Gama quando ambos os parâmetros são desconhecidos.

No contexto da estimação de funções de densidade, especificamente no caso da estimação via *"sieves"* de convolução, foi visto que o problema da estimação não paramétrica de funções de densidade se reduz a um problema de estimação paramétrica. Nas expressões (3.7) e (3.24), no Capítulo III se obtém a forma do estimador que corresponde ao *"sieves"* de convolução.

Na descrição do algoritmo utilizaremos a verossimilhança associada a $\alpha(x|\theta)$, esta será

$$L(\theta) = \prod_{j=1}^{n} \alpha(x_j|\theta), \tag{4.1}$$

ou, equivalentemente, o logaritmo da função de verossimilhança

$$\log L(\theta) = \sum_{j=1}^{n} \log \alpha(x_j|\theta). \tag{4.2}$$

A partir desta função, obtém-se a equação de verossimilhança

$$\frac{\partial}{\partial\theta} \log L(\theta) = \sum_{j=1}^{n} \frac{\partial}{\partial\theta} \log \alpha(x_j|\theta) = 0. \tag{4.3}$$

Trataremos aqui com funções de densidade identificáveis e consideraremos válidas as condições de regularidade(Cramér, 1946) exigidas da função de densidade para se provar a

consistência dos estimadores de máxima verossimilhança. Sob estas condições, temos:

Teorema 4.1: Dadas as condições de regularidade(Cramér, 1946), temos que se θ^* for o verdadeiro valor do vetor de parâmetros θ, com probabilidade 1, para n suficientemente grande, uma única solução $\hat{\theta}$ da equação de verossimilhança (4.3) maximiza a função $logL$ e $\sqrt{n}(\hat{\theta}-\theta^*)$ tem distribuição assintoticamente normal com média zero e matriz de covariâncias sendo a inversa da matriz de informação de Fisher.

Assumiremos funções de densidade nas quais não é possível obter uma forma fechada do estimador $\hat{\theta}$ através da equação de verossimilhança. Uma solução então é utilizar algoritmos numéricos para solucionar esse problema. Com esse objetivo é que estudaremos o algoritmo EM, o qual permitirá estimar o vetor de parâmetros $\theta = \{p_1,\ldots,p_n,y_1,\ldots,y_n,\sigma\}$ que define a função

$$\alpha(x) = \sum_{j=1}^{n} \frac{p_j}{\sqrt{2\pi\sigma}} e^{\frac{(x-y_j)^2}{2\sigma^2}}.$$

Na próxima Seção exporemos as fundamentações teóricas do algoritmo EM. Posteriormente o utilizaremos no caso de uma mistura finita de densidades e na última Seção apresentamos propriedades assintóticas.

4.2 Generalidades

Suponhamos $\theta \in R^s$, que temos uma amostra aleatória $X_1,\ldots,X_n$ da variável X e denotaremos o espaço amostral por Ω. Assumamos também que cada X_i possa ter sido observada de algum dos m espaços amostrais $\Omega_1,\ldots,\Omega_m$, que formam uma partição do espaço amostral, isto é

$$\Omega_i \cap \Omega_j = \varnothing \ \forall i \neq j \quad \text{e} \quad \bigcup_{i=1}^{m} \Omega_i = \Omega.$$

O vetor de variáveis aleatórias (X,I) é chamado de modelo de dados completos, onde X é variável aleatória real e I é o correspondente vetor aleatório indicador de em qual elemento da partição $\Omega_1,\ldots,\Omega_m$ foi observada cada X_i. Se $(X_1,I_1),\ldots,(X_n,I_n)$ é uma amostra aleatória do modelo de dados completos, as X_i constituem uma amostra aleatória da variável real X e cada $I_k = (I_{1k},\ldots,I_{mk})$, onde $I_{rt} \in \{0,1\}\ \forall rt$ e satifaz

$$\sum_{l=1}^{m} I_{lk} = 1.$$

Se no vetor de variáveis aleatórias (X,I), I for não observável, não podemos identificar de qual população os X_i da amostra foram obtidos. Teremos então um modelo de dados incompletos.

Apesar de no vetor (X,I), a variável I não ser contínua, chamaremos função de densidade a função conjunta a $\alpha(x,i|\theta)$, que constitui a função de densidade do modelo de dados completos e assumamos que esta densidade pertence à família exponencial. Sob estas condições teremos

$$\alpha(x,i|\theta) = b(x,i)exp(\theta\, t(x,i)^T)a^{-1}(\theta), \tag{4.4}$$

onde $a(\theta)$ é dada por

$$a(\theta) = \int_{-\infty}^{\infty} \sum_{i\in\{0,1\}^m} b(x,i)\ exp(\theta\, t(x,i)^T)\mathrm{dx}. \tag{4.5}$$

A estatística $t(X,I)$ é um vetor s-dimensional de estatísticas suficientes do modelo de dados completos.

Para obter o estimador de θ através da verossimilhança, primeiramente achamos a função de densidade de probabilidade $\alpha(x|\theta)$. Para isto,

$$\alpha(x|\theta) = \sum_{i\in\{0,1\}^m} \alpha(x,i|\theta) =$$

$$\alpha(x|\theta) = \frac{1}{a(\theta)} \sum_{i\in\{0,1\}^m} b(x,i)exp(\theta\, t(x,i)^T) = \frac{a(\theta|x)}{a(\theta)}. \tag{4.6}$$

A densidade condicional de $(I|X=x,\theta)$ será

$$k(i|x,\theta) = \frac{\alpha(x,i|\theta)}{\alpha(x|\theta)} =$$

$$\frac{b(x,i)\ exp(\theta\, t(x,i)^T)/a(\theta)}{a(\theta|x)/a(\theta)} = b(x,i)exp(\theta\, t(x,i)^T)/a(\theta|x). \tag{4.7}$$

Logo, a função de densidade de probabilidade $k(i|x,\theta)$ também pertence à família exponencial. As funções definidas em (4.4), (4.6) e (4.7) nos permitem escrever a função de log-verossimilhança em (4.2) da seguinte forma,

$$logL(\theta) = \sum_{j=1}^{n} \log a(x_j|\theta) = \sum_{j=1}^{n} \left(\log a(x_j,i|\theta) - \log k(i|X_j = x_j,\theta) \right) =$$

$$= \sum_{j=1}^{n} (-\log\ a(\theta) + \log\ b(x_j,i) + \theta\, t(x_j,i)^T \log\ a(\theta|x_j) - \log\ b(x_j,i) - \theta\, t(x_j,i)^T)$$

$$= \sum_{j=1}^{n} \left(-\log a(\theta) + \log a(\theta|x_j) \right).$$

Portanto

$$logL(\theta) = -\, n \log a(\theta) + \sum_{j=1}^{n} \log a(\theta|x_j). \qquad (4.8)$$

Temos então que

$$\frac{\partial}{\partial\theta} \log L(\hat{\theta}) = -\, n \frac{\partial}{\partial\theta} \log a(\hat{\theta}) + \sum_{j=1}^{n} \frac{\partial}{\partial\theta} \log a(\hat{\theta}|x_j) = 0. \qquad (4.9)$$

Logo

$$\frac{\partial}{\partial\theta} \log a(\hat{\theta}) = \frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial\theta} \log a(\hat{\theta}|x_j).$$

Por propriedade da família exponencial, sabemos que

$$\frac{\partial}{\partial\theta} \log a(\theta) = E(t(X,I)^T|\theta), \qquad (4.10)$$

e que

$$\frac{\partial}{\partial\theta} \log a(\theta|x) = E(t(x,I)^T|X = x,\theta). \qquad (4.11)$$

Substituindo as expressões (4.10) e (4.11) em (4.9), temos

$$\frac{\partial}{\partial\theta} \log L(\hat{\theta}) = -nE(t(X,I)^T|\hat{\theta}) + \sum_{j=1}^{n} E(t(x_j,I)^T|X_j = x_j,\hat{\theta}). \qquad (4.12)$$

Como conseqüência desta relação, temos que a equação de log-verossimilhança (4.9) se reduz a igualdade entre média das esperanças condicional e não condicional da estatística suficiente, isto é

$$E(t(X,I)^T|\hat{\theta}) = \frac{1}{n}\sum_{j=1}^{n} E(t(x_j,I)^T|X_j = x_j,\hat{\theta}). \qquad (4.13)$$

Sabemos que $E(t(x,I)^T|X = x,\hat{\theta})$ não é possível de ser calculada, porém será de muita importância na definição do algoritmo. Para escrever de outra maneira a função de verossimilhança, definamos as funções

$$Q(\theta|\theta') = E(\log f(x,I|\theta)|X=x,\theta'), \qquad (4.14)$$

$$H(\theta|\theta') = E(\log k(I|x,\theta)|X=x,\theta'), \qquad (4.15)$$

para todo par (θ,θ').

Baseados nestas funções escreveremos $logL(\theta)$, definiremos o algoritmo EM e apresentaremos propriedades importantes.

Da expressão (4.14) temos

$$Q(\theta|\theta') = E(\log a(x,I|\theta)|X=x,\theta') = E(-\log a(\theta) + \log b(x,I) + \theta t(x,I)^T|X=x,\theta')$$

$$= -\log a(\theta) + E(\log b(x,I|X=x,\theta') + \theta E(t(x,I)^T|X=x,\theta'). \qquad (4.16)$$

Esta função definirá o algoritmo. A função definida em (4.15) poderá ser escrita como

$$H(\theta|\theta') = E(\log k(I|x,\theta)|X=x,\theta') = E(-\log a(\theta|x) + \log b(x,I) + \theta t(x,I)^T|X=x,\theta')$$

$$= -\log a(\theta|x) + E(\log b(x,I)|X=x,\theta') + \theta E(t(x,I)^T|X=x,\theta'). \qquad (4.17)$$

Escrevamos $logL(\theta)$ como a diferença entre as funções $Q(\theta|\theta')$ e $H(\theta|\theta')$. Observemos primeiramente que

$$Q_j(\theta|\theta') = E(\log a(x_j,I|\theta)|X=x_j,\theta')$$

$$= -\log a(\theta) + E(\log b(x_j,I)|X=x_j,\theta') + \theta E(t(x_j,I)^T|X=x_j,\theta'), \qquad (4.18)$$

e que

$$H_j(\theta|\theta') = E(\log k(I|x_j,\theta)|X=x_j,\theta')$$

$$= - \log a(\theta|x_j) + E(\log b(x_j, I)|X = x_j, \theta') + \theta E(t(x_j, I)^T|X = x_j, \theta'), \qquad (4.19)$$

para $j = 1,2, \ldots, m$. Somando em j as diferenças das funções $Q(\theta|\theta')$ e $H(\theta|\theta')$, temos

$$\sum_{j=1}^{n}\left(Q_j(\theta|\theta') - H_j(\theta|\theta')\right) = \sum_{j=1}^{n}\left(-\log a(\theta) + E(\log b(x_j, I)|X = x_j, \theta') + \theta E(t(x_j, I)^T|X = x_j, \theta') \right)$$

$$- \sum_{j=1}^{n}\left(- \log a(\theta|x_j) + E(\log b(x_j, I)|X = x_j, \theta') + \theta E(t(x_j, I)^T|X = x_j, \theta') \right).$$

Esta soma pode ser reduzida a

$$\sum_{j=1}^{n}\left(Q_j(\theta|\theta') - H_j(\theta|\theta')\right) = \sum_{j=1}^{n}\left(-\log a(\theta) + \log a(\theta|x_j)\right) = - n \log a(\theta) + \sum_{j=1}^{n} \log a(\theta|x_j).$$

Em (4.8), esta última expressão define o $\log L(\theta)$, com isto temos então que

$$\log L(\theta) = \sum_{j=1}^{n}\left(Q_j(\theta|\theta') - H_j(\theta|\theta')\right). \qquad (4.20)$$

Vejamos como as funções $Q(\theta|\theta')$ e $H(\theta|\theta')$ servem para obter a relação (4.13). Derivando $H(\theta|\theta')$, temos

$$\frac{\partial}{\partial\theta} H(\theta|\theta') = - \frac{\partial}{\partial\theta} \log a(\theta|x) + E(t(x, I)^T|X = x, \theta'),$$

e, igualando a zero,

$$E(t(x, I)^T|X = x, \theta') = E(t(x, I)^T|X = x, \theta).$$

Pelas propriedades das funções de densidade pertencentes à família exponencial, temos também que

$$\frac{\partial^2}{\partial\theta^2} \log a(\theta|x) = Var(t(x, I)^T | X = x, \theta).$$

Então

$$\frac{\partial^2}{\partial\theta^2}H(\theta|\theta') = -\frac{\partial^2}{\partial\theta^2}\log a(\theta|x) + \frac{\partial}{\partial\theta}E(t(x,I)^T \mid X=x,\theta)$$

$$= -\frac{\partial^2}{\partial\theta^2}\log a(\theta|x) = -Var(t(x,I)^T \mid X=x,\theta) < 0.$$

Isto é, o valor máximo de $H(\theta|\theta')$ será atingido, pela identificabilidade da função $f(x|\theta)$, quando $\theta = \theta'$. Neste caso

$$\frac{\partial}{\partial\theta}Q(\theta|\theta') = -\frac{\partial}{\partial\theta}\log a(\theta') + E(t(x,I)^T|X=x,\theta') = 0, \qquad (4.21)$$

implicará a igualdade (4.13).

Na demonstração do Lema 4.1 a seguir, utilizaremos a desigualdade *(1e.6.1)*, página 58 em Rao(1973), que afirma: *Se* $\sum_i a_i$ *e* $\sum_i b_i$ *são séries convergentes de números positivos, tais que* $\sum_i a_i \le \sum_i b_i$ *. Então*

$$\sum_i a_i \log\frac{b_i}{a_i} \le 0. \qquad (*)$$

Lema 4.1: Seja $\theta' \in$ R tal que $Q_j(\theta|\theta') \ge Q_j(\theta'|\theta')$. Então $L(\theta) \ge L(\theta')$.

Prova. Provemos primeiramente que $H_j(\theta|\theta') \le H_j(\theta'|\theta')$. As séries

$$\sum_{i\in\{0,1\}^m} k(i|x_j,\theta'), \quad \sum_{i\in\{0,1\}^m} k(i|x_j,\theta)$$

somam 1, logo cumprem as condições para que *(1e.6.1)* seja válida.

Temos então que:

$$H_j(\theta|\theta') - H_j(\theta'|\theta') = E(\log k(I|x_j,\theta)|X=x_j,\theta') - E(\log k(I|x_j,\theta')|X=x_j,\theta')$$

$$= \sum_{i\in\{0,1\}^m} \log k(i|x_j,\theta)k(i|x_j,\theta') - \sum_{i\in\{0,1\}^m} \log k(i|x_j,\theta')k(i|x_j,\theta')$$

$$= \sum_{i\in\{0,1\}^m} \log\frac{k(i|x_j,\theta)}{k(i|x_j,\theta')}k(i|x_j,\theta') \le 0,$$

por (♕), Rao(1973).

Isto implicará que

$$H_j(\theta|\theta') \leq H_j(\theta'|\theta'),$$

e pelas condições do Lema

$$Q_j(\theta|\theta') \geq Q_j(\theta'|\theta').$$

Somando para todo j, temos

$$\sum_{j=1}^{n}\left(Q_j(\theta|\theta') - H_j(\theta|\theta')\right) \geq \sum_{j=1}^{n}\left(Q_j(\theta'|\theta') - H_j(\theta'|\theta')\right)$$

$$\Rightarrow logL(\theta) \geq logL(\theta'). \qquad \blacklozenge$$

Definição 4.1: Algoritmo EM. Suponhamos que $\theta^{(r)} \in R$ é o atual valor de θ depois de r ciclos do algoritmo. O próximo ciclo é descrito pelos seguintes dois passos:

passo E. Avaliar $Q(\theta|\theta^{(r)})$

passo M. Escolher $\theta^{(r+1)} \in arg\ max_{\theta \in R}\ Q(\theta|\theta^{(r)})$.

Pelo Lema 4.1, no algoritmo definido acima obtém-se uma seqüência $\theta^{(1)}, \ldots, \theta^{(r)}, \theta^{(r+1)}, \ldots$ de estimadores do parâmetro θ tal que a seqüência $\left\{L(\theta^{(r)})\right\}$ é não decrescente em r, isto é, $L(\theta^{(r)}) \leq L(\theta^{(r+1)})$ para todo r. Devemos mostrar então que, se a seqüência $\left\{\theta^{(r)}\right\}$ converge, seu limite é o estimador de máxima verossimilhança $\hat{\theta}$.

Resultados sobre a convergência do algoritmo EM foram primeiramente estudados em Dempster *et al*(1977) e, posteriormente, corrigidos por Jeff Wu(1983). Quando aplicado ao caso de misturas de densidades, o algoritmo foi estudado em Peter & Walker(1978a, 1978b) e Redner & Walker(1984). Como o interesse neste trabalho é a aplicação a misturas, os teoremas relativos à convergência serão considerados só para estes modelos.

4.3 Aplicação aos modelos de mistura finita de densidades

Modelos de misturas finitas de densidades são utilizados na análises de *conglomerados, análises discriminante etc.* e as referências básicas são McLachlan & Basford(1988), Titterington *et al*(1985) e Tapia & Thompson(1990). Baseados nestes trabalhos, realizaremos

aqui um estudo destes modelos e do algoritmo EM. Enunciaremos também propriedades assintóticas deste algoritmo.

Definição 4.2: Modelos de mistura finita de densidades. Serão chamadas de mistura finita de densidades as famílias de funções de densidade da forma

$$\alpha(x|\theta)=\sum_{i=1}^{m}\beta_i K_i(x|\phi_i)\ x\in\mathrm{R}, \tag{4.22}$$

onde cada β_i é um número real não negativo. Estes números são chamados de coeficientes de mistura, tais que $\sum_{i=1}^{m}\beta_i=1$, e cada $K_i(x|\phi_i)$ é uma função de densidade parametrizada por $\phi_i\in R^s$. Define-se o parâmetro θ pelo vetor $(\beta_1,\ldots,\beta_m,\phi_1,\ldots,\phi_m)$.

O espaço paramétrico a ser considerado é então o conjunto

$$\Phi=\{\theta=(\beta_1,\ldots,\beta_m,\phi_1,\ldots,\phi_m):\sum_{i=1}^{m}\beta_i=1 \text{ e } \beta_i\geq 0,\ \phi_i\in R^s \text{ para } i=1,\ldots,m\}.$$

Podemos observar que a forma do estimador de densidades via *"sieves"* de convolução, proposto no Capítulo III, se adequa à definição acima, ou seja, é um caso particular da expressão (4.22) com

$$\beta_1=p_1,\ldots,\beta_q=p_q,\phi_1=(y_1,\sigma),\ldots,\phi_q=(y_q,\sigma),$$

e, se trabalhamos com a expressão (3.24) teríamos

$$\beta_1=\pi_1,\ldots,\beta_k=\pi_k,\phi_1=(\lambda_1,m),\ldots,\phi_k=(\lambda_k,m).$$

Em continuação trabalharemos com uma amostra aleatória do modelo de dados incompletos de densidade $\alpha(x|\theta)$. O objetivo então é obter a função $Q(\theta|\theta')$ do **passo E** do algoritmo EM e achar a expressão resultante do **passo M**.

Definamos as variáveis aleatórias $X=(X_1,\cdots,X_n)$ e o vetor real $x=(x_1,\cdots,x_n)$ com a estrutura nos dados definida na Seção 4.2, temos então associada a função de densidade

$$\alpha(x|\theta)=\prod_{j=1}^{n}\alpha(x_j|\theta).$$

A função de densidade conjunta do vetor (X,I) será

$$\alpha(x,i|\theta) = \prod_{j=1}^{n}\left[\beta_1 K_1(x_j|\phi_1)\right]^{i_{1j}}\left[\beta_2 K_2(x_j|\phi_2)\right]^{i_{2j}}\ldots\left[\beta_m K_m(x_j|\phi_m)\right]^{i_{mj}}, \quad (4.23)$$

onde cada $i_{rt} \in \{0,1\}$ e

$$\sum_{r=1}^{m} i_{rt} = 1.$$

Com as funções definidas em (4.22) e (4.23), a densidade condicional $k(i|x,\theta')$ é dada por

$$k(i|x,\theta') = \prod_{j=1}^{n}\frac{\alpha(x_j,i|\theta')}{\alpha(x_j|\theta')}.$$

Substituindo as expressões da distribuição conjunta $\alpha(x,i|\theta)$ e da função de densidade $\alpha(x|\theta)$, obtemos que

$$k(i|x,\theta') = \prod_{h=1}^{m}\frac{\prod_{j=1}^{n}\left[\beta'_h K_h(x_j|\phi'_h)\right]^{i_{hj}}}{\alpha(x_j|\theta')},$$

dado que

$$\log\alpha(x,i|\theta) = \sum_{h=1}^{m}\sum_{j=1}^{n} i_{hj}\log\left(\beta_h K_h(x_j|\phi_h)\right).$$

A função $Q(\theta|\theta')$ é determinada por

$$Q(\theta|\theta') = E\left(\sum_{h=1}^{m}\sum_{j=1}^{n} I_h \log\left(\beta_h K_h(x_j|\phi_h)\right)\middle| X = x,\theta'\right)$$

$$= \sum_{h=1}^{m}\sum_{j=1}^{n}\log\left(\beta_h K_h(x_j|\phi_h)\right)E(I_h|X = x,\theta')$$

$$= \sum_{h=1}^{m}\sum_{j=1}^{n}\log\left(\beta_h K_h(x_j|\phi_h)\right)k(i|x,\theta')$$

$$\phi_h^{(r+1)} \in arg \max_{\phi_h \in R} \sum_{j=1}^{n} log K_h(x_j|\phi_\eta) \frac{\beta_h^{(r)} K_h(x_j|\phi_h^{(r)})}{\alpha(x_j|\theta^{(r)})}, \qquad (4.26)$$

$h=1, \ldots, m$.

Observemos que

$$\frac{\beta_h^{(r)} K_h(x_j|\phi_h^{(r)})}{\alpha(x_j|\theta^{(r)})}$$

é a probabilidade a posteriori de que X_j seja selecionado do h-ésimo espaço amostral Ω_h, dada a atual aproximação $\theta^{(r)}$ do vetor θ.

Para o caso onde os parâmetros $\phi_1,\ldots,\phi_m$ estão relacionados, o segundo problema de maximização seria encontrar

$$\phi^{(r+1)} \in arg \max_{\phi \in R} \sum_{h=1}^{n} \sum_{j=1}^{m} \left[\frac{\beta_h^{(r)} K_h(x_j|\phi_h^{(r)})}{\alpha(x_j|\theta^{(r)})} \right] log K_h(x_j|\phi_h). \qquad (4.27)$$

Para obter a equação de verossimilhança determinada pelas proporções $\phi_1,\ldots,\phi_m$, restritas a serem não negativas e somando 1, definamos $\beta^{(r+1)} = \left(\beta_1^{(r+1)},\ldots,\beta_m^{(r+1)}\right)$ o vetor das próximas estimativas de máxima verossimilhança.

Então

$$g(\beta_1,\ldots,\beta_m) = \sum_{h=1}^{m} \beta_h - 1,$$

é a função das restrições dos coeficientes de mistura que têm a propriedade de $g\left(\beta_1^{(r+1)},\ldots,\beta_m^{(r+1)}\right) = 0$ cujas derivadas com respeito a cada β_h existem.

A função a maximizar neste caso é

$$log L(\beta_1,\ldots,\beta_m) = \sum_{h=1}^{m} \left(\sum_{j=1}^{n} \left[\frac{\beta_h^{(r)} K_h(x_j|\phi_h^{(r)})}{\alpha(x_j|\theta^{(r)})} \right] \right) log \beta_h \quad , \qquad (4.28)$$

com a restrição

$$g\left(\beta_1^{(r+1)},\ldots,\beta_m^{(r+1)}\right)=0 \qquad \Leftrightarrow \qquad \sum_{h=1}^{m}\beta_h - 1 = 0. \tag{4.29}$$

Utilizando Multiplicadores de Lagrange, os estimadores $\beta_1^{(r+1)},\ldots,\beta_m^{(r+1)}$ são obtidos como a solução do sistema de equações,

$$\left(\frac{\partial \log L}{\partial \beta_1},\ldots,\frac{\partial \log L}{\partial \beta_m}\right) + \left(\frac{\partial g}{\partial \beta_1},\ldots,\frac{\partial g}{\partial \beta_m}\right)\lambda = 0 \tag{4.30}$$

$$g\left(\beta_1^{(r+1)},\ldots,\beta_m^{(r+1)}\right)=0.$$

Deste sistema, obtemos as seguintes relações,

$$\frac{\sum_{j=1}^{n}\left[\dfrac{\beta_1^{(r)}K_1(x_j|\phi_1^{(r)})}{\alpha(x_j|\theta^{(r)})}\right]}{\beta_1^{(r+1)}}+\lambda=0,\ \ldots,\ \frac{\sum_{j=1}^{n}\left[\dfrac{\beta_m^{(r)}K_m(x_j|\phi_m^{(r)})}{\alpha(x_j|\theta^{(r)})}\right]}{\beta_m^{(r+1)}}+\lambda=0. \tag{4.31}$$

Temos, portanto, a igualdade

$$\frac{\sum_{j=1}^{n}\left[\dfrac{\beta_1^{(r)}K_1(x_j|\phi_1^{(r)})}{\alpha(x_j|\theta^{(r)})}\right]}{\beta_1^{(r+1)}}=\ \ldots\ =\frac{\sum_{j=1}^{n}\left[\dfrac{\beta_m^{(r)}K_m(x_j|\phi_m^{(r)})}{\alpha(x_j|\theta^{(r)})}\right]}{\beta_m^{(r+1)}},$$

a qual implicará que

$$\begin{aligned}\beta_1^{(r+1)} &= C\sum_{j=1}^{n}\frac{\beta_1^{(r)}K_1(x_j|\phi_1^{(r)})}{\alpha(x_j|\theta^{(r)})}\\ &\vdots\\ \beta_m^{(r+1)} &= C\sum_{j=1}^{n}\frac{\beta_m^{(r)}K_m(x_j|\phi_m^{(r)})}{\alpha(x_j|\theta^{(r)})}\end{aligned},$$

onde C é uma constante de proporcionalidade.

Para determinar esta constante C, utilizaremos a restrição na função $g(\beta_1,\ldots,\beta_m)$ que aparece em (4.29), isto é

$$\sum_{h=1}^{m} C \sum_{j=1}^{n} \frac{\beta_h^{(r)} K_h(x_j|\phi_h^{(r)})}{\alpha(x_j|\theta^{(r)})} = 1, \tag{4.32}$$

e, dado que

$$\sum_{h=1}^{m} \frac{\beta_h^{(r)} K_h(x|\phi_h^{(r)})}{\alpha(x|\theta^{(r)})} = 1,$$

obtemos que $C=1/n$.

Com isto, temos que o valor do estimador dos coeficientes de mistura no próximo passo do algoritmo é dado por

$$\beta_h^{(r+1)} = \frac{1}{n} \sum_{j=1}^{n} \frac{\beta_h^{(r)} K_h(x_j|\phi_h^{(r)})}{\alpha(x_j|\theta^{(r)})} \tag{4.33}$$

$h=1,...,m$.

A teoria desenvolvida até aqui pode ser aplicada quando as densidades componentes da mistura pertencem à família exponencial; esta aplicação permitirá obter uma expressão do estimador do vetor de parâmetros θ em função de estatísticas suficientes, que são de fácil implementação computacional.

As densidades pertencentes à família exponencial, onde o vetor de parâmetros θ aparece linearmente no argumento da função exponencial, são chamadas de densidades com parametrização natural e o vetor θ é chamado de parâmetro natural. Em determinadas condições, estudadas por Barndorff-Nielsen(1978), é permitida uma outra parametrização chamada de parametrização esperada, esta é chamada assim porque as densidades da família exponencial são escritas em termos do parâmetro esperado

$$\gamma = E(t(X)|\phi) = \int_{R^n} t(x) f(x|\theta) dx, \qquad \phi \in \mathrm{R}.$$

Nestas condições,

$$f(x|\phi) = f(x|\gamma(\phi)) = b(x)\, exp(\gamma(\phi)^{\mathrm{T}} t(x)) a^{-1}(\gamma).$$

Assumamos que cada $f_k(x|\phi_k)$ tem *parametrização esperada* da forma

$$f_k(x|\phi_k) = b_k(x)\exp(\gamma_k(\phi_k)^T t_k(x)) / a_k(\phi_k),$$

para a_k, b_k, t_k e γ_k apropriados.

Em Peters & Walker(1978a, 1978b) os autores dedicam-se a obter as expressões que definem o algoritmo EM no caso em que $\phi_1, \ldots, \phi_m$ são não relacionados e as densidades na mistura são gaussianas de médias μ_k e variâncias φ_k. Nesta situação, $\phi_k = (\mu_k, \varphi_k)$ e é mostrado que

$$\mu_k^{(r+1)} = \frac{\sum_{j=1}^{n} x_j \dfrac{\exp\left\{-\dfrac{1}{2}\dfrac{(x_j-\mu_k^{(r)})^2}{\varphi^{(r)}}\right\}}{\sum_{t=1}^{m}\alpha_t^{(r)}\exp\left\{-\dfrac{1}{2}\dfrac{(x_j-\mu_t^{(r)})^2}{\varphi^{(r)}}\right\}}}{\sum_{j=1}^{n} \dfrac{\exp\left\{-\dfrac{1}{2}\dfrac{(x_j-\mu_k^{(r)})^2}{\varphi^{(r)}}\right\}}{\sum_{t=1}^{m}\alpha_t^{(r)}\exp\left\{-\dfrac{1}{2}\dfrac{(x_j-\mu_t^{(r)})^2}{\varphi^{(r)}}\right\}}}, \tag{4.34}$$

e

$$\varphi_k^{(r+1)} = \frac{\sum_{j=1}^{n} (x_j-\mu_k^2)^2 \dfrac{\exp\left\{-\dfrac{1}{2}\dfrac{(x_j-\mu_k^{(r)})^2}{\varphi^{(r)}}\right\}}{\sum_{t=1}^{m}\alpha_t^{(r)}\exp\left\{-\dfrac{1}{2}\dfrac{(x_j-\mu_t^{(r)})^2}{\varphi^{(r)}}\right\}}}{\sum_{j=1}^{n} \dfrac{\exp\left\{-\dfrac{1}{2}\dfrac{(x_j-\mu_k^{(r)})^2}{\varphi^{(r)}}\right\}}{\sum_{t=1}^{m}\alpha_t^{(r)}\exp\left\{-\dfrac{1}{2}\dfrac{(x_j-\mu_t^{(r)})^2}{\varphi^{(r)}}\right\}}}. \tag{4.35}$$

Estes resultados são obtidos maximizando (4.26).

Em geral, se as densidades no modelo de mistura pertencem à família exponencial e $\phi_1, \ldots, \phi_m$ são não relacionados, as expressões que definem o algoritmo seriam

$$\phi_k^{(r+1)} = \frac{\sum_{i=1}^{n} t_k(x_i) \dfrac{\alpha_k^{(r)} f_k(x_i|\phi_k^{(r)})}{f(x_i|\theta^{(r)})}}{\sum_{i=1}^{n} \dfrac{\alpha_k^{(r)} f_k(x_i|\phi_k^{(r)})}{f(x_i|\theta^{(r)})}}. \tag{4.36}$$

Este resultado foi provado em Redner & Walker(1984).

Para conhecer a forma particular do algoritmo EM no caso em que ϕ_1 , ... , ϕ_m estão relacionados, consideremos a função a maximizar em (4.27) como a parametrização natural. Especificamente, seja $\phi_k = (\mu_k, \varphi)$. Derivaremos com respeito a cada ϕ_k, igualaremos a soma a zero e restituiremos a parametrização esperada. Primeiramente, obtemos a expressão de $\mu_k^{(r+1)}$.

Definindo

$$h_k^{(r)}(x_j) = \frac{\alpha_k^{(r)} f_k(x_j|\phi_k^{(r)})}{f(x_j|\theta^{(r)})}, \tag{4.37}$$

temos que o segundo termo em (4.24) pode ser escrito como sendo

$$\sum_{k=1}^{m}\sum_{j=1}^{n} log\, f_k(x_j|\varphi,\mu_k) h_k^{(r)}(x_j) =$$

$$= \sum_{k=1}^{m}\sum_{j=1}^{n} \left\{ log\ b_k(x_j) + \mu_k t_k(x_j,\varphi) - log\ a_k(\mu_k) \right\} h_k^{(r)}(x_j) = \sum_{k=1}^{m} l(\mu_k).$$

Então

$$\frac{\partial}{\partial \mu_s} \sum_{k=1}^{m} l(\mu_k) = \sum_{j=1}^{n} \left\{ \mu_s t_s(x_j,\varphi) - \frac{\partial}{\partial \mu_s} log\ a_s(\mu_s) \right\} h_s^{(r)}(x_j) = 0,$$

$s=1,...,m$.

Resolvendo esta equação, temos

$$\sum_{j=1}^{n} t_s(x_j) h_s^{(r)}(x_j) = \sum_{j=1}^{n} \left\{ \frac{\partial}{\partial \mu_s} log\ a_s(\mu_s) \right\} h_s^{(r)}(x_j).$$

Restituindo agora a parametrização esperada,

$$\frac{\partial}{\partial \mu_s} log\ a_s(\mu_s) = E(t_s(X_j)|\mu_s) = \mu_s$$

e, de (4.34), obtemos

$$\mu_s^{(r+1)} = \frac{\sum_{j=1}^{n} t_s(x_j) h_s(xj)}{\sum_{j=1}^{n} h_s(x_j)} = \frac{\sum_{j=1}^{n} t_s(x_j) \frac{g_s(x_j|\varphi^{(r)}, \mu_s^{(r)})}{g(x_j|\theta^{(r)})}}{\sum_{j=1}^{n} \frac{g_s(x_j|\varphi^{(r)}, \mu_s^{(r)})}{g(x_j|\theta^{(r)})}}. \qquad (4.38)$$

Para obter a expressão de $\varphi^{(r+1)}$, escreveremos o segundo termo de (4.24) da seguinte forma

$$\sum_{k=1}^{m} \sum_{j=1}^{n} log\ g_k(x_j|\phi, \mu_k^{(r)}) \frac{a_k g_k(x_j|\phi^{(r)}, \mu_k^{(r)})}{g(x_j|\theta^{(r)})} =$$

$$= \sum_{k=1}^{m} \sum_{j=1}^{n} \left\{ log\ b_k(x_j) + \varphi t_k(x_j, \mu_k^{(r)}) - log\ a_k(\varphi) \right\} h_k^{(r)}(x_j) = l(\varphi)$$

$$\frac{\partial}{\partial \varphi} l(\varphi) = \sum_{k=1}^{m} \sum_{j=1}^{n} \left\{ t_k(x_j, \mu_k^{(r)}) - \frac{\partial}{\partial \varphi} log\ a_k(\varphi) \right\} h_k^{(r)}(x_j) = 0.$$

Obtém-se então que

$$\sum_{k=1}^{m} \sum_{j=1}^{n} t_k(x_j, \mu_k^{(r)}) h_k^{(r)}(x_j) = \sum_{k=1}^{m} \sum_{j=1}^{n} \left\{ \frac{\partial}{\partial \varphi} log\ a_k(\varphi) \right\} h_k^{(r)}(x_j). \qquad (4.39)$$

Rescrevendo, temos que

$$\frac{\partial}{\partial\varphi}\log a_k(\varphi)=E(t_k(X_j)|\varphi)=\varphi.$$

De (4.39), obtemos

$$\varphi^{(r+1)}=\frac{\sum_{k=1}^{m}\sum_{j=1}^{n}t_k(x_j,\mu_k^{(r)})h_k^{(r)}(x_j)}{\sum_{k=1}^{m}\sum_{j=1}^{n}h_k^{(r)}(x_j)}.$$

Desta expressão, temos

$$\varphi^{(r+1)}=\frac{\sum_{k=1}^{m}\sum_{j=1}^{n}t_k(x_j,\mu_k^{(r)})\frac{g_k(x_j|\varphi^{(r)},\mu_k^{(r)})}{g(x_j|\theta^{(r)})}}{\sum_{k=1}^{m}\sum_{j=1}^{n}\frac{g_k(x_j|\varphi^{(r)},\mu_k^{(r)})}{g(x_j|\theta^{(r)})}}. \qquad (4.40)$$

No caso em que a mistura é de densidades normais, identificando $\varphi = \sigma^2$, temos

$$f_i(x_j|\varphi,\mu_k^{(r)})=\frac{1}{\sqrt{2\pi\varphi}}exp\left\{-\frac{1}{2}\frac{(x_j-\mu_k^{(r)})^2}{\varphi}\right\}. \qquad (4.41)$$

Podemos parametrizá-la em função de $t_k(x_j,\mu_k^{(r)})=(x_j-\mu_k^{(r)})^2$ como

$$f_i(x_j|\varphi,\mu_k^{(r)})=\frac{1}{\sqrt{2\pi\varphi}}exp\left\{-\frac{1}{2}\frac{t_k(x_j,\mu_k^{(r)})}{\varphi}\right\}$$

e, dado que

$$\sum_{k=1}^{m}\sum_{j=1}^{n}\frac{f_k(x_j|\varphi^{(r)},\mu_k^{(r)})}{f(x_j|\theta^{(r)})}=1,$$

a expressão (4.40) será

$$\varphi^{(r+1)}=\frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n}(x_j-\mu_k^{(r)})^2\frac{exp\left\{-\frac{1}{2}\frac{(x_j-\mu_k^{(r)})^2}{\varphi^{(r)}}\right\}}{\sum_{t=1}^{m}\alpha_t^{(r)}\,exp\left\{-\frac{1}{2}\frac{(x_j-\mu_t^{(r)})^2}{\varphi^{(r)}}\right\}}. \tag{4.42}$$

Parametrizando (4.38) em função de $t_k(x_j)=x_j$, temos

$$\mu_k^{(r+1)}=\frac{\sum_{j=1}^{n}x_j\frac{exp\left\{-\frac{1}{2}\frac{(x_j-\mu_k^{(r)})^2}{\varphi^{(r)}}\right\}}{\sum_{t=1}^{m}\alpha_t^{(r)}\,exp\left\{-\frac{1}{2}\frac{(x_j-\mu_t^{(r)})^2}{\varphi^{(r)}}\right\}}}{\sum_{j=1}^{n}\frac{exp\left\{-\frac{1}{2}\frac{(x_j-\mu_k^{(r)})^2}{\varphi^{(r)}}\right\}}{\sum_{t=1}^{m}\alpha_t^{(r)}\,exp\left\{-\frac{1}{2}\frac{(x_j-\mu_t^{(r)})^2}{\varphi^{(r)}}\right\}}}. \tag{4.43}$$

Os coeficientes de mistura em cada passo do algoritmo foram definidos em (4.25) e neste caso assumirão a forma

$$\alpha_k^{(r+1)}=\frac{1}{n}\sum_{j=1}^{n}\frac{\alpha_k^{(r)}\,exp\left\{-\frac{1}{2}\frac{(x_j-\mu_k^{(r)})^2}{\varphi^{(r)}}\right\}}{\sum_{t=1}^{m}\alpha_t^{(r)}\,exp\left\{-\frac{1}{2}\frac{(x_j-\mu_t^{(r)})^2}{\varphi^{(r)}}\right\}}. \tag{4.44}$$

As expressões (4.42), (4.43) e (4.44) definem o algoritmo EM para o caso particular em que as densidades consideradas são normais.

Observemos que, definindo

$$h_k^{(r)}(x) = \exp\left\{-\frac{1}{2}\frac{(x_j-\mu_k^{(r)})^2}{\varphi^{(r)}}\right\} \Big/ \sum_{t=1}^{m}\alpha_t^{(r)}\exp\left\{-\frac{1}{2}\frac{(x_j-\mu_t^{(r)})^2}{\varphi^{(r)}}\right\}$$

podemos escrever (4.42) como

$$\varphi^{(r+1)} = \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n} h_k^{(r)}(x_j)(x_j-\mu_k^{(r)})^2 \tag{4.45}$$

e, (4.35) como sendo

$$\varphi_k^{(r+1)} = \frac{\sum_{j=1}^{n} h_k^{(r)}(x_j)(x_j-\mu_k^{(r)})^2}{\sum_{j=1}^{n} h_k^{(r)}(x_j)}. \tag{4.46}$$

Então, de (4.46),

$$\varphi_k^{(r+1)}\sum_{j=1}^{n} h_k^{(r)}(x_j) = \sum_{j=1}^{n} h_k^{(r)}(x_j)(x_j-\mu_k^{(r)})^2 .$$

Somando em k e dividindo por n, temos

$$\frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n}\varphi_k^{(r+1)}h_k^{(r)}(x_j) = \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n} h_k^{(r)}(x_j)(x_j-\mu_k^{(r)})^2 . \tag{4.47}$$

O segundo termo em (4.47) é exatamente $\varphi^{(r+1)}$ dado em (4.45); logo $\varphi^{(r+1)}$ é uma transformação linear das $\varphi_k^{(r+1)}$, k=1,2, ... , m.

Utilizando as provas da consistência da seqüência $\{\varphi_k^{(r+1)}\}$, que estudaremos na seção a seguir, e resultados da invariância dos estimadores de máxima verossimilhança em Zehna(1966) obteremos uma prova de que a seqüência definida em (4.45) converge ao estimador de máxima verossimilhança de σ^2, isto é, dado que

$$\varphi_k^{(r+1)} \xrightarrow[r\to\infty]{} \hat{\varphi}_k \text{, e } \hat{\varphi} = \sum_{k=1}^{m} \beta_k \hat{\varphi}_k \text{,}$$

a seqüência em (4.45) será consistente para σ^2.

4.4 Propriedades Assintóticas

Estudaremos agora propriedades do algoritmo EM, particularmente no caso de mistura finita de densidades. No trabalho de Peters & Walker(1978b) estuda-se este algoritmo no caso de misturas de densidades normais não relacionadas e, em Redner & Walker(1984), generaliza-se quando a mistura é de densidades não relacionadas pertencentes à família exponencial. Outros exemplos do uso deste algoritmo podem ser achados em Tanner(1991) e Titterington *et al*(1981).

O teorema a seguir mostra que, se a seqüência $\{\theta^{(r)}\}_{r=0,1,2,\ldots}$ gerada pelo algoritmo EM convergir, o limite será o valor $\hat{\theta}$ que maximiza a função de verossimilhança, ou seja, o estimador de máxima verossimilhança.

Lembremos que o espaço paramétrico é

$$\Phi = \left\{\theta = (\beta_1,\ldots,\beta_m,\phi_1,\ldots,\phi_m): \sum_{i=1}^{m}\beta_i = 1 \text{ e } \beta_i \geq 0, \phi_i \in R^s \text{ para } i = 1,\ldots,m\right\}.$$

Teorema 4.2: Suponha que para algum $\theta^{(0)} \in \Phi$, $\{\theta^{(r)}\}_{r=0,1,2,\ldots}$ seja a seqüência em Φ gerada pelo algoritmo EM, isto é

$$\theta^{(r+1)} \in arg\ max_{\theta\in\Phi}\, Q(\theta|\theta^{(r)}), \qquad (4.48)$$

$r=1, 2, \ldots\ldots$ onde $Q(\theta|\theta')$ é a função determinada no passo **E** do algoritmo EM.

Então a função de log-verossimilhança *logL* cresce monotonamente em $\{\theta^{(r)}\}_{r=0,1,2,\ldots}$ para o limite L^* (possivelmente infinito). Ademais, denotando o conjunto de pontos limites da seqüência $\{\theta^{(r)}\}_{r=0,1,2,\ldots}$ em Φ por L' temos

i) Se *logL* é contínua em Φ e $L' \neq \varnothing$, então L^* é finito e $L(\hat{\theta}) = L^* \ \forall\ \hat{\theta} \in L'$.

ii) Se $Q(\theta|\theta')$ e $H(\theta|\theta')$ são contínuas em θ e θ', então $\hat{\theta} \in arg\ max_{\hat{\theta} \in L} Q(\theta|\hat{\theta})$ para cada $\hat{\theta} \in L'$.

iii) Se $Q(\theta|\theta')$ e $H(\theta|\theta')$ são contínuas em θ e θ' e deriváveis, então $logL$ é derivável em $\hat{\theta}$ e a equação de verossimilhança é satisfeita por $\hat{\theta}$, ou seja, $^{d}/_{d\theta}\ L(\hat{\theta}) = 0$.

Prova.

A monotonicidade da função $logL$ foi estabelecida no Lema 4.1, logo L tem limite L^* possivelmente infinito.

i) Dado que o conjunto L' é fechado, então pelo fato de a função $logL$ ser monótona e contínua, temos que L^* é finito e $L(\hat{\theta}) = L^*$ para algum $\hat{\theta} \in L'$.

ii) Suponhamos e que possamos achar um $\hat{\theta} \in L'$ e $\theta \in$R para os quais $Q(\theta|\hat{\theta}) > Q(\hat{\theta}|\hat{\theta})$.

Pelo Lema 1, temos que

$$logL(\theta) > logL(\hat{\theta}) = L^*.$$

Tomando o limite $\lim_{r\to\infty} \theta^{(r)} = \hat{\theta}$ obtemos $L(\theta^{(r)}) \to L^*$, logo $L^* > L^*$, o que constitui uma contradição. Isto mostra que o ponto $\theta \in \Phi$ que faz

$$Q(\theta|\hat{\theta}) > Q(\hat{\theta}|\hat{\theta}),$$

não pode existir.

iii) Foi provado no *item* 4.2 Generalidades, para explicar o algoritmo. ♦

Como o algoritmo só nos garante que a função $logL$ cresce monotonamente em $\{\theta^{(r)}\}$ vejamos agora como as condições sob as quais se obteve esta seqüência nos leva a garantir a convergência em probabilidade. Este resultado é apresentado no Teorema 4.3 a seguir e, o Teorema 4.1 garante que o ponto de convergência será o estimador de máxima verossimilhança. A idéia do Teorema 4.3 é mostrar que a função definida nas iterações (4.25) e (4.36) é contracta (veja Apêndice III).

Seja G: $\Phi \to \Phi$, onde Φ é o espaço paramétrico, a função definida pelas expressões (4.25) e (4.36), isto é,

$$G(\theta) = G(\beta_1,\ldots,\beta_m,\phi_1,\ldots,\phi_m) = \begin{pmatrix} \frac{1}{n}\sum_{j=1}^{n}\frac{\beta_1 K_1(x_j|\phi_1)}{\alpha(x_j|\theta)} \\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n}\frac{\beta_m K_m(x_j|\phi_m)}{\alpha(x_j|\theta)} \\ \sum_{j=1}^{n} t_1(x_j)\frac{K_1(x_j|\phi_1)}{\alpha(x_j|\theta)} \Big/ \sum_{j=1}^{n}\frac{K_k(x_j|\phi_1)}{\alpha(x_j|\theta)} \\ \vdots \\ \sum_{j=1}^{n} t_m(x_j)\frac{K_m(x_j|\phi_m)}{\alpha(x_j|\theta)} \Big/ \sum_{j=1}^{n}\frac{K_m(x_j|\phi_m)}{\alpha(x_j|\theta)} \end{pmatrix}^T .$$

Podemos então escrever a seqüência $\{\theta^{(r)}\}_{r=0,1,2,\ldots}$ da seguinte forma

$$\theta^{(0)}, \quad \theta^{(1)} = G(\theta^{(0)}), \quad \theta^{(2)} = G(\theta^{(1)}) = G(G(\theta^{(0)})), \quad \theta^{(3)} = G(\theta^{(2)}) = G(G(G(\theta^{(0)}))), \ldots .$$

Observemos que a seqüência $\{\theta^{(r)}\}_{r=0,1,2,\ldots}$ gerada por sucessivas aplicações da função G, quando os parâmetros $\phi_1, \ldots, \phi_m$ são não relacionados, convergirá ao ponto fixo $\hat{\theta}$ (veja Apêndice III), se a função G, com a qual é obtido $\theta^{(r+1)}$ a partir de $\theta^{(r)}$, for contracta.

Teorema 4.3: Para $\theta^{(0)} \in \Phi$, seja $\{\theta^{(r)}\}_{r=0,1,2,\ldots}$ a seqüência em Φ gerada pelo algoritmo EM, isto é, a seqüência em Φ gerada pelas iterações definidas em (4.25) e (4.36). Então com probabilidade 1, qualquer que seja n suficientemente grande, a única solução consistente fortemente $\hat{\theta} = (\hat{\beta}_1,\ldots,\hat{\beta}_m,\hat{\phi}_1,\ldots,\hat{\phi}_m)$ da equação de verossimilhança

$$\frac{\partial \log L(\theta)}{\partial \theta} = \frac{\partial \sum_{i=1}^{n} \log f(x_i|\theta)}{\partial \theta} = 0$$

está bem definida e a seqüência $\{\theta^{(r)}\}_{r=0,1,2,\ldots}$ converge linearmente a $\hat{\theta}$, qualquer que seja $\theta^{(0)}$ suficientemente perto de $\hat{\theta}$, isto é, existe uma constante λ, $0 \le \lambda < 1$, para a qual

$$|\theta^{(r+1)} - \hat{\theta}| \le \lambda |\theta^{(r)} - \hat{\theta}|, \qquad r=0, 1, 2, \ldots . \tag{4.49}$$

Prova.

Em Peters & Walker(1978b) prova-se este teorema para o caso em que as misturas são de densidades gaussianas e, em Redner & Walker(1984) prova-se para o caso geral em que estas

pertençam à família exponencial.

Aqui explicaremos os passos desta demonstração e nos referiremos à prova em Redner & Walker(1984).

Se θ^* for o verdadeiro valor do parâmetro, temos, pelo Teorema 4.1, que $\hat{\theta}$ está bem definido e converge com probabilidade 1 a θ^*, quando n é suficientemente grande. É de observar que se a condição (4.49) se cumprir, então a seqüência $\{\theta^{(r)}\}_{r=0,1,2,\ldots}$ converge a $\hat{\theta}$ pelo Teorema do *Ponto Fixo*, veja Anexo III. O funcional G, definido anteriormente, como sendo aquele que $G(\theta^{(r)}) = \theta^{(r+1)}$ tem a $\hat{\theta}$ como *ponto fixo*, ou seja, $\hat{\theta}$ que satisfaz $G(\hat{\theta})=\hat{\theta}$.

Desta forma, pode-se escrever (4.49) como

$$| G(\theta^{(r)}) - G(\hat{\theta}) | \leq \lambda | \theta^{(r)} - \hat{\theta} |, \qquad r=0, 1, 2, \ldots . \tag{4.50}$$

ou seja, o funcional G cumpriria com a definição de *funcional contracto* se existir λ, $0 \leq \lambda < 1$, satisfazendo (4.50). Por conseqüência, o funcional G terá um único *ponto fixo* e já temos observado que a seqüência gerada pelo algoritmo EM pode ser obtida por aplicações consecutivas do funcional G.

Dado que G é derivável, podemos escrever

$$G(\theta^{(r)}) - G(\hat{\theta}) = G(\theta^{(r)}) |\theta^{(r)} - \hat{\theta}| + O(|\theta^{(r)}-\hat{\theta}|^2). \tag{4.51}$$

Devido a $O(|\theta^{(r)} -\hat{\theta}|^2)$ ser estritamente positivo, de (4.51) temos que

$$|G(\theta^{(r)}) - G(\hat{\theta})| \leq |G'(\theta^{(r)})| \, |\theta^{(r)} -\hat{\theta}|. \tag{4.52}$$

Bastaria provar que $|G'(\theta^{(r)})| \leq 1$, com isto teríamos provado que o funcional G é *contracto* e, dado que estamos suficientemente perto de $\hat{\theta}$, este seria o *ponto fixo*. Os detalhes da demonstração de que $|G'(\theta^{(r)})| \leq 1$ assintoticamente acham-se em Redner & Walker(1984).

4.5 Problemas computacionais do *"sieves"* de convolução gaussiano

A utilização prática dos estimadores de Grenander de convolução é difícil devido ao fato de terem $2n+1$ parâmetros a estimar. Uma sugestão é trabalhar com um *"sieves"* reduzido nos pesos $p_1,\ldots,p_n$, considerando-os como iguais a $1/n$.

Desta forma, os estimadores

$$\hat{f}(x)=\sum_{i=1}^{n}\frac{p_i m_n^2}{\sqrt{2\pi}}e^{\frac{-m_n^2(x-y_i)^2}{2}}, \qquad (4.53)$$

se reduzem a

$$\hat{f}(x)=\sum_{i=1}^{n}\frac{1}{n}\frac{m_n^2}{\sqrt{2\pi}}e^{\frac{-m_n^2(x-y_i)^2}{2}}. \qquad (4.54)$$

É importante notar que o estimador de Rosenblatt-Parzen com núcleo gaussiano, isto é, o estimador

$$\tilde{f}(x)=\frac{1}{\sqrt{2\pi}}\frac{1}{nh}\sum_{i=1}^{n}e^{\frac{-(x-X_i)^2}{2h^2}},$$

é da forma (4.53). No entanto, as afirmações no final da Proposição 3.5 indicam que $\tilde{f}$ não pertence ao conjunto $M_{m_n}^n$.

Geman & Hwang(1982) trabalharam com os estimadores da forma (4.53), observando que o número de y_i's distintos é consideravelmente menor de que n. Dados esses resultados, estes autores sugerem definir o parâmetro do *"sieves"* m_n como o número de funções núcleos a serem somadas em (4.54) e, os pesos p_i iguais a $1/m_n$. Com isto, teremos definida a seqüência *"sieves"*

$$S'_{m_n}=\left\{f\in\Theta: f(x)=\frac{1}{m_n}\sum_{i=1}^{m_n}\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-y_i)^2}{2\sigma^2}}\right\}. \qquad (4.55)$$

Geman & Hwang(1982) observaram que o estimador de máxima verossimilhança associado a S'_{m_n} teve uma boa performance em simulações. Eles também provam que $m_n=O(n^{1/5-\varepsilon})$ serve como razão de convergência.

Esta é a idéia do método de estimação dos parâmetros de elementos de seqüências *"sieves"*, proposto neste Capítulo, este algoritmo é mais geral que aquele que seria considerado em (4.55).

A diferença é nos pesos, estes serão considerados diferentes, restritos a serem não negativos e somarem 1. Dadas as boas possibilidades do método de estimação proposto, observou-se que não se precisava afastar-se tanto da forma original do *"sieves"* de convolução gaussiano, definido na Proposição 3.5 como

$$\hat{f}(x)=\sum_{i=1}^{n}\frac{p_i m_n^2}{\sqrt{2\pi}}e^{\frac{-m_n^2(x-y_i)^2}{2}}.$$

Definimos a seguir o conjunto S''_{m_n}. Neste conjunto será obtido o estimador alvo de simulações no Capítulo V,

$$S''_{m_n}=\left\{f\in\Theta: f(x)=\sum_{j=1}^{m_n}\frac{p_j}{\sqrt{2\pi\sigma}}e^{\frac{-(x-y_j)^2}{2\sigma^2}}\right\}. \tag{3.34}$$

Observemos que os $f\in S''_{m_n}$ têm $2m_n+1$ parâmetros a estimar, isto é, a quantidade de parâmetros a estimar depende do número m_n de funções núcleo.

Uma idéia para utilizar o valor de m_n é usar diretamente a relação $m_n=O(n^{1/5-\varepsilon})$. Daqui obtemos que se $n=25, m_n\le 2$, se $n=50, m_n\le 2$ e se $n=100, m_n\le 3$.

Uma forma então de escolher qual o valor de m_n, é avaliar a função de verossimilhança

$$L(m_n)=\prod_{i=1}^{n}f(X_i),$$

para os diferentes valores de m_n, e aquele no qual a função L atinja o máximo será assumido como o número de funções núcleo no estimador $\hat{f}$. Passando posteriormente a estimar os parâmetros $p_1,\ldots,p_{m_n},y_1,\ldots,y_{m_n},\sigma$. Estes parâmetros serão estimados utilizando o algoritmo EM já descrito.

Capítulo V:

Comparações

5.1 Introdução

Para ter idéia do ajuste obtido com cada proposta de estimador da função de densidade, estudados nos Capítulos II e III, escolhemos diferentes funções de densidade e geramos amostras aleatórias destas distribuições. Os estimadores foram comparados de acordo com o Erro Quadrático Médio Integral(*EQMI*).

Dados foram simulados para 10 diferentes funções de densidade. Duas destas possuem geradores de variáveis aleatórias no *SAS* 6.08, nos outros casos teve-se que achar a amostra necessária utilizando a Transformada Integral, o Método de Composição ou o Método de Rejeição.

Os resultados destas comparações serão apresentados em tabelas contendo informações sobre os estimadores e os valores do *EQMI* obtidos quando é utilizado o estimador de Rosenblatt-Parzen e o de Grenander. Gráficos incluídos apresentam a amostra obtida, e a forma da densidade em estudo e a dos estimadores.

Para a realização deste estudo de comparação foram construídos dois tipos de programas, o primeiro tipo de programa foi necessário para gerar as amostras (Apêndice IV), o outro tipo de programa dedica-se à construção dos estimadores. No Apêndice V se mostra o programa para a construção dos estimadores de Rosenblatt-Parzen e Grenander quando a densidade a estimar é normal padrão, este programa faz os cálculos baseado nas expressões (4.42), (4.43) e (4.44).

5.2 Densidades e conceitos gerais

No artigo de Bowman(1985), ele estudou o ajuste aos dados obtido por diferentes funções núcleo no estimador Rosenblatt-Parzen para 6 diferentes funções de densidade. Estas foram incluídas em nossas análises assim como também as densidades propostas por Crain(1974) no seu estudo da estimação de densidades usando expansões ortogonais.

Na tabela a seguir apresentamos as definições das densidades escolhidas para o estudo.

Nome	Definição
Normal Padrão	$f_1(x)=\frac{1}{\sqrt{2\pi}}exp\left(-\frac{1}{2}x^2\right),\ x\in R$
1ªMistura de Normais	$f_2(x)=\frac{1}{2\sqrt{2\pi}}e^{-\frac{1}{2}(x+1.5)^2}+\frac{1}{2\sqrt{2\pi}}e^{-\frac{1}{2}(x-1.5)^2},\ x\in R$
t-Student, 5 g.l.	$f_3(x)=\frac{2}{\sqrt{5\pi}\,\Gamma(2.5)}\left(1+\frac{x^2}{5}\right)^{-3},\ x\in R$
Cauchy Padrão	$f_4(x)=\frac{1}{\pi(1+x^2)},\ x\in R$
Qui-Quadrado, 6 g.l.	$f_5(x)=\frac{1}{16}x^2\,exp\left(-\frac{x}{2}\right),\ x\in[0,\infty)$
Beta, α=2, β=2	$f_6(x)=6x(1-x),\ x\in[0,1]$
Triangular	$f_7(x)=\frac{3}{4}-\frac{\lvert x\rvert}{2},\ x\in[-1,1]$
Trimodal	$f_8(x)=C(cos(10x)+2),\ x\in[-1,1]$ onde $C=\frac{5}{20+sen(10)}$
Uniforme Escada	$f_9(x)=\begin{cases}2/3 & se\ x\in[-1,0]\\ 1/3 & se\ x\in(0,1]\end{cases}$
2ªMistura de Normais	$f_{10}(x)=\frac{1}{2\sqrt{2\pi}0.5}e^{-\frac{1}{2}\left(\frac{x+2}{0.5}\right)^2}+\frac{1}{2\sqrt{2\pi}1.6}e^{-\frac{1}{2}\left(\frac{x-3}{1.6}\right)^2}$

Tabela 5.1

Seja

$$\hat{f}(x)=\begin{cases}\hat{f}(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\sum_{j=1}^{m}p_j\,exp\left(-\frac{1}{2\sigma^2}(x-\mu_j)^2\right)\\ \tilde{f}(x)=\frac{1}{\sqrt{2\pi}}\frac{1}{nh}\sum_{i=1}^{n}exp\left(-\frac{1}{2h^2}(x-X_i)^2\right)\end{cases}\quad ou$$

desta forma temos que $\hat{f}(x)$ pode denotar o estimador Grenander $\hat{f}(x)$ ou o estimador de Rosenblatt-Parzen $\tilde{f}(x)$.

O *EQMI* é dado por

$$EQMI = \int_{-\infty}^{\infty} \left(f(x) - \hat{f}(x)\right)^2 dx = \int_{-\infty}^{\infty} f^2(x)dx - 2\int_{-\infty}^{\infty} f(x)\hat{f}(x)dx + \int_{-\infty}^{\infty} \hat{f}^2(x)dx . \quad (5.1)$$

Em alguns casos utilizaremos métodos numéricos de integração já incorporados no pacote computacional, nas outras situações existem expressões para as integrais.

Inclua-se também um outro estudo comparativo, este baseia-se na expressão

$$\left(\hat{f}(x) - f(x)\right)^2 . \quad (5.2)$$

Apresentam-se, nas diferentes situações, gráficos que mostram o comportamento desta diferença pontualmente.

5.3 Resultados das comparações

Os gráficos das diferentes funções de densidade, assim como do estimador de Rosenblatt-Parzen e de Grenander, apresentam-se a seguir. Inclua-se também a amostra gerada e tabelas com os valores dos estimadores dos parâmetros que definem cada método.

Ainda apresentam-se gráficos do comportamento da diferença em (5.2) para cada tamanho de amostra em cada uma das diferentes funções de densidade relacionadas na Tabela 5.1.

5.3.1 Resultado das comparações na densidade Normal Padrão

As amostras para esta função de densidade de probabilidade foram geradas diretamente pela função SAS, NORMAL ou RANNOR.

Rosenblatt-Parzen	Grenander
$EQMI$ =0.5600876	$EQMI$ =0.046867
$\hat{h}$ = 0.43072	$\hat{\sigma}^2$ = 0.61228
	$\hat{p}_1$=1 $\hat{\mu}_1$ = 0.41471

Tabela 5.2: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=25.

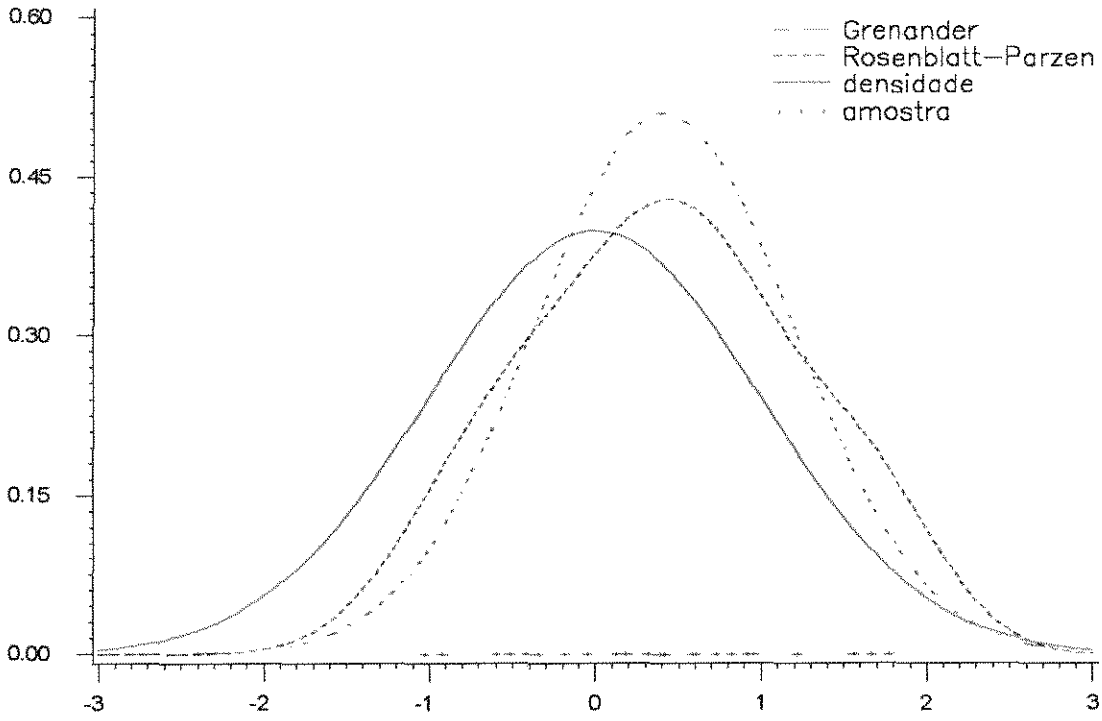


Gráfico 5.1 Função de densidade Normal Padrão e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.2, tamanho de amostra n=25.

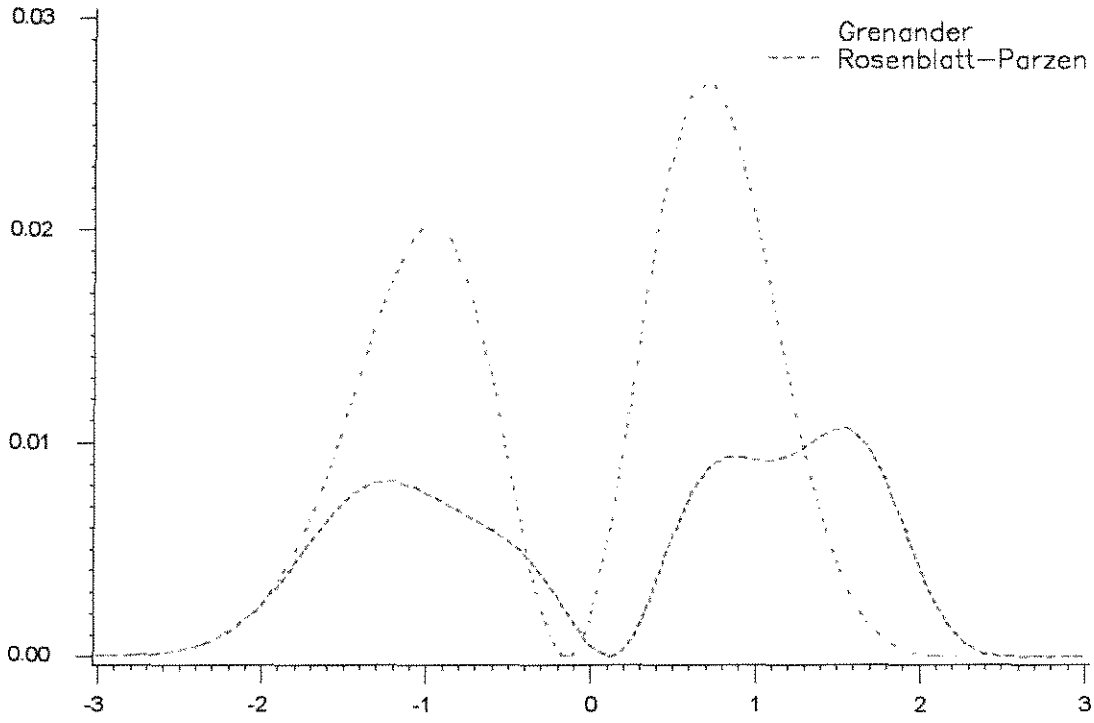


Gráfico 5.2 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Normal Padrão, n=25.

Densidade Normal Padrão, tamanho de amostra n=50.

Rosenblatt-Parzen	Grenander	
$EQMI$ =0.5329381	$EQMI$ =0.0051933	
$\hat{h}$ = 0.45936	$\hat{\sigma}^2$ = 0.63546	
	$\hat{p}_1$ = 0.13767	$\hat{\mu}_1$ = 0.41471
	$\hat{p}_2$ = 0.77695	$\hat{\mu}_2$ = -0.13329
	$\hat{p}_3$ = 0.08539	$\hat{\mu}_3$ = 2.06172

Tabela 5.3: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=50.

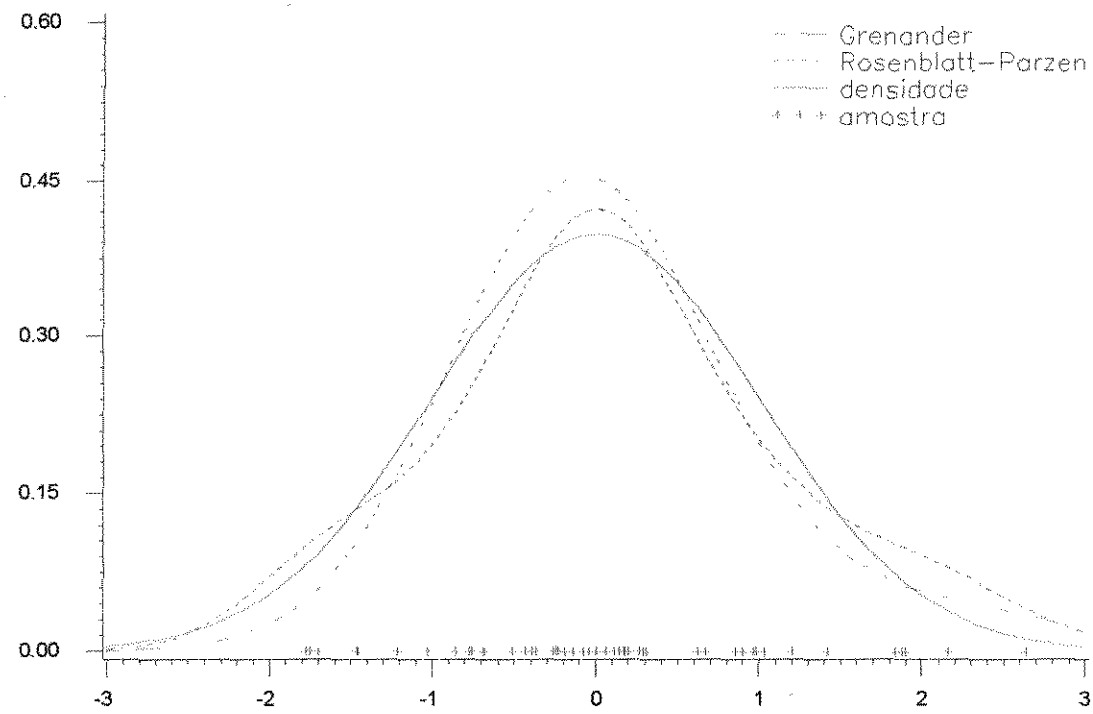


Gráfico 5.3 Função de densidade Normal Padrão e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.3, tamanho de amostra n=50.

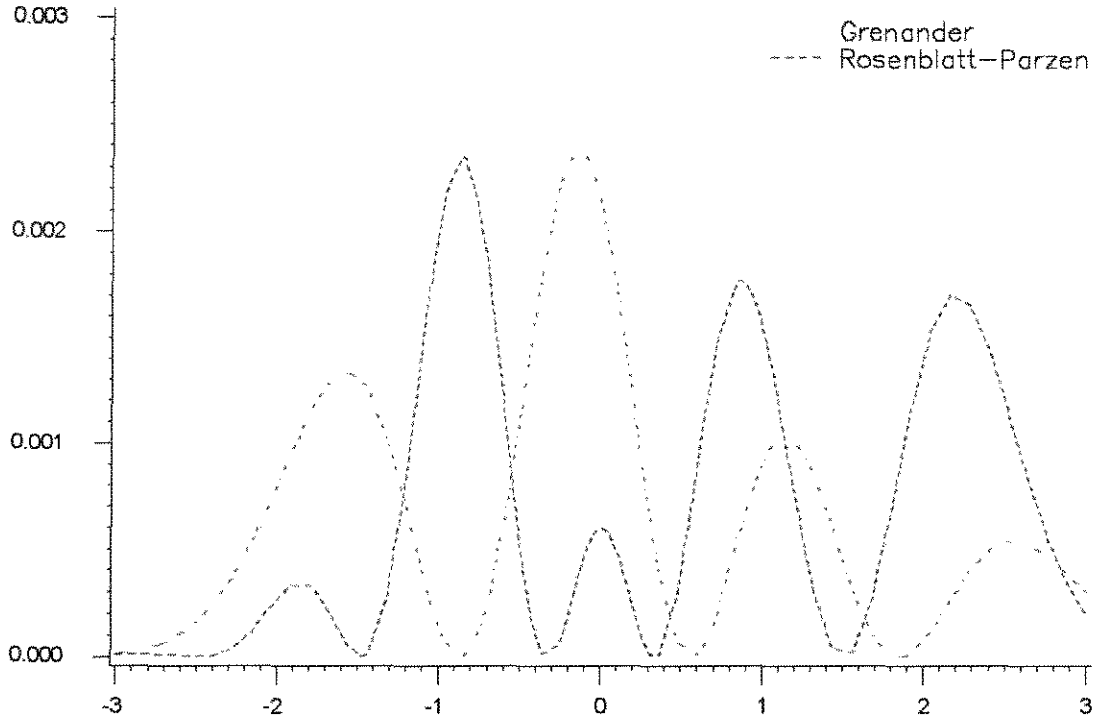


Gráfico 5.4 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Normal Padrão, n=50.

Densidade Normal Padrão, tamanho de amostra n=100.

Rosenblatt-Parzen	Grenander
$EQMI$ =0.5327152	$EQMI$ =0.0004399
$\hat{h}$ = 0.45936	$\hat{\sigma}^2$ = 0.98041
	$\hat{p}_1$ = 1 $\quad$ $\hat{\mu}_1$ = -0.054116

Tabela 5.4: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=100.

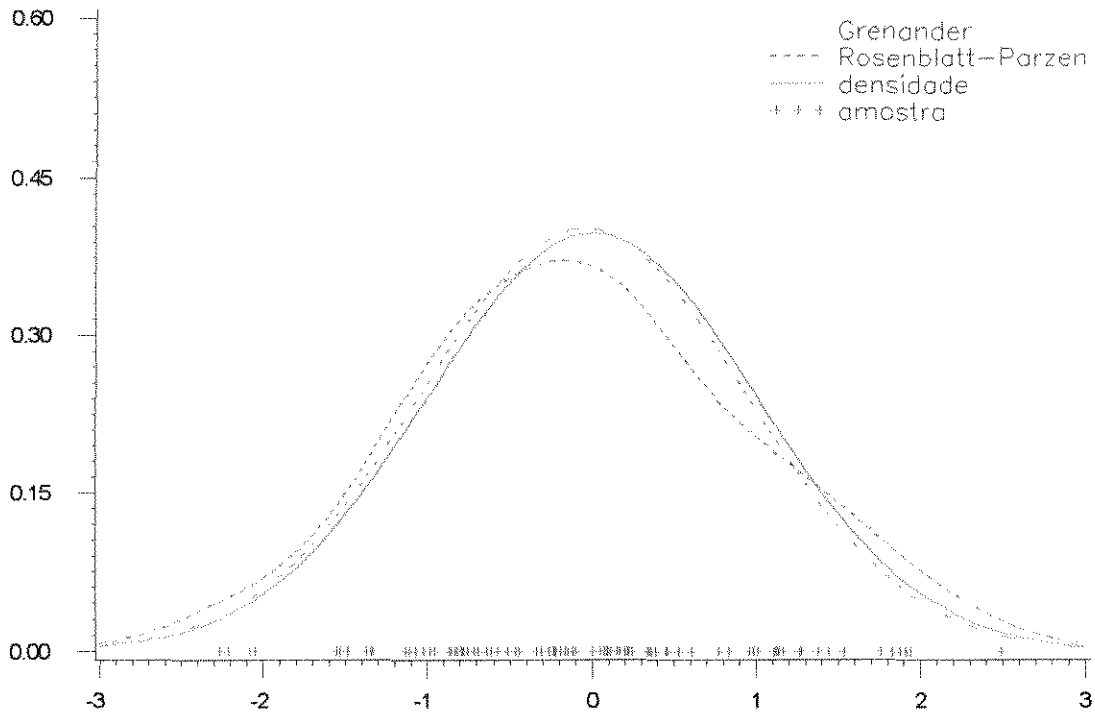


Gráfico 5.5 Função de densidade Normal Padrão e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.4, tamanho de amostra n=100.

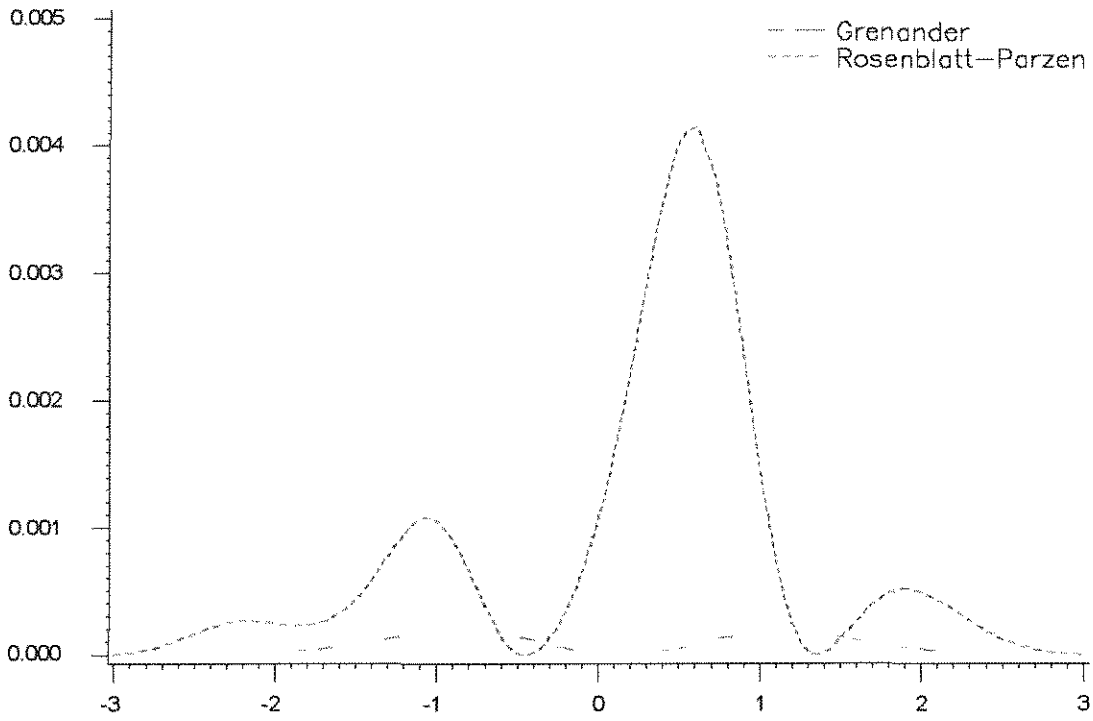


Gráfico 5.6 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Normal Padrão, n=100

5.3.2 Resultado das comparações na densidade 1ª Mistura de Normais

Nesta situação utilizou-se o chamado Método de Composição que consiste em escolher com uma determinada probabilidade uma das densidades na mistura e logo simular um valor da variável correspondente.

Rosenblatt-Parzen	Grenander	
$EQMI$ =0.0378903	$EQMI$ =0.0147441	
$\hat{h}$ = 0.66029	$\hat{\sigma}^2$ = 0.28849	
	$\hat{p}_1$=0.44518	$\hat{\mu}_1$ = -1.38112
	$\hat{p}_2$=0.25629	$\hat{\mu}_2$ = 1.08807
	$\hat{p}_3$=0.29853	$\hat{\mu}_3$ = 2.71556

Tabela 5.5: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros, tamanho de amostra n=25.

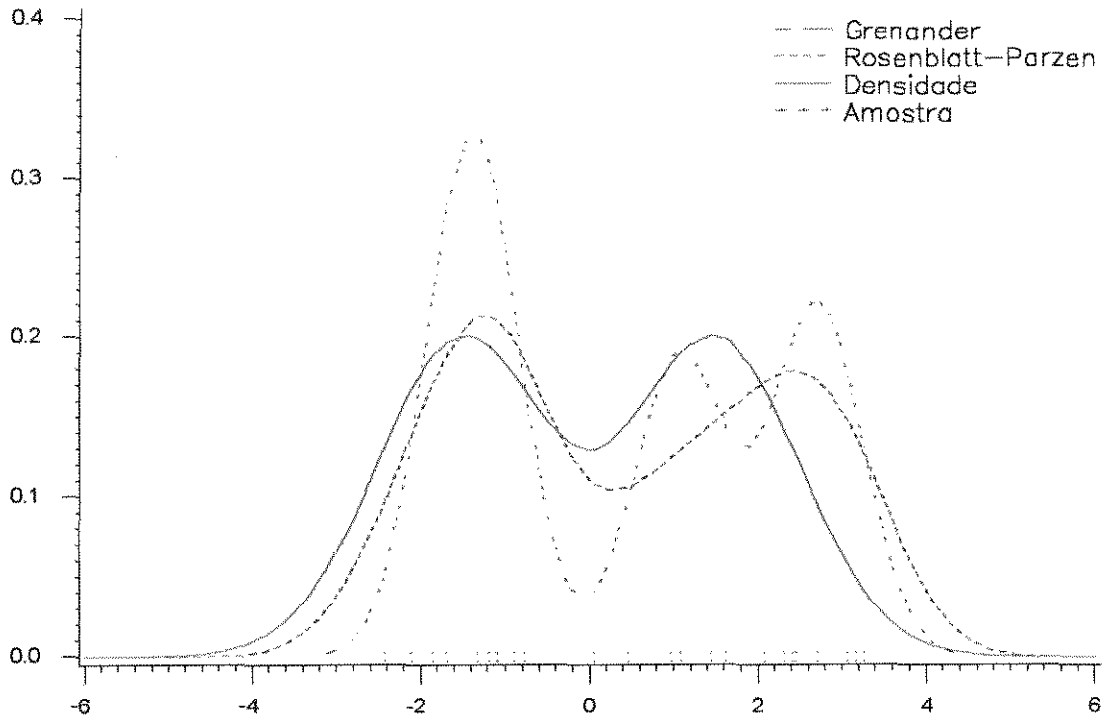


Gráfico 5.7 Função de densidade 1ª Mistura de Normais e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.5, tamanho de amostra n=25.

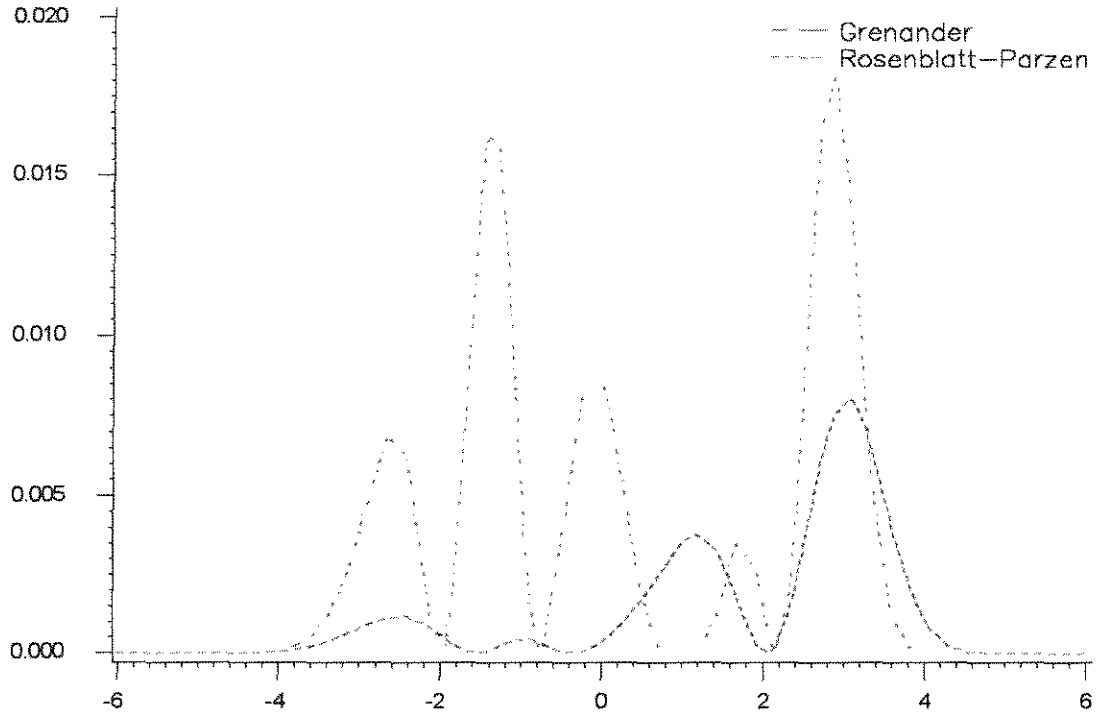


Gráfico 5.8 Comportamento pontual da diferença $\left(\hat{f}(x)-f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade 1ª Mistura de Normais, n=25.

Densidade 1ªMistura de Normais, n=50

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0072809	$EQMI$ =0.0076781
$\hat{h}$ = 0.85225	$\hat{\sigma}^2$ = 1.03704
	$\hat{p}_1$=0.58807 $\hat{\mu}_1$ = -1.34751
	$\hat{p}_2$=0.41193 $\hat{\mu}_2$ = 1.99273

Tabela 5.6: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=50.

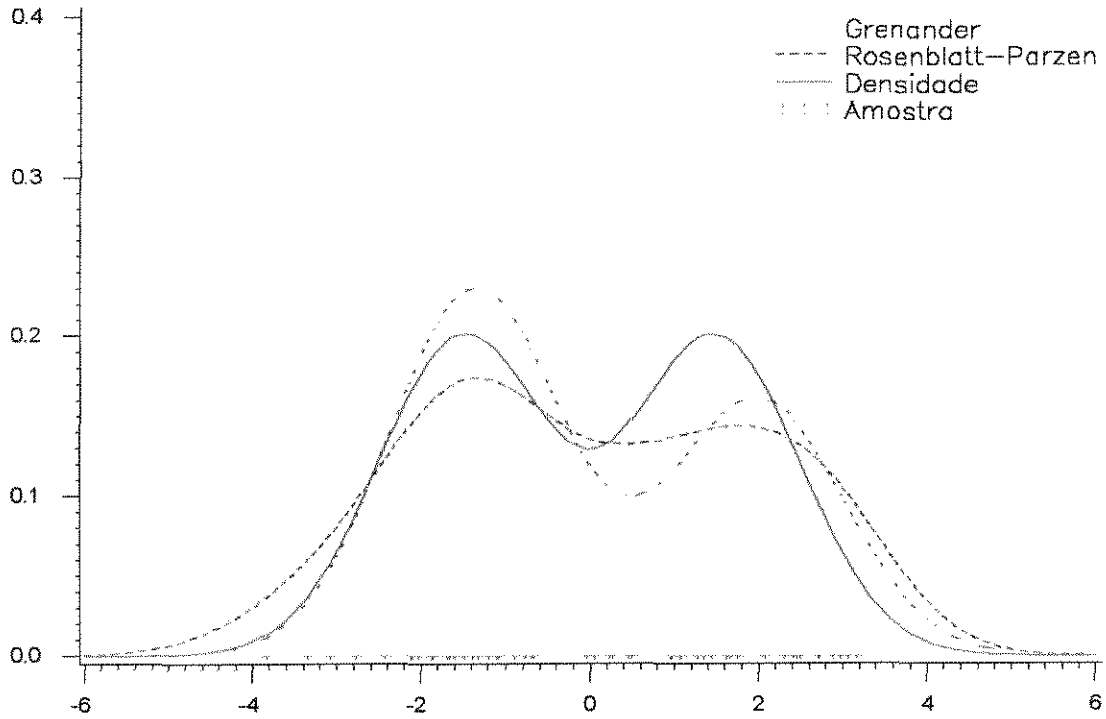


Gráfico 5.9 Função de densidade 1ª Mistura de Normais e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.6, tamanho de amostra n=50.

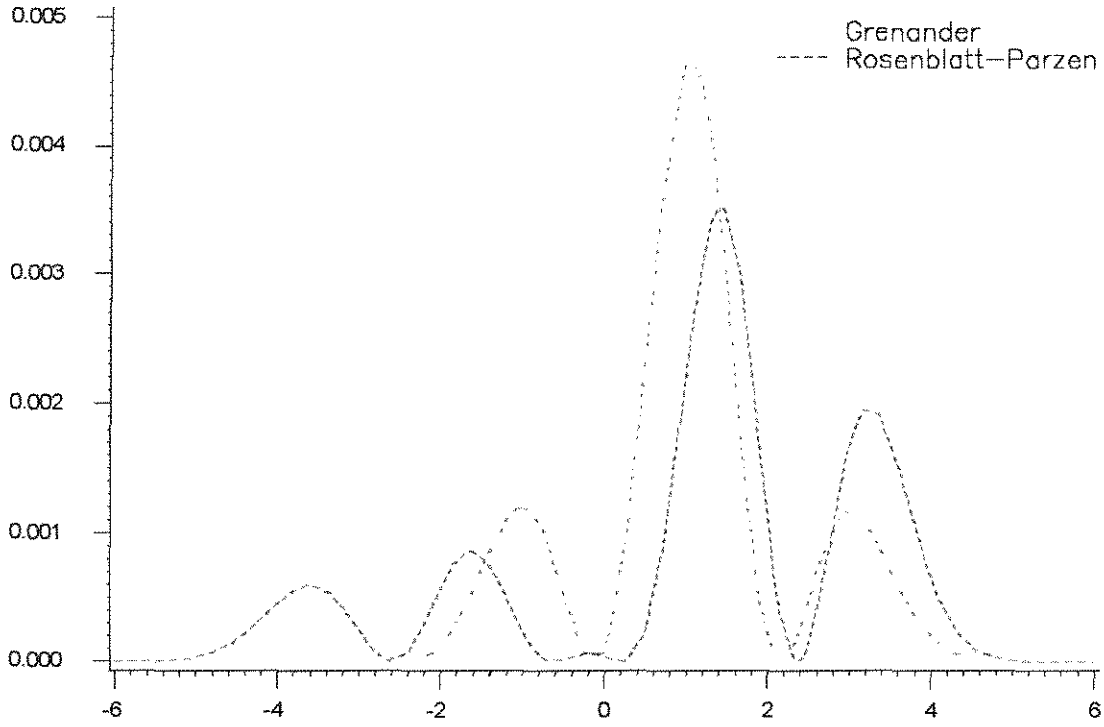


Gráfico 5.10 Comportamento pontual da diferença $\left(\hat{f}(x)-f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade 1ª Mistura de Normais, n=50.

Densidade 1ªMistura de Normais, n=100

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0041481	$EQMI$ =0.0095824
$\hat{h}$ = 0.42032	$\hat{\sigma}^2$ = 0.49203
	$\hat{p}_1$=0.49557 $\quad$ $\hat{\mu}_1$ = -1.6337
	$\hat{p}_2$=0.31346 $\quad$ $\hat{\mu}_2$ = 0.82632
	$\hat{p}_3$=0.19097 $\quad$ $\hat{\mu}_3$ = 2.47937

Tabela 5.7: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=100.

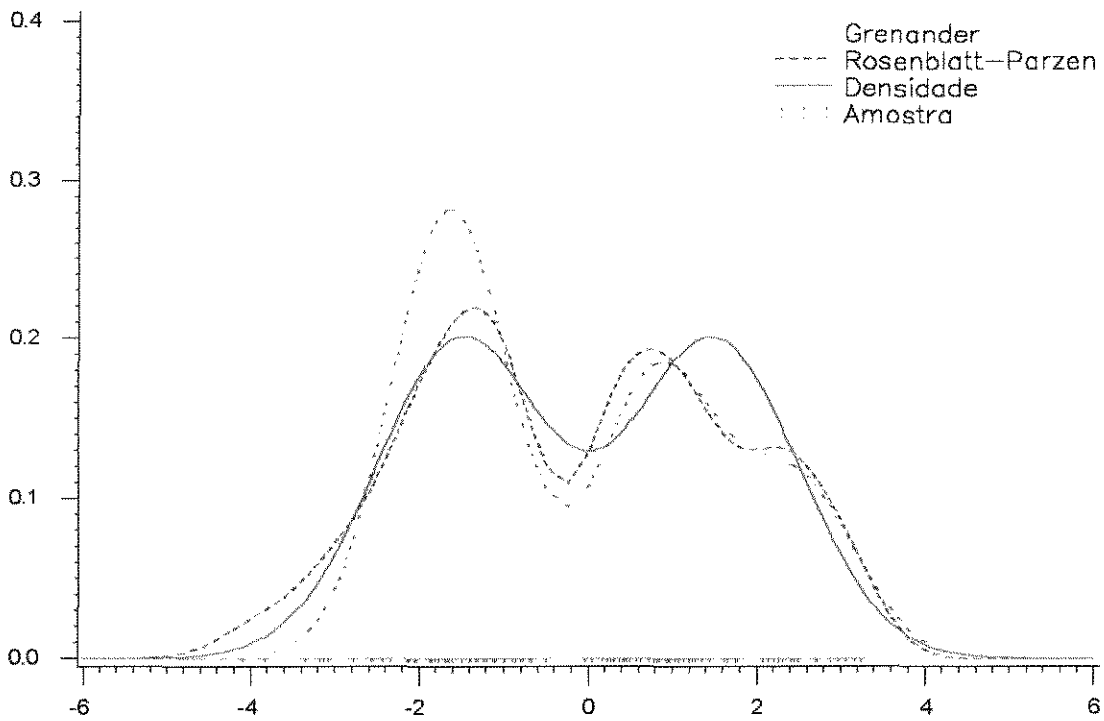


Gráfico 5.11 Função de densidade 1ª Mistura de Normais e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.7, tamanho de amostra n=100.

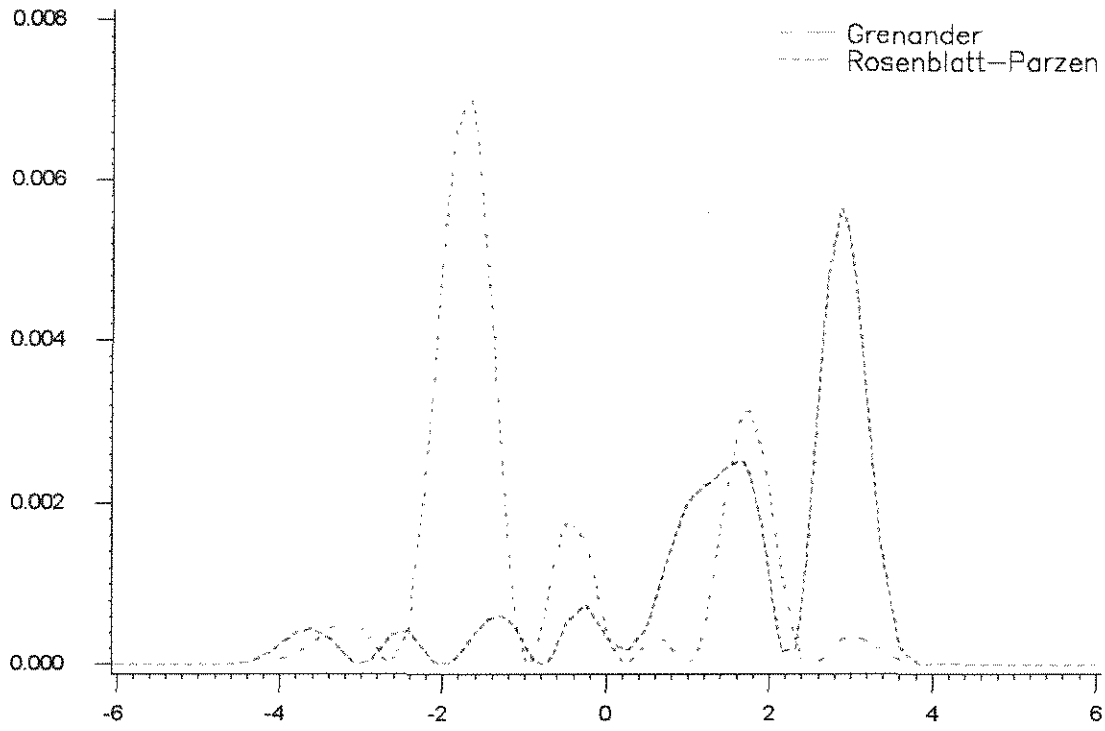


Gráfico 5.12 Comportamento pontual da diferença $\left(\hat{f}(x)-f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade 1ª Mistura de Normais, n=100.

5.3.3 Resultado das comparações na densidade t-Student com 5 g.l.

A variável aleatória da densidade t-Student obteve-se dividindo uma variável com densidade normal padrão pela raiz quadrada de outra variável aleatória com densidade qui-quadrado com 5 graus de liberdade dividida por seus graus de liberdade, ou seja, se $Z \sim N(0,1)$ e $X \sim \chi^2(5)$, então

$$T = \frac{Z}{\sqrt{X/5}} \sim t_5 .$$

Para gerar valores da variável com densidade $\chi^2(5)$, utilizou-se o fato de se Y~Exponencial(½), então

$$Y = \sum_{i=1}^{m} Y_i \sim \chi^2(2m)$$

e

$$X = \sum_{i=1}^{2} Y_i + Z^2 \sim \chi^2(5).$$

Rosenblatt-Parzen	Grenander
$EQMI$ =1.7279656	$EQMI$ =2.035743
$\hat{h}$ = 0.53735	$\hat{\sigma}^2$ = 0.10186
	$\hat{p}_1$=0.11967 $\quad$ $\hat{\mu}_1$ = -1.37079
	$\hat{p}_2$=0.29307 $\quad$ $\hat{\mu}_2$ = 1.08747
	$\hat{p}_3$=0.58726 $\quad$ $\hat{\mu}_3$ = 0.01526

Tabela 5.8: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=25.

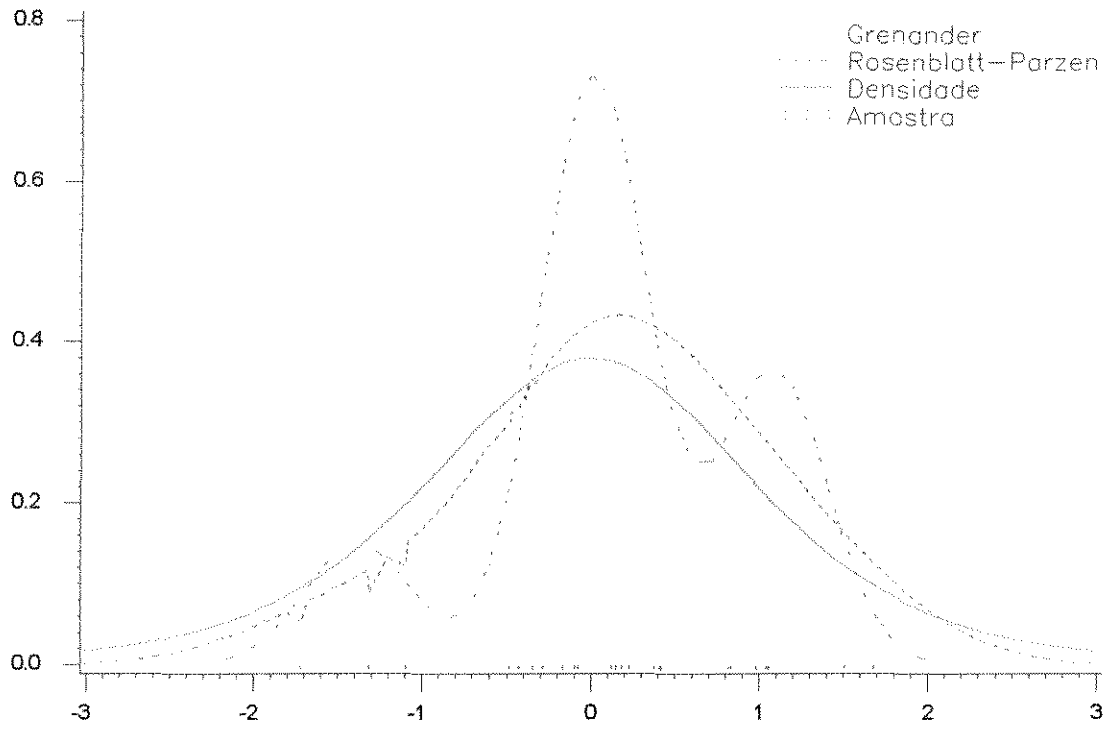


Gráfico 5.13 Função de densidade t-Student com 5 g.l. e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.8, tamanho de amostra n=25.

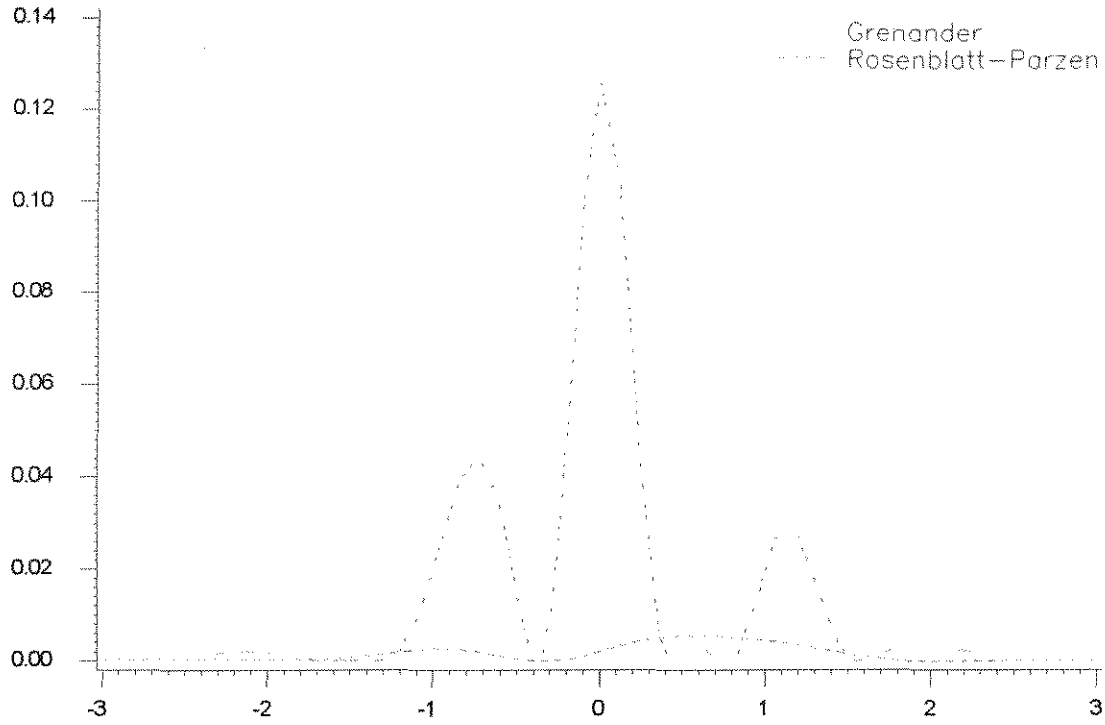


Gráfico 5.14 Comportamento pontual da diferença $\left(\tilde{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade t-Student com 5 g.l., n=25.

Densidade t-Student com 5 g.l., n=50

Rosenblatt-Parzen	Grenander
$EQMI$ =1.7279656	$EQMI$ =1.978845
$\hat{h}$ = 0.53735	$\hat{\sigma}^2$ = 0.64671
	$\hat{p}_1$=0.93474 $\hat{\mu}_1$ = 0.37182
	$\hat{p}_2$=0.06526 $\hat{\mu}_2$ = -1.76381

Tabela 5.9: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=50.

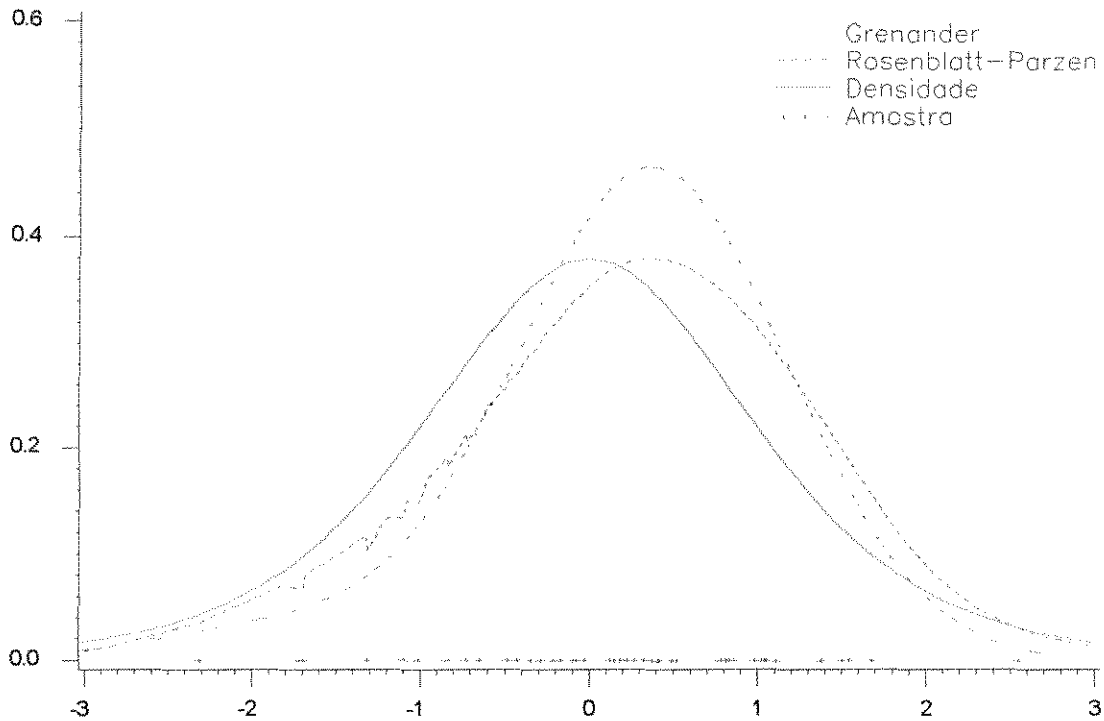


Gráfico 5.15 Função de densidade t-Student com 5 g.l. e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.9, tamanho de amostra n=50.

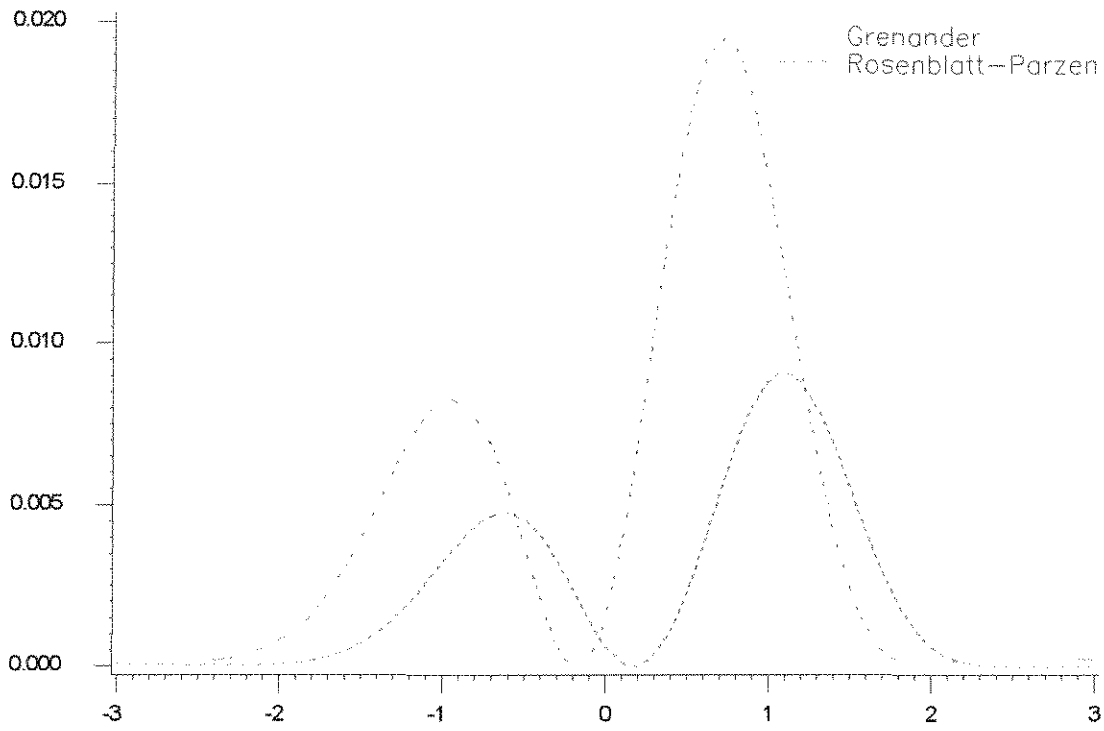


Gráfico 5.16 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade t-Student com 5 g.l., n=50.

Densidade t-Student com 5 g.l., n=100

Rosenblatt-Parzen	Grenander
$EQMI$ =1.7278669	$EQMI$ =1.9559146
$\hat{h}$ = 0.52771	$\hat{\sigma}^2$ = 1.0646
	$\hat{p}_1$=1 $\hat{\mu}_1$ = 0.14651

Tabela 5.10: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=100.

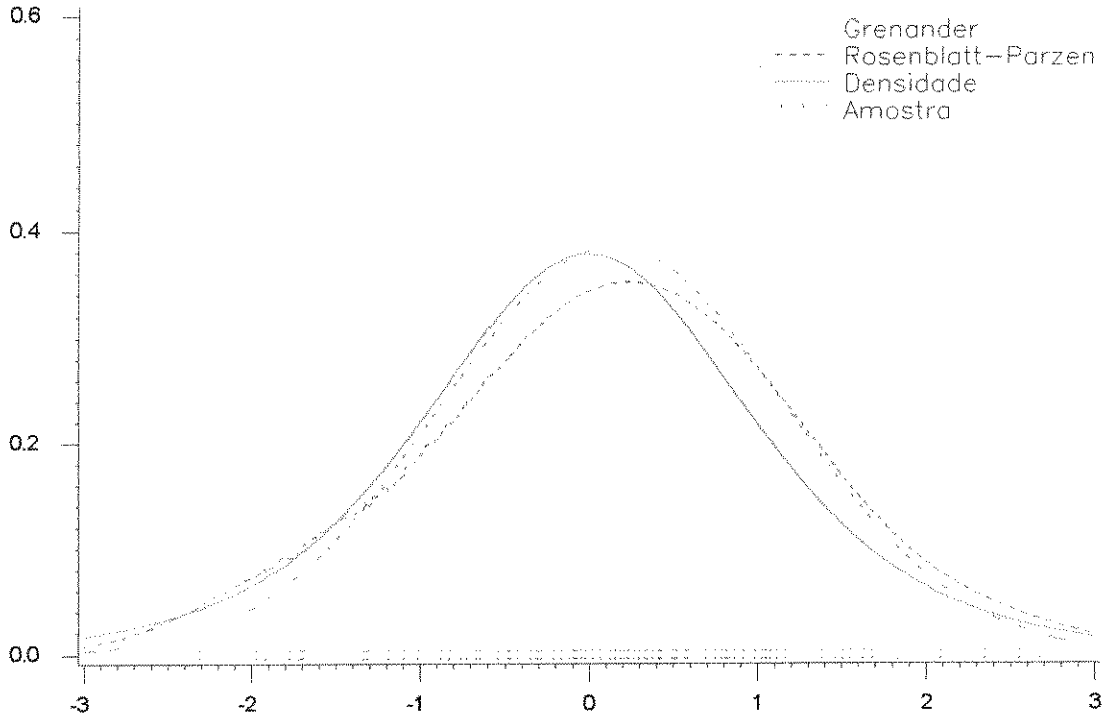


Gráfico 5.17 Função de densidade t-Student com 5 g.l. e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.10, tamanho de amostra n=100.

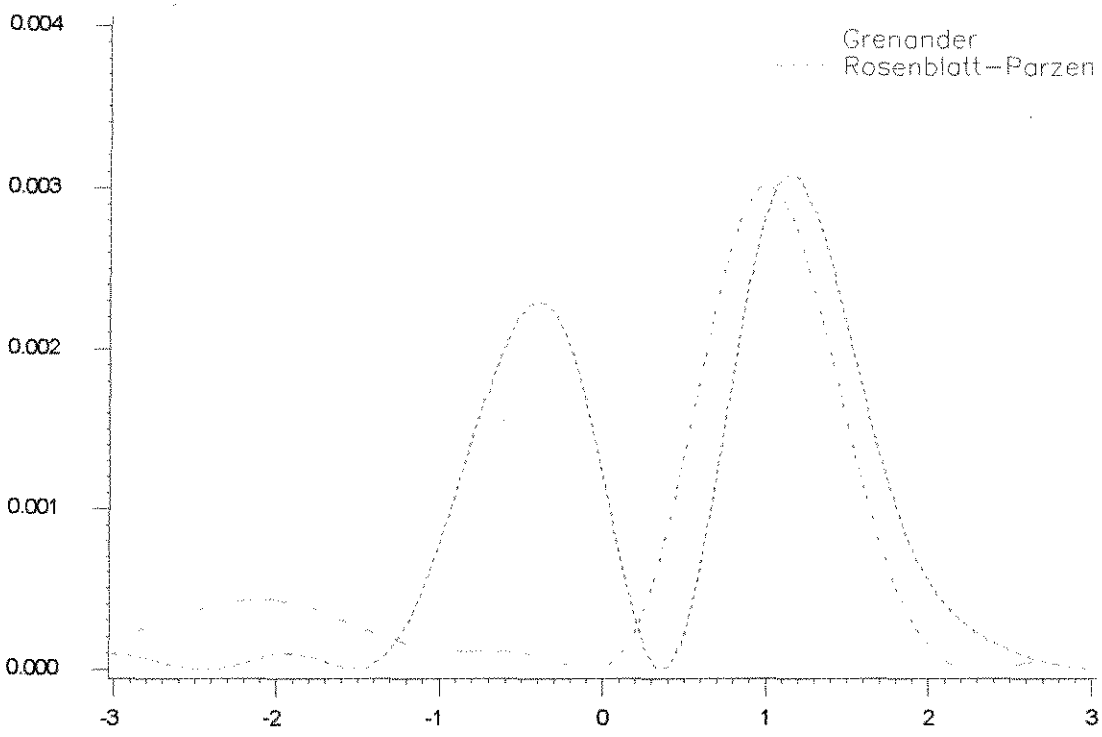


Gráfico 5.18 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade t-Student com 5 g.l., n=100.

5.3.4 Resultado das comparações na densidade Cauchy Padrão

As amostras para esta função de densidade de probabilidade foram geradas diretamente pela função SAS, RANCAU.

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0235206	$EQMI$ =0.0343188
$\hat{h}$ = 1.49338	$\hat{\sigma}^2$ = 2.97013
	$\hat{p}_1$=0.91599 $\hat{\mu}_1$ = -0.84008
	$\hat{p}_2$=0.08421 $\hat{\mu}_2$ = 4.9157

Tabela 5.11: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=25.

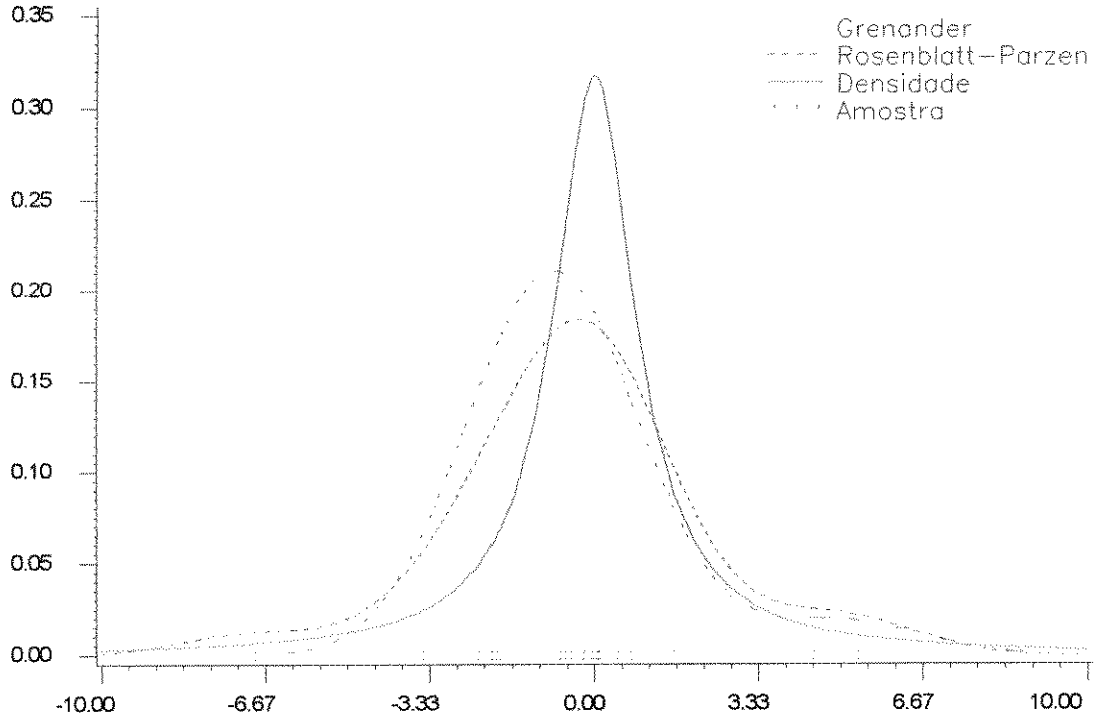


Gráfico 5.19 Função de densidade Cauchy Padrão e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.11, tamanho de amostra n=25.

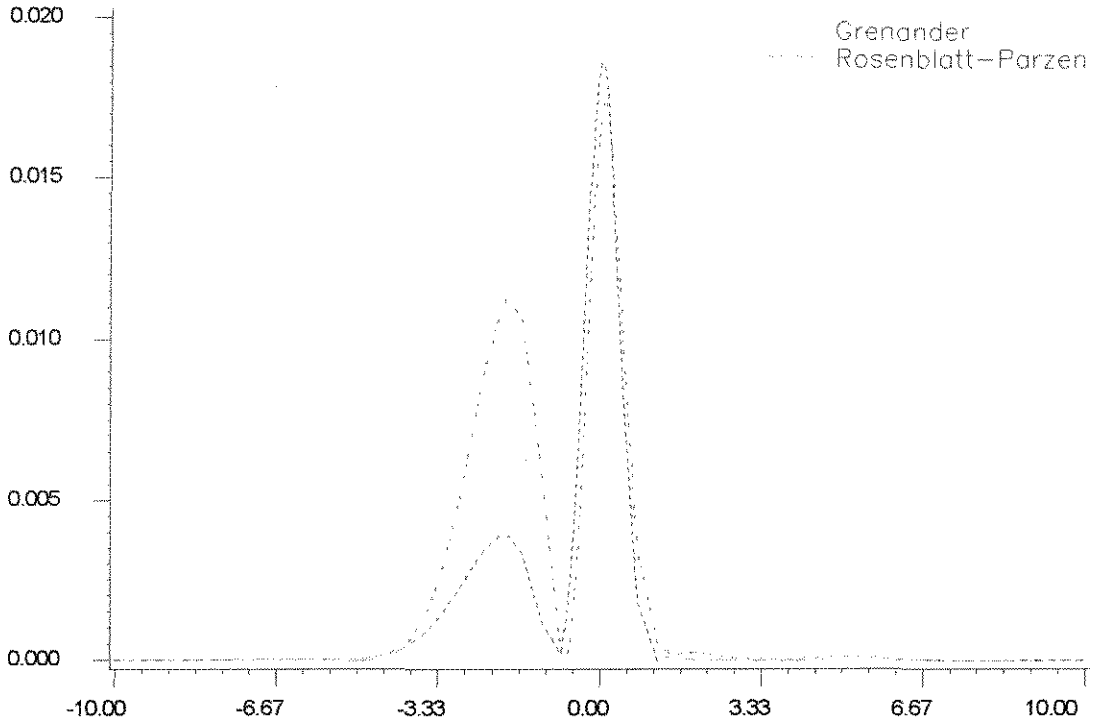


Grafico 5.20 Comportamento pontual da diferença $\left(\hat{f}(x)-f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Cauchy Padrão, n=25.

Densidade Cauchy Padrão, n=50

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0199558	$EQMI$ =0.023294
$\hat{h}$ = 1.06768	$\hat{\sigma}^2$ = 3.9761
	$\hat{p}_1$=0.96735 $\hat{\mu}_1$ = -0.29686
	$\hat{p}_2$=0.03265 $\hat{\mu}_2$ = 4.99337

Tabela 5.12: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=50.

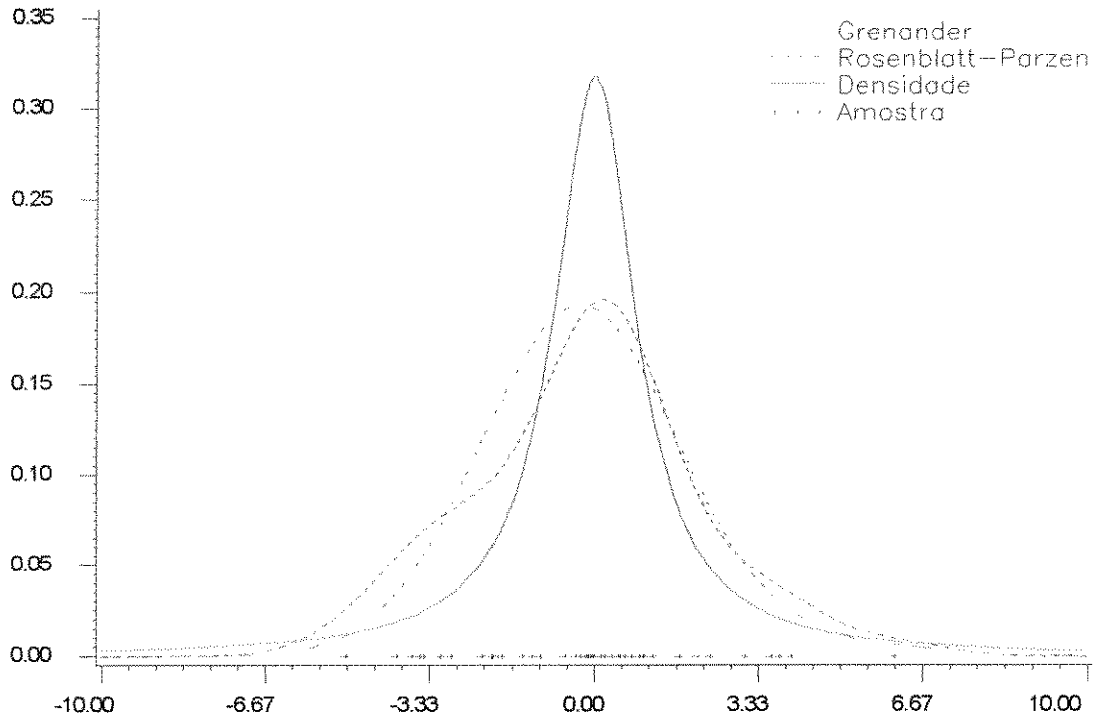


Gráfico 5.21 Função de densidade Cauchy Padrão e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.12, tamanho de amostra n=50.

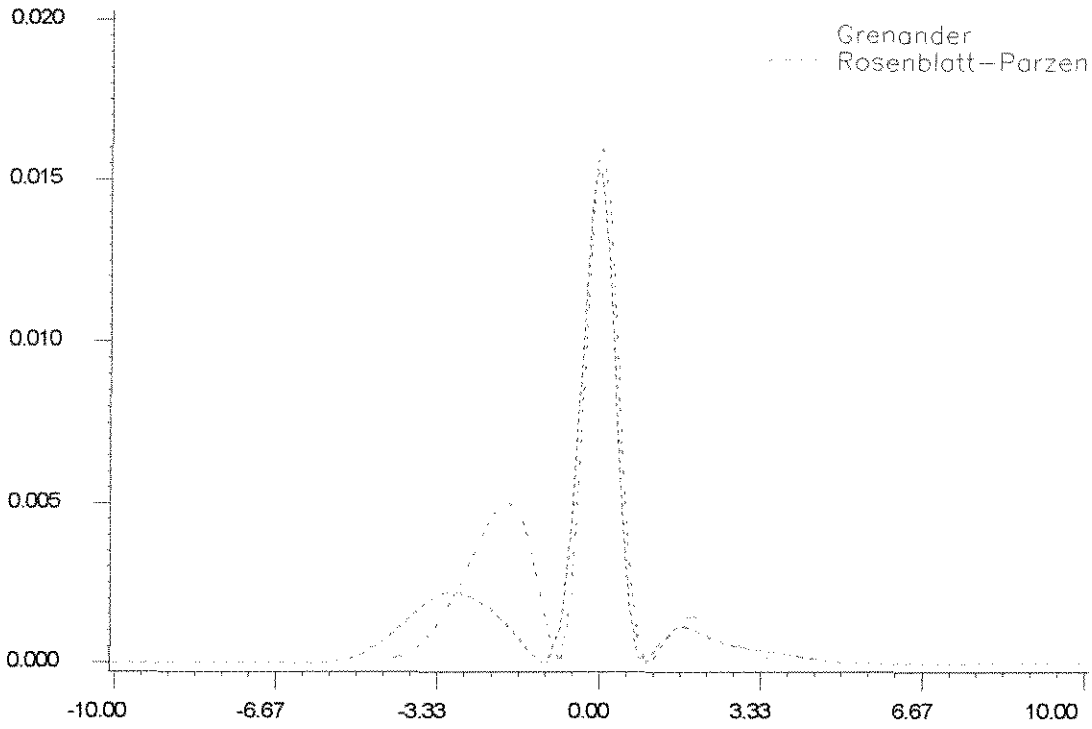


Gráfico 5.22 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Cauchy Padrão, n=50.

Densidade Cauchy Padrão, n=100

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0067652	$EQMI$ =0.0247666
$\hat{h}$ = 0.63053	$\hat{\sigma}^2$ = 4.32318
	$\hat{p}_1$=0.9695 $\hat{\mu}_1$ = -0.2673
	$\hat{p}_2$=0.0300 $\hat{\mu}_2$ = 6.82244

Tabela 5.13: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=100.

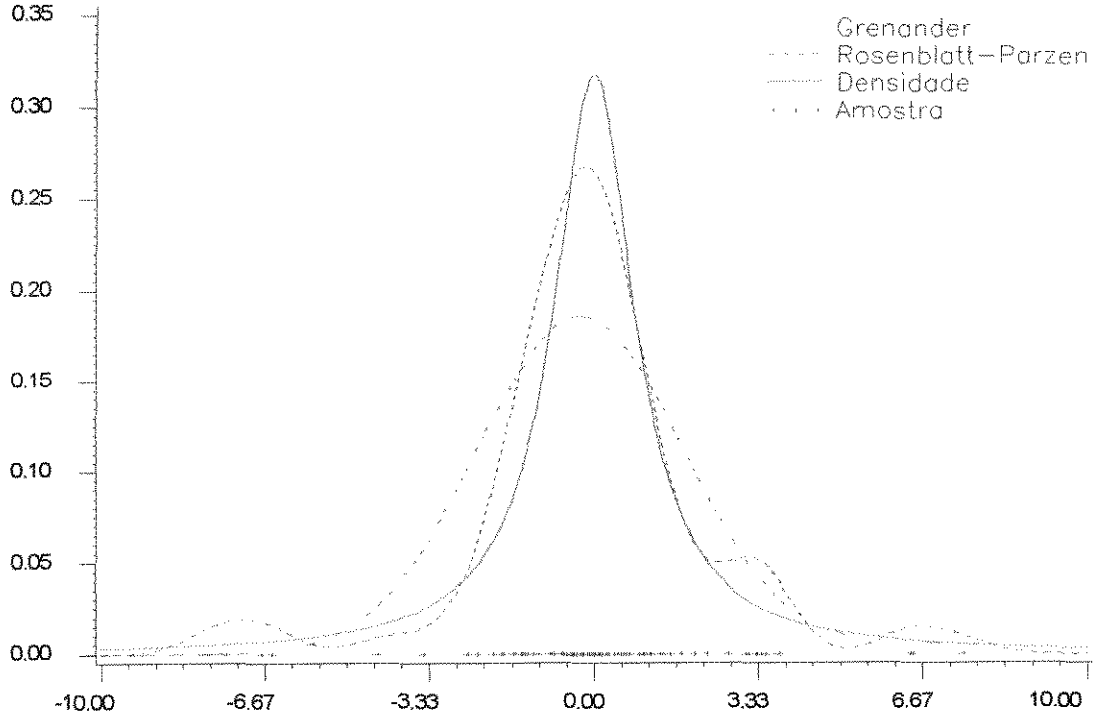


Gráfico 5.23 Função de densidade Cauchy Padrãoe os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.13, tamanho de amostra n=100.

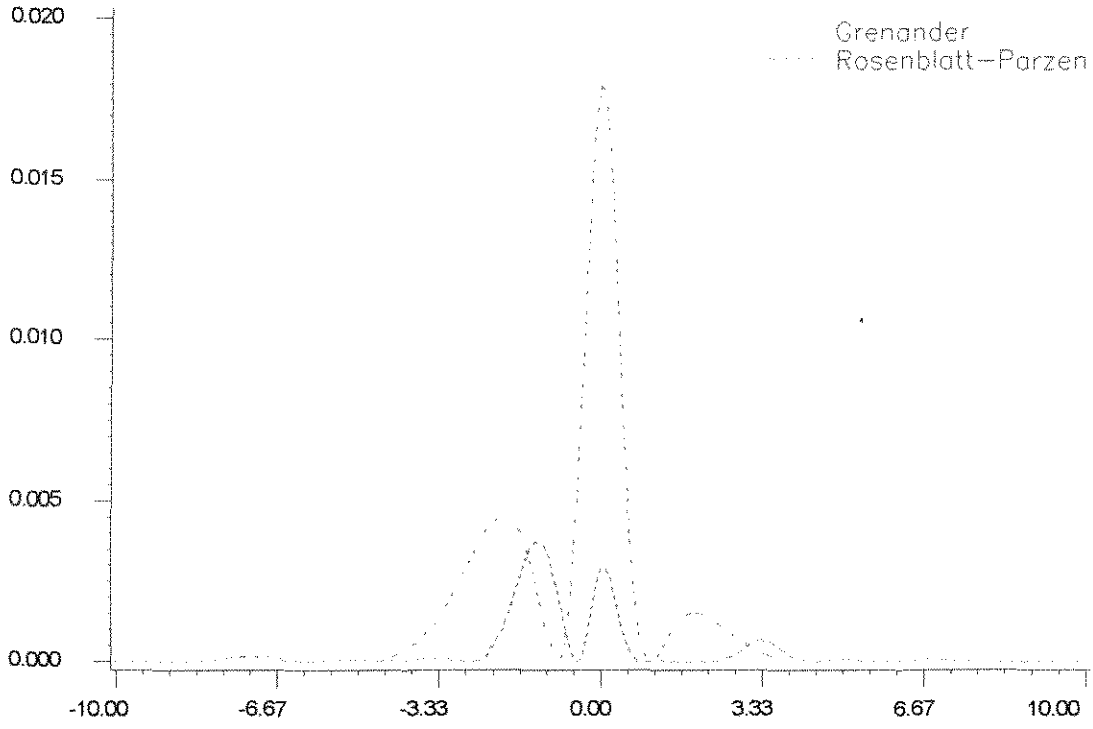


Gráfico 5.24 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Cauchy Padrão, n=100.

5.3.5 Resultado das comparações na densidade Qui-Quadrado com 6 g.l.

O pacote computacional utilizado nas comparações não têm uma função de geração de amostras para o caso da densidade de probabilidade χ^2, e o gerador de variáveis aleatórias Gama só permite um parâmetro nesta densidade. Por essa razão é que se procuraram procedimentos alternativos.

Lembremos que, se $Y_i \sim$ Exponencial(½) então

$$Y = \sum_{i=1}^{m} Y_i \sim \chi^2(2m).$$

Utilizando este resultado, geramos m=3 variáveis com densidade Exponencial(½), obtendo-se a amostra da densidade $\chi^2(6)$. Devido à que a função disponível no pacote computacional, RANEXP só gera amostras da densidade Exponencial(1), multiplicamos por 2 o valor obtido da função RANEXP, deste modo se obtêm um valor da densidade Exponencial(½).

Rosenblatt-Parzen	**Grenander**
$EQMI$ =0.0033452	$EQMI$ =0.0142278
$\hat{h}$ =1.67994	$\hat{\sigma}^2 = 2.28112$
	$\hat{p}_1 = 0.73547$ $\quad \hat{\mu}_1 = 4.6614$
	$\hat{p}_2 = 0.06456$ $\quad \hat{\mu}_2 = 9.0231$
	$\hat{p}_3 = 0.19997$ $\quad \hat{\mu}_3 = 11.4279$

Tabela 5.14: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=25.

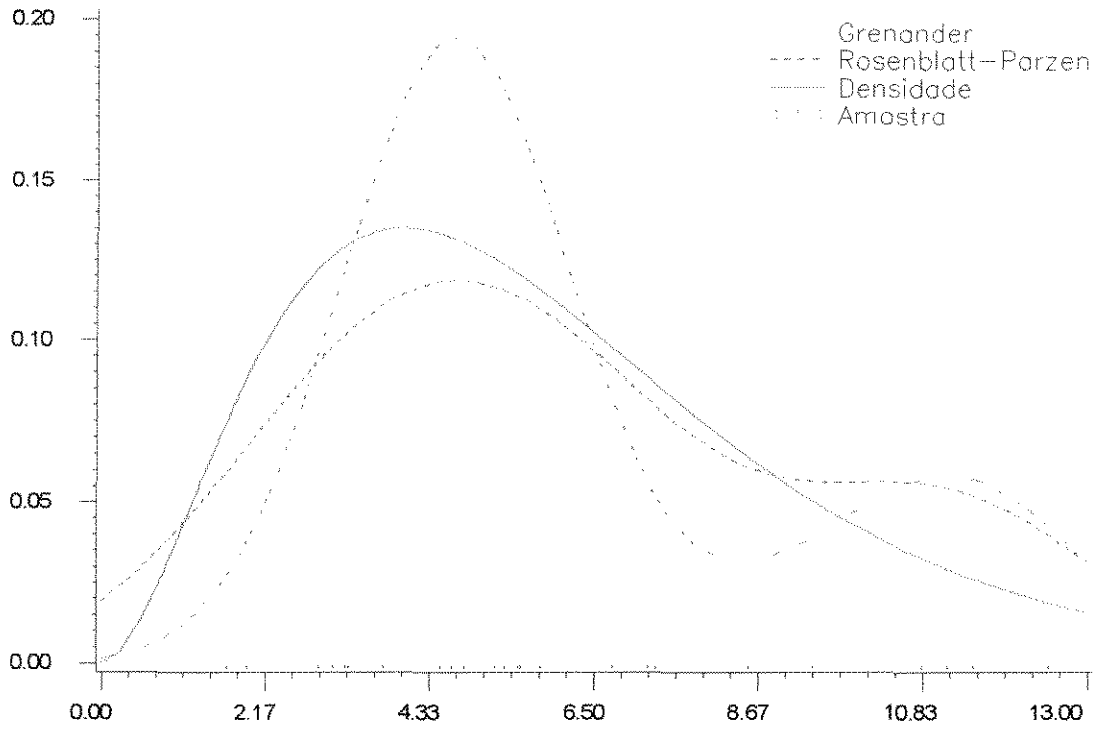


Gráfico 5.25 Função de densidade Qui-Quadrado com 6 g.l. e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.14, tamanho de amostra n=25.

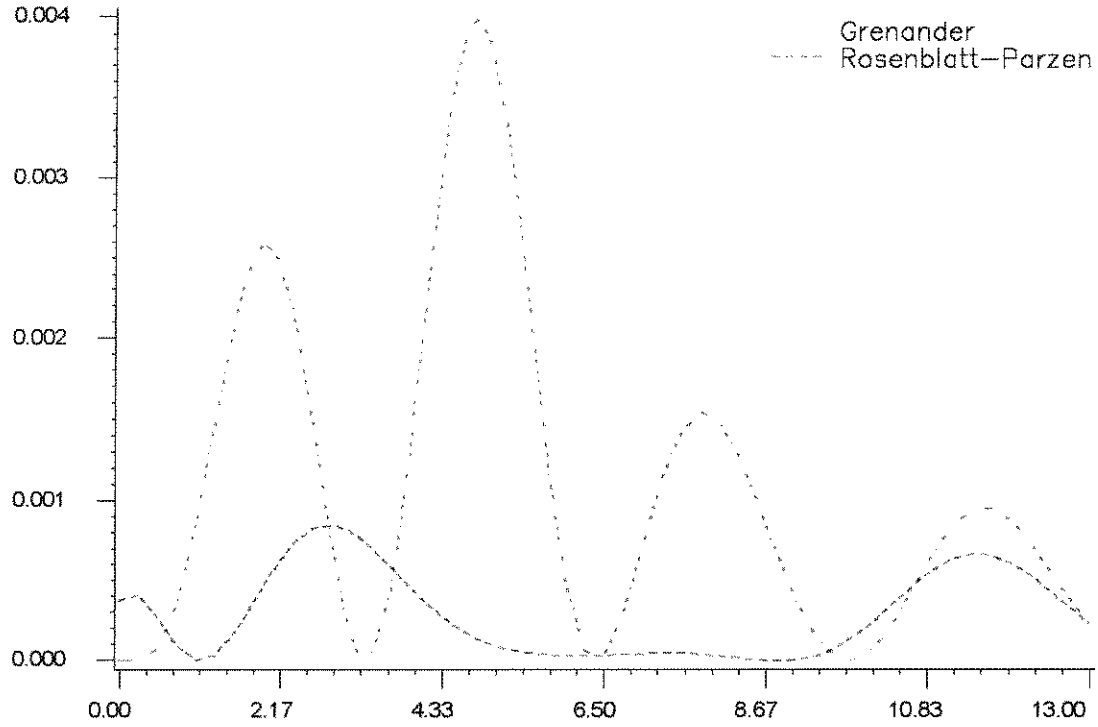


Gráfico 5.26 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Qui-Quadrado com 6 g.l., n=25.

Densidade Qui-Quadrado com 6 g.l., n=50

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0025289	$EQMI$ =0.003852
$\hat{h}$=1.77209	$\hat{\sigma}^2 = 4.8177$
	$\hat{p}_1 = 0.86757$ $\hat{\mu}_1 = 4.9679$
	$\hat{p}_2 = 0.13243$ $\hat{\mu}_2 = 11.5583$

Tabela 5.15: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=50.

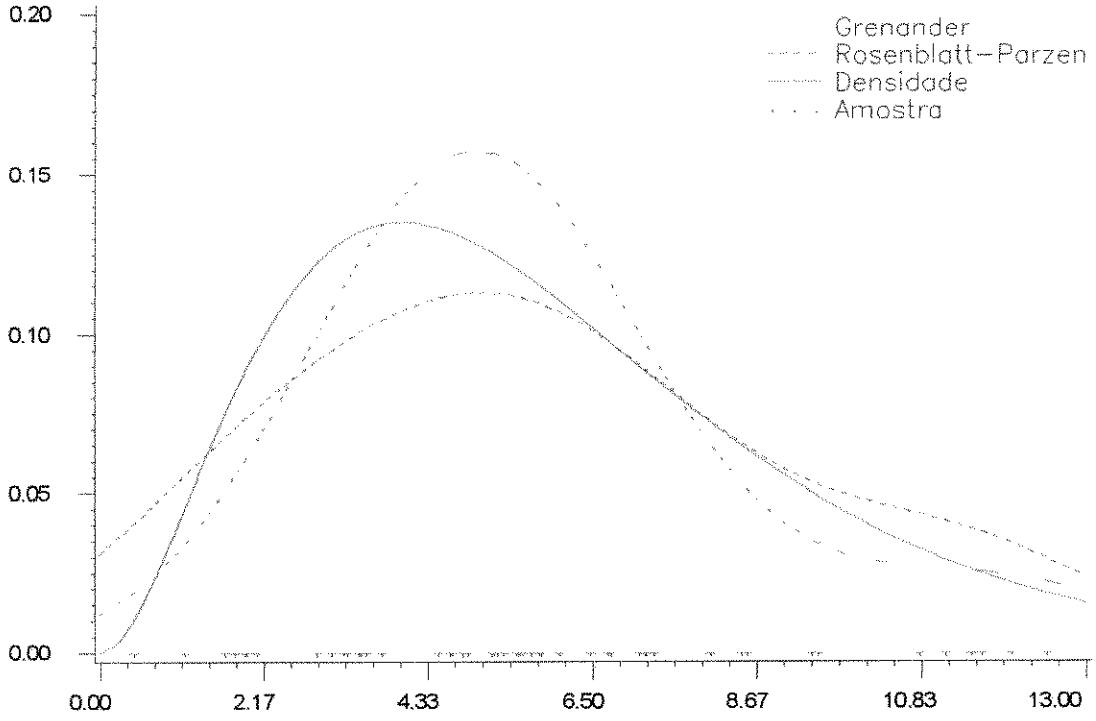


Gráfico 5.27 Função de densidade Qui-Quadrado com 6 g.l. e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.15, tamanho de amostra n=50.

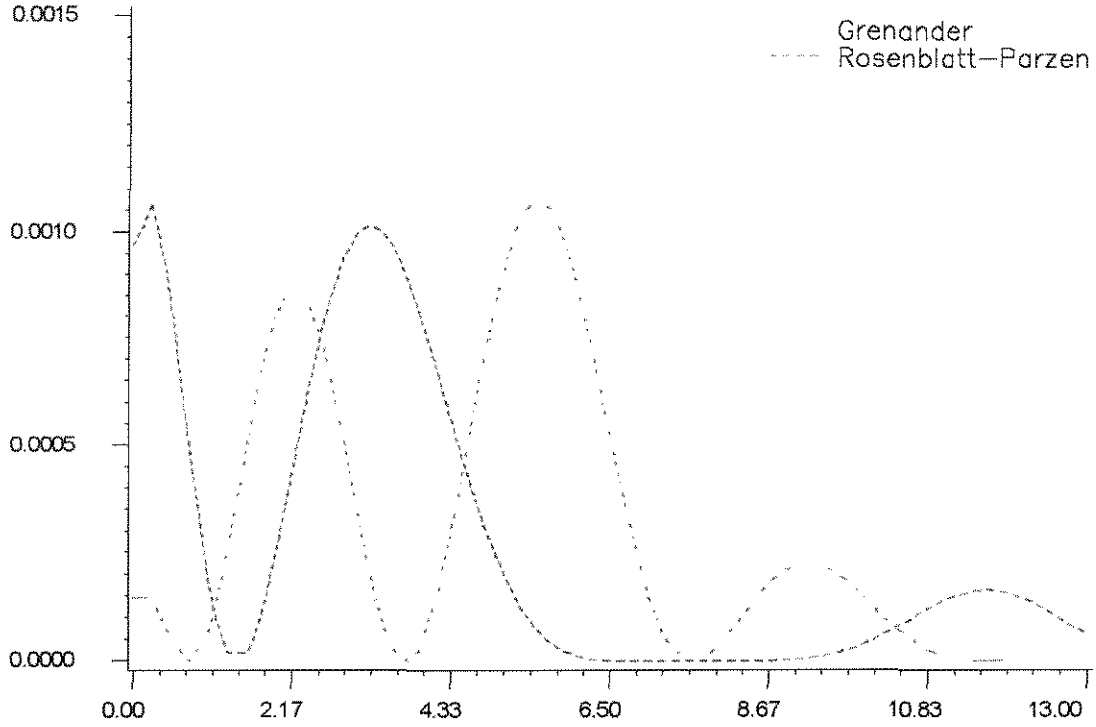


Gráfico 5.28 Comportamento pontual da diferença $\left(\hat{f}(x)-f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Qui-Quadrado com 6 g.l., n=50.

Densidade Qui-Quadrado com 6 g.l., n=100

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0003974	$EQMI$ =0.005346
$\hat{h}$ =1.16574	$\hat{\sigma}^2 = 3.2389$
	$\hat{p}_1 = 0.80871$ $\hat{\mu}_1 = 4.40252$
	$\hat{p}_2 = 0.19129$ $\hat{\mu}_2 = 9.91097$

Tabela 5.16: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=100.

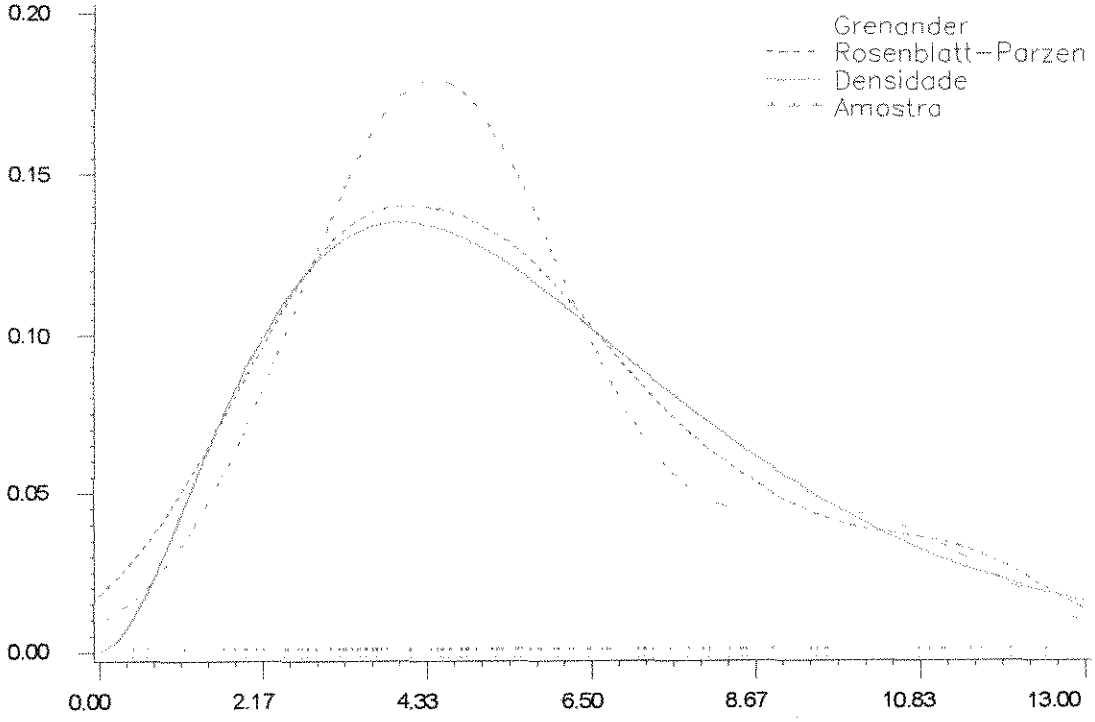


Gráfico 5.29 Função de densidade Qui-Quadrado com 6 g.l. e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.16, tamanho de amostra n=100.

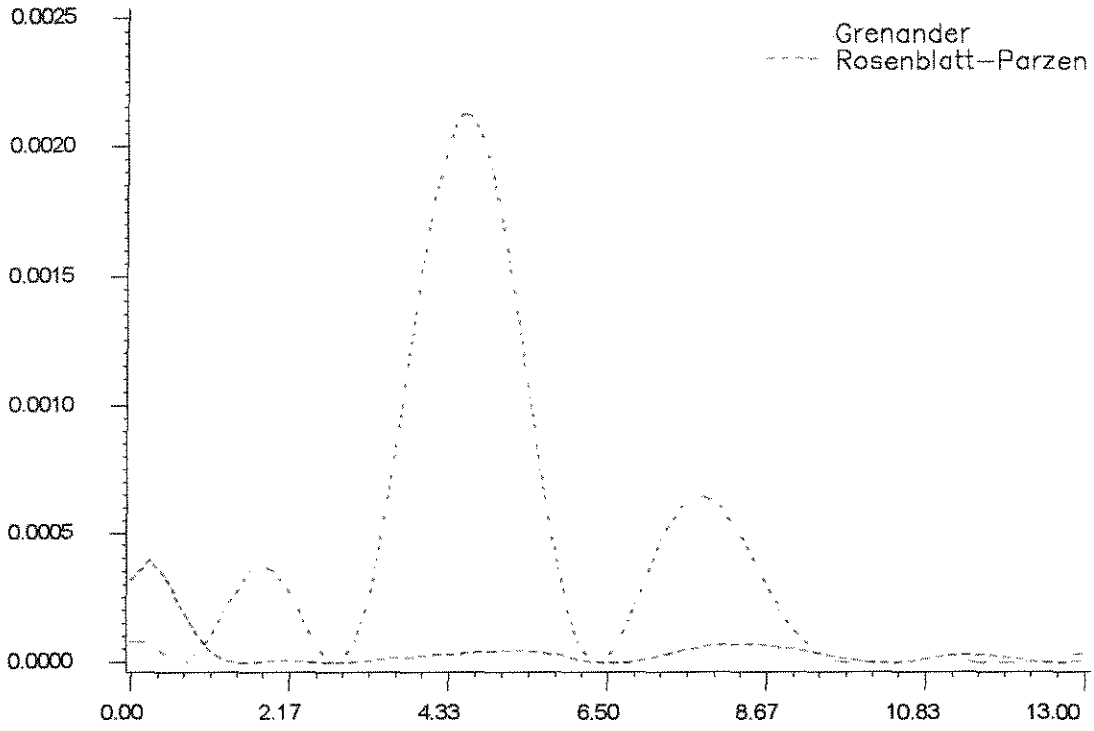


Gráfico 5.30 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Qui-Quadrado com 6 g.l., n=100.

5.3.6 Resultado das comparações na densidade Beta α=2, β=2

Nesta situação, para cada valor a ser gerado com função de densidade Beta(2,2), geraram-se duas variáveis X_1 e X_2 com densidade Gama(2,1), obtidas pela função RANGAM, e posteriormente se utilizou a relação $\frac{x_1}{x_1 + x_2}$ ~Beta(2,2).

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0118454	$EQMI$ =0.0674146
$\hat{h}$=0.18068	$\hat{\sigma}^2$ = 0.042137
	$\hat{p}_1 = 1$ $\hat{\mu}_1 = 0.48669$

Tabela 5.17: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=25.

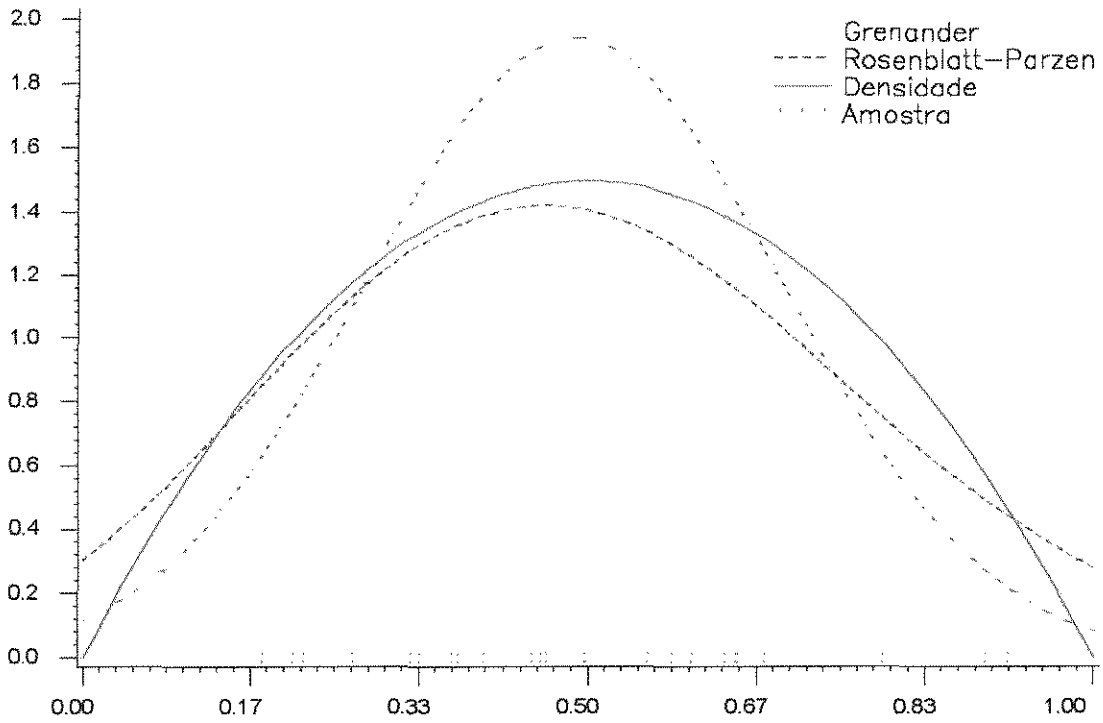


Gráfico 5.31 Função de densidade Beta α=2 β=2 e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.17, tamanho de amostra n=25.

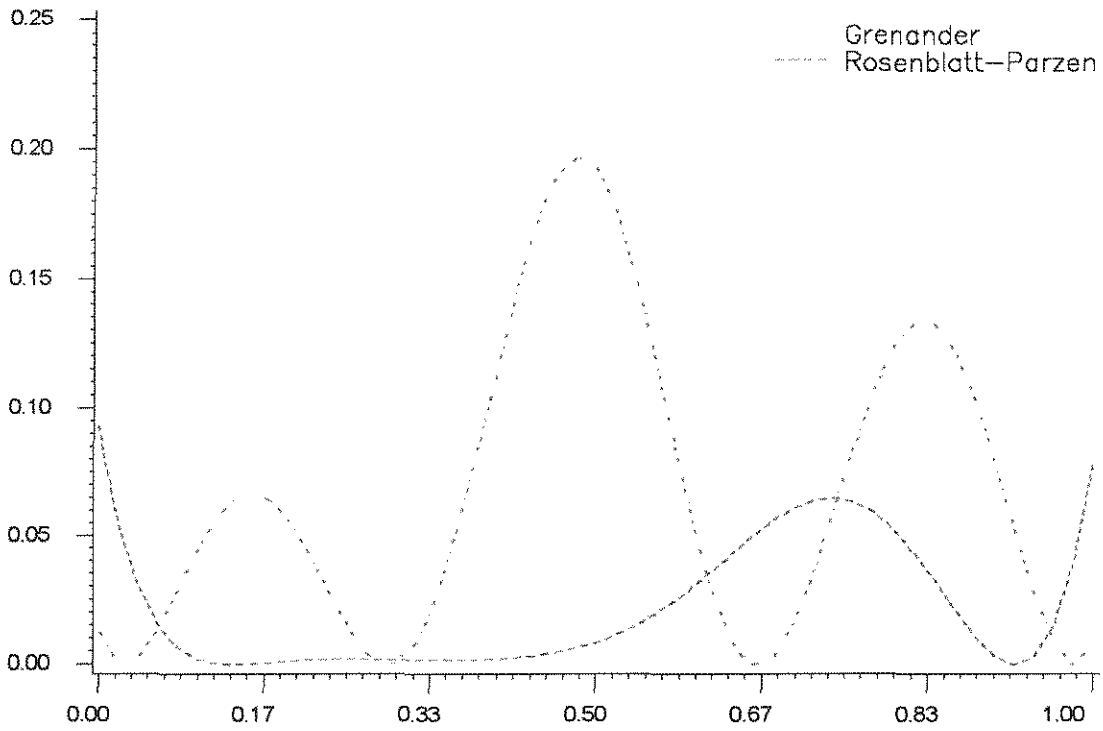


Gráfico 5.32 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenbaltt-Parzen e Grenander da função de densidade Beta α=2 β=2, n=25.

Densidade Beta com α=2 β=2, n=50

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0140144 $\hat{h}$=0.13755	$EQMI$ =0.075389 $\hat{\sigma}^2 = 0.040668$ $\hat{p}_1 = 1$ $\hat{\mu}_1 = 0.49152$

Tabela 5.18: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=50.

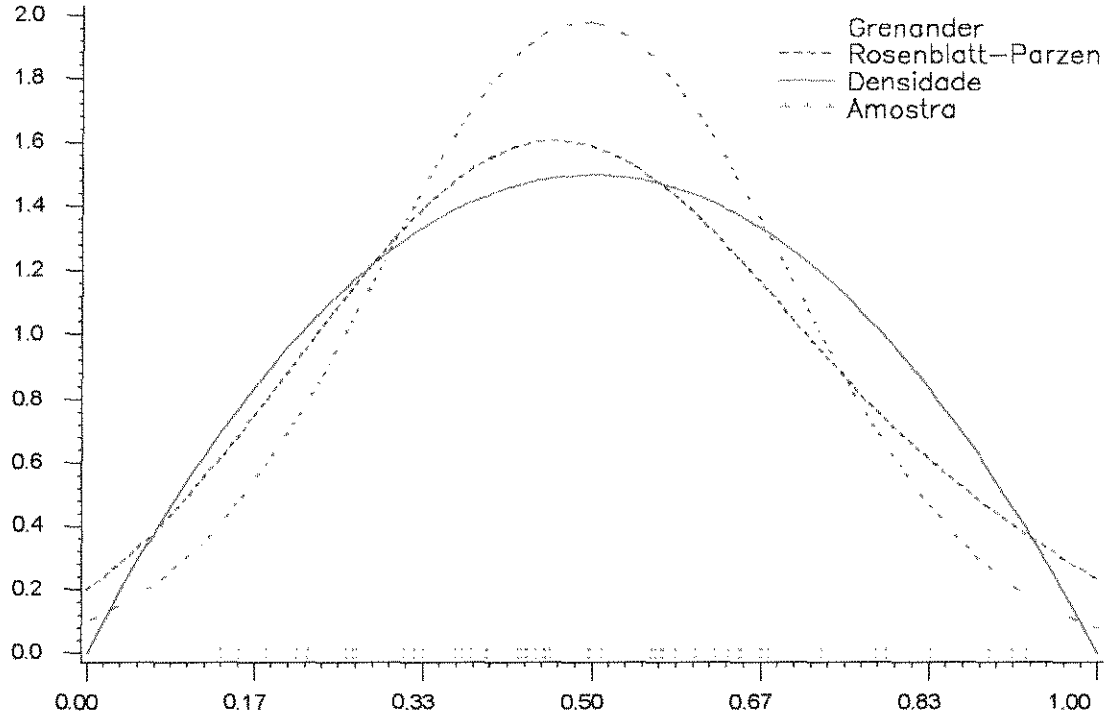


Gráfico 5.33 Função de densidade Beta α=2 β=2 e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.18, tamanho de amostra n=50.

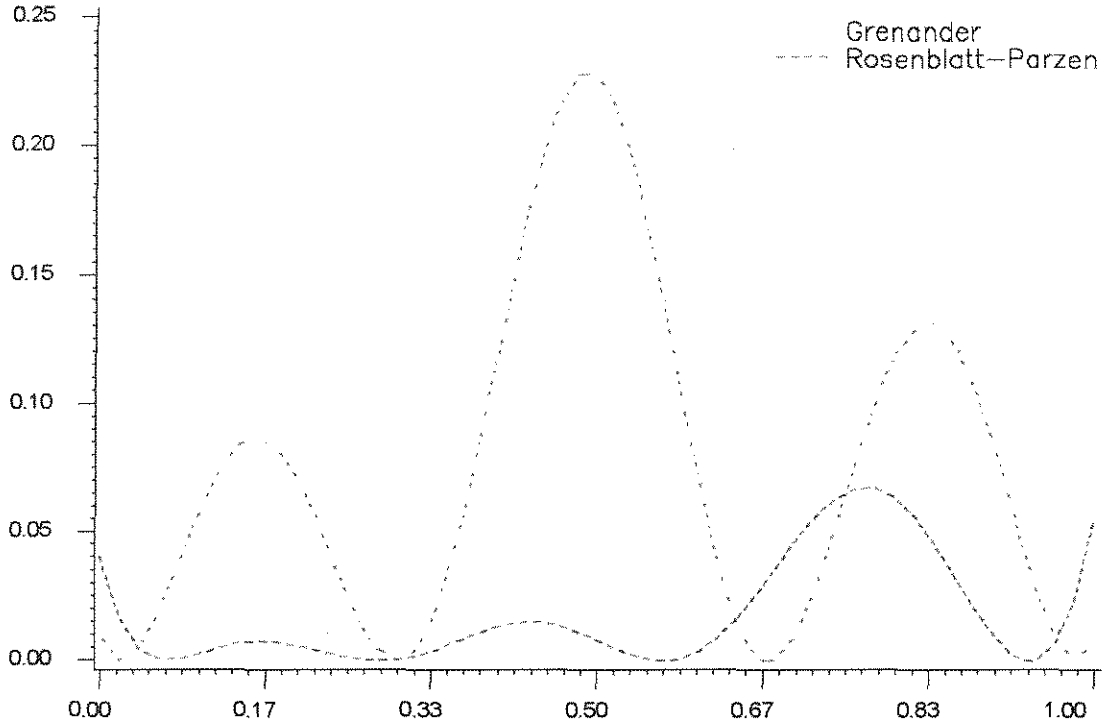


Gráfico 5.34 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenbaltt-Parzen e Grenander da função de densidade Beta α=2 β=2, n=50.

Densidade Beta com α=2 β=2, n=100

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0048005	$EQMI$ =0.0347462
$\hat{h}$=0.11959	$\hat{\sigma}^2$ = 0.023314
	$\hat{p}_1$ = 0.36214 $\quad$ $\hat{\mu}_1$ = 0.29864
	$\hat{p}_2$ = 0.63786 $\quad$ $\hat{\mu}_2$ = 0.60803

Tabela 5.19: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=100.

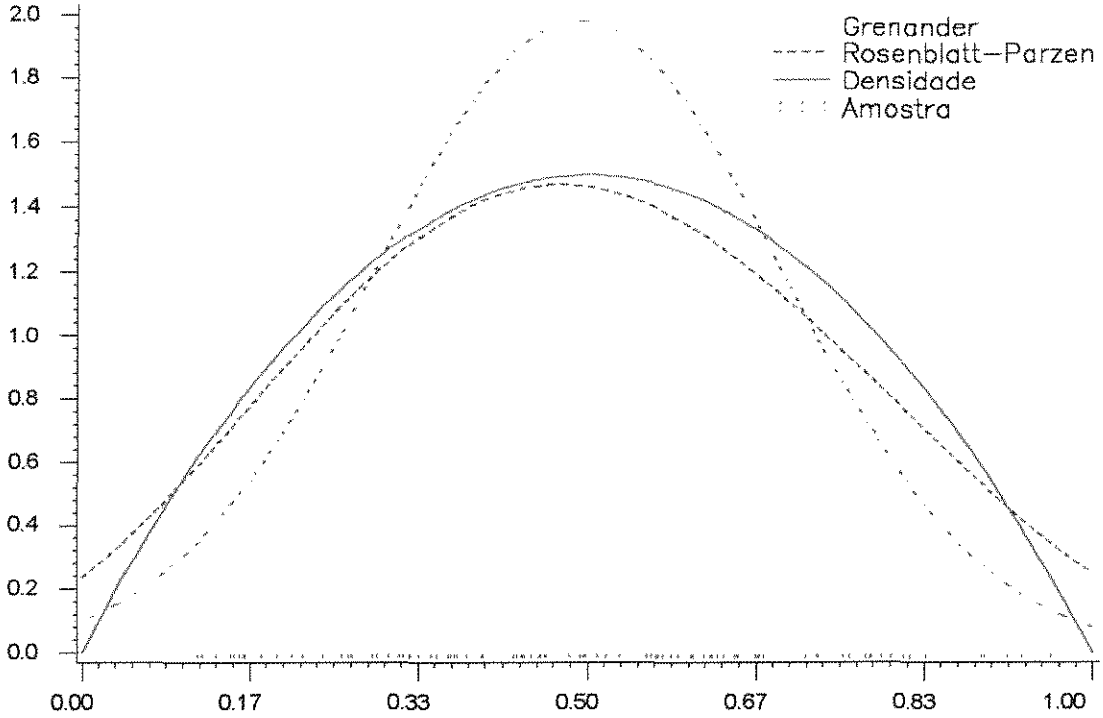


Gráfico 5.35 Função de densidade Beta α=2 β=2 e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.19, tamanho de amostra n=100.

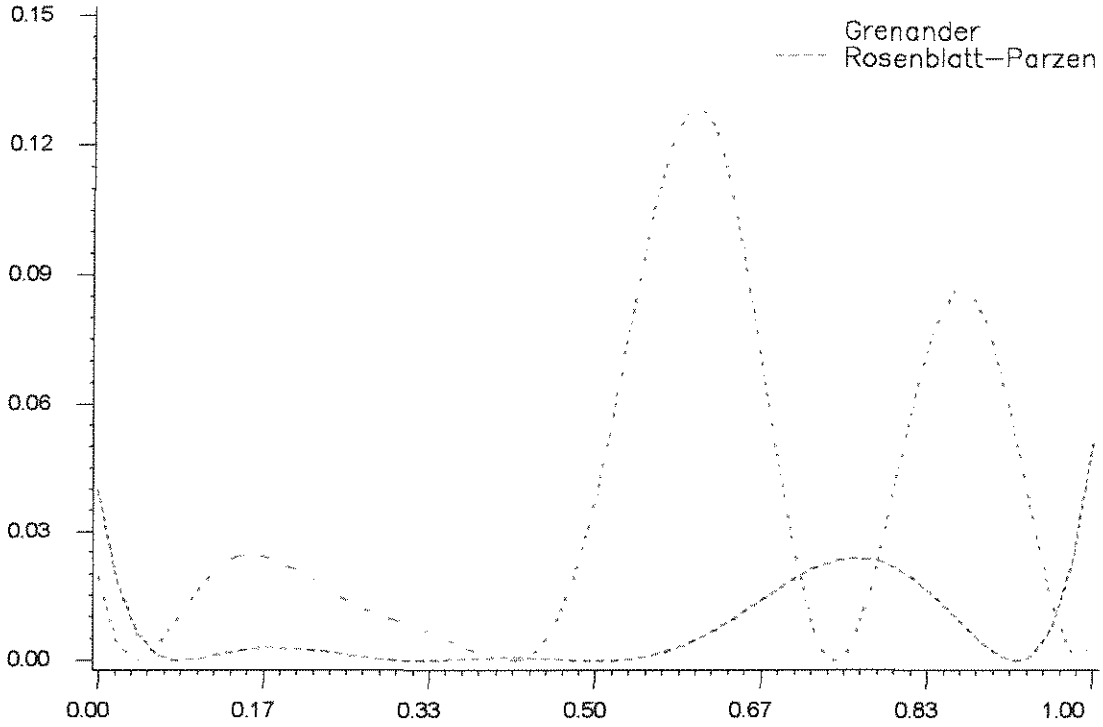


Gráfico 5.36 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Beta α=2 β=2, n=100.

5.3.7 Resultado das comparações na densidade Triangular

A chamada Transformação Integral ou Método de Inversão foi o utilizado para gerar as diferentes amostras. Consiste no seguinte, primeiramente geramos a mostra $U_1,\ldots,U_n \sim U(0,1)$ através da função RANUNI ou UNIFORM, e posteriormente calculamos

$$X_i = \begin{cases} \sqrt{4U_i + 1/4} & \text{se } U_i \le 1/2 \\ 3/2 - \sqrt{17/4 - 4U_i} & \text{se } U_i > 1/2 \end{cases}, \text{ para } i = 1,\ldots,n.$$

Rosenblatt-Parzen	Grenander
$EQMI$ =0.379473	$EQMI$ =0.5053747
$\hat{h}$ =0.51194	$\hat{\sigma}^2$ = 0.089592
	$\hat{p}_1$ = 0.4407 $\quad$ $\hat{\mu}_1$ = -0.36495
	$\hat{p}_2$ = 0.5593 $\quad$ $\hat{\mu}_2$ = 0.42002

Tabela 5.20: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=25.

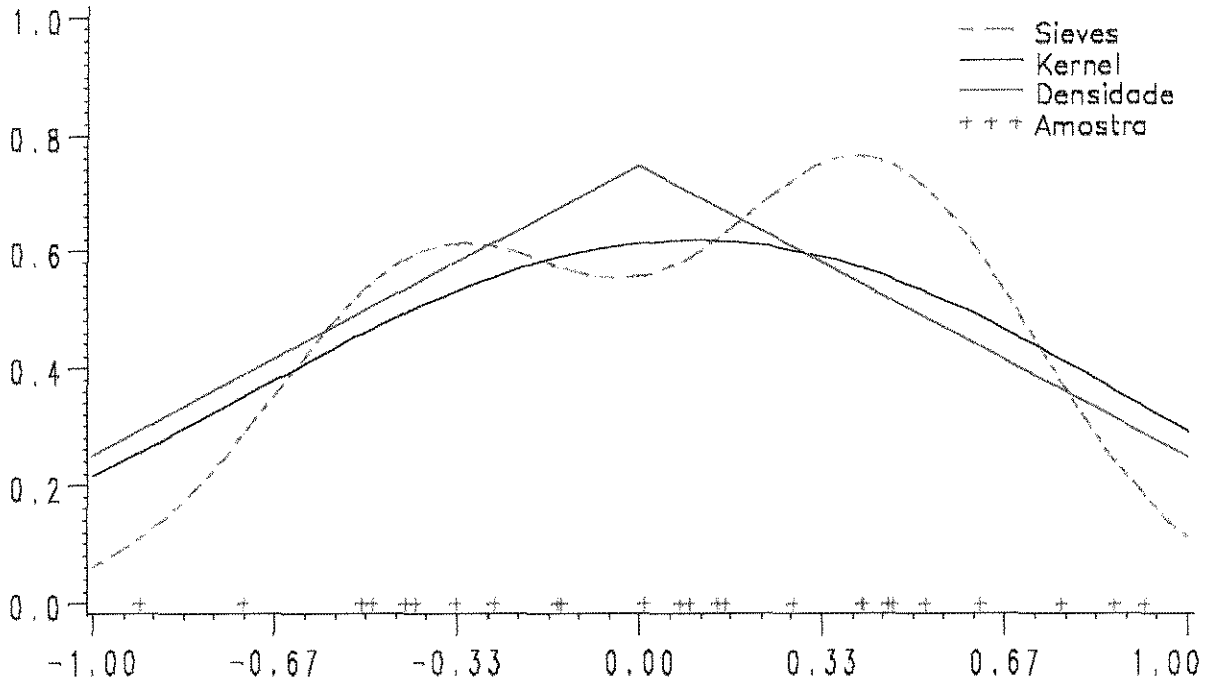


Gráfico 5.37 Função de densidade Triangular e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.20, tamanho de amostra n=25.

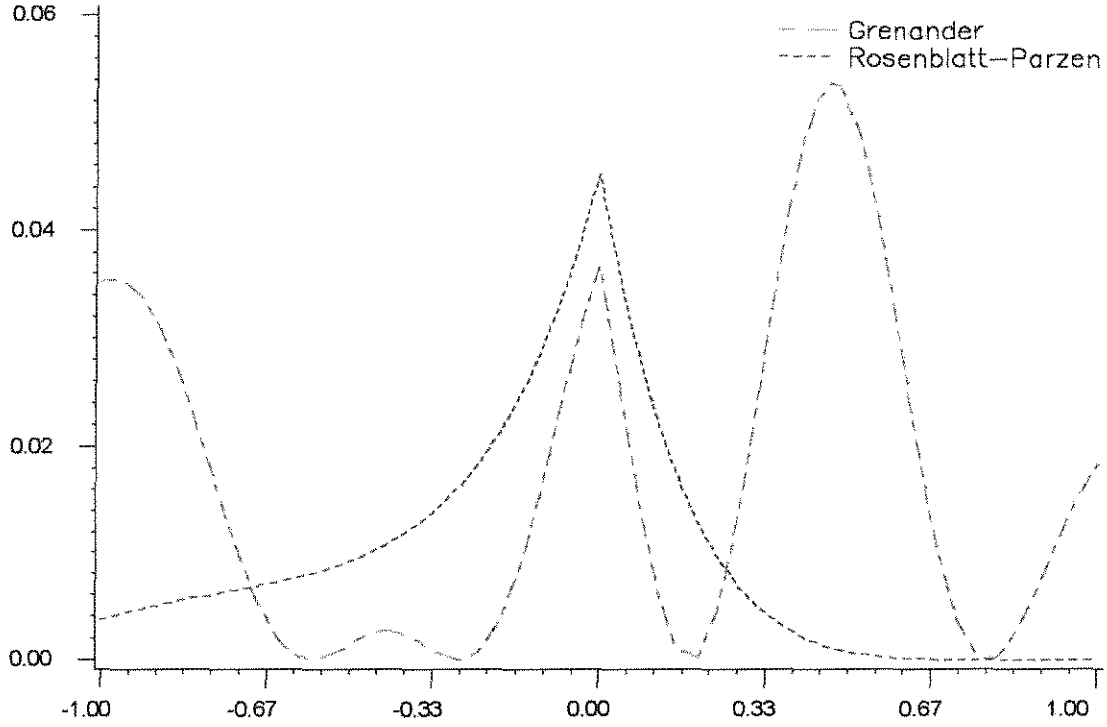


Gráfico 5.38 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Triangular, n=25.

Densidade Triangular, n=50

Rosenblatt-Parzen	Grenander
$EQMI$ =0.4257732 $\hat{h}$=0.42561	$EQMI$ =0.6227975 $\hat{\sigma}^2$ = 0.22786 $\hat{p}_1$ = 1 $\hat{\mu}_1$ = -0.02624

Tabela 5.21: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=50.

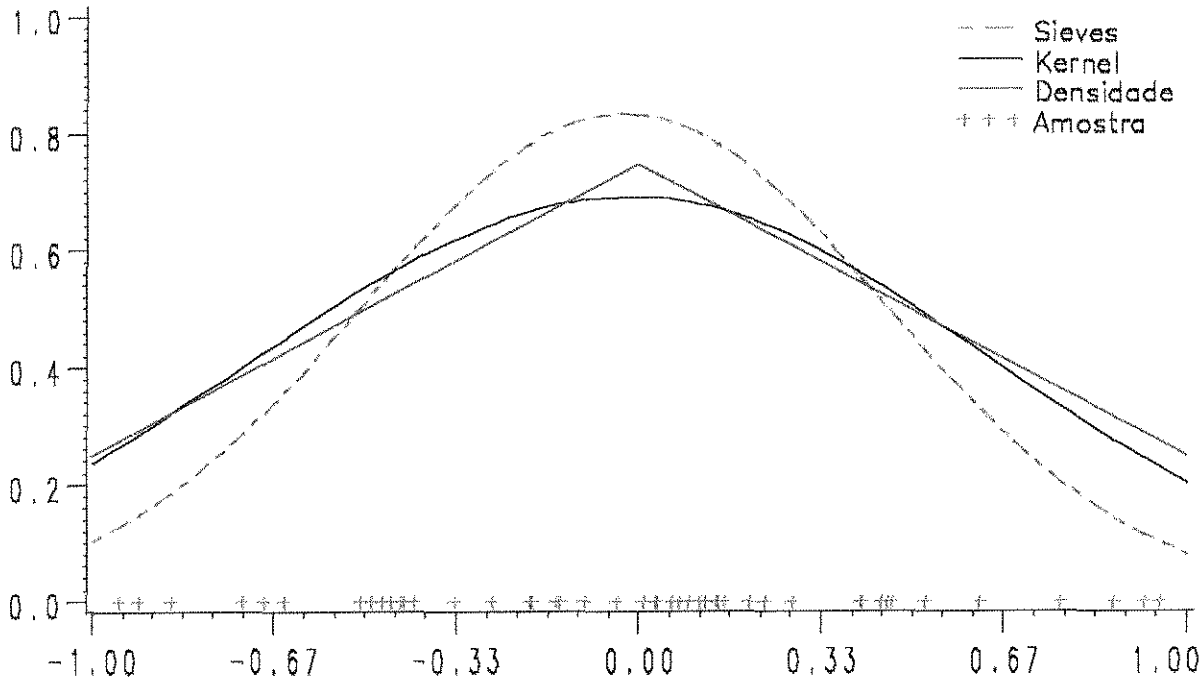


Gráfico 5.39 Função de densidade Triangular e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.21, tamanho de amostra n=50.

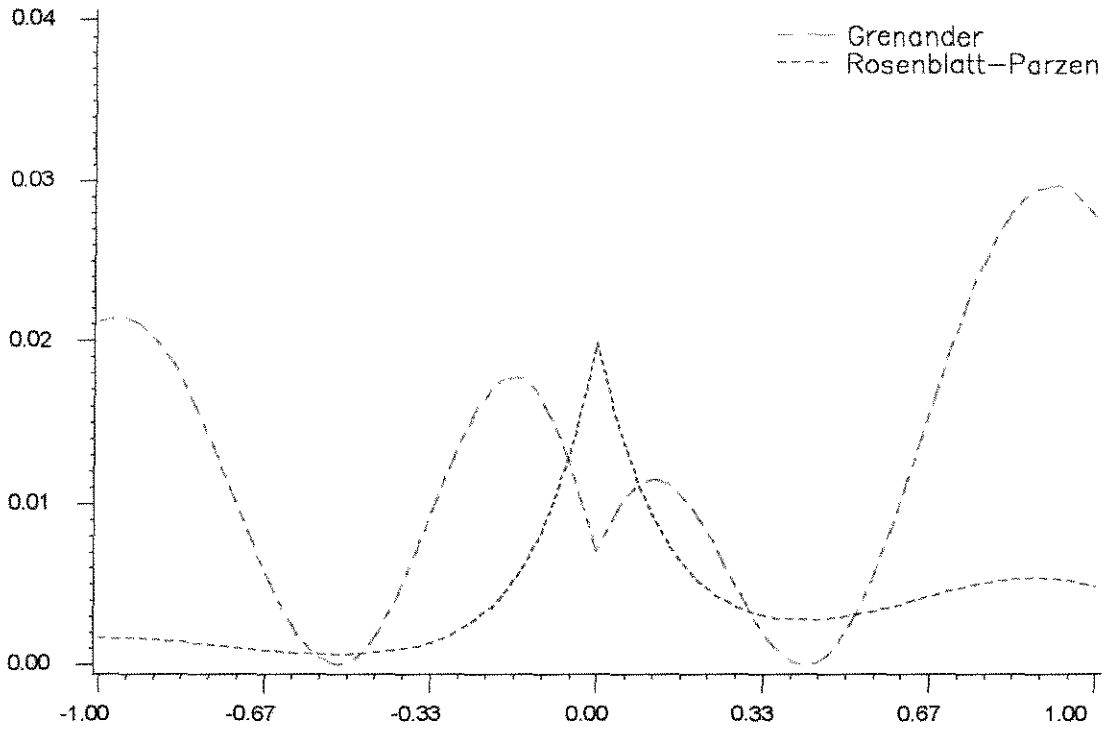


Gráfico 5.40 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Triangular, n=50.

Densidade Triangular, n=100

Rosenblatt-Parzen	Grenander
$EQMI$ =0.4110624	$EQMI$ =0.0.5040943
$\hat{h}$=0.39587	$\hat{\sigma}^2 = 0.077795$
	$\hat{p}_1 = 0.51128$ $\hat{\mu}_1 = -0.33869$
	$\hat{p}_2 = 0.11805$ $\hat{\mu}_2 = 0.07962$
	$\hat{p}_3 = 0.37067$ $\hat{\mu}_3 = 0.55779$

Tabela 5.22: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=100.

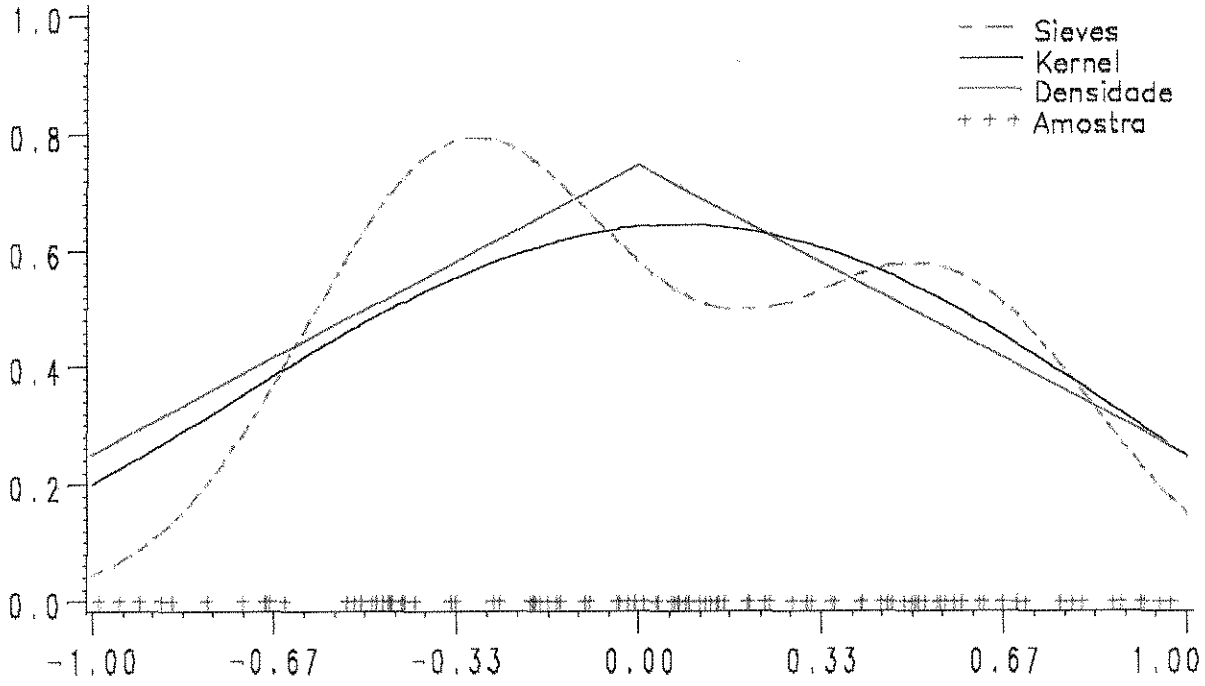


Gráfico 5.41 Função de densidade Triangular e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.22, tamanho de amostra n=100.

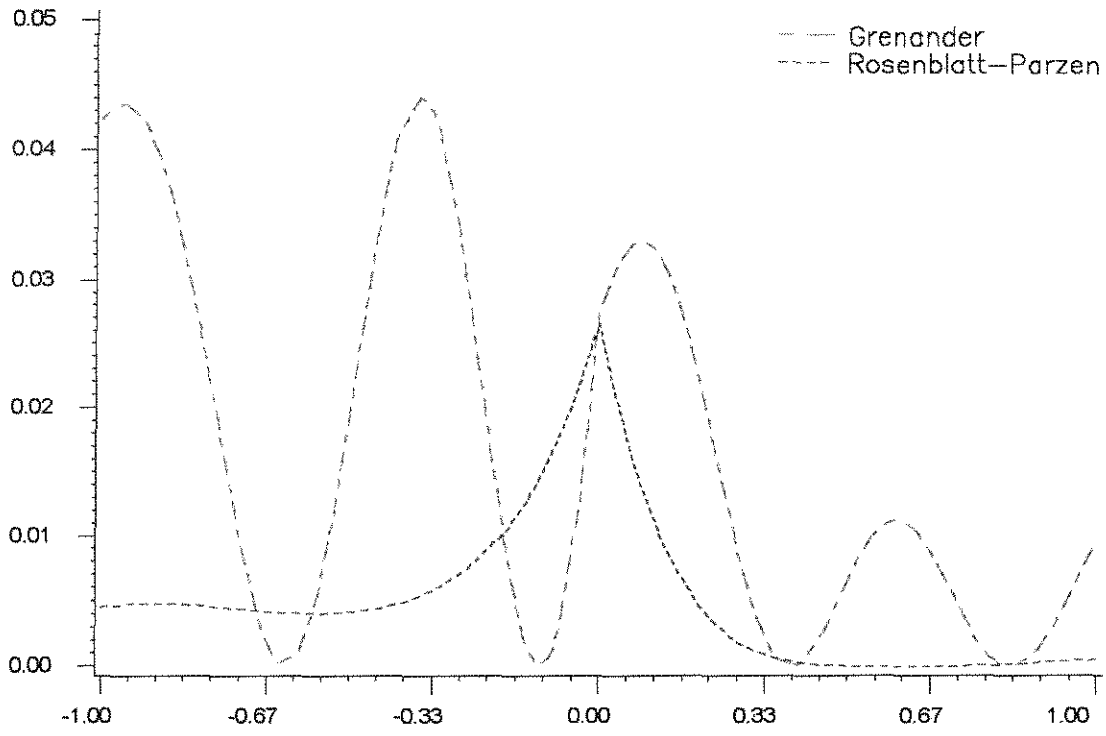


Gráfico 5.42 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Triangular, n=100.

5.3.8 Resultado das comparações da densidade Trimodal

Aqui foi utilizado o Método da Rejeição, este algoritmo consiste no seguinte: para obter uma amostra da variável aleatória X com função de densidade de probabilidade f, dispondo de um gerador da variável Y com função de densidade de probabilidade g e de um gerador da variável $U \sim U(0,1)$, procedemos como a seguir.

Quando $\frac{f(y)}{g(y)} \leq M$ para todo y em R,

1. Gerar um valor y de $U(-1,1)$,
2. Gerar um valor u de U,
3. Se $3u \leq cos(10y) + 2$, definir $x = y$, caso contrário voltar à 1.

Este algoritmo é estudado com maiores detalhes em Bustos & Frery(1992)

Rosenblatt-Parzen	**Grenander**
$EQMI$ =0.1040595	$EQMI$ =0.2574167
$\hat{h}$=0.08152	$\hat{\sigma}^2 = 0.0096161$
	$\hat{p}_1 = 0.44$ $\quad \hat{\mu}_1 = -0.68086$
	$\hat{p}_2 = 0.28$ $\quad \hat{\mu}_2 = -0.03993$
	$\hat{p}_3 = 0.28$ $\quad \hat{\mu}_3 = 0.61846$

Tabela 5.23: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=25.

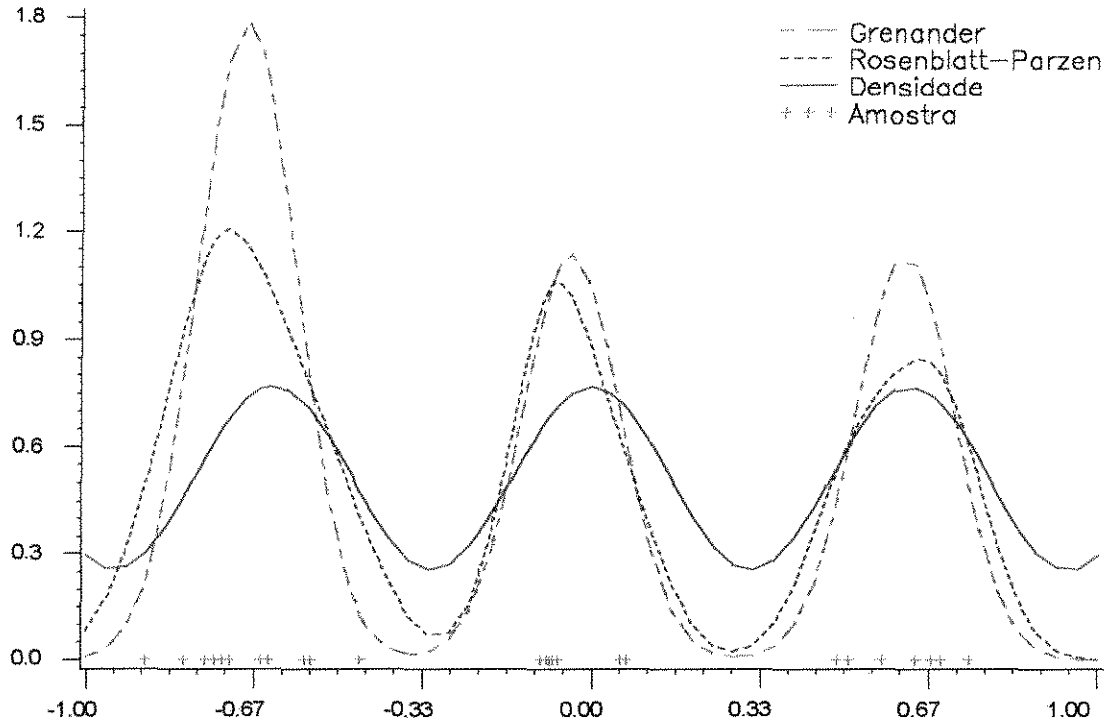


Gráfico 5.43 Função de densidade Trimodal e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.23, tamanho de amostra n=25.

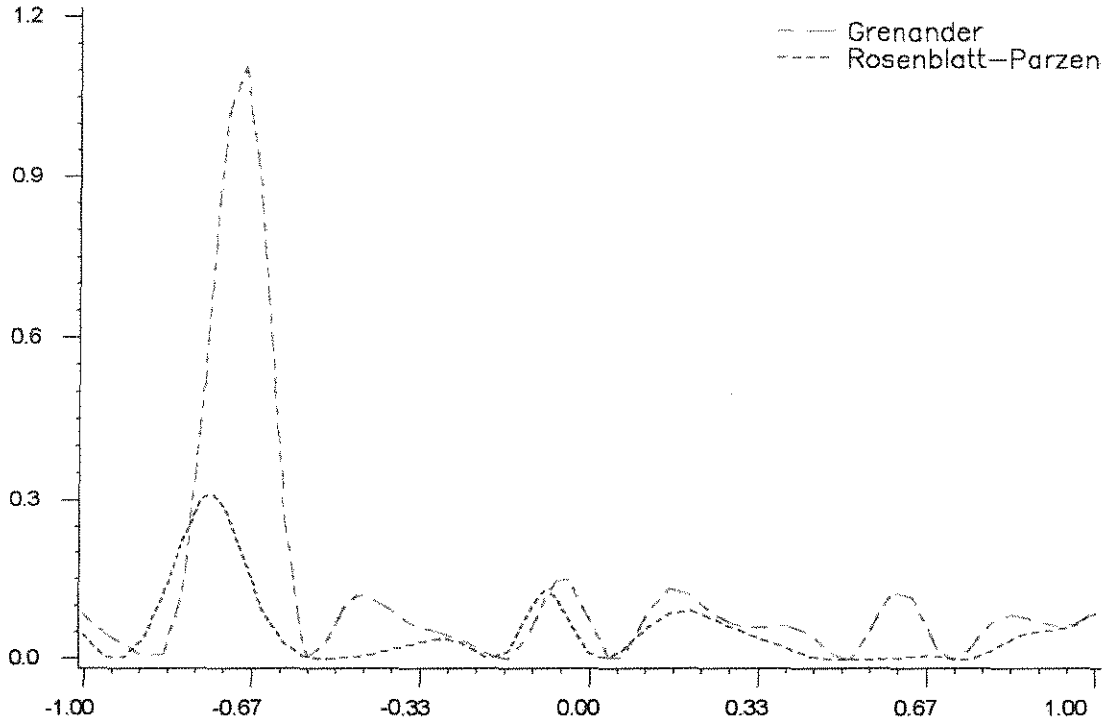


Gráfico 5.44 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Trimodal, n=25.

Densidade Trimodal, n=50

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0609583	$EQMI$ =0.1243947
$\hat{h}$ =0.072791	$\hat{\sigma}^2 = 0.013349$
	$\hat{p}_1 = 0.40003$ $\hat{\mu}_1 = -0.65995$
	$\hat{p}_2 = 0.26009$ $\hat{\mu}_2 = -0.02967$
	$\hat{p}_3 = 0.33988$ $\hat{\mu}_3 = 0.62225$

Tabela 5.24: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=50.

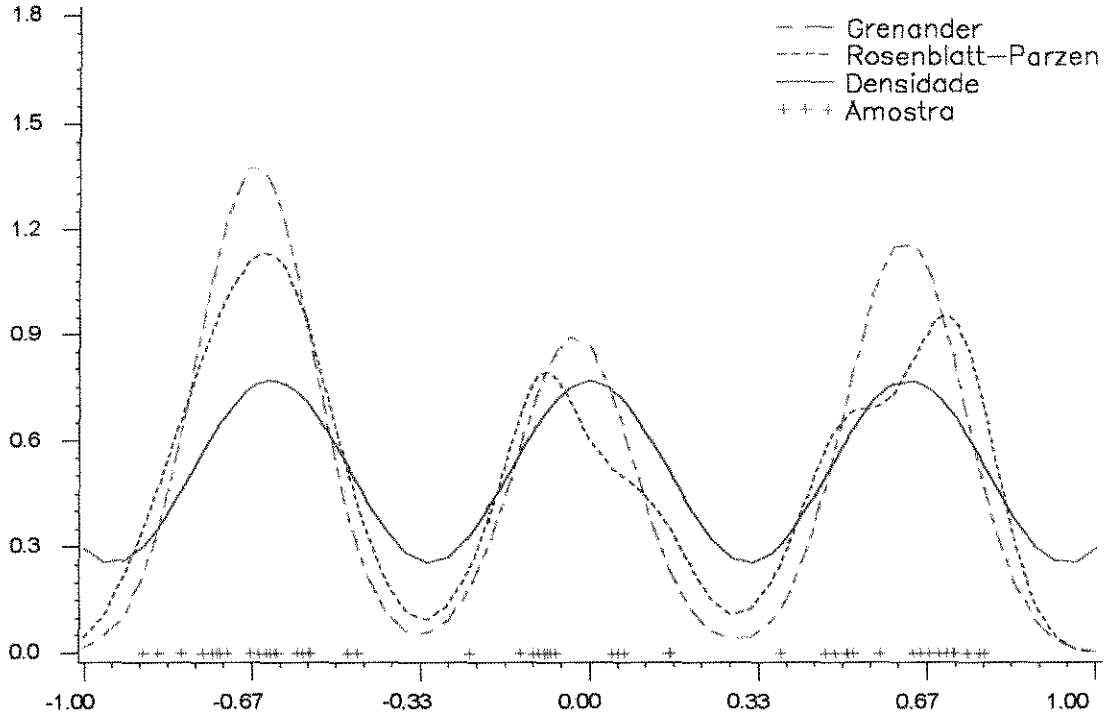


Gráfico 5.45 Função de densidade Trimodal e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.24, tamanho de amostra n=50.

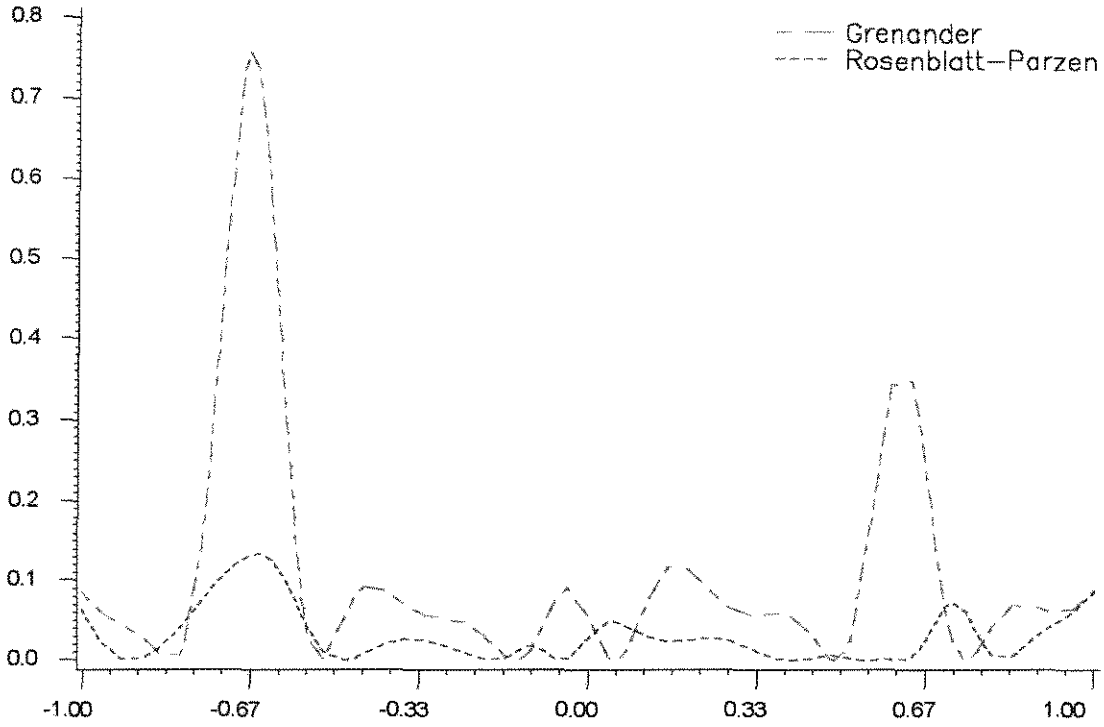


Gráfico 5.46 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Trimodal, n=50.

Densidade Trimodal, n=100

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0333733	$EQMI$ =0.042146
$\hat{h}$=0.095189	$\hat{\sigma}^2 = 0.021208$
	$\hat{p}_1 = 0.40162$ $\hat{\mu}_1 = -0.63465$
	$\hat{p}_2 = 0.29250$ $\hat{\mu}_2 = 0.05667$
	$\hat{p}_3 = 0.30588$ $\hat{\mu}_3 = 0.63482$

Tabela 5.25: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=100.

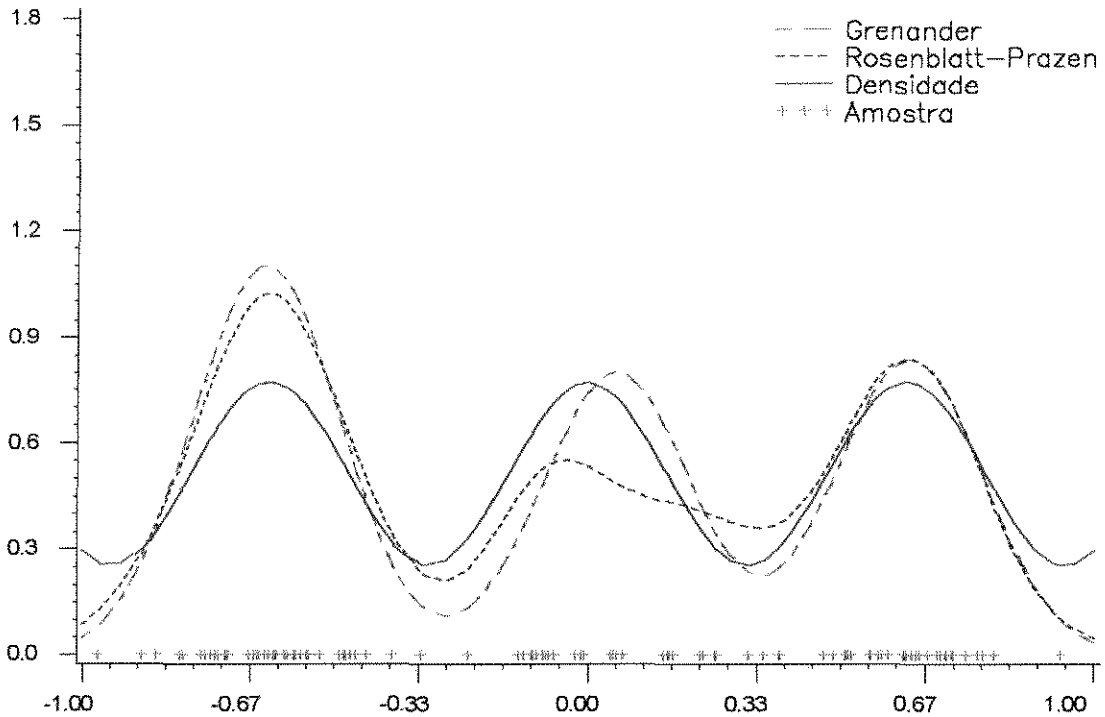


Gráfico 5.47 Função de densidade Trimodal e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.25, tamanho de amostra n=100.

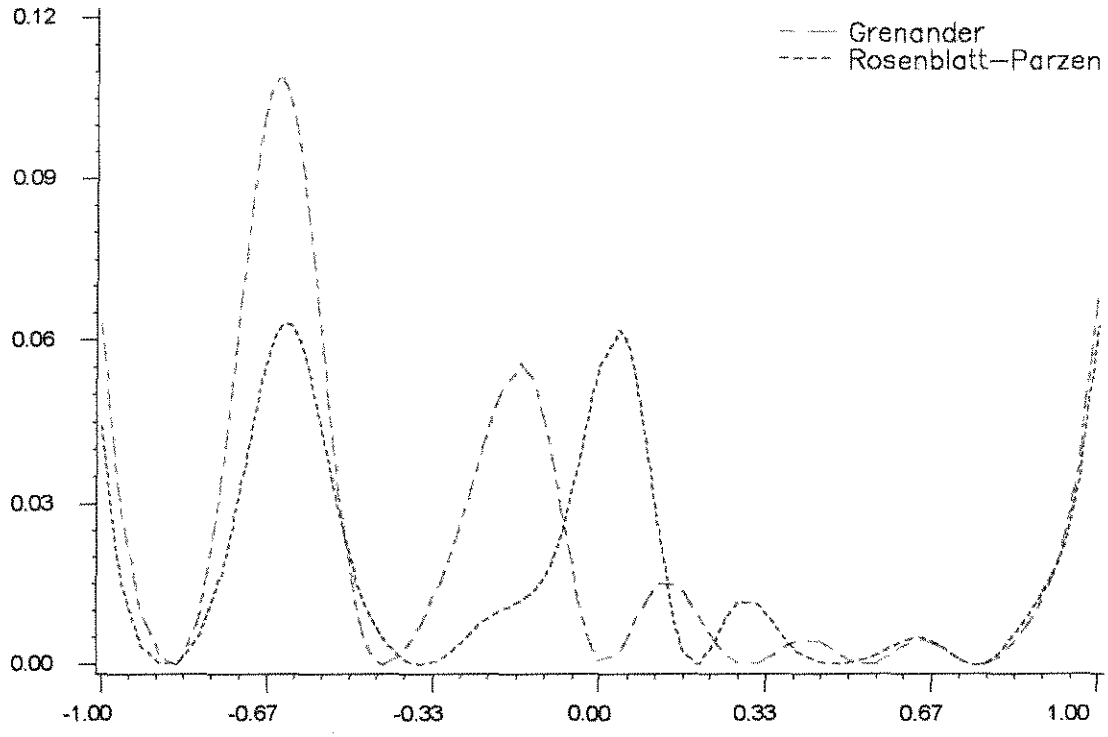


Gráfico 5.48 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Trimodal, n=100.

5.3.9 Resultado das comparações na densidade Uniforme Escada

A chamada Transformação Integral ou Método de Inversão foi o utilizado para gerar as diferentes amostras. Consiste no seguinte, primeiramente geramos a mostra $U_1,\ldots,U_n \sim U(0,1)$ através da função RANUNI ou UNIFORM, e posteriormente calculamos

$$X_i = \begin{cases} \frac{3}{2}U_i - 1 & \text{se } U_i \leq \frac{2}{3} \\ 3U_i - 2 & \text{se } U_i > \frac{2}{3} \end{cases}, \text{ para } i = 1,\ldots,n.$$

Rosenblatt-Parzen	**Grenander**
$EQMI$ =0.0995918	$EQMI$ =0.1596896
$\hat{h}$ =0.23714	$\hat{\sigma}^2$ = 0.039744
	$\hat{p}_1$ = 0.42332 $\quad$ $\hat{\mu}_1$ = -0.60019
	$\hat{p}_2$ = 0.13683 $\quad$ $\hat{\mu}_2$ = -0.06342
	$\hat{p}_3$ = 0.43986 $\quad$ $\hat{\mu}_3$ = 0.57881

Tabela 5.26: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=25.

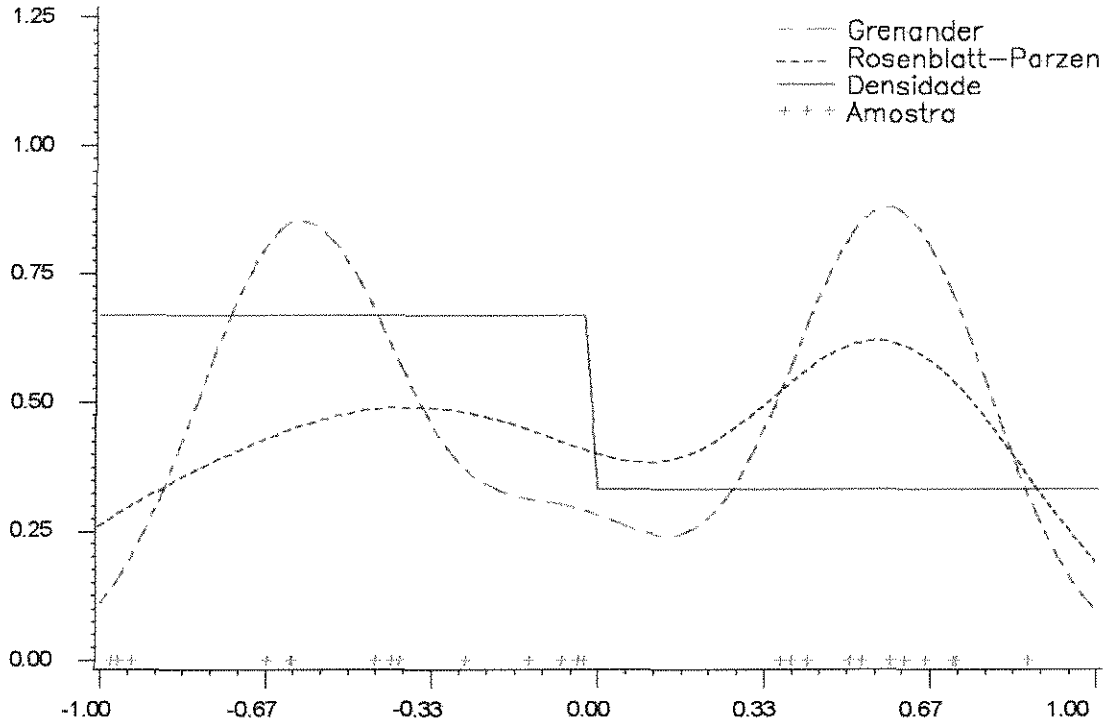


Gráfico 5.49 Função de densidade Uniforme Escada e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.26, tamanho de amostra n=25.

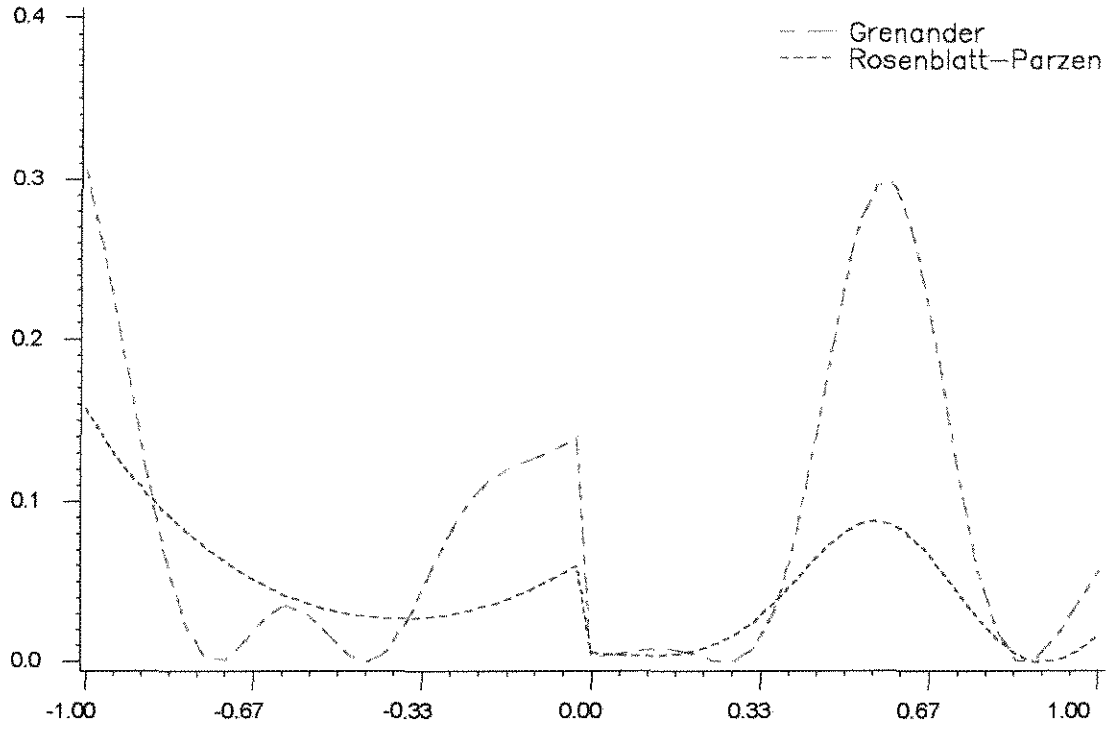


Gráfico 5.50 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Uniform Escada, n=25.

Densidade Uniforme Escada, n=50

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0633534	$EQMI$ =0.1267242
$\hat{h}$ =0.079622	$\hat{\sigma}^2 = 0.042763$
	$\hat{p}_1 = 0.13206$ $\quad \hat{\mu}_1 = -0.95318$
	$\hat{p}_2 = 0.58937$ $\quad \hat{\mu}_2 = -0.38917$
	$\hat{p}_3 = 0.27858$ $\quad \hat{\mu}_3 = 0.60424$

Tabela 5.27: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=50.

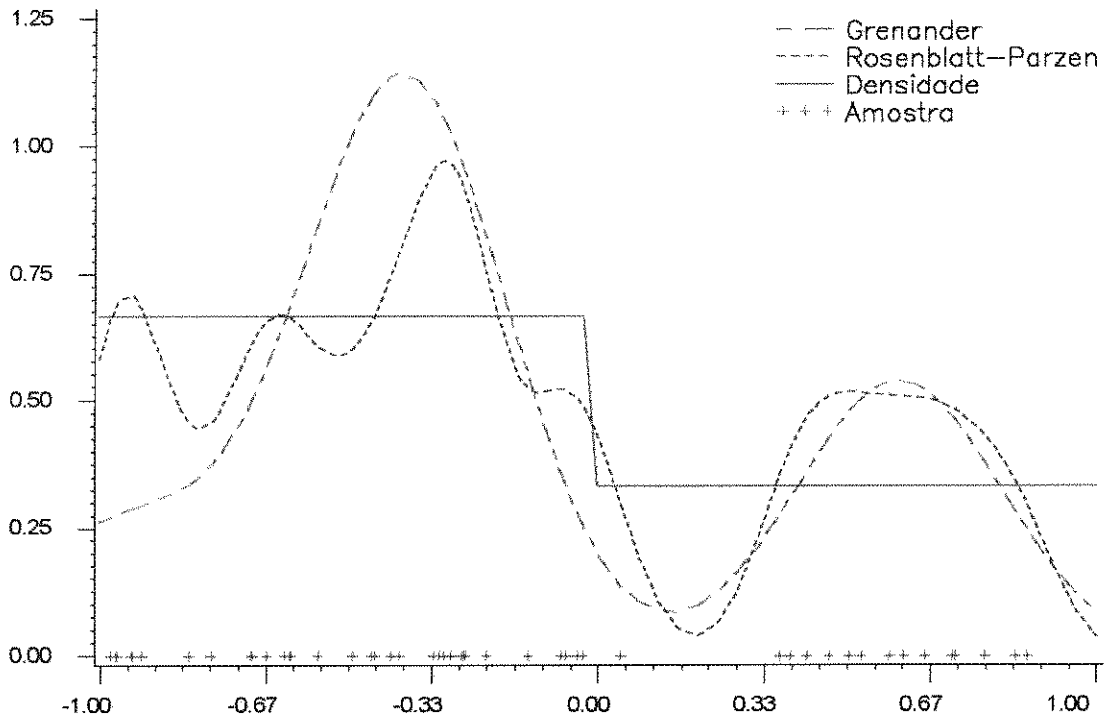


Gráfico 5.51 Função de densidade Uniforme Escada e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.27, tamanho de amostra n=50.

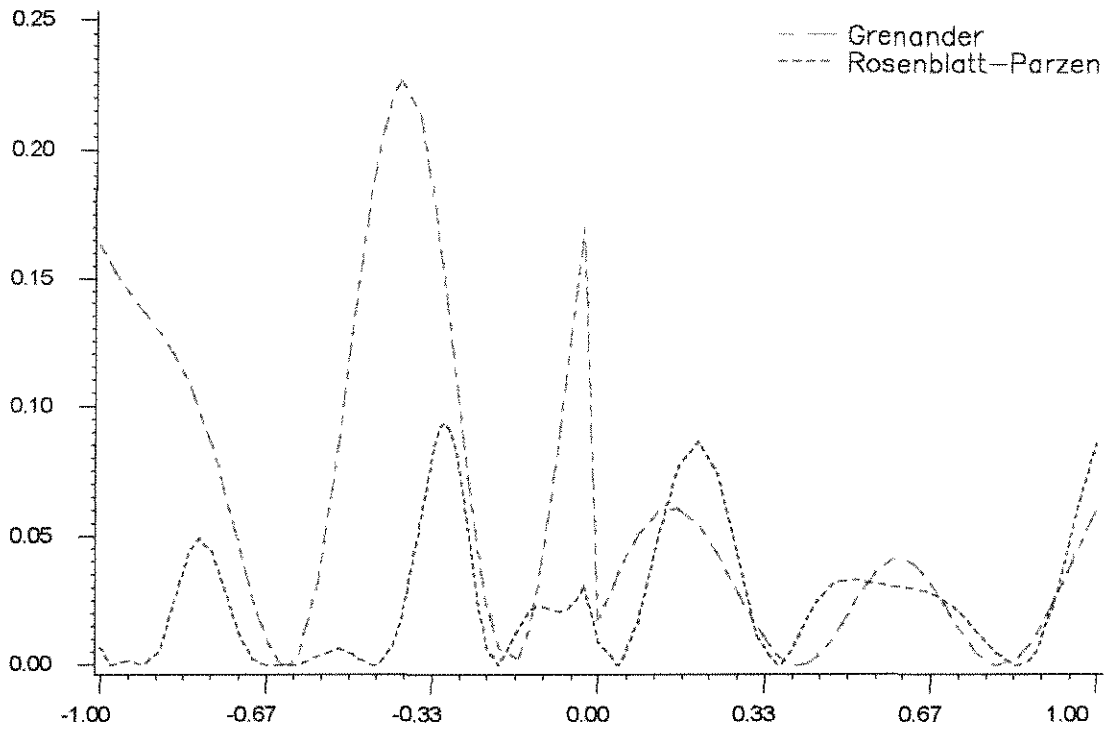


Gráfico 5.52 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Uniform Escada, n=50.

Densidade Uniforme Escada, n=100

Rosenblatt-Parzen	Grenander
$EQMI$ =0.0270741	$EQMI$ =0.072475
$\hat{h}$=0.1546	$\hat{\sigma}^2 = 0.05251$
	$\hat{p}_1 = 0.17868$ $\hat{\mu}_1 = -0.8919$
	$\hat{p}_2 = 0.57474$ $\hat{\mu}_2 = 0.33935$
	$\hat{p}_3 = 0.24658$ $\hat{\mu}_3 = 0.64445$

Tabela 5.28: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=100.

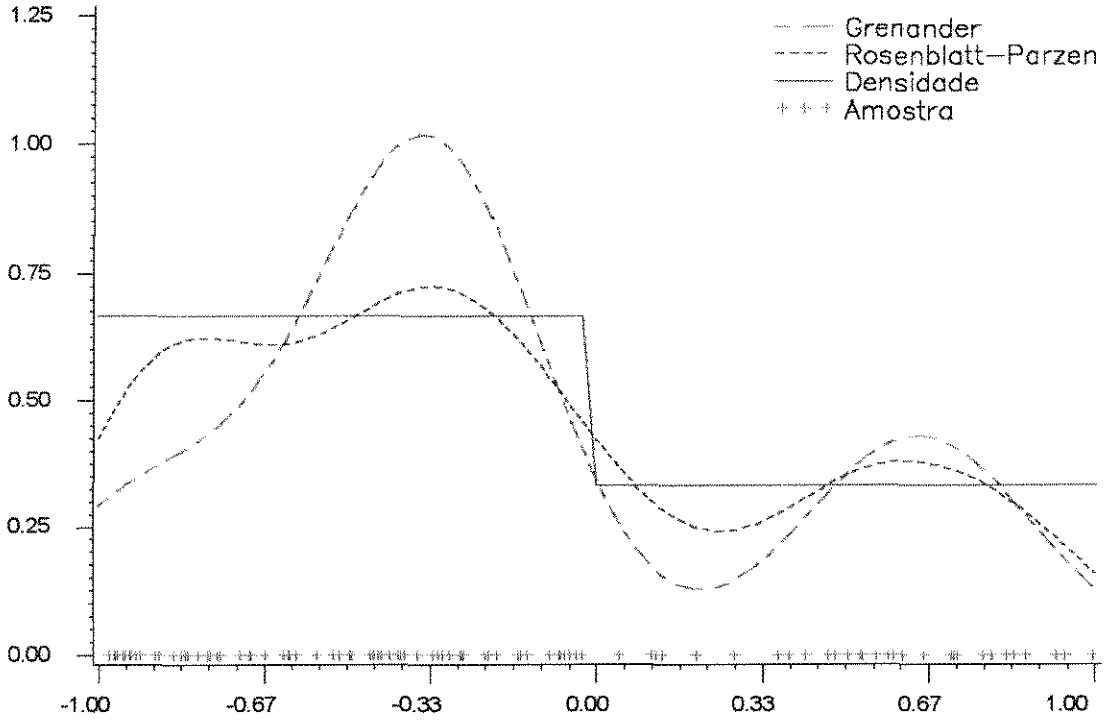


Gráfico 5.53 Função de densidade Uniforme Escada e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.28, tamanho de amostra n=100.

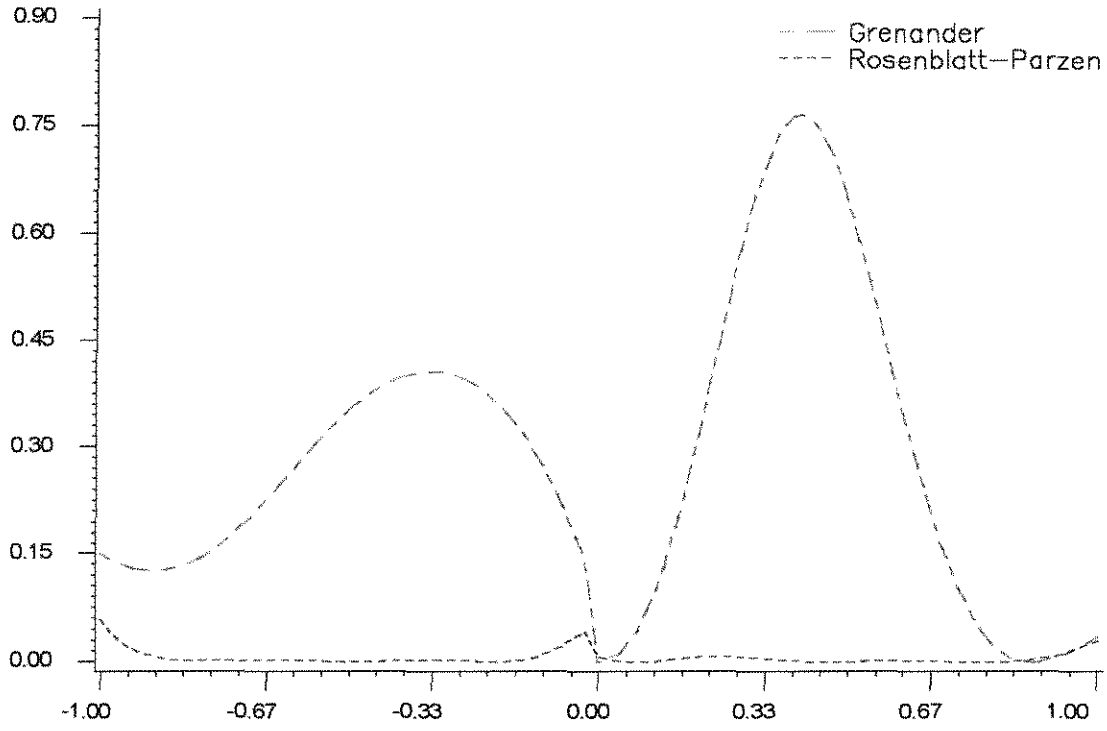


Gráfico 5.54 Comportamento pontual da diferença $\left(\hat{f}(x)-f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade Uniform Escada, n=100.

5.3.10 Resultado das comparações na densidade 2ª Mistura de Normais

Para esta situação utilizou-se o chamado Método de Composição que consiste em escolher com uma determinada probabilidade uma das densidades na mistura e logo simular um valor da variável correspondente.

Rosenblatt-Parzen	Grenander
EQM =0.0671955	$EQMI$ =0.0999838
$\hat{h}$=0.84009	$\hat{\sigma}^2 = 0.84484$
	$\hat{p}_1 = 0.5193$ $\hat{\mu}_1 = -1.87862$
	$\hat{p}_2 = 0.4807$ $\hat{\mu}_2 = 3.35092$

Tabela 5.29: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=25.

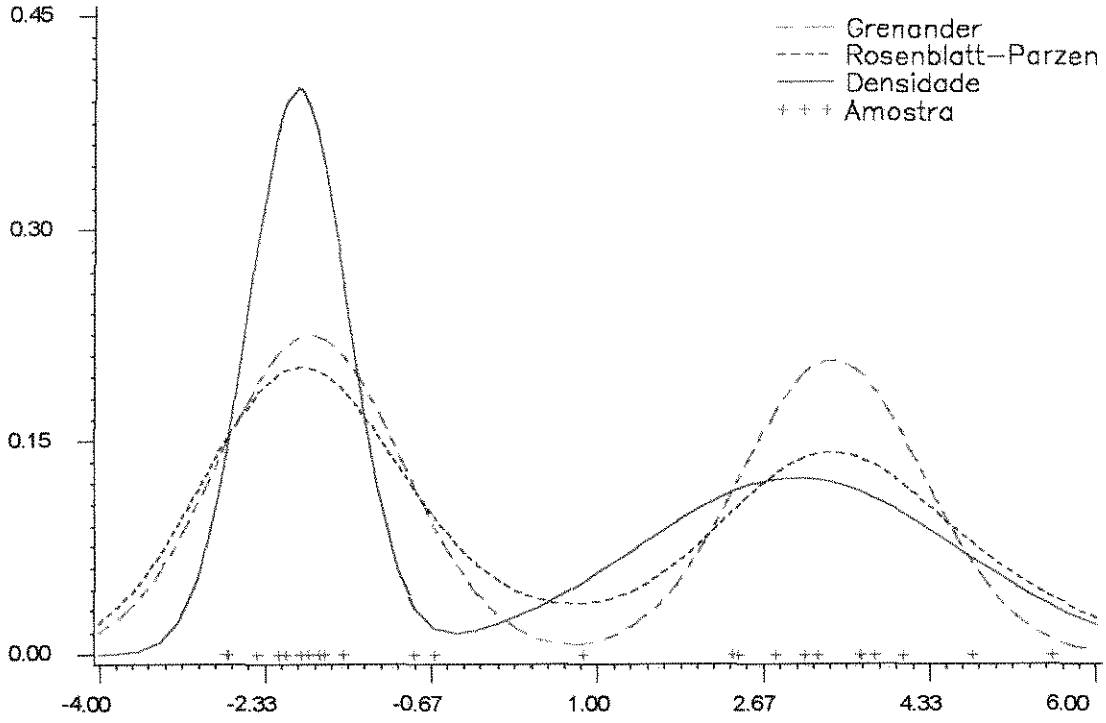


Gráfico 5.55 Função de densidade 2ª Mistura de Normais e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.29, tamanho de amostra n=25.

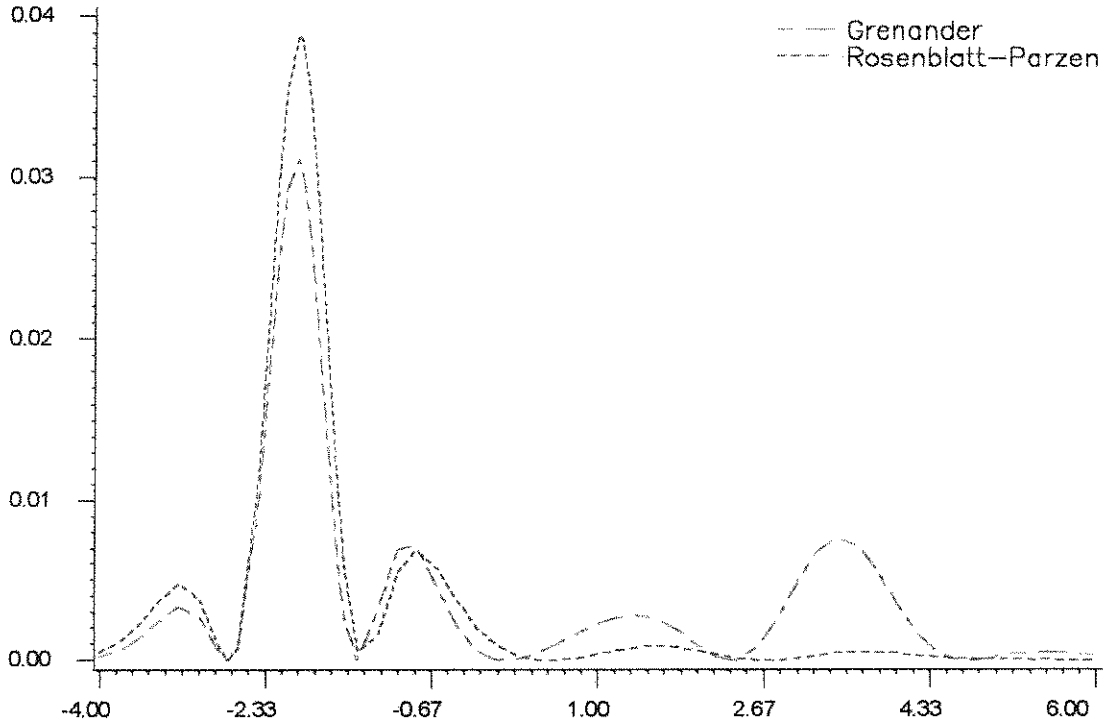


Gráfico 5.56 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade 2ª Mistura de Normais, n=25.

Densidade 2ª Mistura de Normais, n=50

Rosenblatt-Parzen	Grenander
EQM =0.0527703	$EQMI$ =0.0623478
$\hat{h}$ =0.68942	$\hat{\sigma}^2 = 1.22817$
	$\hat{p}_1 = 0.51906$ $\hat{\mu}_1 = -1.92031$
	$\hat{p}_2 = 0.48094$ $\hat{\mu}_2 = 3.00721$

Tabela 5.30: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=50.

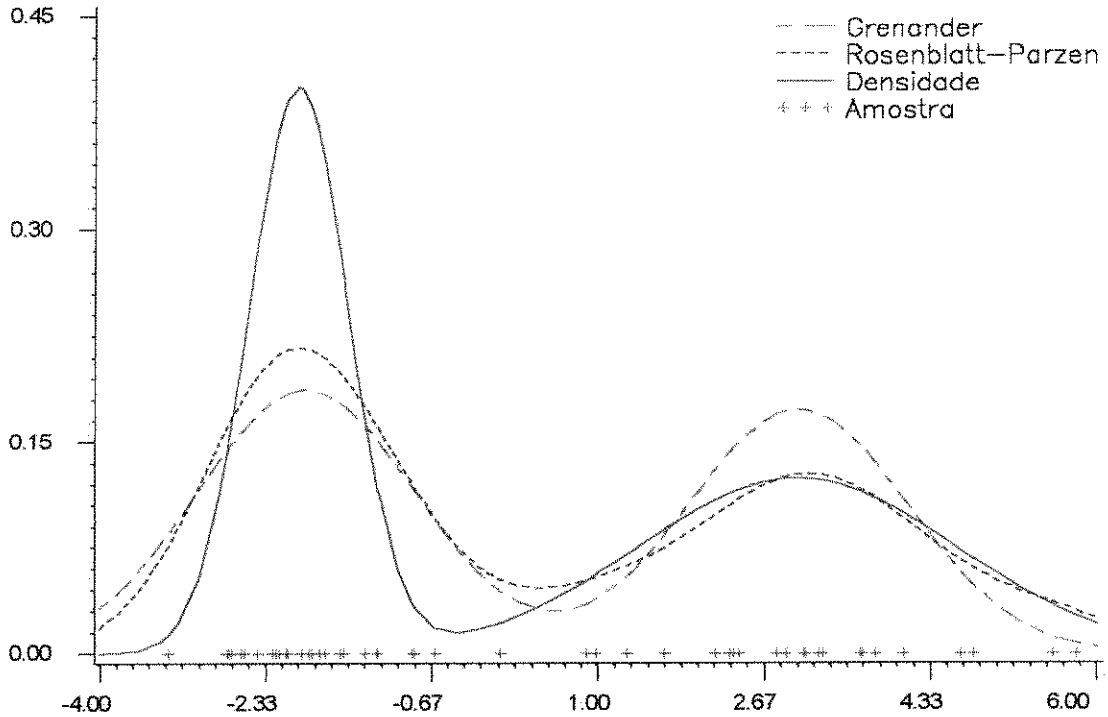


Gráfico 5.57 Função de densidade 2ª Mistura de Normais e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.30, tamanho de amostra n=50.

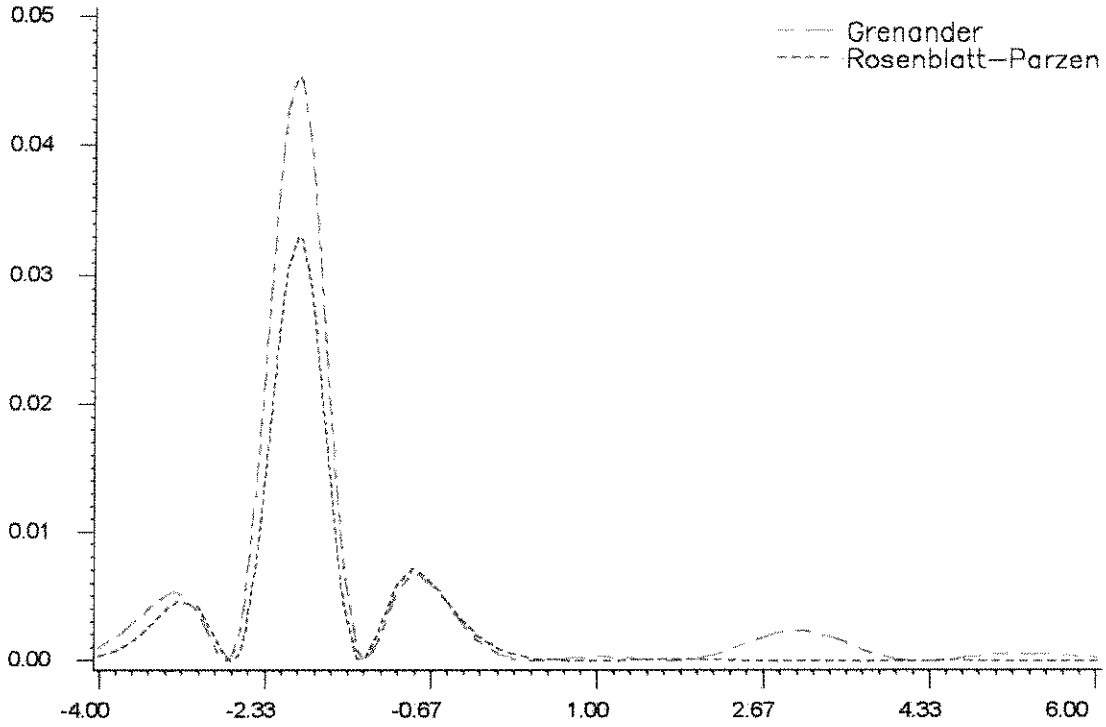


Gráfico 5.58 Comportamento pontual da diferença $\left(\hat{f}(x) - f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade 2ª Mistura de Normais, n=50.

Densidade 2ª Mistura de Normais, n=100

Rosenblatt-Parzen	Grenander
EQM =0.0644018	$EQMI$ =0.0620704
$\hat{h}$ =0.41435	$\hat{\sigma}^2$ = 0.84483
	$\hat{p}_1$ = 0.52266 $\quad$ $\hat{\mu}_1$ = -1.98233
	$\hat{p}_2$ = 0.45116 $\quad$ $\hat{\mu}_2$ = 2.69494
	$\hat{p}_3$ = 0.03619 $\quad$ $\hat{\mu}_3$ = 5.75035

Tabela 5.31: Valores obtidos do $EQMI$ e dos estimadores dos parâmetros. Tamanho de amostra n=100.

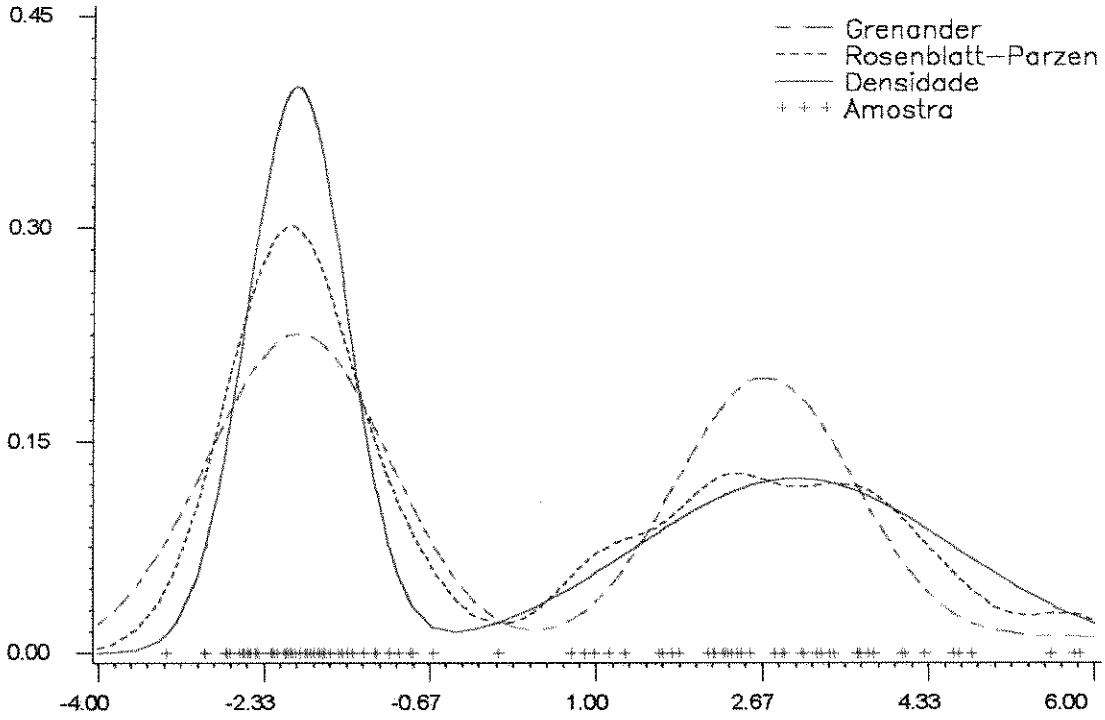


Gráfico 5.59 Função de densidade 2ª Mistura de Normais e os estimadores Grenander e Rosenblatt-Parzen segundo as estimativas dos parâmetros apresentadas na Tabela 5.31, tamanho de amostra n=100.

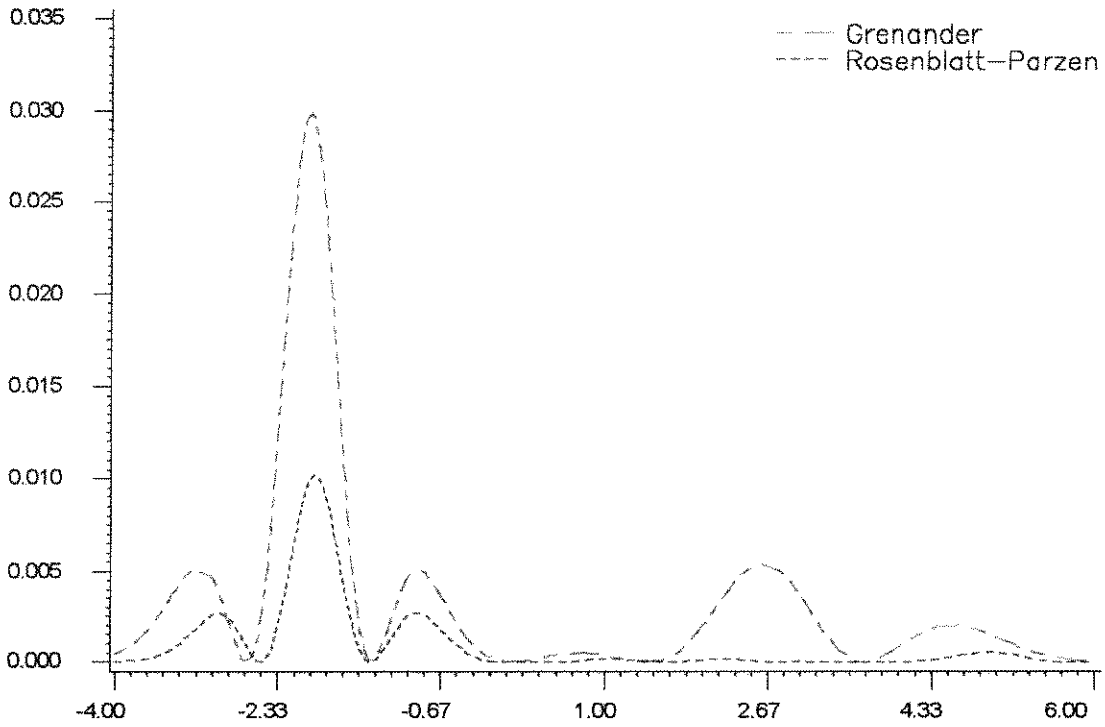


Gráfico 5.60 Comportamento pontual da diferença $\left(\hat{f}(x)-f(x)\right)^2$ para os estimadores de Rosenblatt-Parzen e Grenander da função de densidade 2.ª Mistura de Normais, n=100.

5.4 Conclusões

Estudando o estimador de Grenander de convolução, encontramos que Walter & Blum(1984) observaram que o problema de maximização nesta situação é similar á achar os estimadores num determinado modelo de mistura finita de densidades, para o qual utiliza-se o algoritmo EM. Devido a não termos achado trabalhos justificando esta afirmação, decidimos realizar este estudo.

No Capítulo III, mostramos que o estimador de Grenander de convolução pode ser visto como um modelo de mistura finita de densidades em situações mais gerais a aquela provada por Geman & Hwang(1982).

Naturalmente surge depois o problema de estimar os parâmetros destes modelos quando é reduzido o estimador de Grenander de convolução a uma mistura finita de densidades especifica. No Capítulo IV, provamos que o algoritmo EM pode ser utilizado com este objetivo.

Para termos idéia do comportamento prático desta forma de obter o estimador de Grenander de convolução, realizamos comparações baseados em dados simulados. Nesta simulações chegamos a conclusão que o comportamento obtido pelo estimador de Grenander de convolução é similar ao apresentado pelo estimador de Rosenblatt-Parzen.

Apêndice I

Derivada de Gateaux

No Capítulo III, na demonstração da Proposição 3.2, utiliza-se a derivada de Gateaux, a seguir inclua-se um resumo deste conceito tendo como referência Luenberger(1969).

Assumamos que A é um espaço vetorial, B um espaço normado e T uma transformação definida no domínio $D \subset A$ e tendo valores em $R \subset B$.

Definição I.1: Seja $x \in D \subset A$ e seja h um elemento arbitrário de A. Se o limite

$$\delta T(x;h) = \lim_{\alpha \to 0} \frac{1}{\alpha}[T(x+\alpha h) - T(x)] \qquad \text{(I.1)}$$

existe, é chamado de *diferencial Gateaux de T em x com incremento h*. Se o limite (I.1) existir para cada $h \in A$, a transformação T se diz que é *diferenciável Gateaux em x.*

O limite em (I.1) só terá sentido considerando $x+\alpha h \in D$ para α suficientemente pequena. A aplicação no Capítulo III deste conceito é no caso onde B é a reta real, tendo que a transformação T se reduz a um funcional real em A. Se L for um funcional real em A, o diferencial de Gateaux de L, se existir, é

$$\delta L(x;h) = \frac{d}{d\alpha} L(x+\alpha h)\Big|_{\alpha=0} . \qquad \text{(I.2)}$$

Exemplo I.1: Seja $A = C[0,1]$, ou seja, o conjunto das funções contínuas no intervalo [0,1] e seja $L(x) = \int_0^1 g(x(t),t)dt$, assumiremos que a derivada de g existe e é contínua com respeito a x e t. Temos então

$$\delta L(x;h) = \frac{d}{d\alpha} \int_0^1 g(x(t)+\alpha h(t),t)dt \Big|_{\alpha=0} .$$

Intercâmbiando as operações de diferenciação e integração, que é permitido dadas as condições da função g, temos

$$\delta L(x;h) = \int_0^1 \frac{\partial}{\partial x} g(x,t)h(t)dt . \qquad (I.3)$$

O diferencial de Gateaux generaliza o conceito de derivada direcional e sua existência é uma exigência débil, dado que a definição não exige ter definida norma em A. É relativamente simples aplicar à derivada de Gateaux nos problemas de otimização de funcionais em espaços lineares.

Definição I.2: Seja L um funcional real definido no subconjunto Ω do espaço normado A. O ponto $x_0 \in \Omega$ se diz ser *máximo relativo* de L em Ω se existir um aberto N contendo x_0 tal que $L(x_0) \geq L(x)$ para todo $x \in \Omega \cap N$.

Será usado o termo *extremo* referido ao máximo ou mínimo sobre qualquer conjunto. O Teorema a seguir obtêm a condição necessária de extremo. A pesar da simplicidade deste resultado é de grande utilidade, a demonstração acha-se em Luenberger(1969) assim como algumas generalizações e variados exemplos.

Teorema I.1: Seja L um funcional real diferenciável Gateaux no espaço vetorial A. A condição necessária para L ter extremo no ponto $x_0 \in A$ é que a derivada de Gateaux do funcional L se anula em x_0, isto é $\delta L(x_0;h) = 0$ para todo $h \in A$.

Apêndice II:

Sistema de Tchebycheff *(T-system)*

Os *sistemas de Tchebycheff* (abreviadamente *T-system*) consistem em conjuntos de funções contínuas $\{u_i(t)\}_{i=0,\ldots,n}$ definidas em intervalos reais $[a,b]$ e caracterizadas pela propriedade de que toda combinação linear real não trivial $\sum_{i=1}^{n} a_i u_i(t)$ possui no máximo n zeros distintos em $[a,b]$; a combinação linear anterior se diz real não trivial se os a_i são números reais com a propriedade $\sum_{i=1}^{n} a_i^2 > 0$.

Estes sistemas são utilizados (as vezes indiretamente) na teoria de aproximações, onde as aplicações são os métodos de interpolação, as fórmulas de quadratura e outros.

O exemplo clássico destes sistemas é o *T-system* $\{1, t, t^2, \ldots, t^n\}$, as propriedades deste sistema e suas ramificações se podem estender ao caso mais geral dos *T-system*. Por exemplo, resultados na teoria de aproximações, desigualdades e propriedades de oscilação dos polinômios ordinários podem-se estender aos *polinômios generalizados* formados por *T-system*. Os polinômios generalizados são entendidos como expressões da forma $u(t) = \sum_{i=1}^{n} a_i u_i(t)$, onde os a_i são constantes reais e $\{u_i(t)\}_{i=0,\ldots,n}$ é um *T-system*. Muitas das propriedades dos *sistemas de Tchebycheff* constituem generalizações das propriedades de oscilação e aproximação dos polinômios reais.

Um estudo amplo destes sistemas pode ser encontrado em Karlin & Studden(1966), também estuda-se em Lorentz(1966); aplicações em diferentes áreas acha-se em Anderson *et ali*(1989) e aplicação específica em probabilidade em Johnson *et ali*(1993). A seguir apresentamos definições e resultados importantes que nos ajudarão a compreender estes sistemas de funções e que serão de grande utilidade.

Definição II.1: Sejam $u_0(t), \ldots, u_n(t)$ funções reais continuas definidas no intervalo $[a,b]$. Diz se que estas funções formam um *sistema de Tchebycheff*, abreviadamente *T-system* se o determinante de ordem n+1

$$U\begin{pmatrix} 0,1,\ldots,n \\ t_0,t_1,\ldots,t_n \end{pmatrix} = det\left\| u_i(t_j) \right\|_{i,j=0,\ldots,n} \qquad \text{(II.1)}$$

for estritamente positivo qualquer sejam $a \le t_0 < t_1 < \cdots < t_n \le b$, onde

$$det\left\|u_i(t_j)\right\|_{i,j=0,\ldots,n} = \begin{vmatrix} u_0(t_0) & u_0(t_1) & \cdots & u_0(t_n) \\ u_1(t_0) & u_1(t_1) & \cdots & u_1(t_n) \\ \vdots & \vdots & & \vdots \\ u_n(t_0) & u_n(t_1) & \cdots & u_n(t_n) \end{vmatrix}. \qquad \text{(II.2)}$$

A partir desta definição prova-se o Teorema II.1, que será apresentado posteriormente, nele se justifica que todo polinômio generalizado possui no máximo n zeros.

Definição II.2: As funções $u_o(t),\ldots,u_n(t)$ formam um *sistema estendido de Tchebycheff*, ou *CT-system* se $\{u_o(t),\ldots,u_r(t)\}$ for um *T-system* para cada r =0,1, ... ,n.

O exemplo clássico de *CT-system* é o sistema de funções já referido $u_i(t) = t^i$, i =0,1, ... ,n, definido em qualquer intervalo $[a,b]$. Neste casso o determinante (II.1) se reduz ao determinante de Vandermonde com valor $\prod_{0\le i<j\le n}(t_j - t_i)$.

É importante apreciar que os *T-system* são casos especiais de funções núcleo $K(s,t)$ de duas variáveis possuindo certas propriedades de regularidade de sinal. Consideremos a função real $K(s,t)$ definida em $(s,t)\in S\text{x}T$, onde S e T são subconjuntos ordenados da reta real. No casso quando S é o conjunto finito descrito como {0,1, ... ,n}, o núcleo $K(s,t)$ se reduz a seqüência de funções $K(i,t) = u_i(t)$, i =0,1, ... ,n.

Para uma função núcleo geral o determinante correspondente a (II.1) será definido como

$$K\begin{pmatrix} s_0,s_1,\ldots,s_n \\ t_0,t_1,\ldots,t_n \end{pmatrix} = det\left\|K(s_i,t_j)\right\|_{i,j=0,\ldots,n} \qquad \text{(II.3)}$$

onde $s_0 < s_1 < \cdots < s_n$, $t_0 < t_1 < \cdots < t_n$ e $(s_i,t_j) \in S\text{x}T$.

Definição II.3: O núcleo $K(s,t)$ é chamado *totalmente positivo de ordem k* ou TP_k se para cada r =1,... ,k temos

$$K\begin{pmatrix} s_1,\ldots,s_r \\ t_1,\ldots,t_r \end{pmatrix} \ge 0 \qquad \text{(II.4)}$$

qualquer sejam $s_1 < s_2 < \cdots < s_r$, $t_1 < t_2 < \cdots < t_r$ e $(s_i,t_j) \in S\text{x}T$. Se o determinante (II.4) é estritamente positivo então $K(s,t)$ é chamado de *estritamente totalmente positivo de ordem k* ou

STP_k.

Se a função $K(s,t)$ é suficientemente diferenciável em t, para $s \in S$, e existir valores de t_i iguais, se substituirão as colunas sucessivas, correspondentes aos t_i iguais, pelas suas sucessivas derivadas, isto é, se $s_0 < s_2 < \cdots < s_k$, $a \le t_0 \le t_1 \le \cdots \le t_k \le b$ e $t_{i-1} < t_i = t_{i+1} = \cdots = t_{i+q}$, onde $0 \le q \le p-1$, temos que o determinante $K^*\begin{pmatrix} s_1,\ldots,s_k \\ t_1,\ldots,t_k \end{pmatrix}$ será definido como

$$\begin{vmatrix} K(s_1,t_1) & \cdots & K(s_1,t_i) & \frac{\partial}{\partial_i}K(s_1,t_i) & \cdots & \frac{\partial^q}{\partial_i^q}K(s_1,t_i) & K(s_1,t_{i+q+1}) & \cdots & K(s_1,t_k) \\ K(s_2,t_1) & \cdots & K(s_2,t_i) & \frac{\partial}{\partial_i}K(s_2,t_i) & \cdots & \frac{\partial^q}{\partial_i^q}K(s_2,t_i) & K(s_2,t_{i+q+1}) & \cdots & K(s_2,t_k) \\ \vdots & & \vdots & \vdots & & \vdots & & & \vdots \\ K(s_k,t_1) & \cdots & K(s_k,t_i) & \frac{\partial}{\partial_i}K(s_k,t_i) & \cdots & \frac{\partial^q}{\partial_i^q}K(s_k,t_i) & K(s_k,t_{i+q+1}) & \cdots & K(s_k,t_k) \end{vmatrix}.$$

Pôr exemplo, $s = s_0 = s_1 = \cdots = s_k$ e $t = t_0 = t_1 = \cdots = t_k$, $K(s,t)$ sendo suficientemente diferenciável, temos que

$$K^*\begin{pmatrix} s_1,\ldots,s_k \\ t_1,\ldots,t_k \end{pmatrix} = det\left\| \frac{\partial^{i+j}}{\partial s^i \partial t^j} K(s,t) \right\|_{i,j=0,\ldots,k}$$

seria o Wronskiano da função $K(s,t)$.

Definição II.4: Se a função $K(s,t)$ é suficientemente diferenciável com relação a t, se diz que a função $K(s,t)$ é *estendida totalmente positiva de ordem k com relação a t* ou $ETP_k(t)$, se qualquer seja $r = 1,\ldots,k$ temos

$$K^*\begin{pmatrix} s_1,\ldots,s_r \\ t_1,\ldots,t_r \end{pmatrix} > 0 \qquad \text{(II.5)}$$

qualquer sejam $s_1 < s_2 < \cdots < s_r$, $t_1 \le t_2 \le \cdots \le t_r$ e $(s_i, t_i) \in S \times T$. Definição análoga pode-se aplicar para $ETP_k(s)$.

No casso quando S é finito, escrevendo $u_i(t) = K(i,t)$, a definição II.4 aplicada ao sistema assim obtido será;

Definição II.5: As funções $u_o(t),\ldots,u_n(t)$ se diz formam um *sistema estendido de Tchebycheff de ordem p* ou *ET-system* de ordem p, se $u_i \in C^{p-1}[a,b], i = 0,1,\ldots,n$ e

$$U^*\begin{pmatrix} 0,1,\ldots,n \\ t_0,t_1,\ldots,t_r \end{pmatrix} > 0 \qquad \text{(II.6)}$$

para qualquer seja a escolha de $t_0 \le t_1 \le \cdots \le t_n, t_i \in [a,b]$.

Karlin & Studden(1966) provaram que se o sistema de funções $\{u_i(t)\}_{i=0,\ldots,n}$ é *T-system* e as $u_0(t),\ldots,u_n(t)$ são suficientemente diferenciáveis o determinante (II.6) satisfaz também a desigualdade

$$U^*\begin{pmatrix} 0,1,\ldots,n \\ t_0,t_1,\ldots,t_r \end{pmatrix} \ge 0$$

para $t_0 \le t_1 \le \cdots \le t_n, t_i \in [a,b]$. Este resultado é uma generalização do fato de que se a função $u(t)$ é não-decrescente e diferenciável, então a derivada $u'(t) \ge 0$. O exemplo clássico $u_i(t) = t^i$, $i = 0,1,\ldots,n$ é um *ET-system*.

Exemplo II.1: Funções Potência

Seja $\alpha_0, \alpha_1, \ldots, \alpha_n$ uma seqüência estritamente crescente de números reais, então o sistema $\{t^{\alpha_i}\}_{i=0,\ldots,n}$ é *T-system* em qualquer subintervalo fechado de $(0,\alpha)$. O sistema $\{t^{\alpha_i}\}_{i=0,\ldots,n}$ é *T-system* em $(0,\alpha)$ dado que $t^{\alpha} = exp(\alpha \log t)$ pata $t > 0$ e $det\left\|exp(x_i y_j)_{i,j=0,\ldots,n}\right\| > 0$ qualquer sejam $-\infty < x_0 < \ldots < x_n < \infty$ e $-\infty < y_0 < \ldots < y_n < \infty$. O fato de o determinante anterior ser positivo pode ser provado observando que as funções da forma

$$w_n(y) = \sum_{i=0}^{n} a_i e^{x_i y}, \text{ sendo os } a_i \text{ reais } i = 0,\ldots,n \text{ e } \sum_{i=0}^{n} a_i^2 > 0$$

admitem no máximo n zeros reais distintos. Ver detalhes da prova em Karlin & Studden(1966), este *T-system* também é estudado em Rice(1964). Constitui um caso especial do seguinte resultado provado em Pólya & Szegö(1925).;

Proposição II.1: Os polinômios exponenciais $\sum_{i=1}^{n} q_i(y) e^{c_i y}$, onde $q_i(y)$ são polinômios de grau k_i, admitem no máximo $\left[\sum_{i=1}^{n}(k_i + 1)\right] - 1$ zeros contando as multiplicidades.

Exemplo II.2: Operações elementares que preservam *T-system*

(a) se $r(t)$ é uma função positiva e contínua em $[a,b]$ e $\{u_i\}_{i=0,\ldots,n}$ é um *T-system*, então as funções $v_i(t) = r(t)u_i(t)$, $i = 0,\ldots,n$ constituem um *T-system* pois

$$V\begin{pmatrix} 0,1,\ldots,n \\ t_0,t_1,\ldots,t_n \end{pmatrix} = \left(\prod_{i=0}^{n} r(t_i)\right) U\begin{pmatrix} 0,1,\ldots,n \\ t_0,t_1,\ldots,t_n \end{pmatrix}$$

(b) se $r(t)$ é estritamente crescente e contínua definida em $[c,d]$ com valores em $[a,b]$, então o sistema de funções $v_i(t) = u_i(r(t))$, $i = 0,\ldots,n$ é um *T-system* no intervalo $[c,d]$, se $\{u_i\}_{i=0,\ldots,n}$ for um *T-system* em $[a,b]$.

Exemplo II.3: *T-system* gerados por *TP* núcleos

(a) se $K(s,t)$ é um núcleo $STP_{n+1}(t)$ o sistema $\{K(s_i,t)\}_{i=0,\ldots,n}$ é um *CT-system* onde $s_0 < s_2 < \cdots < s_n$ e $K(s_i,t)$ é uma função contínua em t para cada s_i, $i = 0,\ldots,n$.

(b) se $K(s,t)$ é um núcleo $ETP_{n+1}(t)$ o sistema $\{K(s_i,t)\}_{i=0,\ldots,n}$ é um *ET-system*.

Os exemplos II.2 e II.3 fornecem uma grande variedade de *T-system* a partir de sistemas clássicos, como é o mencionado no exemplo II.1 e aquele mencionado no exemplo a seguir.

Exemplo II.4: O núcleo Cauchy

O núcleo $K(s,t)=\dfrac{1}{s+t}$ é *STP* de todas as ordens para $s>0$, $t>0$; neste caso temos que

$$K\begin{pmatrix} s_0,\ldots,s_n \\ t_0,\ldots,t_n \end{pmatrix} = \frac{\prod_{j>i}^{n}(s_j - s_i)(t_j - t_i)}{\prod_{i,j=0}^{n}(s_i + t_j)}.$$

O sistema $u_i(t)=\dfrac{1}{s_i + t}$ para $i = 0,\ldots,n$ onde $s_0 < s_2 < \cdots < s_n$ é *CT-system* em qualquer subintervalo de $(0,\alpha)$.

Exemplo II.5: O núcleo gaussiano

O núcleo $K(s,t)=exp[-(s-t)^2]$ é $STP(t)$ de todas as ordens em $S=T=(0,\alpha)$; ou seja, é ETP núcleo em ambas as variáveis. Para provar esto à que observar que e^{st} é STP em $(-\infty,\infty)$ pelo exemplo II.1, escrevendo $K(s,t)=exp(-s^2)exp(2st)exp(-t^2)$, utilizando o exemplo II.2*(a)* vemos que $K(s,t)$ é STP. Será de interesse provar que as funções da forma $v_n(t)=\sum_{i=1}^{n} a_i\, exp[-(s_i-t)^2]$, têm no máximo n zeros reais contando as multiplicidades se $s_0<s_2<\cdots<s_n$ e $\sum_{i=0}^{n} a_i^2>0$. Para provar isto vejamos que

$$v_n(t)=\sum_{i=1}^{n} a_i\, exp[-(s_i-t)^2]=e^{-t^2}\sum_{i=1}^{n} a_i e^{-s_i^2} e^{2s_i t}=e^{-t^2}\sum_{i=1}^{n} b_i e^{c_i t}$$

e pela Proposição II.1, no exemplo II.1 obtemos que $v_n(t)$ têm no máximo n zeros.

Denotemos como $Z(f)$ o número de zeros distintos no intervalo $[a,b]$ da função contínua f. Se f for um polinômio real, temos o seguinte Teorema sobre o valor assumido por $Z(f)$, veja Kochendörffer(1972).

Teorema II.1: Qualquer polinômio real de grau m têm no máximo m zeros. Cada zero é contado de acordo à multiplicidade.

Utilizaremos a continuação os *T-system* para obter uma generalização do Teorema acima.

Definição II.6: Seja o sistema $\{u_i\}_{i=0,\ldots,n}$ definido em algum intervalo $[a,b]$. A função $u=\sum_{i=1}^{n} a_i u_i$ onde os a_i são números reais é chamado u-polinômio. O polinômio se diz não trivial se $\sum_{i=1}^{n} a_i^2>0$.

Teorema II.2: Se $\{u_i\}_{i=0,\ldots,n}$ é um *T-system* então $Z(u)\leq n$ para qualquer u-polinômio não trivial. E se o sistema de funções contínuas $\{u_i\}_{i=0,\ldots,n}$ em $[a,b]$ satisfaz $Z(u)\leq n$ qualquer seja $u(t)\neq 0$, temos que $\{u_i\}_{i=0,\ldots,n}$ é *T-system* exceto possivelmente pelo sinal de alguma destas funções.

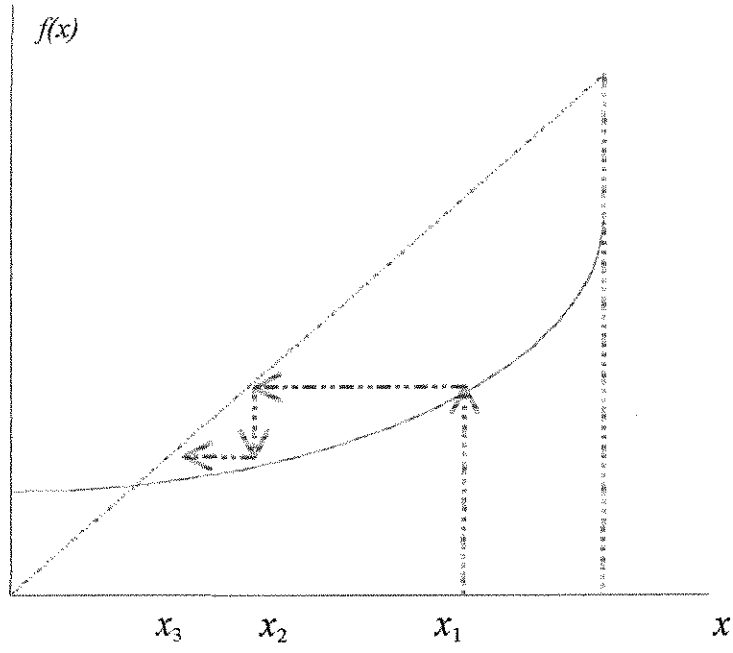
É de observar que o Teorema II.2 assume que as funções $u_o(t),\ldots,u_n(t)$ sejam continuas e não necessariamente diferenciáveis, a demonstração pode ser vista em Karlin & Studden(1966).

Apêndice III:

Aproximações sucessivas

O método de *aproximações sucessivas* é uma das váriadas formas para resolver equações via métodos iterativos de optimização, é aplicado a equações da forma $G(x)=x$. A solução desta equação é chamada de *ponto fixo* da função G desde que G seja invariante em x. Para achar o ponto fixo por aproximações sucessivas, se começa por um valor inicial x_1 e se calcula $x_2=G(x_1)$. Continuando desta maneira iterativamente, se acharão sucessivos valores $x_{n+1}=G(x_n)$. Em determinadas condições a sequëncia $\{x_n\}$ converge à solução da equação original.

A Figura a seguir representa o processo de solução da equação $x=f(x)$, no diagrama esto é equivalente à achar o ponto de interseção de f com a reta $f(x)=x$ que pasa pelo origem. Se começa pelo ponto x_1, se obtem $x_2=f(x_1)$ moviëndo-se ao longo da curva. Na figura, a função f têm intersecto menor que a unidade, isto é o conceito de *contracta*.



Definição III.1: Função contracta

Se dirá que a função real G é *contracta* se existir uma constante $\lambda \in R$, $0 \leq \lambda < 1$, que verifique $|G(x) - G(y)| \leq \lambda \, | x - y |$.

Definição III.2: Seja G uma função real, o ponto x se chamará *ponto fixo* de $G(x)$ se $G(x)=x$.

A continuação inclúem-se o teorema do *ponto fixo* e sua demonstração pela importância que eles têm para provar a convergência da seqüência gerada pelo algoritmo EM, este resultado aparece assim em Luenberger(1969).

Teorema III.1: Ponto Fixo

Seja G uma função real *contracta*, então existe um único vetor $\hat{x}\in R$ satisfazendo $G(\hat{x}) = \hat{x}$. O *ponto fixo* $\hat{x}$ pode ser obtido pelo método de *aproximações sucessivas* a partir de um vetor inicial arbitrário em R.

Prova. Selecionem-mos um vetor arbitrário $x_1\in R$. Definam-mos a seqüência $\{x_n\}$ pela fórmula $G(x_n) = x_{n+1}$. Então $|x_{n+1} - x_n| = |G(x_n) - G(x_{n-1})| \leq \lambda |x_n - x_{n-1}|$. Então, aplicando n - 1 vezes a relação anterior se obtêm que $|x_{n+1} - x_n| \leq \lambda^{n-1} |x_2 - x_1|$.

Este resultado nos permitirá provar que a seqüência $\{x_n\}$ é de Cauchy;

$$|x_{n+p} - x_n| \leq |x_{n+p} - x_{n+p-1}| + |x_{n+p-1} - x_{n+p-2}| + \ldots + |x_{n+1} - x_n|$$

$$\leq (\lambda^{n+p-2} + \lambda^{n+p-3} + \ldots + \lambda^{n-1}) |x_2 - x_1|$$

$$\leq (\lambda^{n-1} \sum_{k=1}^{\infty} \lambda^k) |x_2 - x_1| = \frac{\lambda^{n-1}}{1-\lambda} |x_2 - x_1|.$$

O que prova, que a seqüência $\{x_n\}$ é de Cauchy, logo existirá $\hat{x}\in R$ tal que $x_n \to \hat{x}$. Mostremos agora que $G(\hat{x}) = \hat{x}$. Temos que

$$|\hat{x} - G(\hat{x})| = |\hat{x} - x_n + x_n - G(\hat{x})| \leq |\hat{x} - x_n| + |x_n - G(\hat{x})|,$$

pela desigualdade triangular, e

$$\leq |\hat{x} - x_n| + |x_n - G(\hat{x})| = |\hat{x} - x_n| + |G(x_{n-1}) - G(\hat{x})|,$$

e pela definição de função *contracta* temos que $\exists\lambda\in[0,1)$ tal que

$$|G(x_{n-1}) - G(\hat{x})| \leq \lambda |x_{n-1} - \hat{x}|,$$

logo

$$|\hat{x} - G(\hat{x})| \leq |\hat{x} - x_n| + \lambda |x_{n-1} - \hat{x}|,$$

dado a convergência da seqüência $\{x_n\}$ a $\hat{x}$ a direita da desigualdade anterior é tão pequena quanto se queser, e então

$$|\hat{x} - G(\hat{x})| = 0 \Rightarrow G(\hat{x}) = \hat{x}.$$

Provemos a unicidade, assumamos que $\hat{x}$ e $\hat{y}$ são *pontos fixos*, então

$$|\hat{x} - \hat{y}| = |G(\hat{x}) - G(\hat{y})| \leq \lambda |\hat{x} - \hat{y}|, \text{ logo } \hat{x} = \hat{y} \blacklozenge$$

Apêndice IV

Programa *SAS* utilizado para a geração das amostras de todas as densidades. Incluam-se só aquelas programações desenvolvidas para gerar 25 dados.

```
/* 1- Normal Padrão: N(0,1)                                          */
data datos.n25;
  do i=1 to 25;
    x=normal(10000);
    output datos.n25;
  end;
  keep x;
/* 2- Mistura de Normais: 0.5N(-1.5,1)+0.5N(1.5,1)                   */
data datos.m25;
  do i=1 to 12;
    x=normal(10000)-1.5;
    output datos.m25;
    x=normal(2500)+1.5;
    output datos.m25;
  end;
  y=uniform(800);
  if y<0.5 then x=normal(20000)-1.5;
           else x=normal(40000)+1.5;
  output datos.m25;
  keep x;
/* 3- t-Student, com 5 graus de liberdade                            */
data datos.t25;
  do i=1 to 25;
    x1=2*ranexp(2000);
    x2=2*ranexp(4000);
    y=x1+x2+normal(9000)**2;
    x=normal(10000)/sqrt(y/5);
    output datos.t25;
  end;
  keep x;
/* 4- Cauchy Padrão C(0,1)                                           */
data datos.c25;
  do i=1 to 25;
    x=rancau(40000);
    output datos.c25;
  end;
  keep x;
/* 5- Chi-quadrada, com 6 graus de liberdade                         */
data datos.ch25;
  do i=1 to 25;
    x1=2*ranexp(2000);
    x2=2*ranexp(4000);
    x3=2*ranexp(7000);
    x=x1+x2+x3;
```

```
    output datos.ch25;
  end;
  keep x;
/* 6- Beta(2,2)                                                    */
data datos.b25;
  do i=1 to 25;
    x1=rangam(10000,2);
    x2=rangam(30000,2);
    x=x1/(x1+x2);
    output datos.b25;
  end;
  keep x;
/* 7- Densidade 7                                                  */
data datos.d725;
  do i=1 to 25;
    y=uniform(1000);
    if y<=0.5 then x=sqrt(4*y+0.25)-1.5; else x=1.5-sqrt(17/4-4*y);
    output datos.d725;
  end;
  keep x;
/* 8- Densidade 8                                                  */
data datos.d825;
  total=0;
  do while(total<25);
    signo=uniform(800);
    y=uniform(10000);
    if signo<0.5 then y=-y;
    u=uniform(30500);
    if 3*u<=cos(10*y)+2 then do;
                               x=y;
                               output datos.d825;
                               total=total+1;
                             end;
  end;
  keep x;
/* 9- Densidade 9                                                  */
data datos.d925;
  do i=1 to 25;
    y=uniform(2900);
    if y<=2/3 then x=1.5*y-1; else x=3*y-2;
    output datos.d925;
  end;
  keep x;
```

Apêndice V

Programa *SAS* para a estimação da função de densidade Normal(0,1) pelos métodos Grenander de convolução e Rosenblatt-Parzen. Com este programa se obtêm os estimadores dos parâmetros nos métodos de estimação referidos assim como o Gráfico 5.1.

```
%let min=-3;
%let max=3;
%let c=sqrt(2);
%let pi=sqrt(3.14159265359);
%let sv=3;          /*  Número de parâmetros a estimas no sieves      */
data dados;
  set datos.n25;
  if &min<=x<=&max then output;
data _null_;
  set dados end=eof;
  call symput('x'||left(_n_),x);
  if eof then call symput('n',_n_);
/* ESTIMAÇÃO SIEVES pelo ALGORITMO EM                                   */
proc iml symsize=640;
  p=j(&sv,&sv+1,1);
  m=j(&sv,&sv+1,0);
  s=j(&sv,&sv+1,1);
  var=j(&sv,1,0);
  somas=j(&sv,1,-10000);
  error=j(&sv,3,0);
  start fun(x,m,s);
    ff=exp(-(x-m)**2/(2*s))/sqrt(s);
    return(ff);
  finish;
  do nn=1 to &sv;
    total=nn*(nn+1)/2;
    do i=1 to nn;
      m[i,nn]=&min+i*(&max-&min)/nn;
      p[i,nn]=i/total;
    end;
    %Macro parm;
      erro=1;
      do while( erro>0.00001 );
        do k=1 to nn;
          sum1=0;  sum2=0;
          %do j=1 %to &n;
            sum=0;
            do t=1 to nn;
              sum=sum+p[t,nn]*fun(&&x&j,m[t,nn],s[t,nn]);
            end;
            sum1=sum1+&&x&j*fun(&&x&j,m[k,nn],s[k,nn])/sum;
            sum2=sum2+fun(&&x&j,m[k,nn],s[k,nn])/sum;
```

```
        %end;
        p[k,&sv+1]=p[k,nn]*sum2/&n;
        m[k,&sv+1]=sum1/sum2;
        sum3=0;
        %do j=1 %to &n;
          sum=0;
          do t=1 to nn;
            sum=sum+p[t,nn]*fun(&&x&j,m[t,nn],s[t,nn]);
          end;
          sum3=sum3+fun(&&x&j,m[k,nn],s[k,nn])*(&&x&j-m[k,nn])**2/sum;
        %end;
        s[k,&sv+1]=sum3/sum2;
      end;
      do j=1 to nn;
        error[j,1]=abs(p[j,nn]-p[j,&sv+1]);
        error[j,2]=abs(m[j,nn]-m[j,&sv+1]);
        error[j,3]=abs(s[j,nn]-s[j,&sv+1]);
      end;
      erro=max(error);
      do j=1 to nn;
        p[j,nn]=p[j,&sv+1];
        m[j,nn]=m[j,&sv+1];
        s[j,nn]=s[j,&sv+1];
      end;
    end;
    sum2=0;
    do k=1 to nn;
      sum1=0;
      %do j=1 %to &n;
        sum=0;
        do t=1 to nn;
          sum=sum+p[t,nn]*fun(&&x&j,m[t,nn],s[t,nn]);
        end;
        sum1=sum1+p[k,nn]*fun(&&x&j,m[k,nn],s[k,nn])/sum;
      %end;
      sum2=sum2+sum1*s[k,nn];
    end;
    var[nn]=sum2/&n;
    somas[nn]=0;
    %do i=1 %to &n;
      sum1=0;
      do j=1 to nn;
        sum1=sum1+p[j,nn]*fun(&&x&i,m[j,nn],var[nn]);
      end;
      somas[nn]=somas[nn]+log(sum1);
    %end;
    somas[nn]=somas[nn]-log(&c*&pi);
  %Mend;
  %parm;
 end;
nn=max(somas);
 do i=1 to &sv;
  if somas[i]=nn then nn=i;
 end;
 p1=j(nn,1,0);
 m1=j(nn,1,0);
```

```
  do k=1 to nn;
    p1[k]=p[k,nn];
    m1[k]=m[k,nn];
  end;
  s1=var[nn];
  create parm1 from p1[ colname={ p }];
  append from p1;
  create parm2 from m1[ colname={ m }];
  append from m1;
  create parm3 from s1[ colname={ s }];
  append from s1;
quit;
/* ESTIMAÇÃO KERNEL por Máxima Verossimilhaça Validação Cruzada      */
proc iml symsize=640;
  %Macro soma3;
    start logver(h);
      sum=0;
      %do j=1 %to &n;
        sum1=0;
        %do i=1 %to &j-1;
          sum1=sum1+exp(-((&&x&j-&&x&i)/h)**2/2);
        %end;
        %do i=&j+1 %to &n;
          sum1=sum1+exp(-((&&x&j-&&x&i)/h)**2/2);
        %end;
        sum=sum+log(sum1/((&n-1)*h*&c*&pi))+
            probnorm((&min-&&x&j)/h)-probnorm((&max-&&x&j)/h);
      %end;
      return(sum);
    finish;
    h=1;
    opt={1};
    call nlpdqn(rc,hres,"logver",h,opt);
  %Mend;
  %soma3;
  create result from hres;
  append from hres;
quit;
data null;
  set result;
  call symput('h',COL1);
data _null_;
  set parm1 end=eof;
  call symput('p'||left(_n_),p);
  if eof then call symput('total',_n_);
data _null_;
  set parm2;
  call symput('m'||left(_n_),m);
data _null_;
  set parm3;
  call symput('s',s);
/* Gráfico das funções de densidade estimadas e a teórica           */
proc iml symsize=60;
  f=j(51+&n,5,0);
  cont=0;
  do y=&min to &max by (&max-&min)/50;
```

```
    cont=cont+1;
    sum=0; sum1=0;
    %Macro somas;
      %do j=1 %to &total;
        sum=sum+&&p&j*exp(-(y-&&m&j)**2/(2*&s));
      %end;
      %do j=1 %to &n;
        sum1=sum1+exp(-0.5*((y-&&x&j)/&h)**2);
      %end;
      f[cont,1]=sum/(&c*&pi*sqrt(&s));
      f[cont,2]=(1/(&c*&pi))*exp(-0.5*y**2);
      f[cont,3]=sum1/(&c*&pi*&h*&n);
      f[cont,4]=y;
      f[cont,5]=.;
    %Mend;
    %somas;
  end;
  %Macro somas1;
    %do i=1 %to &n;
      cont=cont+1;
      sum=0; sum1=0;
      %Macro somas2;
        %do j=1 %to &total;
          sum=sum+&&p&j*exp(-(&&x&i-&&m&j)**2/(2*&s));
        %end;
        %do j=1 %to &n;
          sum1=sum1+exp(-0.5*((&&x&i-&&x&j)/&h)**2);
        %end;
      %Mend;
      %somas2;
      f[cont,1]=sum/(&c*&pi*sqrt(&s));
      f[cont,2]=(1/(&c*&pi))*exp(-0.5*(&&x&i)**2);
      f[cont,3]=sum1/(&c*&pi*&h*&n);
      f[cont,4]=&&x&i;
      f[cont,5]=0;
    %end;
  %Mend;
  %somas1;
  create result1 from f[ colname={ fest1 fteo fest2 x pontos }];
  append from f;
quit;
proc sort data=result1;
  by x;
/* Érro quadrático Médio para a estimação Sieves                    */
proc iml symsize=640;
  sum1=0;
  %Macro soma1;
    %do i=1 %to &total;
      sum1=sum1+(&&p&i)*exp(-(&&m&i)**2/(2*(&s+1)));
    %end;
  %Mend;
  sum2=0; sum3=0;
  %Macro soma2;
    %do i=1 %to &total;
      sum2=sum2+(&&p&i)**2;
    %end;
```

```
    %do j=1 %to &total;
      %do l=1 %to &j-1;
         sum3=sum3+(&&p&j)*(&&p&l)*exp(-(&&m&j-&&m&l)**2/(4*&s));
      %end;
    %end;
  %Mend;
  %soma1;
  %soma2;
  eqm=(0.5/&pi)*(1-4*sum1/(&c*sqrt(&s+1))+(sum2+2*sum3)/sqrt(&s));
  print "Erro Quadrático Medio Sieves =" eqm;
quit;
/* Érro quadrático Médio para a estimação Kernel                    */
proc iml symsize=640;
  sum1=0;
  %Macro soma2;
    %do i=1 %to &n;
      sum1=sum1+exp(-(&&x&i)**2/(2*((&h)**2+1)));
    %end;
  %Mend;
  sum2=0;
  %Macro soma2;
    %do j=1 %to &n;
      %do l=1 %to &j-1;
         sum2=sum2+exp(-(&&x&j-&&x&l)**2/(4*(&h)**2));
      %end;
    %end;
  %Mend;
  %soma1;
  %soma2;
  eqm=(0.5/&pi)*(1-4*sum1/(&c*sqrt((&h)**2+1)*&n)+(1+2*sum2/&n)/(&n*&h));
  print "Erro Quadrático Medio Kernel =" eqm;
quit;
symbol1 c=red i=join v=point l=25;
symbol2 c=black i=join v=point l=2;
symbol3 c=green i=join v=point l=1;
symbol4 c=red i=none v=plus;
axis1 order=(0 to 0.6 by 0.15) value=(height=1.5)
      major=(height=1.5) label=none;
axis2 value=(height=1.5) major=(number=7 height=1) label=none;
legend1 label=none down=4 position=(top right inside) mode=protect
        value=(height=1.5 font=simplex "Sieves" "Kernel"
                       "Densidade" "Amostra");
proc gplot data=result1;
  plot fest1*x=1 fest2*x=2 fteo*x=3 pontos*x=4 /
                    overlay legend=legend1 haxis=axis2 vaxis=axis1;
run;
quit;
```

Referências

_ Anderson, T.W., Athjereya, K.B., Iglehart, D.L. (1989) *Probability, Statistics, And Mathematics: Papers In Honor Of Samuel Karlin.* Academic Press.

_ Antoniadis, A. (1988) Parametric estimation for the mean of a gaussian process by the methods of sieves. *Journal of Multivariate Analysis*, Vol.26, 1-15.

_ Barndorff-Nielsen, O. (1978) *Information And Exponential Families Is Statistical Theory.* John Wiley & Sons.

_ Billingsley, P. (1979) *Probability and Measure.* John Wiley & Sons.

_ Bowman, A.W. (1984) An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, Vol.71, 353-360.

_ Bowman, A.W. (1985) A Compartive Study Of Some Kernel-Based Nonparametric Density Estimators. *Journal Of Statistical Computations And Simulations*, Vol.21, 313-327.

_ Bustos, O.H., Frery, A.C. (1992) *Simulação Estocástica: Teoria E Algoritmos.* 10° Simpósio Nacional De Probabilidade E Estatística.

_ Chapman, D.G. (1956) Estimating The Parameters Of A Truncated Gamma Distribution. *The Annals Of Mathematical Statistics*, Vol.27, No.2, 498-506.

_ Chow, Y.-S., Geman, S., Wu, L.-D. (1983) Consistent Cross-Validated Density Estimation. *The Annals Of Statisticcs*, Vol.11, No.1, 25-38.

_ Cline, D.B.H. (1988) Admissible Kernel Estimators Of A Multivariate Density. *The Annals Of Statistics*, Vol.16, No.4, 1421-1427.

_ Crain, B.R. (1974) Estimation Of Distributions Using Orthogonal Expansions. *The Annals Of Statistics*, Vol.2, No.3, 454-463.

_ Cramér, H. (1946) *Mathematical Methods of Statistics*, Princeton University Press.

_ De Montricher, G.F., Tapia, R.A., Thompson, J.R. (1975) Nonparametric Maximum Likelihood Estimation Of Probability Densities By Penalty Function Methods. *The Annals Of Statistics*, Vol.3, No.6, 1329-1348.

Deheuvels, P. (1977) Estimation Nonparametrique De La Densité Par Histogrammes Generalisés. *Revue De Statistique Appliquée*, Vol. XXV, No.3, 5-42.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977) Maximum Likelihood From Incomplete Data Via The EM Algorithm. *Journal Of The Royal Statistical Society*, Series B, Vol.39, No.1, 1-38.

Devroye, L. (1987) *A Course In Density Estimation*. Birkhäuser.

Devroye, L., Gyorfi, L. (1985) *Nonparametric Density Estimation: The L_1 View*. John Wiley.

Donoho, D.L., Johnstone, I.M. (1989) Projection-Based Approximation And A Duality With Kernel Methods. *The Annals Of Statistics*, Vol. 17, 58-106.

Duin, R.P.W. (1976) On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Transactions on Computers*. C-125, 1175-1179.

Efron, B. (1967) *The Two Sample Problem With Censored Data*. Proceedings Of The Fifth Berkeley Simposium On Mathematical Statistics And Probability, IV. Berkeley: University Of California Press, 831-853.

Epanechnikov, V.A. (1969) Non-Parametric Estimation Of A Multivariate Probability Density. *Theory Of Probability And Its Applications*, Vol.14, 153-158.

Essen, C. (1945) Fourier Analysis Of Distribution Functions. A Mathematical Study Of The Laplace-Gaussian Law. *Acta Mathematica*, Vol.77, 1-125.

Fryer, M.J. (1976) Some Errors Associated With The Nonparametric Estimation Of Density Function. *Journal Of The Institute Of Mathematics And Its Application*, Vol. 18, 371-380.

Geman, S. (1981) *Sieves For Nonparametric Estimation Of Densities And Regressions*. Reports In Pattern Analysis No.99, Division Of Applied Mathematics, Brown University.

Geman, S. (1982) *Method Of Sieves*. Reports On Pattern Analysis No.125, Division Of Applied Mathematics, Brown University.

Geman, S., Hwang, C. (1982) Nonparametric Maximum Likelihood Estimation By The Method Of Sieves. *The Annals Of Statistics*, Vol.10, No.2, 401-414.

Gimenez, P.C. (1993) *Estimação De Máxima Verossimilhança Não Paramétrica Pelos Métodos De Grenander E De Máxima Verossimilhança Penalizada*. Dissertação De Mestrado Apresentada Ao IMECC, UNICAMP.

Good, I.J., Gaskins, R.A. (1971) Nonparametric Roughness Penalties For Probability Densities. *Biometrika*, Vol.58, 255-277.

Grenander, U. (1981) *Abstract Inference*. John Wiley & Sons.

Habbema, J.D.F., Hermans, J., Van Den Broek, K. (1974) *A Stepwise Discrimination Analysis Program Using Density Estimation*. In Compstat 1974: Proceeding In Computational Statistics. (G. Buckman, Editor), 101-110. Physica Verlag, Vienna.

Hall, P. (1982) Cross-Validation In Density Estimation. *Biometrika*, Vol.69, No.2, 383-390.

Hall, P. (1983) Large Sample Optimality Of Least Squares Cross-Validation In Density Estimation. *The Annals Of Statistics*, Vol. 11, 1156-1174.

Hall, P., Marron, J.S. (1987a) Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probability Theory and Related Fields*, Vol.74, 567-581.

Hall, P., Marron, J.S. (1987b) On the amount of noise inherent in bandwidth selection for a Kernel density estimation. *Annals of Statistics*. Vol.15, 163-181.

Hasselblad, V. (1966) Estimation Of Parameters For A Mixture Of Normal Distributions. *Technometrics*, Vol. 8, 431-444.

Hasselblad, V. (1969) Estimation Of Finite Mixtures Of Distributions From The Exponencial Family. *Journal Of The American Statistical Assiciation*, Vol. 64, 1459-1471.

Hodges, J.L., Lehmann, E.L. (1956) The Efficiency Of Some Nonparametric Competitors To The T-Test. *The Annals Of Mathematical Statistics*, Vol.13, 324-335.

Jeff Wu, C.F. (1983) On The Convergence Properties Of The EM Algorithm. *The Annals Of Statistics*, Vol.11, No.1, 95-103.

Johnson, N., Kotz, S. (1988) *Encyclopedia of Statistical Science*. John Wiley.

Johnson, M.A., Taaffe, M.R. (1993) Tchebycheff Systems For Probabilistics Analysis. *American Journal Of Mathematical And Management Sciences*, Vol.13, 83-111.

Karlin, S., Studden, W.J. (1966) *Tchebycheff System: With Application In Analysis And Statistics*. John Wiley & Sons.

Karr, A. (1987) Maximum likelihood estimation in the multivariate intensity model via sieves. *Annals of Statistics*, Vol.15, 473-490.

Kochendörffer, R. (1972) *Introduction To Algebra*. Wolters-Noordhoff Publishing.

Kullback, S., Leibler, R.A. (1951) On Information And Sufficiency. *The Annals Of Mathematical Statistics*, Vol.22, No.1, 79-86.

Laird, N. (1978) Nonparametric Maximum Likelihood Estimation Of A Mixing Distribution. *Journal Of The American Statistical Association*, Vol.73, No.364, 805-811.

Lorentz, G.G. (1966) *Aproximation Of Functions*. Holt, Rinchart And Winston.

Luenberger, D.G. (1969) *Optimization By Vector Space Methods*. John Wiley & Sons.

Marron, J.S. (1983) Optimal rates of convergence to Bayes risk in nonparametric discrimination. *Annals of Statistics*, Vol.11, 1142-1155.

Marron, J.S. (1985) An Assymptotically Efficient Solution To The Bandwidth Problem Of Kernel Density Estimation. *The Annals Of Statistics*, Vol.13, No.2, 1011-1023.

McKeague, I.W. (1986) Estimation for a semimartingale regression model using the method of sieves. *Annals of Satatistics*, Vol.14, 579-589.

McLahchlan, G.J., Basford, K.E. (1988) *Mixture Models: Inference And Applications To Clustering*. Marcel Dekker, Inc.

Moulin, P., O'Sullivan, J., Snyder, D.L. (1992) A method of sieves for multiresolution spectrum estimation and radar imaging. *IEEE Transactions on Information Theory*, Vol.38, 801-813.

Nguyen, H.T. (1982) Identification of nonstationary diffusion model by the method of sieves. *SIAM Journal of Control and Optimization*, Vol.20, 603-611.

Park, B.U., Marron, J.S. (1990) Comparison Of Data-Driven Bandwidth Selectors. *Journal Of The American Statistical Association*, Vol.85, 66-72.

Parzen, E. (1962) On Estimation Of A Probability Density Function And Mode. *The Annals Of Mathematical Statistics*, Vol.33, 1065-1076.

Peters, C.(Jr), Walker, H.F. (1978a) The Numerical Evaluation Of The Maximum-Lekelihood Estimate Of A Subset Of Mixture Proportions. *SIAM Journal Of Applied Mathematics*, Vol.35, No.3, 447-452.

Peters, C.(Jr), Walker, H.F. (1978b) An Iterative Procedure For Obtaining Maximum-Likelihood Estimates Of The Parameters For A Mixture Of Normal Distributions. *SIAM Journal Of Applied Mathematics*, Vol.35, No.2, 362-378.

Pólya, G., Szegö, G. (1925) *Aufgaben Und Lehrsätze Aus Der Analysis*, Band I, II. Springer.

Rao, C.R. (1973) *Linear Statistical Inference And Its Applications.* John Wiley & Sons.

Redner, R.A., Walker, H.F. (1984) Mixture Densities, Maximum Likelihood And The EM Algorithm. *SIAM Review*, Vol.26, No.2, 195-239.

Rice, J.R. (1964) *The Approximation Of Functions*. Volume 1. *Linear Theory*. Addison-Wesley Publishing Company, INC.

Robbins, H. (1964) The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics*, Vol.35, 1-20.

Robertson, T. (1967) On estimating a density which is measurable with respect to a s-lattice. *Annals of Mathematical Statistics*, Vol.38, 482-493.

Rosenblatt, M. (1956) Remarks On Some Nonparametric Estimate Of A Density Function. *The Annals Of Mathematical Statistics*, Vol. 27, 832-837.

Rudemo, M. (1982) Empirical Choise Of Histograms And Kernel Density Estimators. *Scandinavian Journal Of Statistics,* Vol.9, 65-78.

Scott, D.W., Terrell, G.R. (1987) Biased And Unbiased Cross-Validation In Density Estimation. *Journal Of The American Statistical Association*, Vol.82, 1131-1146.

Shen, X., Hung Wing, W. (1994) Convergence Rate Of Sieve Estimates. *The Annals Of Statistics,* Vol.22, No.2, 580-615.

Silverman, B.W. (1978) Choosing the window width when estimating a density. *Biometrika*, Vol.65, 1-11.

_ Silverman, B.W. (1986) *Density Estimation For Statistical And Data Analysis*. Chapman & Hall.

_ Stone, C.J. (1984) An asymtotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, Vol.12, 1285-1297.

_ Tanner, M.A. (1991) *Tools For Statistical Inference*. Lecture Notes In Statistical. Springer-Verlag, Vol.61.

_ Tapia, R.A., Thompson, J.R. (1990) *Nonparametric Function Estimation, Modeling And Simulation*. SIAM.

_ Titterington, D.M., Smith, A.F.M., Markov, U.E. (1985) *Statistical Analysis Of Finite Mixture Distributions*. John Wiley & Sons.

_ Turnbull, B.W. (1974) Nonparametric Estimation Of A Survivorship Function With Doubly Censored Data. *Journal Of The American Statistical Association*, Vol. 69, 169-173.

_ Turnbull, B.W. (1976) The Empirical Distribution Function With Truncated Data. *Journal Of The Royal Statistical Society*, Series B, Vol. 38, 290-295.

_ Wahba, G. (1981) Data-based optimal smoothing of orthogonal series density estimates. *Annals of Statistics*, Vol.9, 146-156.

_ Wald, A. (1949) Note On The Consistency Of The Maximum Likelihood Estimate. *The Annals Of Mathematical Statistics*, Vol.20, 595-601.

_ Walter, G., Blum, R. (1984) A Simple Solution To A Nonparametric Maximum Likelihood Estimation Problem. *The Annals Of Statistics*,Vol.12, No.1, 372-379.

_ Wand, M.P., Jones, M.C. (1995) *Kernel Smoothing*. Chapman & Hall.

_ Wegman, E.J. (1975) Maximum Likelihood Estimation Of A Probability Density Function. *Sankyä*, Series A, Vol.37, Part.2, 211-224.

_ Wheeden, R., Zygmund, A. (1977) *Measure and Integral: an Introduction to Real Analysis*. Marcel dekker.

_ Zehna, P.W. (1966) Invariance Of Maximum Likelihood Estimators. *The Annals Of Mathematical Statistics*, Vol.37, 744.