



UNIVERSIDADE ESTADUAL DA PARAÍBA
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA

André Pereira do Nascimento

**Estimação do Tipo Kernel para Distribuições
Simétricas Usando Técnica de Monte Carlo**

CAMPINA GRANDE - PB
FEVEREIRO DE 2013.

André Pereira do Nascimento

Estimação do Tipo Kernel para Distribuições Simétricas Usando Técnica de Monte Carlo

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador:

Prof. Msc. Kleber Napoleão Nunes de Oliveira Barros

CAMPINA GRANDE - PB

FEVEREIRO DE 2013.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL – UEPB

N244e Nascimento, André Pereira do.
Estimação tipo *Kernel* para distribuições simétricas usando técnica de Monte Carlo [manuscrito] / André Pereira do Nascimento. – 2013.
49 f. : il. color.

Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2014.

“Orientação: Prof. Me. Kleber Napoleão Nunes de Oliveira Barros, Departamento de Estatística”.

1. Método Kernel. 2. Distribuições simétricas. 3. Histograma.
I. Título.

21. ed. CDD 310

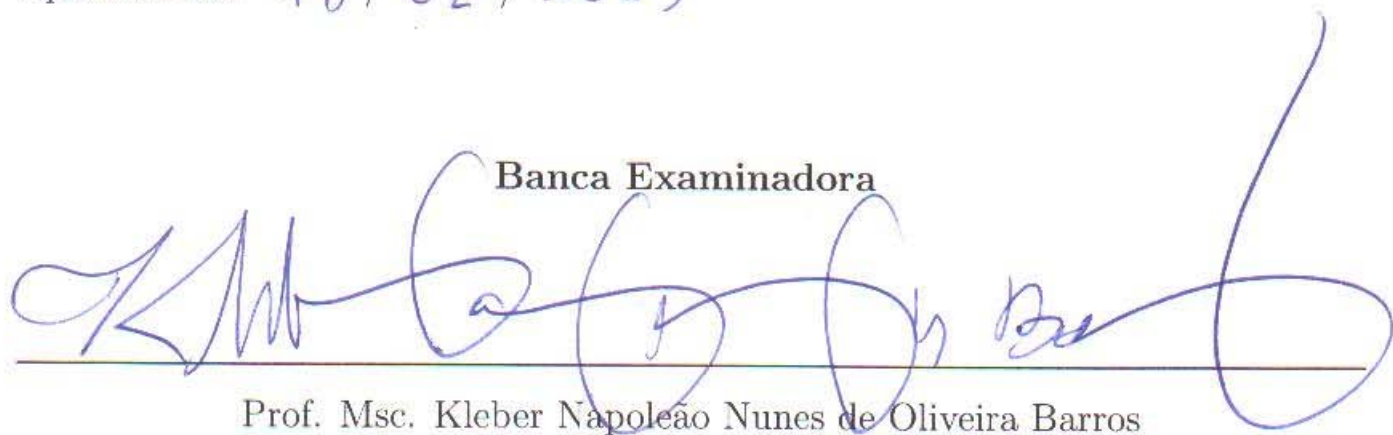
André Pereira do Nascimento

Estimação Tipo Kernel para Distribuições Simétricas Usando Técnica Monte Carlo

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Aprovada em: 18 / 02 / 2013

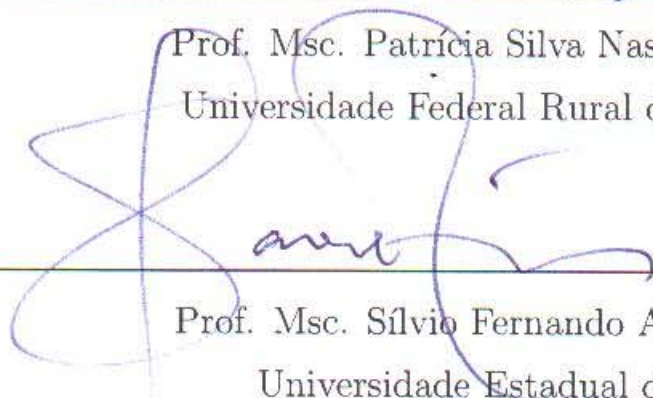
Banca Examinadora



Prof. Msc. Kleber Napoleão Nunes de Oliveira Barros
Universidade Estadual da Paraíba - Orientador

Patrícia Silva Nascimento Barros

Prof. Msc. Patrícia Silva Nascimento Barros
Universidade Federal Rural de Pernambuco



Prof. Msc. Sílvia Fernando Alves Xavier Jr.
Universidade Estadual da Paraíba

Dedicatória

A minha esposa, Eunice Minervina do Nascimento, pela dedicação, companheirismo e amizade e ao meu filho André Felipe Minervino do Nascimento pelos momentos de alegria.

Agradecimentos

Ao meu Deus que me deu força saúde e sabedoria durante toda minha vida, aos meus pais, que me direcionaram ao estudo como também todos os professores que nos acompanhou durante esta jornada.

Resumo

O presente trabalho mostra de uma forma simples e objetiva algumas distribuições estatísticas simétricas que são aquelas que se caracterizam por terem médias, modas e medianas similares. Através da construção de um gráfico do tipo histograma, esta característica fica bem visível, neste caso é necessário observar se as caldas dos lados esquerdo e direito do gráfico se equivalem. O objetivo principal do estudo é mostrar um tipo de estimação não paramétrico chamado de *kernel* que se caracteriza por não ter parâmetros do tipo média e variância que frequentemente estão presentes em estimações. O estimador de densidades tipo *kernel* se baseia em um núcleo K , que é uma equação ponderada por um número positivo h . Se o valor desse h for grande, fará com que esse estimador seja sempre suave.

Palavras-chave: Método Kernel, Estimação, Distribuições Simétricas.

Abstract

This paper presents a simple and objective way some statistical distributions that are symmetrical ones that are characterized by averages, medians and similar mode through the construction of a histogram chart type it all becomes clear in this case it is necessary to note that the tails of the left and right sides of the graph are equivalent. The main objective of the study is to show a kind of nonparametric estimation called *kernel* which is characterized by not having parameters like average and variation that are frequently present in much of the estimates, however, the *kernel* type density estimator is based a nucleus in which K in equation weighted by a positive number h if the value of h is large, will cause this estimator is always mild.

Keywords: Kernel Methods, Pets, Symmetric Distributions.

Sumário

Dedicatória	
Agradecimentos	
Resumo	
Abstract	
1 Introdução	p. 5
2 Revisão Bibliográfica	p. 8
2.1 Método Kernel	p. 8
2.2 Integração Monte Carlo	p. 11
2.3 Funções de Kernel Frequentemente Utilizadas	p. 12
2.4 Distribuições Simétricas	p. 14
2.5 Testes de Ajustamento	p. 21
3 Metodologia	p. 23
4 Resultados e Discussão	p. 25
5 Conclusão	p. 35
Referências Bibliográficas	p. 36

1 Introdução

Em estatística, assimetria é o grau de afastamento de uma distribuição em torno de um valor central. De um modo geral, quando os valores da média, da moda e da mediana coincidem, diz-se que os dados exibem simetria ou são assimétricos. É possível observar este comportamento ao se desenhar o histograma. Se a calda direita e a esquerda são idênticas quanto ao formato, então os dados são simétricos. No entanto, quando o valor da moda é inferior ao da mediana que, por sua vez, possui um valor menor que a média, a distribuição diz-se assimétrica positiva ou assimétrica à direita. Já quando o valor da moda é superior ao da mediana que, por sua vez, possui um valor superior ao da média, a distribuição diz-se assimétrica negativa ou assimétrica à esquerda. A assimetria é fácil de determinar graficamente. Podendo dizer se uma distribuição é simétrica ou assimétrica (positiva ou negativa) pelo aspecto do seu histograma. Quando não se dispõem de meios gráficos o grau de assimetria de uma distribuição pode ser medido utilizando um indicador que é o coeficiente de assimetria. A modelagem estatística de dados utiliza quase sempre a distribuição normal. Ela é uma das mais importantes distribuições da estatística, pois além de descrever uma série de fenômenos, possui grande uso inferencial. É inteiramente descrita por seus parâmetros de média e desvio padrão, ou seja, conhecendo os dois consegue-se determinar qualquer probabilidade em uma distribuição normal. Os métodos paramétricos baseiam-se na suposição de que os dados observados na amostra são provenientes de uma população com distribuição teoricamente conhecida. A suposição de que os dados seguem uma distribuição normal é assumida para a maioria dos métodos estatísti-

cos. Este fato somado a resultados teóricos fundamentais faz com que a distribuição normal seja a distribuição teórica mais importante em estatística. Um motivo pelo qual se usa tanto a distribuição normal é que ela serve de aproximação para o cálculo de outras distribuições quando o número de observações fica grande. Essa importante propriedade provém do Teorema Central do Limite que diz que toda soma de variáveis aleatórias independentes de média finita e variância limitada é aproximadamente normal, desde que o número de termos da soma seja suficientemente grande. Alguns tipos de dados podem apresentar uma dispersão maior que a admitida pela distribuição normal (apresentam caudas mais “pesadas”). Por exemplo, numa empresa podem ser encontrados salários bastante discrepantes em relação à média. Assim, Novos métodos têm sido desenvolvidos, recentemente, tais como modelos simétricos transformados utilizando outras distribuições tais como: Cauchy, Exponencial potência, logística tipo I e II, t de Student, entre outras, inclusive a normal.

Este trabalho disserta sobre estimação não paramétrica, em contraposição a estimação paramétrica. A estatística paramétrica é uma área da Estatística que procura relacionar certas variáveis de interesse com outras variáveis, ditas auxiliares por meio de estimativas (pesos) de parâmetros e admitindo um modelo probabilístico subjacente. São campos da estatística paramétrica: amostragem, inferência, teoria da regressão entre outras, podemos ter grande precisão dependendo da coleta dos dados, como também do uso adequado dos softwares. Em um exemplo geral, é que podemos prever a população futura de uma cidade, estudando o crescimento dessa mesma população no passado. A estatística não paramétrica por sua vez, procura não utilizar estimativas de parâmetros, nem supor uma distribuição para a variável de interesse. Este método estatístico permite que a forma funcional de um ajuste para os dados a serem obtidos na ausência de qualquer orientação ou limitações teóricas. Como resultado, os procedimentos de estimação não paramétrica não têm parâmetros significativos associados. Podemos citar dois tipos de técnicas não paramétricas, o primeiro se refere as redes neurais artificiais e o segundo é a estimação tipo *kernel*. As redes neurais artificiais modelam uma

função desconhecida, expressada como uma soma ponderada de curvas logísticas, cada uma delas é uma função de todas as variáveis relevantes explicativas. Isso se resume em uma forma extremamente flexível funcional para os quais exige uma estimativa não linear de mínimos quadrados iterativo algoritmo de busca baseado em gradientes. O método de estimação de densidade *kernel* é um estimador probabilístico não paramétrico que não utiliza média e desvio padrão como parâmetro e não segue uma distribuição normal ou não tem elementos suficientes para afirmar que seja uma normal. Em distribuições não paramétricas, um *kernel* é uma função de ponderação utilizada em técnicas de estimação. Unidades de medidas são utilizados na estimativa de densidade para estimarmos funções de densidade de variáveis aleatórias, ou em regressão *kernel* para estimar a esperança condicional de uma variável aleatória. Núcleos também são usados em série de tempo, para estimar-se a densidade espectral. Uma utilização adicional é na estimativa de um tempo que varia de intensidade para um processo de ponto. É também uma técnica estatística de interpolação, exploratória que mostra o padrão de distribuição de pontos gerando uma superfície de densidade com identificação visual de áreas com maior intensidade da ocorrência de um evento. Esse tipo de estimador trabalha em parceria com algumas distribuições como, uniforme, triangular, Epanechnikov, quártica (biweight), tricube, triweight, normal, e cosseno. Em consequência desta parceria, obten-se funções estimadoras de *kernel*. No presente estudo, pretende-se simular, a partir de diversas distribuições conhecidas, dados com vários comportamentos; e ajustar estimadores do tipo *kernel* utilizando como núcleo diversas distribuições simétricas, comparando os resultados das distribuições amplamente empregadas para verificar qual o melhor ajuste produzido.

2 Revisão Bibliográfica

2.1 Método Kernel

O método do Kernel é um método não paramétrico para estimação de curvas de densidades onde cada observação é ponderada pela distância em relação a um valor central, o núcleo. Histogramas são descontínuos. Assim, estimadores de densidade de kernel, que são mais suaves, podem ser empregados pois estes convergem mais rápido para a verdadeira densidade de histogramas.

Sejam X_1, X_2, \dots, X_n uma amostra aleatória observada de uma função f . Um núcleo é definido como sendo qualquer função suave K tal que

- (i) $K(x) \geq 0$;
- (ii) $\int K(x)dx = 1$;
- (iii) $\int xK(x)dx = 0$ e
- (iv) $\sigma^2 = \int x^2K(x)dx > 0$.

Um exemplo de Kernel bastante utilizado é o Gaussiano ou normal, dado por

$$K(x) = (2\pi)^{-1/2}e^{-x^2/2}$$

Na Figura (2.1) é mostrado um exemplo de estimador tipo *kernel*. Note que o estimador de densidade tipo kernel (linha preta) é a sobreposição das densidades, centradas em X_i e com desvio h .

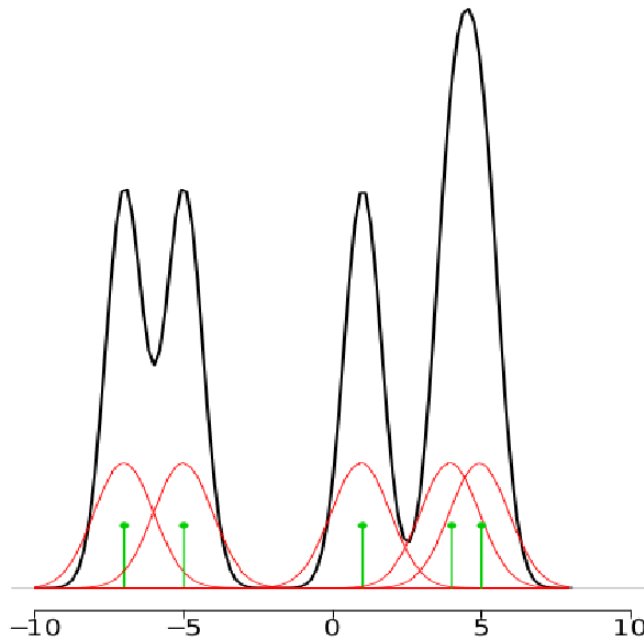


Figura 2.1: Um estimador de densidade por *kernel* \hat{f} . Para cada ponto x , $\hat{f}(x)$ é uma média dos “*kerneis*” centrados sobre os pontos X_i . Os dados são indicados por barras verticais.

Fonte: Wasserman (2004).

O estimador de densidades tipo kernel (KDE, em inglês) se baseia num núcleo (kernel, em inglês) K e um número positivo h , é denominado bandwidth (largura de banda), então o KDE é definido por:

$$\hat{f}_h = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - X_i}{h}\right) \quad (2.1)$$

Assim, um valor h maior fará o KDE ser mais suave. Uma pergunta importante é então: qual o valor de h que deve ser utilizado para o KDE se ajustar de forma plausível aos dados?

Assumindo pressupostos fracos sobre F e K , pode-se definir a seguinte função de risco

$$R(f, \hat{f}_n) \approx \frac{1}{4} \sigma_K^4 h^4 \int (f''(x))^2 + \frac{\int K^2(x) dx}{nh}$$

onde $\sigma_K^2 = \int K^2(x) dx$. A largura ótima de banda é

$$h^* = \frac{c_1^{-2/5} c_2^{1/5} c_3^{-1/5}}{n^{1/5}} \quad (2.2)$$

em que $c_1 = \int x^2 k(x) dx$, $c_2 = \int k(x)^2 dx$ e $c_3 = \int (f''(x))^2 dx$. Com esta escolha para a largura de banda,

$$R(f, \hat{f}_n) \approx \frac{c_4}{n^{4/5}}$$

para alguma constante $c_4 > 0$.

Conforme Wasserman (2004), pode ser mostrado que o kernel Epanechnikov (ver Tabela 2.1) é ótimo, no sentido de dar o menor erro quadrático médio assintótico, mas é realmente a escolha de largura de banda que é crucial.

Note que a equação subentende a determinação da função desconhecida f dos dados, sendo portanto pouco prática. Uma alternativa apresentada por Rudemo (1982) é utilizar a função

$$\hat{J}(h) = \int [\hat{f}(x)]^2 dx - \frac{2}{n(n-1)} \sum_{i \neq j} \frac{1}{h} K(X_i, X_j) \quad (2.3)$$

em que $K(u, v) = K\left(\frac{u-v}{h}\right)$. O melhor valor para h , será o mínimo de $\hat{J}(h)$. Porém a função (2.3) é de avaliação complexa, pois envolve uma integral do quadrado de uma soma. Não havendo resolução analítica, então deve-se utilizar uma técnica de integração numérica.

2.2 Integração Monte Carlo

Uma alternativa para a resolução da equação 2.4 é utilizar uma técnica de integração numérica, conhecida como integração Monte Carlo. Suponha uma função qualquer $H(x)$ e uma densidade $p(x|X)$ qualquer. Esta técnica garante que:

$$\int H(x)p(x|X)dx \approx \frac{1}{m} \sum_{k=1}^m H(x)$$

desde que m seja suficientemente grande e x seja simulado da distribuição $p(x|X)$, isto é, $x \sim p(x|X)$.

Suponha que se pretende resolver a expressão (2.3) para K seguindo a distribuição normal padrão. Então, a integral em $\hat{J}(h)$ fica:

$$\begin{aligned} \int [\hat{f}(x)]^2 dx &= \int_{-\infty}^{\infty} \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{(x-X_i)^2}{2h^2}} \right)^2 dx \\ &= \frac{1}{n^2} \sum_{j=1}^n \int_{-\infty}^{\infty} \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{(x-X_i)^2}{2h^2}} \right) \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{(x-X_j)^2}{2h^2}} dx \\ &= \frac{1}{n^2} \sum_{j=1}^n \int H(x)p(x|X)dx \end{aligned} \quad (2.4)$$

Assim, para resolver a equação (2.3) basta simular $x \sim N(X_i, h^2)$ e escrever

$$H(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{(x-X_i)^2}{2h^2}}.$$

O processo é idêntico para outras distribuições. Basta simular $x \sim D(X_i, h^2)$, da distribuição requerida e fazer $H(x) = 1/n \sum_{i=1}^n K((x-X_i)/h)$. O cálculo do (2.3) exige algum esforço computacional, porém com computadores cada vez mais rápidos não se leva mais que alguns segundos para ser realizado.

2.3 Funções de Kernel Frequentemente Utilizadas

Na literatura (ZUCCHINI, 2003; COMANICIU; MEER, 2002) são encontradas diversas distribuições que se utilizam do núcleo da estimação por *Kernel*. Entre elas estão: uniforme, triangular, Epanechnikov, quadrática, tricúbica, triweight, normal e cosseno. A Tabela 2.1 a seguir mostra as distribuições mais utilizadas em KDE:

Tabela 2.1: Principais distribuições utilizadas em KDE.

Kernel	$K(x)$
Uniforme	$\frac{1}{2}I_{ x \leq 1}$
Triangular	$(1 - x)I_{ x \leq 1}$
Epanechnikov	$\frac{3}{4}(1 - x^2)I_{ x \leq 1}$
Quártica	$\frac{15}{16}(1 - x^2)^2I_{ x \leq 1}$
Triweight	$\frac{35}{32}(1 - x^2)^3I_{ x \leq 1}$
Tricúbica	$\frac{70}{81}(1 - u ^3)^3I_{ x \leq 1}$
Normal	$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$
Cosseno	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}x\right)I_{ x \leq 1}$

em que

$$I_A = \begin{cases} 1, & \text{se } x \in A, \\ 0, & \text{caso contrário.} \end{cases}$$

é a função indicadora. A Figura 2.2 a seguir mostra os gráficos para as funções de Kernel da Tabela 2.1.

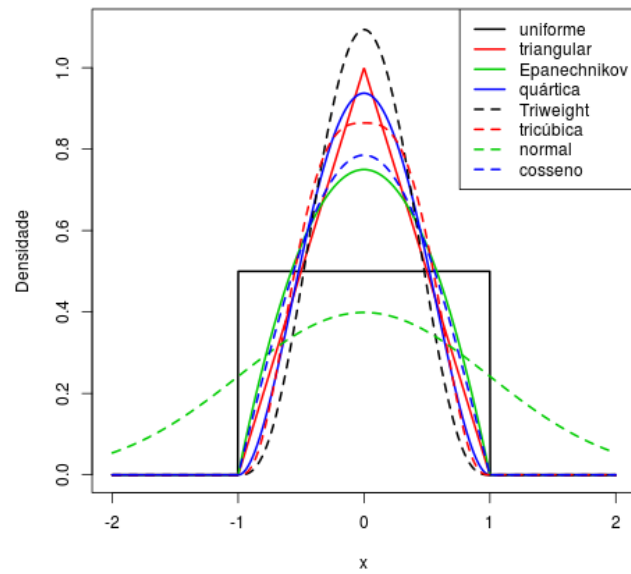


Figura 2.2: Densidades das principais distribuições utilizadas em KDE.

Observe que todas as distribuições acima são simétricas em torno do zero. A seguir propõe-se uma classe de distribuições, conhecidas como distribuições simétricas (ou elípticas univariadas) para substituir as distribuições acima na estimação de densidades por Kernel.

2.4 Distribuições Simétricas

Uma distribuição diz-se simétrica quando os valores são relativamente uniformemente distribuídos em torno da média. É geralmente o formato de uma curva de sino quando representado num gráfico, e ainda, se uma linha é desenhada desse gráfico, um lado vai espelhar o outro.

Diz-se que a variável aleatória X tem distribuição simétrica, com suporte em \mathbb{R} , com parâmetros de localização $\mu \in \mathbb{R}$ e de escala $\phi > 0$, se sua função densidade é da forma

$$f(x; \mu, \phi) = \frac{1}{\sqrt{\phi}} g \left[\left(\frac{x - \mu}{\sqrt{\phi}} \right)^2 \right], \quad x \in \mathbb{R} \quad (2.5)$$

para alguma função $g(\cdot)$ denominada função geradora de densidade, com $g(u) > 0$, para $u > 0$ e $\int_0^\infty u^{-1/2} g(u) du = 1$ (CYSNEIROS *et al.*, 2005). Essa condição é necessária e suficiente para que $f(x; \mu, \phi)$ seja uma função densidade de probabilidade. Denota-se $X \sim S(\mu, \phi)$ e denomina-se X de variável aleatória simétrica.

As distribuições mais comuns pertencentes a esta família de distribuições simétricas são: normal, Cauchy, laplace, lógica tipo I e II, t de Student.

Distribuição Normal

A distribuição normal é a distribuição de probabilidade contínua conhecida por alguns nomes, tais como distribuição Gaussiana, distribuição de Gauss ou curva de sino (LIMA-FILHO, 2009). Sua densidade tem uma forma característica, conhecida como forma de sino

E pode ser escrita como:

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty \quad (2.6)$$

O parâmetro μ é a média e σ^2 é a variância. O parâmetro σ é conhecido como o

desvio-padrão. A distribuição com $\mu = 0$ e $\sigma^2 = 1$ é chamado a distribuição normal padrão. É frequentemente utilizada como a primeira aproximação para descrever o valor das variáveis aleatórias que estão em torno da média. A distribuição normal dentre as demais é a mais utilizada por diversos fatores, em primeiro lugar, a distribuição normal surge do teorema central do limite (TCL), que estabelece que, sob condições de regularidade, a média de um grande número de variáveis aleatórias tiradas da mesma distribuição é aproximadamente normalmente distribuída, independentemente da forma da distribuição original. Em segundo lugar, a distribuição normal é tratável analiticamente, isto é, um número elevado de resultados que envolvem essa distribuição podem ser derivada de forma explícita. Por estes e outros motivos a distribuição normal é tão utilizada na pratica das diversas áreas de estudo. A Figura (2.3) mostra o gráfico da normal padrão.

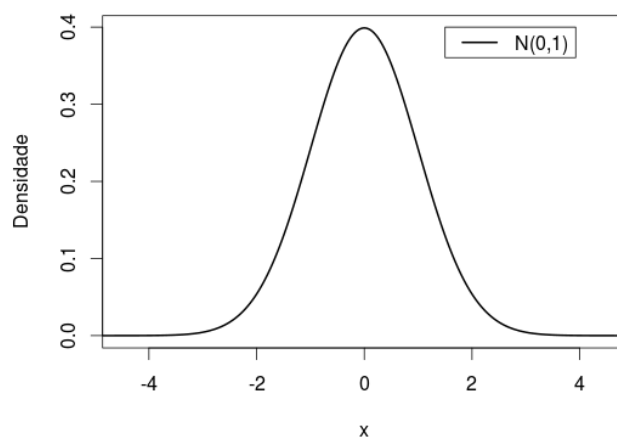


Figura 2.3: Gráfico da função densidade de probabilidade da normal padrão.

Distribuição Cauchy

A distribuição de Cauchy recebeu este nome em homenagem Augustin Cauchy que foi um matemático francês que se destacou como um dos pioneiros em análise matemática. É uma distribuição contínua de probabilidade que tem como característica, uma variável aleatória que é a razão entre dois padrões normais independentes de variáveis aleatórias. Uma particularidade desta distribuição é que não possui função geradora de momentos. Sua f.d.p.(função de densidade de probabilidade) é dada por:

$$f(x; \mu, \sigma) = \left\{ \pi \sigma \left[1 + \left(\frac{x - \mu}{\sigma} \right)^2 \right] \right\}^{-1}, \quad -\infty < x < \infty \quad (2.7)$$

em que μ é um parâmetro de locação e $\sigma > 0$. Outra importante característica da distribuição de Cauchy é que ela não possui momentos finitos e, portanto, não tem valor esperado e variância (CYSNEIROS *et al.*, 2005). Não se deve, portanto, confundir momentos com os parâmetros de locação e escala. A distribuição de Cauchy possui caudas mais pesadas do que a distribuição normal, conforme Figura 2.4 abaixo:

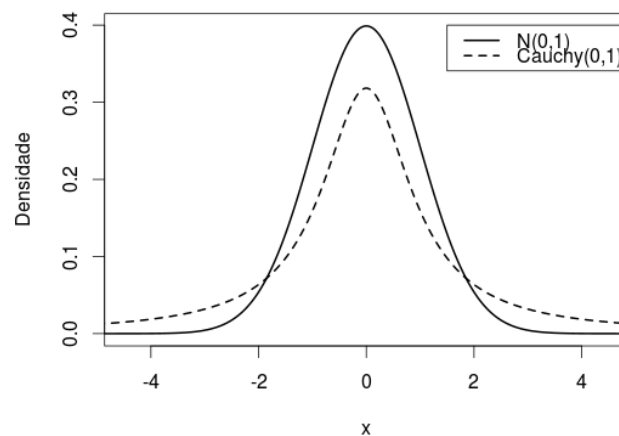


Figura 2.4: Gráfico comparando densidade da normal padrão com a densidade de Cauchy.

Quando os valores dos parâmetros de locação e de escala são iguais a 0 e 1, respectivamente, a distribuição Cauchy passa a ser chamada como Cauchy padrão ou t de Student central com um grau de liberdade.

Distribuição Laplace

Como as anteriores, a distribuição de Laplace também é contínua e recebeu este nome em homenagem ao matemático, astrônomo e físico francês Pierre-Simon de Laplace, algumas literaturas denominam esta distribuição como exponencial dupla (ED), pois pode ser considerada como duas distribuições exponenciais. A diferença entre as duas variáveis aleatórias independentes identicamente distribuídas exponenciais é gerada por uma distribuição de Laplace, que é um movimento browniano avaliado em um momento exponencialmente aleatório e distribuído, sua f.d.p.(função de densidade de probabilidade) é dada por:

$$f(x; \mu, \sigma) = \frac{1}{2\sigma} \exp\left(-\left|\frac{x-\mu}{\sigma}\right|\right), \quad -\infty < x < \infty \quad (2.8)$$

em que μ é um parâmetro de localização e $\sigma > 0$, o que é por vezes referida como

a diversidade, é um parâmetro de escala. Se $\mu = 0$ e $\sigma = 1$, a meia-linha positiva é exatamente uma distribuição exponencial dimensionada por $1/2$, a função de densidade de probabilidade da distribuição de Laplace é também uma lembrança da distribuição normal, no entanto, ao passo que a distribuição normal é expressa em termos da diferença ao quadrado da média μ , a densidade de Laplace é expressa em termos da diferença absoluta em relação à média. Por consequência, a distribuição de Laplace também tem caudas mais pesadas do que a distribuição normal (veja Figura 2.5).

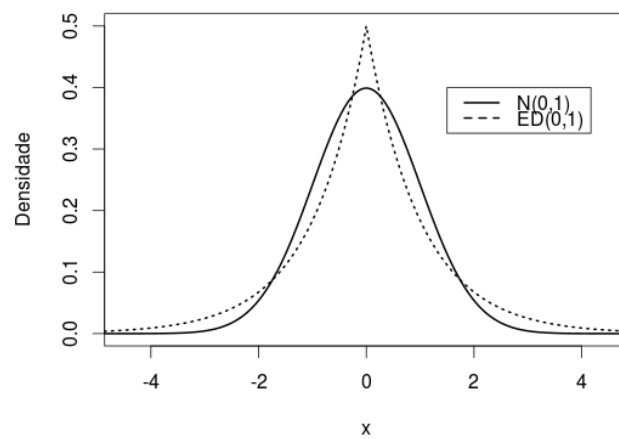


Figura 2.5: Gráfico comparando densidade da normal padrão com a densidade de Laplace.

Note que a distribuição de Laplace é não diferenciável em μ (zero, neste caso) por causa do módulo no expoente da função.

Distribuição Logística Tipo I

Dois grandes aplicações interessantes da distribuição logística são nas áreas de análise de sobrevivência e em modelagem de distribuição de renda. Segundo Sakanoue (2007) essa distribuição foi utilizada por Verhulst em 1838 para o ajuste de curvas de crescimento demográfico. Sua função densidade é expressa por:

$$f(x; \mu, \sigma) = \frac{c e^{-(x-\mu)^2/\sigma^2}}{\sigma^2 (1 + e^{-(x-\mu)^2/\sigma^2})^2}, \quad -\infty < x < \infty \quad (2.9)$$

em que $c \approx 1.484300027$ é uma constante normalizadora. Sua função de distribuição acumulada é frequentemente aplicada em teoria da regressão (regressão logística) e redes neurais. A Figura 2.6 a seguir faz uma projeção entre os gráficos da distribuição normal e a distribuição logística tipo I.

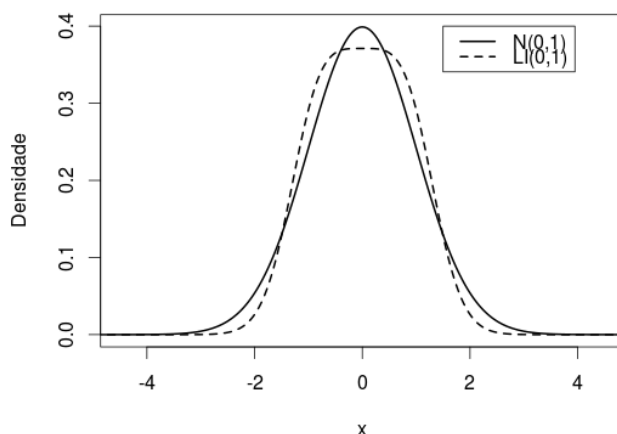


Figura 2.6: Gráfico comparando densidade da normal padrão com a densidade logística tipo I.

Distribuição Logística Tipo II

Uma grande diferença entre as distribuições logísticas é que, se X tem distribuição logística tipo I, então $-X$ tem distribuição logística tipo II. Sua função

de densidade é dada por:

$$f(x; \mu, \sigma) = \frac{1}{\sigma} \frac{e^{-(x-\mu)/\sigma}}{[1 + e^{-(x-\mu)/\sigma}]^2} \quad (2.10)$$

Graficamente, a distribuição também possui caudas mais pesadas do que a distribuição normal, conforme Figura 2.7 abaixo:

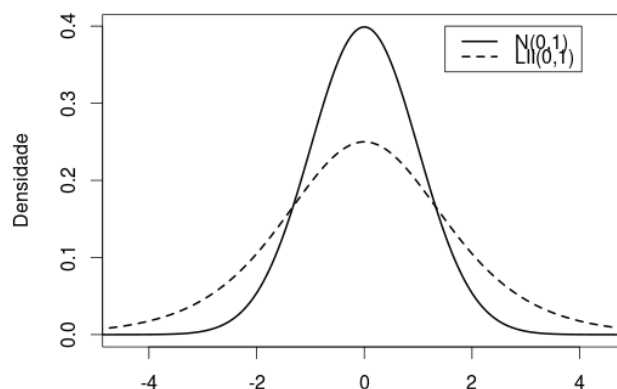


Figura 2.7: Gráfico comparando densidade da normal padrão com a densidade logística tipo II.

Distribuição t de Student

A variável aleatória $X \sim t(\mu, \phi, \nu)$ tem distribuição t de Student com ν graus de liberdade se sua função densidade $f(\cdot)$ é expressa por

$$f(x; \mu, \sigma, \nu) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left[1 + \left(\frac{x-\mu}{\sigma} \right)^2 / \nu \right]^{-\frac{n+1}{2}}, \quad \nu > 0,$$

em que $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$ é a função gama. A distribuição t é simétrica em torno de μ . Quando $\nu \rightarrow \infty$, a distribuição de t tende para a distribuição normal com média μ e variância σ^2 . Quando $\nu = 1$, a distribuição se reduz à distribuição Cauchy com parâmetros μ e σ . A distribuição t com ν graus de liberdade foi originada da razão $t_\nu = N\left(\frac{\chi_\nu^2}{\nu}\right)^{-1/2}$, em que N é a variável aleatória normal padrão

e χ_v^2 é uma variável aleatória qui-quadrado com ν graus de liberdade, sendo ambas independentes (LIMA-FILHO, 2009).

A distribuição t de Student é utilizada para modelar o comportamento de dados provenientes de uma distribuição com caudas mais pesadas que a normal, permitindo reduzir a influência de observações aberrantes.

LANGE *et al.* (1989) propõem o modelo t de Student como uma extensão paramétrica robusta do modelo normal, já que a t de Student é uma distribuição que permite ajustar a curtose da distribuição dos dados através do parâmetro ν .

Na Figura 2.8, observa-se a forma da distribuição t de Student para alguns valores de ν comparando-os com a distribuição normal padrão.

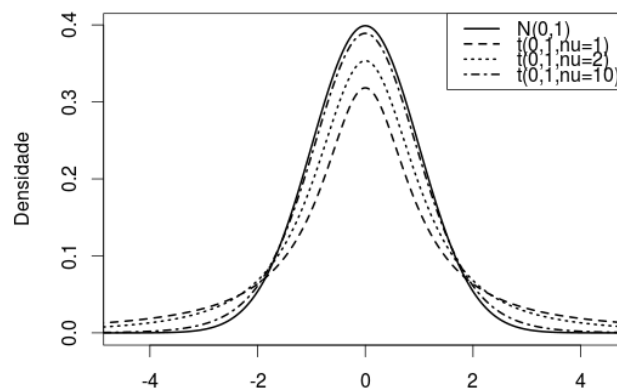


Figura 2.8: Gráfico comparando densidade da normal padrão versus a distribuição t para vários valores de graus de liberdade.

2.5 Testes de Ajustamento

Para verificar a qualidade do ajuste serão utilizados dois critérios, o **Erro Quadrático Médio** (EQM), definido da seguinte forma

$$EQM = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$$

e **Erro Médio Percentual Absoluto** (EMPA) definido por

$$EMPA = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{Y}_i - Y_i}{\hat{Y}_i} \right|$$

em que Y_i é a densidade pontual observada no histograma e \hat{Y}_i é a densidade predita pelo método implementado. Estes dois critérios devem apontar o melhor modelo pelo seu valor mínimo. Isto é, quanto menor o valor do EQM e do EMPA, melhor será o modelo.

3 Metodologia

Na próxima seção serão aplicadas as densidades da família (2.5), fazendo $\mu = 0$ e $\sigma^2 = 1$ no estimador de densidade por kernel (2.6) através da função $K(x)$. O melhor valor para h será estimado pela minimização da função (2.7) com o auxílio da técnica de integração monte carlo utilizando $m = 1000$ simulações. Para os ajustes serão utilizados três conjuntos de dados. O primeiro sendo proveniente de 50 valores gerados de uma distribuição $N(100,25)$, o segundo simulado a partir de 100 valores de uma distribuição $G(3,10)$ e o terceiro é um conjunto de dados reais obtidos de tempos de vida de 100 componentes de alumínio exposto a 31000 psi (BIRNBAUM; SAUNDERS, 1969). Os dados são mostrados no Apêndice A. Na Tabela 3.1 segue um resumo com média, variância, coeficiente de variação percentual, mínimo, mediana e máximo para cada um dos dados são mostrados.

Tabela 3.1: Algumas estatísticas descritivas para os dados utilizados no trabalho.

	\bar{x}	S^2	$CV\%$	Min	Med	Max
Dados 1	99,862	21,599	4,654	88,300	99,750	107,070
Dados 2	27,882	273,371	59,300	3,400	24,050	78,700
Dados 3	134,110	490,301	16,511	70,000	133,500	212,000

Para se escolher o melhor modelo, isto é, a função kernel que melhor explica os dados através do histograma, se utilizará os critérios EQM e EMPA, discutidos anteriormente.

Como para se utilizar a técnica de integração monte carlo é necessário se gerar valores pseudo-aleatórios, se optará pelo software R (R Development Core

Team, 2011) no qual estão implementados geradores de números aleatórios das distribuições normal, Cauchy, Laplace, logística tipo II e t de Student. Portanto, estas ser ao as distribuições testadas. No Apêndice B, são mostrados os códigos utilizados neste trabalho.

4 Resultados e Discussão

Para o conjunto de dados 1, simulados a partir da distribuição normal, minimizou-se a função (2.3), para os os núcleos das distribuições normal, Cauchy, Laplace e Logística tipo II. O valores ótimos de h , denotados por h^* e os respectivos valores de $\hat{J}(h^*)$ (mínimos da função $\hat{J}(h)$) são mostrados na Tabela 4.1.

Tabela 4.1: Valores ótimos de h e respectiva função $\hat{J}(h^*)$ para dados simulados a partir da distribuição normal.

Densidade	h^*	$\hat{J}(h^*)$
Normal	1,912	-0,056
Cauchy	2,818	-0,072
Laplace	3,120	-0,059
Logística II	3,724	-0,064

Na Figura 4 são mostrados os gráficos de $\hat{J}(h^*)$ versus h para os quatro núcleos distintos. Observe que o mínimo da função $\hat{J}(h)$ ocorre em h^* .

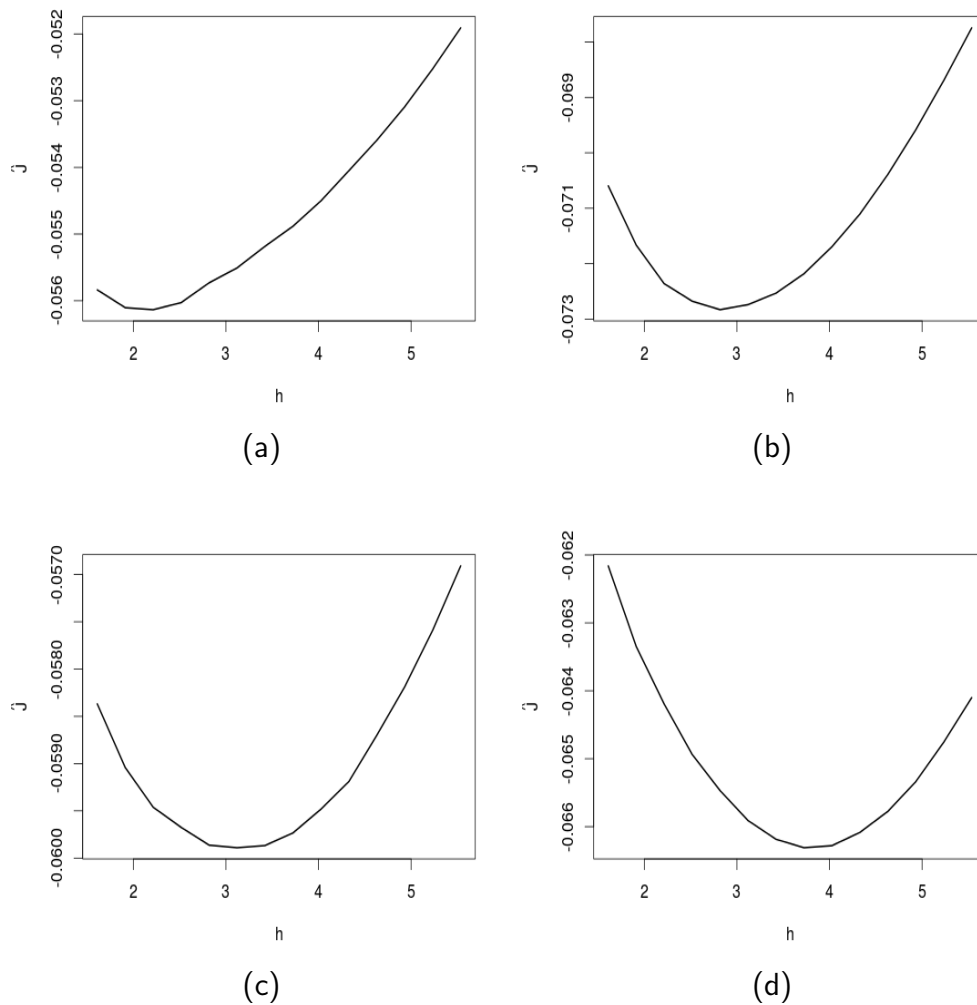


Figura 4.1: Gráficos de $\hat{J}(h^*)$ versus h para os núcleos normal (a), Cauchy (b), Laplace (c) e Logística tipo II (d) para dados simulados a partir da distribuição normal.

Na Figura 4.2 foram plotados conjuntamente o histograma dos dados e os ajustes por estimador kernel para K normal, Cauchy, Laplace e logística tipo II. Pelo gráfico, aparentemente o estimador com kernel normal tem um melhor ajuste, pois segue melhor as características do histograma.

Para se ter uma ideia mais acurada dos ajustes que a simples inspeção visual,

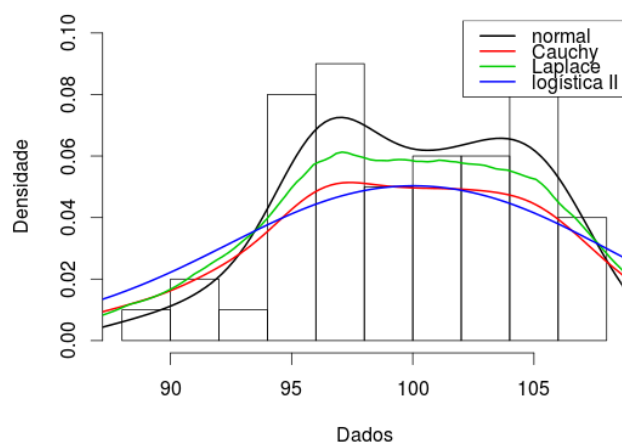


Figura 4.2: Histograma e ajustes por KDE.

utilizou-se os criterios EQM e EMPA. A seguir na Tabela 4.2 é possível se observar o EQM e EMPA de cada modelo para decidir qual dentre os núcleos melhor descreve o histograma dos dados. Os valores mínimos de EQM e EMPA são observados para o kernel normal, confirmando as suspeitas observadas indiretamente pelo gráfico.

Tabela 4.2: EQM e EMPA para núcleos normal, Cauchy, Laplace e logístico tipo II para o conjunto de dados 1

Densidade	EQM	EMPA
Normal	$3,83 \times 10^{-5}$	0,82
Cauchy	$9,75 \times 10^{-5}$	4,39
Laplace	$6,32 \times 10^{-5}$	1,16
Logística II	$1,17 \times 10^{-4}$	4,07

Para o conjunto de dados 2, simulados a partir da distribuição gama, minimizou-se a função (2.3), para os os núcleos das distribuições normal, Cauchy, Laplace e Logística tipo II. O valores ótimos de h , denotados por h^* e os respectivos valores de $\hat{J}(h^*)$ (mínimos da função $\hat{J}(h)$) são mostrados na Tabela 4.3.

Tabela 4.3: Valores ótimos de h e respectiva função $\hat{J}(h^*)$ para dados simulados a partir da distribuição gama.

Densidade	h^*	$\hat{J}(h^*)$
Normal	5,234	-0,018
Cauchy	7,348	-0,022
Laplace	7,348	-0,017
Logística II	8,859	-0,020

Na Figura 4 são mostrados os gráficos de $\hat{J}(h^*)$ versus h para os quatro núcleos distintos.

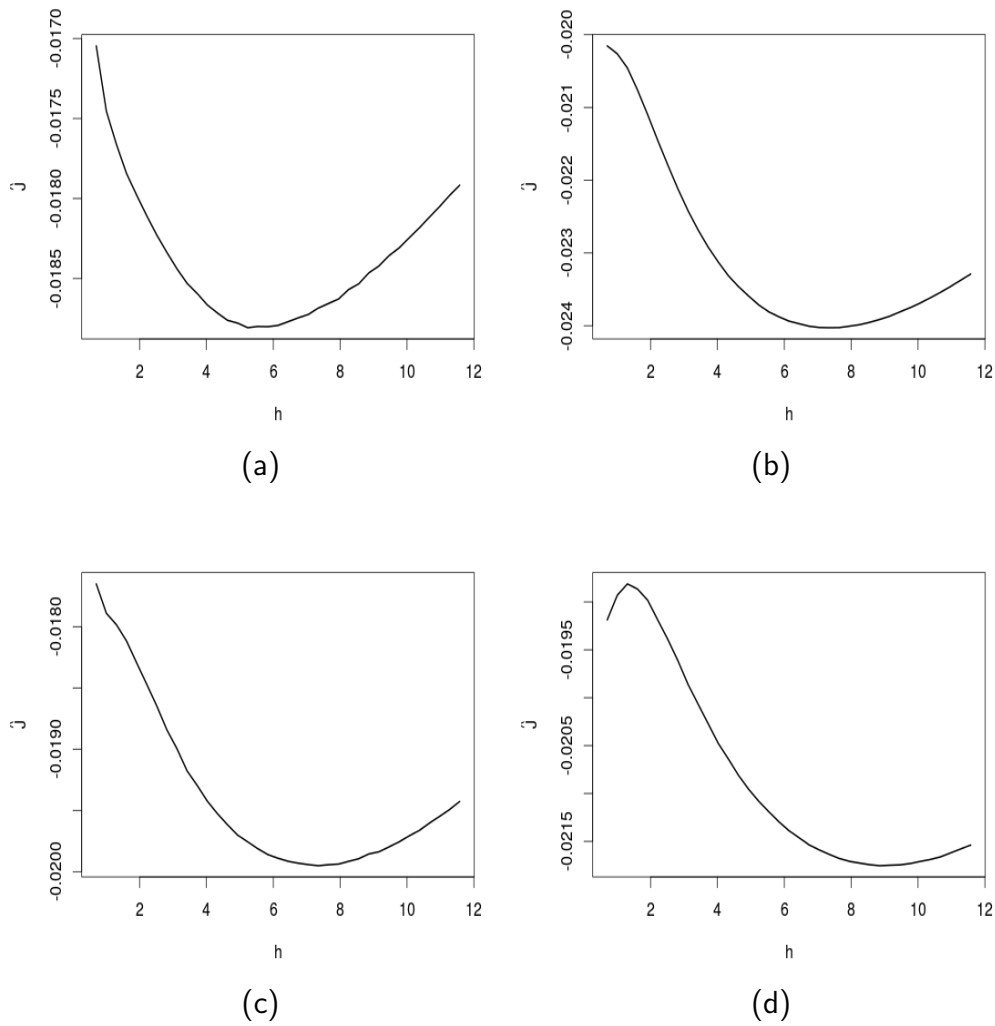


Figura 4.3: Gráficos de $\hat{J}(h^*)$ versus h para os núcleos normal (a), Cauchy (b), Laplace (c) e Logística tipo II (d) para dados simulados a partir da distribuição normal.

Na Figura 4.4 foram plotados conjuntamente o histograma dos dados simulados a partir da distribuição gama e os ajustes por estimador kernel para K normal, Cauchy, Laplace e logística tipo II. Pelo gráfico, aparentemente o estimador com kernel normal tem um melhor ajuste.

A seguir na Tabela 4.4 é possível se observar o EQM e EMPA de cada mo-

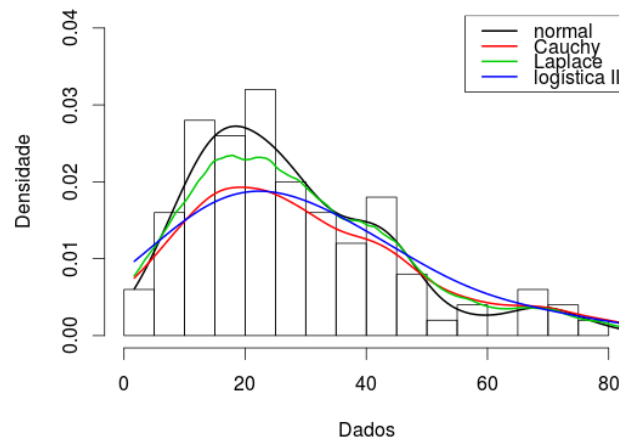


Figura 4.4: Histograma da distribuição gama e ajustes por KDE.

delo para decidir qual dentre os núcleos melhor descreve o histograma dos dados simulados a partir da distribuição gama. Os valores mínimos de EQM e EMPA são observados para o kernel normal, confirmando as suspeitas observadas indiretamente pelo gráfico.

Tabela 4.4: EQM e EMPA para núcleos normal, Cauchy, Laplace e logístico tipo II para o conjunto de dados 2

Densidade	EQM	EMPA
Normal	$1,22 \times 10^{-6}$	4,52
Cauchy	$4,45 \times 10^{-6}$	5,98
Laplace	$2,27 \times 10^{-6}$	5,01
Logística II	$5,14 \times 10^{-6}$	6,60

Para o conjunto de dados 3, obtidos de tempos de vida de 100 componentes de alumínio exposto a 31000 psi, minimizou-se a função (2.3), para os os núcleos das distribuições normal, Cauchy, Laplace e Logística tipo II. O valores ótimos de h , denotados por h^* e os respectivos valores de $\hat{J}(h^*)$ são mostrados na Tabela 4.5.

Tabela 4.5: Valores ótimos de h e respectiva função $\hat{J}(h^*)$ para dados reais.

Densidade	h^*	$\hat{J}(h^*)$
Normal	8,859	-0,0131
Cauchy	12,483	-0,0167
Laplace	13,388	-0,0138
Logística II	15,201	-0,0150

Na Figura 4 são mostrados os gráficos de $\hat{J}(h^*)$ versus h para os quatro núcleos distintos gerados a partir dos dados reais.

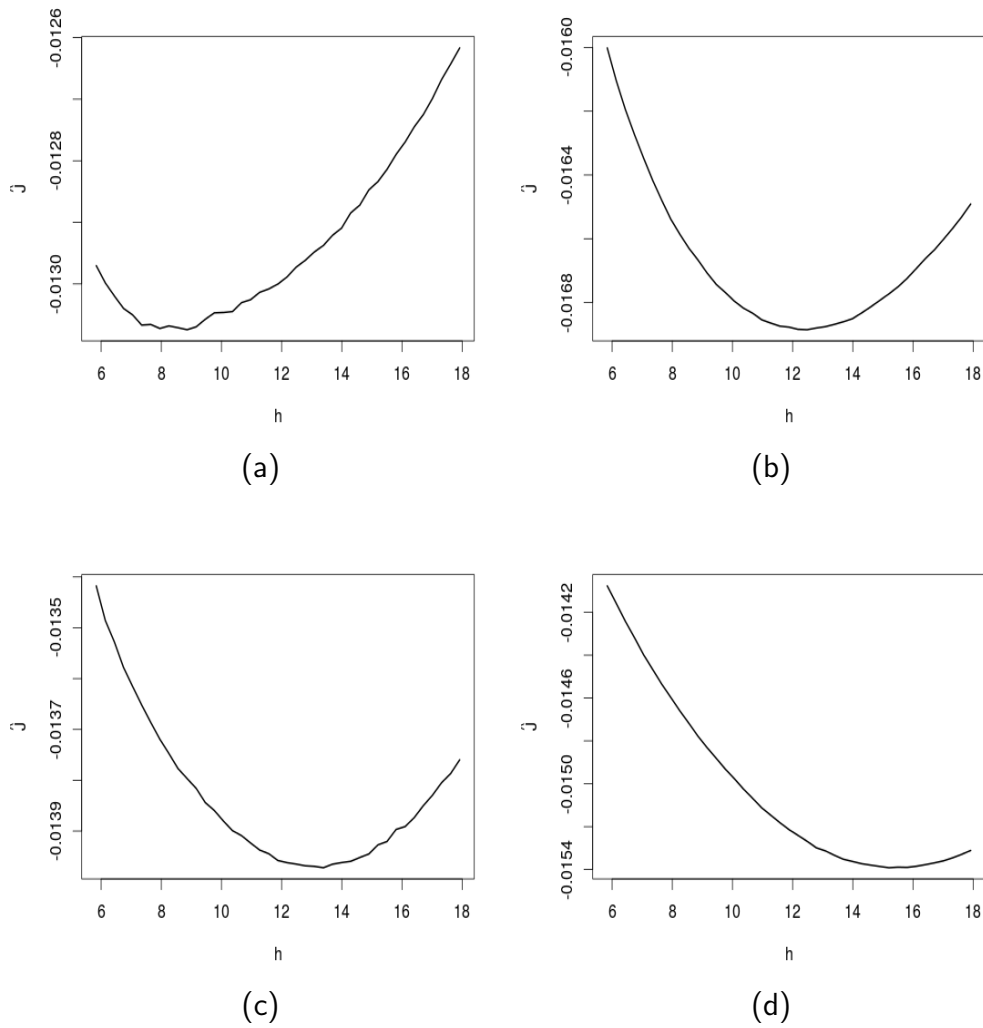


Figura 4.5: Gráficos de $\hat{J}(h^*)$ versus h para os núcleos normal (a), Cauchy (b), Laplace (c) e Logística tipo II (d) para dados reais.

Na Figura 4.6 foram plotados conjuntamente o histograma dos dados de tempos de vida de 100 componentes de alumínio exposto a 31000 psi e os ajustes por estimador kernel para K normal, Cauchy, Laplace e logística tipo II. Pelo gráfico, aparentemente o estimador com kernel normal tem um melhor ajuste.

A seguir na Tabela 4.6 é possível se observar o EQM e EMPA de cada mo-

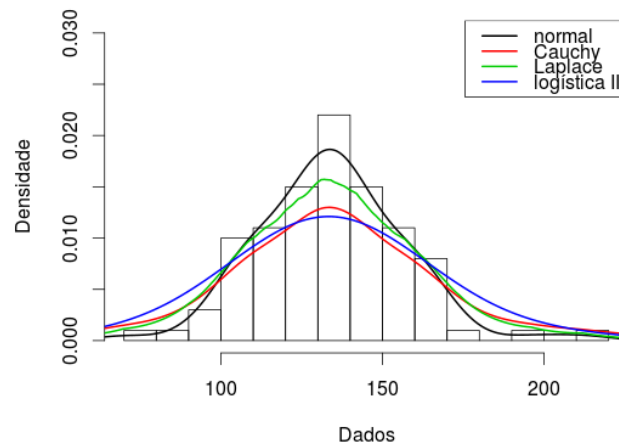


Figura 4.6: Histograma da distribuição gama e ajustes por KDE.

delo para decidir qual dentre os núcleos melhor descreve o histograma dos dados reais. Os valores mínimos de EQM e EMPA são observados para o kernel normal, confirmando as suspeitas observadas no gráfico.

Tabela 4.6: EQM e EMPA para núcleos normal, Cauchy, Laplace e logístico tipo II para o conjunto de dados reais.

Densidade	EQM	EMPA
Normal	$4,54 \times 10^{-5}$	5,20
Cauchy	$9,64 \times 10^{-5}$	8,03
Laplace	$6,23 \times 10^{-5}$	5,76
Logística II	$1,16 \times 10^{-4}$	9,61

Como resultados, verificou-se que dentre as distribuições simétricas utilizadas a distribuição normal foi aquela que mais se adequou aos dados tanto simulados quanto reais. A distribuição de Laplace foi aquela que mais se adequou aos dados depois da normal. Para todos os dados foi também testada a distribuição t de Student, a qual se esperava que pudesse competir com a normal, por se tratar de uma distribuição flexível, na qual se poderia ajustar os graus de liberdade antes de rodar o algoritmo principal. No entanto, o valor mínimo de $\hat{J}(h)$ ocorria sempre quando $h = 0$, o que não é permitido, uma vez que por restrição h deve ser positivo.

5 Conclusão

O presente trabalho abordou algumas distribuições simétricas que serviram de apoio para estudarmos um tipo de estimação pouco conhecido, no entanto, tem suas particularidades e benefícios no meio acadêmico. O estimador tipo Kernel nos mostra que podemos estimar um dado sem termos a priori uma referência que em nosso caso chamamos de parâmetro que é um valor extraído da população através de uma função estimativa. Varias distribuições utilizam o núcleo da estimação kernel para definição dos resultados, dentre elas se destaca a mais conhecida e usada, que é a distribuição normal, por outro lado, é preciso saber qual o melhor modelo a ser aplicado e isso é possível através de um ajuste que geralmente é obtido no uso de dois testes, o EQM (Erro Quadrático Médio) e o EMPA (Erro Médio Percentual Absoluto), estes quando calculado devem apontar o melhor modelo pelo seu valor mínimo, ou seja, quanto melhor o valor do EQM e do EMPA, melhor será o modelo. Como em quase tudo feito em estatística, a distribuição normal e a que tem maior uso e destaque, em nosso trabalho não foi diferente, através dos testes feitos anteriormente, vimos que de todas as distribuições aqui apresentadas, a normal foi aquela que melhor se ajustou aos dados aqui presentes.

Referências Bibliográficas

- BIRNBAUM, Z.; SAUNDERS, S. A new family of life distributions. *Journal of Applied Probability*, v. 6, p. 319–327, 1969.
- COMANICIU, D.; MEER, P. Mean shift: A robust approach towards feature space analysis. *IEEE Transactions on Patt*, v. 24, p. 603–619, 2002.
- CYSNEIROS, F. J.; PAULA, G. A.; GALEA, M. *Modelos Simétricos Aplicados*. [S.l.]: Associação Brasileira de Estatística., 2005.
- LANGE, K. L.; LITTLE, R. J. A.; TAYLOR, J. M. G. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, v. 84, p. 881–896, 1989.
- LIMA-FILHO, L. M. A. *Modelos Simétricos transformados não-lineares com diferentes distribuições dos erros: aplicações em ciências florestais*. Tese (Doutorado) — Universidade Federal Rural de Pernambuco, 2009.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2011. ISBN 3-900051-07-0. Disponível em: <<http://www.R-project.org/>>.
- RUDEMO, M. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, v. 9, p. 65–78, 1982.
- SAKANOUE, S. Extended logistic model for growth of single-species populations. *Ecol. Modelling*, v. 205, p. 159–168, 2007.
- WASSERMAN, L. *All of Statistics: a concise course in statistics inference*. 2. ed. Heidelberg: Springer, 2004.
- ZUCCHINI, W. *Applied Smoothing Techniques - Part 1: Kernel Density Estimation*. 2003. Disponível em: <<http://isc.temple.edu/economics/Econ616>>.

APÊNDICE

APÊNDICE A

Conjunto de dados 1 - 50 valores gerados de uma distribuição $N(100,25)$.

97,1 95,9 105,5 95,2 88,3 103,0 93,0 96,9 105,1 107,1 97,3 100,4 95,9 102,7 91,8
 99,9 94,9 101,1 97,7 107,7 94,4 103,8 99,6 97,2 96,9 91,0 98,0 105,0 104,2 99,5
 96,0 104,5 103,9 105,1 105,3 105,6 103,1 101,3 101,0 106,6 95,8 98,7 106,7 101,7 95,1
 102,4 101,9 96,2 99,2 96,9

Conjunto de dados 2 - 100 valores simulados de uma distribuição $G(3,10)$.

31,4 18,1 12,1 15,5 6,4 39,3 14,8 21,8 15,5 41,2 12,3 24,2 12,4 47,4 45,9 39,0 8,8 28,8
 18,0 33,7 4,4 41,6 3,4 39,2 32,9 24,5 29,1 27,3 17,1 22,5 8,4 21,7 69,1 9,8 19,1 24,0
 14,8 66,8 65,3 24,1 8,7 33,8 42,0 78,7 22,6 43,4 23,3 45,5 15,0 41,3 70,9 29,3 41,4 33,5
 23,7 11,7 15,0 14,2 45,7 56,4 57,4 37,6 14,3 38,2 3,4 8,0 28,5 50,5 30,4 18,7 8,6 20,5
 17,6 31,1 16,2 44,5 44,8 14,5 24,7 26,3 26,6 22,0 11,1 14,1 23,1 32,9 16,8 44,8 28,1 12,7
 17,8 16,9 23,4 21,3 71,7 6,6 28,5 18,0 38,4 29,8

Conjunto de dados 3 - dados reais obtidos de tempos de vida de 100 componentes de alumínio exposto a 31000 psi (BIRNBAUM; SAUNDERS, 1969)

70 90 97 99 100 103 104 104 105 107 108 108 108 109 109 112 112 113 114 114 114
 116 119 120 120 120 121 121 123 124 124 124 124 124 128 128 129 129 130 130 130
 131 131 131 131 131 132 132 132 133 134 134 134 134 134 136 136 137 138 138 138
 139 139 141 141 142 142 142 142 142 142 144 144 145 146 148 148 149 151 151 152
 155 156 157 157 157 157 158 159 162 163 163 164 166 166 168 170 174 196 212

APÊNDICE B

```
setwd('C:\\CAMINHO\\...')
```

```
#X=round(rnorm(50,100,5),1)
```

```
X=c(97.1, 95.9, 105.5, 95.2, 88.3, 103.0, 93.0, 96.9, 105.1, 107.1,
97.3, 100.4, 95.9, 102.7, 91.8, 99.9, 94.9, 101.1, 97.7, 107.7,
94.4, 103.8, 99.6, 97.2, 96.9, 91.0, 98.0, 105.0, 104.2, 99.5,
96.0, 104.5, 103.9, 105.1, 105.3, 105.6, 103.1, 101.3, 101.0, 106.6,
95.8, 98.7, 106.7, 101.7, 95.1, 102.4, 101.9, 96.2, 99.2, 96.9)
```

```
round(c(mean(X),var(X),sd(X)/mean(X)*100,min(X),median(X),max(X)),3)
```

```
#X=round(rgamma(100,3,1/10),1)
```

```
X=c(31.4,18.1,12.1,15.5,6.4,39.3,14.8,21.8,15.5,41.2,12.3,24.2,12.4,
47.4,45.9,39.0,8.8,28.8,18.0,33.7, 4.4,41.6,3.4,39.2,32.9,24.5,29.1,
27.3,17.1,22.5,8.4,21.7,69.1,9.8,19.1,24.0,14.8,66.8,65.3,24.1, 8.7,
33.8,42.0,78.7,22.6,43.4,23.3,45.5,15.0,41.3,70.9,29.3,41.4,33.5,
23.7,11.7,15.0,14.2,45.7,56.4,57.4,37.6,14.3,38.2,3.4,8.0,28.5,
50.5,30.4,18.7,8.6,20.5,17.6,31.1,16.2,44.5,44.8,14.5,24.7,26.3,
26.6,22.0,11.1,14.1,23.1,32.9,16.8,44.8,28.1,12.7,17.8,16.9,23.4,
21.3,71.7,6.6,28.5,18.0,38.4,29.8)
```

```
round(c(mean(X),var(X),sd(X)/mean(X)*100,min(X),median(X),max(X)),3)
```

```
#Tempos de vida de componentes de aluminio exposto a 31000 psi
```

```
#Birnbaum,Z.W.and Saunders, S.C.()1969a)
```

```

X=c(70,90,97,99,100,103,104,104,105,
    107,108,108,108,109,109,112,112,113,
    114,114,114,116,119,120,120,120,121,
    121,123,124,124,124,124,124,128,128,
    129,129,130,130,130,131,131,131,131,
    131,132,132,132,133,134,134,134,134,
    134,136,136,137,138,138,138,139,139,
    141,141,142,142,142,142,142,142,144,
    144,145,146,148,148,149,151,151,152,
    155,156,157,157,157,157,158,159,162,
    163,163,164,166,166,168,170,174,196,
    212)
round(c(mean(X),var(X),sd(X)/mean(X)*100,min(X),median(X),max(X)),3)

source('kernelmontecarlo.txt')

kernelmontecarlo(X,1000,30,6,16,1)
kernelmontecarlo(X,1000,30,6,16,2)
kernelmontecarlo(X,1000,30,6,16,3)
kernelmontecarlo(X,1000,30,6,16,4)
#kernelmontecarlo(X,1000,30,1,12,5)

```

O código a seguir deve ser colocado num arquivo chamado 'kernelmontecarlo.txt':

```
kernelmontecarlo=function(X,m,bmax,a,b,flag){
```

```

n=length(X)

K=function(x) 1/sqrt(2*pi)*exp(-x^2/2)

f=function(x){
  1/n*sum(1/h*K((x-X)/h))
}

#----- Fun??o Monte Carlo -----#

Integral=function(h){
  Int=S=matrix(0,n,n)

  for(i in 1:n){
    for(j in 1:n){

      #m=1000

      # Normal
      if(flag==1){
        #x=rnorm(m,X[j],h)
        x=X[j]+h*rnorm(m,0,1)

        H=function(x){
          u=(x-X[i])/h
          1/sqrt(2*pi*h^2)*sum(exp(-u^2/2))
        }
      }
    }
  }
}

```

```
}
```

```
}
```

```
# Cauchy
```

```
if(flag==2){
```

```
x=rcauchy(m,X[j],h)
```

```
H=function(x){
```

```
u=(x-X[i])/h
```

```
1/(pi*h*(1+u^2))
```

```
}
```

```
}
```

```
# Laplace
```

```
if(flag==3){
```

```
v=runif(m, -.5, .5)
```

```
x=X[j] - h*sign(v)*log(1-2*abs(v))
```

```
H=function(x){
```

```
u=(x-X[i])/h
```

```
1/(2*h)*exp(-abs(u))
```

```
}
```

```
}
```

```
# Log?stica II
```

```
if(flag==4){
x=rlogis(m,X[j],h)

H=function(x){
u=(x-X[i])/h
1/h * exp(u) / ( (1+exp(u))^2 )
}
}

df=50

# t-Student
if(flag==5){
x=X[j]+h*rt(m,df)

H=function(x){
u=(x-X[i])/h
gamma((df+1)/2)/(sqrt(df*pi)*gamma(df/2))*(1+u^2/df)^(-(df+1)/2)
}
}

aa=3
# Gama
if(flag==6){
x=X[j]+rgamma(m,aa,h)
```

```

H=function(x){
u=(x-X[i])/h
u^(aa-1)*(1/h)^aa*exp(-u)/gamma(aa)
}
}

Int[i,j] = 1/m*sum(H(x)) / (n^2)

if(i!=j) S[i,j]=K((X[i]-X[j])/h)
}}
Int=sum(Int)
S=2/(n*(n-1)*h)*sum(S)

return(Int-S)
}
#-----#

Erro=0
h=seq(.1,bmax,l=100)
a=max(100*a/bmax,1)
bmax=10*bmax
b=min(1000*b/bmax,100)
for(i in 1:length(h)) Erro[i]=Integral(h[i])
plot(h[a:b],Erro[a:b],type='l',xlab='h',ylab=expression(hat(J)),lwd=2)
#return(a)
hh=h[a:b]

```



```
Er=Erro[a:b]
hh=min(hh[Er==min(Er)])
print(hh)
print(Er[hh])
}
```