Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Estimação da Entropia Diferencial pelo Método do Kernel de Caudas Pesadas

por

Carolina Gomes Casado de Carvalho

Brasília, 30 de Julho de 2025

Estimação da Entropia Diferencial pelo Método do Kernel de Caudas Pesadas

por

Carolina Gomes Casado de Carvalho

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília, como requisito parcial para obtenção do título de Mestre em Estatística.

Orientador: Raul Yukihiro Matsushita

Brasília, 30 de Julho de 2025

Dissertação submetida ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade de Brasília como parte dos requisitos para a obtenção do título de Mestre em Estatística.

Texto aprovado por:

Prof. Dr. Raul Yukihiro Matsushita
Orientador, PPGEST/UnB

Profa. Dra. Regina Célia Bueno da Fonseca
Membro titular, PPGTGS/IFG

Prof. Dr. Felipe Sousa Quintino
Membro titular, PPGEST/UnB

Prof. Dr. Roberto Vila Gabriel
Suplente, PPGEST/UnB

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

A entropia diferencial é uma medida essencial para quantificar a incerteza associada a variáveis aleatórias absolutamente contínuas. Sua estimação é particularmente desafiadora quando a distribuição populacional apresenta caudas pesadas, assimetria ou múltiplos modos. Este trabalho investiga estimadores de entropia baseados em métodos não paramétricos, com ênfase na estimativa de densidade por kernel (KDE) utilizando funções núcleo de caudas pesadas. Destacam-se dois kernels: o t de *Student*, com graus de liberdade ajustáveis, e o kernel de Pareto simétrico, que incorpora explicitamente um parâmetro de forma que controla a cauda da densidade. São realizados estudos de simulação de Monte Carlo para avaliar o desempenho desses estimadores em diferentes cenários de caudas. Os resultados mostram que o uso de kernels com caudas pesadas, em especial o kernel de Pareto, permite maior estabilidade e precisão na estimação da entropia, mesmo em amostras pequenas. A aplicação empírica com dados financeiros ilustra o potencial do método para identificar padrões de incerteza e estruturas informacionais em séries temporais.

Palavras-chave: entropia diferencial; estimação de densidade por kernel; caudas pesadas; kernel t-Student; kernel de Pareto simétrico.

Abstract

Differential entropy is a fundamental measure to quantify uncertainty in continuous random variables. Its estimation becomes particularly challenging in heavy-tailed, asymmetric, or multimodal distributions. This study explores nonparametric entropy estimators based on kernel density estimation (KDE), focusing on heavy-tailed kernels. Two kernels are considered: the Student's t kernel, with adjustable degrees of freedom, and the symmetric Pareto kernel, which explicitly incorporates a tail parameter to control density behavior. Monte Carlo simulation studies are conducted to evaluate the performance of these estimators across various tail scenarios. Results indicate that heavy-tailed kernels—especially the Pareto kernel—enhance the accuracy and stability of entropy estimates, even with small samples. An empirical application using financial return data demonstrates the method's capacity to capture uncertainty and informational structure in time series data.

Keywords: differential entropy; kernel density estimation; heavy tails; Student's t kernel; symmetric Pareto kernel.

Sumário

Abreviações e Siglas

CV	validação cruzada
FD	função densidade de probabilidade
FDA	função de distribuição acumulada
GL	graus de liberdade
KDE	estimação de densidade por kernel (*kernel density estimation*)
KL	divergência de Kullback–Leibler
MC	simulação de Monte Carlo
MSE	erro quadrático médio
MISE	erro quadrático médio integrado
NND	método do vizinho mais próximo (*nearest neighbor distance*)
R	linguagem de programação R
SPK	kernel de Pareto simétrico (*symmetric Pareto kernel*)
SS	espaçamento amostral (*sample spacing*)
VAR	variância

Lista de Símbolos e Notações

α	parâmetro de forma do kernel de Pareto
a	parâmetro de forma da distribuição de Pareto
b	parâmetro de escala
d	dimensão populacional
E	esperança
Eff	eficiência
EffR	eficiência relativa
$f(x)$	função de densidade de probabilidade da variável X
$\hat{f}_h(x)$	estimativa da função de densidade de probabilidade ($f(x)$)
$F(x)$	função de distribuição acumulada (FDA)
H	entropia diferencial
$\hat{H}$	estimativa da entropia diferencial
h	parâmetro de suavização (largura de banda)
K	função kernel
κ^2	integral do quadrado do kernel (2.6)
n	tamanho da amostra
ψ	função digama
ρ_i	menor distância entre x_i e os demais pontos (usado em NND)
U	variável aleatória uniforme em $[0, 1]$
X	variável aleatória
γ	constante de Euler-Mascheroni
bias	viés da estimativa da entropia
Var	variância

Capítulo 1

Introdução

1.1 Considerações iniciais

A entropia diferencial é uma medida fundamental para quantificar a incerteza associada a variáveis aleatórias contínuas. Sua estimativa representa um desafio, especialmente quando a função de densidade de probabilidade subjacente $f(x)$ apresenta características como multimodalidade, assimetria, caudas pesadas e multidimensionalidade.

Em muitos problemas práticos, a forma funcional de $f(x)$ é desconhecida, sendo objeto do próprio estudo estatístico. Isso justifica a utilização de métodos não paramétricos que buscam construir uma aproximação da densidade a partir da amostra observada. Dentre esses métodos, destaca-se o estimador de densidade por kernel (*Kernel Density Estimation* – KDE) (Rosenblatt, 1956; Parzen, 1962a), o qual permite a construção de estimadores suaves e adaptáveis de $f(x)$.

Tradicionalmente, o método KDE envolve a escolha de dois componentes fundamentais, a função kernel $K(\cdot)$ e a largura de banda h, que controla o grau de suavização. Matsushita et al. (2024) introduziram um terceiro componente, o índice de cauda da função kernel. Isso porque a acurácia do estimador de entropia diferencial calculado a partir de uma densidade estimada por KDE depende dessas escolhas, sendo especialmente sensível nas regiões de baixa densidade, como nas caudas da distribuição. E essa acurácia é sensível ao peso da cauda da distribuição

populacional (Hall, 1987; Hall et al., 1993b; Matsushita et al., 2024).

Distribuições com aspectos de caudas pesadas são comuns em dados financeiros e ambientais. Nessas situações, as funções kernel tradicionais, como a gaussiana e a quadrática, tendem a subestimar a probabilidade de observações extremas, comprometendo a qualidade da estimativa da entropia. Por esse motivo, funções kernel com caudas pesadas têm sido propostas, como o kernel *t-Student* (Hall et al., 1993b) e, mais recentemente, o de Pareto simétrico centrado em zero (Matsushita et al., 2024).

Hall et al. (1993b) discutem que a eficiência estatística do estimador da entropia, do tipo *plug-in* (baseada em métodos não paramétricos, como histogramas ou funções kernel), é fortemente influenciada pelo comportamento das caudas da distribuição subjacente e pela própria escolha da função kernel. Essencialmente, eles explicam que isso ocorre por causa da sensibilidade do logaritmo da densidade a regiões de baixas frequências. Por isso, esses autores recomendam a aplicação de uma função kernel que também possua caudas pesadas — como o kernel t de *Student* — pois eles conferem maior estabilidade às estimativas ao atribuir maior peso a observações extremas. Sua forma é dada pela expressão

$$K(u) = \frac{1}{\sqrt{\nu}\,\mathrm{B}\left(\frac{1}{2}, \frac{\nu}{2}\right)} \left(1 + \frac{u^2}{\nu}\right)^{-(\nu+1)/2}, \tag{1.1}$$

na qual $u \in \mathbb{R}$, B denota a função Beta, e $\nu > 0$ representa o número de graus de liberdade. À medida que ν aumenta, o peso da cauda diminui, aproximando-se da gaussiana. Porém, Hall et al. (1993b) não tratam do uso simultâneo de h e ν como dois parâmetros livres para a estimação da função de densidade. Eles assumem o kernel com ν fixado, e ajustam apenas a largura de banda (h) com base na amostra dos dados.

Já o kernel de Pareto introduzido por Matsushita et al. (2024) apresenta um hiperparâmetro adicional de forma α que regula explicitamente a cauda da função kernel com base na amostra, conferindo maior flexibilidade ao estimador. Assim, essa proposta permite o ajuste da cauda do kernel aos dados observados, o que é particularmente vantajoso na estimação da en-

tropia diferencial em distribuições com observações extremas. Além disso, esse kernel admite a estimação conjunta de α e h por validação cruzada, minimizando a distância de Kullback-Leibler entre $f(x)$ e sua estimativa $\hat{f}(x)$ para todo x. Como resultado desse processo, o estimador da entropia diferencial coincide com estimador de momentos da entropia cruzada $H(f||\hat{f}) = \int \ln \hat{f}(x) f(x) dx$, ou seja, $\hat{H}(f) = \sum_{i=1}^{n} \ln \hat{f}(X_i)/n$.

Neste trabalho, como uma complementação do trabalho de Matsushita et al. (2024), propõe-se aplicar o kernel t de *Student* para a estimação da entropia diferencial, considerando seu número de graus de liberdade ν como um hiperparâmetro adicional a ser ajustado com base nos dados, além da habitual largura de banda (h) encontrada no método KDE.

1.2 Objetivos

O objetivo principal é estudar as propriedades do estimador da entropia diferencial de distribuições univariadas com distribuição contínua, utilizando a função kernel de caudas pesadas como a de Pareto simétrico e a t de *Student*.

Os objetivos específicos são:

i) Revisar os conceitos teóricos da estimação de densidade por kernel, incluindo propriedades, escolhas de kernel e métodos de seleção da largura de banda.

ii) Apresentar estimadores clássicos da entropia diferencial no caso univariado.

iii) Revisitar a construção do kernel de Pareto simétrico, suas propriedades e seu uso na estimação da entropia (Matsushita et al., 2024);

iv) Desenvolver estudos de Monte Carlo para avaliar o desempenho do KDE com o kernel t de *Student*.

v) Desenvolver aplicação voltada à modelagem de dados financeiros, com foco na detecção de regimes de cauda em séries temporais, utilizando estimadores de entropia diferencial baseados em kernels de caudas pesadas.

1.3 Perguntas da pesquisa

Com respeito a esse último objetivo específico, as perguntas dessa investigação sob a perspectiva metodológica são:

- O KDE com kernel t de *Student* produz resultados satisfatórios para amostras pequenas (*por exemplo*, $n = 30$), assim como no caso do kernel de Pareto (Matsushita et al., 2024)?
- Caso a pergunta anterior seja negativa, qual seria o tamanho mínimo da amostra para alcançá-los?
- Há correspondência entre os hiperparâmetros ν do kernel t de *Student* com o expoente α do kernel de Pareto?
- A estimativa do hiperparâmetro ν propiciaria *insights* em uma análise de dados financeiros?

1.4 Dados

Para responder às questões propostas, foram utilizados tanto dados sintéticos quanto dados reais de séries financeiras.

No estudo com dados sintéticos, revisita-se a construção dos kernels baseados nas distribuições Pareto simétrica centrada em zero e *t*-Student, explorando suas propriedades e aplicabilidade na estimação da entropia diferencial. A distribuição Pareto simétrica é parametrizada por um parâmetro de forma $\alpha > 0$, responsável pelo grau de cauda, e por um parâmetro de escala $b > 0$, fixado em $b = 1$ em todos os experimentos. Já a distribuição *t*-Student é controlada pelo parâmetro de forma $\nu > 0$, que regula a espessura das caudas, e pelo parâmetro de suavização $h > 0$, associado à largura de banda do estimador.

A geração das amostras foi realizada por meio da inversão das funções de distribuição acumulada de ambas as distribuições, cobrindo diferentes regimes de cauda, leves e pesadas, por

Programa 1.1: Exemplo de coleta de dados do Yahoo Finance: obtenção dos retornos diários do fechamento do Ibovespa, de 1/jan/1998 a 31/jan/2025

```
# ======================================================
# Get daily data from Yahoo Finance
# ======================================================
if(!require(quantmod)){
  install.packages("quantmod")
  library(quantmod)
}
dfbvsp <- as.data.frame(getSymbols('^BVSP',periodicity='daily',
          from='1998-01-01',to='2025-01-31'))
dfbvsp <- data.frame(dia = as.Date(time(BVSP)), ibov = BVSP$BVSP.Close)
X      <- data.frame(day.x = dfbvsp$dia, x = as.numeric(dfbvsp$BVSP.Close))
X      <- na.omit(X)
x.var  <- diff(log(X$x)) # Retorno do Ibovespa
```

meio da variação dos parâmetros α e ν. Foram considerados tamanhos amostrais de $n = 30, 60$ e 200, com 500 replicações para cada cenário. Essa configuração permite uma avaliação empírica detalhada das propriedades estatísticas dos estimadores, como viés, variância e erro quadrático médio (MSE).

Além do estudo com dados simulados, foi feita uma aplicação prática com séries temporais reais de ativos listados na B3 (Bolsa de Valores do Brasil). Os dados de preços diários foram coletados via *web scraping* na plataforma Yahoo! Finance, utilizando o pacote `quantmod` do software R. O processo de coleta e cálculo dos retornos logarítmicos diários do índice Ibovespa se encontra exemplificado no Programa 1.1.

1.5 Esboço do trabalho

Esta dissertação está estruturada da seguinte forma:

- O Capítulo 2 apresenta os fundamentos da estimação de densidade por kernel no caso univariado, suas propriedades estatísticas e critérios de escolha do parâmetro de suavização.

- O Capítulo 3 trata da estimação da entropia diferencial univariada, com revisão de méto-

dos clássicos e detalhamento do estimador baseado no kernel de Pareto.

- O Capítulo 4 é dedicado à condução dos estudos de simulação. Investiga o comportamento do estimador de entropia usando o kernel de Pareto e do kernel t de Student.
- O Capítulo 5 traz uma aplicação da metodologia proposta, utilizando séries temporais financeiras de empresas listadas na B3, com o objetivo de identificar regimes de cauda ao longo do tempo.
- Por fim, o Capítulo 6 apresenta as conclusões do trabalho, com reflexões sobre os resultados obtidos.
- Embora este trabalho se limite ao caso univariado, o Anexo A aborda sobre a extensão da KDE ao caso multivariado, incluindo uma breve descrição de suas propriedades.

Capítulo 2

Estimação de densidade por kernel (KDE)

2.1 Introdução

O estimador de uma função de densidade constitui ferramenta fundamental no âmbito da análise estatística, permitindo descrever e evidenciar características relevantes da distribuição de probabilidade subjacente a um conjunto de dados. Nesse sentido, a escolha criteriosa da abordagem de estimação revela-se imprescindível (Beirlant et al., 1997).

Dado um conjunto de observações considerado como uma amostra proveniente de uma função de densidade de probabilidade cuja forma funcional seja desconhecida, o objetivo central consiste em estimar tal função de densidade a partir dessas informações amostrais. No âmbito da estimação de densidade, destacam-se duas abordagens principais: a paramétrica e a não paramétrica (Silverman, 2018).

A abordagem paramétrica fundamenta-se na hipótese de que os dados seguem determinada distribuição de probabilidade previamente conhecida, havendo, portanto, uma forma funcional específica para a função de densidade que depende de parâmetros desconhecidos. Dessa forma, as estimativas desses parâmetros podem ser obtidos por meio de métodos como o da máxima verossimilhança ou o dos momentos. Embora essa abordagem seja eficiente sob suposições adequadas, seu desempenho será insatisfatório caso a forma real da distribuição difira daquela

assumida.

Por outro lado, a abordagem não paramétrica não impõe restrições quanto à forma da distribuição subjacente, oferecendo, portanto, maior flexibilidade ao permitir que a estrutura da função de densidade seja determinada diretamente pelos dados. Dentre os métodos não paramétricos, destaca-se o histograma, considerado o estimador de densidade mais antigo e conceitualmente simples. Nele, o espaço amostral é particionado em intervalos, e a frequência de observações em cada intervalo é observada. Apesar de sua simplicidade e fácil implementação, o histograma apresenta limitações importantes, como a dependência do resultado em relação à amplitude dos intervalos de classe e a descontinuidade nas bordas dos mesmos. Essas limitações fazem com que o histrograma seja um estimador ineficiente da densidade populacional (Beirlant et al., 1997).

Visando proporcionar formas suavizadas, contínuas e aderentes às características da distribuição subjacente aos dados observados, entre outros métodos, destacam-se os estimadores de densidade via funções kernel. Esses estimadores são formulados sob uma perspectiva não paramétrica, construídos a partir da soma de funções de ponderação, denominadas kernels, atribuídas a cada observação da amostra. Cada observação recebe um peso calculado com base em sua distância em relação ao ponto onde se deseja estimar a densidade. A abordagem envolve centralizar cada ponto x onde se deseja estimar a densidade, utilizando uma janela h que define a vizinhança de x e os pontos amostrais que contribuem para a estimação (Silverman, 2018). Introduzido por Rosenblatt (1956), e posteriormente refinado por Parzen (1962a), esse procedimento confere maior flexibilidade à modelagem da distribuição de probabilidade, permitindo que a estimativa se ajuste de maneira eficiente a distintos padrões e estruturas presentes nos dados, inclusive em casos de distribuições assimétricas, multimodais e de caudas pesadas.

Formalmente, dada uma amostra aleatória simples $X_1, X_2, ..., X_n$ retirada de uma população com densidade desconhecida $f(x)$, define-se o estimador da densidade de probabilidade no

ponto $x \in \mathbb{R}$ como

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

em que X_i é a amostra de dados $i = 1, 2...$, K é a *função kernel*, $h > 0$ é o parâmetro de suavização (ou largura de banda) que permite controlar o grau de suavização da curva estimada, e n é o número de observações na amostra. A função kernel atua como um suavizador local, atribuindo maior peso às observações x_i mais próximas do ponto x em que se deseja estimar a densidade. O kernel K deve satisfazer algumas propriedades fundamentais, que garantem que ela possa ser interpretada como uma densidade de probabilidade com média nula. Ou seja, para todo $u \in \mathbb{R}$, exige-se

$$\begin{aligned} &K(u) \geq 0, \\ &K(-u) = K(u), \\ &\int_{-\infty}^{+\infty} K(u)\, du = 1, \\ &\int_{-\infty}^{+\infty} uK(u)\, du = 0. \end{aligned} \tag{2.2}$$

Portanto, a função kernel apresenta as mesmas propriedades de uma função de densidade de uma variável aleatória simétrica em torno de zero.

Além da função densidade $f(x)$, também é possível estimar a função de distribuição acumulada da variável aleatória a partir do mesmo princípio de suavização. A função de distribuição acumulada estimada por kernel é expressa por

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{(x-X_i)/h} K(u)du, \tag{2.3}$$

em que a integral representa a contribuição acumulada da função kernel centrada em cada ponto x_i, após escalonamento pela largura de banda h.

2.2 Propriedades para o caso de caudas leves

A avaliação da eficiência estatística desses estimadores requer o exame criterioso de diversas propriedades teóricas, as quais são determinantes para a compreensão de seu desempenho em diferentes contextos amostrais.

O vício associado ao estimador de densidade via kernel quantifica a discrepância entre o valor esperado da estimativa $\hat{f}(x)$ e o valor verdadeiro da função de densidade $f(x)$. Uma aproximação para esse vício é dada por (Silverman, 2018)

$$E[\hat{f}(x)] - f(x) \approx \frac{h^2}{2} f^{(2)}(x) \sigma_K^2, \tag{2.4}$$

no qual h representa o parâmetro de suavização (largura de banda), $f^{(2)}(x)$ é a segunda derivada da função de densidade $f(x)$, e $\sigma_K^2 = \int_{\mathbb{R}} y^2 K(y)\, dy$ é uma medida relacionada à função kernel K.

Nota-se que o vício é influenciado simultaneamente pelo parâmetro h, pela concavidade ($f^{(2)}(x) > 0$) ou convexidade ($f^{(2)}(x) < 0$) da função densidade $f(x)$, sendo $f^{(2)}(x)$ a *segunda derivada de* $f(x)$, e pela escolha da função kernel. Como $h > 0$ e $\sigma_K^2 > 0$, o vício é nulo apenas se $f^{(2)}(x) = 0$, ou seja, em regiões localmente lineares da densidade.

Quanto à variância do estimador kernel de $f(x)$, $Var[\hat{f}(x)]$, ela pode ser obtida com base na aproximação de ordem zero da função de densidade, para uma amostra suficientemente grande (Silverman, 2018) dada por

$$\text{Var}[\hat{f}(x)] \approx \frac{1}{nh} f(x) \kappa^2, \tag{2.5}$$

sendo que

$$\kappa^2 = \int_{\mathbb{R}} K^2(y)\, dy. \tag{2.6}$$

Assim, tanto σ_k^2 quanto κ^2 atuam como indicadores da qualidade da função de kernel. Nessas expressões para $E[\hat{f}(x)]$ e $Var[\hat{f}(x)]$, observa-se que a redução de h tende a diminuir o vício, mas ao custo de aumento da variância do estimador, estabelecendo um dilema clássico entre precisão e variabilidade. Por outro lado, o aumento do tamanho amostral n contribui para a redução da variância, o que implica uma relação indireta entre o vício e n.

Há diversas funções de densidade simétricas em torno de zero que podem desempenhar o papel de função de kernel. Dentre elas, as que comumente são utilizadas para essa finalidade estão destacadas na Tabela 2.1. Em geral, essas diferentes funções geralmente produzem resultados similares, sugerindo que a escolha da função $K(u)$ pode não ser tão crucial quanto a determinação do parâmetro de suavização h (Silverman, 2018; Mammen et al., 1992). No entanto, para aplicações inferenciais ou análises mais refinadas, pode ser necessário considerar funções kernel assimétricas ou de caudas pesadas, para capturarem melhor as características específicas dos dados (Hall, 1987; Matsushita et al., 2024).

Tabela 2.1: Exemplos de funções kernel

nomes	kernel	suporte
retangular	$K(u) = \frac{1}{2}$	$\lvert u\rvert \leq 1$
triangular	$K(u) = (1 - \lvert u\rvert)$	$\lvert u\rvert \leq 1$
Epanechnikov	$K(u) = \frac{3}{4}(1 - u^2)$	$\lvert u\rvert \leq 1$
logística	$K(u) = \frac{1}{2+\mathrm{e}^u + \mathrm{e}^{-u}}$	$u \in \mathbb{R}$
sigmoidal	$K(u) = \frac{2}{\pi(\mathrm{e}^u + \mathrm{e}^{-u})}$	$u \in \mathbb{R}$
gaussiano	$K(u) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-u^2/2}$	$u \in \mathbb{R}$

A eficiência de uma função kernel K pode ser formalmente definida como:

$$\text{Eff}(K) = \kappa^2\sqrt{\sigma_K^2}, \tag{2.7}$$

em que $\sigma_K^2 = \int_{\mathbb{R}} y^2 K(y)\,dy$ e κ^2 definida conforme (2.6). Já a eficiência relativa entre duas

funções de kernel K_1 e K_2 é dada por

$$\text{EffR}(K_1, K_2) = \frac{\text{Eff}(K_1)}{\text{Eff}(K_2)}. \tag{2.8}$$

Assim, a qualidade de uma função kernel pode ser medida com base em σ_K^2 e κ^2, e sua eficiência é definida como o produto dessas medidas. Dentre as funções kernel exemplificadas na Tabela 2.1, a mais eficiente é o de Epanechnikov (Tabela 2.2). No entanto, essa métrica não pode ser utilizada para funções kernel de cauda pesada com $\sigma_K^2 = \infty$. Para essa situação, Hall (1987) e Matsushita et al. (2024) sugerem a utilização da distância de Kullback-Leibler e do erro quadrático médio como medidas alternativas.

Tabela 2.2: Eficiências das funções kernel da Tabela 2.1.

kernels	σ_K^2	κ^2	Eff(K)	EffR(K_1, K)
Epanechnikov	$\frac{1}{5}$	$\frac{3}{5}$	0,2683	100,0%
triangular	$\frac{1}{6}$	$\frac{2}{3}$	0,2722	98,6%
gaussiano	1	$\frac{1}{2\sqrt{\pi}}$	0,2821	95,1%
retangular	$\frac{1}{3}$	$\frac{1}{2}$	0,2887	92,9%
logística	$\frac{\pi^2}{3}$	$\frac{1}{6}$	0,3023	88,7%
sigmoidal	$\frac{\pi^2}{4}$	$\frac{2}{\pi^2}$	0,3183	84,3%

2.3 Determinação do parâmetro de suavização

A seleção apropriada do valor para o parâmetro h é uma consideração crítica, tanto do ponto de vista descritivo quanto inferencial, merecendo uma atenção especial. Uma abordagem comum consiste em minimizar uma função de perda que quantifique a diferença entre a estimativa $\hat{f}(x)$ e a densidade real $f(x)$. Uma métrica clássica nesse contexto para o desempenho local do estimador é o erro quadrático médio (MSE - Mean Squared Error),

$$\text{MSE}[\hat{f}, f] = E[(\hat{f}(x) - f(x)]^2, \tag{2.9}$$

que pode ser decomposto em termos de variância (2.5) e vício (2.4) como

$$\text{MSE}[\hat{f}, f] = Var[\hat{f}(x)] + E[f(x), \hat{f}(x)], \tag{2.10}$$

no qual

$$E[f(x), \hat{f}(x)] = \{E[\hat{f}(x)] - f(x)\}^2 \tag{2.11}$$

representa o vício do estimador. Assim das equações (2.5) e (2.4), o erro quadrático médio (MSE) é dada por

$$\text{MSE}[\hat{f}, f] \approx \frac{1}{nh} f(x)\kappa^2 + \frac{h^4\sigma_K^4}{4}[f^{(2)}(x)]^2, \tag{2.12}$$

em que $\sigma_K^2 = \int_{\mathbb{R}} y^2 K(y)dy$ e $\kappa^2 = \int_{\mathbb{R}} K^2(y)dy$. Outra medida de erro utilizada para medir o desempeho global do estiomador é o erro quadrático médio integrado (MISE - Mean Integrated Squared Error), definido por

$$\text{MISE}[\hat{f}, f] = \int_{-\infty}^{+\infty} E[\hat{f}(x) - f(x)]^2 dx. \tag{2.13}$$

Substituindo a Eq. (2.11) no integrando da Eq. (2.12) obtém-se a aproximação correspondente

$$\text{MISE}[\hat{f}, f] \approx \frac{\kappa^2}{nh} + \frac{h^4\sigma_K^4}{4}\int_{-\infty}^{+\infty} [f^{(2)}(x)]^2 dx. \tag{2.14}$$

Para obter o valor ótimo estimado d parâmetro h, minimiza-se o MISE (Eq. (2.14)), ou seja,

$$\hat{h} = \arg\min_{h>0} \text{MISE}[f]. \tag{2.15}$$

Para isso, derivando MISE$[f]$ em relação a h e iguala a zero, obtendo

$$\frac{d}{dh}\text{MISE}[\hat{f}, f] \approx \frac{\kappa^2}{nh^2} + 4h^3\sigma_K^4 \int_{\mathbb{R}} [f^{(2)}(x)]^2 dx = 0, \tag{2.16}$$

da qual resulta

$$\hat{h}^5 = \frac{-\kappa^2}{n\sigma_K^4 \int_{\mathbb{R}} [f^{(2)}(x)]^2 dx}. \tag{2.17}$$

No entanto, esse resultado depende de $f^{(2)}$, que de fato é desconhecido. Na prática, existem algumas regras empíricas que podem ser úteis como ponto de partida para a escolha de h. Para densidade e kernel gaussianos, Silverman (2018) sugere $h = 0,9\min(s, IQ/1,34)^{-1/5}$, em que IQ representa o intervalo interquartílico e s é o desvio padrão amostral. Já Terrell et al. (1985) propõem $h \leq 2,44sn^{-1/5}$ para o kernel quadrático, e no caso de dados gaussianos, $h = 2,34sn^{-1/5}$.

Outra abordagem relevante é o método de validação cruzada, que visa encontrar o valor de h que minimize a estimativa da MISE com base nos próprios dados (Bowman, 1984; Silverman, 1986) :

$$\hat{h} = \arg\min_{h>0} \left\{ \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} (K * K)\left(\frac{X_i - X_j}{h}\right) - \frac{1}{n} \sum_{i=1}^{n} \hat{f}_i(X_i) \right\}, \tag{2.18}$$

em que $\hat{f}_i(X_i)$ denota a estimativa de densidade obtida excluindo a i-ésima observação, e

$$(K * K)\left(\frac{X_i - X_j}{h}\right) = \int_{\mathbb{R}} K(u) K\left(\frac{X_i - X_j}{h} - u\right) du \tag{2.19}$$

representa uma convolução entre kernels.

2.4 Estimação de densidades com caudas pesadas

Uma variável aleatória X com distribuição de caudas pesada é caracterizada por um decaimento lento da função de densidade à medida que seu quantil x se afasta da posição central. Isso implica probabilidades maiores de ocorrência de eventos extremos em comparação com distribuições com caudas leves, como a normal e a exponencial. Além disso, a partir de alguma ordem r_0, o decaimento lento da cauda produz integrais divergentes no cálculo de momentos de ordens $r > r_0$. Como exemplos desse tipo de distribuições, encontram-se a distribuição t de Student, a de Pareto, e a de Cauchy. As distribuições de caudas pesadas são comuns em diversas áreas aplicadas, como finanças e climatologia (Embrechts et al., 1997; Resnick, 2007) .

A estimação da função de densidade de probabilidade em contextos com caudas pesadas merece atenção especial, pois a qualidade dos resultados dependem da escolha da função kernel apropriada (Hall et al., 1993b; Matsushita et al., 2024). Nesse contexto, o risco de Kullback-Leibler (KL) tem sido útil para o processo de estimação, visto que ele mede o desvio ao se substituir a densidade verdadeira $f(x)$ pela sua estimativa $\hat{f}(x)$ (Hall, 1987). Embora os estimadores de kernel sejam flexíveis e não exijam suposições sobre a forma da distribuição, seu desempenho tende a piorar nas regiões da distribuição onde há poucos dados, justamente as caudas da distribuição, sendo muito sensíveis à escolha do núcleo e, especialmente, da largura de banda (Hall, 1987; Hall et al., 1993b).

Hall (1987) explora de forma teórica como o risco de Kullback-Leibler (KL) se comporta em cenário não paramétricos, com ênfase no desempenho dos estimadores kernel. Essa medida é natural quando se deseja estimar a *entropia diferencial*, uma vez que ambas envolvem integrais com o logaritmo da densidade.

A divergência KL entre a densidade real f e sua estimativa $\hat{f}$ é dada por

$$\mathrm{KL}(f \parallel \hat{f}) = \int f(x) \log \frac{f(x)}{\hat{f}(x)}\, dx. \tag{2.20}$$

De fato, $\mathrm{KL}(f \parallel \hat{f})$ — que mede a discrepância informacional entre duas distribuições — é sensível a erros em regiões onde $f(x)$ é pequeno (Hall, 1987). No contexto do método de kernel, no qual a estimativa $\hat{f}(x)$ é construída a partir de uma média ponderada local das observações, tal sensibilidade torna-se crítica em distribuições de cauda pesada (Matsushita et al., 2024). Como a entropia diferencial também é definida por uma integral que envolve o logaritmo da densidade,

$$H(f) = -\int f(x) \log f(x)\, dx, \tag{2.21}$$

fica claro que análise do risco de KL está diretamente relacionada à precisão da estimativa da entropia diferencial quando substituímos $f(x)$ por$\hat{f}(x)$. Essa substituição produz uma medida denominada *entropia cruzada*.

Ao contrário da versão discreta da entropia — entropia de Shannon (Cover et al., 2006) —, a entropia diferencial não mede a quantidade absoluta de informação, mas sim a densidade local de incerteza. Cover et al. (2006) ressaltam que essa medida assume valores mais altos quanto mais dispersa for a densidade e pode inclusive ser negativa, dependendo da concentração da distribuição.

Hall (1987) mostra que o risco de KL é sensível a erros na estimativa de $\hat{f}(x)$ em regiões onde $f(x)$ é pequena, como ocorre nas caudas. Nesses pontos, o termo logarítmico amplifica os erros, o que pode comprometer de maneira desproporcional a qualidade da estimativa global da entropia. Além disso, ele estabelece que, sob hipóteses regulares, a esperança da divergência de KL pode ser decomposta em termos do viés e variância,

$$\mathrm{E}\left[\mathrm{KL}(f \parallel \hat{f})\right] = b\left[f(\mathrm{x}), \hat{f}(\mathrm{x})\right] + \mathrm{Var}\left[\hat{f}(\mathrm{x})\right] + o(1), \tag{2.22}$$

em que o vício e a variância são influenciados por três fatores principais: a suavidade da função f, seu comportamento nas caudas e a escolha da largura de banda h. A escolha usual $h \sim n^{-1/5}$

pode ser inadequada em regiões nos quais $f(x) \to 0$, como é típico em distribuições com caudas pesadas, muitas vezes descritas por $f(x) \sim |x|^{-a}$, com $a > 1$. Nesses casos, o erro nas extremidades pode dominar o risco total, expondo os limites do uso tradicional do método de kernel. Esse comportamento foi observado experimentalmente por Matsushita et al. (2024), comparando-se os resultados proporcionados pelas funções kernel de Laplace, Cauchy, Pareto e o gaussiano.

Apesar dessas dificuldades, algumas estratégias podem ser empregadas para aprimorar o desempenho de estimadores de densidade em distribuições com caudas pesadas. Uma delas refere-se à escolha adaptativa da largura de banda. Hall (1987) argumenta que utilizar uma largura de banda constante ao longo de todo o domínio da variável pode ser inadequado, especialmente nas extremidades da distribuição. Nesses pontos, o uso de larguras de banda localmente ajustadas, geralmente mais amplas, ajuda a suavizar oscilações causadas pela escassez de dados, reduzindo o erro relativo da estimativa.

Outra alternativa consiste na aplicação de transformações de variáveis como o logaritmo ou o inverso, para reduzir a assimetria e suavizar as caudas. Isso pode tornar a densidade mais regular nas extremidades e facilitar o uso do estimador kernel, com ganhos tanto na estimação de $f(x)$ quanto na entropia diferencial.

Além disso, quando os dados extremos são muito escassos ou ruidosos, a censura ou truncamento controlado pode ser uma alternativa. Isso envolve o truncamento da amostra além de certo ponto, com a modelagem paramétrica da cauda baseada em distribuições adequadas (como Pareto ou t de Student). Essa abordagem semiparamétrica mantém a flexibilidade na parte central da distribuição, ao mesmo tempo em que lida melhor com as caudas.

Essas estratégias são discutidas e comparadas por Madukaife et al. (2024), que conduziram uma extensa comparação entre estimadores de entropia diferencial, incluindo métodos baseados em kernel. Destacam que, em distribuições com caudas pesadas, a performance desses estimadores depende fortemente da adoção de técnicas adaptativas, como largura de banda variável e transformações que suavizem o comportamento das caudas.

Hall et al. (1993b) e Matsushita et al. (2024) discutem que a escolha apropriada da função kernel também seria uma alternativa possível. Enquanto Hall et al. (1993b) sugere o kernel t de Student, Matsushita et al. (2024) propõem o kernel de Pareto, que oferece maior flexibilidade e precisão em cenários como a análise de dados financeiros. O kernel de Pareto possui um parâmetro adicional de suavização, o expoente de Pareto, permitindo a adaptação entre distribuições de caudas leves e pesadas. A novidade dessa abordagem é que o índice de Pareto da função Kernel permite estabelecer uma inferência qualitativa sobre o tipo de cauda (pesada, leve, muito leve) da distribuição dos dados, sem necessidade de grandes amostras.

Capítulo 3

Estimação da Entropia Diferencial

3.1 Introdução

A entropia diferencial é uma generalização contínua do conceito clássico de entropia de Shannon (Shannon, 1948). Enquanto a entropia discreta quantifica a incerteza associada a uma variável aleatória discreta, a entropia diferencial aplica-se a variáveis aleatórias contínuas, funcionando como uma medida de dispersão ou imprevisibilidade da densidade de probabilidade.

Seja X uma variável aleatória contínua com função densidade de probabilidade (f.d.p.) $f(x)$, a entropia diferencial $H(X)$ é definida como:

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx, \tag{3.1}$$

desde que a integral exista e seja finita. Para isso, é necessário que $f(x) \log f(x)$ seja absolutamente integrável, ou seja, $f(x) \log f(x) \in L^1(\mathbb{R})$. O logaritmo é geralmente tomado na base e, com a entropia expressa em *nats*, embora também se use a base 2, resultando em *bits*.

Diferentemente da entropia discreta, que é sempre não negativa, a entropia diferencial pode assumir valores negativos. Isso ocorre porque, em distribuições altamente concentradas (como a normal com pequena variância), os valores de $\log f(x)$ podem ser suficientemente grandes para tornar a integral negativa.

Além disso, a entropia diferencial não é invariante a transformações de escala, o que a distingue da entropia de Shannon. Por exemplo, para $Y = aX$, com $a \in \mathbb{R} \setminus \{0\}$, tem-se:

$$H(Y) = H(X) + \log |a|. \tag{3.2}$$

Isso revela que a entropia depende da unidade de medida usada para X, refletindo sua natureza densitária.

Na teoria da informação, a entropia diferencial é interpretada como a quantidade média de informação contida na observação de uma variável contínua. No entanto, como argumentado por Cover et al. (2006), essa interpretação deve ser feita com cautela. A entropia diferencial não corresponde diretamente à quantidade de informação transmitida por um canal, pois a noção de "bits por símbolo"não é bem definida no contexto contínuo.

3.2 Estimação

Como a densidade $f(x)$ da variável contínua X é geralmente desconhecida, torna-se necessário construir estimadores baseados em amostras finitas. Diversas abordagens têm sido propostas para esse fim, que podem ser agrupadas em quatro classes principais: estimadores baseados em histogramas, funções kernel, vizinhos mais próximos (*k*-NN) e métodos bayesianos ou espectrais (Silverman, 1986; Devroye et al., 1985; Beirlant et al., 1997).

Os primeiros métodos foram baseados em histogramas, que particionam o suporte da variável em subintervalos (ou *bins*) e estimam a densidade de forma constante em cada intervalo. Embora simples, essa abordagem é sensível à escolha do número de bins, o que pode introduzir viés significativo.

Para contornar essas limitações, Vasicek (1976) propôs um estimador baseado em estatísticas de ordem, denominado espaçamentos amostrais (*sampling-spacing*) de ordem m, que pode ser interpretado como uma suavização local das distâncias entre amostras ordenadas. Ele se

define como (Beirlant et al., 1997)

$$\hat{H}_{SS(m)} = \frac{1}{n}\sum_{i=1}^{n} \log\Big(\frac{n}{m}(X_{(i+m)} - X_{(i-m)})\Big) - \psi(m) + \ln m, \tag{3.3}$$

em que m é um parâmetro de janela e $X_{(i)}$ representa a i-ésima estatística de ordem e ψ é a função digamma. Este estimador é assintoticamente consistente, apresentando boas propriedades estatisticas sob diferentes cenários (Beirlant et al., 1997; Matsushita et al., 2024).

Outra classe de estimadores são os baseados em vizinhos mais próximos, que utilizam a distância entre pontos amostrais para inferir a densidade local. Ele se define como (Kozachenko et al., 1987; Beirlant et al., 1997) como

$$\hat{H}_{NND} = \frac{1}{n}\sum_{i=1}^{n} \ln(n\rho_{n,i}) + \log 2 + \gamma, \tag{3.4}$$

em que

$$\rho_{n,i} = \min_{j\neq i, j\leq n} ||X_i - X_j||$$

representa a distância de vizinho mais próximo entre X_i e X_j, e $\gamma = -\int_0^\infty e^{-t}\ln t dt$ é a constante de Euler. Esse estimador também é consistente e não requer suposições explícitas sobre a forma de $f(x)$ (Beirlant et al., 1997). Singh et al. (2003) investigaram sua eficiência e estenderam sua aplicação a dados multivariados.

Os estimadores bayesianos modelam $f(x)$ como uma variável aleatória, incorporando conhecimento a priori sobre sua forma. O estimador NSB (I.Nemenman, F.Shafee, W. Bialek), proposto por Nemenman et al. (2002), é robusto a amostras pequenas e foi adaptado ao caso contínuo via quantização e regularização.

Já Paninski (2003) analisaram abordagens espectrais, como a projeção da densidade sobre bases de Fourier, e mostraram que não existem estimadores imparciais da entropia diferencial sem restrições fortes sobre $f(x)$.

Estudos comparativos, como o de Holmes et al. (2019), demonstraram que estimadores baseados em k-NN tendem a superar os métodos baseados em histogramas e kernel, especialmente em amostras pequenas e distribuições multimodais. A escolha do método ideal, entretanto, depende de fatores como o tamanho da amostra, conhecimento prévio sobre a densidade e restrições computacionais.

No entanto, para situações de caudas pesadas, Matsushita et al. (2024) observaram que - embora apresentem vícios baixos, os estimadores k-NN apresentam maior variância e erro quadrático médio em comparação com o método sampling-spacing de primeira ordem e o estimador da entropia por meio de densidade por kernel (KDE).

O estimador da entropia diferencial baseados em densidade kernel (KDE) é definido como

$$\hat{H}_K = -\int \hat{f}_h(x) \log \hat{f}_h(x)\, dx, \tag{3.5}$$

em que

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{3.6}$$

Moon et al. (1995) demonstraram que a escolha do parâmetro de suavização h afeta significativamente o desempenho do estimador. Beirlant et al. (1997) apresentaram uma revisão crítica das propriedades assintóticas desses métodos.

3.3 O kernel de Pareto

No âmbito da estimação da entropia diferencial para distribuições de caudas pesadas, a estimação da entropia diferencial por meio do método KDE depende fortemente da escolha da função kernel (Hall et al., 1993b; Matsushita et al., 2024). A entropia é bastante sensível a erros em regiões onde a densidade é baixa. Por isso, a seleção do kernel se torna ainda mais relevante em distribuições com caudas pesadas. Funções kernel tradicionais, como a Gaussiana, assumem caudas leves e simétricas, o que pode comprometer a qualidade da estimativa justamente onde

há maior ocorrência de observações extremas.

Nesse contexto, destaca-se o *kernel de Pareto*, uma função núcleo inspirada na distribuição de Pareto simétrica centrada em zero, amplamente conhecida por modelar fenômenos com caudas longas, como rendas, riscos e eventos extremos (Matsushita et al., 2024). Em métodos de estimação kernel, cada ponto da amostra contribui com uma função centrada sobre si, e a escolha da forma do kernel determina como as observações mais distantes influenciam a estimativa. O kernel de Pareto se diferencia por atribuir maior peso a observações afastadas, sendo controlado por um parâmetro de cauda α, que regula a taxa de decaimento da função.

A forma típica do kernel de Pareto pode ser expressa como:

$$\kappa_\alpha(u) = \frac{\alpha}{2}\left(1 + |u|^{\alpha+1}\right)^{-1},$$

na qual u representa a distância padronizada entre o ponto avaliado e os dados, e $\alpha > 0$ determina o comportamento da cauda. Valores pequenos de α resultam em uma cauda mais pesada, atribuindo maior influência a observações extremas. Já valores grandes fazem o kernel decair rapidamente, concentrando a massa em torno da média. Essa estrutura permite que o kernel seja adaptável tanto a distribuições de caudas leves ou pesadas.

Do ponto de vista prático, essa flexibilidade é fundamental para contextos nos quais os dados apresentam comportamento turbulento, como no caso de distribuições com transições abruptas de regime de caudas, que são frequentes em séries temporais financeiras e fenômenos climáticos, por exemplo. Além disso, ao controlar simultaneamente a suavização e a influência de valores extremos, esse kernel se mostra promissor na estimação da entropia diferencial, justamente por atuar com mais precisão nas regiões de baixa densidade onde os erros impactam mais fortemente o valor do funcional.

Os parâmetros α e a largura de banda h são otimizados por validação cruzada, minimizando-se a entropia cruzada. Nesse processo, estima-se a entropia diferencial com base na função

densidade kernel $\hat{f}_h(x)$, dada por:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \kappa_\alpha \left(\frac{x - X_i}{h} \right),$$

onde $h > 0$ é o parâmetro de suavização (largura de banda) e κ_α é a função kernel parametrizada pelo expoente de Pareto (α). A função κ_α é parametrizada de forma que o comportamento da cauda varie de acordo com o valor de α, podendo assumir cauda leve quando $\alpha > 2$ ou cauda pesada quando $\alpha \leq 2$.

A entropia é então estimada por meio do estimador *plug-in*:

$$\hat{H}(\hat{f}_h) = -\frac{1}{n} \sum_{j=1}^{n} \ln \hat{f}_h(X_j),$$

o qual corresponde à entropia cruzada entre a densidade verdadeira f e a estimada $\hat{f}_h$. Como esse estimador tende a superestimar $H(f)$, utiliza-se a função de perda por validação cruzada, definida por:

$$CV(h, \alpha) = -\frac{1}{n} \sum_{j=1}^{n} \ln \left[\frac{1}{(n-1)h} \sum_{i \neq j} \kappa_\alpha \left(\frac{X_j - X_i}{h} \right) \right],$$

e os parâmetros ótimos são obtidos resolvendo-se:

$$(\hat{h}, \hat{\alpha}) = \arg \min_{h>0, \alpha>0} CV(h, \alpha).$$

Esse procedimento torna o estimador sensível à estrutura empírica dos dados, ajustando simultaneamente a suavização e o comportamento da cauda. Como consequência, o método é capaz de adaptar-se a situações com alta variabilidade, reduzindo viés e variância em regiões críticas da densidade.

Adicionalmente, a função kernel de Pareto é suficientemente geral para incluir, como casos particulares, situações de cauda leve, pesada e até mesmo comportamentos quase determinísticos (quando $\alpha \to \infty$, aproximando-se da função delta de Dirac). Essa versatilidade torna o

estimador particularmente útil em aplicações onde a estrutura de cauda não é conhecida inicialmente, permitindo uma abordagem orientada pelos dados.

3.4 O kernel t de *Student*

Além do kernel de Pareto, outro candidato para a estimação da entropia diferencial em distribuições com caudas pesadas é o kernel baseado na distribuição t de *Student*. Essa proposta foi originalmente apresentada por Hall et al. (1993b), que sugeriu o uso de kernels com caudas mais pesadas do que as do kernel Gaussiano clássico para lidar com situações em que os dados apresentam maior propensão a valores extremos.

A densidade da distribuição t de *Student*, com $\nu > 0$ graus de liberdade, é dada por:

$$\kappa_\nu(u) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{u^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

onde $u \in \mathbb{R}$ e $\Gamma(\cdot)$ denota a função gama. Essa função kernel é simétrica, com cauda controlada pelo parâmetro ν. Para valores pequenos de ν, a cauda é mais pesada; à medida que $\nu \to \infty$, o kernel converge para o kernel Gaussiano.

A densidade estimada com base no kernel t de *Student* é dada por:

$$\hat{f}_{h,\nu}(x) = \frac{1}{nh} \sum_{i=1}^{n} \kappa_\nu \left(\frac{x - X_i}{h}\right),$$

em que $h > 0$ é o parâmetro de suavização e ν é o parâmetro de cauda.

Nesta dissertação, propõe-se uma extensão ao procedimento tradicional de estimação, estimando simultaneamente os dois parâmetros (h, ν).

A vantagem dessa abordagem está na flexibilidade adicional oferecida pela estimativa de ν, permitindo que o kernel se adapte dinamicamente à estrutura de cauda dos dados observados. Tal como o kernel de Pareto, o kernel t de *Student* oferece robustez em regiões de baixa densidade, essenciais para a estimação precisa da entropia diferencial.

Capítulo 4

Simulações

4.1 Introdução

Para a estimação da entropia diferencial, Hall et al. (1993b) propuseram a função kernel de distribuições de caudas pesadas, Hall et al. (1993b) propuseram a utilização do kernel *t-Student*, enquanto Matsushita et al. (2024) introduziram o kernel de Pareto. No kernel *t-Student* com ν graus de liberdade, o parâmetro de suavização h controla também o efeito da cauda da distribuição, já que seu parâmetro de escala se relaciona com ν. Já no kernel de Pareto, há dois parâmetros: o de suavização e o índice de Pareto (α) que ajusta a cauda da distribuição. Nesse caso, embora o hiperparâmetro α não indique diretamente o valor do índice caudal da distribuição populacional, ele permite avaliação qualitativa sobre o tipo de cauda da distribuição populacional, se pesada ou leve (Matsushita et al., 2024).

Os estudos de Monte Carlo efetuados por Matsushita et al. (2024) foram realizados sobre uma população *t-Student*. Neste capítulo apresentamos uma replicação desse estudo de Monte Carlo para estimação da entropia diferencial de uma distribuição de Pareto simétrica, mas considerando a própria distribuição de Pareto simétrica centrada em zero como modelo populacional.

Será proposto também implementação do estimador de entropia diferencial baseado na função kernel da distribuição *t-Student*. Este kernel é controlado por dois parâmetros: o grau de

liberdade ν, que ajusta a espessura da cauda, e o parâmetro de suavização h, que regula a largura do kernel.

Uma das propriedades interessantes desse kernel é que, à medida que $\nu \to \infty$, ele converge para o kernel gaussiano, tornando-se mais leve. Assim, o kernel *t-Student* oferece flexibilidade ao se adaptar tanto a distribuições com caudas pesadas quanto a distribuições mais regulares, dependendo da escolha de ν.

4.2 A distribuição de Pareto simétrica em zero

Define-se a função de densidade acumulada (FDA) de uma distribuição de Pareto simétrica centrada em 0 (X) como

$$F(x) = \mathrm{P}[X \leq x] = \begin{cases} \frac{1}{2}\left(\frac{b}{b-x}\right)^a, & \text{para } x < 0; \\ 1 - \frac{1}{2}\left(\frac{b}{b+x}\right)^a, & \text{para } x \geq 0, \end{cases} \tag{4.1}$$

em que $b > 0$ é o parâmetro de escala e $a > 0$ denota o parâmetro de forma (índice de Pareto). A função de densidade de probabilidade (PDF) associada a essa distribuição é dada por

$$f(x) = \frac{a}{2}\left(\frac{b}{b+|x|}\right)^{a+1}, \quad \text{para } x \in \mathbb{R}.$$

O índice de Pareto a controla a cauda da distribuição, sendo mais pesada à medida que a diminui. O peso da cauda reflete nos momentos de ordem r centrados na origem da distribuição, de modo que $\mathrm{E}(X^r) = \infty$ para $a \leq r$.

4.2.1 Entropia

Partindo da definição da entropia diferencial,

$$H(X) = -\int_{-\infty}^{\infty} f(x) \ln f(x)\, dx,$$

devido à simetria da função densidade da distribuição de Pareto simétrica, a integral que define a entropia diferencial teórica pode ser reescrita de forma mais simples como

$$H(X) = -2\int_0^\infty f(x)\ln f(x)\,dx.$$

Para $x \geq 0$, substituímos a densidade:

$$f(x) = \frac{a}{2}\left(\frac{b}{b+x}\right)^{a+1},$$

$$\ln f(x) = \log\left(\frac{a}{2}\right) + (a+1)\ln\left(\frac{b}{b+x}\right).$$

Logo,

$$f(x)\ln f(x) = \frac{a}{2}\left(\frac{b}{b+x}\right)^{a+1}\left[\ln\left(\frac{a}{2}\right) + (a+1)\ln\left(\frac{b}{b+x}\right)\right].$$

Substituindo na integral da entropia,

$$H(X) = -2\int_0^\infty \frac{a}{2}\left(\frac{b}{b+x}\right)^{a+1}\left[\ln\left(\frac{a}{2}\right) + (a+1)\ln\left(\frac{b}{b+x}\right)\right]dx.$$

Fazendo a substituição $u = \frac{b}{b+x}$, $x = b - \frac{b}{u}$, de modo que $dx = \frac{b}{u^2}du$, obtemos:

$$H(X) = -ab\int_0^1 u^{a-1}\left[\ln\left(\frac{a}{2}\right) + (a+1)\ln u\right]du.$$

Separando os termos,

$$H(X) = -ab\ln\left(\frac{a}{2}\right)\int_0^1 u^{a-1}du - ab(a+1)\int_0^1 u^{a-1}\ln u\,du.$$

Sabendo que

$$\int_0^1 u^{a-1}du = \frac{1}{a}, \qquad \int_0^1 u^{a-1}\ln u\,du = -\frac{1}{a^2},$$

Conclui-se, portanto, que a entropia diferencial da distribuição Pareto simétrica centrada em zero é dada por:

$$H(X) = -b \ln\left(\frac{a}{2}\right) + b \cdot \frac{a+1}{a}. \tag{4.2}$$

A seguir apresenta-se a estimativa da entropia diferencial com base no kernel *t-Student*, a partir de uma amostra $\{x_1, x_2, \ldots, x_n\}$, dada por:

$$\hat{H} = -\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{1}{(n-1)h}\sum_{\substack{j=1 \\ j\neq i}}^{n} \kappa_{h,\nu}(x_i - x_j)\right],$$

em que $\kappa_{h,\nu}(\cdot)$ corresponde à função kernel da distribuição *t-Student* com parâmetros de suavização h e grau de liberdade ν, centrada em zero. Essa função kernel é expressa como,

$$\kappa_{h,\nu}(u) = \frac{1}{h\sqrt{\nu}\,\mathrm{B}\left(\frac{1}{2}, \frac{\nu}{2}\right)}\left(1 + \frac{u^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

em que $u = (x_i - x_j)/h$ e $\mathrm{B}(\cdot,\cdot)$ denota a função beta de Euler. A flexibilidade do kernel t permite que ele se adapte a diferentes comportamentos de cauda da distribuição populacional, ajustando-se por meio do parâmetro ν.

A estimação conjunta dos parâmetros h e ν é feita por meio de minimização do valor estimado da entropia, usando a função `optim()`. O valor de ν é restrito ao domínio positivo para garantir validade estatística.

O programa 4.1 ilustra o procedimento de otimização em linguagem R.

Programa 4.1: Otimização conjunta de h e ν para o kernel *t-Student*

```
h.T <- optim(Entropy.T, par = c(2,1), x.var = x)
nu <- h.T$par[1]
h_opt <- h.T$par[2]
```

Programa 4.2: Geração de amostras da distribuição de Pareto simétrica centrada em zero via inversa da CDF

```
rpareto_sym <- function(n, a, b) {
  U <- runif(n)
  X <- numeric(n)

  X[U < 0.5] <- b * (1 - (2 * U[U < 0.5])^(-1 / a))
  X[U > 0.5] <- b * ((2 * (1 - U[U > 0.5]))^(-1 / a) - 1)

  return(X)
}

# Exemplo de uso:
set.seed(123)
samples <- rpareto_sym(n = 10000, a = 2.5, b = 1)
hist(samples, breaks = 100, main = "Pareto", col = "lightblue")
```

4.2.2 Simulação

Podemos gerar amostras de uma distribuição de Pareto simétrica centrada em 0 invertendo a função de distribuição acumulada. Para isso, devemos expressar X como uma função de U, em que U é uma variável aleatória uniformemente distribuída no intervalo $[0, 1]$. De imediato, podemos encontramos

$$X = \begin{cases} b\left\{1 - [2U]^{-1/a}\right\}, & \text{para } 0 < U < 0.5, \\ b\left\{[2(1-U)]^{-1/a} - 1\right\}, & \text{para } 0.5 < U < 1. \end{cases} \tag{4.3}$$

4.3 Experimento

O estudo de Monte Carlo realizado consiste em uma série de simulações para estimar a entropia diferencial de uma distribuição de Pareto simétrica, centrada em zero, utilizando diferentes métodos de suavização. Foram comparadas seis abordagens diferentes para a estimativa de entropia diferencial: Gaussiano, que é um estimador de entropia diferencial baseado na função kernel Gaussiana (Silverman, 1986); Laplace, que utiliza a função kernel Laplaciana (Parzen,

1962b); Cauchy, que é baseado na função kernel de Cauchy (Epanechnikov, 1969); Pareto, que utiliza a função kernel de Pareto (Matsushita et al., 2024); NND, que é um estimador de entropia diferencial baseado nas distâncias do vizinho mais próximo (Kozachenko et al., 1987); SS, que se baseia nos espaçamentos amostrais de primeira ordem (Vasicek, 1976); e o kernel *t-Student*, que emprega a função núcleo t com controle de cauda via graus de liberdade (Hall et al., 1993a). Consideramos nesse estudo apenas o caso de amostras pequenas, 30 e 60, mantendo o número de replicações fixo em 500.

Cada simulação utilizou o parâmetro de escala b igual a 1, que define o ponto a partir do qual a distribuição de Pareto é aplicada. Além disso, foram testados diversos valores do parâmetro de forma, $a \in \{0.5, 1, 1.5, 2, 3, 10\}$, que controla a cauda da distribuição: valores menores de α indicam caudas mais pesadas, significando uma maior probabilidade de eventos extremos, enquanto valores maiores de α indicam caudas mais leves, com menos eventos extremos. A entropia teórica também foi quantificada. Posteriormente, o viés, a variância e o erro quadrático médio (MSE) de cada estimador foram computados.

A Tabela 4.1 apresenta os resultados do experimento de Monte Carlo para amostras de tamanho 30, considerando diferentes valores do parâmetro de forma a da distribuição de Pareto simétrica centrada em zero. Foram calculados o erro quadrático médio (MSE), a variância e o viés para cada método de estimação da entropia diferencial. A partir desses resultados, é possível avaliar o desempenho relativo de cada estimador em diferentes regimes de cauda.

Observa-se que, para valores baixos de a, como $a = 0.5$ e $a = 1$, que correspondem a distribuições com caudas pesadas, os métodos Gaussiano e o Laplace, apresentam altos valores de MSE e viés, evidenciando fraco desempenho desses métodos nesse contexto. Nos casos mais extremos de cauda pesada, o kernel *t-Student* apresenta valores de MSE e variância consideravelmente inferiores aos dos kernels Gaussiano e Laplace, e praticamente equivalentes aos do kernel de Pareto. Para $a = 0{,}5$, por exemplo, o MSE do estimador t (0,42) é próximo ao do kernel de Pareto (0,42) e menor que o do Cauchy (0,63), enquanto os métodos Gaussiano e Laplace apresentam valores muito elevados (7,10 e 15,13, respectivamente). O viés também

Tabela 4.1: Comparação entre diferentes estimadores de entropia diferencial usando múltiplas amostras de tamanho 30 da distribuição de Pareto simétrica.

a	**Medida**	**Gaussiano**	**Laplace**	**Cauchy**	**Pareto**	**t-Student**	**NND**	**SS**
	MSE	7,10	15,13	0,63	0,42	0,42	0,39	0,38
0.5	Var	4,09	8,87	0,49	0,37	0,37	0,38	0,33
	Viés	-1,74	-2,50	-0,37	-0,22	-0,21	-0,09	0.23
	MSE	2,23	1,24	0,20	0,18	0,18	0,21	0,19
1	Var	1,16	0,80	0,18	0,17	0,17	0,21	0,16
	Viés	-1,03	-0,66	-0,15	-0,13	-0,12	0,05	0,18
	MSE	0,85	0,35	0,13	0,13	0,13	0,17	0,14
1.5	Var	0,52	0,25	0,12	0,12	0,12	0,16	0,12
	Viés	-0,58	-0,32	-0,11	-0,10	-0,09	0,04	0,16
	MSE	0,43	0,19	0,10	0,10	0,10	0,15	0,12
2	Var	0,28	0,15	0,09	0,10	0,10	0,15	0,10
	Viés	-0.38	-0,21	-0,10	-0,08	-0,08	0,03	0,14
	MSE	0,20	0,11	0,08	0,08	0,08	0,13	0,10
3	Var	0,14	0,09	0,07	0,07	0,08	0,13	0,08
	Viés	-0,23	-0,13	-0,09	-0,07	-0,06	-0,02	-0,13
	MSE	0,07	0,05	0,06	0,05	0,05	0,11	0,08
10	Var	0,06	0,05	0,05	0,05	0,05	0,11	0,06
	Viés	-0,10	-0,07	-0,09	-0,05	-0,04	0,02	0,12

permanece controlado, sendo -0,21 no kernel t, contra -1,74 no Gaussiano e -2,50 no Laplace, evidenciando a sensibilidade desses últimos às caudas pesadas.

À medida que o valor de a aumenta, tornando a cauda da distribuição menos pesada, a diferença de desempenho entre os estimadores se reduz. Todos os estimadores começam a apresentar viés e variância mais controlados. O desempenho do estimador com kernel t continua estável, para $a = 1{,}5$, $a = 2$ e $a = 3$, o MSE se mantém baixo, com valores entre 0,08 e 0,13, e o viés reduzido, indicando que o estimador t é capaz de se adaptar automaticamente ao tipo de cauda da distribuição populacional. Essa capacidade de adaptação está relacionada ao ajuste conjunto do grau de liberdade ν e do parâmetro de suavização h, otimizados via validação cruzada em cada simulação.

No cenário de cauda leve ($a = 10$), os resultados do kernel t tornam-se praticamente indistinguíveis dos dos demais métodos bem ajustados. O MSE é de 0,05, valor semelhante ao observado para os estimadores de Pareto e Laplace, e o viés permanece moderadamente baixo (-0,04).

Outro ponto a destacar é que, em todos os cenários, o estimador baseado em espaçamento amostral (SS) manteve viés consistentemente baixo e variância pequena, sendo, portanto, uma opção confiável e simples de implementar. Por outro lado, o método do vizinho mais próximo (NND) apresentou leve aumento de variância para valores intermediários de a, o que pode ser atribuído à sua sensibilidade à dispersão local das amostras.

Esses resultados reforçam as conclusões de Matsushita et al. (2024), demonstrando que o kernel de Pareto oferece uma alternativa flexível e eficaz para a estimação da entropia diferencial, especialmente em distribuições com caudas pesadas, ao passo que estimadores tradicionais podem falhar nesses contextos. Além disso, os métodos baseados em distâncias e espaçamentos amostrais mostraram-se competitivos, com a vantagem de não dependerem da escolha de uma função kernel ou de um parâmetro de suavização.

De modo geral, os resultados evidenciam que o estimador com kernel *t-Student* apresenta desempenho robusto nos diferentes cenários analisados, competindo com os melhores estimadores em termos de erro quadrático médio e viés. Sua capacidade de adaptação à espessura da cauda da distribuição torna-o uma alternativa eficaz e prática, sobretudo quando há incerteza sobre o comportamento da cauda da distribuição populacional.

A Tabela 4.2 apresenta os resultados do experimento de Monte Carlo para amostras de tamanho $n = 60$, complementando a análise realizada anteriormente para $n = 30$. Com o aumento do tamanho amostral, observa-se, de forma geral, uma melhora no desempenho de todos os estimadores, refletida na redução dos valores de MSE, variância e viés. No entanto, essa melhora é particularmente notável no estimador com kernel *t-Student*, que mantém resultados bastante competitivos ao longo de todos os cenários analisados.

No caso extremo de caudas pesadas, como em $a = 0{,}5$, o estimador *t-Student* obteve MSE

Tabela 4.2: Comparação entre diferentes estimadores de entropia diferencial usando múltiplas amostras de tamanho 60 da distribuição de Pareto simétrica.

a	**Medida**	**Gaussiano**	**Laplace**	**Cauchy**	**Pareto**	**t-Student**	**NND**	**SS**
	MSE	5,38	10,67	0,37	0,22	0,22	0,19	0,18
0.5	Var	2,25	4,53	0,24	0,18	0,18	0,19	0,16
	Viés	-1,77	-2,48	-0,36	-0,21	-0,20	-0,03	-0,14
	MSE	2,18	1,16	0,11	0,10	0,10	0,11	0,09
1	Var	0,84	0,60	0,09	0,08	0,08	0,11	0,08
	Viés	-1,16	-0,75	-0,14	-0,12	-0,12	-0,01	-0,10
	MSE	0,91	0,35	0,07	0,07	0,06	0,08	0,07
1.5	Var	0,47	0,23	0,06	0,06	0,06	0,08	0,06
	Viés	-0,66	-0,35	-0,10	-0,09	-0,09	0,01	0,09
	MSE	0,40	0,15	0,05	0,05	0,05	0,08	0,06
2	Var	0,23	0,10	0,04	0,05	0,05	0,08	0,05
	Viés	-0,41	-0,22	-0,09	-0,08	-0,08	0,01	0,08
	MSE	0,17	0,07	0,04	0,04	0,04	0,07	0,05
3	Var	0,11	0,05	0,03	0,04	0,04	0,07	0,04
	Viés	-0,23	-0,13	-0,08	-0,06	-0,06	0,00	-0,07
	MSE	0,04	0,03	0,03	0,03	0,03	0,06	0,04
10	Var	0,03	0,03	0,02	0,02	0,02	0,06	0,03
	Viés	-0,09	-0,06	-0,07	-0,05	-0,04	0,00	-0,06

de 0,22, valor consideravelmente inferior ao observado nos kernels Gaussiano (5,38) e Laplace (10,67), e praticamente equivalente ao kernel de Pareto (0,22). Além disso, seu viés (-0,20) é ligeiramente menor que o do kernel de Pareto (-0,21), e sua variância (0,18) está entre as menores do conjunto. O bom desempenho do kernel t persiste ao longo dos demais valores de a, mantendo baixos níveis de viés e variância. Para $a = 1$, o MSE do estimador t foi de 0,10, igual ao do Pareto e inferior ao do Cauchy (0,11), enquanto o viés permaneceu controlado (-0,12).

À medida que o valor de a cresce, indicando distribuições com caudas mais leves, o desempenho do estimador t converge para o dos demais métodos bem ajustados. Para $a = 2$, por exemplo, o MSE foi de 0,05, o mesmo valor obtido pelos estimadores de Pareto e Cauchy. Para

$a = 10$, cenário que representa uma distribuição com cauda leve, o kernel *t-Student* alcançou MSE de 0,03, empatando com todos os melhores métodos e com viés próximo de zero (-0,04), demonstrando sua adaptabilidade.

De maneira geral, observa-se que o aumento do tamanho amostral impacta positivamente a precisão da estimação da entropia diferencial, e o estimador com kernel *t-Student* continua exibindo desempenho robusto. Ele se mostra eficaz tanto em distribuições com caudas pesadas quanto leves, reforçando sua versatilidade. Os resultados reafirmam o potencial do kernel t como um intermediário entre estimadores clássicos e métodos especializados em caudas pesadas.

4.4 Aplicação do kernel *t-Student* em amostras da Pareto simétrica

Os resultados com o kernel *t-Student* evidenciam três pontos principais. Primeiro, a capacidade do parâmetro de cauda ν em modular a robustez do estimador de entropia nas regiões de baixa densidade: valores menores de ν (caudas mais pesadas) reduziram o viés relativo em cenários com maior incidência de observações extremas, ao custo de um leve aumento de variância; à medida que ν cresce, a resposta se aproxima da do kernel Gaussiano, com caudas mais leves e suavização proporcionalmente mais concentrada no centro da distribuição.

Segundo, a escolha conjunta (h, ν) mostrou-se determinante para o desempenho global, confirmando que o parâmetro de suavização h e o índice de cauda ν operam de forma complementar: h controla a suavização local enquanto ν regula a influência de valores afastados. Na prática, a validação cruzada sobre a entropia cruzada favoreceu pares (h, ν) que equilibram viés e variância, sobretudo em amostras pequenas (por exemplo, $n = 30$), em que escolhas muito conservadoras de ν (caudas leves) tendem a superestimar a entropia.

Terceiro, quando comparamos o kernel *t-Student* com alternativas de cauda pesada, observa-se que o kernel t se beneficia de uma interpretação direta do hiperparâmetro ν para diagnóstico de regimes de cauda: valores mais baixos de ν foram selecionados em cenários com maior

frequência de extremos, aproximando o ajuste ao comportamento observado nas caudas sem comprometer excessivamente a estabilidade em regiões centrais. Esse padrão é particularmente útil para aplicações financeiras, em que transições de regime alteram a espessura de cauda ao longo do tempo e podem ser capturadas por trajetórias estimadas de ν.

Em síntese, os achados apoiam o uso do kernel *t-Student* quando há suspeita de caudas espessas, com seleção conjunta de (h, ν) por validação cruzada. Para dados com caudas mais leves, a escolha de ν maiores conduz a resultados similares aos de kernels clássicos, preservando a estabilidade do estimador.

4.5 Aplicação do kernel *t-Student* em amostras *t-Student*

Nesta seção, replicamos o estudo de Monte Carlo para a estimação da entropia diferencial, agora assumindo que a população segue uma distribuição *t-Student* com diferentes graus de liberdade ν, variando entre valores baixos, representando caudas pesadas, e altos, caudas leves. O objetivo é avaliar o desempenho dos estimadores de entropia diferencial em diferentes regimes de cauda, considerando os estimadores Gaussiano, Laplace, Cauchy, Pareto, *t-Student*, NND e SS.

4.5.1 A distribuição *t-Student*

A distribuição *t-Student* é uma distribuição de probabilidade simétrica e de cauda pesada. Sua forma é controlada por um parâmetro positivo $\nu(>0)$, que representa os graus de liberdade. A densidade de probabilidade (PDF) é dada por:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}, \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad \text{para } x \in \mathbb{R}, \tag{4.4}$$

na qual $\Gamma(\cdot)$ representa a função gama.

O parâmetro ν controla o peso das caudas da distribuição. Quanto menor for o valor de ν, mais pesada será a cauda da distribuição. Quando ν tende ao infinito, a distribuição *t-Student*

Programa 4.3: Geração de amostras da distribuição *t-Student* centrada em zero via inversa da CDF

```
rt_sym <- function(n, nu) {
  u <- runif(n)           # Amostras uniformes
  x <- qt(u, df = nu)     # Inversa da CDF da \emph{t-Student}
  return(x)
}

# Exemplo de uso:
set.seed(123)
samples <- rt_sym(n = 10000, nu = 2)
hist(samples, breaks = 100, main = "t-Student", col = "lightblue")
```

converge para a distribuição normal padrão.

4.5.2 Entropia

Partindo da definição da entropia diferencial de uma variável contínua X com função densidade $f(x)$,

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx.$$

considerando a função densidade da distribuição *t-Student* com ν graus de liberdade, pode-se expressar a entropia da distribuição *t-Student* como:

$$H(X) = \log\left(\sqrt{\nu} \cdot \mathrm{B}\left(\frac{\nu}{2}, \frac{1}{2}\right)\right) + \frac{\nu+1}{2}\left[\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right], \tag{4.5}$$

4.5.3 Simulação

Podemos gerar amostras da distribuição *t-Student* simétrica utilizando a inversa da sua função de distribuição acumulada (CDF). Para isso, basta gerar uma variável $U \sim \mathcal{U}(0,1)$ e aplicar a função inversa da CDF da *t-Student* com ν graus de liberdade, denotada por $F_\nu^{-1}(u) =$

$\texttt{qt}(u, df = \nu)$ (Programa 4.3).

As simulações foram realizadas com tamanhos amostrais $n = 30$ e $n = 60$, mantendo o número de replicacões igual a 500. Foram considerados os seguintes valores para os graus de liberdade da distribuição t: $\nu \in 0.5, 1, 3, 5, 10, 30$, de forma a representar distribuições com caudas pesadas (valores pequenos de ν) e caudas leves (valores grandes de ν).

4.5.4 Experimento

As Tabelas 4.3 e 4.4 apresentam os resultados do experimento de Monte Carlo para estimação da entropia diferencial com amostras provenientes da distribuição *t-Student*, considerando tamanhos amostrais $n = 30$ e $n = 60$, respectivamente. Foram avaliados sete métodos de estimação: os kernels tradicionais (Gaussiano, Laplace e Cauchy), o kernel de Pareto, o kernel, *t-Student*, e os métodos não paramétricos NND (vizinho mais próximo) e SS (espaçamento amostral). O foco da análise recai sobre os kernels de Pareto e t-Student, pois ambos possuem mecanismos de controle de cauda, via α e ν, permitindo avaliar sua correspondência empírica.

Observa-se, de maneira geral, que os estimadores baseados em kernels com caudas pesadas, em particular o kernel *t-Student*, o de Pareto e o de Cauchy — apresentam desempenho superior nos cenários com caudas mais pesadas (valores baixos de ν). Por outro lado, os métodos baseados em kernels mais leves (Gaussiano e Laplace) apresentaram fraco desempenho nessas situações, com altos valores de viés e MSE.

Na Tabela 4.3, os resultados indicam que, para $\nu = 0{,}5$, os métodos Gaussiano e Laplace apresentam MSEs elevados (5,38 e 8,70, respectivamente), enquanto os kernels de Pareto, *t-Student* e Cauchy apresentam os menores erros. O kernel *t-Student* iguala o Pareto em termos de MSE (0,35) e possui viés controlado (-0,14), sugerindo excelente adequação às caudas extremamente pesadas.Essa proximidade sugere uma equivalência entre os parâmetros α (Pareto) e ν (t-Student), reforçando a ideia de que ambos desempenham papéis semelhantes na suavização em regiões de baixa densidade. À medida que ν aumenta, todos os métodos mostram melhoria no desempenho. No entanto, o kernel *t-Student* se mantém competitivo em todas as situações,

Tabela 4.3: Comparação entre diferentes estimadores de entropia diferencial, incluindo o kernel *t-Student*, usando múltiplas amostras de tamanho 30 da distribuição *t-Student*.

ν	**Medida**	**Gaussiano**	**Laplace**	**Cauchy**	**Pareto**	**t-Student**	**NND**	**SS**
	MSE	5,38	8,70	0,50	0,35	0,35	0,35	0,36
0.5	Var	2,78	5,11	0,43	0,33	0,33	0,33	0,28
	Viés	-1,61	-1,89	-0,26	-0,15	-0,14	-0,13	0,27
	MSE	1,85	0,98	0,14	0,14	0,14	0,18	0,16
1	Var	1,16	0,73	0,14	0,14	0,14	0,18	0,13
	Viés	-0,83	-0,50	-0,08	-0,07	-0,06	0,08	0,19
	MSE	0,09	0,06	0,05	0,05	0,05	0,11	0,08
3	Var	0,09	0,06	0,05	0,06	0,05	0,11	0,06
	Viés	-0,09	-0,05	-0,07	-0,03	-0,02	0,04	0,13
	MSE	0,04	0,04	0,04	0,04	0,04	0,10	0,07
5	Var	0,04	0,04	0,04	0,04	0,04	0,10	0,05
	Viés	-0,04	-0,03	-0,07	-0,02	-0,01	0,03	0,12
	MSE	0,03	0,03	0,04	0,03	0,03	0,09	0,06
10	Var	0,03	0,03	0,03	0,03	0,03	0,09	0,05
	Viés	-0,03	-0,03	-0,08	-0,03	-0,01	0,02	0,11
	MSE	0,02	0,02	0,03	0,02	0,02	0,09	0,05
30	Var	0,02	0,02	0,03	0,02	0,02	0,09	0,04
	Viés	-0,02	-0,03	-0,09	-0,03	-0,02	0,02	0,10

apresentando MSEs baixos e viés moderado mesmo para valores intermediários como $\nu = 3$ ou $\nu = 5$. Para $\nu = 30$, as diferenças entre os métodos tornam-se sutis. O MSE do kernel *t-Student* (0,02) é igual ao de outros bons estimadores, como o de Pareto e o Gaussiano, mas o método SS apresenta ligeiro viés positivo (0,10). Mesmo em cenários de caudas leves, o kernel t-Student mantém desempenho comparável ao Pareto, confirmando sua versatilidade. Esses resultados mostram que o kernel *t-Student* se adapta bem à estrutura da distribuição, mesmo com um tamanho amostral pequeno.

A Tabela 4.4 mostra que o aumento no tamanho amostral impacta positivamente todos os estimadores, mas tanto o Pareto quanto o t-Student preservam desempenho superior nas regiões de cauda pesada, enquanto kernels leves ainda falham em capturar corretamente a estrutura da

Tabela 4.4: Comparação entre diferentes estimadores de entropia diferencial, incluindo o kernel *t-Student*, usando múltiplas amostras de tamanho 60 da distribuição *t-Student*.

ν	**Medida**	**Gaussiano**	**Laplace**	**Cauchy**	**Pareto**	**t-Student**	**NND**	**SS**
	MSE	4,11	7,69	0,29	0,19	0,18	0,17	0,17
0.5	Var	1,21	2,89	0,21	0,16	0,16	0,17	0,14
	Viés	-1,70	-2,19	-0,28	-0,16	-0,15	0,06	0,16
	MSE	1,93	0,98	0,08	0,07	0,07	0,09	0,08
1	Var	0,86	0,59	0,07	0,07	0,07	0,09	0,06
	Viés	-1,03	-0,62	-0,09	-0,08	-0,08	0,04	0,12
	MSE	0,06	0,03	0,03	0,02	0,02	0,05	0,04
3	Var	0,05	0,03	0,02	0,02	0,02	0,05	0,03
	Viés	-0,11	-0,06	-0,06	-0,03	-0,03	0,02	0,08
	MSE	0,02	0,02	0,02	0,02	0,02	0,05	0,03
5	Var	0,02	0,042	0,02	0,02	0,02	0,05	0,03
	Viés	-0,05	-0,04	-0,06	-0,03	-0,02	0,01	0,07
	MSE	0,01	0,01	0,02	0,01	0,01	0,04	0,02
10	Var	0,01	0,01	0,01	0,01	0,01	0,04	0,02
	Viés	-0,02	-0,03	-0,07	-0,03	-0,02	0,01	0,06
	MSE	0,02	0,02	0,03	0,02	0,02	0,09	0,05
30	Var	0,02	0,02	0,03	0,02	0,02	0,09	0,04
	Viés	-0,02	-0,03	-0,09	-0,03	-0,02	0,02	0,10

distribuição. Em $\nu = 0{,}5$, o MSE do kernel *t-Student* (0,18) é inferior ao do Cauchy (0,29), do Gaussiano (4,11) e do Laplace (7,69). O viés também é pequeno (-0,15), reforçando sua robustez em distribuições com caudas extremamente pesadas. Para $\nu = 1$, o MSE do kernel *t-Student* é 0,07, empatado com o Pareto e melhor que os demais. O viés também é reduzido (-0,08), indicando que o estimador mantém bom desempenho mesmo quando a cauda ainda é moderadamente pesada. A partir de $\nu = 3$, as diferenças entre os estimadores tornam-se pequenas, e o kernel *t-Student* continua apresentando viés e variância baixos, confirmando sua estabilidade também em cenários com caudas moderadas e leves. No caso mais leve, $\nu = 30$, o estimador com kernel *t-Student* apresenta MSE de 0,02, igual aos melhores métodos da comparação. O viés também é muito próximo de zero (-0,02), demonstrando que, além de

robusto, o estimador é versátil e eficiente em uma ampla gama de estruturas de cauda.

Os resultados obtidos para a distribuição *t-Student* confirmam o desempenho robusto do estimador baseado no kernel t, que se adapta bem a diferentes regimes de cauda graças à flexibilidade de seu parâmetro de forma ν. Quando comparado aos demais métodos, apresenta desempenho comparável ao kernel de Pareto mesmo nas situações de caudas extremas, além de superar os kernels tradicionais, como o Gaussiano e o Laplace, em todos os cenários com caudas pesadas. Também compete com os métodos não paramétricos NND e SS, mantendo bom equilíbrio entre viés e variância. Dessa forma, o kernel t-Student se mostra mais adaptável a diferentes regimes de cauda, ao passo que o Pareto pode ser preferível em análises em que a modelagem explícita da cauda seja prioritária. A estabilidade do estimador se mantém mesmo em tamanhos amostrais reduzidos. Esses achados reforçam que o kernel *t-Student* representa uma alternativa versátil e eficaz na estimação da entropia diferencial, especialmente útil quando há incerteza quanto ao grau de cauda da distribuição populacional.

Capítulo 5

Detecção de Regimes de Cauda em Séries Financeiras do Setor Farmacêutico

5.1 Introdução

Com base na entropia diferencial estimada, busca-se avaliar o grau de eficiência informacional de mercados financeiros. A aplicação do estimador com kernel t-Student permite detectar, baseado nos dados, mudanças estruturais e transições entre regimes de cauda ao longo do tempo.

Nesse contexto, séries com janelas classificadas como de cauda pesada, associadas a valores baixos de ν, indicam maior concentração local de probabilidade e, consequentemente, menor entropia diferencial. Essa baixa entropia sugere maior dependência entre as observações, refletindo uma estrutura mais previsível nos retornos. Já séries com janelas de cauda leve, associadas a valores mais altos de ν, implicam uma distribuição mais dispersa, refletindo maior aleatoriedade no comportamento da série.

Por isso, a análise baseada na entropia diferencial estimada com kernel t-Student se mostra útil para revelar padrões de comportamento que indicam instabilidade, choques informacionais, especulação excessiva ou outros fatores que comprometem a eficiência do mercado. Essa leitura oferece uma nova perspectiva sobre a dinâmica dos retornos, podendo ser utilizada como

uma ferramenta para avaliação da eficiência dos mercados financeiros, com potencial de complementar indicadores.

O estimador de entropia diferencial baseado na função kernel da distribuição t de Student foi utilizado para investigar a presença de regimes de cauda leve e pesada nas séries temporais dos retornos logarítmicos diários das ações de quatro empresas do setor farmacêutico: Hypera, Biomm, Pfizer e Amgen.

O procedimento adotado baseia-se na estimação móvel da entropia diferencial, utilizando uma janela deslizante centrada de tamanho 61 dias. Para cada janela,foram estimados os parâmetros ν (graus de liberdade) e h (largura de banda), e classificados como cauda pesada se $\log(\nu) < 5$ e cauda leve se $\log(\nu) \geq 5$.

Essa abordagem permite identificar transições dinâmicas no comportamento de cauda ao longo do tempo, revelando potenciais mudanças estruturais nas séries de retornos.

5.2 Resultados

A Figura 5.1 apresenta os gráficos com os retornos logarítmicos diários de cada ação, sobrepostos pelas classificações de regime de cauda: os períodos em vermelho indicam janelas classificadas como cauda pesada, enquanto os períodos em preto correspondem a cauda leve.

A série temporal dos retornos logarítmicos da Hypera revela uma dinâmica marcada por alternância entre regimes de cauda leve e pesada ao longo de todo o período analisado. Observa-se que, embora o regime de cauda leve seja predominante em partes da série, há janelas bem definidas em que o regime de cauda pesada (vermelho) se manifesta de forma recorrente. Tais episódios ocorrem, por exemplo, nos anos de 2009-2011, 2015 e, com destaque particular, entre 2019 e 2021. Esses períodos coincidem com eventos de maior volatilidade no mercado financeiro brasileiro, como crises macroeconômicas, instabilidades políticas internas e, mais recentemente, os efeitos da pandemia de COVID-19. A cauda pesada indica o aumento da probabilidade de retornos extremos, refletindo maior concentração de entropia diferencial (valores

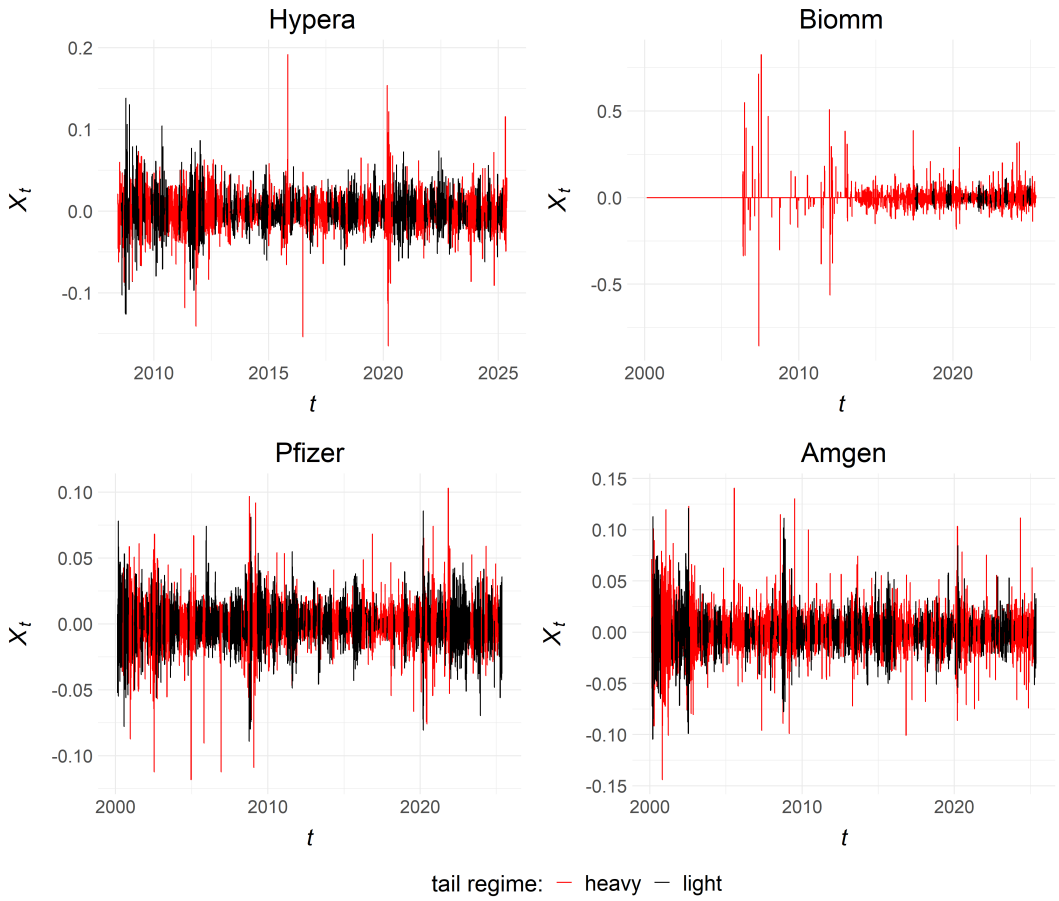

Figura 5.1: Retornos logarítmicos diários com identificação de regimes de cauda (cauda pesada evcauda leve).

mais baixos) e, portanto, menor previsibilidade e maior risco nos retornos. A presença desses períodos sugere que, mesmo sendo uma empresa de grande porte e relativa estabilidade no setor, a Hypera esteve exposta a choques específicos do mercado ou do setor farmacêutico nacional.

A série de retornos logarítmicos da Biomm apresenta um comportamento de alta volatilidade, forte assimetria e presença recorrente de valores extremos. Como evidenciado no gráfico, há uma clara predominância do regime de cauda pesada ao longo de quase toda a série histórica, indicando elevada frequência de episódios de alta entropia nos retornos da ação. Entre 2005 e 2012, nota-se a ocorrência de oscilações expressivas, com episódios de retornos positivos e negativos, reflexo do estágio inicial da empresa no mercado e da dependência de eventos específicos. Mesmo após esse período, a instabilidade persiste. Um novo momento de volatilidade é visível entre 2019 e 2021, período marcado pelos efeitos da pandemia de COVID-19. Apesar de haver alguns trechos pontuais classificados como regime de cauda leve, o comportamento dominante continua sendo de cauda pesada, denotando entropia diferencial elevada.

A série de retornos da Pfizer apresenta um comportamento notavelmente mais estável em comparação às demais empresas analisadas. Observa-se uma predominância clara do regime de cauda leve ao longo de praticamente todo o período analisado, com menores oscilações e baixa frequência de episódios extremos. Tal comportamento é coerente com o perfil da empresa uma grande farmacêutica global com ampla diversificação de portfólio, atuação consolidada em diversos mercados e forte capacidade de absorção de choques. Ainda assim, o gráfico revela alguns episódios pontuais de regime de cauda pesada, especialmente próximos ao ano de 2020, período coincidente com a pandemia de COVID-19. Esse comportamento pode estar associado à intensa oscilação de expectativas em torno do desenvolvimento da vacina da empresa em parceria com a BioNTech, o que gerou especulação e maior volume de negociações no mercado.De forma geral, a Pfizer demonstra um perfil de risco mais controlado, e a entropia diferencial estimada ao longo da série confirma seu comportamento típico de empresa mais madura, com prevalência de retornos em regime de cauda leve.

A trajetória dos retornos da Amgen revela um comportamento intermediário entre estabili-

dade e volatilidade, com alternância visível entre períodos classificados como de cauda leve e episódios de cauda pesada. Essa oscilação indica que, embora a empresa apresente fundamentos sólidos e atuação consolidada no setor biofarmacêutico, seu desempenho no mercado financeiro é afetado por eventos específicos que geram impactos pontuais na distribuição dos retornos. Ao longo da série, é possível identificar episódios de aumento na entropia que refletem momentos de maior incerteza e dispersão nos retornos. Esses períodos podem estar relacionados à divulgação de resultados clínicos de produtos em desenvolvimento, a decisões regulatórias por parte de agências como o FDA, ou ainda a movimentos estratégicos da empresa, como fusões, aquisições ou lançamentos comerciais relevantes. Também se observa a influência do contexto da pandemia de COVID-19, que afetou todo o setor de saúde, trazendo consigo tanto oportunidades de valorização quanto choques de incerteza.

Em síntese, a análise dos retornos logarítmicos das ações da Hypera, Biomm, Pfizer e Amgen evidencia padrões distintos de comportamento dinâmico de caudas, refletindo tanto a estrutura das empresas quanto os choques conjunturais do mercado farmacêutico e da economia global. A detecção de regimes de cauda leve e pesada foi realizada por meio da estimação da entropia diferencial com kernel t-Student, cuja flexibilidade permite capturar distribuições com caudas mais espessas, sendo especialmente útil para identificar episódios de retornos extremos.

Empresas mais consolidadas e com maior diversificação global, como Pfizer e Amgen, apresentaram predominância do regime de cauda leve, sugerindo maior estabilidade e menor sensibilidade a choques de mercado. Em contrapartida, companhias como Biomm, de menor porte e com histórico mais volátil, revelaram elevada frequência de janelas classificadas como cauda pesada. A Hypera, embora seja uma empresa de grande porte no contexto nacional, também apresentou episódios recorrentes de regime de cauda pesada, particularmente em períodos associados a crises macroeconômicas, instabilidades políticas e, notadamente, durante a pandemia de COVID-19.

A metodologia empregada fornece uma ferramenta robusta para avaliação não apenas da intensidade da variabilidade dos retornos, mas da sua estrutura informacional. A utilização da

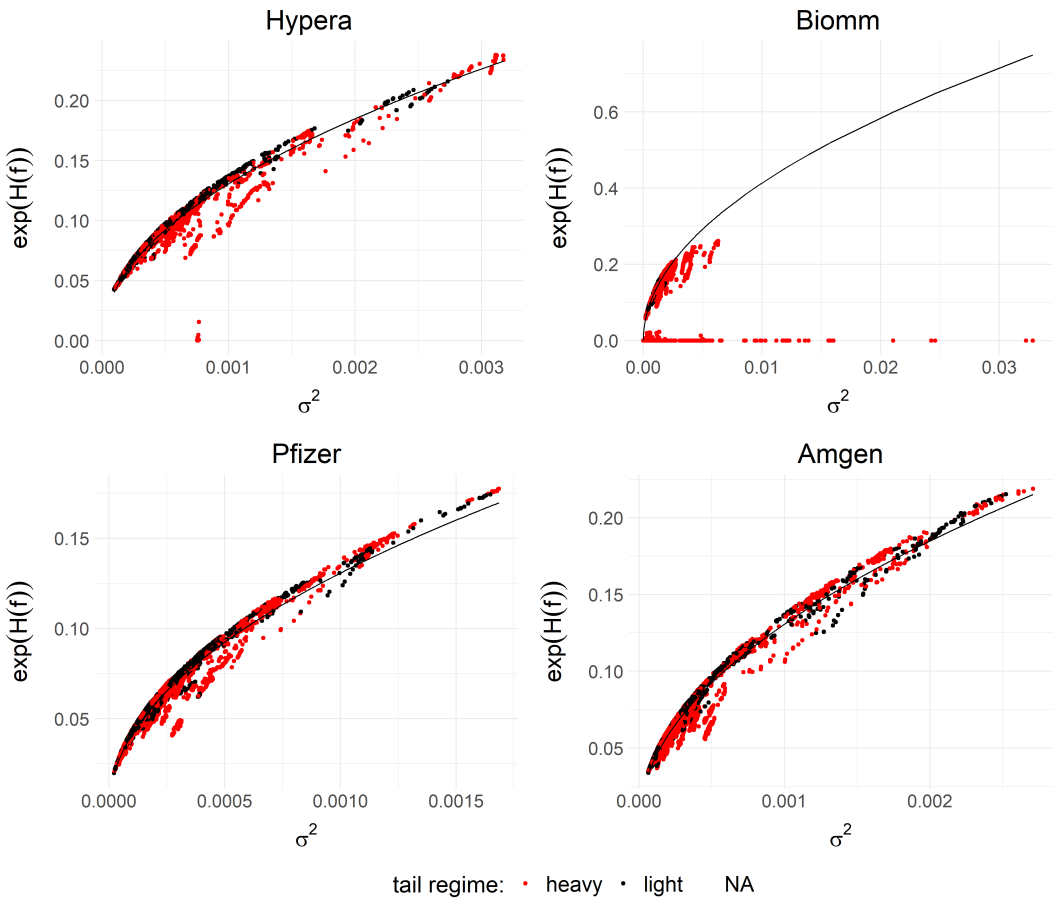

Figura 5.2: Relação entre a variância amostral dos retornos (σ^2) e a entropia diferencial estimada ($\exp(H(f))$), para diferentes regimes de cauda.

entropia diferencial estimada via kernel t-Student permite capturar alterações na distribuição de probabilidade fazer suposições.

A Figura 5.2 apresenta a relação entre a variância amostral dos retornos e os valores da entropia diferencial estimada, em escala exponencial, para cada uma das quatro empresas. Os pontos vermelhos indicam janelas classificadas como regime de cauda pesada, enquanto os pontos pretos representam janelas de cauda leve. A curva em cada gráfico representa a referência teórica esperada sob a suposição de distribuição normal dos retornos.

Observa-se que, para todas as empresas, existe uma relação positiva entre variância e entropia, o que é esperado dado que distribuições com maior dispersão tendem a apresentar maior

incerteza. No entanto, os pontos associados a janelas de cauda pesada se afastam mais da curva de referência, especialmente no caso da Biomm, onde diversas janelas apresentam entropia diferencial significativamente inferior ao esperado para os respectivos níveis de variância. Esse comportamento reflete a presença de retornos extremos que geram distribuições com caudas espessas e alta concentração de densidade, reduzindo a entropia.

No caso da Hypera, a maioria das janelas segue a tendência teórica, embora janelas com cauda pesada apresentem entropia relativamente mais baixa para a mesma variância. Isso sugere uma quebra de simetria na distribuição e maior concentração de probabilidade em torno de valores específicos. Para a Pfizer e Amgen, os resultados mostram maior aderência à curva de referência, com predominância de janelas classificadas como cauda leve, refletindo maior estabilidade na distribuição dos retornos.

Dessa forma, a entropia diferencial se mostra uma métrica complementar relevante na análise de séries financeiras, ao fornecer uma visão mais refinada sobre a organização e previsibilidade dos retornos ao longo do tempo.

A Figura 5.3 apresenta a relação entre a entropia diferencial estimada $\exp(H(\mathcal{J}))$ e o logaritmo natural do parâmetro ν do kernel t-Student para quatro empresas do setor farmacêutico: Hypera, Biomm, Pfizer e Amgen. Cada ponto representa uma janela móvel de tamanho fixo aplicada à série de retornos diários da respectiva empresa. As cores distinguem os regimes de cauda classificados a partir de ν: vermelho para cauda pesada ($\nu < 5$) e azul para cauda leve ($\nu > 5$). A linha tracejada vertical em $\ln(\nu) = \ln(5)$ demarca essa transição.

Observa-se que todas as séries apresentam momentos classificados como pertencentes aos dois regimes. No entanto, há variações importantes no padrão e na frequência desses regimes entre as empresas. A Hypera e a Biomm mostram uma predominância significativa de janelas com caudas pesadas, associadas a maior incerteza ou instabilidade nos retornos, o que se reflete nos altos valores de entropia diferencial. Por outro lado, Pfizer e Amgen, empresas multinacionais, demonstram uma distribuição mais equilibrada entre os dois regimes, com maior estabilidade no regime de cauda leve.

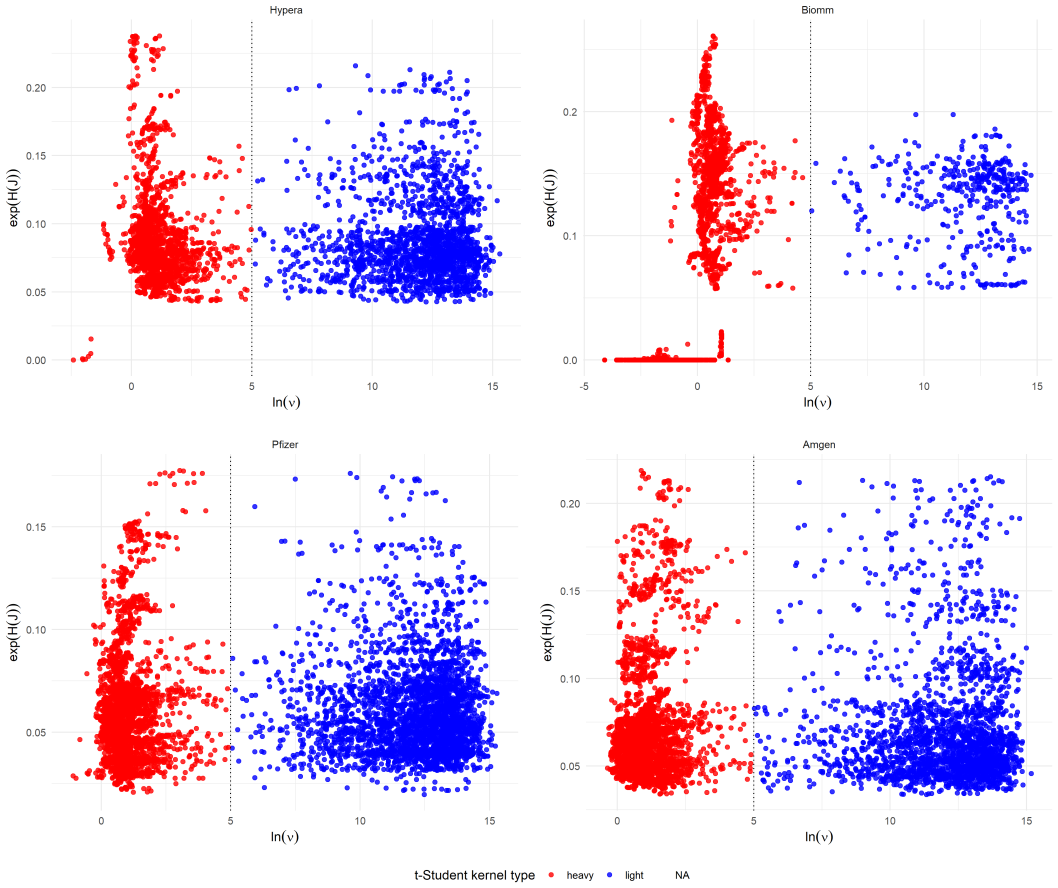

Figura 5.3: Entropia diferencial estimada em função de $\ln(\nu)$ para janelas móveis das séries de retornos de quatro empresas farmacêuticas: Hypera, Biomm, Pfizer e Amgen.

Em todas as séries, os pontos classificados como pertencentes ao regime de cauda leve tendem a se concentrar em regiões de entropia mais baixas e menos dispersas, indicando períodos mais previsíveis e menos voláteis. Já os regimes de cauda pesada estão associados a maior variabilidade na entropia estimada, o que pode refletir momentos de estresse de mercado ou eventos específicos que impactaram as ações.

5.3 Comparação entre os hiperparâmetros dos kernels t-Student e Pareto

Com o objetivo de investigar uma possível correspondência entre os hiperparâmetros ν, do kernel t de Student, e α, do kernel de Pareto, ambos aplicados à estimação de entropia diferencial em distribuições com caudas pesadas, realizou-se um estudo empírico com base em séries temporais financeiras.

Esses dois kernels surgem como alternativas aos tradicionais (como o Gaussiano), oferecendo maior flexibilidade para modelar comportamentos extremos. No kernel t de Student, o hiperparâmetro ν regula simultaneamente o grau de suavização e o peso das caudas da função densidade estimada. De forma semelhante, no kernel de Pareto, o parâmetro α influencia o decaimento das caudas, podendo ser interpretado como um indicativo indireto da intensidade das caudas da distribuição amostral.

Para explorar essa possível relação entre ν e α, aplicou-se a metodologia de janelas móveis sobre séries de preços das empresas farmacêuticas Hypera (Brasil) e Pfizer (EUA). Em cada janela deslizante, estimaram-se separadamente os hiperparâmetros ν e α, mantendo constantes o tamanho da janela e os critérios de suavização.

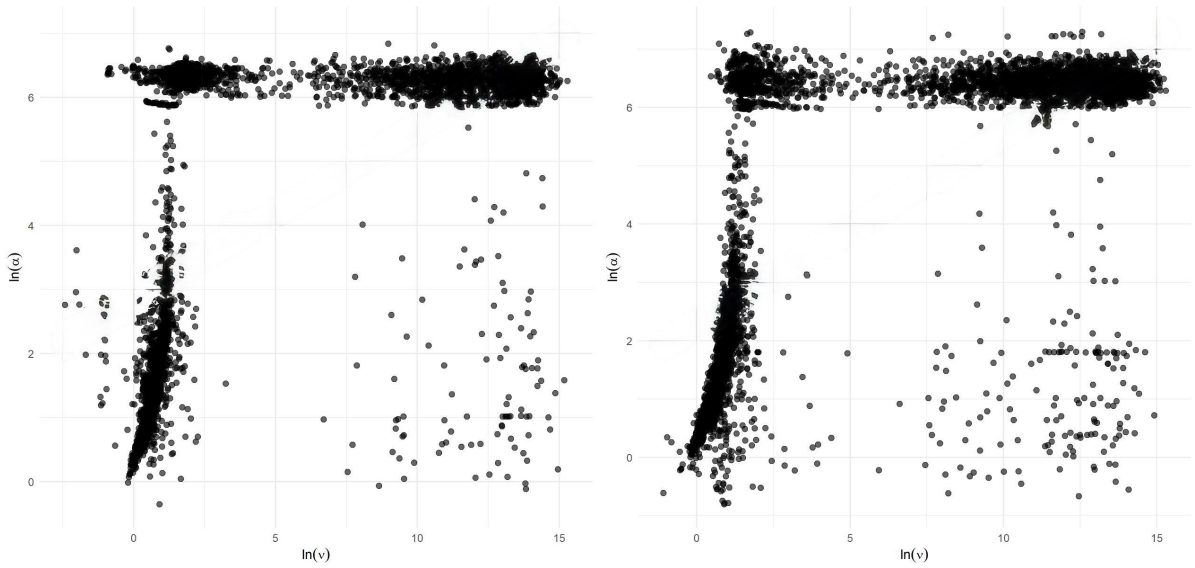

Figura 5.4: Relação entre os hiperparâmetros $\log(\nu)$ e $\log(\alpha)$ estimados via janelas móveis para as séries Hypera e Pfizer.

A Figura 5.4 ilustra a relação entre os logaritmos dos parâmetros estimados para as duas séries. Ela evidencia uma correspondência aproximadamente linear entre $\log(\nu) < 5$ e $\log(\alpha) < 6$. Essa região coincide com a detecção de casos de caudas pesadas.

No entanto, para $\log(\alpha) > 5$, observa-se uma quebra estrutural, na qual os valores de $\log(\alpha)$ ficam concentrados no patamar $\log(\alpha) \approx 6,5$, havendo grande incidência de casos $\log(\nu) < 5$ com $\log(\alpha) > 5$. Por outro lado, há menor incidência de casos $\log(\nu) > 5$ com $\log(\alpha) < 5$.

Isso significa que essas duas funções kernel possuem diferentes sensibilidades na detecção de caudas pesadas. Há uma massa maior de casos de caudas pesadas detectadas pelo kernel *t-Student* do que pelo de Pareto. Ao mesmo tempo, não se sabe se essa sensibilidade maior do kernel *t* produz maior exposição a falsos positivos.

Esses resultados sugerem que os hiperparâmetros ν e α carregam informações semelhantes sobre o regime de cauda da distribuição amostral, mas com sensibilidades (ou graus de robustez) distintas nessa detecção.

Capítulo 6

Conclusão

Este trabalho como objetivo principal investigar a estimação da entropia diferencial para distribuições univariadas de caudas pesadas, utilizando métodos baseados em estimadores de densidade por kernel. A motivação decorreu do fato de que estimadores tradicionais, como o kernel Gaussiano, apresentam desempenho limitado em contextos com observações extremas, comuns em aplicações financeiras e ambientais.

O estudo enfatizou dois tipos de funções kernel com caudas pesadas: o de Pareto simétrico e o t de *Student*. O primeiro, proposto por Matsushita et al. (2024), apresenta um parâmetro explícito de cauda (α), enquanto o segundo, introduzido por Hall et al. (1993b), possui controle indireto de cauda via os graus de liberdade (ν). Ambos permitem maior flexibilidade na estimação da entropia diferencial, sobretudo em regiões de baixa densidade.

Inicialmente, foram apresentadas definições formais da entropia diferencial e dos principais estimadores não paramétricos existentes na literatura, incluindo métodos baseados em distâncias (NND) e em espaçamentos amostrais (SS).

No desenvolvimento teórico, discutiram-se as propriedades dos estimadores kernel, com destaque para o papel crítico da escolha da função núcleo na qualidade da estimativa. Nesse contexto, analisaram-se funções kernel com caudas pesadas, como as de Pareto e t-Student, que se mostraram particularmente adequadas para estimar a entropia de distribuições com compor-

tamento extremo.

Através de simulações Monte Carlo, conduzidas com amostras geradas de distribuições com diferentes níveis de cauda (Pareto simétrica e t-Student), avaliou-se o desempenho de múltiplos estimadores em termos de viés, variância e erro quadrático médio (MSE). Os resultados evidenciaram que os estimadores baseados em kernels com caudas pesadas, especialmente o kernel t-Student, apresentam desempenho superior nos casos de distribuições com caudas espessas, mantendo viés reduzido e variância controlada mesmo em amostras de pequeno porte. Além disso, observaram-se resultados competitivos também em cenários com caudas leves, demonstrando a versatilidade e adaptabilidade do estimador.

As perguntas de pesquisa que motivaram este trabalho também foram respondidas de forma clara. Verificou-se que o estimador KDE com kernel t-Student apresenta bom desempenho mesmo em amostras pequenas, como no caso de $n = 30$, alcançando resultados comparáveis ao kernel de Pareto em contextos de caudas pesadas, com viés e MSE reduzidos. Observou-se ainda uma correspondência entre os parâmetros ν (do kernel t) e α (do kernel Pareto), já que ambos atuam de forma semelhante no controle da cauda, sugerindo que existe uma equivalência conceitual entre eles. Finalmente, a análise mostrou que a estimação do parâmetro ν pode fornecer informações relevantes em aplicações financeiras: valores baixos de ν se associaram a menor entropia diferencial, indicando maior concentração de probabilidade e sugerindo a possibilidade de identificação de regimes extremos.

Conclui-se, portanto, que a adoção de funções kernel com caudas pesadas é uma estratégia promissora para a estimação da entropia diferencial, em especial quando a distribuição subjacente apresenta valores extremos. O kernel t-Student, em particular, mostrou-se um estimador eficiente ao longo dos diferentes cenários simulados, o que reforça sua aplicabilidade prática em contextos reais, como no estudo de regimes financeiros e modelagem de incertezas.

Como perspectivas futuras, propõe-se investigar extensões multivariadas dos estimadores estudados, bem como sua aplicação em dados reais de alta dimensão, o que pode ampliar significativamente o escopo e a relevância das abordagens aqui discutidas.

Apêndice A

KDE multivariada

A.1 Introdução

A estimação de densidade desempenha um papel fundamental na análise de dados multivariados. Em casos bidimensionais, visualizações por meio de gráficos de contorno e representações tridimensionais facilitam a interpretação da estrutura dos dados. Contudo, à medida que a dimensionalidade aumenta, mesmo ferramentas computacionais avançadas enfrentam limitações na representação visual. Scott et al. (1983) discutem representações em até cinco dimensões, embora a interpretação gráfica continue desafiadora. Ainda assim, a estimação de densidade continua sendo um recurso valioso em análises estatísticas, oferecendo informações relevantes independentemente das restrições impostas pela visualização.

Além disso, a estimativa de densidade por kernel (KDE) multivariado é empregada em regressão não paramétrica, fornecendo uma base flexível para modelar relações complexas entre variáveis sem impor estruturas funcionais rígidas (Simonoff, 1996). Outro uso relevante está no campo das medidas de informação, especialmente na estimação da entropia diferencial, em que estimativas de densidade suavizadas são utilizadas para quantificar a incerteza associada a distribuições contínuas (Cover et al., 2006).

Tais aplicações destacam a importância da KDE multivariada não apenas como ferramenta

exploratória, mas também como componente essencial em métodos inferenciais modernos.

A.2 A extensão multivariada

A definição do estimador kernel para o caso multivariado pode ser generalizada naturalmente como uma soma de curvas suavizadoras multivariadas centradas na amostra. Nessa situação, suponha que $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ represente o conjunto de dados multivariados cuja densidade subjacente que se deseja estimar, no qual a i-ésima cópia amostral $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d})'$ é um vetor de dimensão d.

O estimador de densidade kernel no ponto $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, com kernel multivariado K_d pode ser definido como (Simonoff, 1996; Scott, 2006; Silverman, 2018) :

$$\hat{f}(\mathbf{x}) = \frac{1}{n|H|} \sum_{i=1}^{n} K_d \left(H^{-1}(\mathbf{x} - \mathbf{X}_i) \right), \tag{A.1}$$

no qual H é uma matriz (não-singular) de suavização $d \times d$ e $|H|$ representa o valor absoluto do seu determinante, e K_d é uma função kernel que satisfaz a condição de normalização

$$\int_{\mathbb{R}^d} K(\mathbf{u})\, d\mathbf{u} = 1,$$

para $\mathbf{u} \in \mathbb{R}^d$. Como no caso univariado, K_p possui a forma de uma função de densidade de probabilidade multivariada de uma variável aleatória contínua, geralmente simétrica em torno de zero e unimodal, como a densidade normal multivariada padrão.

Considerando H como matriz diagonal, $H = \text{diag}(h_1, \ldots, h_d)$, obtém-se uma versão simplificada do kernel multivariado na forma de um produto de funções univariadas, ou seja,

$$K_d(\mathbf{u}) = \prod_{j=1}^{d} K(u_j). \tag{A.2}$$

Nesse caso, o estimador de densidade kernel no ponto $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ com kernel

multiplicativo (A.2) e larguras de janela $h_1, \ldots, h_d$ pode ser escrito como

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_1 \cdots h_d} \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{x_j - X_{i,j}}{h_j}\right). \tag{A.3}$$

A.3 Algumas propriedades

Silverman (2018) apresenta expressões aproximadas para o vício e a variância do KDE multivariado com base na expansão de Taylor multivariada, considerando $h_j \equiv h$ constante $\forall j$. Assim, esses resultados podem não ser úteis para orientar a escolha adequada da função kernel e as larguras das janelas de suavização, já que a condição de uma janela de suavização h comum a todos os elementos de **x** é muito restritiva.

Para o propósito deste trabalho, efetuaremos estudos de Monte Carlo para a avaliação das propriedades do KDE multivariado. Mesmo assim, estudaremos as propriedades propostas por Silverman (2018) como referencial teórico.

Supondo que a densidade populacional f possua derivadas segundas, limitadas e contínuas, defina A e B tais que

$$A = \int u_1^2 K_d(\mathbf{u})\, d\mathbf{u}$$

e

$$B = \int K_d(\mathbf{u})^2\, d\mathbf{u}.$$

Com isso, obtêm-se os valores aproximados para o vício,

$$b\left[f(\mathbf{x}), \hat{f}(\mathbf{x})\right] \approx \frac{1}{2} h^2 A \nabla^2 f(\mathbf{x}),$$

e para a variância,

$$\mathrm{Var}\left[\hat{f}(\mathbf{x})\right] \approx \frac{1}{n} h^{-d} B f(\mathbf{x}).$$

Combinando essas expressões, obtêm-se o erro quadrático médio integrado (MISE) aproxi-

mado,

$$\text{MISE} \approx \frac{1}{4}h^4A^2 \int \{\nabla^2 f(\mathbf{x})\}^2 \, d\mathbf{x} + \frac{1}{n}h^{-d}B.$$

Assim, Silverman (2018) obtém uma aproximação para o parâmetro de suavização ótimo que minimiza o erro quadrático médio integrado,

$$\hat{h} = \frac{dB}{nA^2} \left\{ \int (\nabla^2 f)^2 \right\}^{-1},$$

que depende diretamente de propriedades desconhecidas da densidade f. À medida que n aumenta, $\hat{h}$ tende para 0 lentamente, com taxa $n^{-1/(d+4)}$.

Além das regras heurísticas determinação de h, há métodos mais sofisticados que buscam otimizar diretamente o desempenho do estimador de densidade. Dois desses métodos amplamente utilizados são a validação cruzada por mínimos quadrados e a validação cruzada por máxima verossimilhança, ambos extensíveis ao contexto multivariado sem modificações conceituais significativas.

No caso da validação cruzada por mínimos quadrados, o critério a ser minimizado na versão multivariada assume a seguinte forma (Silverman, 2018)

$$CV_{MQ}(h) = n^{-2}h^{-4} \sum_i \sum_j K * K \left\{ h^{-1}(\mathbf{X}_i - \mathbf{X}_j) \right\} + 2n^{-1}h^{-d}K(0).$$

Esse critério avalia o desempenho do estimador em prever os próprios dados, penalizando larguras de banda que produzem estimativas excessivamente suaves ou irregulares. A operação $K * K$ representa a convolução do kernel com ele mesmo, e $K(0)$ é o valor da função kernel no ponto zero.

Um aspecto relevante desse método é que seu custo computacional cresce com o número de observações, mas a dimensionalidade d dos dados afeta principalmente a complexidade dos cálculos das distâncias entre os pontos. Isso torna o método viável mesmo em espaços moderadamente dimensionais, embora a carga computacional possa se tornar significativa para grandes

amostras.

Em dimensões mais elevadas, surgem características específicas que tornam a estimação de densidade significativamente mais desafiadora (Silverman, 2018; Simonoff, 1996). Por exemplo, regiões de baixa densidade desempenham papel muito mais crítico no caso multivariado do que no univariado. Mesmo em distribuições bem comportadas, como a normal multivariada, uma grande proporção das observações pode se concentrar em áreas onde a densidade é bastante pequena, algo que contraria a intuição obtida em uma ou duas dimensões. Esse fenômeno implica que truncar as caudas ou desconsiderar regiões de baixa densidade pode levar a distorções nas estimativas. Além disso, há o chamado fenômeno do espaço vazio, no qual mesmo em regiões com alta densidade teórica, pode não haver observação real em uma amostra de tamanho moderado. Isso evidencia a dificuldade prática de estimar densidades multivariadas sem contar com amostras suficientemente grandes. Mesmo em um cenário ideal, o número necessário de observações cresce rapidamente à medida que d aumenta.

A.4 Nota sobre o caso $d = 2$

Para o caso bivariado, Duong et al. (2003) propõem uma alternativa explícita à forma padrão diagonal para a matriz de suavização H. Em contraste com abordagens tradicionais que restringem H a matrizes diagonais ou proporcionais à identidade, eles desenvolvem um método *plug-in* para selecionar matrizes de suavização completas (*full bandwidth matrices*), permitindo orientação arbitrária do kernel e maior flexibilidade na adaptação à estrutura dos dados. O foco dessa abordagem está na minimização do erro quadrático médio integrado assintótico (AMISE), na estabilidade numérica das estimativas e na parsimônia computacional, por meio da escolha conjunta de parâmetros piloto e uso de transformações como esferização. Embora o método seja geral e aplicável a uma ampla gama de distribuições bivariadas contínuas, Duong et al. (2003) não abordam explicitamente sobre distrbuições de caudas pesadas. Além disso, seus experimentos computacionais limitam-se a misturas de normais, e o kernel adotado é o normal

bivariado, o que favorece distribuições com suporte concentrado e caudas leves. Assim, embora teoricamente aplicável, a adequação do método a distribuições com caudas pesadas não é discutida nem avaliada, o que representa uma limitação potencial em contextos onde esse tipo de comportamento marginal seja relevante.

Referências Bibliográficas

Beirlant, Jan et al. (1997). "Nonparametric entropy estimation: An overview". *International Journal of Mathematical and Statistical Sciences* 6, pp. 17–39.

Bowman, Adrian W (1984). "An alternative method of cross-validation for the smoothing of density estimates". *Biometrika* 71.2, pp. 353–360.

Cover, Thomas M e Thomas, Joy A (2006). *Elements of Information Theory*. 2nd. Wiley-Interscience.

Devroye, Luc e Györfi, László (1985). *Nonparametric Density Estimation: The L_1 View*. New York: John Wiley & Sons.

Duong, Tarn e Hazelton, Martin L. (2003). "Plug-in bandwidth matrices for bivariate kernel density estimation". *Nonparametric Statistics* 15.1, pp. 17–30. DOI: 10.1080/10485250210000497711.

Embrechts, Paul, Klüppelberg, Claudia e Mikosch, Thomas (1997). *Modelling Extremal Events for Insurance and Finance*. Berlin: Springer.

Epanechnikov, V. A. (1969). "Non-parametric estimation of a multivariate probability density". *Theory of Probability & Its Applications* 14.1, pp. 153–158.

Hall, P. e Morton, S. C. (1993a). "On the estimation of entropy". *Annals of the Institute of Statistical Mathematics* 45.1, pp. 69–88.

Hall, P. e Morton, S.C. (1993b). "On the estimation of entropy". *Annals of the Institute of Statistical Mathematics* 45, pp. 69–88.

Hall, Peter (1987). "On Kullback-Leibler Loss and Density Estimation". *The Annals of Statistics* 15.4, pp. 1491–1519.

Holmes, Clara M. e Nemenman, Ilya (2019). "Estimators for entropy and information: a comparison and new methods". *Entropy* 21.12, p. 1155.

Kozachenko, L e Leonenko, N (1987). "Sample estimate of the entropy of a random vector". *Problems of Information Transmission* 23.1, pp. 95–101.

Madukaife, Mbanefo S. e Phuc, Ho Dang (2024). "Estimation of Shannon differential entropy: An extensive comparative review". *Entropy* 26.2, p. 266. DOI: 10.3390/e26020266.

Mammen, Enno, Marron, James S e Fisher, Nick I (1992). "Some asymptotics for multimodality tests based on kernel density estimates". *Probability Theory and Related Fields* 91.1, pp. 115–132.

Matsushita, Raul et al. (2024). "Differential entropy estimation with a Paretian kernel: Tail heaviness and smoothing". *Physica A: Statistical Mechanics and its Applications* 646, p. 129850. DOI: 10.1016/j.physa.2024.129850.

Moon, Yong e Hero, Alfred O (1995). "Estimation of entropy using kernel density estimators". *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 3, pp. III–1804.

Nemenman, Ilya, Shafee, Fariel e Bialek, William (2002). "Entropy and inference, revisited". *Advances in Neural Information Processing Systems (NeurIPS)* 14, pp. 471–478.

Paninski, Liam (2003). "Estimation of entropy and mutual information". *Neural Computation* 15.6, pp. 1191–1253.

Parzen, Emanuel (1962a). "On Estimation of a Probability Density Function and Mode". *The Annals of Mathematical Statistics* 33.3, pp. 1065–1076. DOI: 10.1214/aoms/1177704472.

— (1962b). "On estimation of a probability density function and mode". *The Annals of Mathematical Statistics* 33.3, pp. 1065–1076.

Resnick, Sidney I (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. New York: Springer.

Rosenblatt, Murray (1956). “Remarks on Some Nonparametric Estimates of a Density Function”. *The Annals of Mathematical Statistics* 27.3, pp. 832–837. DOI: 10.1214/aoms/1177728190.

Scott, David W. (2006). *Multivariate Density Estimation: Theory, Practice, and Visualization*. 2ª ed. Hoboken, NJ: John Wiley & Sons. ISBN: 978-0-471-68475-0.

Scott, D.W. e Thompson, L.A. (1983). “Multivariate Density Estimation: A Statistical Review”. *Biometrika* 70.2, pp. 297–305. DOI: 10.1093/biomet/70.2.297.

Shannon, Claude E (1948). “A mathematical theory of communication”. *Bell System Technical Journal* 27.3, pp. 379–423.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall/CRC.

Silverman, Bernard W (2018). *Density estimation for statistics and data analysis*. Routledge.

Simonoff, Jeffrey S. (1996). *Smoothing Methods in Statistics*. Springer Series in Statistics. New York: Springer. ISBN: 978-0-387-94640-7.

Singh, S., Póczos, B. e Wasserman, L. (2003). “Nearest neighbor estimates of entropy”. *Annals of Statistics* 41.2, pp. 697–724.

Terrell, G. R. e Scott, D. W. (1985). “Oversmoothed Nonparametric Density Estimates”. *Journal of the American Statistical Association* 80.389, pp. 209–214. DOI: 10.1080/01621459.1985.10477141.

Vasicek, Oldrich (1976). “A test for normality based on sample entropy”. *Journal of the Royal Statistical Society: Series B* 38.1, pp. 54–59.