MLKDE: Estimação de Densidade por Kernel Baseada em Máximo-Verossimilhança

Kelvin Brenand da Silva

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, PB
Agosto - 2024

Kelvin Brenand da Silva

MLKDE: Estimação de Densidade por Kernel Baseada em Máximo-Verossimilhança

Artigo apresentado ao curso Ciência da Computação do Centro de Informática, da Universidade Federal da Paraíba, como requisito para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Leandro Carlos de Souza

João Pessoa, PB
Agosto- 2024

Catalogação na publicação
Seção de Catalogação e Classificação

S586m Silva, Kelvin Brenand da.
MLKDE: estimação de densidade por kernel baseada em máximo-verossimilhança / Kelvin Brenand da Silva. - João Pessoa, 2024.
23 f. : il.

Orientação: Leandro Carlos de Souza.
TCC (Graduação) - UFPB/CI.

1. Estimativa de Densidade Kernel. 2. Estimativa de Largura de Banda. 3. Maximo-verossimilhança. I. Souza, Leandro Carlos de. II. Título.

UFPB/CI CDU 004.62

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Trabalho de Conclusão de Curso de Ciência da Computação intitulado **MLKDE: Estimação de Densidade por Kernel Baseada em Máximo-Verossimilhança** de autoria de Kelvin Brenand da Silva, aprovada pela banca examinadora constituída pelos seguintes professores:

_____

Prof. Dr. Leandro Carlos de Souza
Universidade Federal da Paraíba

_____

Prof. Dr. Gustavo Henrique Matos Bezerra Motta
Universidade Federal da Paraíba

_____

Prof. Dr. Natasha Correia Queiroz Lino
Universidade Federal da Paraíba

João Pessoa, 27 de Agosto de 2024

Centro de Informática, Universidade Federal da Paraíba
Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600
Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

MLKDE: Estimação de Densidade por Kernel Baseada em Máximo-Verossimilhança

Kelvin B. Silva[1], **Leandro C. Souza**[2]

[1]Centro de Informática – Universidade Ferderal da Paraíba (UFPB)
João Pessoa – PB – Brazil

[2]Departamento de Informática – Universidade Ferderal da Paraíba (UFPB)
João Pessoa – PB – Brazil

kelvinsilva@cc.ci.ufpb.br, leandro@ci.ufpb.br

Abstract. *In this paper, we propose to solve the problem of estimating the probability density function of a random variable using Kernel density estimation, which depends on the bandwidth value, wich is the method parameter. Variations in the value of this parameter can generate different shapes for the associated distributions. Therefore, we propose a method to determine the best value to be used for bandwidth for Kernel density estimation, considering data sets and using the maximum-likelihood function as a reference. The results show that the proposed method achieves suitable results for different scenarios of using probability distributions. We evaluate the performance of the proposed method using a metric that selects the largest absolute difference between the points of the true probability density function and that estimated by the methods.*

Resumo. *Neste trabalho, abordamos o problema de estimar a função densidade de probabilidade de uma variável aleatória utilizando a estimação por densidade por Kernel, que depende do valor de largura de banda, que é um parâmetro do método. Variações no valor desse parâmetro podem gerar formatos variados para as distribuições associadas. Assim, propomos um método para determinar o melhor valor a ser usado para a largura de banda para a estimação de densidade por Kernel, considerando conjuntos de dados e utilizando a função de máximo-verossimilhança como referência. Os resultados mostram que o método proposto alcança resultados adequados para diferentes cenários de uso de distribuições de probabilidade. Nós avaliamos o desempenho do método proposto usando uma métrica que seleciona a maior diferença absoluta entre os pontos da função densidade de probabilidade verdadeira e a estimada pelos métodos.*

Palavras-chave: *Estimativa de Densidade Kernel*; *Estimativa de Largura de Banda*; *Máximo-verossimilhança*.

1. Introdução

A estimação da probabilidade de dados é utilizada em diversos métodos de tomada de decisão e de Aprendizagem de Máquina. Em geral, estas estimações são obtidas de aproximações de distribuições de probabilidades, que requerem a existência de certas

propriedades nos dados para serem aplicadas, e de métodos numéricos que calculam esta aproximação [Parzen 1962].

Um método genérico bastante utilizado é a estimação de probabilidades por meio da estimação de densidade por Kernel (*Kernel Density Estimation* - KDE) [Parzen 1962]. O KDE é um método não-paramétrico usado para estimar a função densidade de uma amostra aleatória baseado-se no uso de uma função kernel. Esse procedimento é especialmente usado para suavização de dados [Turlach 1999].

O KDE utiliza os dados amostrais e uma largura de banda como parâmetros para o cálculo das estimações. A largura de banda controla o nível de suavização que será aplicado nos dados. Valores muito pequenos ou muito grandes para a largura de banda podem comprometer as estimativas obtidas [Sheather and Jones 1991] [Jones et al. 1996]. Neste sentido, existem estudos para a determinar a melhor largura de banda que possa ser utilizada, considerando um conjunto de dados de interesse [Duong 2004].

Dois métodos principais são utilizados para a determinação da largura de banda que proporcione o melhor ajuste da probabilidades do KDE: o *Least-Squares Cross-Validation* (LSCV) e o *Plug-in* [Scott and Terrell 1987]. O LSCV provê uma estimação não enviesada e é caracterizado por realizar uma minimização da soma dos erros quadrados. Já a ideia principal do *Plug-in* consiste em fazer uma suposição sobre a distribuição desconhecida. A partir disso é criada uma estimativa piloto que será plugada, ou associada, ao Erro Quadrado Integrado Médio Assintótico [Chu et al. 2015].

Este trabalho propõe uma nova abordagem para a estimação do parâmetro de largura de banda, utilizado na estimação de densidade via Kernel. Essa nova forma baseia-se em execuções de Monte Carlo do KDE e na seleção da largura de banda correspondente à maior máximo-verosimilhança. Experimentos foram conduzidos de modo a avaliar o desempenho do método proposto. Foram realizados 13 testes - utilizando diferentes parâmetros - com as distribuições gaussiana, beta, pareto e weibull. As métricas escolhidas para fazer as avaliações foram: O Erro Absoluto Máximo (EAM), *Mean Squareed Error* (MSE) e *Mean Absolute Error* (MAE).

Além desta seção de introdução, este trabalho apresenta as seguintes seções: Seção 2, que expõe de forma mais detalhada o KDE e os métodos LSCV e Plug-in, utilizados na determinação da largura de banda; Seção 3, em que o método proposto, o MLKDE, é apresentado; Seção 4, que trata da metodologia adotada na realização da avaliação, os resultados obtidos e discussões dos resultados; A seção 5 apresenta as conclusões e os trabalhos futuros.

2. Estimação de densidade por kernel

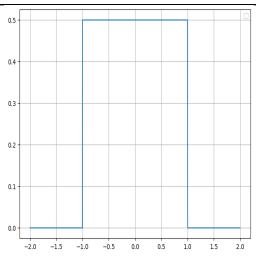
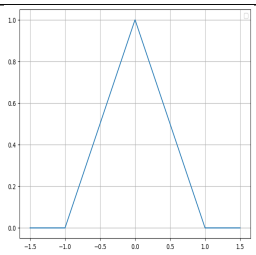
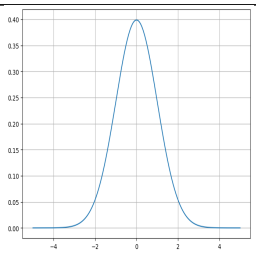
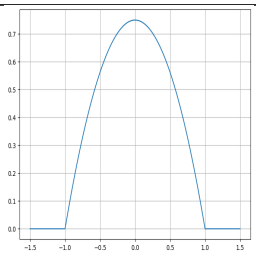
Dada uma variável aleatória unidimensional $X = \{x_1, \cdots, x_n\}$, a estimação de uma Função Densidade de Probabilidade (FDP) $\hat{f}_h$ utilizando densidade de Kernel pode ser encontrada através da Equação (1),

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

em que h é o parâmetro de suavização e $h > 0$, chamado de largura de banda, e K

é uma função kernel [Sheather and Jones 1991]. Na Tabela 1 é possível observar alguns dos kernels disponíveis na literatura [Scott 1992] [Silverman 1986] [Chen 2017] [Simonoff 1996].

Tabela 1. Alguns dos kernels mais relevantes, suas respectivas fórmulas e gráficos.

Kernel	**Fórmula** $K(x) =$	**Gráfico**
Uniforme	$U(-1, 1)$	
Triangular	$(1 - \|x\|)$	
Gaussiano	$e^{-\|\|x\|\|^2}$	
Epanechnikov	$\frac{3}{4}(1 - x^2)$	

Na Figura 1 é apresentado diferentes estimativas de densidade utilizando os kernels uniforme, triangular, gaussiano e epanechnikov aplicados em uma distribuição bimodal com cem amostras. Já na Figura 2 é apresentado o impacto que diferentes valores de largura de banda possuem ao serem utilizados nos kernels mencionados.

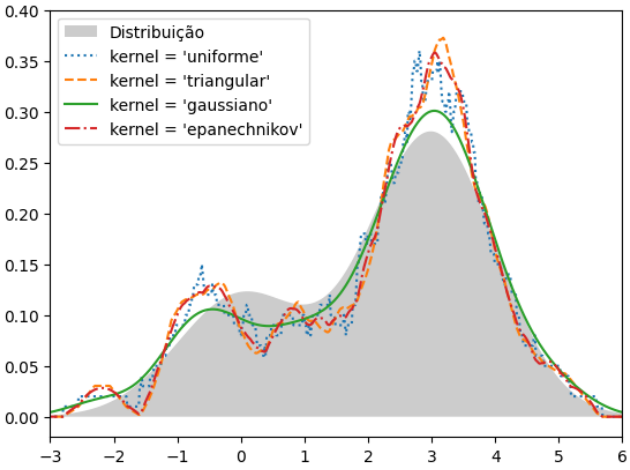


Figura 1. Comparação de estimativas de densidade utilizando diferentes kernels em relação a uma distribuição bimodal. Fonte: Autoria própria (2024).

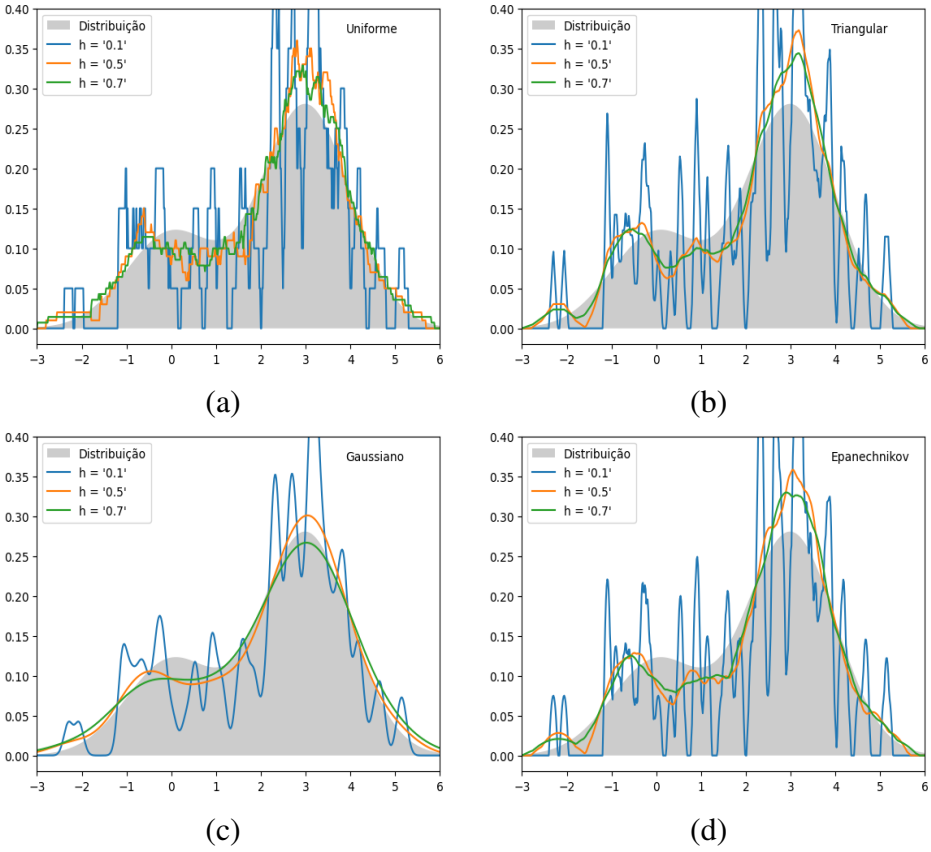


Figura 2. Impacto de diferentes larguras de banda em relação ao kernel utilizado. Em (a) temos o kernel uniforme, (b) o kernel triangular, em (c) o kernel gaussiano e em (d) o epanechnikov. Fonte: Autoria própria (2024).

A qualidade de uma estimativa de densidade é determinada principalmente pela escolha de um bom parâmetro de suavização, sendo o efeito da escolha do kernel de menor escala [Silverman 1986]. Como já mencionado, o uso de uma boa largura de banda

é crucial para a obtenção de um KDE ótimo. Nesse sentido, a literatura dispõe de dois métodos para alcançar esse fim: O LSCV e o *Plug-in*.

2.1. Least-squares cross validation

O LSCV tem como principio de funcionamento minimizar o *Integrated Square Error* (ISE) entre a distribuição $\hat{f}$ que foi estimada e a distribuição real [Silverman 1986]. Assim, o método tenta minimizar a Equação (2)

$$\text{ISE} = \int \{\hat{f}_h(x) - f(x)\}^2 dx \tag{2}$$

Como a distribuição real é desconhecida, Silverman [Horne and Garton 2006](48-49) derivou a equação do LSCV(h) [Siloko et al. 2019] como sendo a Equação (3)

$$\text{LSCV(h)} = \int \hat{f}^2 - 2n^{-1} \sum \hat{f_{-i}}(X_i) \tag{3}$$

em que n representa o número de observações e $\hat{f_{-i}}$ a estimativa de densidade sem o elemento X_i.

2.2. Plug-in

Outro método que também se baseia na redução do erro entre a função de densidade estimada e a função que gera os dados, que é desconhecida é o plug-in. O método plug-in, diferentemente do LSCV utiliza o *Asymptotic Mean Integrated Squared Error* (AMISE), como especificado na Equação (4),

$$\textbf{AMISE(h)} = n^{-1}h^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f'') \tag{4}$$

em que n é o tamanho da amostra, $R(g)$ é especificado na Equação (5) para uma função g, $\mu_2(K)$, definido pela Equação (6), é a aspereza da FDP desconhecida [Siloko et al. 2019]. Por sua vez, a FDP é representada por $\int_{-\infty}^{\infty} f''(x)^2 dx$ e K é o kernel [Eliason 1993].

$$R(g) = \int_{-\infty}^{\infty} g(x)^2 dx \tag{5}$$

$$\mu_2(K) = \int_{-\infty}^{\infty} x^2 K(x) dx \tag{6}$$

Neste caso, como requisitos tem-se que h tenda a zero, a medida que n tende ao infinito [Sheather and Jones 1991]. Assim, o h ideal seria aquele que minimiza a AMISE, cujo cálculo depende do conhecimento prévio da função geradora. Logo, o parâmetro de suavização que minimiza a AMISE do KDE [Scott and Terrell 1987] tem a forma indicada na Equação (7).

$$\hat{h}_{PI} = \left[\frac{R(K)}{\mu_2(K)^2 R(f'')n}\right]^{\frac{1}{5}} \tag{7}$$

3. MLKDE: Estimador de densidade de kernel baseado em máximo-verossimilhança

O Estimador de Densidade de Kernel de Máximo-Verossimilhança (MLKDE) tenta encontrar uma largura de banda que maximize a verosimilhança dos dados, considerando as estimações de probabilidade obtidas através da Equação (1). Para isso, o MLKDE retira B valores aleatórios para a largura de banda, em um intervalo de busca, e para cada valor, estima a probabilidade, de acordo com a Equação (1) utilizando o KDE.

O limite esquerdo do intervalo é um parâmetro do método. Já o limite direito é calculado pelo método. No algoritmo 1 o limite esquerdo é representado por ϵ e o limite direito representado por D. O cálculo de D consiste de, para cada ponto do conjunto de dados, definir a distância euclidiana em relação aos demais pontos. A maior distância obtida será então o limite direito do intervalo de buscas utilizado pelo MLKDE. Isso garante encontrar uma largura de banda que se adeque as configurações de vizinhança de qualquer conjunto de dados utilizado. A largura de banda que obtiver o KDE com maior máximo-verosimilhança será o selecionado.

A máximo-verossimilhança pode ser calculada a partir da fórmula [Held and Bové 2014]:

$$\hat{\ell}(\theta|x_1, \ldots, x_n) = \ln L = \sum_{i=1}^{n} \ln f(x_i|\theta) \tag{8}$$

O método da máximo-verossimilhança estima θ_0 buscando o valor de θ que maximiza $\hat{\ell}(\theta \mid x)$. Este é o chamado Estimador de Máximo-Verossimilhança (MLE) de θ_0 [Chu et al. 2015]:

$$\hat{\theta}_{mle} = \arg\max_{\theta \in \Theta} \ell(\hat{\theta}|x_1, ..., x_n) \tag{9}$$

Além disso, os KDEs calculados para subsequente utilização na máximo-verossimilhança possuem uma particularidade que é o uso da técnica Validação Cruzada *Leave-One-Out* (LOO). Essa técnica consiste em retirar um elemento por vez do conjunto de dados e calcular o KDE com um dado valor de h. O LOO é utilizado por possuir um baixíssimo viés e possibilitar o uso máximo do conjunto de dados, já que, durante o cálculo do KDE, utiliza-se o conjunto de dados uma vez para cada elemento que for retirado [Molinaro et al. 2005]. O Algoritmo 1 apresenta o pseudocódigo do MLKDE.

Algoritmo 1 Pseudocódigo do MLKDE

Entrada: X: Conjunto de pontos de um conjunto de dados; ϵ: Limite inferior do intervalo; B: número de buscas realizadas

Saída: Estimativa de largura de banda (h) para o KDE.

1: Para cada ponto do conjunto X, determinar a distância a todos os outros;
2: Coloque na variável **D** a maior distância dentre todas obtidas no passo 1.
3: **para** $b = 1$ até B **faça**
4: $h_b \leftarrow$ um valor aleatório no intervalo $[\epsilon, D]$.
5: **para** cada ponto x_i de X **faça**
6: $p_i \leftarrow$ A probabilidade de x_i, usando a Equação (1) e os outros pontos de X.
7: **fim para**
8: $MLE_b \leftarrow$ Máximo-verossimilhança, usando a Equação (8) e as probabilidades calculadas no passo 6.
9: **fim para**
10: **retorna** Determine o maior MLE, dentre os calculados no passo 8 e retorne o h correspondente.

4. Avaliação Experimental

Esta seção descreve como os experimentos para validação do método MLKDE foram conduzidos. Para realizar a avaliação do método proposto em relação aos demais métodos, foram empregados dados sintéticos provenientes de quatro distribuições probabilísticas distintas: Gaussiana, Beta, Pareto e Weibull. Cada uma dessas distribuições foi explorada com variações em seus parâmetros, a fim de criar cenários representativos para a aplicação do método em estudo. O kernel escolhido para ser usado foi o gaussiano.

O algoritmo proposto neste trabalho foi desenvolvido utilizando a linguagem de programação Python 3, escolhida devido à sua ampla utilização e popularidade nas áreas de machine learning e estatística. A linguagem Python é conhecida por sua versatilidade e eficiência no desenvolvimento de algoritmos complexos, tornando-a uma escolha ideal para este projeto. Além disso, A escolha dessa linguagem também permite uma integração mais fácil com outras bibliotecas e ferramentas comumente utilizadas em projetos de análise de dados.

A métrica escolhida para avaliar o desempenho dos métodos consiste de selecionar o erro absoluto máximo entre os pontos da função densidade de probabilidade verdadeira de cada distribuição e os pontos estimados por cada método. Essa métrica é dada pela fórmula (10).

$$EAM = \max\left(\left|f - \hat{f}\right|\right) \tag{10}$$

A partir dessa métrica é possível observar a magnitude do maior erro de cada método em relação aos dados verdadeiros. Isso nos fornece uma visão detalhada das discrepâncias entre os métodos utilizados e os dados reais. Além dessa métrica, também foram utilizadas duas das mais populares métricas de erro de uso geral: A *mean squared error* e *mean absolute error*. As representações matemáticas da MSE e da MAE podem ser observadas nas equações (11) e (12) respectivamente [Botchkarev 2018]. Para as três métricas

utilizadas, quanto menor o valor obtido, melhor o desempenho do método. Para facilitar a compreensão dos resultados obtidos, esse valor também estará em destaque negrito nas tabelas que seguem.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{11}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \tag{12}$$

4.1. Avaliação com distribuição gaussiana

Esta subseção descreve como o experimento utilizando distribuição gaussiana foram realizados, seus resultados e quais conclusões podem ser extraídas. Foi realizado um experimento utilizando distribuição gaussiana, onde os parâmetros utilizados serão descritos na subsubseção que segue.

4.1.1. Experimento com distribuição gaussiana

Neste experimento, foram retiradas 500 amostras de uma distribuição gaussiana com média igual a zero e variância igual a um. Ou seja, $N(\mu = 0, \sigma^2 = 1)$. O valor para ϵ utilizado foi 1e-3 e para B foi 100. Na Tabela 2 e na Figura 3 são apresentados os resultados desse experimento. Para as três métricas utilizadas, é possível notar que o método proposto obteve um desempenho superior em relação aos demais métodos. Na Figura 3 observa-se que o MLKDE conseguiu se aproximar mais da função densidade de probabilidade real, ao passo que o LSCV e o Plug-in desempenharam de forma similar e inferior ao MLKDE.

Tabela 2. Resultados do experimento utilizando uma distribuição N(μ=0,σ^2=1). Os parâmetros do MLKDE foram ϵ = 1e-3 e B = 100.

Método	EAM	MSE	MAE	Largura de banda
LSCV	0.056157	0.001299	0.031568	0.3315
Plug-in	0.054967	0.001245	0.030921	0.3220
MLKDE	**0.035628**	**0.000423**	**0.017828**	**0.3691**

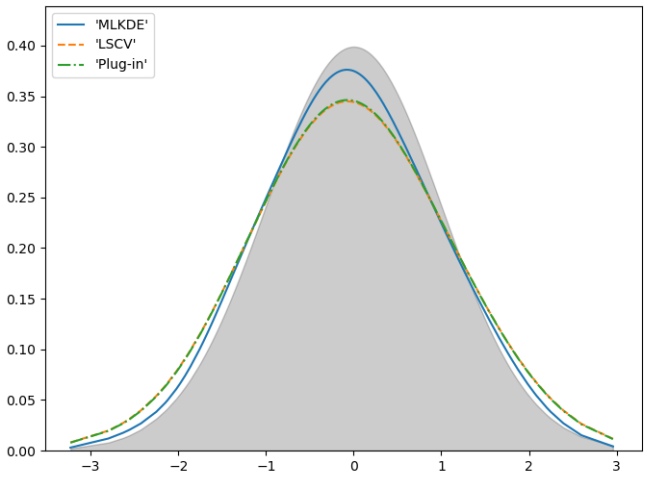


Figura 3. Representação gráfica dos resultados da Tabela 2. Para esse experimento, foi utilizado uma N(μ=0,σ^2=1). No MLKDE, os valores dos parâmetros utilizados foram ϵ = 1e-3 e B = 100. Fonte: Autoria própria (2024).

4.2. Avaliação com distribuições beta

Esta subseção descreve como os experimentos utilizando distribuições beta foram realizados, seus resultados e quais conclusões podem ser extraídas. Foram realizados quatro experimentos distintos utilizando distribuições beta, onde os parâmetros utilizados serão descritos em cada subsubseção.

4.2.1. Primeiro experimento com distribuições beta

Para o primeiro experimento, foram retiradas 1000 amostras de uma distribuição beta com $\alpha = 2$ e $\beta = 10$. Ou seja, $Beta(2, 10)$. O valor para ϵ utilizado foi 1e-3 e para B foi utilizado 100. Na Tabela 3 e na Figura 4 são apresentados os resultados desse experimento. Para as três métricas utilizadas, é possível notar que o método proposto obteve um desempenho consideravelmente superior em relação aos demais métodos. A Figura 4 demonstra isso mais claramente.

Tabela 3. Resultados do primeiro experimento utilizando uma distribuição Beta(2,10). Os parâmetros do MLKDE foram ϵ = 1e-3 e B = 100.

Método	EAM	MSE	MAE	Largura de banda
LSCV	2.006552	1.667754	1.091353	0.0176
Plug-in	2.087372	1.835004	1.148853	0.0198
MLKDE	**0.492856**	**0.055592**	**0.196259**	0.0116

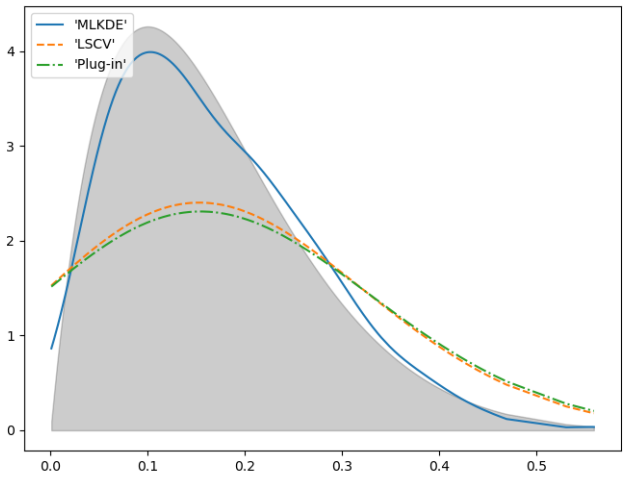


Figura 4. Representação gráfica dos resultados da Tabela 3. Para esse experimento, foi utilizado uma Beta(2,10). No MLKDE, os valores dos parâmetros utilizados foram ϵ = 1e-3 e B = 100. Fonte: Autoria própria (2024).

4.2.2. Segundo experimento com distribuições beta

Para o segundo experimento, foram retiradas 1000 amostras de uma distribuição beta com $\alpha = 1$ e $\beta = 2$. Ou seja, $Beta(1, 2)$. O valor para ϵ utilizado foi 1e-3 e para B foi utilizado 100. Na Tabela 4 e na Figura 5 são apresentados os resultados desse experimento. Para as três métricas utilizadas, é possível observar que tanto o método proposto quanto o método LSCV obtiveram bons desempenhos na realização da tarefa, ao passo que o Plug-in não foi tão bem. A Figura 5 fornece mais detalhes sobre o teste.

Tabela 4. Resultados do segundo experimento utilizando uma distribuição Beta(1,2). Os parâmetros do MLKDE foram ϵ = 1e-3 e B = 100.

Método	EAM	MSE	MAE	Largura de banda
LSCV	1.048395	0.116629	0.181216	0.0124
Plug-in	1.128056	0.220706	0.320132	0.0399
MLKDE	**0.422637**	**0.02693**	**0.133783**	0.0097

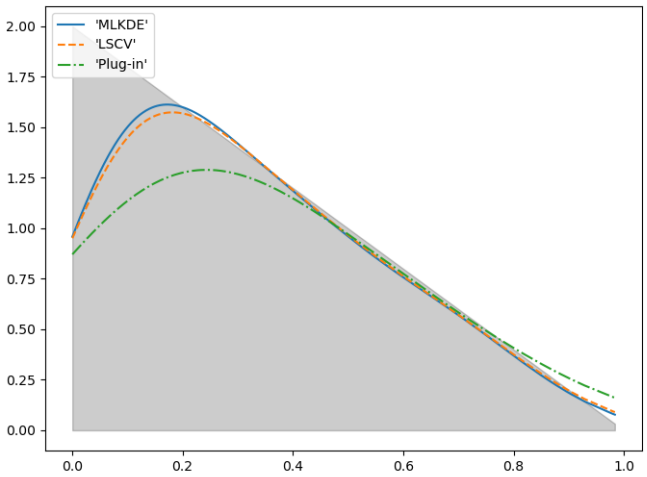


Figura 5. Representação gráfica dos resultados da Tabela 4. Para esse experimento, foi utilizado uma Beta(1,2). No MLKDE, os valores dos parâmetros utilizados foram ϵ = 1e-3 e B = 100. Fonte: Autoria própria (2024).

4.2.3. Terceiro experimento com distribuições beta

Para o terceiro experimento, foram retiradas 1000 amostras de uma distribuição beta com $\alpha = 2$ e $\beta = 2$. Ou seja, $Beta(2, 2)$. O valor para ϵ utilizado foi 1e-3 e para B foi utilizado 100. Na Tabela 5 são apresentados os resultados desse experimento. Para as três métricas utilizadas, é possível notar que o método proposto mais uma vez superou os outros dois métodos. Na Figura 6 observa-se que o MLKDE conseguiu se aproximar mais da função densidade de probabilidade real, ao passo que o LSCV e o Plug-in tiveram dificuldade em realizar essa tarefa.

Tabela 5. Resultados do terceiro experimento utilizando uma distribuição Beta(2,2). Os parâmetros do MLKDE foram ϵ = 1e-3 e B = 100.

Método	EAM	MSE	MAE	Largura de banda
LSCV	0.388338	0.074325	0.253969	0.0637
Plug-in	0.378818	0.065616	0.239071	0.0585
MLKDE	**0.200366**	**0.008132**	**0.070029**	0.0194

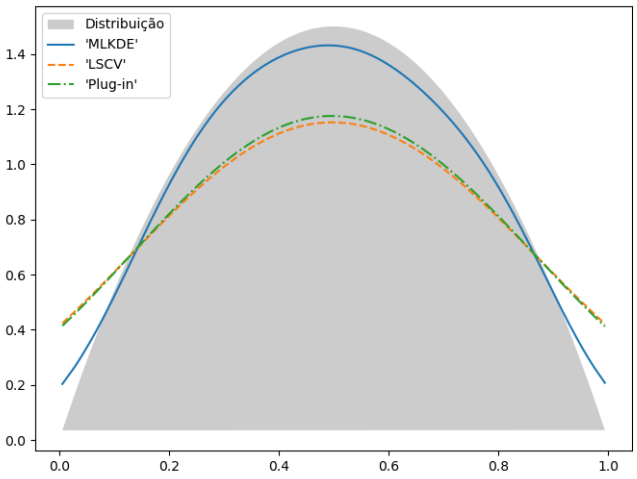


Figura 6. Representação gráfica dos resultados da Tabela 5. Para esse experimento, foi utilizado uma Beta(2,2). No MLKDE, os valores dos parâmetros utilizados foram ϵ = 1e-3 e B = 100. Fonte: Autoria própria (2024).

4.2.4. Quarto experimento com distribuições beta

Para o quarto experimento, foram retiradas 1000 amostras de uma distribuição beta com $\alpha = 2$ e $\beta = 5$. Ou seja, $Beta(2, 5)$. O valor para ϵ utilizado foi 1e-3 e para B foi utilizado 100. Na Tabela 6 e na Figura 7 são apresentados os resultados do experimento. O comportamento dos métodos permaneceu constante em relação aos outros experimentos com a distribuição beta: O MLKDE conseguiu superar o LSCV e o Plug-in tanto para métrica EAM quanto para as métricas MSE e MAE.

Tabela 6. Resultados do quarto experimento utilizando uma distribuição Beta(2,5). Os parâmetros do MLKDE foram ϵ = 1e-3 e B = 100.

Método	LSCV	Plug-in	MLKDE	Largura de banda
LSCV	0.95289	0.393999	0.539881	0.0374
Plug-in	0.926929	0.369683	0.521967	0.035
MLKDE	**0.355012**	**0.022364**	**0.126821**	0.0129

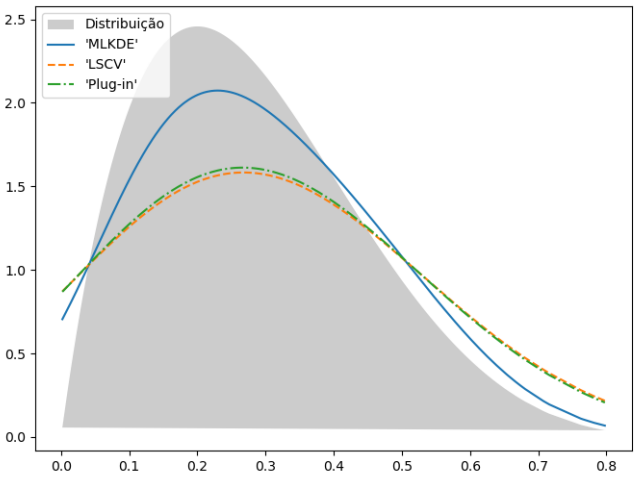


Figura 7. Representação gráfica dos resultados da Tabela 6. Para esse experimento, foi utilizado uma Beta(2,5). No MLKDE, os valores dos parâmetros utilizados foram ϵ = 1e-3 e B = 100. Fonte: Autoria própria (2024).

4.3. Avaliação com distribuições pareto

Esta subseção descreve como os experimentos utilizando distribuições pareto foram realizados, seus resultados e quais conclusões podem ser extraídas. Foram realizados quatro experimentos distintos utilizando distribuições pareto, onde os parâmetros utilizados serão descritos em cada subsubseção.

4.3.1. Primeiro experimento com distribuições pareto

Para o primeiro experimento, foram retiradas 1000 amostras de uma distribuição pareto com $\alpha = 5$. O valor para ϵ utilizado foi 1e-3 e para B foi utilizado 100. Na Tabela 7 e na Figura 8 são apresentados os resultados desse experimento. Para as três métricas utilizadas, é possível notar que o método proposto obteve um desempenho consideravelmente superior em relação aos demais métodos. Na Figura 8 isso é demonstrado mais claramente.

Tabela 7. Resultados do primeiro experimento utilizando uma distribuição pareto com α = 5. Os parâmetros do MLKDE foram ϵ = 1e-3 e B = 100.

Método	EAM	MSE	MAE	Largura de banda
LSCV	3.197188	1.180124	0.620866	0.0058
Plug-in	3.555613	2.127115	0.982356	0.0221
MLKDE	**3.192304**	**1.162836**	**0.612793**	**0.0748**

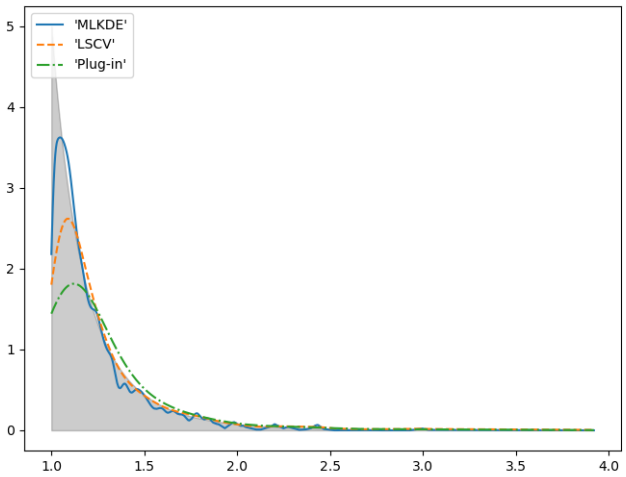


Figura 8. Representação gráfica dos resultados da Tabela 7. Para esse experimento, foi utilizado uma distribuição pareto com α = 5. No MLKDE, os valores dos parâmetros utilizados foram ϵ = 1e-3 e B = 100. Fonte: Autoria própria (2024).

4.3.2. Segundo experimento com distribuições pareto

Para o segundo experimento, foram retiradas 1000 amostras de uma distribuição pareto com $\alpha = 10$. O valor para ϵ utilizado foi 1e-3 e para B foi utilizado 100. Na Tabela 8 e na Figura 9 são apresentados os resultados desse experimento. Para as três métricas utilizadas, é possível notar que o método proposto obteve um desempenho consideravelmente superior em relação aos demais métodos. Na Figura 9 isso é demonstrado mais claramente.

Tabela 8. Resultados do segundo experimento utilizando uma distribuição pareto com α = 10. Os parâmetros do MLKDE foram ϵ = 1e-3 e B = 100.

Método	EAM	MSE	MAE	Largura de banda
LSCV	6.632118	6.447306	1.609385	0.0028
Plug-in	7.476074	11.4627	2.472358	0.0106
MLKDE	**6.119227**	**3.640273**	**1.02295**	**0.0279**

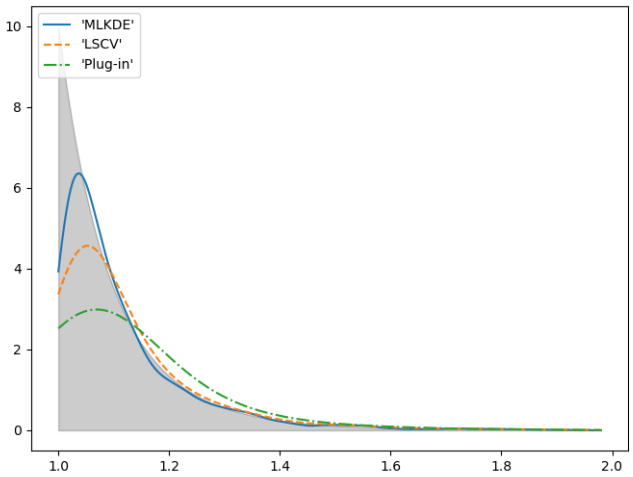


Figura 9. Representação gráfica dos resultados da Tabela 8. Para esse experimento, foi utilizado uma distribuição pareto com α = 10. No MLKDE, os valores dos parâmetros utilizados foram ϵ = 1e-3 e B = 100. Fonte: Autoria própria (2024).

4.3.3. Terceiro experimento com distribuições pareto

Para o terceiro experimento, foram retiradas 1000 amostras de uma distribuição pareto com $\alpha = 15$. O valor para ϵ utilizado foi 1e-3 e para B foi utilizado 100. Na Tabela 9 e na Figura 10 são apresentados os resultados desse experimento. Para as três métricas utilizadas, é possível notar que o método proposto obteve um desempenho consideravelmente superior em relação aos demais métodos. A Figura 10 demonstra isso mais claramente. Neste teste observa-se que todos os métodos, em especial o LSCV e Plug-in, apresentam uma notável dificuldade de se ajustar aos dados. Isso é especialmente claro em vista das métricas EAM e MSE.

Tabela 9. Resultados do terceiro experimento utilizando uma distribuição pareto com α = 15. Os parâmetros do MLKDE foram ϵ = 1e-3 e B = 100.

Método	EAM	MSE	MAE	Largura de banda
LSCV	10.252768	17.5764	2.817748	0.0019
Plug-in	11.590022	30.223529	4.163242	0.007
MLKDE	**9.009154**	**7.038164**	**1.379848**	**0.0153**

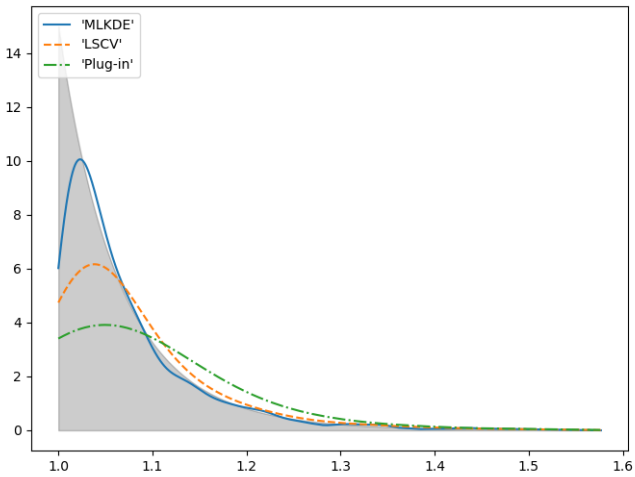


Figura 10. Representação gráfica dos resultados da Tabela 9. Para esse experimento, foi utilizado uma distribuição pareto com α = 15. No MLKDE, os valores dos parâmetros utilizados foram ϵ = 1e-3 e B = 100. Fonte: Autoria própria (2024).

4.3.4. Quarto experimento com distribuições pareto

Para o quarto experimento, foram retiradas 1000 amostras de uma distribuição pareto com $\alpha = 20$. O valor para ϵ utilizado foi 1e-3 e para B foi utilizado 100. Na Tabela 10 e na Figura 11 são apresentados os resultados desse experimento. Para as três métricas utilizadas, é possível notar que o método proposto obteve um desempenho consideravelmente superior em relação aos demais métodos. Neste teste observa-se que todos os métodos, em especial o LSCV e Plug-in, apresentam uma notável dificuldade de se ajustar aos dados. Isso é especialmente claro em vista das métricas EAM e MSE.

Tabela 10. Resultados do quarto experimento utilizando uma distribuição pareto com α = 20. Os parâmetros do MLKDE foram ϵ = 1e-3 e B = 100.

Método	EAM	MSE	MAE	Largura de banda
LSCV	13.995738	35.62474	4.154902	0.0014
Plug-in	15.821382	59.542864	5.968571	0.0052
MLKDE	**11.936579**	**11.885863**	**1.780406**	**0.0107**

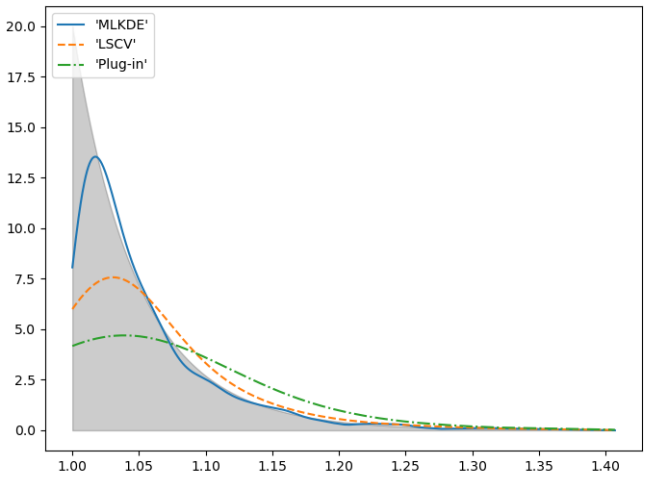


Figura 11. Representação gráfica dos resultados da Tabela 10. Para esse experimento, foi utilizado uma distribuição pareto com α = 20. No MLKDE, os valores dos parâmetros utilizados foram ϵ = 1e-3 e B = 100. Fonte: Autoria própria (2024).

4.4. Avaliação com distribuições weibull

Esta subseção descreve como os experimentos utilizando distribuições weibull foram realizados, seus resultados e quais conclusões podem ser extraídas. Foram realizados quatro experimentos distintos utilizando distribuições weibull, onde os parâmetros utilizados serão descritos em cada subsubseção.

4.4.1. Primeiro experimento com distribuições weibull

Para o primeiro experimento, foram retiradas 1000 amostras de uma distribuição weibull com $\alpha = 1.5$. O valor para ϵ utilizado foi 1e-3 e para B foi utilizado 100. Na Tabela 11 e na Figura 12 são apresentados os resultados desse experimento. Para as três métricas utilizadas, é possível notar que o método proposto obteve um desempenho superior, porém todos os métodos desempenharam bem. Na Figura 12 observa-se esse bom comportamento dos métodos.

Tabela 11. Resultados do primeiro experimento utilizando uma distribuição weibull com α = 1.5. Os parâmetros do MLKDE foram ϵ = 1e-3 e B = 100.

Método	EAM	MSE	MAE	Largura de banda
LSCV	0.173804	0.007315	0.060676	0.0885
Plug-in	0.19138	0.009547	0.070647	0.1145
MLKDE	**0.118789**	**0.003241**	**0.044358**	**0.1813**

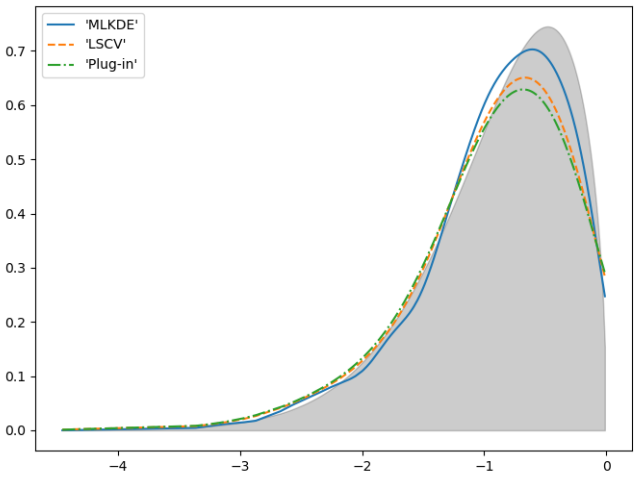


Figura 12. Representação gráfica dos resultados da Tabela 11. Para esse experimento, foi utilizado uma distribuição weibull com α = 1.5. No MLKDE, os valores dos parâmetros utilizados foram ϵ = 1e-3 e B = 100. Fonte: Autoria própria (2024).

4.4.2. Segundo experimento com distribuições weibull

Para o segundo experimento, foram retiradas 1000 amostras de uma distribuição weibull com $\alpha = 5$. O valor para ϵ utilizado foi 1e-3 e para B foi utilizado 100. Na Tabela 12 e na Figura 13 são apresentados os resultados desse experimento. Diferentemente do teste anterior, neste teste, para as três métricas utilizadas, é possível notar que o método proposto obteve um desempenho consideravelmente superior ao LSCV e ao Plug-in. Na Figura 13 observa-se esse de forma mais clara o comportamento dos métodos.

Tabela 12. Resultados do segundo experimento utilizando uma distribuição weibull com α = 5. Os parâmetros do MLKDE foram ϵ = 1e-3 e B = 100.

Método	EAM	MSE	MAE	Largura de banda
LSCV	0.627549	0.179646	0.372327	0.0575
Plug-in	0.586882	0.156132	0.346815	0.0514
MLKDE	**0.110538**	**0.003224**	**0.04946**	**0.0622**

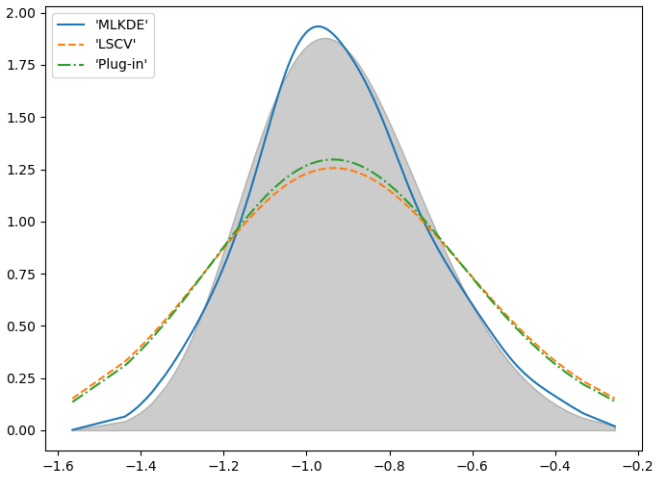


Figura 13. Representação gráfica dos resultados da Tabela 12. Para esse experimento, foi utilizado uma distribuição weibull com α = 5. No MLKDE, os valores dos parâmetros utilizados foram ϵ = 1e-3 e B = 100. Fonte: Autoria própria (2024).

4.4.3. Terceiro experimento com distribuições weibull

Para o terceiro experimento, foram retiradas 1000 amostras de uma distribuição weibull com $\alpha = 10$. O valor para ϵ utilizado foi 1e-3 e para B foi utilizado 100. Na Tabela 13 e na Figura 14 são apresentados os resultados desse experimento. Neste teste o método proposto também obteve um desempenho consideravelmente superior ao LSCV e ao Plug-in. Na Figura 14 observa-se de forma mais clara o comportamento dos métodos.

Tabela 13. Resultados do terceiro experimento utilizando uma distribuição weibull com α = 10. Os parâmetros do MLKDE foram ϵ = 1e-3 e B = 100.

Método	EAM	MSE	MAE	Largura de banda
LSCV	1.736213	1.376514	1.023987	0.0282
Plug-in	1.669367	1.262617	0.979685	0.0256
MLKDE	**0.252744**	**0.011756**	**0.087745**	**0.0319**

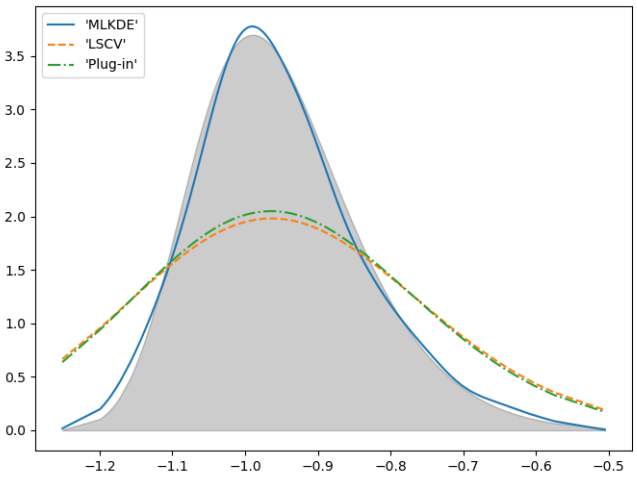


Figura 14. Representação gráfica dos resultados da Tabela 13. Para esse experimento, foi utilizado uma distribuição weibull com α = 10. No MLKDE, os valores dos parâmetros utilizados foram ϵ = 1e-3 e B = 100. Fonte: Autoria própria (2024).

4.4.4. Quarto experimento com distribuições weibull

Para o último experimento, também foram retiradas 1000 amostras de uma distribuição weibull com $\alpha = 15$. O valor para ϵ utilizado foi 1e-3 e para B foi utilizado 100. Os resultados desse experimento são apresentados na 14 e na Figura 15. Neste teste o método proposto continuou a obter um desempenho dominante para com os demais métodos avaliados. Esse comportamento é facilmente demonstrado na Figura 15.

Tabela 14. Resultados do quarto experimento utilizando uma distribuição weibull com α = 15. Os parâmetros do MLKDE foram ϵ = 1e-3 e B = 100.

Métrica	EAM	MSE	MAE	Largura de banda
LSCV	3.011563	4.214557	1.790626	0.0187
Plug-in	2.920974	3.934284	1.728211	0.017
MLKDE	**0.393747**	**0.026127**	**0.125819**	**0.022**

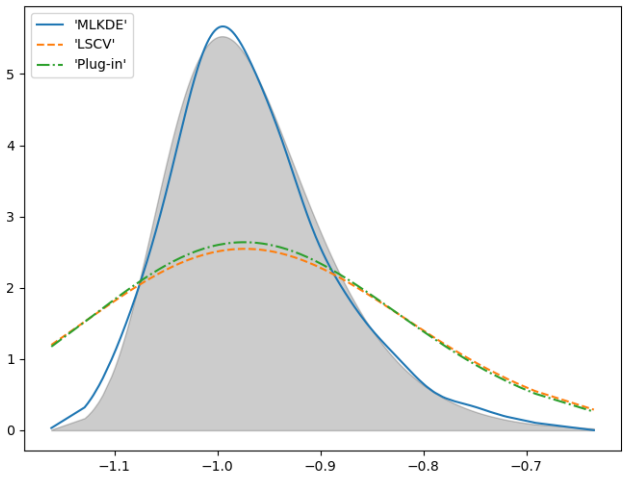


Figura 15. Representação gráfica dos resultados da Tabela 14. Para esse experimento, foi utilizado uma distribuição weibull com α = 15. No MLKDE, os valores dos parâmetros utilizados foram ϵ = 1e-3 e B = 100. Fonte: Autoria própria (2024).

5. Conclusões e trabalhos futuros

A função de densidade de probabilidade é uma ferramenta fundamental na teoria das probabilidades e estatística. Ela descreve a verossimilhança de uma variável aleatória contínua fique entre um determinado intervalo de valores [Devore and Berk 2007]. Foi apresentado um novo método eficiente e simples para estimar a FDP utilizando estimativa de densidade por kernel. Foram realizados experimentos usando dados sintéticos de modo a definir o desempenho do método proposto.

Os experimentos conduzidos consistiram de comparar o desempenho do método proposto em relação a dois métodos da literatura - o LSCV e o Plug-in - na tarefa de gerar a estimativa de densidade que melhor se ajustasse aos dados utilizados. Quatro conjuntos de dados sintéticos foram produzidos, utilizando parâmetros diferentes para cada uma das seguintes distribuições: Beta, pareto e weibull e um experimento para a distribuição gaussiana. Totalizando 13 experimentos, portanto. Para avaliar os métodos, foram utilizadas as métricas EAM, MSE e MAE, onde as duas últimas são métricas clássicas da área.

Nos testes realizados utilizando as distribuições pareto, a performance do método proposto foi bem melhor que a dos demais métodos. Em média, 16.81% para a EAM, 69.0% para a MSE e 66.48% para a MAE. Para as distribuições beta, a performance do MLKDE também foi superior aos demais métodos. Para a EAM o resultado foi, em média, 38.25% melhor. 74.38% melhor para a MSE e 63.89% para a MAE. Por fim, para as distribuições weibull, o método proposto performou em média 73.57% melhor que os demais métodos para a Métrica EAM. Para a MSE e MAE os valores foram 90.59% e 77.37%, respectivamente.

O MLKDE demonstrou ser capaz de obter um ótimo desempenho em diferentes cenários de uso de distribuições de probabilidade. Isso significa que ele se mostrou

robusto e adaptável, independentemente das características específicas dos dados analisados. Desta maneira, alcançamos um resultado adequado em relação ao que o MLKDE se dispõe, que é a determinação da melhor largura de banda a ser utilizada na estimação de densidade de dados por estimador de densidade de kernel.

Como trabalhos futuros, propõe-se avaliar o desempenho do método proposto em relação a ainda mais métodos da literatura, além de se utilizar outras métricas para avaliar tais métodos. Dado que a aplicação direta do método proposto é a sua utilização no cálculo mais preciso de probabilidades, também seria importante verificar o impacto de sua performance em modelos que possuem uma alta demanda de cálculo de probabilidades, como é o caso de redes Bayesianas gerais. O uso de programação concorrente também poderia ser empregada numa versão futura de sua implementação de modo a otimizá-lo e aumentar o desempenho computacional das buscas.

Referências

Botchkarev, A. (2018). Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio.

Chen, Y. C. (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187.

Chu, C., Henderson, D., and Parmenter, C. (2015). Plug-in bandwidth selection for kernel density estimation with discrete data. *Econometrics*, 3(2):199–214.

Devore, J. L. and Berk, K. N. (2007). *Modern Mathematical Statistics with Applications*. Thomson Brooks/Cole.

Duong, T. (2004). *Bandwidth Selectors For Multivariate Kernel Density Estimation*. PhD thesis, Bulletin of The Australian Mathematical Society.

Eliason, S. R. (1993). *Maximum Likelihood Estimation: Logic and Practice*. Sage Publications, Inc., Newbury Park, CA, 1st edition. Quantitative Applications in the Social Sciences, No. 07-096.

Held, L. and Bové, D. S. (2014). *Applied Statistical Inference Likelihood and Bayes*. Springer, 1 edition.

Horne, J. S. and Garton, E. O. (2006). Likelihood cross-validation versus least squares cross-validation for choosing the smoothing parameter in kernel home-range analysis. *Journal of Wildlife Management*, 70(1):641–648.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407.

Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076.

Scott, D. and Terrell, G. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82(1):1131–1146.

Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization*. Wiley Series in Probability and Statistics. Wiley.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, 53(3):683–690.

Siloko, I. U., Siloko, E. A., Ikpotokin, O., Ishiekwene, C. C., and Afere, B. A. (2019). On asymptotic mean integrated squared error's reduction techniques in kernel density estimation. *International Journal of Computational and Theoretical Statistics*, 6(1):5.

Silverman, B. (1986). Density estimation for statistics and data analysis. routledge, abingdon.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer Series in Statistics. Springer.

Turlach, B. (1999). Bandwidth selection in kernel density estimation: A review. Technical report.