Universidade Federal de Juiz de Fora
Engenharia
Bacharelado em Engenharia Elétrica

Rafael Mascarenhas Costa

Estudo de diferentes métodos de discretização aplicados a estimadores de densidade

Trabalho de Conclusão de Curso

Juiz de Fora
2018

Rafael Mascarenhas Costa

Estudo de diferentes métodos de discretização aplicados a estimadores de densidade

Trabalho Final de Curso em Engenharia Elétrica, área de concentração: Sistemas Eletrônicos, da Faculdade de Engenharia da Universidade Federal de Juiz de Fora como requisito para obtenção do grau de Bacharel em Engenharia Elétrica.

Orientadores: Prof. Rafael Antunes Nóbrega, D.Sc.
Igor Abritta Costa

Juiz de Fora
2018

Costa, Rafael Mascarenhas

Estudo de diferentes métodos de discretização aplicados a estimadores de densidade/ Rafael Mascarenhas Costa. - 2018.

107 f. : il.

Dissertação (Trabalho de Conclusão de Curso) - Universidade Federal de Juiz de Fora, 2018

1. Identificação de Elétrons. 2. *Likelihood*. 3. KDE Multivariado I. Título.

CDU 621.3.0

Rafael Mascarenhas Costa

Estudo de diferentes métodos de discretização aplicados a estimadores de densidade

Trabalho Final de Curso em Engenharia Elétrica, área de concentração: Sistemas Eletrônicos, da Faculdade de Engenharia da Universidade Federal de Juiz de Fora como requisito para obtenção do grau de Bacharel em Engenharia Elétrica.

Aprovada em 06 de Setembro de 2018.

BANCA EXAMINADORA:

Prof. Rafael Antunes Nóbrega, D.Sc.
Universidade Federal de Juiz de Fora, UFJF
Orientador

David Melo Souza, M.Sc.
Universidade Federal de Juiz de Fora, UFJF

Prof. Luciano Manhaes Andrade, D.Sc.
Universidade Federal de Juiz de Fora, UFJF

Aos meus pais, familiares e amigos.

AGRADECIMENTOS

O homem não é nada em si mesmo. Não passa de uma probabilidade infinita. Mas ele é o responsável infinito dessa probabilidade.

Albert Camus

RESUMO

Palavras-chave:

ABSTRACT

Keywords:

LISTA DE ILUSTRAÇÕES

Figura 1 Caso representativo de estimação de PDF utilizando 25 pontos para dois intervalos diferentes (-4,4) e (-10,10) 22

Figura 2 Ilustração da curva Gaussiana com média $\mu = 0$ e desvio padrão $\sigma = 1$. 23

Figura 3 Ilustração do método *Linspace* aplicado à uma distribuição normal. 24

Figura 4 Ilustração da discretização da distribuição Gaussiana baseada em sua CDF. 25

Figura 5 Ilustração da discretização da distribuição Gaussiana baseada em sua PDF. 25

Figura 6 PDF Gaussiana e sua primeira derivada à esquerda. Ilustração da discretização da distribuição Gaussiana baseada na CDF da sua primeira derivada à direita. 26

Figura 7 PDF Gaussiana e sua segunda derivada à esquerda. Ilustração da discretização da distribuição Gaussiana baseada na CDF da sua segunda derivada à direita. 27

Figura 8 Caso representativo, variáveis de identificação de elétrons, do experimento CMS: variável $\Delta\eta$, forma do chuveiro $\sigma_{\eta\eta}$ e distribuição de energia-momento $1/E_{SC} - 1/p$. Extraído de (COLLABORATION et al., 2015). 29

Figura 9 Caso representativo, variáveis de identificação de elétrons, do experimento ATLAS, no calorímetro, formato do chuveiro, apresentados separadamente para sinal e os vários tipos de ruídos de fundo. As variáveis apresentadas são: (a) vazamento hadrônico R_{had}, (b) de largura em eta no segundo W_2 amostragem, (c) R_η, (d) largura em η nas $w_{s,tot}$, pequeno, e (e) E_{ratio}. Extraído de (ALISON, 2014). 30

Figura 10 Ilustração das curvas Lognormais construidas com diferentes parâmetros: (a) possui $\sigma = 0.01$; (b) possui $\sigma = 0.25$; (c) possui $\sigma = 0.5$; (d) possui $\sigma = 1$; (e) possui $\sigma = 1.25$; e (f) possui $\sigma = 1.5$ 31

Figura 11 Diagrama de blocos do algoritmo para validação dos métodos de discretização. 31

Figura 12 Discretização utilizando o método de *Linspace*: (a) N(0,1) com $N = 15$, (b) N(0,1) com $N = 25$, (c) L(0,1) com $N = 15$ e (a) L(0,1) com $N = 25$. 33

Figura 13 Discretização utilizando o método de *CDFm*: (a) N(0,1) com $N = 15$, (b) N(0,1) com $N = 25$, (c) L(0,1) com $N = 15$ e (a) L(0,1) com $N = 25$. 34

Figura 14 Discretização utilizando o método de *PDFm*: (a) N(0,1) com $N = 15$, (b) N(0,1) com $N = 25$, (c) L(0,1) com $N = 15$ e (a) L(0,1) com $N = 25$. 35

Figura 15 Discretização utilizando o método de *iPDF1*: (a) N(0,1) com $N = 15$, (b) N(0,1) com $N = 25$, (c) L(0,1) com $N = 15$ e (a) L(0,1) com $N = 25$. 36

Figura 16 Discretização utilizando o método de *iPDF2*: (a) N(0,1) com $N = 15$, (b) N(0,1) com $N = 25$, (c) L(0,1) com $N = 15$ e (a) L(0,1) com

$N = 25$. 37

Figura 17 Gráfico do tempo de processamento de um algoritmo de estimação de densidades baseado em KDE quando aumenta-se o número de eventos a serem estimados e o número de pontos de estimação. 38

Figura 18 Ilustração da medida de erro entre a PDF Real e a Estimada com 20 regiões de interesse. 39

Figura 19 Caso representativo com 200 pontos, 100 Regiões de Interesse (do inglês, *Regions of Interest*) (RoI) e usando a interpolação pelo vizinho mais próximo. 41

Figura 20 Caso representativo com 200 pontos, 100 RoI e usando a interpolação linear. 42

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CDF Função de Distribuição Cumulativa (do inglês, *Cumulative Distribution Function*)

CDFm Método CDF (do inglês, *CDF method*)

CERN Centro Europeu de Pesquisa Nuclear, (do francês, *Conseil Européen pour la Recherche Nucléaire*)

IAE Erro Absoluto Integrado (do inglês,*Integrated Absolute Error*

iPDF1 Integral da distribuição da primeira derivada da PDF

KDE Estimação de Densidade de Núcleo, (do inglês, *Kernel Density Estimation*)

LHC Grande Colisor de Hádrons (do inglês, *Large Hadron Collider*)

MVA Análise Multivariada, (do inglês, *Multivariate Analysis*)

PDF Função de Densidade de Probabilidade (do inglês, *Probability Density Function*)

PDFm Método PDF (do inglês, *PDF method*)

RoI Regiões de Interesse (do inglês, *Regions of Interest*)

SUMÁRIO

1 INTRODUÇÃO **17**

1.1 Motivação 18

1.2 O que foi feito 19

1.3 Estrutura do Trabalho 20

2 DISCRETIZAÇÃO **21**

2.1 Métodos de Discretização 22

2.1.1 *Linspace* 23

2.1.2 *CDFm* 24

2.1.3 *PDFm* 25

2.1.4 *iPDF1* 26

2.1.5 *iPDF2* 27

3 DESENVOLVIMENTO **28**

3.1 Conjunto de dados 28

3.2 Algoritmo 31

3.3 Demostração dos métodos de discretização 32

3.3.1 Método *Linspace* 32

3.3.2 Método *CDFm* 34

3.3.3 Método *PDFm* 35

3.3.4 Método *iPDF1* 36

3.3.5 Método *iPDF2* 37

3.4 Custo Computacional 37

3.5 Ambiente de Análise 38

4 Resultados **40**

4.1 Estimação de erro pela interpolação do vizinho mais próximo 40

4.2 Estimação de erro pela interpolação linear 42

4.3 Estimação de erro considerando *outliers* 42

5 Conclusão **43**

Referências **44**

Apêndice A – Lista de Publicações **46**

A.1 Publicações em Anais de Congresso Internacional 46

1 INTRODUÇÃO

A crescente evolução tecnológica vem possibilitando o desenvolvimento de muitas áreas do conhecimento, sendo uma delas a engenharia elétrica, mais precisamente a análise multivariada (VICINI; SOUZA, 2005) que torna-se cada vez mais uma ferramenta importante para a solução de problemas ligados à estimação de densidades e seleção de eventos, tanto no ambiente industrial quanto em laboratórios de pesquisa. Entretanto, tais problemas podem ocorrem em outras áreas do conhecimento, sendo assim, o esforço em prol da otimização dessas ferramentas de maneira multidisciplinar é de grande interesse.

Nas últimas décadas, a importância de uma modelagem estocástica por Função de Densidade de Probabilidade (do inglês, *Probability Density Function*) (PDF), utilizando-se de métodos não paramétricos teve um crescimento considerável devido ao fato de que vários experimentos geradores de enorme quantidade de dados foram iniciados. Os experimentos ligados ao Grande Colisor de Hádrons (do inglês, *Large Hadron Collider*) (LHC) representam alguns deles. Desde a criação do Centro Europeu de Pesquisa Nuclear, (do francês, *Conseil Européen pour la Recherche Nucléaire*) (CERN), físicos e engenheiros de diferentes países têm trabalhado em conjunto para investigar questões referentes ao estado da arte da ciência fundamental relacionada à física de altas energias, usando instrumentos científicos complexos para estudar os constituintes básicos da matéria e suas interações. No complexo principal do CERN, o LHC, prótons são colocados em um acelerador que os faz colidir quase à velocidade da luz. Este processo permite estudar como as partículas interagem e fornece uma visão das leis fundamentais da natureza (CERN, 2015).

Atualmente, a física experimental de altas energias é um ramo da ciência em progressiva expansão e pode ser considerada um dos campos científicos mais exigentes em termos de processamento de sinal, esse fato é explicado devido aos eventos de interesse serem raros e contaminados com alto nível de ruído de fundo, demandando sistemas cada vez mais otimizados no que diz respeito a tempo de processamento, eficiência de

detecção de sinal e rejeição de ruído.

Com o objetivo de observar os subprodutos dessas colisões, é necessário usar detectores; basicamente, sensores que, trabalhando em conjunto, são capazes de medir algumas características dos subprodutos das colisões e transformá-los em sinais elétricos que podem ser armazenados e utilizados em estudos relacionados a física de altas energias.

Em geral, para problemas cujas variáveis podem ser modeladas, a estimação das mesmas se torna paramétrica. No entanto, é muito importante enfatizar que, devido à complexidade do problema, suas variáveis podem não ser descritas com as funções de densidade de probabilidade conhecidas na literatura. Sendo assim, a aplicação de métodos não paramétricos se espalhou consideravelmente nos últimos anos devido às ferramentas recentemente desenvolvidas para análise estatística. Tais métodos fornecem um caminho alternativo a estimação paramétrica e possibilita o estudo de grandes quantidades de dados, essa linha de pesquisa torna-se objeto significativo de estudo, uma vez que contempla pesquisas teóricas e práticas com relação direta a temas como regressão, discriminação e reconhecimento de padrões.

Neste contexto, o presente trabalho visa avaliar os erros inseridos pelo processo de discretização propondo diferentes métodos e olhando diretamente à sua performance de estimação, considerando as interpolações pelo vizinho mais próximo e linear. O impacto pelos pontos longe da região de alta probabilidade também são avaliados uma vez que é um problema comum na estimação de PDF.

1.1 MOTIVAÇÃO

A reconstrução e seleção de elétrons é de grande importância em muitas análises realizadas em experimentos de Física de Partículas, como é o caso do detector CMS e ATLAS que utilizam o LHC; o Belle (HANAGAKI et al., 2002) que usa o KEK-B (KEK, 2018); entre outros. No caso dos experimentos ATLAS e CMS a identificação correta dos elétrons é exigida para medições precisas do modelo padrão (RONQUI, 2015), medições e pesquisas em busca do Bóson de Higgs, e pesquisas de processos que vão além do modelo padrão. Essas análises científicas exigem uma excelente resolução de momento e pequenas incertezas sistemáticas. É possível alcançar um alto nível de desempenho desses algoritmos fazendo uso de algumas etapas de processamento, evoluindo a partir dos algoritmos iniciais de reconstrução eletrônica desenvolvidos no contexto da seleção *online* até algoritmos mais complexos no contexto de seleção *offline*.

Os princípios básicos da reconstrução de elétrons nesses detectores dependem de uma combinação da energia medida no calorímetro eletromagnético e o impulso medido no detector de traço.

Diversas estratégias podem ser usadas para identificar elétrons isolados (chamados de sinal), e separá-los de fontes de ruído de fundo, originários principalmente de conversões de fótons, jatos erroneamente identificados como elétrons, ou elétrons de decaimentos semi-leptônicos de quarks b e c. Algoritmos simples e robustos são desenvolvidos para aplicar seleções sequenciais em um conjunto discriminantes. Algoritmos mais complexos combinam variáveis em uma análise de Análise Multivariada, (do inglês, *Multivariate Analysis*) (MVA) para alcançar uma melhor discriminação. E uma das abordagens que tem ganhado força, nesse ambiente que requer cada vez mais precisão e desempenho, é o uso de métodos de classificação via probabilidade estatística, sendo que esses métodos tem como maior desafio a estimação de densidades que não podem ser parametrizadas pelas funções já conhecidas na literatura.

Portanto, na última década muitos trabalhos relacionados ao tema de otimização da estimação de densidade não paramétrica, tanto numérica (SCHINDLER, 2012) quanto computacional (GRAMACKI, 2017), foram publicados, bem como sobre o tema de discretização em processamento de sinais, mostrando que este é um tema que mesmo sendo discutido há décadas ainda é muito utilizado, explorado e em desenvolvimento. Portanto, abre-se a possibilidade de contribuir nessa área no que diz respeito a otimização da estimação de densidades usando como base de desenvolvimento um sistema altamente complexo e distribuições com características bastante distintas, como ocorre com os experimentos do LHC.

1.2 O QUE FOI FEITO

Este trabalho se concentra na otimização da etapa de discretização do processo estimação de densidades não-paramétricas Abordando o problema do ponto de vista somente de estimação, avaliando o erro inserido nesse processo e buscando garantir uma otimização do *trade-off* entre custo computacional e erro inserido. O desempenho dos algoritmos propostos foram avaliados em detalhe e comparados com outros métodos.

1.3 ESTRUTURA DO TRABALHO

Este documento está organizado da seguinte maneira: o Capítulo **??** apresenta uma revisão bibliográfica do tema de discretização e introduz a matemática dos métodos que serão utilizados nesse trabalho. O Capítulo 3 faz uma ambientação do meio onde está inserido esse trabalho e detalha o funcionamento do algoritmo de avaliação dos métodos propostos. O Capítulo 4 traz os resultados utilizando-se os métodos propostos, as comparações e a analise do que foi pesquisado. Por fim, as conclusões e os próximos passos para a continuidade desse trabalho serão apresentados no Capítulo 5.

2 DISCRETIZAÇÃO

A estimação de densidade via Estimação de Densidade de Núcleo, (do inglês, *Kernel Density Estimation*) (KDE) de uma série de medidas contínuas, por razões computacionais, é geralmente representado na forma discreta. Consequentemente, estimações diretas acontecem apenas para valores discretizados (JONES, 1989) e interpolação é usado para solucionar qualquer outro valor que possa vir fora durante as mediações. Este processo insere erros de estimação os quais podem ser minimizados incrementando o número de pontos a serem estimados, buscando um equilíbrio entre otimização computacional e performance de estimação.

Vários autores seguem a mesma abordagem, como em (JONES, 1989), explorando os diferentes aspectos do processo de discretização e propondo novos métodos no intuito de minimizar as adversidades relatadas. Por exemplo, em (FAYYAD; IRANI, 1993) o método bem conhecido Ent-MDLP é proposto; em (FRIEDMAN; GOLDSZMIDT et al., 1996) é sugerido um algoritmo de discretização baseado em Redes Bayesianas; em (BIBA et al., 2007) os autores propõem um método não supervisionado para discretização utilizando-se o KDE; também usando o método não supervisionado, os autores de (SCHMIDBERGER; FRANK, 2005) apresentam um estudo de discretização aplicado à estimação de densidade baseado em árvore; e em (ZHANG et al., 2007) um algoritmo de aprendizagem de máquina baseado-se no critério de *Gini* foi estudado.

Estes trabalhos geralmente possuem foco em algoritmos de aprendizagem de máquina ou minimização dos critérios selecionados a fim de otimizar os vários atributos existentes, que como consequência, tendem a ter um alto custo computacional quando submetidos a uma grande quantidade de dados. Além do mais, tais estudos abordam a performance da discretização através do prisma da classificação e alguns como forma de preprocessamento do conjunto de dados.

O método de discretização mais aplicado atualmente é o baseado em espaçamento uniforme entre os pontos estimados. Isso trata de maneira igualitária todas as densidades de região (e. g. a função de densidade nas regiões de baixa probabilidade é

discretizada com a mesma resolução das regiões de alta probabilidade) levando a um erro de estimação que tende a não ser uniforme ao longo de todas as regiões de função de densidade de probabilidade. A figura 1 mostra um exemplo de quando a região de baixa probabilidade é grande devido a eventos fora da curva, neste caso, uma discretização baseada em um espaçamento uniforme pode colocar um grande número de pontos desnecessários nessa região, fazendo com que, para minimizar o erro de estimação, o número de pontos a ser estimado seja maior a fim de representar bem a região de alta probabilidade.

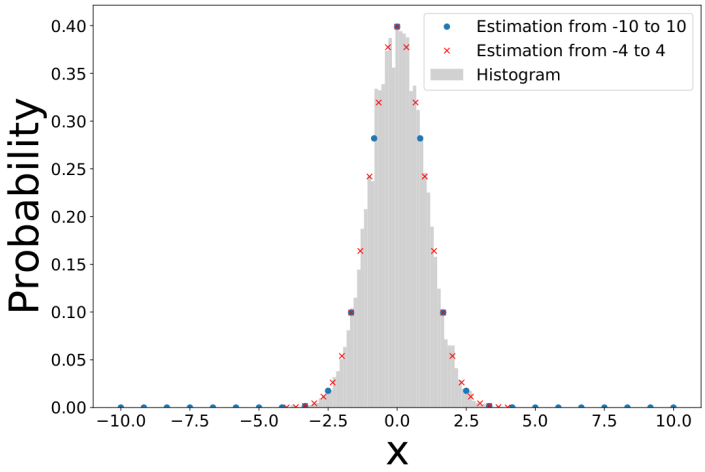


Figura 1: Caso representativo de estimação de PDF utilizando 25 pontos para dois intervalos diferentes (-4,4) e (-10,10)

2.1 MÉTODOS DE DISCRETIZAÇÃO

Para se estudar os efeitos da discretização no processo de estimação de PDF, a performance de cinco diferentes métodos serão confrontados, como listados abaixo:

- *Linspace*;
- *CDFm*;
- *PDFm*;
- *iPDF1*;
- *iPDF2*.

Estes cinco métodos serão demostrados a priori utilizando-se uma distribuição Gaussiana com média $\mu = 0$ e desvio padrão $\sigma = 1$, cuja PDF pode ser descrita

pela Equação (3.1) e ilustrada na Figura 2 com o numero de pontos $N = 25$ para uma melhor visualização.

$$f_X(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.1}$$

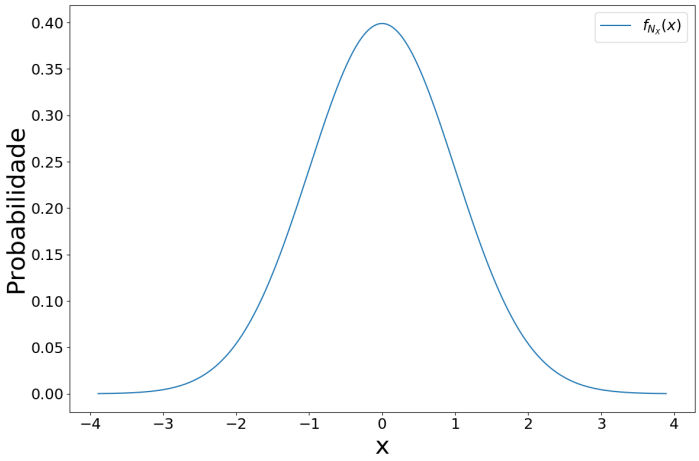


Figura 2: Ilustração da curva Gaussiana com média $\mu = 0$ e desvio padrão $\sigma = 1$.

2.1.1 LINSPACE

O método *Linspace* é caracterizado por amostrar de maneira uniforme a variável aleatória, representada pelo eixo das abscissas de uma PDF qualquer. Após, o eixo x terá N pontos igualmente espaçados entre dois valores predefinidos que definem os parâmetros de início e término da distribuição. Este método é o mais utilizado na literatura devido a sua simplicidade. A Figura 3 ilustra o método *Linspace* para a distribuição Normal, limitando o eixo horizontal à uma área de probabilidade de $99,99\%$.

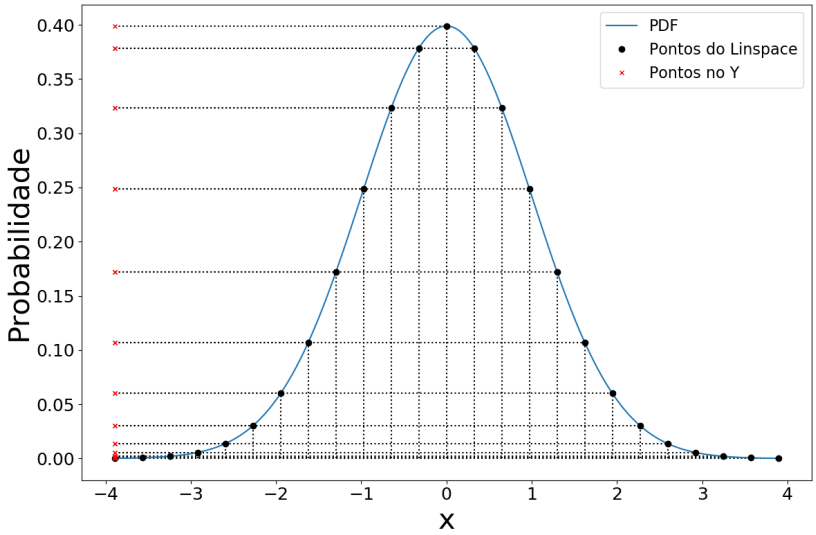


Figura 3: Ilustração do método *Linspace* aplicado à uma distribuição normal.

2.1.2 CDFM

O método denominado nesse trabalho de Método CDF (do inglês, *CDF method*) (CDFm) representa a discretização baseada na Função de Distribuição Cumulativa (do inglês, *Cumulative Distribution Function*) (CDF) descrita pela Equação (2.2) para uma variável contínua e (2.3) para o caso de uma variável discreta possuindo valores em b. Para este método, no caso de se ter a função geradora, primeiramente calcula-se a CDF da distribuição e então faz-se uma distribuição linear de pontos no eixo y e encontra-se os seus respectivos valores para o eixo x fazendo sua função inversa, conforme mostra a Equação (2.4) e é ilustado pela Figura 4.

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \tag{2.2}$$

$$P(X = b) = F_X(b) - \lim_{x \to b^-} F_X(x) \tag{2.3}$$

$$CDFm(y) = F_X^{-1}(x) \tag{2.4}$$

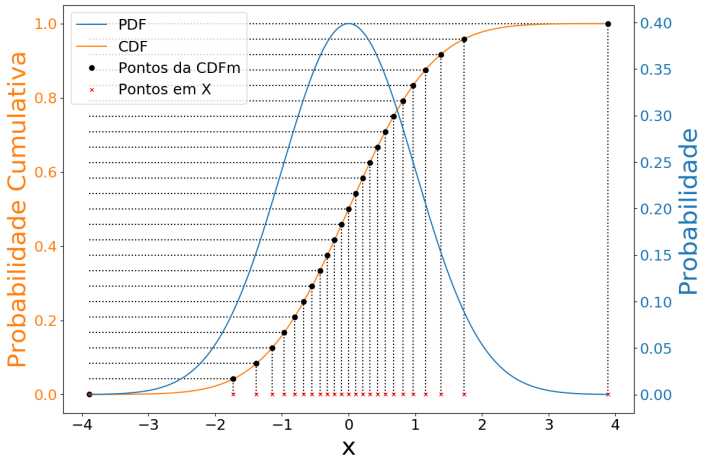


Figura 4: Ilustração da discretização da distribuição Gaussiana baseada em sua CDF.

É possível notar que, quanto maior a probabilidade da CDF, maior o número de pontos em sua região.

2.1.3 PDFM

Este método, denominado de Método PDF (do inglês, *PDF method*) (PDFm) também usa a técnica de reflexão aplicada ao método da *CDFm*, mas a função de referência é a própria PDF, ao invés da sua CDF, com isso, os pontos de intercessão da curva com os valores em y são calculados. A Figura 5 mostra como este método funciona.

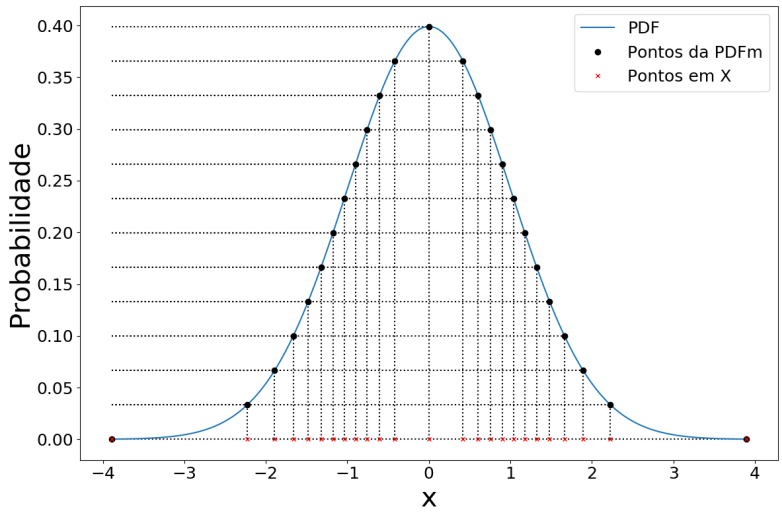
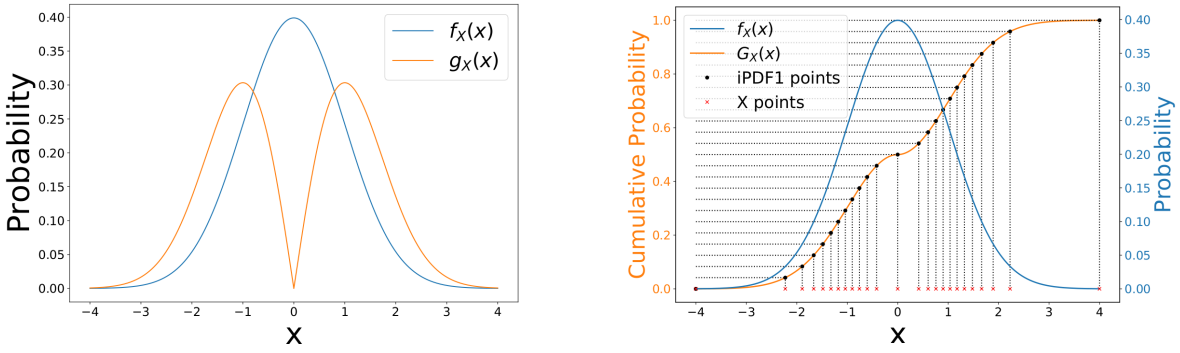


Figura 5: Ilustração da discretização da distribuição Gaussiana baseada em sua PDF.

Ela possui o efeito de incrementar o número de pontos estimados onde a inclinação da curva é mais acentuada.

2.1.4 IPDF1

O método da Integral da distribuição da primeira derivada da PDF (iPDF1) reflete os valores verticais para o eixo horizontal usando a CDF da primeira derivada da PDF como uma transformação de base, como é ilustrado nas Figuras 6a e 6b



(a) (b)

Figura 6: PDF Gaussiana e sua primeira derivada à esquerda. Ilustração da discretização da distribuição Gaussiana baseada na CDF da sua primeira derivada à direita.

As equações (2.5) e (2.6) descrevem este método matematicamente.

$$\begin{aligned} \zeta(x) &= \frac{|\mu - x|}{\sigma^3\sqrt{2\pi}} \cdot e^{\left(\frac{-(\mu-x)^2}{2\sigma^2}\right)} \\ \int_{-\infty}^{\infty} \zeta(x) \cdot dx &= c_1 \\ g_X(x) &= \frac{\zeta(x)}{c_1} \end{aligned} \tag{2.5}$$

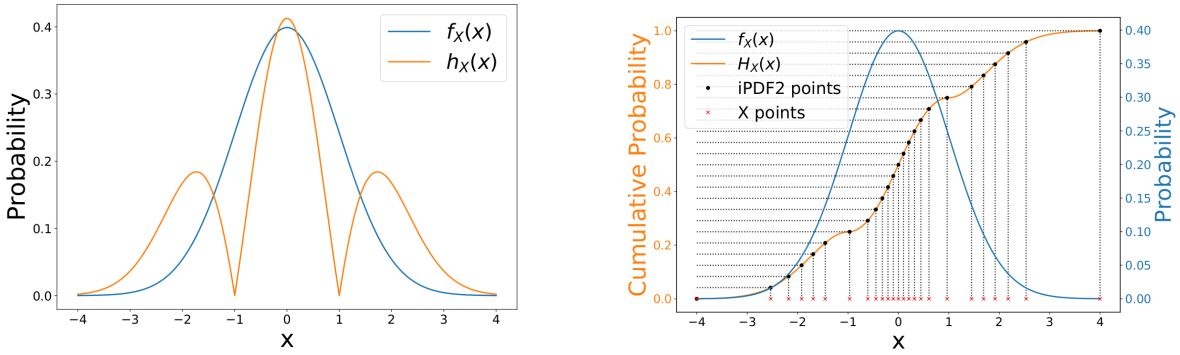
onde ζ é a equação da distribuição da derivada da distribuição normal, μ é a média, σ o desvio padrão, x a variável aleatória, c_1 é a área abaixo da curva ζ, e g_X é a versão normalizada. A CDF de g_X ($G_X(x)$) é usada para transferir os valores da abscissa ao eixo da ordenada como mostra a Figura 6b.

$$G_X(x) = \int_{-\infty}^{x} g_X(y) \cdot dy \tag{2.6}$$

É possível notar que este método consegue fazer uma melhor estimação nas regiões em que a primeira derivada de sua PDF são maiores.

2.1.5 *IPDF2*

Este método é construído da mesma maneira da *IPDF1* mas usando a segunda derivada ao invés da primeira, como é mostrado na Figura 7a e 7b. Suas equações são mostradas em (2.7) e (2.8).



(a) (b)

Figura 7: PDF Gaussiana e sua segunda derivada à esquerda. Ilustração da discretização da distribuição Gaussiana baseada na CDF da sua segunda derivada à direita.

$$\eta(x) = \frac{|\sigma^2 - (\mu - x)^2|}{\sigma^5\sqrt{2\pi}} \cdot e^{-\frac{(\mu-x)^2}{2\sigma^2}}$$
$$\int_{-\infty}^{\infty} \eta(x) \cdot dx = c_2 \tag{2.7}$$
$$h_X(x) = \frac{\eta(x)}{c_2}$$

onde η é a equação de distribuição de segunda derivada da distribuição Normal, c_2 é a área abaixo da curva desta distribuição, e h_X é sua versão normalizada. Finalmente, $H_X(x)$ é a CDF de h_X, dada por (2.8).

$$H_X(x) = \int_{-\infty}^{x} h_X(y) \cdot dy \tag{2.8}$$

Como é possível notar, há uma maior concentração de pontos onde a segunda derivada é maior.

Uma análise mais afunda sobre estes métodos será mostrada nos capítulos abaixo.

3 DESENVOLVIMENTO

Neste capítulo, serão descritos alguns detalhes da construção dos métodos e do algoritmo para avaliação de sua performance, além disso, serão mostradas as dificuldades encontradas na aplicação prática desses métodos. A discretização da estimação de densidades de variáveis discriminantes pode influenciar de forma direta na tarefa de classificação, entretanto, este trabalho se concentrou somente no impacto desses métodos na estimação, entendendo que um menor erro de estimação pode levar a uma melhor classificação.

3.1 CONJUNTO DE DADOS

Um dos objetivos desse trabalho é otimizar o desempenho dos algoritmos de estimação de densidade via KDE que fazem uso de métodos de discretização e posteriormente essas estimações serão usadas para a identificação e classificação de eventos. Portanto o conjunto de dados aqui escolhido tem por base as estimações que podem ser encontradas em alguns dos experimentos de física de partículas mais importantes atualmente, como o ATLAS e CMS.

Nas Figuras 8 e 9 são mostrados casos representativos de variáveis usadas para a identificação de elétrons nos experimentos CMS e ATLAS, respectivamente. Para o CMS, pode-se ver o perfil dos elétrons e do ruído de fundo, bem como a diferença entre dados reais e simulados. Já na Figura 9 é mostrado também as varias formas de ruído de fundo para elétrons no ATLAS.

Pode-se considerar que a identificação de elétrons em experimentos de física de altas energias é de certa forma similar, respeitando as particularidades de cada detector. Portanto, é de se esperar que a otimização de algoritmos de identificação de elétrons estudada em um conjunto de dados possa ser reproduzida, considerando as especificidades de cada experimento, em um outro conjunto de dados, fazendo com que o estudo para melhoria do desempenho desse processo seja de suma importância.

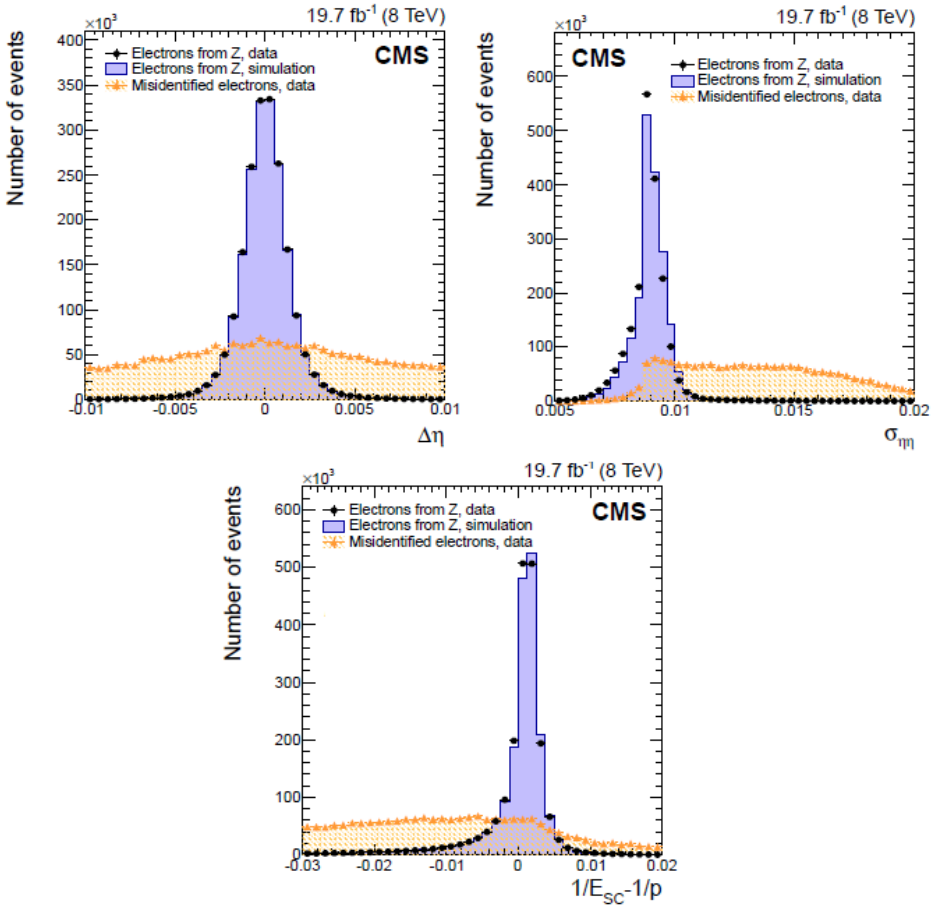


Figura 8: Caso representativo, variáveis de identificação de elétrons, do experimento CMS: variável $\Delta\eta$, forma do chuveiro $\sigma_{\eta\eta}$ e distribuição de energia-momento $1/E_{SC} - 1/p$. Extraído de (COLLABORATION et al., 2015).

Pode-se observar que as variáveis discriminantes destes experimentos apresentam distribuições que muitas vezes se assemelham à algumas distribuições conhecidas na literatura, como a distribuição *Gaussiana* e a *Lognormal*. Sendo assim, com o intuito de realizar as análises em um ambiente controlado foram escolhidas as distribuições *Gaussiana* e *Lognormal*, sendo que a última parametrizada de 6 maneiras diferentes.

A distribuição normal foi definida com média $\mu = 0$ e desvio padrão $\sigma = 1$, ilustrada na Figura 2, devido ao fato de tais parâmetros não interferirem na forma desta distribuição.

$$f_{N_X}(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.1}$$

Já a distribuição *Lognormal* com média $\mu = 0$ e desvio padrão σ com valores $0.01, 0.25, 0.5, 1, 1.25$ e 1.5, descrita pela Equação (3.2) e ilustrada na Figura 10.

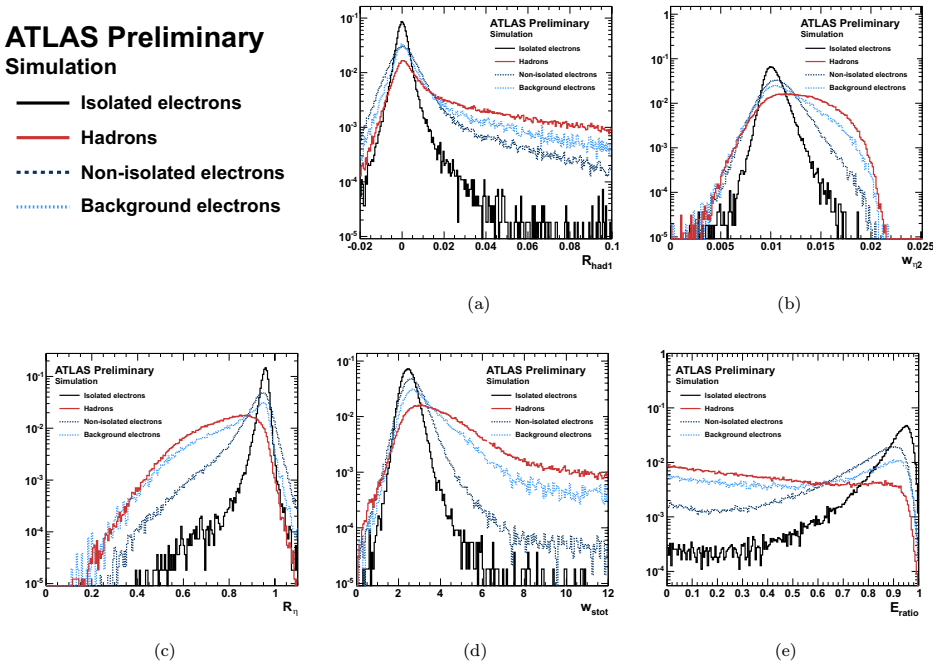


Figura 9: Caso representativo, variáveis de identificação de elétrons, do experimento ATLAS, no calorímetro, formato do chuveiro, apresentados separadamente para sinal e os vários tipos de ruídos de fundo. As variáveis apresentadas são: (a) vazamento hadrônico R_{had}, (b) de largura em eta no segundo W_2 amostragem, (c) R_η, (d) largura em η nas $w_{s,tot}$, pequeno, e (e) E_{ratio}. Extraído de (ALISON, 2014).

$$f_{L_X}(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \cdot e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \tag{3.2}$$

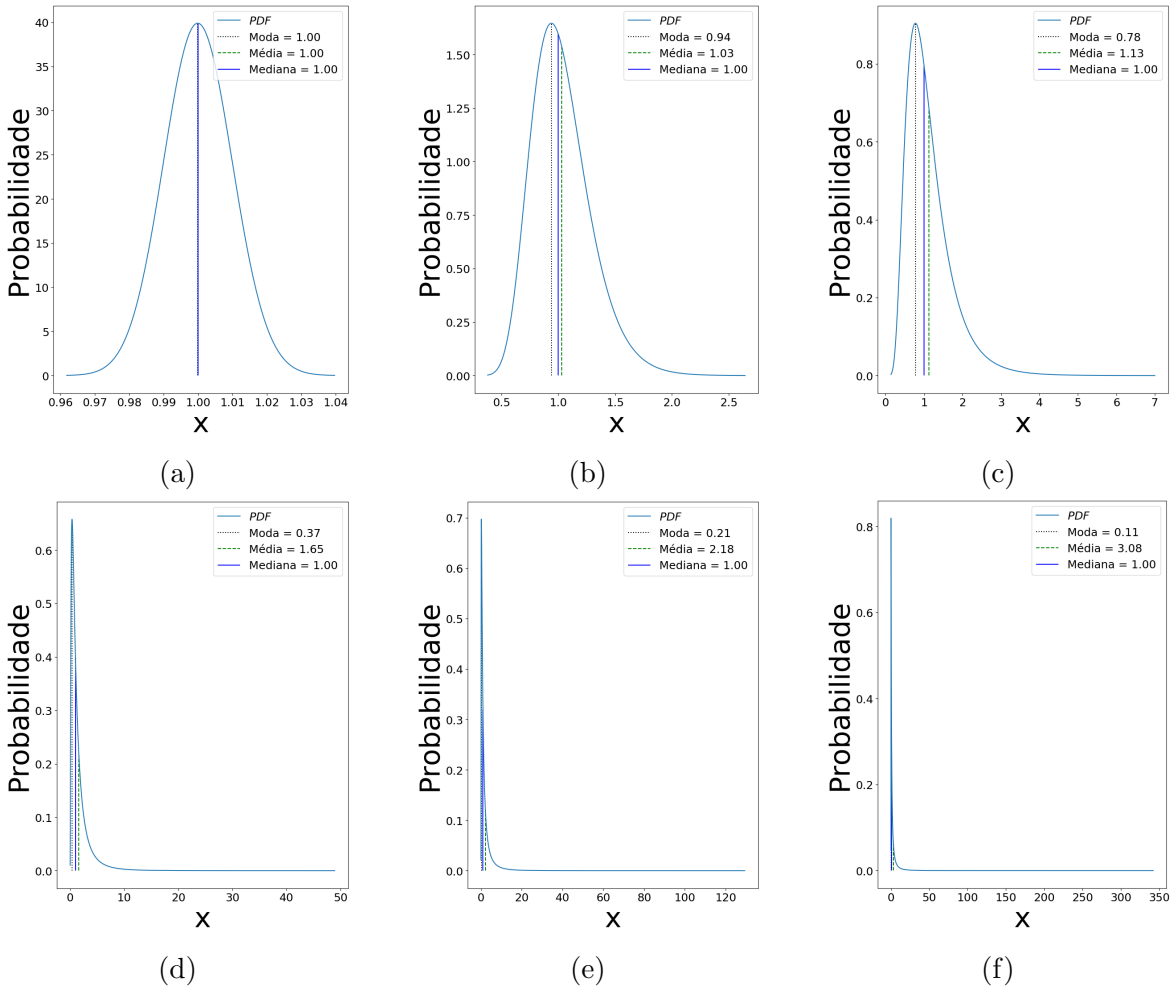


Figura 10: Ilustração das curvas Lognormais construidas com diferentes parâmetros: (a) possui $\sigma = 0.01$; (b) possui $\sigma = 0.25$; (c) possui $\sigma = 0.5$; (d) possui $\sigma = 1$; (e) possui $\sigma = 1.25$; e (f) possui $\sigma = 1.5$

3.2 ALGORITMO

O algoritmo para comparação e validação dos métodos de discretização de estimação construído nesse trabalho pode ser resumido pelo diagrama de blocos da Figura 11, sendo testado de duas maneiras diversas: utilizando somente a função analítica e usando os dados gerados a partir das funções geradoras.

Figura 11: Diagrama de blocos do algoritmo para validação dos métodos de discretização.

Função Analítica Inicialmente uma função geradora é escolhida, sua discretização é feita utilizando os métodos apresentados, essa discretização é utilizada para

calcular todos os pontos da curva original (utilizando interpolação) e por fim faz-se o cálculo da distância entre as duas curvas (analítica e discretizada) utilizando a métrica de distancia chamada L1, que é um caso particular da Equação (3.3).

$$L_p = \left(\int |f - g|^p \right)^{1/p} \tag{3.3}$$

Onde p é o parâmetro a ser escolhido. No caso mais simples, $p = 1$, a equação 3.3 se torna a equação 3.4.

$$L_1 = \int |f - g| \tag{3.4}$$

A equação 3.4 é chamado de Erro Absoluto Integrado (do inglês,*Integrated Absolute Error* (IAE) ou distância L1.

Dados gerados Nesse caso, única diferença é que o calculo da discretização é feito usando como base a distribuição de eventos aleatórios geradas pela função geradora escolhida.

3.3 DEMOSTRAÇÃO DOS MÉTODOS DE DISCRETIZAÇÃO

Com o intuito de validar os algoritmo e os métodos de discretização serão apresentados nessa seção o funcionamento desses métodos para a função gaussiana e a função *lognormal* com $\sigma = X$. E, com o objetivo de demostrar de forma mais clara o efeito desses métodos e sua dependência ao número de pontos de estimação escolhidos, os métodos serão avaliados variando o número de estimação para $N = 15$ e $N = 25$.

3.3.1 MÉTODO LINSPACE

De acordo com o princípio de funcionamento do método de discretização *Linspace* espera-se que este entregue resultados satisfatórios em distribuições com variações mais lentas, como mostrado nas Figuras 12a e 12b. Já para conjunto de dados que apresenta variações mais rápidas, como mostrado nas Figuras 12c e 12d, este método não alcança uma boa representação da PDF.

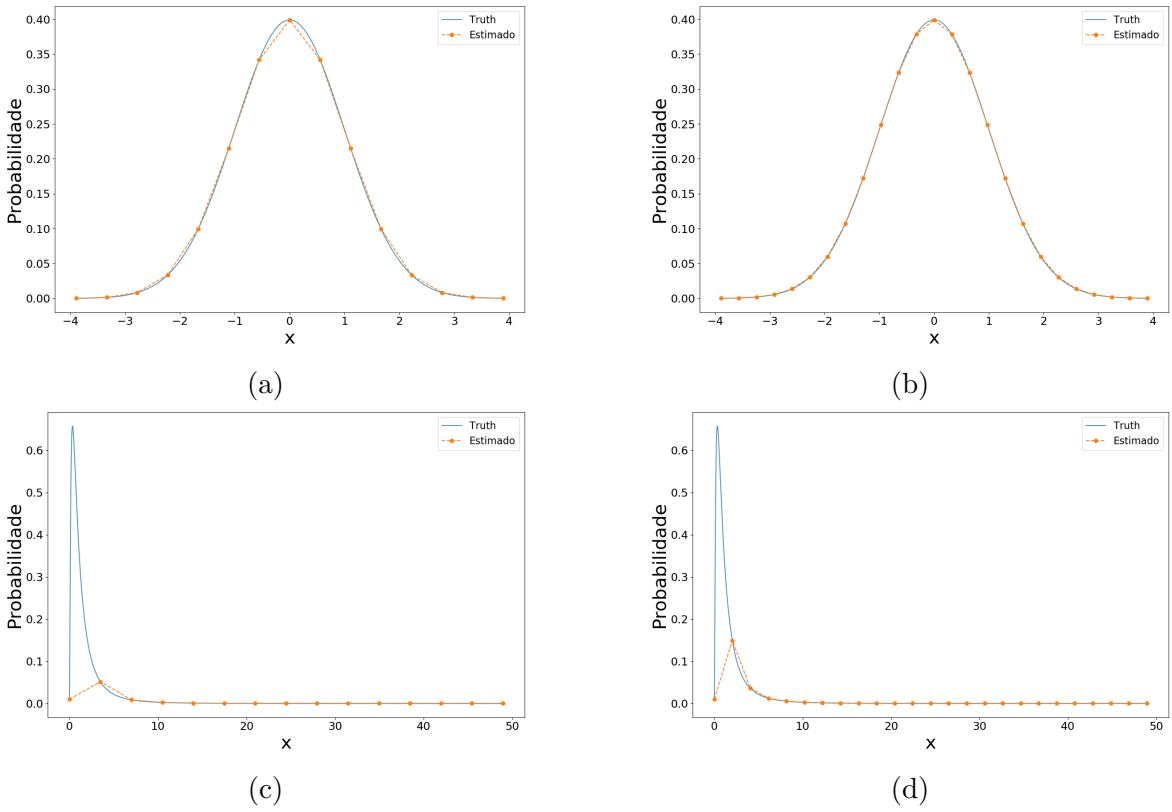


Figura 12: Discretização utilizando o método de *Linspace*: (a) N(0,1) com $N = 15$, (b) N(0,1) com $N = 25$, (c) L(0,1) com $N = 15$ e (a) L(0,1) com $N = 25$.

Entretanto, pode-se perceber que com o aumento do número de pontos de estimação ($N = 15$ para $N = 25$) o desempenho deste método apresenta uma melhora significativa, portanto, fica claro que é possível alcançar uma boa performance com o método de *Linspace*. Mas, há de se comentar que o aumento do número de pontos de estimação é um fator decisivo no custo computacional desses algoritmos, além disso, aumenta a quantidade de informação a ser armazenada ou transmitida.

3.3.2 *MÉTODO CDFM*

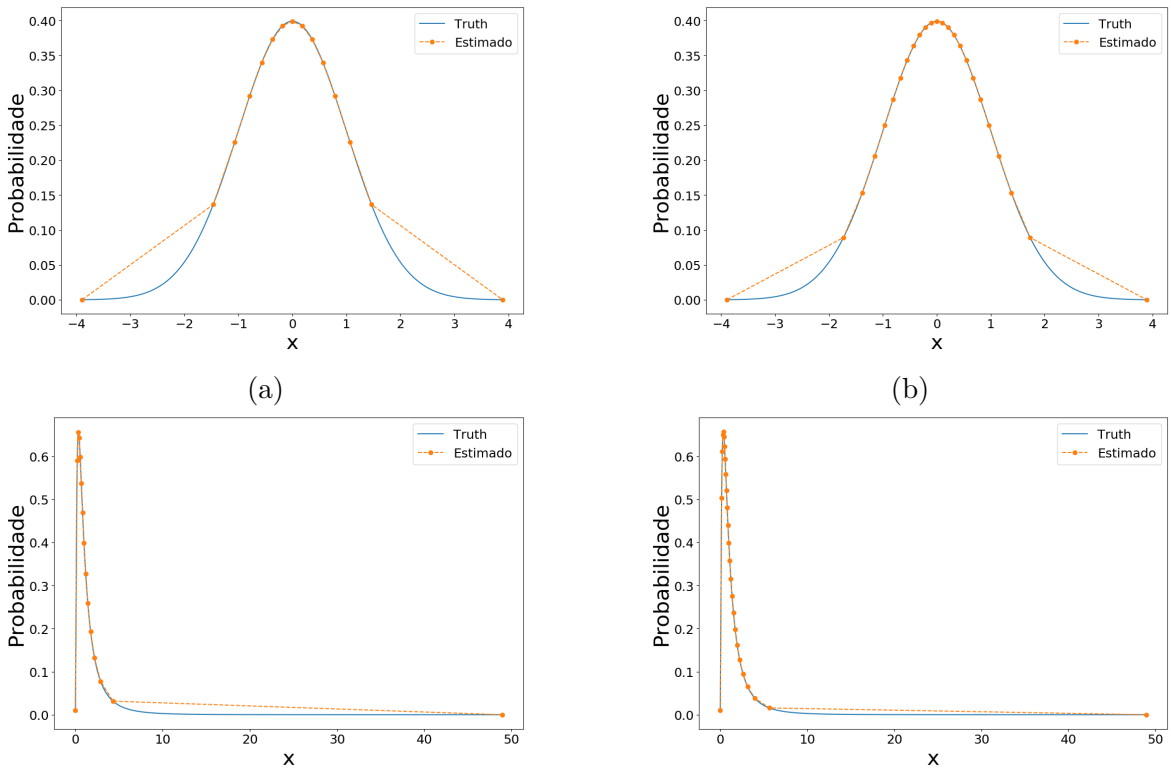


Figura 13: Discretização utilizando o método de *CDFm*: (a) N(0,1) com $N = 15$, (b) N(0,1) com $N = 25$, (c) L(0,1) com $N = 15$ e (a) L(0,1) com $N = 25$.

Vemos que a região de alta probabilidade é representada com um menor erro do que no método *Linspace* mas, em contrapartida, a região de baixa probabilidade necessitaria de um número muito maior de pontos para possuir o mesmo erro do método anterior.

3.3.3 *MÉTODO PDFM*

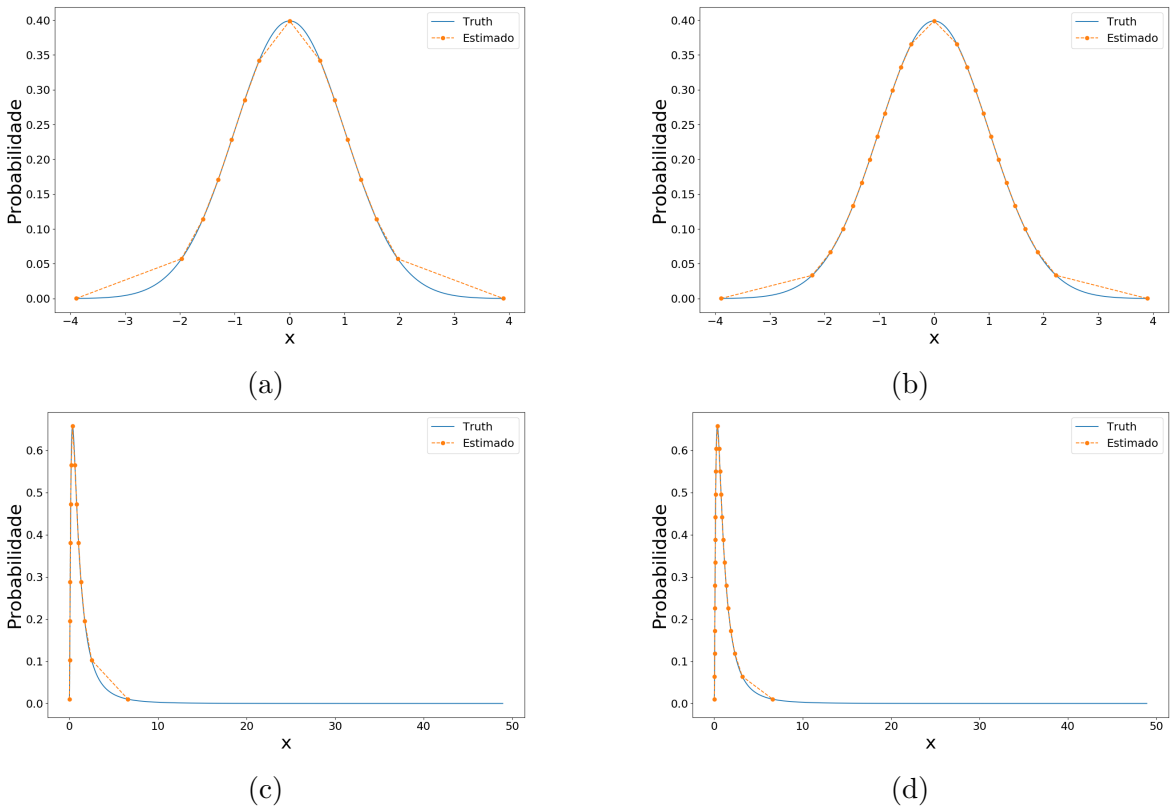


Figura 14: Discretização utilizando o método de *PDFm*: (a) N(0,1) com $N = 15$, (b) N(0,1) com $N = 25$, (c) L(0,1) com $N = 15$ e (a) L(0,1) com $N = 25$.

Para a distribuição Normal, este método apresenta um maior erro na região de alta probabilidade do que o método CDFm mas nas regiões de baixa probabilidade o erro de estimação é menor do que o visto anteriormente, fazendo assim uma combinação dos métodos *Linspace* e *CDFm*.

3.3.4 *MÉTODO IPDF1*

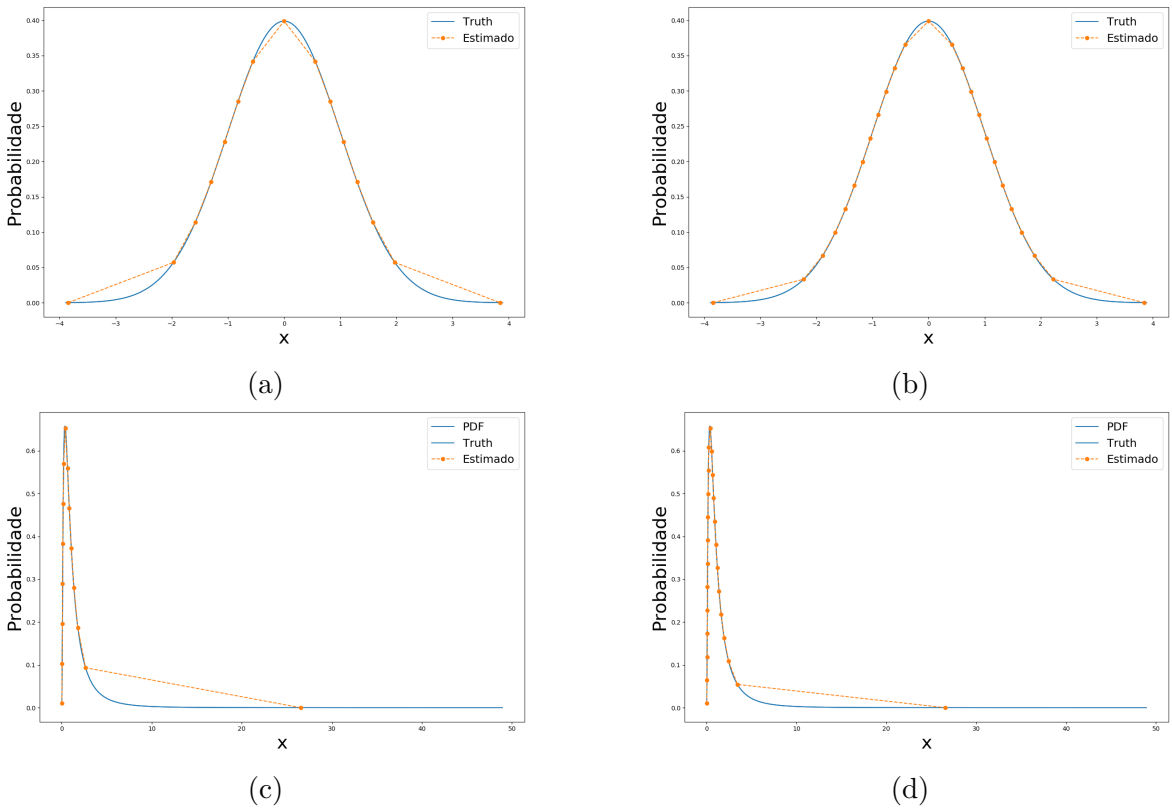


Figura 15: Discretização utilizando o método de *iPDF1*: (a) N(0,1) com $N = 15$, (b) N(0,1) com $N = 25$, (c) L(0,1) com $N = 15$ e (a) L(0,1) com $N = 25$.

3.3.5 MÉTODO IPDF2

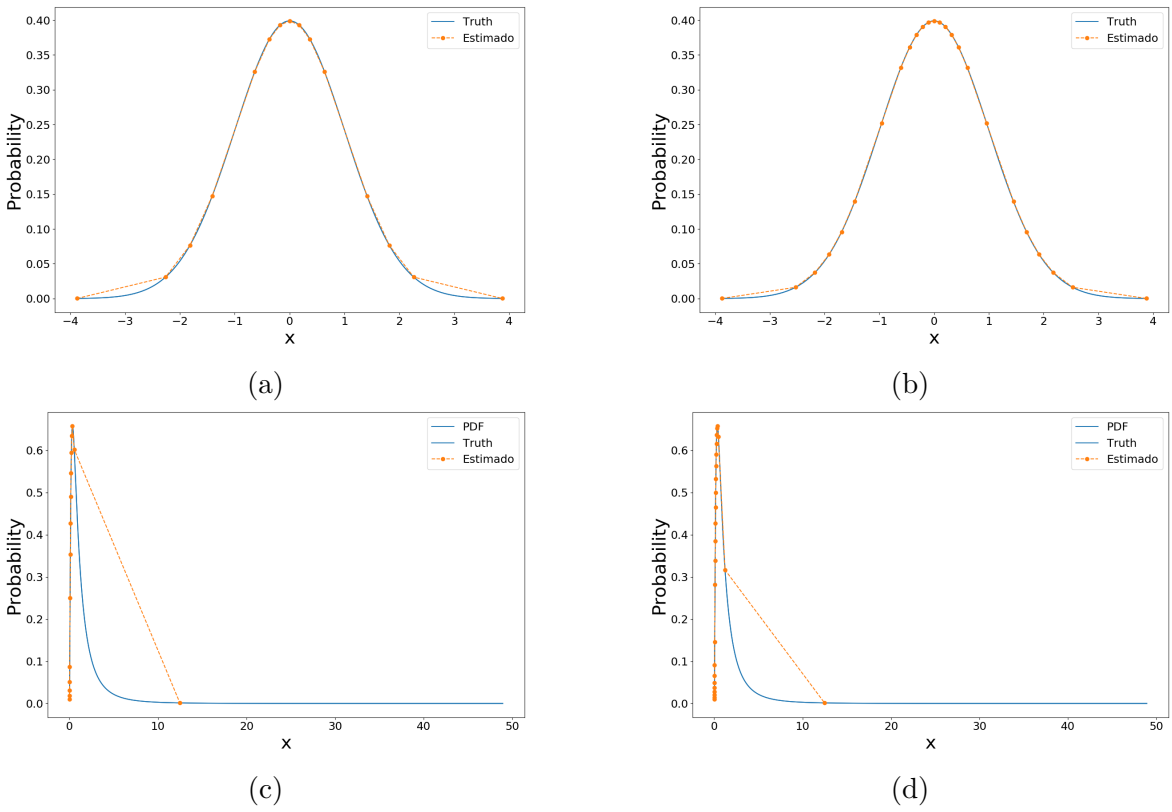


Figura 16: Discretização utilizando o método de *iPDF2*: (a) N(0,1) com $N = 15$, (b) N(0,1) com $N = 25$, (c) L(0,1) com $N = 15$ e (a) L(0,1) com $N = 25$.

3.4 CUSTO COMPUTACIONAL

Os algoritmo de *Kernel* são muito utilizados na literatura no contexto de análise de dados ou modelagem de dados, entretanto é sabido que esse método é computacionalmente mais lento em comparação com outros. Por isso muitos pesquisadores fazem uso de algoritmos que efetuam aproximações matemáticas no intuito de ganhar em custo computacional, chamados de *FastKDE*, ou seja, existe um *trade-off* entre estabilidade numérica e economia computacional. Entretanto, como já mencionado, o tempo de processamento esta diretamente ligado ao número de eventos da distribuição e ao número de pontos a serem estimados.

Portanto, no intuito de ilustrar a consequência de se aumentar o número de pontos de estimação a Figura 17 apresentada o tempo de processamento de um algoritmo matricial de *FastKDE* utilizado para a estimação de uma distribuição gaussiana $N(0, 1)$ ao se variar o número de eventos e número de pontos de estimação.

Pode-se observar que o tempo de processamento para $N = 1024$ é aproximada-

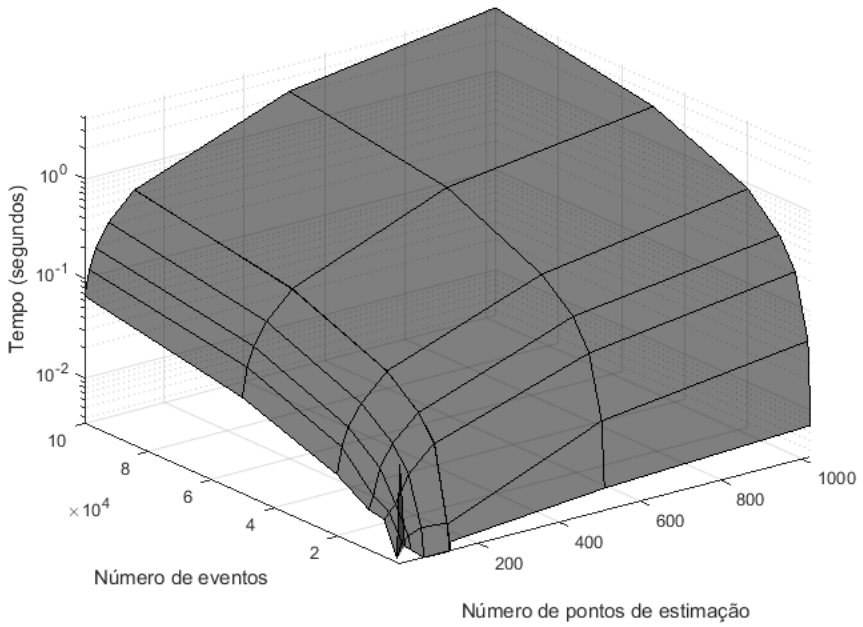


Figura 17: Gráfico do tempo de processamento de um algoritmo de estimação de densidades baseado em KDE quando aumenta-se o número de eventos a serem estimados e o número de pontos de estimação.

mente 75% maior que para $N = 128$ quando o número de eventos é igual a 10^5 e aproximadamente 67% maior para número de eventos igual a 10^4. Ou seja, um método de discretização capaz de apresentar o mesmo erro de estimação mesmo com menos pontos de estimação pode trazer benefícios importantes em ambientes de alta exigência.

3.5 AMBIENTE DE ANÁLISE

Para analisar as diferenças entre a PDF real e estimada ao longo de toda a extensão do eixo das abscissas, a área entre as duas PDFs será usada como medida da estimação de erro. Além do mais, o eixo das abscissas foi dividida em N regiões de mesmo tamanho, chamado RoI (RON, 1999). Essas regiões são compreendidas entre valores máximos e mínimos predefinidos do eixo horizontal. A Figura 18 mostra este processo quando a abscissa é dividida em 20 regiões, todas compreendidas entre os valores -4 e 4 do eixo x.

A maneira que a RoI é usada neste trabalho permitirá avaliar o erro de estimação em função de quatro diferentes parâmetros: Probabilidade; Eixo das abscissas; Primeira e Segunda Derivada. Para estimar os valores entre os pontos discretos, dois métodos de interpolação serão usados: interpolação pelo Vizinho Mais Próximo e Linear. 200

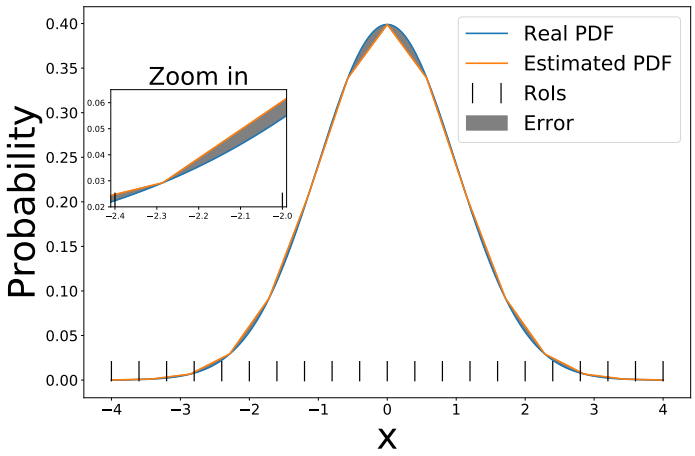


Figura 18: Ilustração da medida de erro entre a PDF Real e a Estimada com 20 regiões de interesse.

amostras serão usadas no processo de discretização. O erro de estimação tende a melhoras conforme o número de amostras aumenta mas sua característica geral não muda. Este último é a principal preocupação deste trabalho.

4 RESULTADOS

Nessa sessão, os resultados serão dados em termos da área medida entre a diferença da PDF real e a estimada, como previamente mencionado. Nas figuras que seguem , os valores serão mostrados no eixo vertical, nomeado de *Erro*. O eixo horizontal mostra quatro perspectivas diferentes: Probabilidade, Eixo x (que seria os valores aleatórios da variável), Primeira derivada, e segunda derivada. Essas perspectivas diferentes irão permitir um melhor entendimento das características de cada método. Cada gráfico será completado com 100 pontos de dados, cada um representando o erro medido encontrado em cada RoI (Contudo, nessa sessão, serão considerados 100 RoI ao invés de 20 como mostrado na Figura 18). Essa sessão será será dividida em três análises diferentes. Sessão 4.1 analisa a estimação de erro quando a interpolação pelo vizinho mais próximo é usada; Sessão 4.2 avalia para a interpolação linear; e a Sessão 4.1 insere problemas de *outliers*. Quando *outliers* são gerados, o desempenho de alguns métodos de discretização podem ser altamente degradados em comparação com outros, sendo uma questão importante a ser analisada.

4.1 ESTIMAÇÃO DE ERRO PELA INTERPOLAÇÃO DO VIZINHO MAIS PRÓXIMO

A interpolação pelo vizinho mais próximo basicamente atribui o valor vizinho mais próximo ao valor da probabilidade da variável aleatória que será estimada. Portanto, o erro de estimação será proporcional à sua distância da amostra mais próxima. Tal método de interpolação produz um erro diretamente proporcional à primeira derivada (GUREVICH et al., 1966). Analisando as Figura 19a e 19b pode-se inferir que:

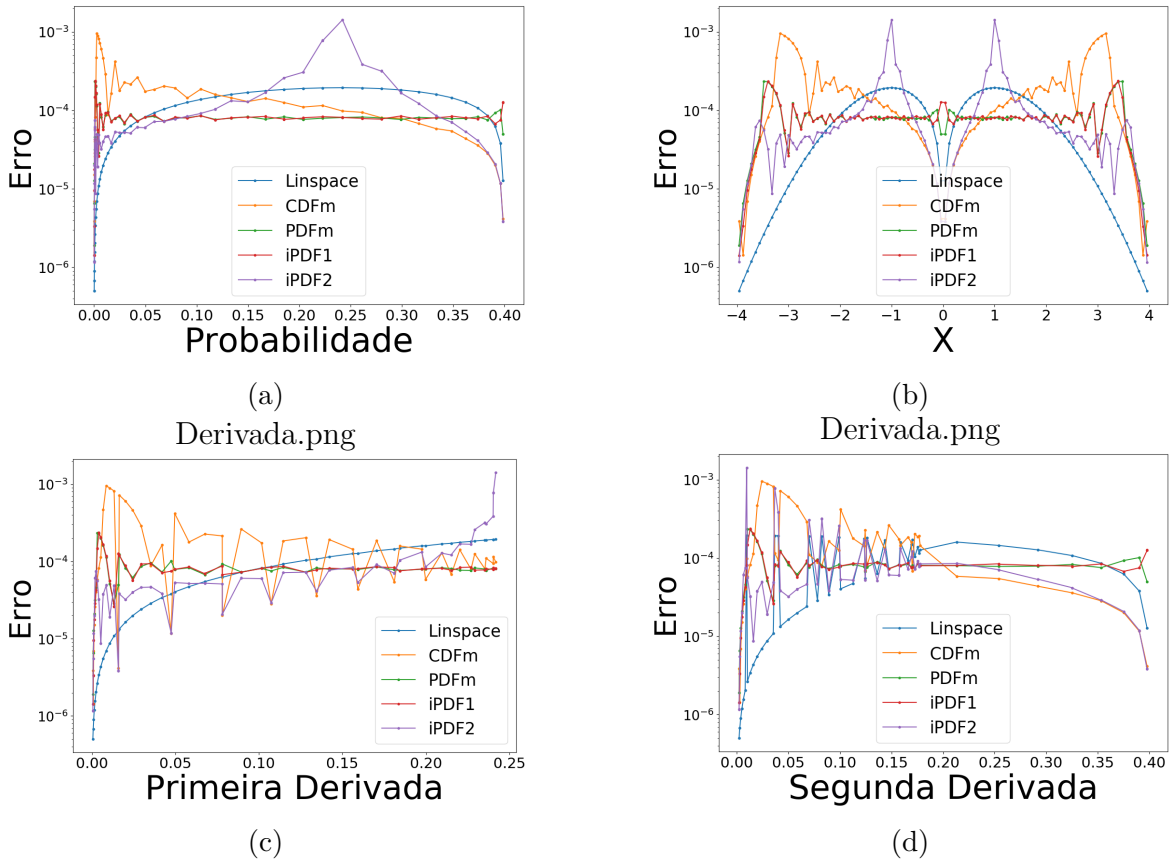
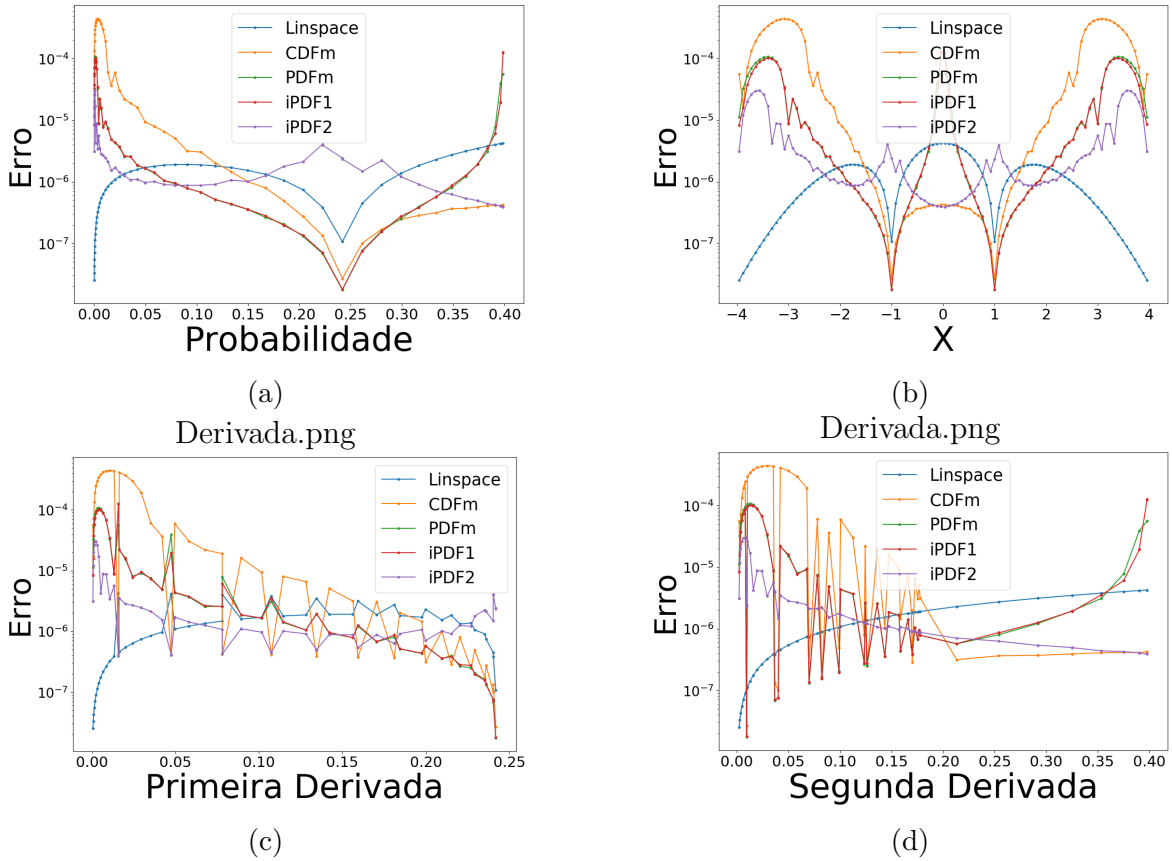


Figura 19: Caso representativo com 200 pontos, 100 RoI e usando a interpolação pelo vizinho mais próximo.

4.2 ESTIMAÇÃO DE ERRO PELA INTERPOLAÇÃO LINEAR



(a)

(b)

Derivada.png

Derivada.png

(c)

(d)

Figura 20: Caso representativo com 200 pontos, 100 RoI e usando a interpolação linear.

4.3 ESTIMAÇÃO DE ERRO CONSIDERANDO OUTLIERS

5 CONCLUSÃO

REFERÊNCIAS

ALISON, J. *The road to discovery: Detector alignment, electron identification, particle misidentification, ww physics, and the discovery of the Higgs Boson*. [S.l.]: Springer, 2014.

BIBA, M. et al. Unsupervised discretization using kernel density estimation. In: *IJCAI*. [S.l.: s.n.], 2007. p. 696–701.

CERN. *About CERN*. 2015. Disponível em: <http://home.web.cern.ch/about>.

COLLABORATION, C. et al. Performance of electron reconstruction and selection with the cms detector in proton-proton collisions at sqrt (s)= 8 tev. *arXiv preprint arXiv:1502.02701*, 2015.

FAYYAD, U.; IRANI, K. Multi-interval discretization of continuous-valued attributes for classification learning. 1993.

FRIEDMAN, N.; GOLDSZMIDT, M. et al. Discretizing continuous attributes while learning bayesian networks. In: *ICML*. [S.l.: s.n.], 1996. p. 157–165.

GRAMACKI, A. *Nonparametric Kernel Density Estimation and Its Computational Aspects*. [S.l.]: Springer, 2017.

GUREVICH, B. L. et al. *Integral, measure, and derivative: a unified approach*. [S.l.]: Courier Corporation, 1966.

HANAGAKI, K. et al. Electron identification in belle. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Elsevier, v. 485, n. 3, p. 490–503, 2002.

JONES, M. C. Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 84, n. 407, p. 733–741, 1989.

KEK, H. E. A. R. O. *Super KEKB Ring*. 2018. Disponível em: <https://www.kek.jp/en/Facility/ACCL/SuperKEKBRing/>.

RON, B. *The Art and Science of Digital Compositing*. [S.l.]: Morgan Kaufmann Verlag, 1999.

RONQUI, C. M. *Modelo Padrão*. 2015. Accessed: 2015-12-22.

SCHINDLER, A. Bandwidth selection in nonparametric kernel estimation. 2012.

SCHMIDBERGER, G.; FRANK, E. Unsupervised discretization using tree-based density estimation. In: SPRINGER. *European Conference on Principles of Data Mining and Knowledge Discovery*. [S.l.], 2005. p. 240–251.

VICINI, L.; SOUZA, A. M. Análise multivariada da teoria à prática. *Santa Maria: UFSM, CCNE*, 2005.

ZHANG, X.-H. et al. A discretization algorithm based on gini criterion. In: IEEE. *Machine Learning and Cybernetics, 2007 International Conference on*. [S.l.], 2007. v. 5, p. 2557–2561.

APÊNDICE A – LISTA DE PUBLICAÇÕES

A.1 PUBLICAÇÕES EM ANAIS DE CONGRESSO INTERNACIONAL

1.COSTA, R. M., SOUZA, D. M., COSTA, I. A., NÓBREGA, R. A. "Study of the Discretization Process applied to Continuous Random Variables in the Density Estimation Context."Instrumentation Systems, Circuits and Transducers (INSCIT), 2018 3rd International Symposium on IEEE, 2018.

Ultimamente, com o surgimento de grandes experimentos geradores de dados, há uma demanda crescente para otimizar os algoritmos responsáveis por interpretar esse volume de informações, de modo que ele use o mínimo de dados possível para realizar a operação desejada. Este trabalho permeia esse contexto, propondo alternativas em uma das escolhas mais elementares em algoritmos de estimação/classificação: a discretização de uma determinada variável. Este artigo propõe avaliar as características de diferentes métodos de discretização aplicados à estimação da função de densidade de probabilidade considerando o trade-off entre desempenho e simplicidade, bem como a suscetibilidade a *outliers*. Além disso, este trabalho analisa as vantagens e desvantagens de cada método e indica possíveis formas de ampliar o conhecimento sobre o assunto abordado.