



Universidade Federal de Juiz de Fora
Faculdade de 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 de Conclusão de Curso apresentado
ao programa de Bacharelado em Engenharia
Elétrica - Habilitação em Sistemas Eletrôni-
cos da Universidade Federal de Juiz de Fora,
como requisito parcial para obtenção do tí-
tulo de Bacharel em Engenharia Elétrica -
Habilitação em Sistemas Eletrônicos.

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

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. Estimação de Densidade. 2. Discretização. 3. Estimação Não Paramétrica I. Título.

CDU 621.3.0

Rafael Mascarenhas Costa

ESTUDO DE DIFERENTES MÉTODOS DE DISCRETIZAÇÃO APLICADOS A
ESTIMADORES DE DENSIDADE

Trabalho de Conclusão de Curso apresentado
ao programa de Bacharelado em Engenharia
Elétrica - Habilitação em Sistemas Eletrônicos
da Universidade Federal de Juiz de Fora,
como requisito parcial para obtenção do tí-
tulo de Bacharel em Engenharia Elétrica -
Habilitação em Sistemas Eletrônicos.

Aprovada em 13 de Dezembro de 2018.

BANCA EXAMINADORA:

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

Igor Abritta Costa, M.Sc.
Universidade Federal de Juiz de Fora, UFJF
Coorientador

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

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

Aos meus pais, familiares e amigos.

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, por me darem plenas condições de estudo, além de todo apoio moral nessa jornada, aos meus amigos e colegas pelas risadas e choros durante a vida acadêmica.

Aos meus nobres professores pelos ensinamentos que contribuiram com meu crescimento profissional e pessoal. Em especial ao meu orientador, Rafael e coorientador, Igor, por acreditarem em mim e me guiarem na vida acadêmica. Meus sinceros agradecimentos.

Aos companheiros do NIPS e amigos de Laboratório. Pelas conversas, ajudas e bom humor. É sempre bom ter alguém para compartilhar alegrias e desespero.

Aos meus amigos do ramo IEEE por toda motivação e crescimento e por me mostrar que estou no curso certo.

Finalmente, agradeço à Universidade Federal de Juiz de Fora e à Faculdade de Engenharia por todo o suporte e pelas ferramentas necessárias ao desenvolvimento deste trabalho.

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

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. Esta pesquisa propõe avaliar as características de diferentes métodos de discretização aplicados à estimação da função 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.

Palavras-chave: Estimação de Densidade, Discretização, Estimação Não Paramétrica.

ABSTRACT

Lately, with the emergence of large data-generating experiments, there is a growing demand to optimize the algorithms responsible for interpreting this volume of information so that it uses as little data as possible to perform the desired operation. This work permeates this context, proposing alternatives in one of the most elementary choices in estimation/classification algorithms: discretization of a given variable. This research proposes to evaluate the characteristics of different discretization methods applied to probability density function estimation considering the trade-off between performance and simplicity, as well as susceptibility to outliers. In addition, this work analyzes the advantages and disadvantages of each method and indicates possible ways to extend the knowledge about the addressed subject.

Keywords: Density Estimation, Discretization, Non-parametric Estimation.

LISTA DE ILUSTRAÇÕES

Figura 1	Caso representativo de estimativa de PDF utilizando 25 pontos para dois intervalos diferentes (-4,4) e (-10,10)	23
Figura 2	Ilustração da curva Gaussiana $N(0,1)$	24
Figura 3	Ilustração do método <i>Linspace</i> 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.	26
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). ...	30
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). 31

Figura 10 Ilustração das curvas Lognormais construídas com diferentes parâmetros: (a) $L(0,0.01)$; (b) $L(0,0.25)$; (c) $L(0,0.25)$; (d) $L(0,1)$; (e) $L(0,1.25)$; e (f) $L(0,1.5)$ 32

Figura 11 Histograma dos dados gerados sendo eles: (a) Gaussiana com $\mu = 0$ e $\sigma = 1$; (b) Gaussiana com $\mu = 0$, $\sigma = 1$ e *outlier* em ± 25 ; (c) Lognormal com $\mu = 0$ e $\sigma = 1$ 33

Figura 12 Diagrama de blocos do algoritmo para validação dos métodos de discretização. 33

Figura 13 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 (d) $L(0,1)$ com $N = 25$ 35

Figura 14 Discretização com os dados gerados utilizando o método de *Linspace*: (a) $\text{randN}(0,1)$ com $N = 15$, (b) $\text{randN}(0,1)$ com $N = 25$, (c) $\text{randN}(0,1)$ com $N = 15$ e *outlier* em ± 25 , (d) $\text{randN}(0,1)$ com $N = 25$ e *outlier* em ± 25 , (e) $\text{randL}(0,1)$ com $N = 15$ e (f) $\text{randL}(0,1)$ com $N = 25$ 35

Figura 15 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 (d) $L(0,1)$ com $N = 25$ 37

Figura 16 Discretização com os dados gerados utilizando o método CDF (do inglês, *CDF method*) (*CDFm*): (a) $\text{randN}(0,1)$ com $N = 15$, (b) $\text{randN}(0,1)$ com $N = 25$, (c) $\text{randN}(0,1)$ com $N = 15$ e *outlier* em ± 25 , (d) $\text{randN}(0,1)$ com $N = 25$ e *outlier* em ± 25 , (e) $\text{randL}(0,1)$ com $N = 15$ e (f) $\text{randL}(0,1)$ com $N = 25$ 38

Figura 17 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 (d) $L(0,1)$ com $N = 25$ 39

Figura 18 Discretização com os dados gerados utilizando o método PDF (do inglês, *PDF method*) (*PDFm*): (a) $\text{randN}(0,1)$ com $N = 15$, (b) $\text{randN}(0,1)$ com $N = 25$, (c) $\text{randN}(0,1)$ com $N = 15$ e *outlier* em ± 25 , (d) $\text{randN}(0,1)$ com $N = 25$ e *outlier* em ± 25 , (e) $\text{randL}(0,1)$ com $N = 15$ e (f) $\text{randL}(0,1)$ com $N = 25$ 40

Figura 19 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 (d) $L(0,1)$ com $N = 25$ 41

Figura 20 Discretização com os dados gerados utilizando o método Integral da distribuição da primeira derivada da PDF (*iPDF1*): (a) $\text{randN}(0,1)$ com $N = 15$, (b) $\text{randN}(0,1)$ com $N = 25$, (c) $\text{randN}(0,1)$ com $N = 15$ e *outlier* em ± 25 , (d) $\text{randN}(0,1)$ com $N = 25$ e *outlier* em ± 25 , (e) $\text{randL}(0,1)$ com $N = 15$ e (f) $\text{randL}(0,1)$ com $N = 25$ 42

Figura 21 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 (d) $L(0,1)$ com $N = 25$ 43

Figura 22 Discretização com os dados gerados utilizando o método Integral da distribuição da segunda derivada da PDF (*iPDF2*): (a) $\text{randN}(0,1)$ com $N = 15$, (b) $\text{randN}(0,1)$ com $N = 25$, (c) $\text{randN}(0,1)$ com $N = 15$ e *outlier* em ± 25 , (d) $\text{randN}(0,1)$ com $N = 25$ e *outlier* em ± 25 , (e) $\text{randL}(0,1)$ com $N = 15$ e (f) $\text{randL}(0,1)$ com $N = 25$ 44

Figura 23 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. 45

Figura 24 Ilustração da medida de erro entre a PDF Real e a Estimada com 20

regiões de interesse.	46
Figura 25 Caso representativo da distribuição Gaussiana com 200 pontos, 100 Regiões de Interesse (do inglês, <i>Regions of Interest</i>) (RoI) e usando a interpolação pelo vizinho mais próximo.	48
Figura 26 Caso representativo da distribuição Lognormal com desvio padrão $\sigma = 1$ com 200 pontos, 100 RoI e usando a interpolação pelo vizinho mais próximo.	49
Figura 27 Erro total de estimativa para a interpolação pelo vizinho mais próximo variando-se o números de pontos a se estimar: (a) $N(0,1)$ e (b) $L(0,1)$.	50
Figura 28 Erro total de estimativa para a interpolação pelo vizinho mais próximo variando-se o números de pontos a se estimar: (a) $\text{randN}(0,1)$ e (b) $\text{randL}(0,1)$	51
Figura 29 Caso representativo da distribuição Gaussiana com 200 pontos, 100 RoI e usando a interpolação linear.	52
Figura 30 Caso representativo da distribuição Lognormal com desvio padrão $\sigma = 1$ com 200 pontos, 100 RoI e usando a interpolação linear.	53
Figura 31 Erro total de estimativa para a interpolação linear variando-se o números de pontos a se estimar: (a) $N(0,1)$ e (b) $L(0,1)$	54
Figura 32 Erro total de estimativa para a interpolação linear variando-se o números de pontos a se estimar: (a) $\text{randN}(0,1)$ e (b) $\text{randL}(0,1)$	54
Figura 33 Análise de <i>outlier</i> usando 100 Rois e interpolação linear.	55
Figura 34 Erro total de estimativa para a distribuição Normal com outlier em 25: (a) para a interpolação pelo vizinho mais próximo, (b) para a interpolação	

linear. 56

Figura 35 Erro total de estimacao para a distribuiao $\text{randN}(0,1)$ com outlier em 25:
(a) para a interpolaao pelo vizinho mais proximo, (b) para a interpolaao
linear. 57

Figura 36 Histograma representativo de uma realidade com 100 mil eventos 57

Figura 37 Erro total de estimacao para a distribuiao da Figura 36: (a) para a
interpolaao pelo vizinho mais proximo, (b) para a interpolaao linear. 58

LISTA DE TABELAS

Tabela 1	Erro de estimativa média usando a distribuição normal e interpolação do vizinho mais próximo com 100 pontos de estimação.	55
Tabela 2	Erro de estimação média usando a distribuição normal e interpolação linear com 100 pontos de estimação.	55
Tabela 3	Equivalência de pontos para o mesmo erro em diferentes métodos utilizando-se a interpolação linear.	58
Tabela 4	Equivalência de pontos para o mesmo erro em diferentes métodos utilizando-se a interpolação pelo vizinho mais próximo.	58

LISTA DE ABREVIATURAS E SIGLAS

- ATLAS** Dispositivo Instrumental Toroidal para o LHC (do inglês, *A Toroidal LHC Apparatus*)
- 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*)
- CMS** Solenoide de Múon Compacto (do inglês, *Compact Muon Solenoid*)
- FD** Estimador Freedman–Diaconi
- IAE** Erro Absoluto Integrado (do inglês, *Integrated Absolute Error*)
- iPDF1** Integral da distribuição da primeira derivada da PDF
- iPDF2** Integral da distribuição da segunda 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	19
1.1 Motivação	20
1.2 O que foi feito	21
1.3 Estrutura do Trabalho	21
2 DISCRETIZAÇÃO	22
2.1 Métodos de Discretização	23
2.1.1 <i>Linspace</i>	24
2.1.2 <i>CDFm</i>	25
2.1.3 <i>PDFm</i>	25
2.1.4 <i>iPDF1</i>	26
2.1.5 <i>iPDF2</i>	27
3 DESENVOLVIMENTO	29
3.1 Conjunto de dados	29
3.2 Algoritmo	33
3.3 Demostraçāo dos métodos de discretizaçāo	34
3.3.1 Método <i>Linspace</i>	34
3.3.2 Método <i>CDFm</i>	36
3.3.3 Método <i>PDFm</i>	39
3.3.4 Método <i>iPDF1</i>	40
3.3.5 Método <i>iPDF2</i>	42
3.4 Custo Computacional	45

3.5	Ambiente de Análise	46
4	RESULTADOS	47
4.1	Estimação de erro pela interpolação do vizinho mais próximo	47
4.2	Estimação de erro pela interpolação linear	51
4.3	Estimação de erro considerando <i>outliers</i>	54
4.4	Exemplo Prático	57
5	Conclusão	60
5.1	Próximos Passos	60
Referências		61
Apêndice A – Lista de Publicações		63
A.1	Publicações em Anais de Congresso Internacional	63

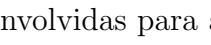
1 INTRODUÇÃO

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 estimativa de densidades e seleção de eventos, tanto no ambiente industrial quanto em laboratórios de pesquisa. Entretanto, tais problemas podem ocorrer 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. Este 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 à 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 a  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 estimativa, 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  a estimativa 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 Solenoide de Múon Compacto (do inglês, *Compact Muon Solenoid*) (CMS) e do Dispositivo Instrumental Toroidal para o LHC (do inglês, *A Toroidal LHC Apparatus*) (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), ~~medidas 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 estimativa 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 estimativa 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 estimativa 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 estimativa, 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 2 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 análise 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 estimativa 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, estimativas 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 estimativa os quais podem ser minimizados incrementando o número de pontos a serem estimados, buscando um equilíbrio entre otimização computacional e performance de estimativa.

Vários autores seguem a mesma abordagem, como em (JONES, 1989), explorando os diferentes aspectos do processo de discretização e propõendo 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 estimativa de densidade baseado em árvore; e em (ZHANG et al., 2007) um algoritmo de aprendizagem de máquina seado-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. 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 estimativa 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 estimativa, o número de pontos a ser estimado seja maior a fim de representar bem a região de alta probabilidade.

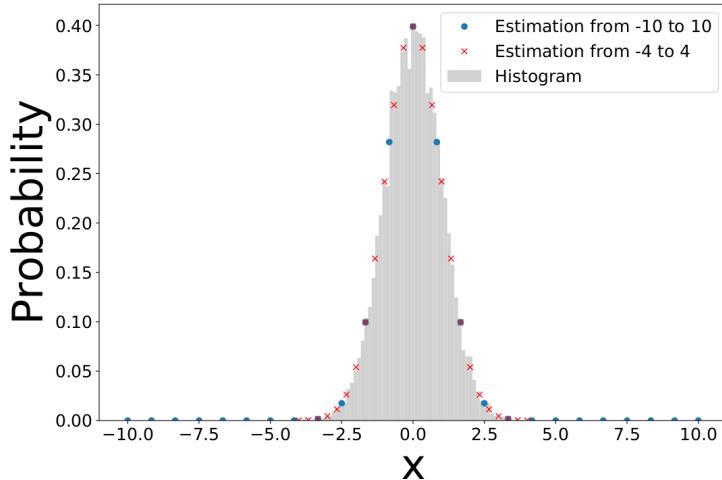


Figura 1: Caso representativo de estimativa de PDF utilizando 25 pontos para dois intervalos diferentes (-4,4) e (-10,10)

2.1 MÉTODOS DE DISCRETIZAÇÃO

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

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

Estes cinco métodos serão demonstrados a priori utilizando-se uma distribuição Gaussiana com média $\mu = 0$ e desvio padrão $\sigma = 1$, representado por $N(0,1)$, cuja PDF pode ser descrita pela Equação (3.1) e ilustrada na Figura 2 com o número de pontos $N = 25$

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

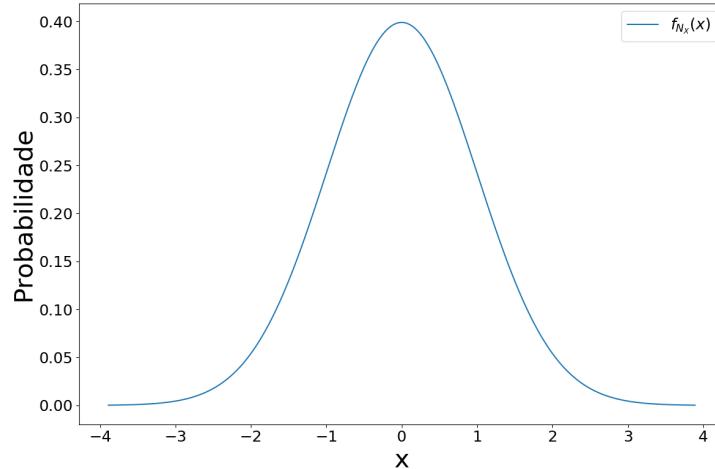


Figura 2: Ilustração da curva Gaussiana $N(0,1)$.

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%.

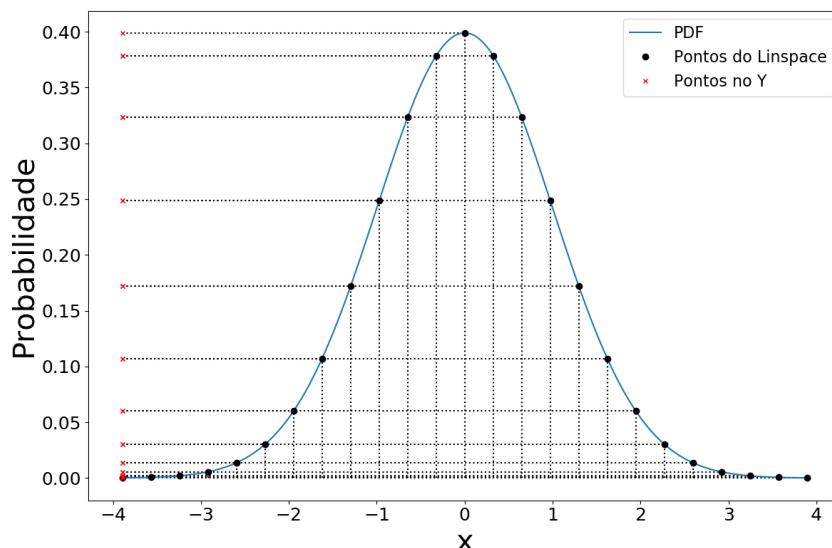


Figura 3: Ilustração do método *Linspace* aplicado à uma distribuição normal.

2.1.2 CDFM

O método denominado nesse trabalho de 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 ilustrado pela Figura 4.

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

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

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

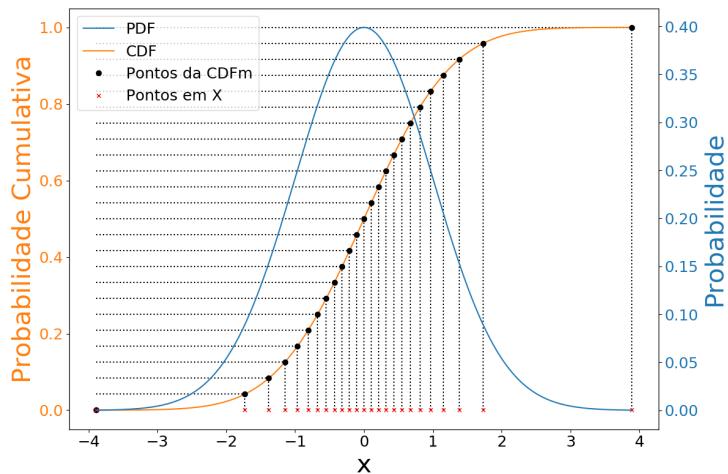


Figura 4: Ilustração da discretização da distribuição Gaussiana baseada em sua CDF.

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 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.

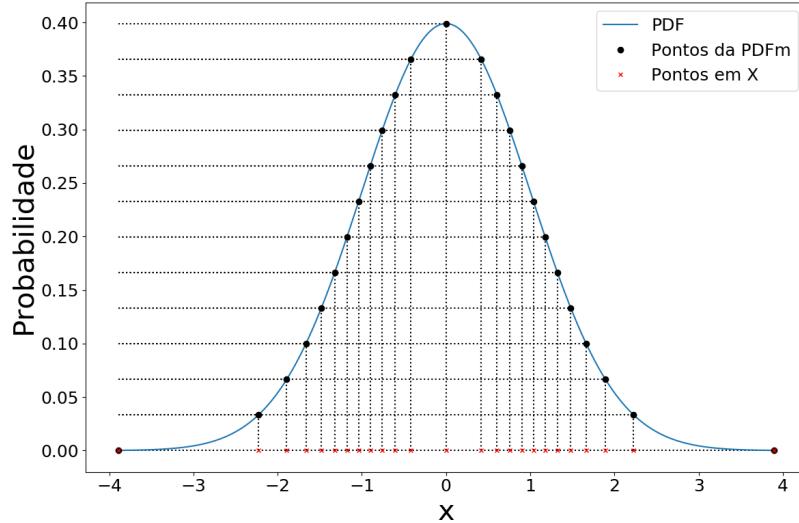


Figura 5: Ilustração da discretização da distribuição Gaussiana baseada em sua PDF.

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 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

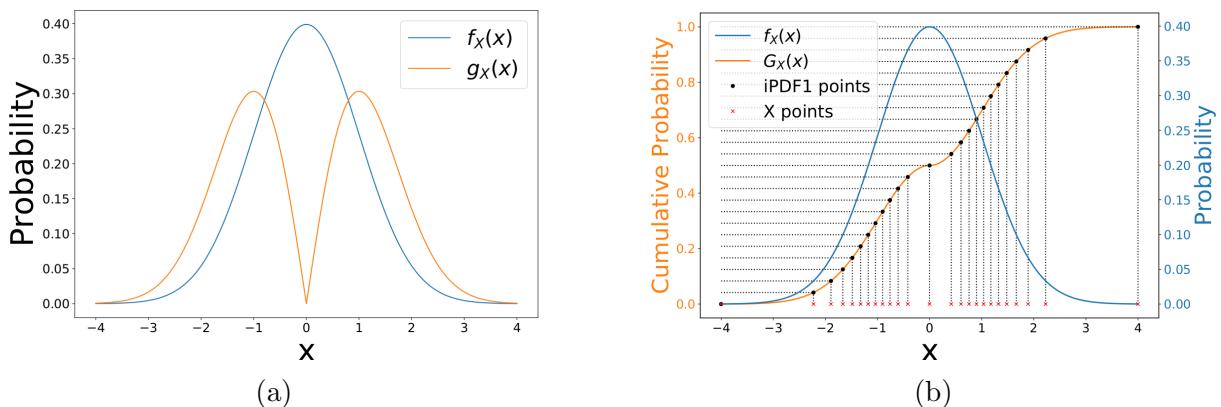


Figura 6: PDF Gaussiana e sua primeira derivada à esquerda. Ilustração da discretização da distribuição Gaussiana baseada na CDF da sua primeira derivada à direita.

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

$$\zeta(x) = \frac{|\mu - x|}{\sigma^3 \sqrt{2\pi}} \cdot e^{\left(\frac{-(\mu-x)^2}{2\sigma^2}\right)}$$

$$\int_{-\infty}^{\infty} \zeta(x) \cdot dx = c_1 \quad (2.5)$$

$$g_X(x) = \frac{\zeta(x)}{c_1}$$

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.

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

É possível notar que este método consegue fazer uma melhor estimativa 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).

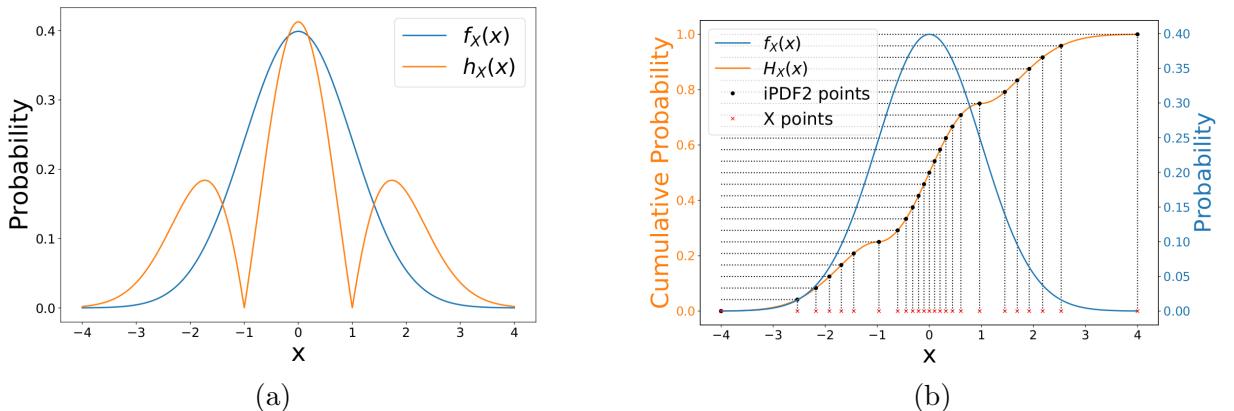


Figura 7: PDF Gaussiana e sua segunda derivada à esquerda. Ilustração da discretização da distribuição Gaussiana baseada na CDF da sua segunda derivada à direita.

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

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 estimativas serão usadas para a identificação e classificação de eventos. Portanto o conjunto de dados aqui escolhido tem por base as estimativas 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  mostrado também as  varias formas de ruído de fundo para elétrons no ATLAS.

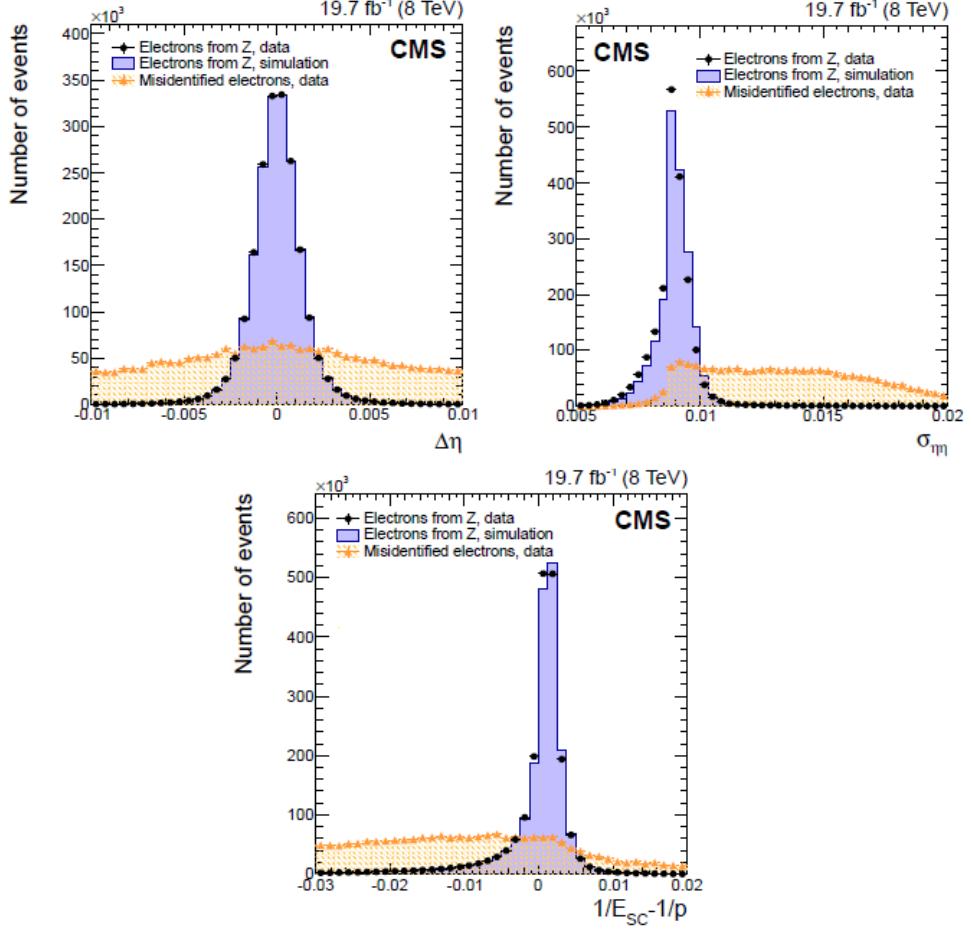


Figura 8: Caso representativo, variáveis de identificação de elétrons, do experimento CMS: variável $\Delta\eta$, forma do chuveiro $\sigma_{\eta\eta}$ e distribuição de energia-momento $1/E_{SC} - 1/p$. Extraído de (COLLABORATION et al., 2015).

Pode-se 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.

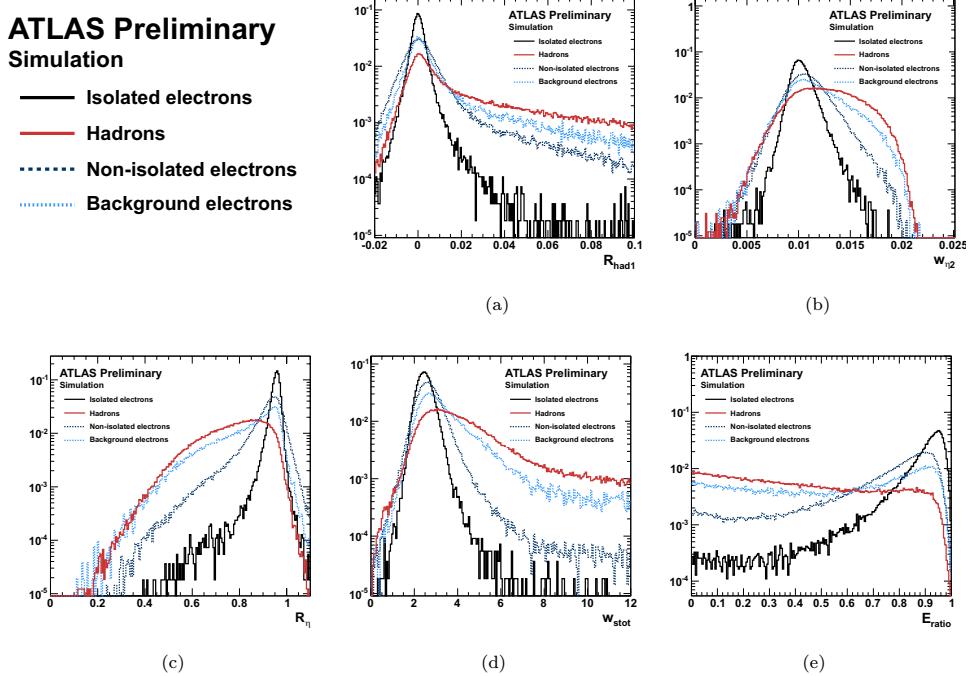


Figura 9: Caso representativo, variáveis de identificação de elétrons, do experimento ATLAS, no calorímetro, formato do chuveiro, apresentados separadamente para sinal e os vários tipos de ruídos de fundo. As variáveis apresentadas são: (a) vazamento hadrônico R_{had} , (b) de largura em eta no segundo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

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 *Gaussian* e a *Lognormal*. Sendo assim, com o intuito de realizar as análises em um ambiente controlado foram escolhidas as distribuições analíticas *Gaussian* e *Lognormal*, sendo que a última parametrizada de 6 maneiras diferentes e três *datasets* discretos diferentes cuja *binagem* foi feita usando o método Estimador Freedman–Diaconi (FD).

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.

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

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, representada

por $L(\mu, \sigma)$.

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

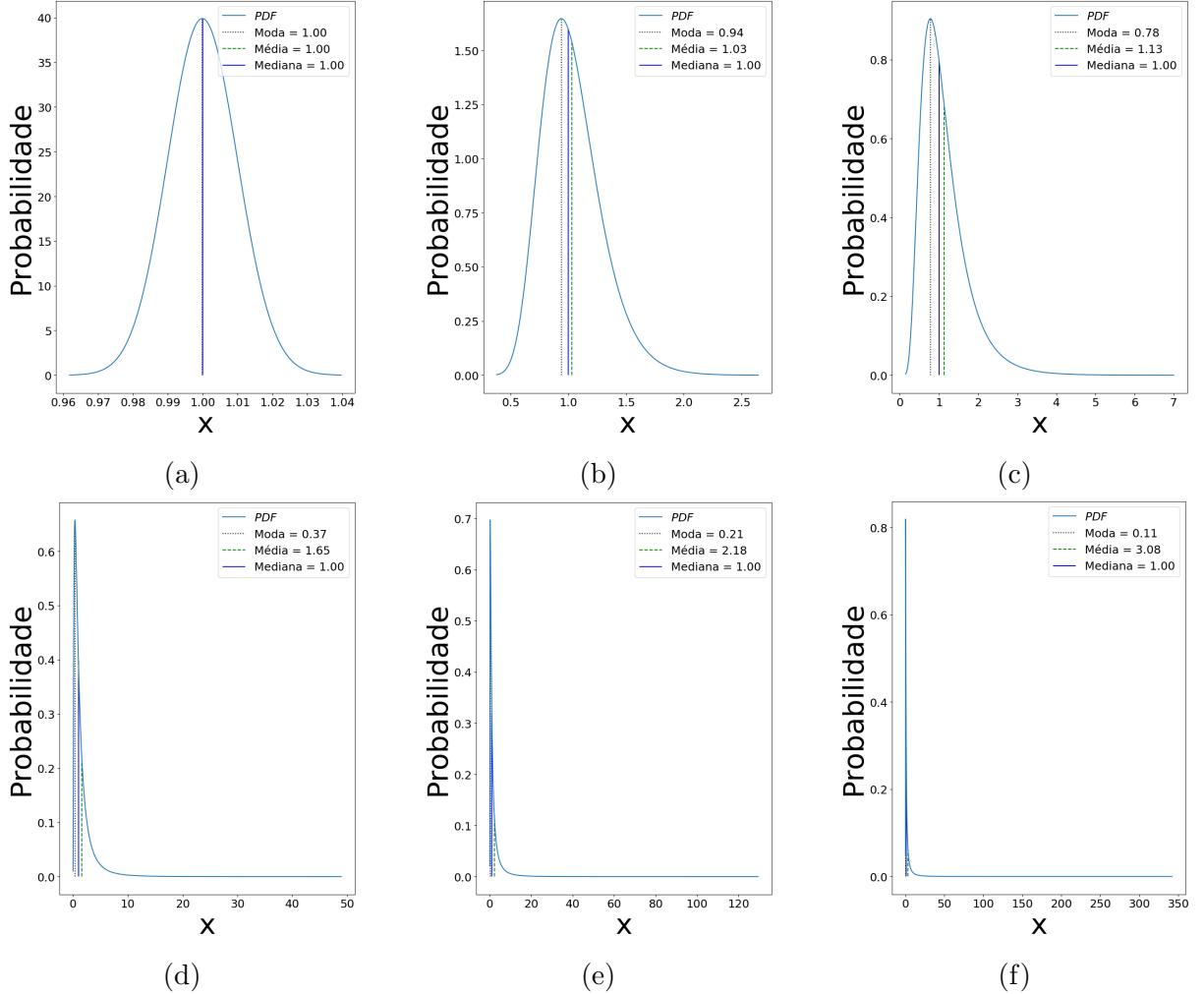


Figura 10: Ilustração das curvas Lognormais construídas com diferentes parâmetros: (a) $L(0,0.01)$; (b) $L(0,0.25)$; (c) $L(0,0.25)$; (d) $L(0,1)$; (e) $L(0,1.25)$; e (f) $L(0,1.5)$

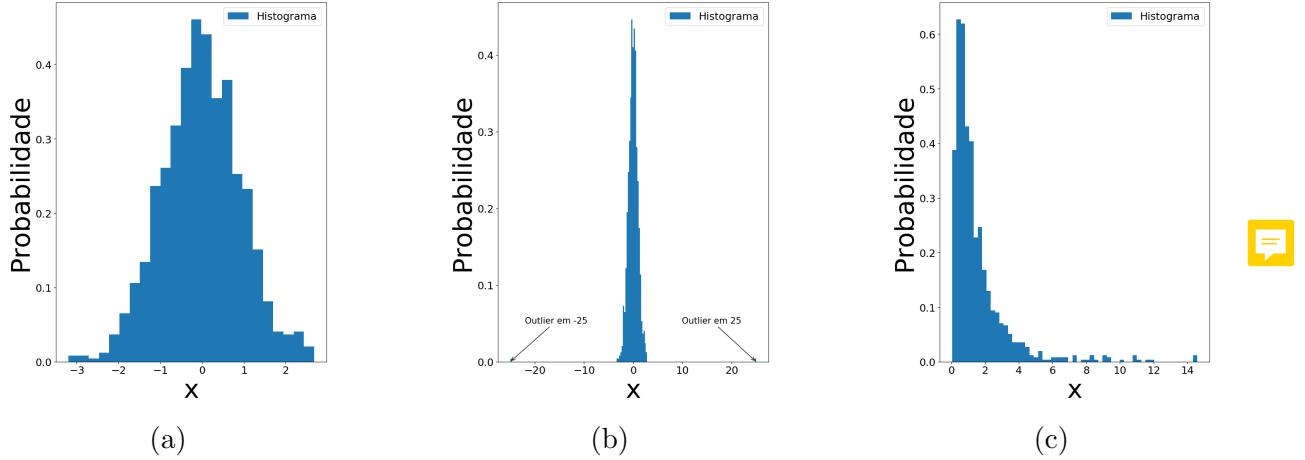


Figura 11: Histograma dos dados gerados sendo eles: (a) Gaussiana com $\mu = 0$ e $\sigma = 1$; (b) Gaussiana com $\mu = 0$, $\sigma = 1$ e *outlier* em ± 25 ; (c) Lognormal com $\mu = 0$ e $\sigma = 1$.

Os *datasets* escolhidos possuem mil eventos, média $\mu = 0$ e são baseados na distribuição Gaussiana com desvio padrão $\sigma = 1$ representado por $\text{randN}(\mu, \sigma)$ sem *outliers* ilustrado na Figura 11a e com *outliers* em ± 25 ilustrado na Figura 11b e, por fim, baseado na distribuição Lognormal com desvio padrão $\sigma = 1$ representado por $\text{randL}(\mu, \sigma)$ ilustrado na Figura 11c.

3.2 ALGORITMO

O algoritmo para comparação e validação dos métodos de discretização de estimativa construído nesse trabalho pode ser resumido pelo diagrama de blocos da Figura 12, sendo testado de duas maneiras diversas: utilizando somente a função analítica e usando os dados gerados a partir das funções geradoras.



Figura 12: 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

distância chamada L1, que é um caso particular da Equação (3.3).

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

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

$$L_1 = \int |f(x) - g(x)| \cdot dx \quad (3.4)$$

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

Dados gerados Nesse caso,  a diferença é que o  critério 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, função *lognormal* e com dados gerados,  ~~presentados pelas Figuras 2, 10 e 11 respectivamente~~. E, com o objetivo de demonstrar 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 13a, 13b, 14a e 14b. Já para conjunto de dados que apresenta variações mais rápidas ou *outliers*, como mostrado nas Figuras 13c, 13d, 14c e 14d, este método não alcança uma boa representação da PDF.

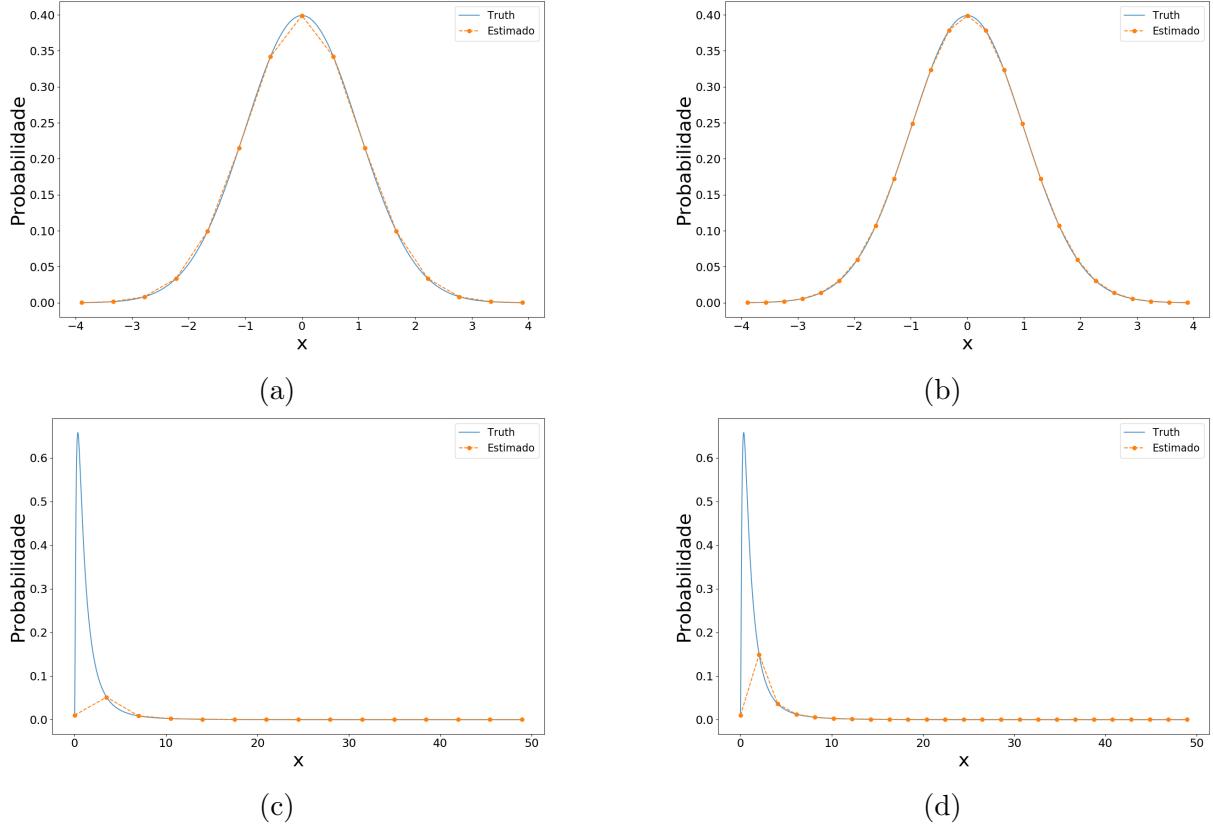
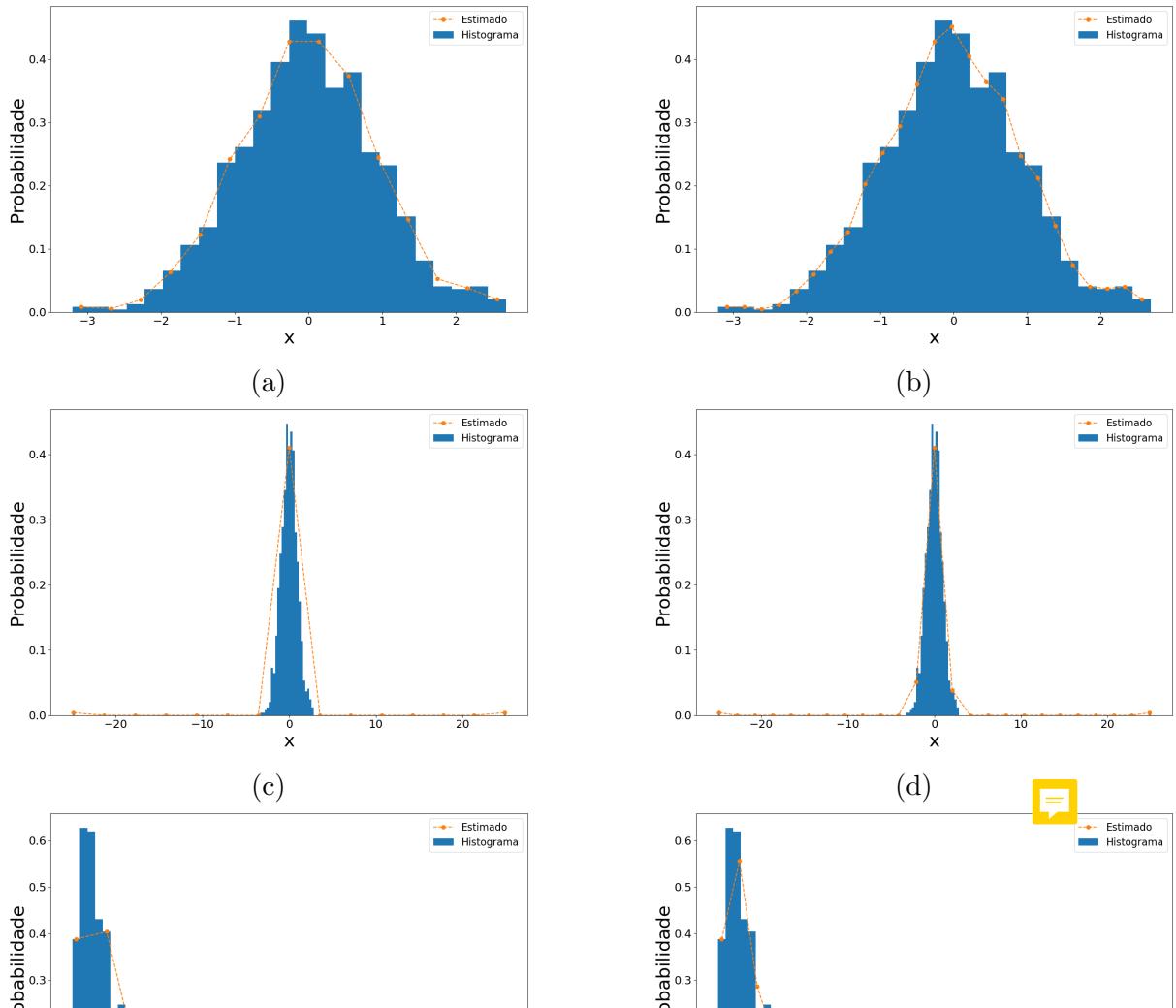


Figura 13: Discretização utilizando o método de *Linspace*: (a) $N(0,1)$ com $N = 15$, (b) $N(0,1)$ com $N = 25$, (c) $L(0,1)$ com $N = 15$ e (d) $L(0,1)$ com $N = 25$.



Entretanto, pode-se perceber que com o aumento do número de pontos de estimação ($M = 15$ 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  Ém disso, aumenta a quantidade de informação a ser armazenada ou transmitida. Apesar da distribuição `randL(0,1)` ser baseada na $L(0,1)$, está apresenta uma melhor discretização devido ao fato da sua extensão no eixo x ser menor. As distribuições $L(0,0.01)$, $L(0,0.25)$ e $L(0,0.5)$ possuem um comportamento semelhante à $N(0,1)$, enquanto as distribuições $L(0,1.25)$ e $L(0,1.5)$ semelhantes à $L(0,1)$.

3.3.2 MÉTODO CDFM

Pelo método CDFm é esperado um acúmulo maior de pontos nas regiões em que a probabilidade é maior, como é ilustrado nas Figuras 15 e 16, porem, é possível perceber que quanto mais rápida é a variação, melhor a sua discretização como é notado comparando as Figuras 15a e 15c e que a adição de *outliers* não impactam tanto na discretização, mostrado nas Figuras 16c e 16f. Esse mesmo padrão se repete tanto com os dados gerados mostrado na Figura 16 quanto nas outras distribuições baseadas na função *Lognormal*.

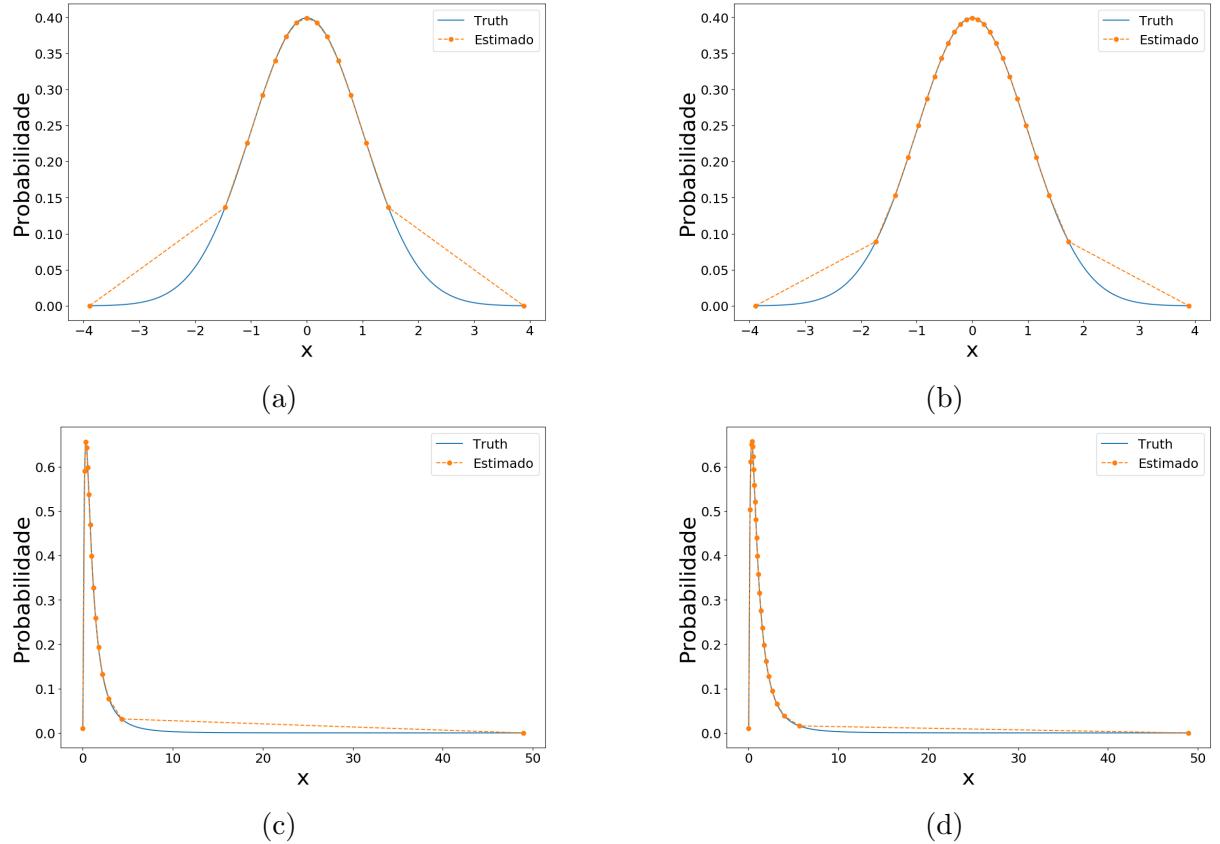


Figura 15: Discretização utilizando o método de $CDFm$: (a) $N(0,1)$ com $N = 15$, (b) $N(0,1)$ com $N = 25$, (c) $L(0,1)$ com $N = 15$ e (d) $L(0,1)$ com $N = 25$.

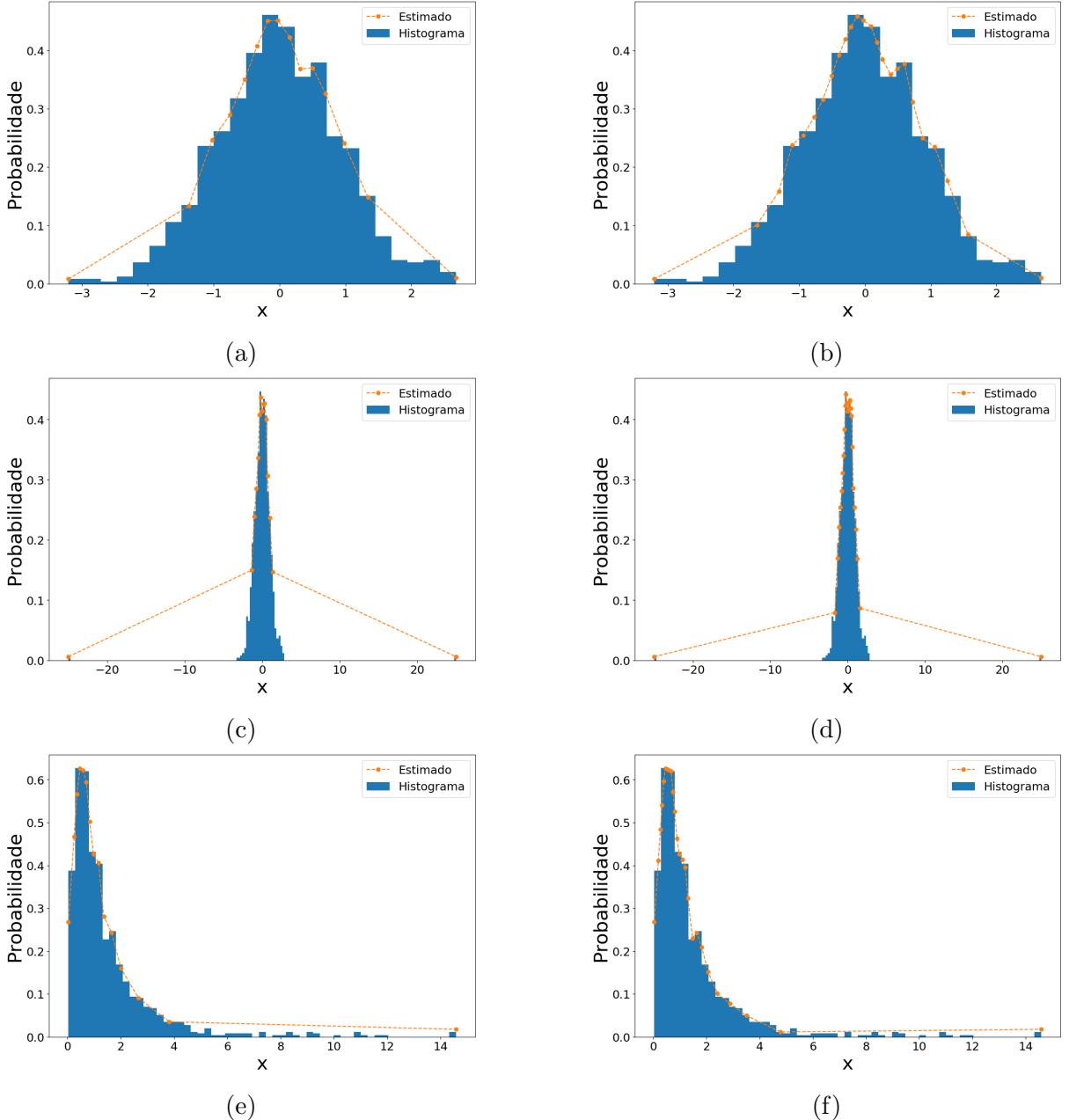


Figura 16: Discretização com os dados gerados utilizando o método CDFm: (a) $\text{randN}(0,1)$ com $N = 15$, (b) $\text{randN}(0,1)$ com $N = 25$, (c) $\text{randN}(0,1)$ com $N = 15$ e *outlier* em ± 25 , (d) $\text{randN}(0,1)$ com $N = 25$ e *outlier* em ± 25 , (e) $\text{randL}(0,1)$ com $N = 15$ e (f) $\text{randL}(0,1)$ com $N = 25$.

Este método, se comparado ao *Linspace* , apresenta um erro muito menor nas regiões de alta probabilidade em contrapartida, ele não deve ser indicado quando o objeto de estudo é a análise de eventos em regiões de baixa probabilidade, pois o mesmo teria que ter um número de pontos superior para o estudo dessas regiões.

3.3.3 MÉTODO PDFM

Como este método divide o eixo y de maneira uniforme, ele tende a penalizar menos as regiões de baixa probabilidade e que sejam simétricas se comparado com o método CDFm como pode ser visto nas Figuras 17a, 17b, 18a, 18b, 18c e 18d, já para o caso destas distribuições não serem simétricas, como mostra as Figuras 17c, 17c, 18e e 18f a região de baixa probabilidade é mais afetada que o método anterior. Este padrão se repete para desvios padrões diferentes da distribuição Lognormal.

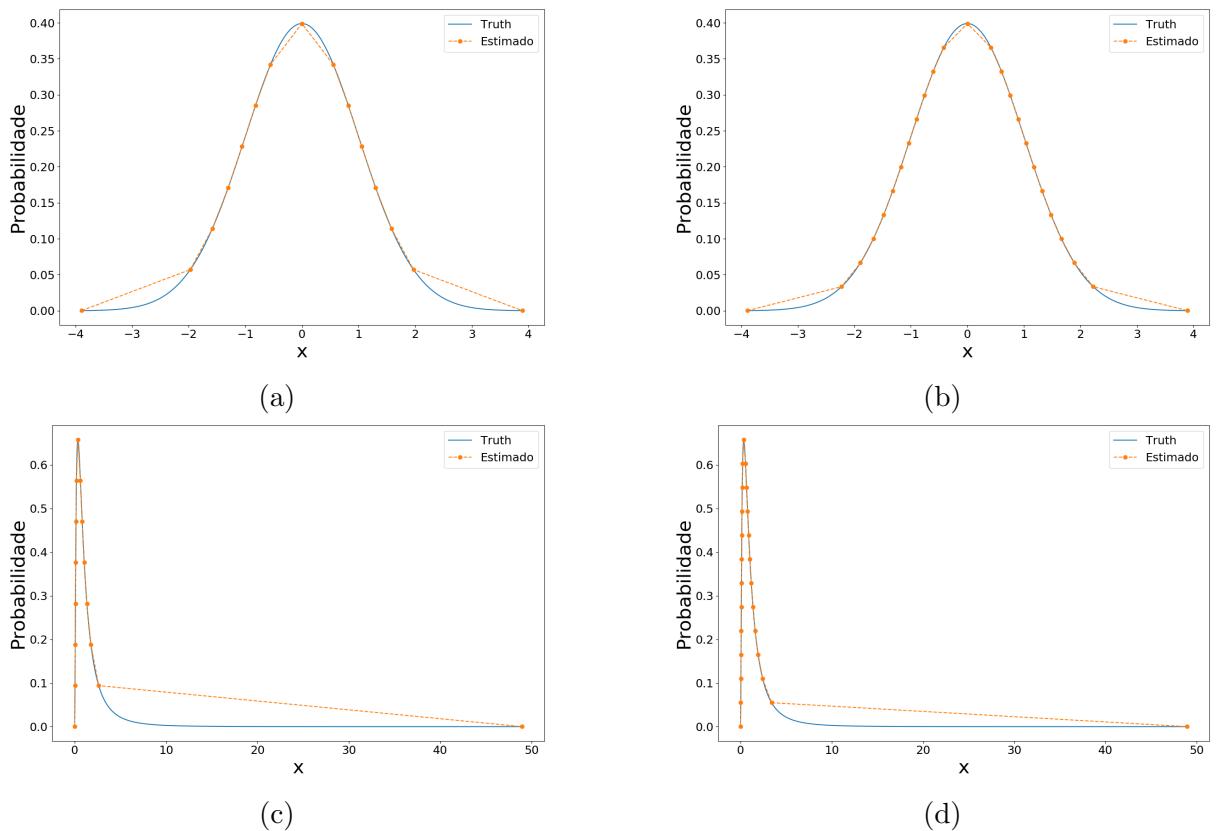


Figura 17: Discretização utilizando o método de *PDFM*: (a) $N(0,1)$ com $N = 15$, (b) $N(0,1)$ com $N = 25$, (c) $L(0,1)$ com $N = 15$ e (d) $L(0,1)$ com $N = 25$.

Deste modo, este método relaciona os pontos positivos da CDFm para regiões de alta probabilidade sendo que seu erro para as regiões de baixa probabilidade é um pouco menor.

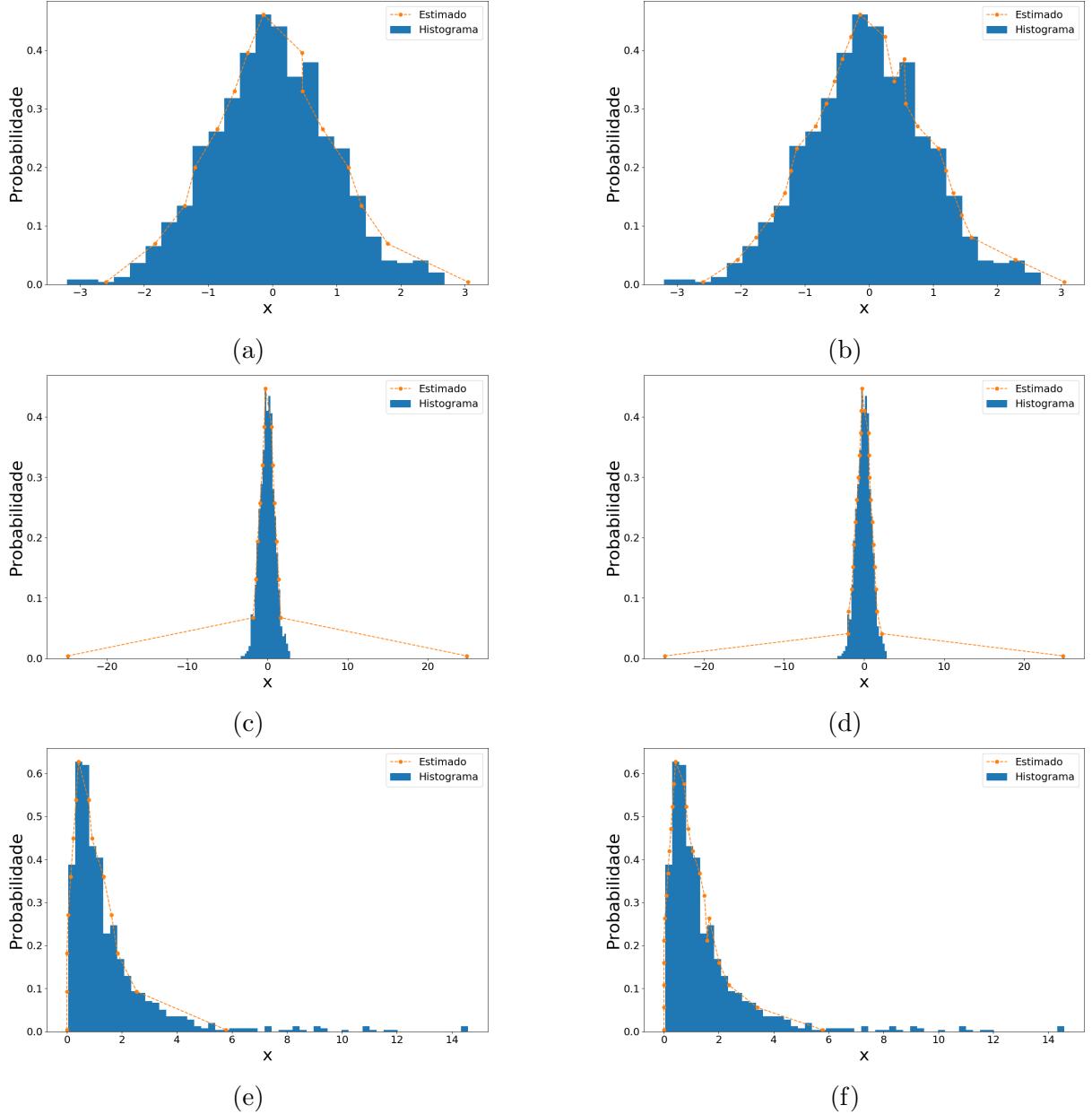


Figura 18: Discretização com os dados gerados utilizando o método PDFm: (a) $\text{randN}(0,1)$ com $N = 15$, (b) $\text{randN}(0,1)$ com $N = 25$, (c) $\text{randN}(0,1)$ com $N = 15$ e *outlier* em ± 25 , (d) $\text{randN}(0,1)$ com $N = 25$ e *outlier* em ± 25 , (e) $\text{randL}(0,1)$ com $N = 15$ e (f) $\text{randL}(0,1)$ com $N = 25$.

3.3.4 MÉTODO IPDF1

Este método ~~tem como base fazer~~ a discretização baseada na primeira derivada, com isso é esperado uma melhor resolução nos pontos em que a variação da distribuição é maior, como pode ser verificado na Figura 19.

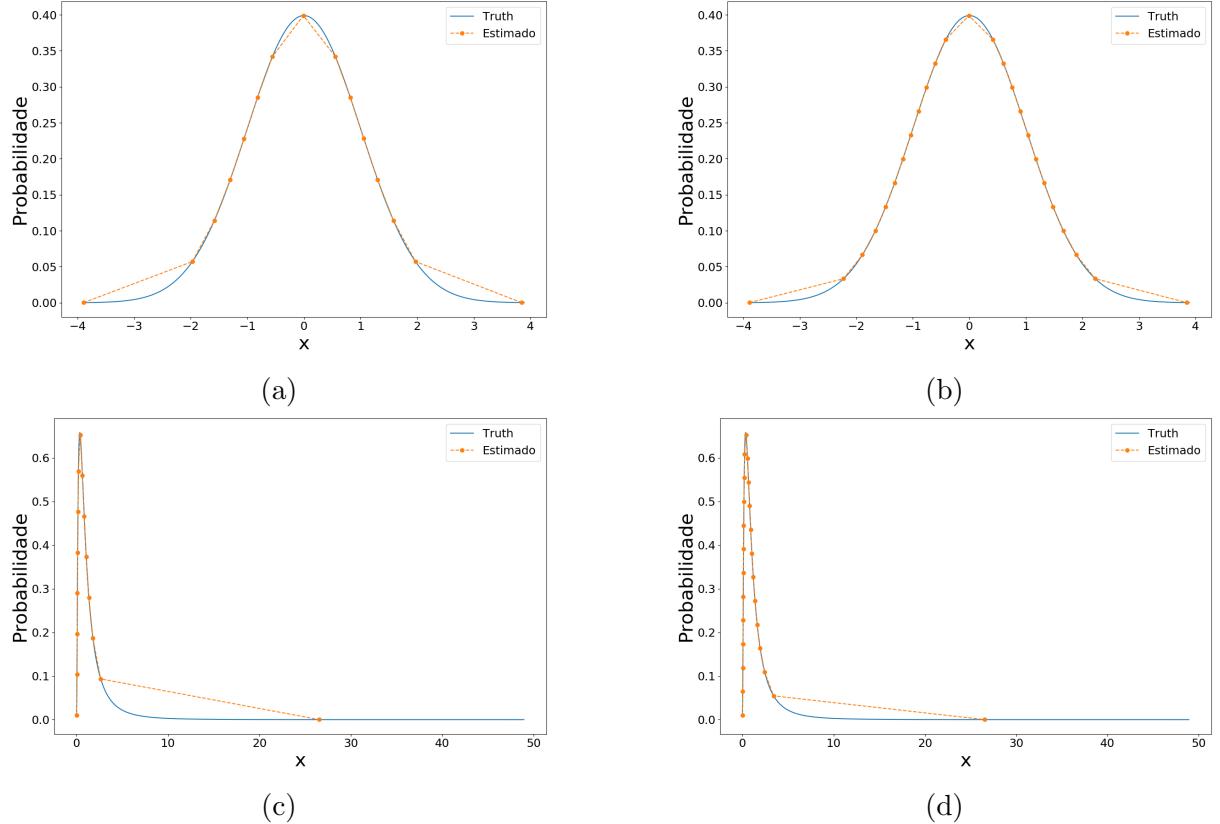


Figura 19: Discretização utilizando o método de *iPDF1*: (a) $N(0,1)$ com $N = 15$, (b) $N(0,1)$ com $N = 25$, (c) $L(0,1)$ com $N = 15$ e (d) $L(0,1)$ com $N = 25$.

Já para os dados gerados, devido ao fato de sua derivada discreta ~~inserir ruído~~, sua CDF acaba sendo suavizada, fazendo com que este coloque mais pontos na região de baixa probabilidade, como pode ser visto na Figura 20 e, em especial nas Figuras 20c e 20d em que o erro na calda é baixo.

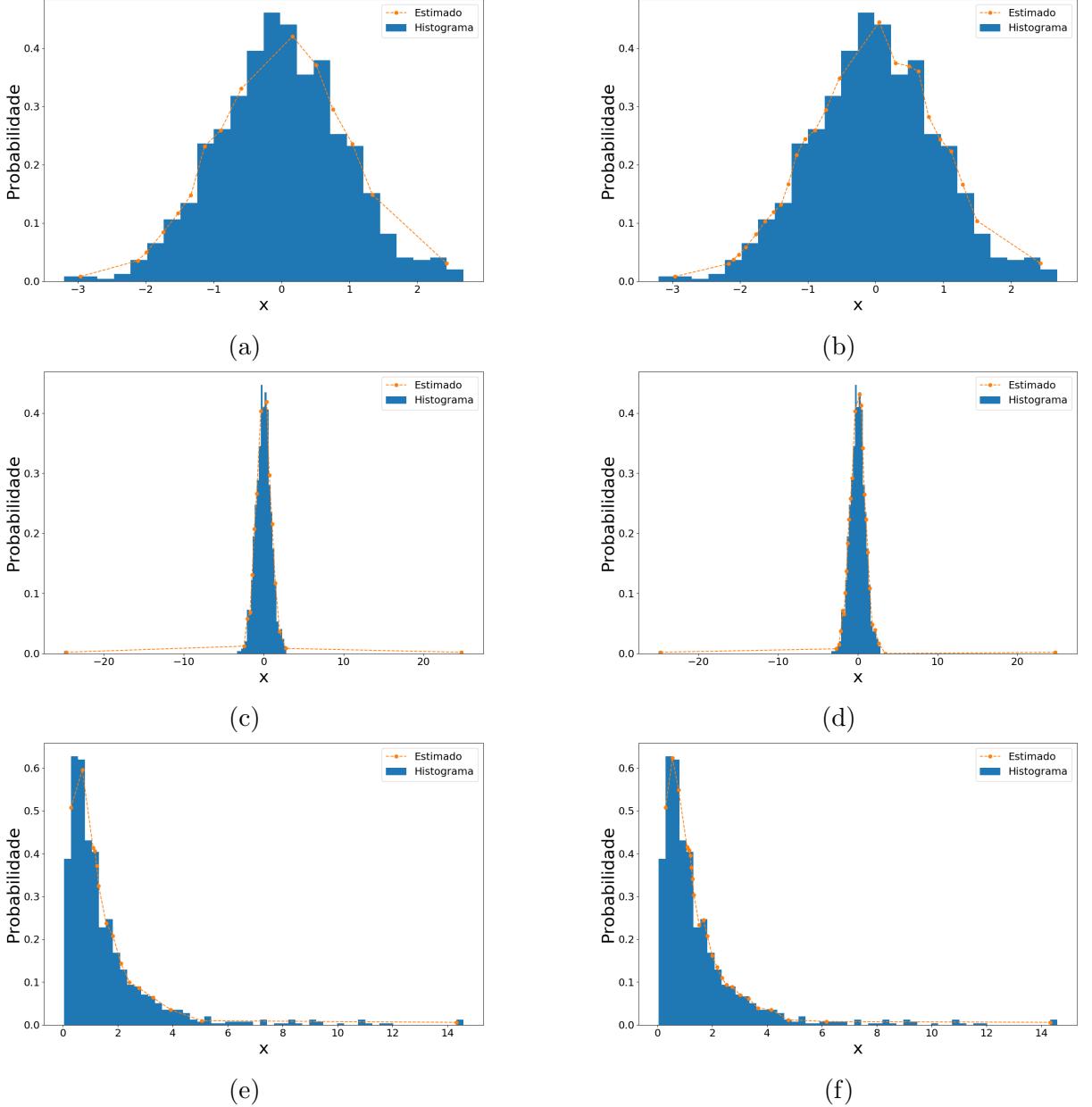


Figura 20: Discretização com os dados gerados utilizando o método iPDF1: (a) $\text{randN}(0,1)$ com $N = 15$, (b) $\text{randN}(0,1)$ com $N = 25$, (c) $\text{randN}(0,1)$ com $N = 15$ e *outlier* em ± 25 , (d) $\text{randN}(0,1)$ com $N = 25$ e *outlier* em ± 25 , (e) $\text{randL}(0,1)$ com $N = 15$ e (f) $\text{randL}(0,1)$ com $N = 25$.

3.3.5 MÉTODO IPDF2

A iPDF2 se baseia na segunda derivada, com isso, é esperado que o número de pontos seja maior onde haja uma variação de sua derivada e menor nos pontos de inflexão, como é mostrado na Figura 21a e 21b. O mesmo acontece para as Figuras 22e e 22f, mas, como a taxa de variação é muito maior na porção esquerda da distribuição, a quantidade de pontos do lado direito fica reduzido, fazendo com que o erro de estimativa aumente.

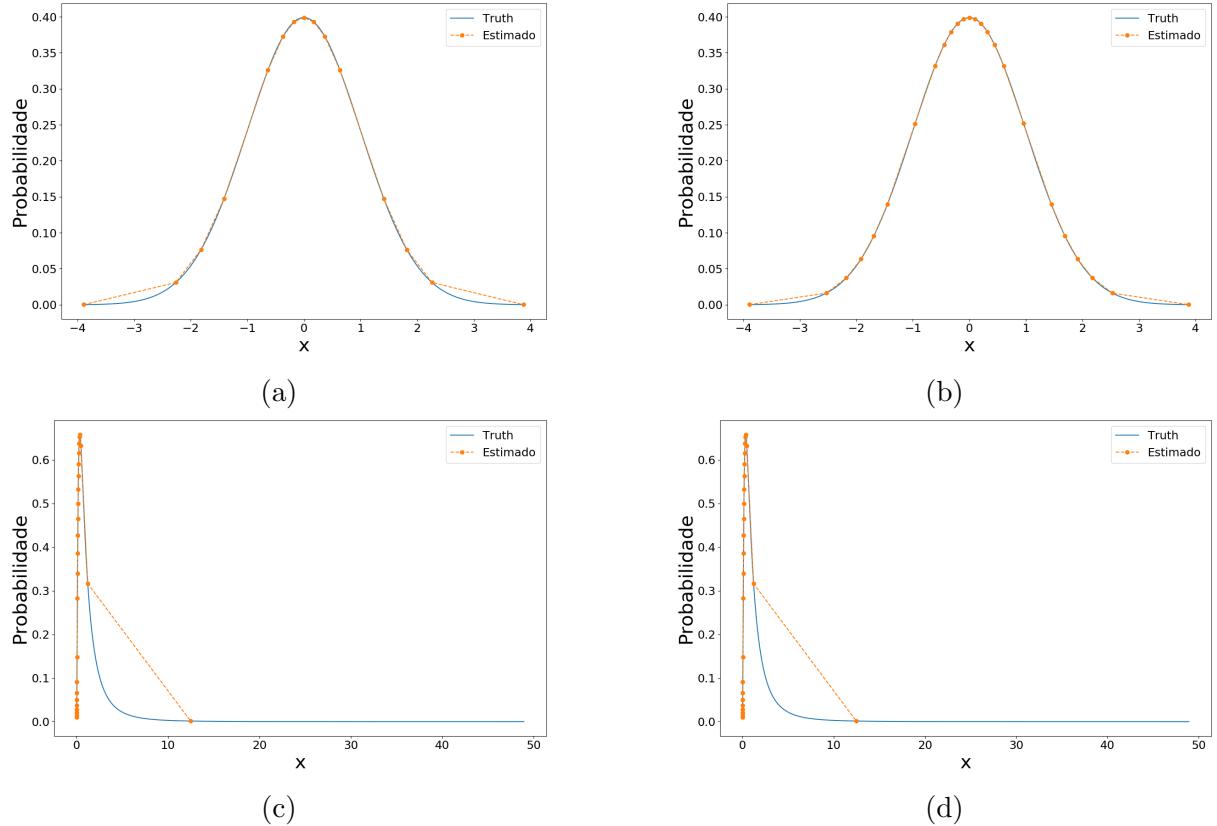


Figura 21: Discretização utilizando o método de *iPDF2*: (a) $N(0,1)$ com $N = 15$, (b) $N(0,1)$ com $N = 25$, (c) $L(0,1)$ com $N = 15$ e (d) $L(0,1)$ com $N = 25$.

Para os dados gerados, o mesmo problema ocorre no método iPDF1 só que com mais ruído, devido ao fato de ser a segunda derivada discreta, com isso a sua CDF é ainda mais suave, fazendo com que haja mais pontos na região de baixa probabilidade, diminuindo assim seu erro nessa região, mas aumentando na região onde a probabilidade é maior, conforme pode ser visto na Figura 22.

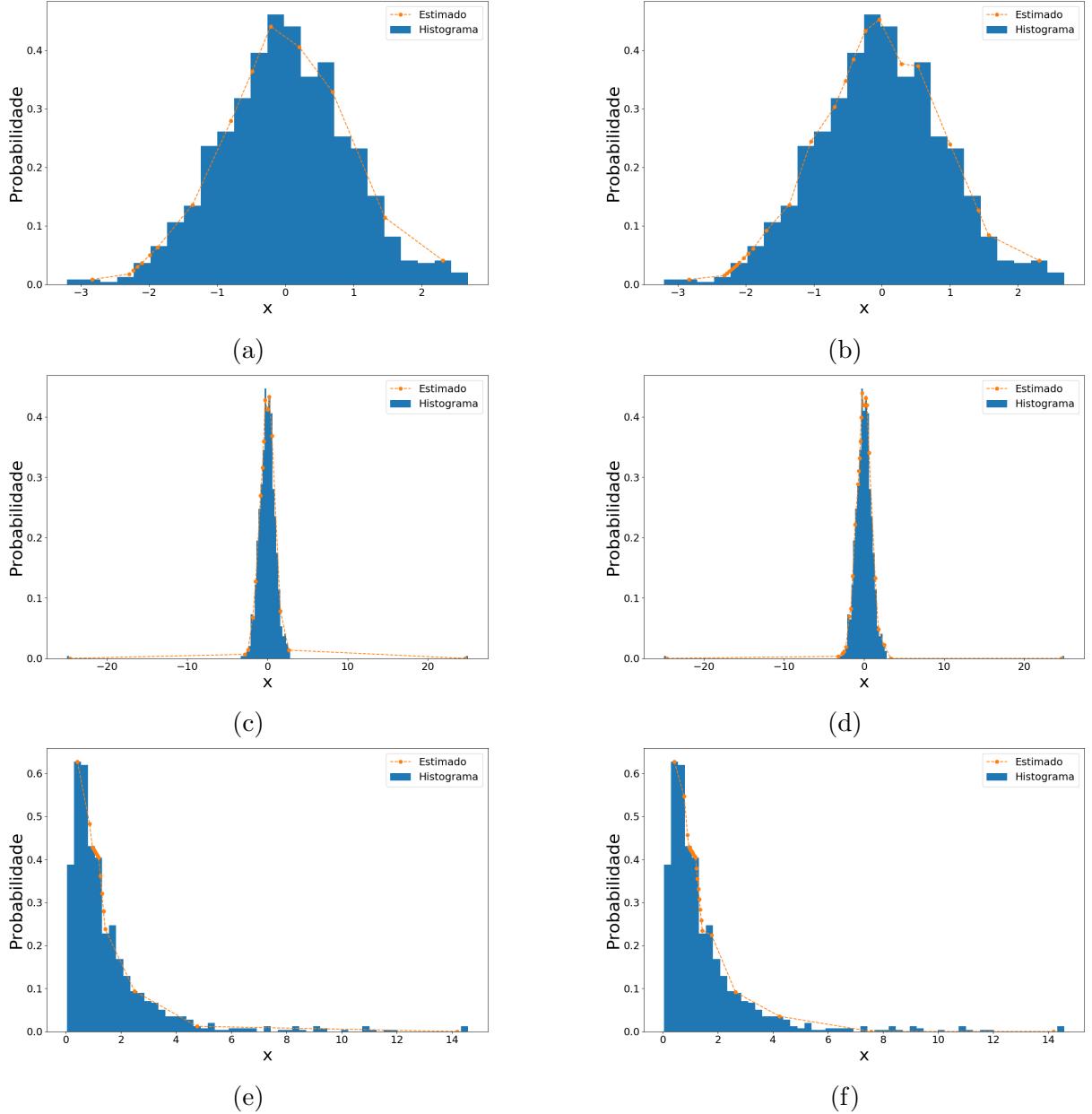


Figura 22: Discretização com os dados gerados utilizando o método iPDF2: (a) $\text{randN}(0,1)$ com $N = 15$, (b) $\text{randN}(0,1)$ com $N = 25$, (c) $\text{randN}(0,1)$ com $N = 15$ e *outlier* em ± 25 , (d) $\text{randN}(0,1)$ com $N = 25$ e *outlier* em ± 25 , (e) $\text{randL}(0,1)$ com $N = 15$ e (f) $\text{randL}(0,1)$ com $N = 25$.

Podemos ver que cada método se comporta melhor em uma determinada região, botando mais ou menos pontos nas regiões de interesse, ~~deste modo podemos analisar o custo computacional envolvido para cada método que será explicado na seção seguinte.~~

3.4 CUSTO COMPUTACIONAL

Os algoritmos **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 23 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.

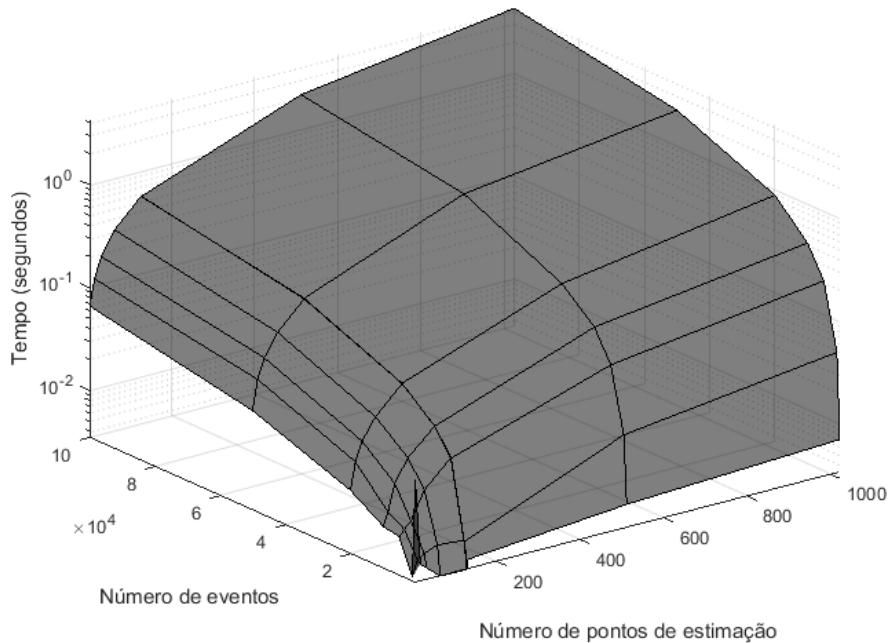


Figura 23: Gráfico do tempo de processamento de um algoritmo de estimação de densidades baseado em KDE quando aumenta-se o número de eventos a serem estimados e o número de pontos de estimação.

Pode-se observar que o tempo de processamento para $N = 1024$ é aproximadamente 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 estimativa de erro. Além disso, o eixo das abscissas foi dividido 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 24 mostra este processo quando a abscissa é dividida em 20 regiões, todas compreendidas entre os valores -4 e 4 do eixo x .

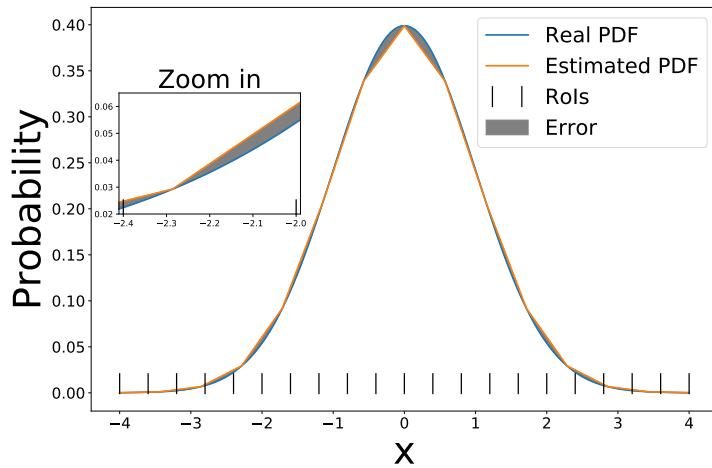


Figura 24: Ilustração da medida de erro entre a PDF Real e a Estimada com 20 regiões de interesse.

A maneira que a RoI é usada neste trabalho permitirá avaliar o erro de estimativa 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 amostras serão usadas no processo de discretização. O erro de estimativa tende a melhorias 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. Ém disso há outra figura que mostra como o método se comporta quando o número de pontos de estimativa varia sendo que, para os dados gerados, foram pegos 10 amostras que são representadas com a sua média e seu erro como o desvio padrão dessa média sobre a raiz quadrada do número de amostras. Essas perspectivas diferentes irão permitir um melhor entendimento das características de cada método. Essa sessão será dividida em três análises diferentes. Sessão 4.1 analisa a estimativa de erro quando a interpolação pelo vizinho mais próximo é usada; Sessão 4.2 avalia para a interpolação linear; e a Sessão 4.3 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 conhecido mais próximo ao valor da probabilidade da variável aleatória que será estimada. Portanto, o erro de estimativa 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 25a, 25b, 26a e 26b pode-se inferir que:

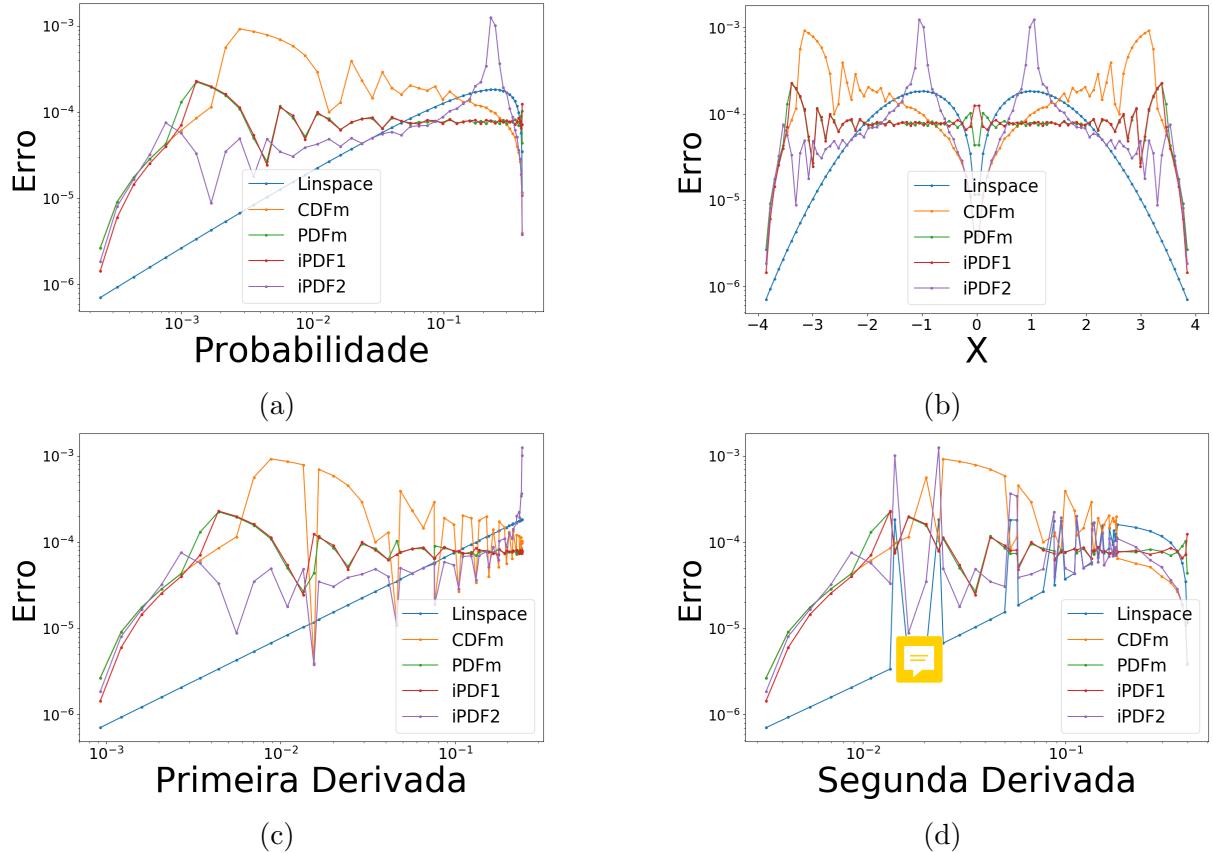


Figura 25: Caso representativo da distribuição Gaussiana com 200 pontos, 100 RoI e usando a interpolação pelo vizinho mais próximo.

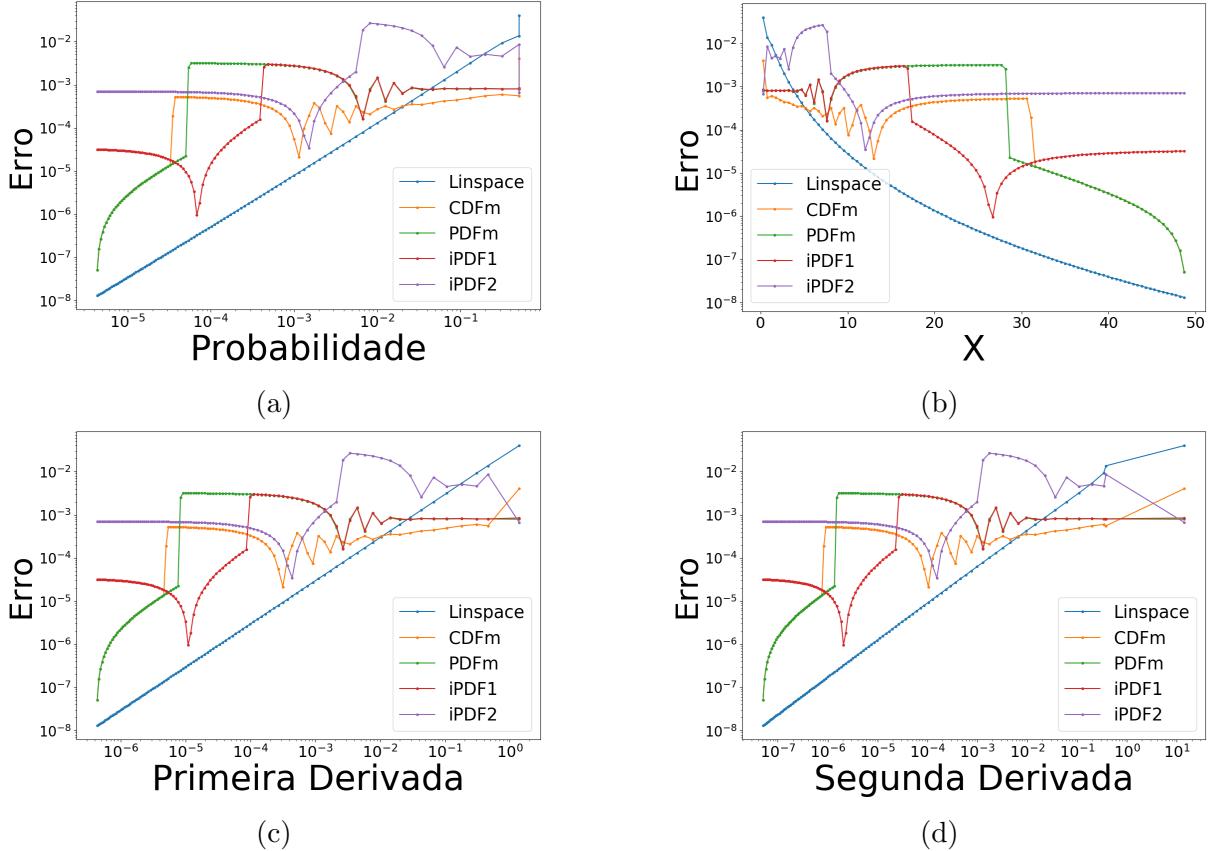


Figura 26: Caso representativo da distribuição Lognormal com desvio padrão $\sigma = 1$ com 200 pontos, 100 RoI e usando a interpolação pelo vizinho mais próximo.

Linspace erro de estimação aumenta com a 1^a derivada da função que está sendo estimada;

CDFm erro de estimação diminui com o aumento da probabilidade;

PDFm e **iPDF1** tendem a equalizar o erro de estimativa aumentando a densidade de pontos discretos nas regiões de derivadas mais altas;

iPDF2 diminui o erro com o aumento da probabilidade, no entanto, apresenta um aumento de erro próximo ao ponto de inflexão.

Na Figura 25c e 26c é possível perceber que:

Linspace apresenta um aumento do erro de estimação diretamente proporcional ao valor da 1^a derivada;

CDFm apresenta maior densidade de pontos na região de alta probabilidade e poucos pontos na região de baixa probabilidade. No entanto, do ponto de vista da 1^a

derivada, o erro de estimativa oscila entre valores altos e baixos para uma distribuição mais lenta;

PDFm e iPDF1 tendem a manter constante o valor do erro de estimativa, uma vez que se coloca mais pontos em regiões com derivadas maiores, onde o erro do método de discretização pelo vizinho mais próximo é maior. Porém, como mencionado anteriormente, as caudas apresentam os menores erros, o que justifica a diminuição do erro nas regiões de baixa derivada;

iPDF2 aumenta o erro de estimativa de acordo com o aumento da 1^a derivada, uma vez que tais regiões tendem a receber menos pontos estimados.

Finalmente, quando a Figura 25d e 26d são observadas (lembrando que valores baixos de 2^a derivada representam regiões intercaladas de baixa probabilidade e inflexão, de modo que, estas são as situações onde o erro causado pela interpolação é menor) ela pode ser inferida que:

Linspace e iPDF2 exibem comportamento similar, aumentam o erro diretamente proporcional à 2^a derivada, porém, ao se aproximarem da região de alta probabilidade, o erro tende a cair;

CDFm tende a diminuir em  a medida que a segunda derivada aumenta;

PDFm e iPDF1 não parecem sofrer  com a variação da 2^a derivada, exceto em regiões de valores baixos, onde o erro flutua.

Podemos também realizar uma análise de desempenho verificando a equivalência do erro de estimativa para quando o número de pontos varia, conforme mostra a Figura 27.

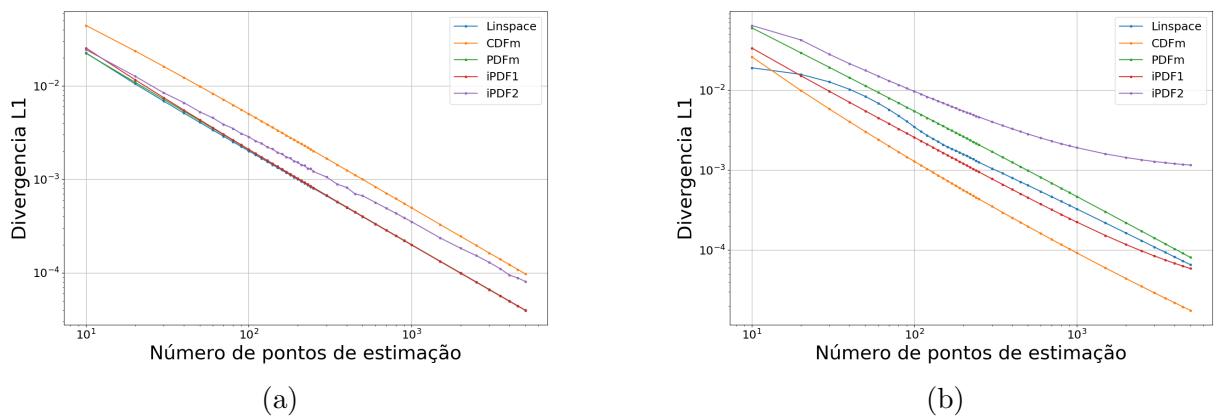


Figura 27: Erro total de estimativa para a interpolação pelo vizinho mais próximo variando-se o números de pontos a se estimar: (a) $N(0,1)$ e (b) $L(0,1)$

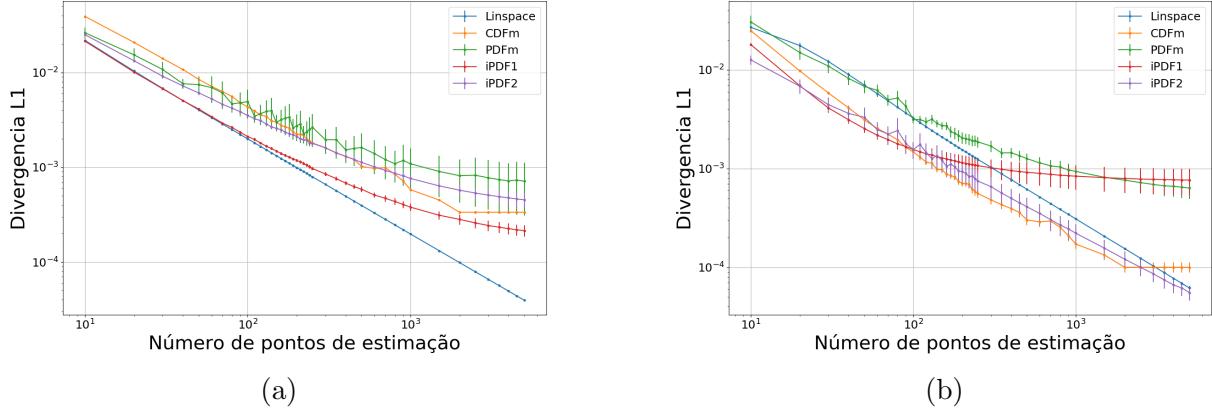


Figura 28: Erro total de estimação para a interpolação pelo vizinho mais próximo variando-se o números de pontos a se estimar: (a) $\text{randN}(0,1)$ e (b) $\text{randL}(0,1)$

Podemos perceber que na Figura 27a, devido ao fato de possuir uma **derivada mais lenta**, o método *Linspace* é o que possui o menor erro, juntamente com os métodos *PDFm* e *iPDF1*. Por outro lado, o método *CDFm* é o que possui o pior desempenho, fazendo com que se necessite de mais pontos de estimação para possuir o mesmo erro total que os outros métodos. Já quando a distribuição possui uma **variação maior**, a cena se inverte, como é mostrada na Figura 27b em que a *CDFm* é a que possui o menor erro, sendo necessário, por exemplo possuir apenas 120 pontos de estimação enquanto o *Linspace* necessita de 300 pontos para manter o mesmo erro.

4.2 ESTIMAÇÃO DE ERRO PELA INTERPOLAÇÃO LINEAR

O método de interpolação linear naturalmente tende a apresentar erros de estimação mais altos em regiões com a 2^a derivada maior se os pontos discretos forem distribuídos uniformemente ao longo do eixo horizontal, como pode ser notado pelo método *Linspace* da Figura 29d e 30d. Adicionalmente, analisando as Figuras 29a, 29b, 30a e 30b é possível observar o seguinte:

Linspace erro de estimação diminui em regiões de baixa probabilidade e perto dos pontos de inflexão;

CDFm erro de estimação diminui com o aumento da probabilidade, e nos pontos de inflexão há uma melhoria ainda maior;

PDFm e iPDF1 erro aumenta em regiões de baixa e alta probabilidade e diminui próximo aos pontos de inflexão;

iPDF2 apresenta menor densidade de pontos em regiões de baixa probabilidade e regiões de inflexão, levando a um comportamento inverso quando comparado ao método *Linspace*.

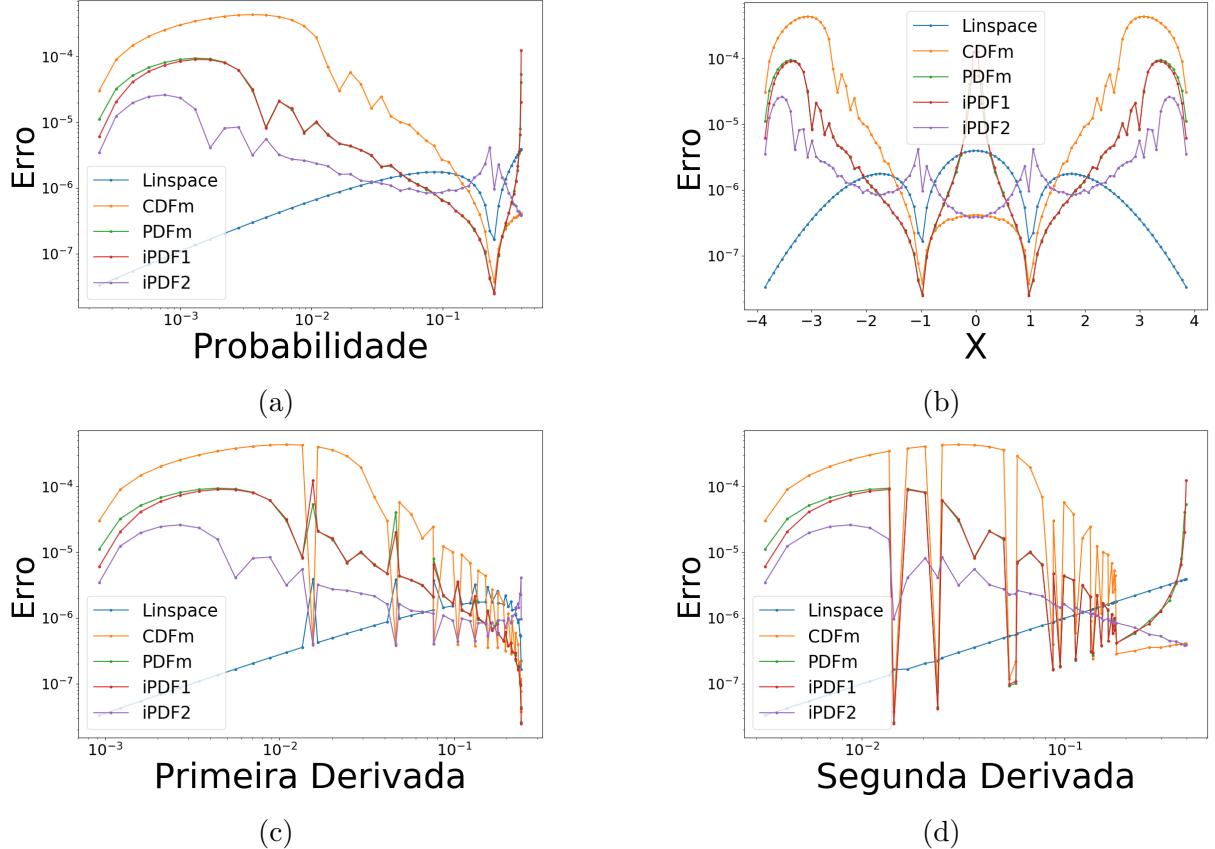


Figura 29: Caso representativo da distribuição Gaussiana com 200 pontos, 100 RoI e usando a interpolação linear.

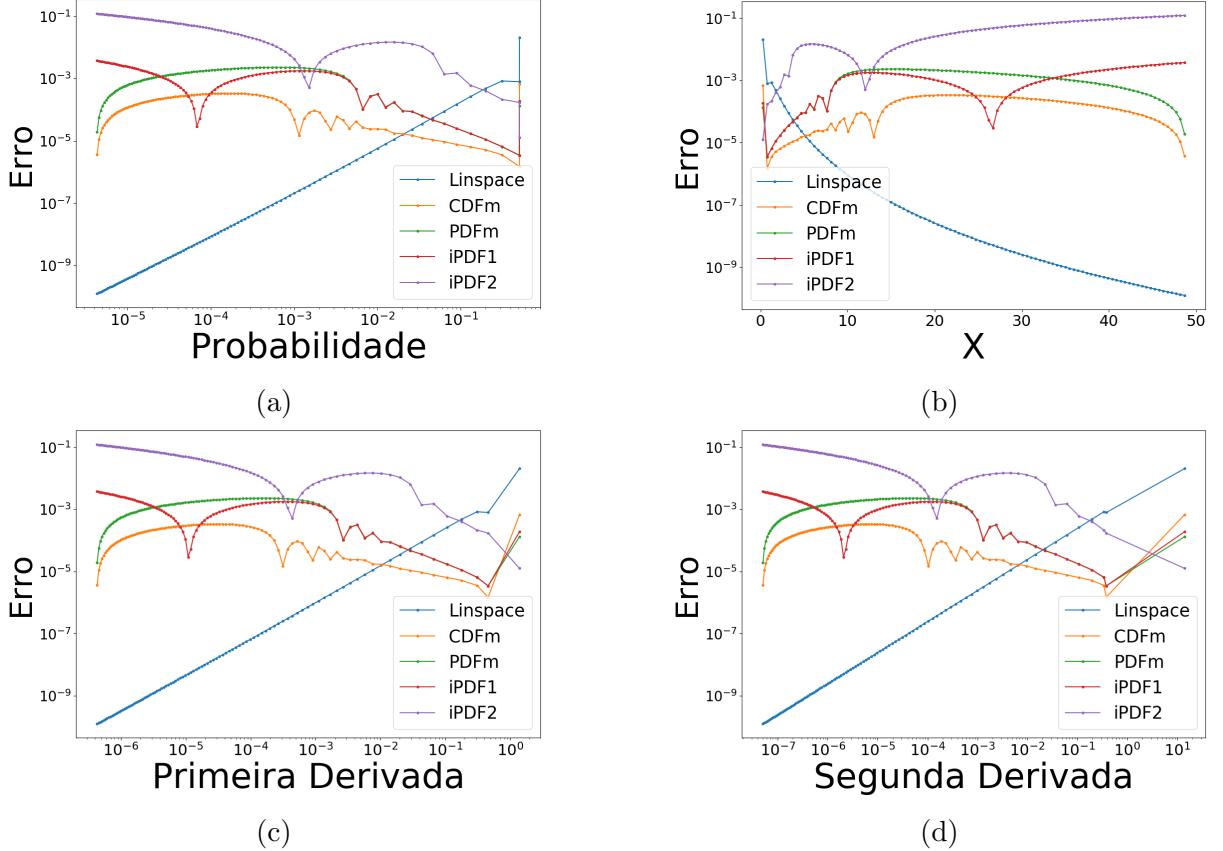


Figura 30: Caso representativo da distribuição Lognormal com desvio padrão $\sigma = 1$ com 200 pontos, 100 RoI e usando a interpolação linear.

Avaliando as Figuras 29c e 30c é possível confirmar que os métodos *Linspace* e *iPDF2* têm comportamentos opostos. Os outros métodos se comportam de maneira semelhante, pois reduzem o erro de estimativa quando a 1^a derivada aumenta. Nas Figuras 29d e 30d, é possível notar que o método *Linspace* mostra a mesma tendência em relação à 2^a derivada para o caso da interpolação linear da apresentada em relação à 1^a derivada para o caso da interpolação pelo vizinho mais próximo. A flutuação de erro observada no método *CDFm* é causada pela variação entre as regiões de alta e baixa probabilidade ao longo do eixo da 2^a derivada, no entanto, uma vez que as regiões de alta probabilidade estão associadas à regiões de segunda derivada alta, esta oscilação tende a desaparecer com o aumento desta derivada. Isso também explica as flutuações nos métodos *PDFm* e *iPDF1*, porém esses métodos apresentam desempenho degradado em regiões de alta probabilidade. Finalmente, o método *iPDF2* mantém o comportamento inverso ao método *Linspace*.

Ao mudarmos para a interpolação linear, é possível perceber que o erro diminui de maneira considerável, como pode ser visto na Figura 31 em que para a distribuição Normal, o método *Linspace* se sobressai a todos os outros e, como na interpolação pelo vizinho mais próximo o método *CDFm* é o pior, como ilustra a Figura 31a, já para a distribuição

Lognormal, o método CDFm se sobressai até os 1000 pontos de estimativação, após o método *Linspace* se torna mais eficiente como é ilustrado na Figura 31b devido ao fato de ser o único método que estima com maior precisão as regiões de baixa probabilidade.

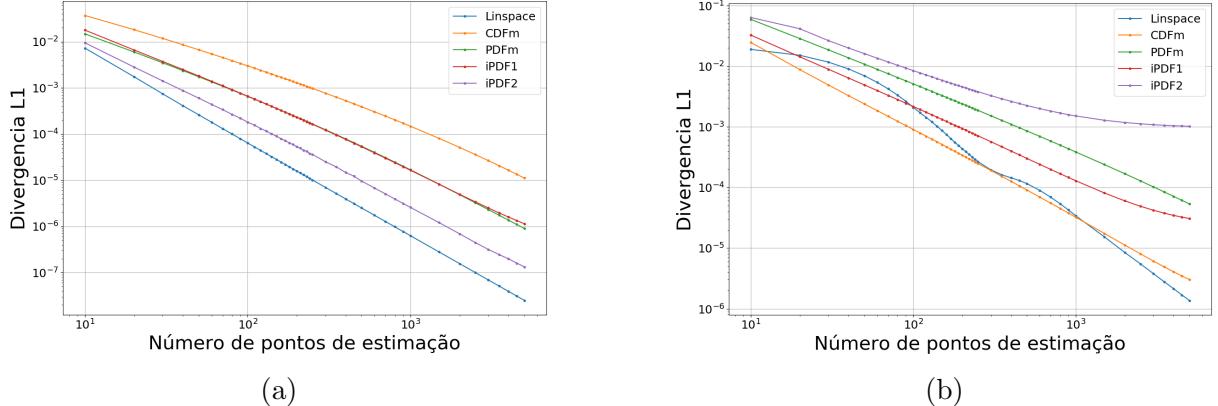


Figura 31: Erro total de estimação para a interpolação linear variando-se os números de pontos a se estimar: (a) $N(0,1)$ e (b) $L(0,1)$

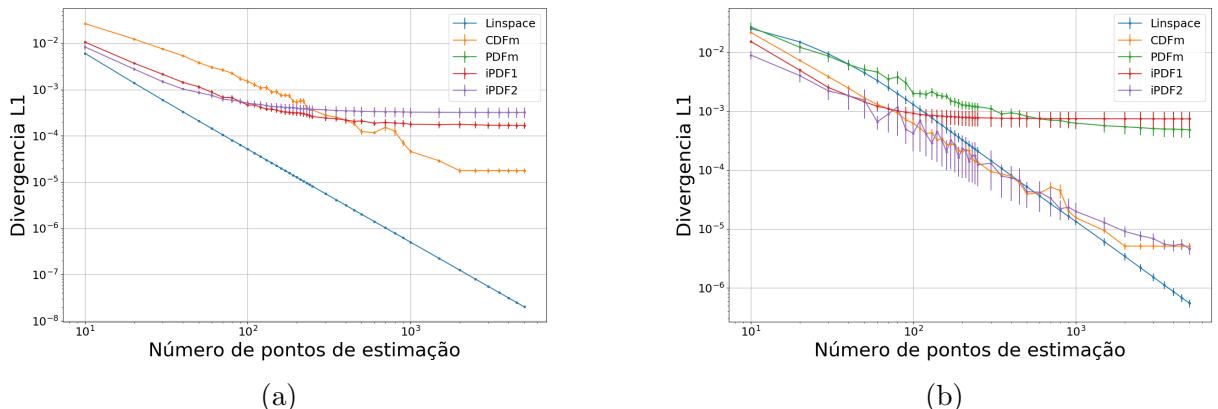


Figura 32: Erro total de estimação para a interpolação linear variando-se os números de pontos a se estimar: (a) $\text{randN}(0,1)$ e (b) $\text{randL}(0,1)$

4.3 ESTIMAÇÃO DE ERRO CONSIDERANDO OUTLIERS

A fim de verificar o comportamento dos métodos de discretização em uma realidade onde existem conjuntos de dados com *outliers*, como é comum em experimentos reais, *outliers* foram inseridos nos dados gerados. As posições *outliers* foram varridas até o valor de 50. O problema dos *outliers* pode ser visto de outra perspectiva, relacionado à definição dos limites do eixo horizontal, que é geralmente um pré-requisito para aplicar algoritmos de estimativa de PDF. Além disso, o número de pontos estimados foi varrido para 1000. Esta análise pode ser vista na Figura 33.

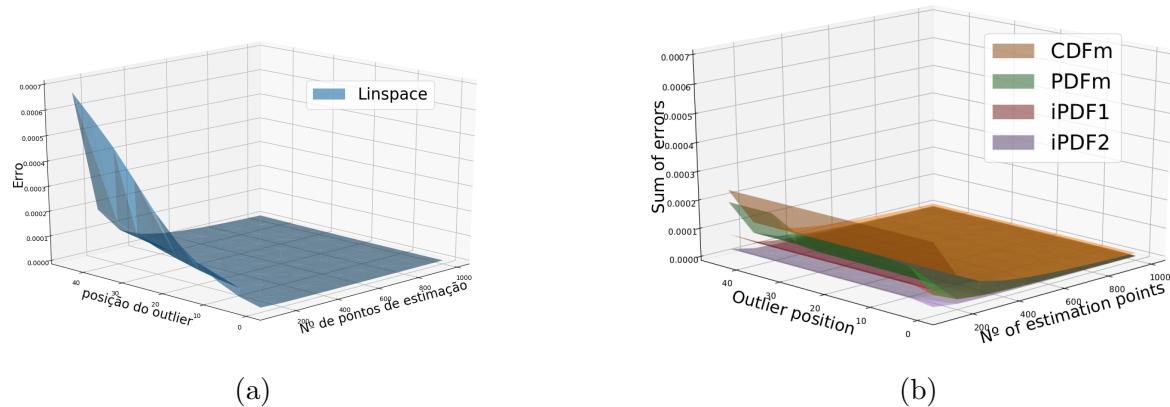


Figura 33: Análise de *outlier* usando 100 Rois e interpolação linear.

A figura 33a mostra que o método *Linspace* aumenta consideravelmente seu erro de estimação quando *outliers* são inseridos; quanto mais distantes os *outliers*, maior o erro. Este efeito é mitigado pelo aumento do número de pontos estimados. A Figura 33b mostra que os métodos propostos são menos sensíveis aos *outliers* (ou à escolha dos limites do eixo horizontal). As tabelas 1 e 2 mostram o erro médio dos métodos para três posições de *outliers* (20 e 50) e 100 pontos de estimação), para interpolações do vizinho mais próximo e linear, respectivamente.

Tabela 1: Erro de estimativa média usando a distribuição normal e interpolação do vizinho mais próximo com 100 pontos de estimação.

Média					
#Outlier	Linspace	CDFm	PDFm	iPDF1	iPDF2
0	8.02e-5	1.96e-4	8.26e-5	8.19e-5	1.20e-4
20	4.82e-4	1.96e-4	1.14e-4	9.22e-5	1.25e-4
50	1.09e-3	1.96e-4	1.14e-4	9.22e-5	1.25e-4

Tabela 2: Erro de estimacao media usando a distribuiao normal e interpolaao linear com 100 pontos de estimacao.

Média					
#Outlier	Linspace	CDFm	PDFm	iPDF1	iPDF2
0	1.30e-6	1.01e-4	2.14e-5	2.07e-5	4.97e-6
20	4.66e-5	1.01e-4	5.61e-5	3.01e-5	8.35e-6
50	2.31e-4	1.01e-4	6.38e-5	3.00e-5	8.35e-6

Para a interpolação do vizinho mais próximo, o método iPDF1 apresentou o melhor

desempenho enquanto para a interpolação Linear, o iPDF2 foi o melhor se o erro de estimação e a sensibilidade de *outliers* forem considerados. O CDFm mostrou-se praticamente imune a *outliers*. Quando nenhum valor discrepante está presente, o Linspace atinge o melhor resultado seguido de perto pelo PDFm e iPDF1 para o caso do vizinho mais próximo e pelo iPDF2 para o caso Linear. No entanto, é influenciado pela escolha arbitrária de 99.99% da área como limites do eixo horizontal padrão para caracterizar a ausência de *outliers*.

A Figura 34 mostra como o erro varia conforme é aumentado a quantidade de pontos de estimação, sendo que, como é mostrado na Figura 34a, há uma mudança na performance de cada método, fazendo com que alguns se sobressaiam sobre outros conforme é aumentado o número de pontos, para este caso, utilizando a interpolação pelo vizinho mais próximo, os métodos iPDF1 e iPDF2 foram os que tiveram o menor erro até os 5000 pontos estimados. Já para a interpolação Linear, mostrada na Figura 34b os métodos iPDF1, iPDF2 e CDFm estagnam por volta do 10^0 ponto de estimação, não alterando mais tanto o valor do erro após isso, em contrapartida, o método *Linspace* continua a melhorar conforme o número de pontos aumenta, chegando ser o melhor a partir do 1000^0 ponto.

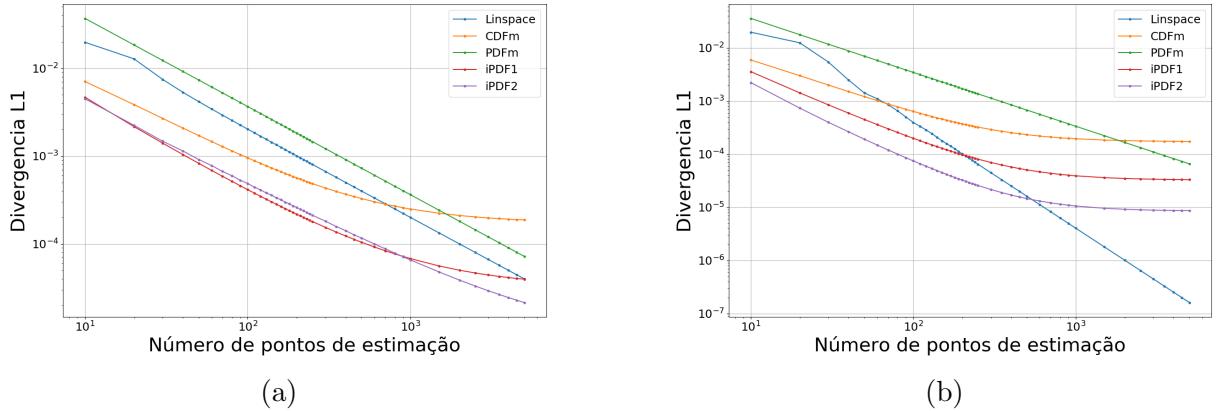


Figura 34: Erro total de estimação para a distribuição Normal com outlier em 25: (a) para a interpolação pelo vizinho mais próximo, (b) para a interpolação linear.

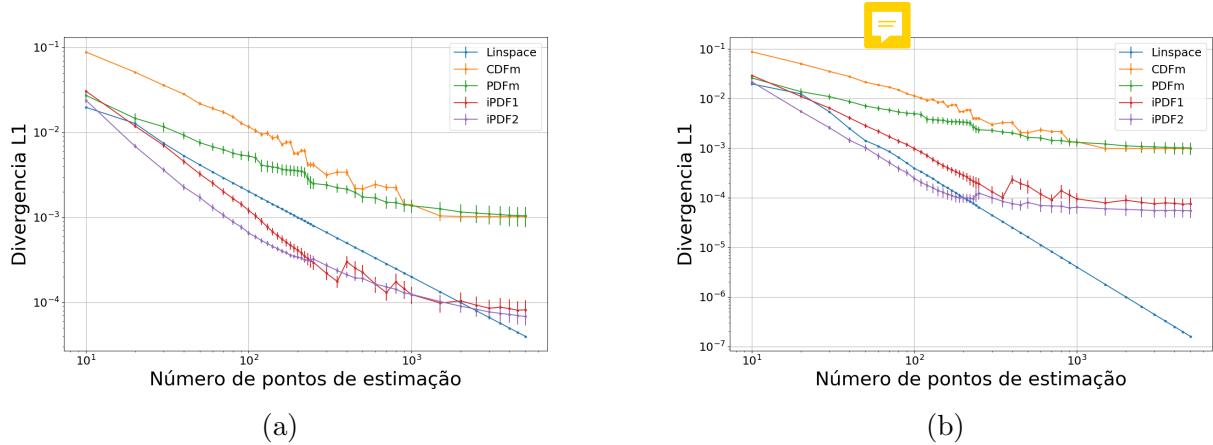


Figura 35: Erro total de estimacao para a distribuiao $\text{randN}(0,1)$ com outlier em 25: (a) para a interpolaao pelo vizinho mais proximo, (b) para a interpolaao linear.

4.4 EXEMPLO PRÁTICO

Vamos agora supor que um *dataset* qualquer vindo de um sensor possui 100 mil eventos e você medido 25 vezes, e que estes geraram o histograma da Figura 36

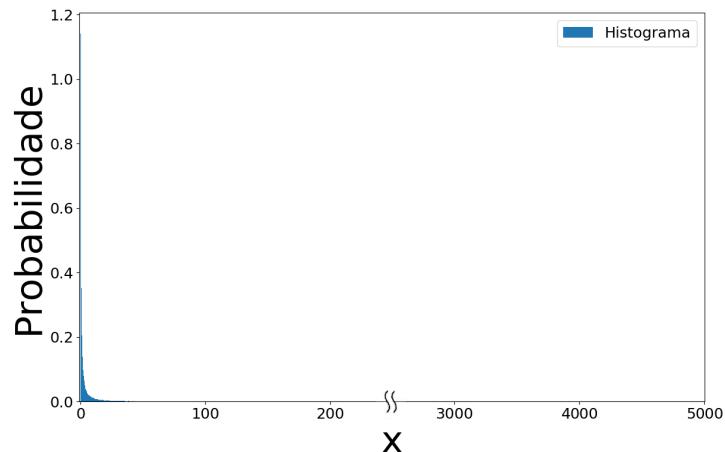


Figura 36: Histograma representativo de uma realidade com 100 mil eventos

Como pode ser visto, este histograma possui uma variação muito rápida, o que torna difícil de ser discretizada pelo método *Linspace*, visto que o mesmo irá necessitar de muitos pontos de estimação para que o erro permaneça pequeno 

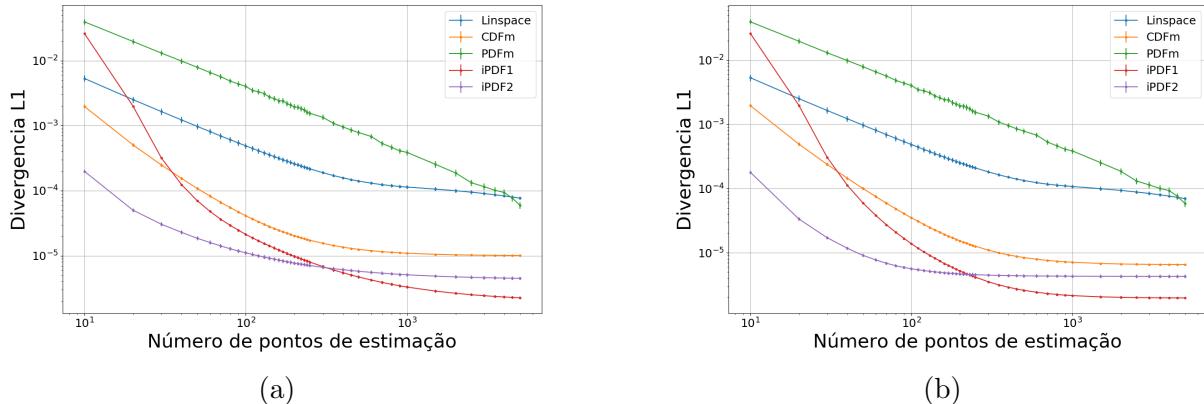


Figura 37: Erro total de estimativa para a distribuição da Figura 36: (a) para a interpolação pelo vizinho mais próximo, (b) para a interpolação linear.

Devido ao fato da distribuição possuir valores muito distantes e a variação muito rápida, o método de interpolação já não faz tanto efeito, como pode-se ver pela Figura 37. Entretanto, para este caso, o método iPDF2 foi o que obteve o melhor resultado que pode ser visto nas Tabelas 3 e 4 mostrando a diferença do número de pontos de estimativa para cada estimativa possuindo o mesmo erro.

Tabela 3: Equivalência de pontos para o mesmo erro em diferentes métodos utilizando-se a interpolação linear.

	Número de Pontos de Discretização				
Divergência L1	Linspace	CDFm	PDFm	iPDF1	iPDF2
2e-4	260	35	1900	35	10
1e-5	>5000	350	>5000	120	45
6.2e-6	>5000	>5000	>5000	170	85

Tabela 4: Equivalência de pontos para o mesmo erro em diferentes métodos utilizando-se a interpolação pelo vizinho mais próximo.

	Número de Pontos de Discretização				
Divergência L1	Linspace	CDFm	PDFm	iPDF1	iPDF2
2e-4	300	35	1900	35	10
1e-5	>5000	4480	>5000	200	120
6.2e-6	>5000	>5000	>5000	340	400

As distribuições que são baseadas em derivadas levam uma vantagem em relação às outras quando se trata de dados gerados, devido ao fato de suas distribuições ficarem mais

ruidosas que o normal, o que suaviza sua CDF fazendo com que coloque mais pontos ao longo de toda distribuição ao invés de se concentrar mais na região de alta probabilidade, como é o caso do método CDFm.

Os resultados apresentados mostraram que os métodos propostos são mais resilientes aos *outliers* e destacaram as vantagens e desvantagens de cada um, mostrando que o processo de discretização é de fundamental importância para minimizar o erro de estimativa. Além disso, este trabalho fornece uma base sólida de conhecimento sobre esta questão, tornando este estudo uma possível base para futuras investigações  profundadas.

5 CONCLUSÃO

A discretização contínua de dados representa uma importante tarefa de pré-processamento no contexto de estimação de dados. Até agora, muito trabalho foi feito para melhorar a discretização, a fim de reduzir as informações redundantes ou desconectadas.

Este trabalho abordou o problema de discretização no cenário de estimação de densidades, fazendo uma avaliação cuidadosa do assunto na literatura, propondo quatro diferentes métodos de discretização e comparando-os com o método *Linspace*, amplamente utilizado na literatura. Em particular, o início e o final da região da variável aleatória devem ser definidos antes de aplicar a discretização, cujo desempenho é altamente sensível a esses parâmetros. Nesse contexto, pode ser importante procurar métodos mais resilientes.

Vale ressaltar que estes métodos se comportaram de maneiras diferentes quando submetidos a funções analíticas ou com dados gerados, sendo que este *ultimo* se agrava em distribuições assimétricas ou com caldas muito longas devido ao ruído que é inserido quando se deriva uma função discreta que já é ruidosa, e, curiosamente, este ruído contribuiu para que estes métodos (iPDF1 e iPDF2) apresentassem resultados melhores que os esperados, fazendo-os uma boa opção de discretização apesar de seu algoritmo ser um pouco mais complexo de se implementar se comparado aos métodos *Linspace* e CDFm.

5.1 PRÓXIMOS PASSOS

Ao final desse estudo, fica claro que existem possibilidades de melhorias nos métodos apresentados, tendo a necessidade de uma análise mais profunda para outras distribuições além das Normais e Lognormais além de se buscar outros métodos *complementares* aos já abordados ou até mesmo algum método que junte dois ou mais métodos, fazendo com que se pegue a melhor região para cada um deles.

REFERÊNCIAS

- ALISON, J. *The road to discovery: Detector alignment, electron identification, particle misidentification, $w w$ 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 estimativa 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.