Universidade Federal de Juiz de Fora
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

Qualificação apresentada ao Programa de Pós–Graduação em Engenharia Elétrica, área de concentração: Sistemas Eletrônicos, da Faculdade de Engenharia da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do grau de Doutor.

Orientadores: Prof. Rafael Antunes Nóbrega, D.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. 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

Qualificação apresentada ao Programa de Pós–Graduação em Engenharia Elétrica, área de concentração: Sistemas Eletrônicos, da Faculdade de Engenharia da Universidade Federal de Juiz de Fora como requisito parcial para obtenção da graduação.

Aprovada em 06 de Setembro de 2018.

BANCA EXAMINADORA:

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

Prof. Ernesto Kemp, D.Sc.
Universidade Estadual de Campinas, Unicamp

Prof. Leandro Rodrigues Manso Silva, D.Sc.
Universidade Federal de Juiz de Fora, UFJF

Aos meus pais, meus irmãos, minha namorada, aos meus familiares, aos meus amigos.

AGRADECIMENTOS

Não vemos as coisas como elas são, mas como nós somos.

Anaïs Nin

RESUMO

A identificação de partículas é de fundamental importância para os experimentos de física de altas energias desenvolvidos ao redor do mundo. Nesse ambiente de física de partículas, a probabilidade de ocorrência de partículas relevantes aos estudos propostos são baixíssimas em relação às partículas que formam o ruído de fundo, exigindo algoritmos com índices de eficiência de detecção dos sinais de interesse e rejeição de ruído de fundo cada vez melhores. Muitos métodos de identificação de partículas fazem uso da técnica de verossimilhança, esse método utiliza a densidade de probabilidade das variáveis para criar um discriminador. Sendo assim, para garantir a performance da classificação faz-se necessário uma boa qualidade de estimação das distribuições em questão, sendo estas comumente diferentes das distribuições conhecidas e parametrizadas na literatura. Nessa tese, os métodos aplicados à estimação de densidade não-paramétrica serão revisados e possíveis otimizações serão avaliadas a partir dos dados produzidos por um dos maiores experimentos do CERN, o ATLAS. Concentrado no contexto *offline*, o trabalho reproduz o método baseado em verossimilhança e propõe algumas melhorias com o uso de algoritmos de processamento mais otimizados, estatística robusta e técnicas de discretização diferentes das utilizadas na literatura. Além disso, esse trabalho se propõe a expandir as ferramentas utilizadas na estimação univariada para o ambiente de estimação de densidades multivariada, conhecida como MKDE (do inglês, *Multivariate Kernel Density Estimation*), que pode ser capaz de mitigar o erro inserido na consideração de independência entre as variáveis discriminantes inserida pelo método de *Likelihood* atualmente em uso por vários experimentos desse tipo. Inicialmente, este trabalho se propõe a implementar o método de verossimilhança baseando-se na estimação de densidades univariadas usadas na reconstrução da densidade conjunta das variáveis discriminantes e a estudar o impacto de possíveis parâmetros relacionados à implementação do algoritmo de estimação de densidades univariadas. Em uma segunda etapa, a implementação do MKDE é inserida através de uma comparação direta com o método univariado.

Palavras-chave: Estimação de Densidades, Verossimilhança, *FastKDE*, Discretização.

ABSTRACT

The particle identification has fundamental importance to the high-energy physics experiments developed around the world. In this environment of particle physics, the particles occurrence probability relevant to the proposed studies is very low in relation to the particles that form the background noise, requiring algorithms with efficiency indices of interest signals detection and background noise rejection each time best. Many methods of identifying particles using the likelihood technique, which uses the probability distribution of variables to create a discriminator. Therefore, to guarantee the performance of the classification, a good quality of estimation of the distributions in question is necessary, being these commonly different from the distributions known and parameterized in the literature. In this thesis, the methods applied to the estimation of non-parametric density will be reviewed and possible optimizations will be evaluated from the data produced by one of the largest experiments of CERN, the ATLAS. Concentrated in the offline context, the paper reproduces the likelihood-based method and proposes some improvements with the use of more optimized processing algorithms, robust statistics and discretization techniques different from those used in the literature. In addition, this work proposes to expand the tools used in the univariate estimation for the multivariate density estimation environment, known as MKDE (Multivariate Kernel Density Estimation), which may be able to mitigate the inserted error in the consideration of independence between the discriminant variables inserted by the method of Likelihood currently in use by several experiments of this type. Initially, this work proposes to implement the likelihood method based on the estimation of univariate densities used in the reconstruction of the joint density of the discriminant variables and to study the impact of possible parameters related to the implementation of the algorithm for estimating univariate densities. In a second step, the implementation of MKDE is inserted through a direct comparison with the univariate method.

Keywords: Density estimation, Likelihood, FastKDE, Discretization.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

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

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

LHC do inglês, *Large Hadron Collider*

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

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

FD Estimador Freedman–Diaconi

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

SUMÁRIO

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 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 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

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 os temas de discretização, estatística robusta e medidas de distância, mostrando que são temas que mesmo sendo discutidos há décadas ainda estão sendo utilizados, explorados e em desenvolvimento. Além disso, experimentos complexos de Big Data, como os do LHC, têm aplicado análises usando estimação de densidade em conjunto com técnicas de verossimilhança empregados em problemas de identificação de partículas obtendo resultados relevantes, mesmo utilizando uma simplificação da formulação matemática desse método, assumindo independência entre variáveis. Portanto, abre-se a possibilidade de contribuir nessa área no que diz respeito a otimização da estimação de densidades e suas nuances, usando como base de desenvolvimento um sistema altamente complexo,

com grande número de variáveis e distribuições com características bastante distintas, como ocorre com os experimentos do LHC.

1.2 ESTRUTURA DO TRABALHO

Este documento está organizado da seguinte maneira: XXXXXXXX

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, enquanto regiões de alta probabilidade terão que usar mais pontos a fim de minimizar o erro de estimação.

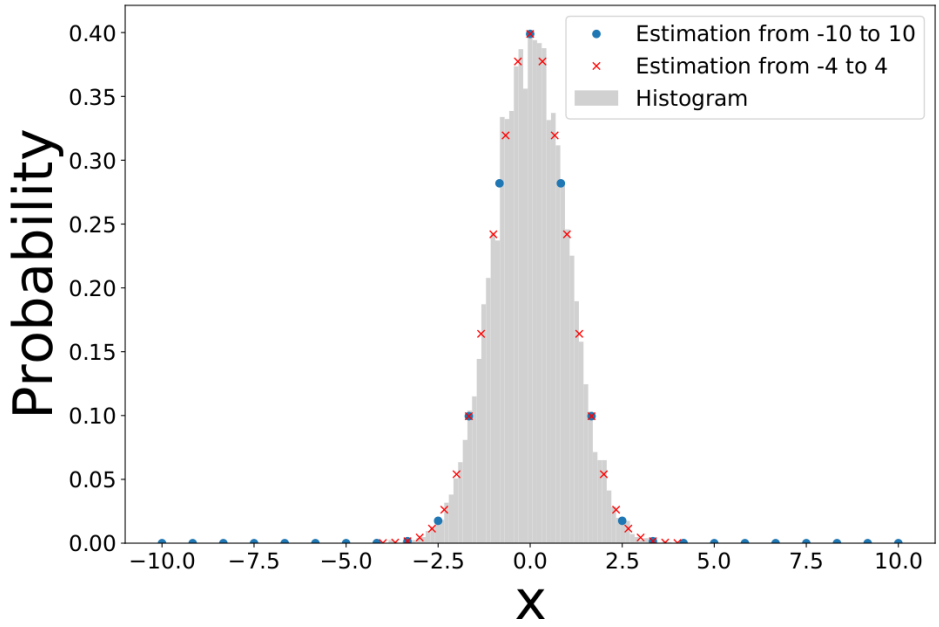


Figura 1: Caso representativo de estimação de PDF utilizando 25 pontos para dois intervalos diferentes (-4,4) e (-10,10)

3 DESENVOLVIMENTO

Neste capítulo, será apresentado as propostas de métodos de discretização, sua descrição e equacionamento. Além disso, o contexto em que estes métodos serão avaliados também será mostrado.

Como o objetivo deste trabalho é validar apenas os efeitos da discretização no contexto da estimação de PDF, o erro de estimação causado por este processo é medido entre a saída do processo de discretização em si e a função usada para gerar os dados. As funções aqui testadas serão baseadas em distribuições Gaussianas ou Lognormais com diferentes variâncias, bem como suas funções analíticas ou através de dados gerados.

3.1 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 mostrados utilizando PDFs analíticas Gaussiana, com média $\mu = 0$ e desvio padrão $\sigma = 1$, descrita pela Equação (3.1) e ilustrada na Figura 2, Lognormal, com $\mu = 0$ 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 3 e com três *Datasets* diferentes, o primeiro sendo de uma distribuição normal com média nula e desvio padrão unitário representado na Figura 4a, o segundo será também uma distribuição normal com os

mesmos parâmetros da primeira mas com a diferença que este irá possuir *outliers*, ou seja, alguns pontos distantes da região de interesse ilustrado pela Figura 4b e, por fim, a ultima será uma distribuição Lognormal com média nula e desvio padrão de 0.5, ilustrado pela Figura 4c. Todos os *data sets* possuem mil eventos e foram gerados utilizando a biblioteca *numpy* do *software Python* e com o número de *binagem* definido pelo Estimador Freedman–Diaconi (FD) que é um estimador robusto que leva em conta a variabilidade dos dados e o tamanho dos mesmos. Todos os métodos testados terão o número de estimação $N = 25$ para uma melhor visualização.

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

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

onde μ é a média da distribuição, σ o desvio padrão e x a variável aleatória.

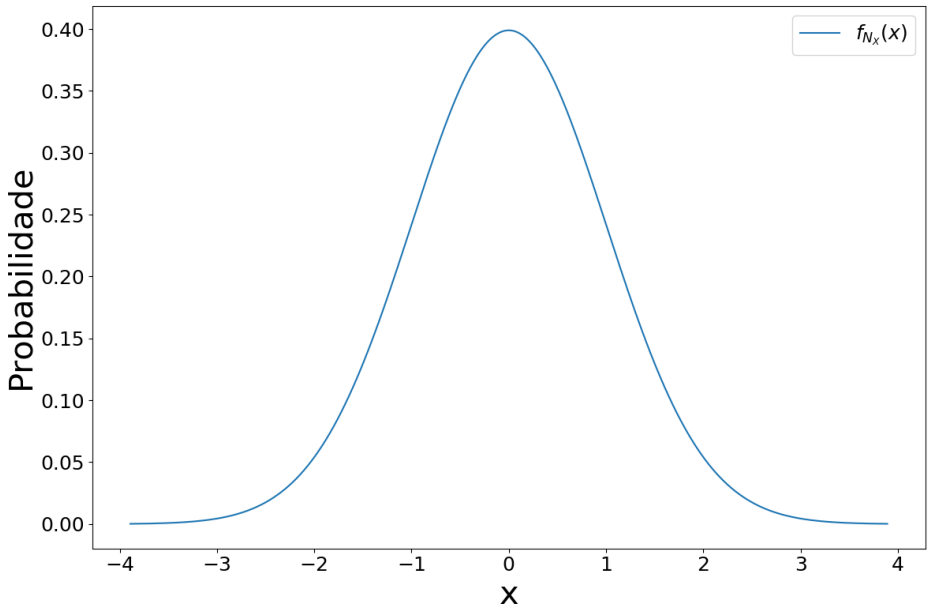


Figura 2: Ilustração da curva Gaussiana com média $\mu = 0$ e desvio padrão $\sigma = 1$.

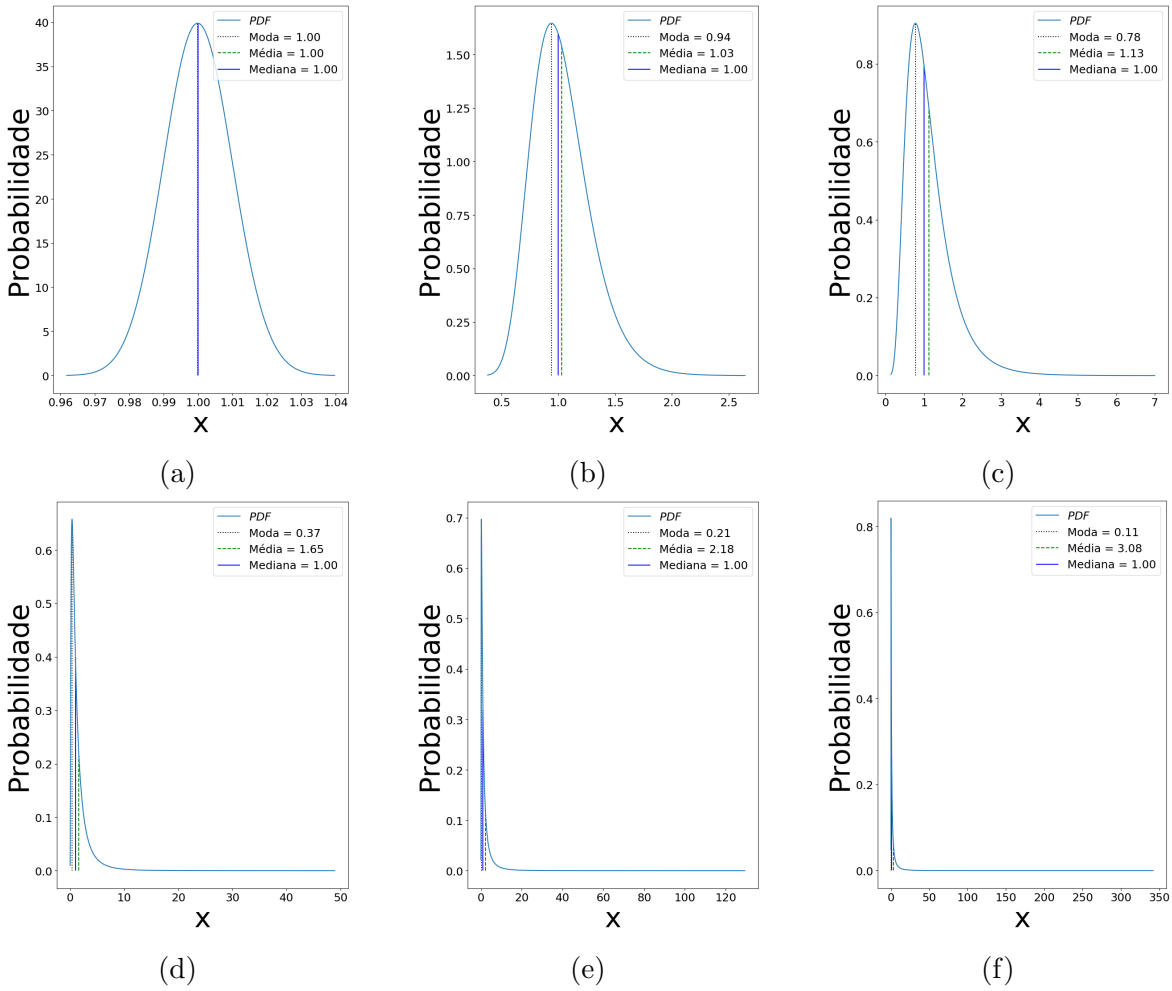


Figura 3: Ilustração das curvas Lognormais em que: (a) possui $\sigma = 0.01$; (b) possui $\sigma = 0.25$; (c) possui $\sigma = 0.5$; (d) possui $\sigma = 1$; (e) possui $\sigma = 1.25$; e (f) possui $\sigma = 1.5$

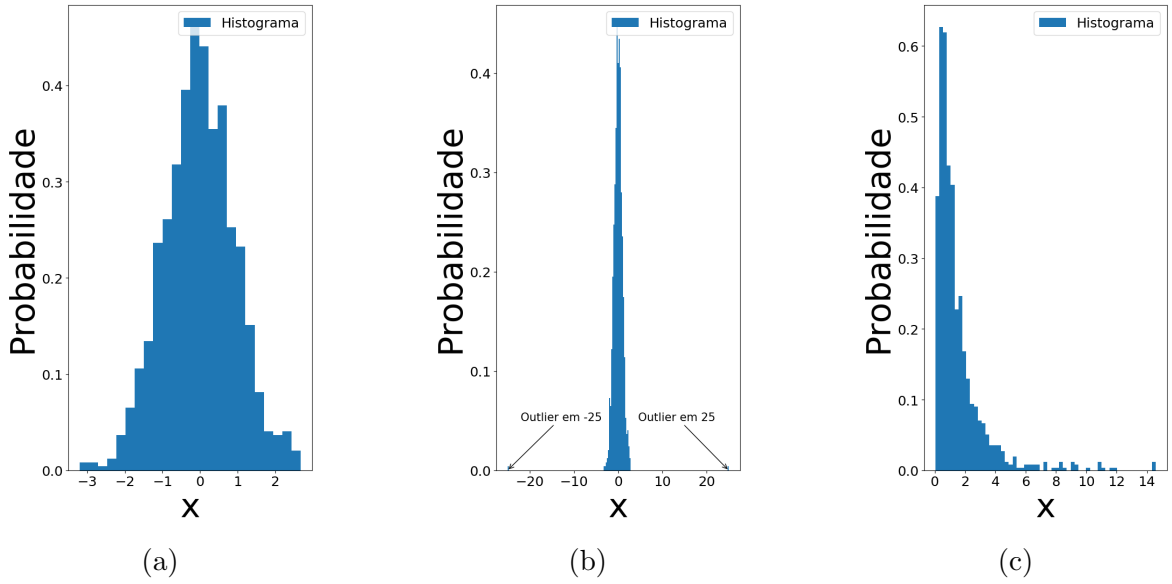


Figura 4: Histograma dos dados gerados sendo eles: (a) Gaussiana com $\mu = 0$ e $\sigma = 1$; (b) Gaussiana com $\mu = 0$, $\sigma = 1$ e *outlier* em ± 25; (c) Lognormal com $\mu = 0$ e $\sigma = 0.5$.

3.1.1 LINSPACE

O método *Linspace* é caracterizado por amostrar de maneira uniforme a variável aleatória, representada pelo eixo das abscissas de uma PDF dada. Após, o eixo horizontal 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 5 ilustra o método *Linspace* para a distribuição Normal e a Figura 7 ilustra o mesmo método para uma distribuição Lognormal com diferentes desvios padrões, limitando o eixo horizontal à uma área de probabilidade de 99.99%. A Figura 6 mostra as distribuições geradas, conforme é mostrada na Figura 4, discretizadas utilizando este método.

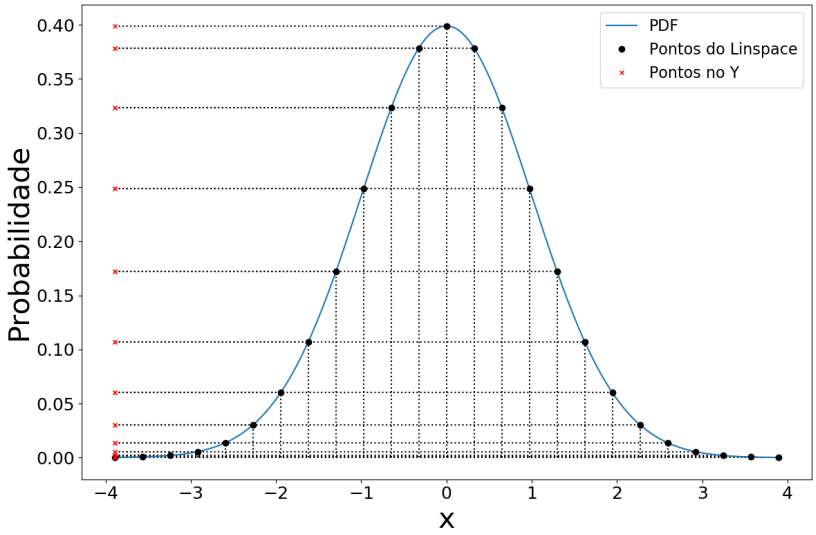
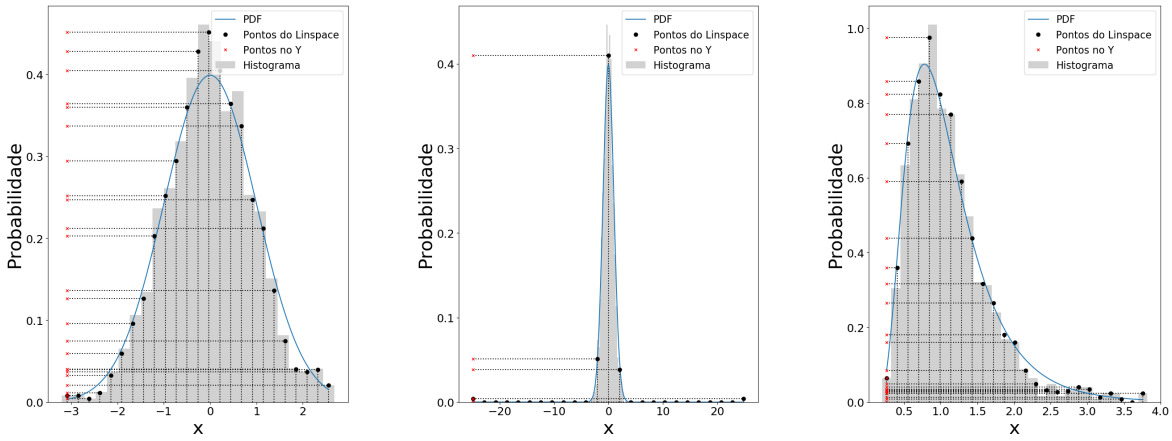
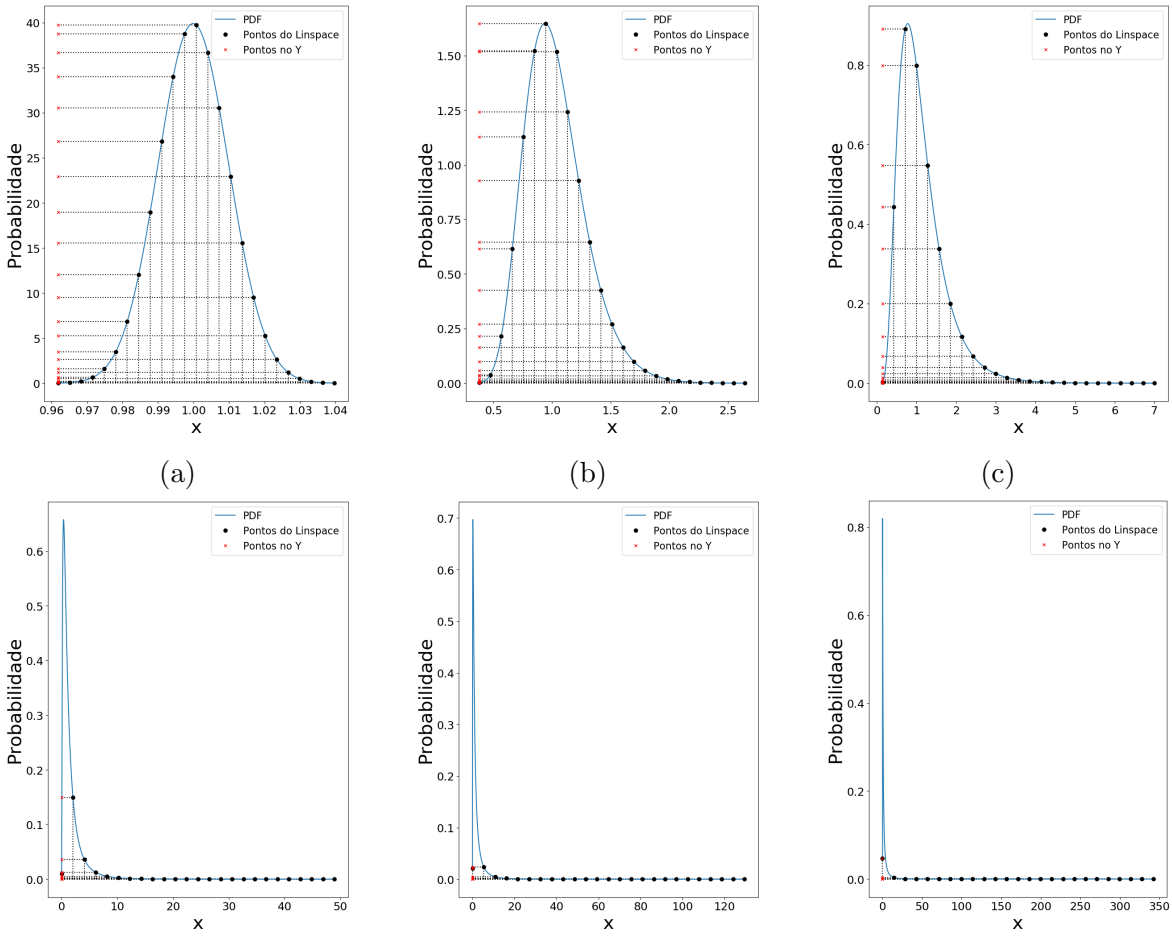


Figura 5: Ilustração do método Linspace aplicado à uma distribuição normal.



(a) (b) (c)

Figura 6: Histograma dos dados gerados utilizando a discretização pelo método *Linspace* 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 $\sigma = 0.5$.



(a) (b) (c)

(d) (e) (f)

Figura 7: Ilustração do método Linspace aplicado à uma distribuição lognormal em que: (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 $\sigma = 1.5$

É possível perceber que este método atende de forma satisfatória distribuições que não possuem derivadas muito altas como é ilustrado na figura 5 e nas figuras 7a à 7c, embora nas Figuras 6a e 6c há um erro de estimação por devida à quantidade de eventos simulados. Nas figuras 7d à 7f e 6b o método em questão já não consegue descrever a curva, colocando um número insuficientes de pontos na região de alta probabilidade e um número superior de pontos na região de alta probabilidade.

3.1.2 CDFM

Ela representa a discretização baseada na Função de Distribuição Cumulativa (do inglês, *Cumulative Distribution Function*) (CDF). Para este método, a discretização baseada no espaçamento uniforme é aplicada ao eixo vertical e então os relativos valores horizontais são encontrados refletindo todos os valores, como mostra a Figura 9 para a distribuição Normal, a Figura 8 para a distribuição Lognormal e a Figura XXXXX para o *dataset*. Note que, quanto maior a probabilidade da função de densidade, maior o número de pontos na sua região e que os *outliers* não fazem mais tanto efeito, como é o cada da *Linspace*.

COLOCAR GRÁFICO 2X3

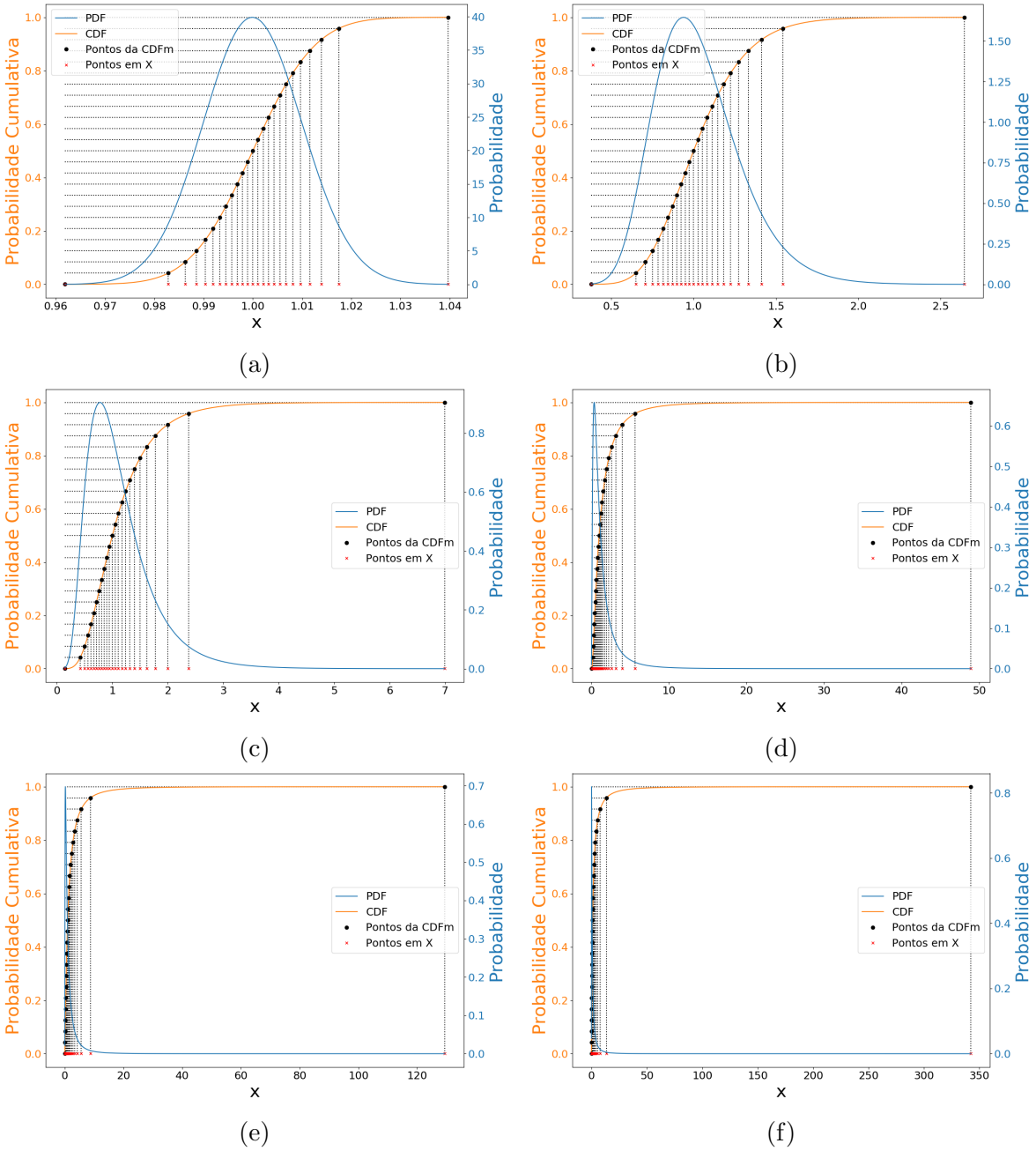


Figura 8: Ilustração do método *CDFm* aplicado à uma distribuição lognormal em que: (a) possui $\sigma = 0.01$; (b) possui $\sigma = 0.25$; (c) possui $\sigma = 0.5$; (d) possui $\sigma = 1$; (e) possui $\sigma = 1.25$; e (f) possui $\sigma = 1.5$

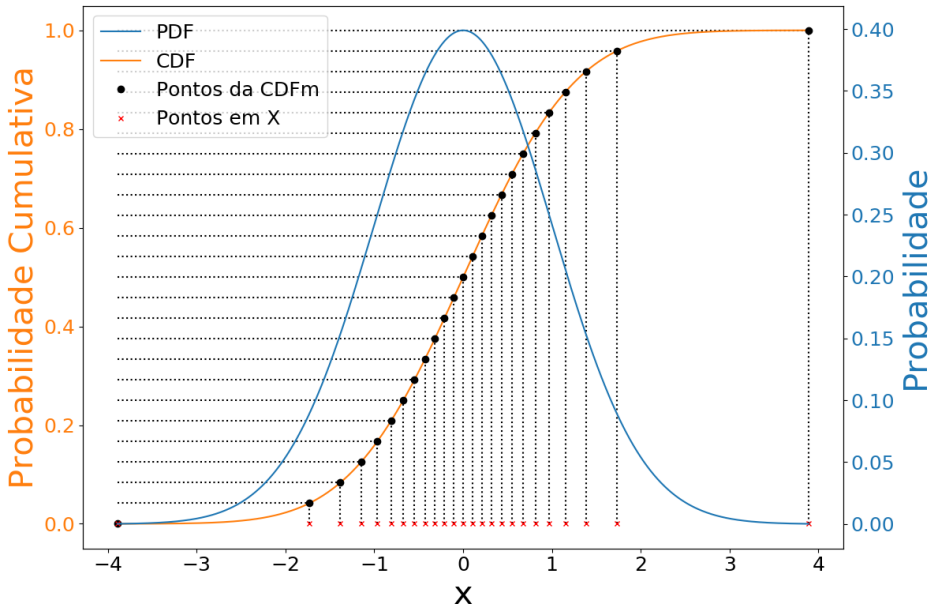


Figura 9: Ilustração da discretização da distribuição Gaussiana baseada em sua CDF.

COLOCAR AQUI A FIGURA COM DATASET P/ CDFM

Como este método baseia-se na CDF, quando maior a variação na PDF, mais rápido a CDF sobe, fazendo assim com que este método coloque mais pontos nas regiões de alta probabilidade e poucos pontos nas regiões de baixa probabilidade, sendo assim um método imune à *outliers*.

3.1.3 PDFM

Este método também usa a técnica de reflexão aplicada ao método da *CDFm*, mas função de referência a própria PDF, ao invés da sua CDF. A Figura 10 mostra como este método funciona para a distribuição Normal e a Figura 11 para a distribuição Lognormal. Ela possui o efeito de incrementar o número de pontos onde a primeira derivada da PDF é maior.

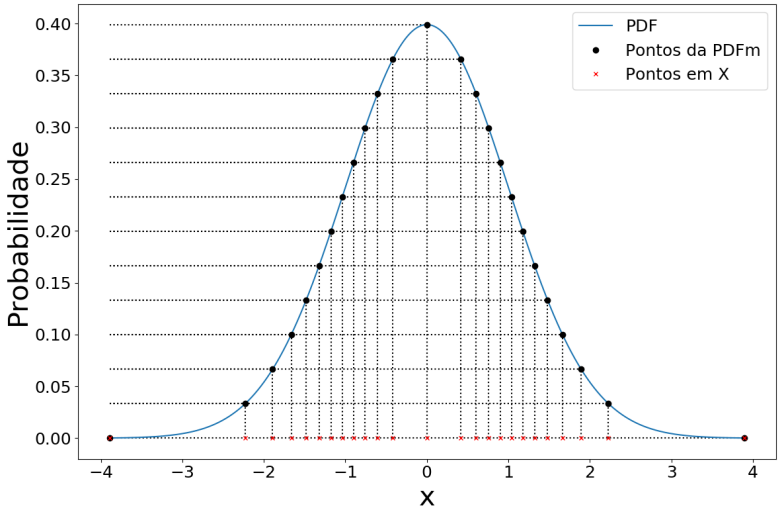


Figura 10: Ilustração da discretização da distribuição Gaussiana baseada em sua PDF.

COLOCAR GRÁFICO 2X3

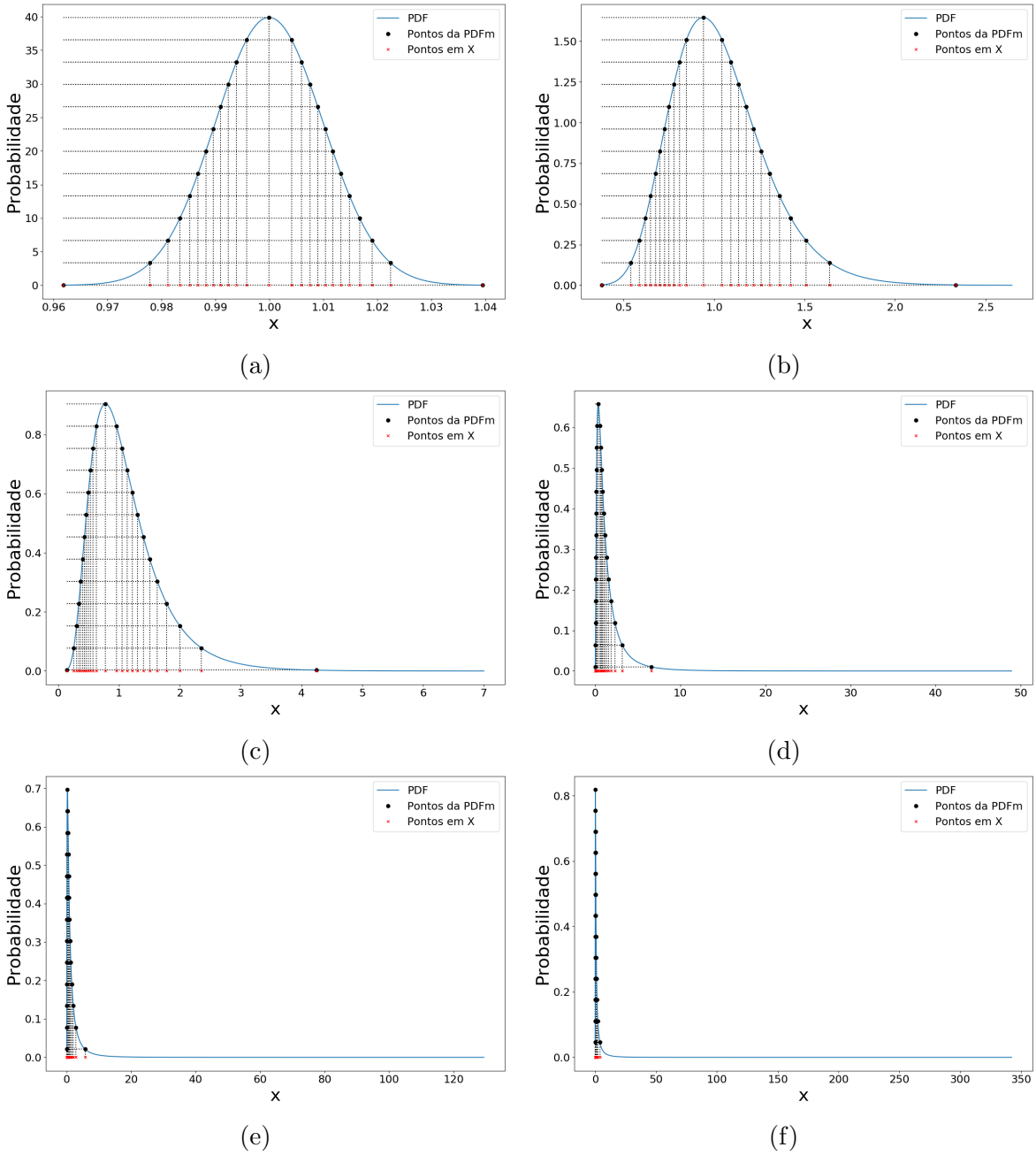


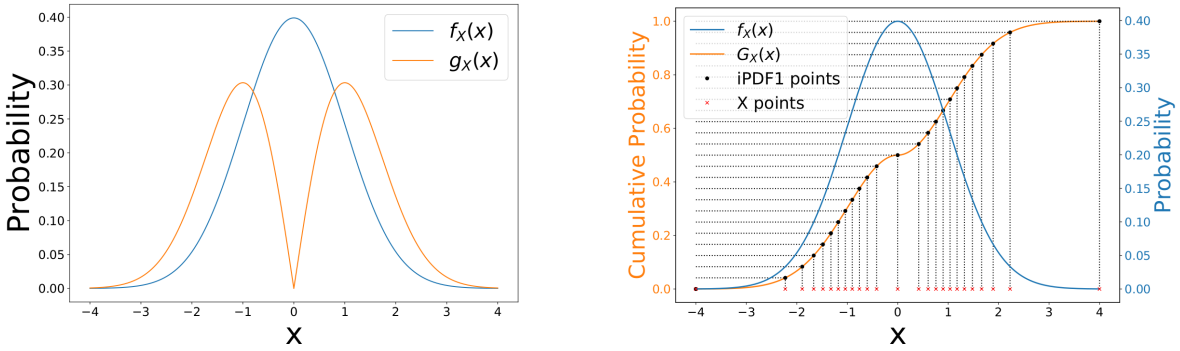
Figura 11: Ilustração do método *PDFm* aplicado à uma distribuição Lognormal em que: (a) possui $\sigma = 0.01$; (b) possui $\sigma = 0.25$; (c) possui $\sigma = 0.5$; (d) possui $\sigma = 1$; (e) possui $\sigma = 1.25$; e (f) possui $\sigma = 1.5$

COLOCAR GRÁFICO DOS DATASETS

3.1.4 IPDF1

Este método 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 12a e 12b COLOCAR O GRÁFICO DA DERIVADA EM ANEXO E AQUI O

GRÁFICO DA CDF PARA A LOGNORMAL E COM OS DATASETS



(a) (b)

Figura 12: 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 (3.3) e (3.4) descrevem este método matematicamente.

COLOCAR AQUI TB A EQUAÇÃO PRA LOGNORMAL

$$\begin{aligned} \zeta(x) &= \frac{|\mu - x|}{\sigma^3\sqrt{2\pi}} \cdot e^{\left(\frac{-(\mu-x)^2}{2\sigma^2}\right)} \\ \int_{-\infty}^{\infty} \zeta(x) \cdot dx &= c_1 \\ g_X(x) &= \frac{\zeta(x)}{c_1} \end{aligned} \tag{3.3}$$

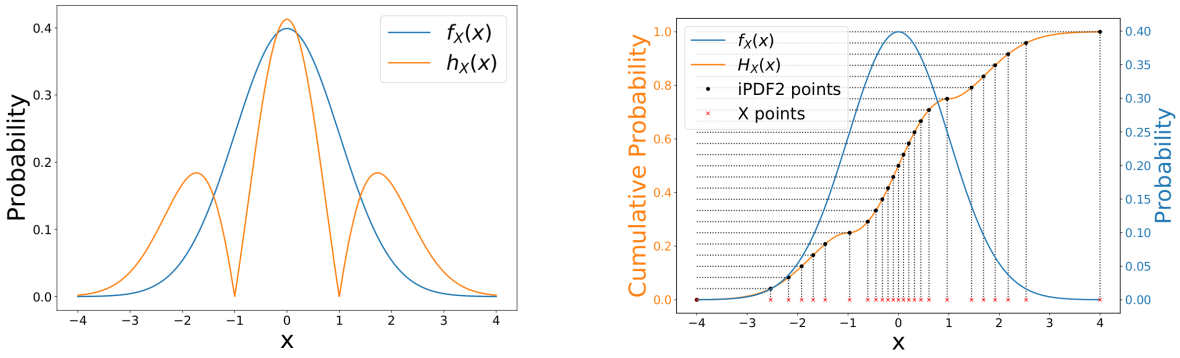
onde ζ é a equação da distribuição da derivada da distribuição normal, μ é a média, σ o desvio padrão, x a variável aleatória, c_1 é a área abaixo da curva ζ, e g_X é a versão normalizada. A CDF de g_X ($G_X(x)$) é usada para transferir os valores da abscissa ao eixo da ordenada como mostra a Figura 12b.

$$G_X(x) = \int_{-\infty}^{x} g_X(y) \cdot dy \tag{3.4}$$

3.1.5 IPDF2

Este método é construído da mesma maneira da *IPDF1* mas usando a segunda detivada ao invés da primeira, como é mostrado na Figura 13a e 13b. Suas equações são mostradas em (3.5) e (3.6).

COLOCAR O GRÁFICO DA DERIVADA EM ANEXO E AQUI O GRÁFICO DA CDF PARA A LOGNORMAL E COM OS DATASETS



(a) (b)

Figura 13: PDF Gaussiana e sua segunda derivada à esquerda. Ilustração da discretização da distribuição Gaussiana baseada na CDF da sua segunda derivada à direita.

COLOCAR AQUI TB A EQUAÇÃO PRA LOGNORMAL

$$\eta(x) = \frac{|\sigma^2 - (\mu - x)^2|}{\sigma^5\sqrt{2\pi}} \cdot e^{-\frac{(\mu-x)^2}{2\sigma^2}}$$

$$\int_{-\infty}^{\infty} \eta(x) \cdot dx = c_2 \tag{3.5}$$

$$h_X(x) = \frac{\eta(x)}{c_2}$$

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 (3.6).

$$H_X(x) = \int_{-\infty}^{x} h_X(y) \cdot dy \tag{3.6}$$

3.2 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 Regiões de Interesse (do inglês, *Regions of Interest*) (RoI) (RON, 1999). Essas regiões são compreendidas entre valores máximos e mínimos predefinidos do eixo horizontal. A Figura XXXXX mostra este processo quando a abscissa é dividida em 20 regiões, todas compreendidas entre os valores -4 e 4 do eixo x.

COLOCAR AQUI O GRÁFICO COM O ERRO QUE TEM ZOOM

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 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

Conforme apresentado na Seção 3,

5 CONCLUSÃO

REFERÊNCIAS

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>.

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.

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.

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

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.