



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, familiares e amigos e ao Amaro.

AGRADECIMENTOS

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

Anaïs Nin

RESUMO

Palavras-chave:

ABSTRACT

Keywords:

LISTA DE ILUSTRAÇÕES

Figura 1	Caso representativo de estimação de PDF utilizando 25 pontos para dois intervalos diferentes $(-4,4)$ e $(-10,10)$	21
Figura 2	Ilustração da curva Gaussiana com média $\mu = 0$ e desvio padrão $\sigma = 1$	22
Figura 3	Ilustração do método <i>Linspace</i> aplicado à uma distribuição normal.	23
Figura 4	Ilustração da discretização da distribuição Gaussiana baseada em sua CDF.	24
Figura 5	Ilustração da discretização da distribuição Gaussiana baseada em sua PDF.	24
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.	25
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.	26
Figura 8	Ilustração das curvas Lognormais construídas com diferentes parâmetros: (a) possui $\sigma = 0.01$; (b) possui $\sigma = 0.25$; (c) possui $\sigma = 0.5$; (d) possui $\sigma = 1$; (e) possui $\sigma = 1.25$; e (f) possui $\sigma = 1.5$	28
Figura 9	Histograma dos dados gerados sendo eles: (a) Gaussiana com $\mu = 0$ e σ	

	= 1; (b) Gaussiana com $\mu = 0$, $\sigma = 1$ e <i>outlier</i> em ± 25 ; (c) Lognormal com $\mu = 0$ e $\sigma = 0.5$	29
Figura 10	Ilustração do método Linspace aplicado à uma distribuição normal.	30
Figura 11	Histograma dos dados gerados utilizando a discretização pelo método <i>Linspace</i> sendo eles: (a) Gaussiana com $\mu = 0$ e $\sigma = 1$; (b) Gaussiana com $\mu = 0$, $\sigma = 1$ e <i>outlier</i> em ± 25 ; (c) Lognormal com $\mu = 0$ e $\sigma = 0.5$	30
Figura 12	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 $\sigma = 1.5$	31
Figura 13	Ilustração do método <i>CDFm</i> 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 $\sigma = 1.5$	32
Figura 14	Ilustração da discretização da distribuição Gaussiana baseada em sua CDF.	33
Figura 15	Ilustração da discretização da distribuição Gaussiana baseada em sua PDF.	34
Figura 16	Ilustração do método <i>PDFm</i> 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 $\sigma = 1.5$	35
Figura 17	PDF Gaussiana e sua primeira derivada à esquerda. Ilustração da discretização da distribuição Gaussiana baseada na CDF da sua primeira derivada à direita.	36
Figura 18	PDF Gaussiana e sua segunda derivada à esquerda. Ilustração da dis-	

cretização da distribuição Gaussiana baseada na CDF da sua segunda derivada à direita.	37
--	----

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

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

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

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

FD Estimador Freedman–Diaconi

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*)

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

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

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

SUMÁRIO

1	INTRODUÇÃO	17
1.1	Motivação	18
1.2	Estrutura do Trabalho	19
2	DISCRETIZAÇÃO	20
2.1	Métodos de Discretização	21
2.1.1	<i>Linspace</i>	22
2.1.2	<i>CDFm</i>	23
2.1.3	<i>PDFm</i>	24
2.1.4	<i>iPDF1</i>	24
2.1.5	<i>iPDF2</i>	25
3	DESENVOLVIMENTO	27
3.1	Métodos de discretização	27
3.1.1	<i>Linspace</i>	29
3.1.2	<i>CDFm</i>	31
3.1.3	<i>PDFm</i>	33
3.1.4	<i>iPDF1</i>	35
3.1.5	<i>iPDF2</i>	37
3.2	Ambiente de Análise	38
4	Resultados	39
5	Conclusão	40

Referências	41
Apêndice A – Lista de Publicações	42
A.1 Publicações em Anais de Congresso Internacional	42

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 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, esse fato é explicado devido aos eventos de interesse serem raros e contaminados com alto nível de ruído de fundo, demandando sistemas cada vez mais otimizados no que diz respeito a tempo de processamento, eficiência de

detecção de sinal e rejeição de ruído.

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

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

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

1.1 MOTIVAÇÃO

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 pré-processamento do conjunto de dados.

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

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

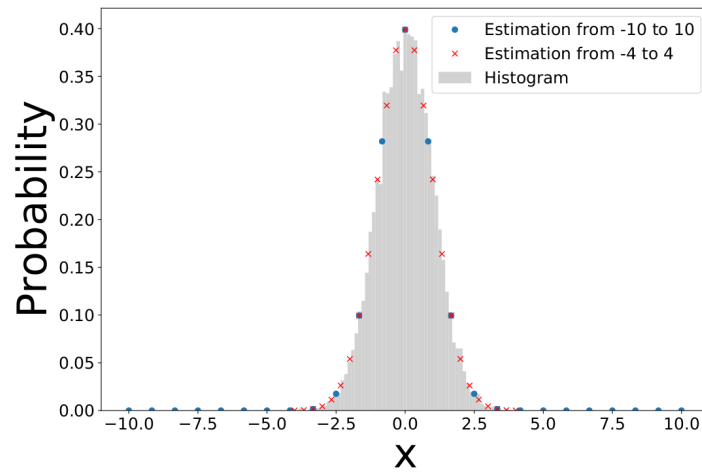


Figura 1: Caso representativo de estimação de PDF utilizando 25 pontos para dois intervalos diferentes $(-4,4)$ e $(-10,10)$

2.1 MÉTODOS DE DISCRETIZAÇÃO

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

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

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

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

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

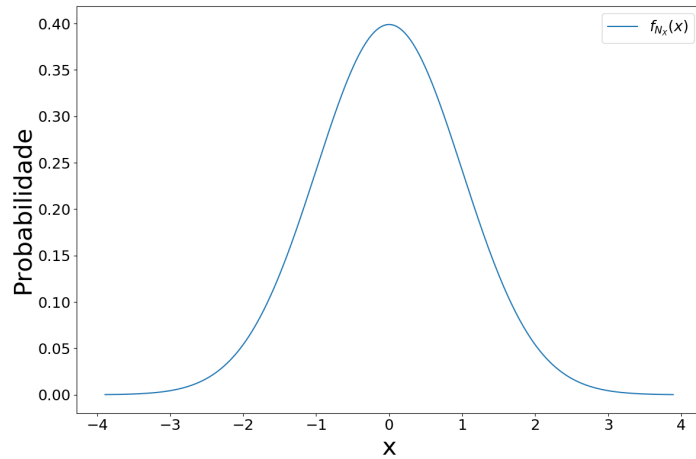


Figura 2: Ilustração da curva Gaussiana com média $\mu = 0$ e desvio padrão $\sigma = 1$.

2.1.1 *Linspace*

O método *Linspace* é caracterizado por amostrar de maneira uniforme a variável aleatória, representada pelo eixo das abscissas de uma PDF. 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 10 ilustra o método *Linspace* para a distribuição Normal, limitando o eixo horizontal à uma área de probabilidade de 99,99%.

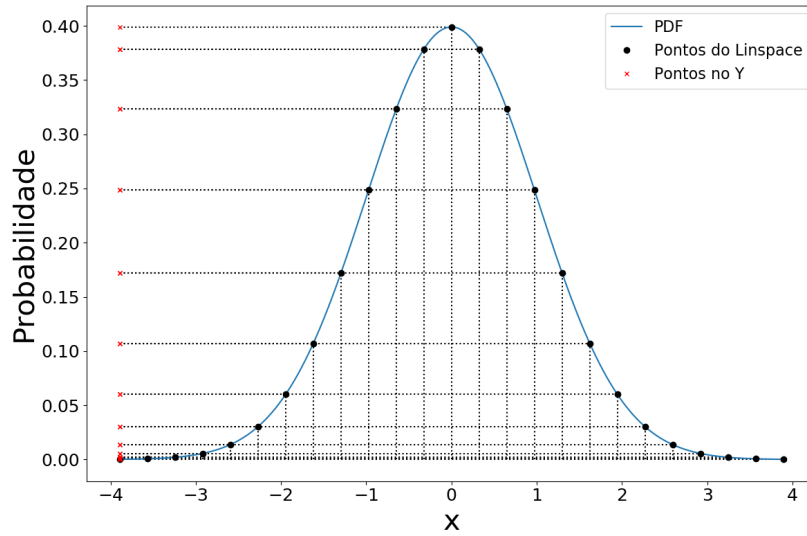


Figura 3: Ilustração do método *Linspace* aplicado à uma distribuição normal.

2.1.2 CDFM

O método denominado nesse trabalho de Método CDF (do inglês, *CDF method*) (CDFm) representa a discretização baseada na Função de Distribuição Cumulativa (do inglês, *Cumulative Distribution Function*) (CDF) descrita pela Equação (2.2) para uma variável contínua e (2.3) para o caso de uma variável discreta possuindo valores em b . Para este método, no caso de se ter a função geradora, primeiramente calcula-se a CDF da distribuição e então faz-se uma distribuição linear de pontos no eixo y e encontra-se os seus respectivos valores para o eixo x , conforme mostra a Figura 14. Note que, quanto maior a probabilidade da CDF, maior o número de pontos em sua região.

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

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

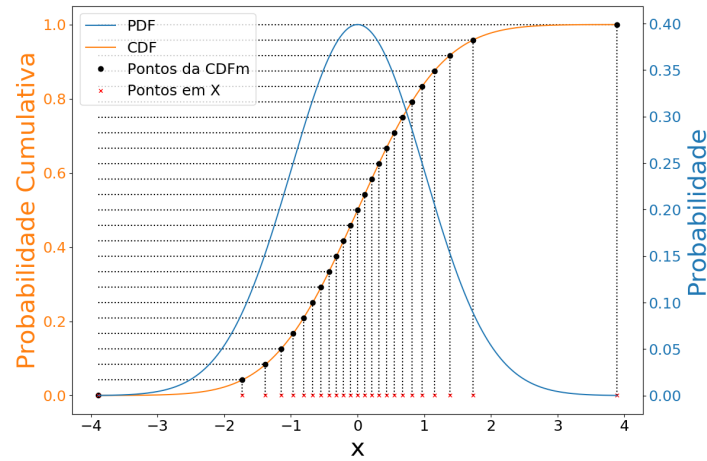


Figura 4: Ilustração da discretização da distribuição Gaussiana baseada em sua CDF.

2.1.3 PDFM

Este método, denominado de Método PDF (do inglês, *PDF method*) (PDFm) também usa a técnica de reflexão aplicada ao método da *CDFm*, mas a função de referência é a própria PDF, ao invés da sua CDF. A Figura 15 mostra como este método funciona.

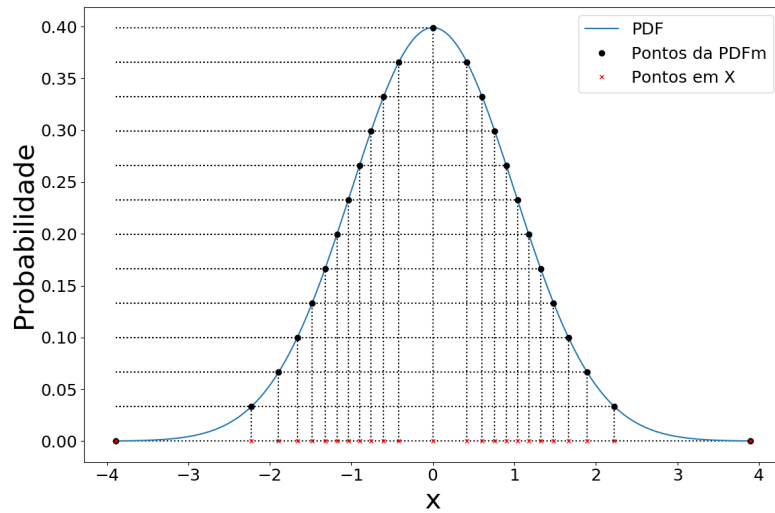


Figura 5: Ilustração da discretização da distribuição Gaussiana baseada em sua PDF.

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

2.1.4 IPDF1

O método da Integral da distribuição da primeira derivada da PDF (iPDF1) reflete os valores verticais para o eixo horizontal usando a CDF da primeira derivada da PDF

como uma transformação de base, como é ilustrado nas Figuras 17a e 17b

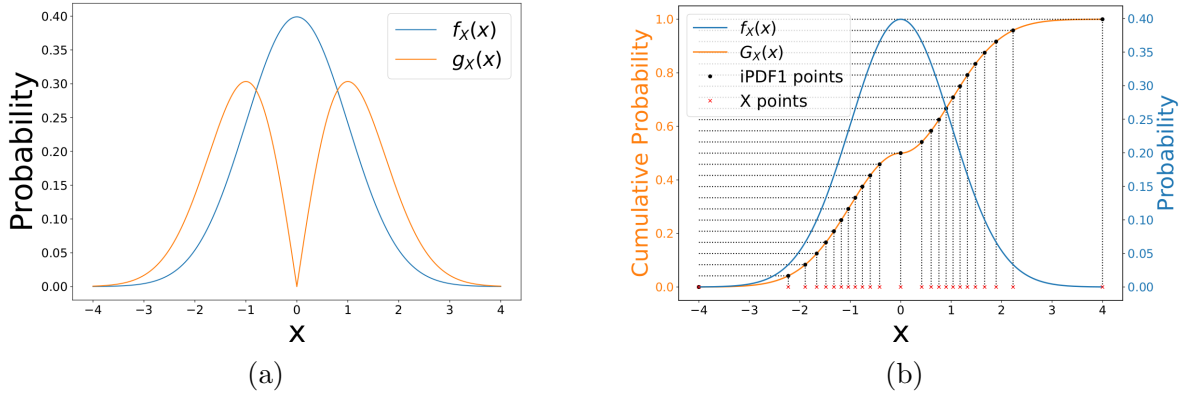


Figura 6: PDF Gaussiana e sua primeira derivada à esquerda. Ilustração da discretização da distribuição Gaussiana baseada na CDF da sua primeira derivada à direita.

As equações (3.3) e (2.5) descrevem este método matematicamente.

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

$$\int_{-\infty}^{\infty} \zeta(x) \cdot dx = c_1 \quad (2.4)$$

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

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 17b.

$$G_X(x) = \int_{-\infty}^x g_X(y) \cdot dy \quad (2.5)$$

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 18a e 18b. Suas equações são mostradas em (3.5) e (3.6).

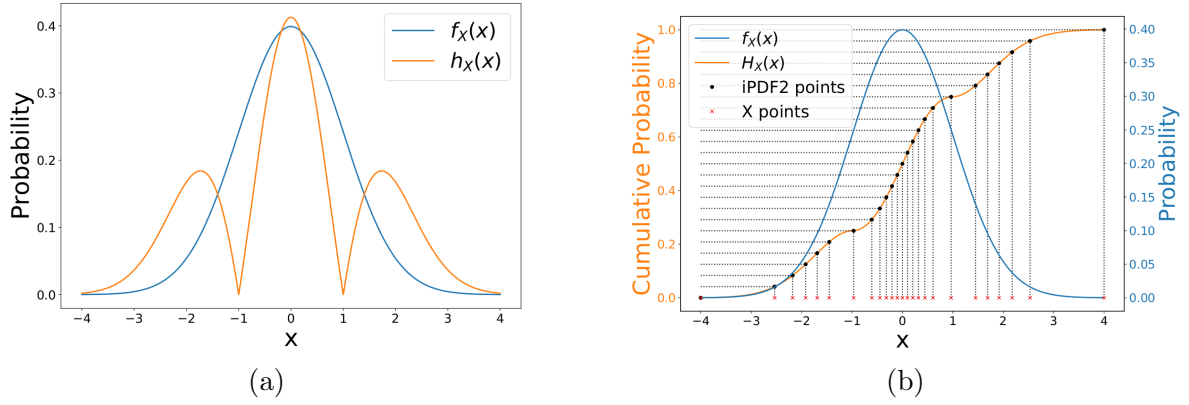


Figura 7: PDF Gaussiana e sua segunda derivada à esquerda. Ilustração da discretização da distribuição Gaussiana baseada na CDF da sua segunda derivada à direita.

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

$$\int_{-\infty}^{\infty} \eta(x) \cdot dx = c_2 \quad (2.6)$$

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

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).

$$H_X(x) = \int_{-\infty}^x h_X(y) \cdot dy \quad (2.7)$$

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, 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 Gaussianas, 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 8 e com três *Datasets* diferentes, o primeiro sendo de uma distribuição normal com média nula e desvio padrão unitário representado na Figura 9a, 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 9b e, por fim, a ultima será uma distribuição Lognormal com média nula e desvio padrão de 0.5, ilustrado pela Figura 9c. 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.

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

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

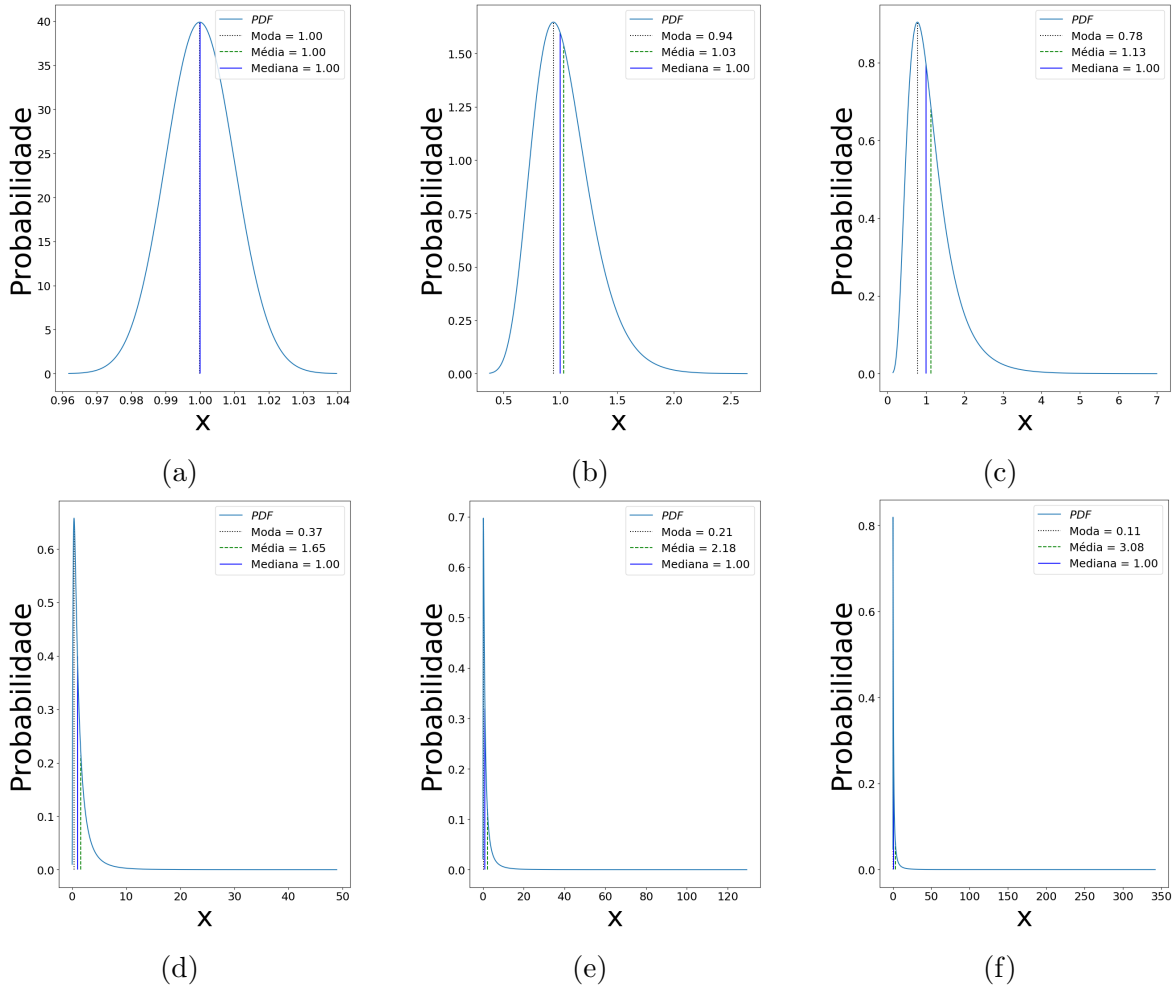


Figura 8: Ilustração das curvas Lognormais construídas com diferentes parâmetros: (a) possui $\sigma = 0.01$; (b) possui $\sigma = 0.25$; (c) possui $\sigma = 0.5$; (d) possui $\sigma = 1$; (e) possui $\sigma = 1.25$; e (f) possui $\sigma = 1.5$

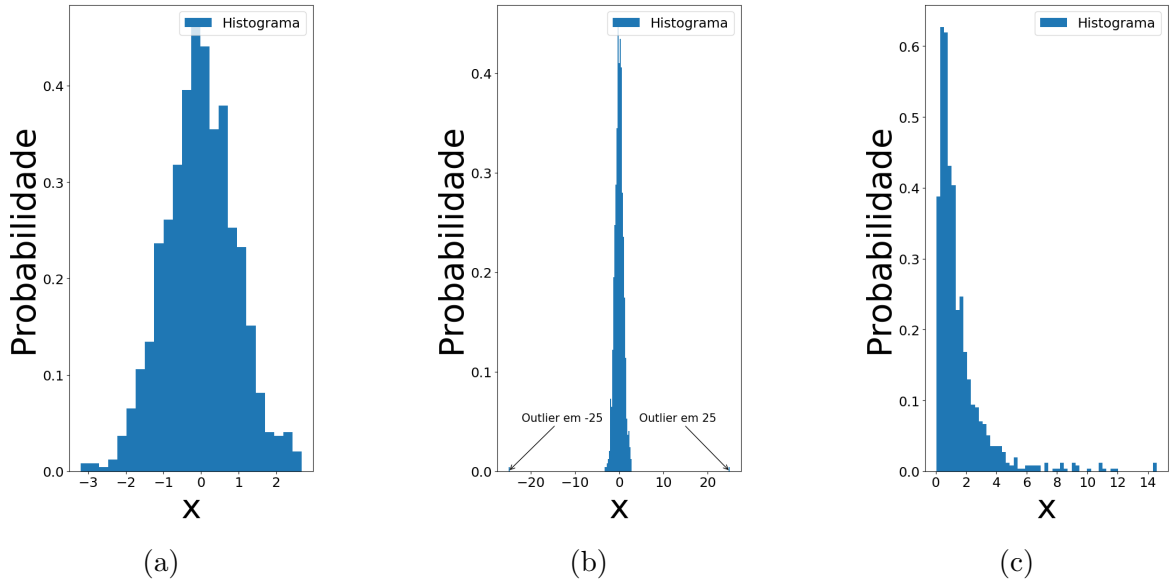


Figura 9: Histograma dos dados gerados sendo eles: (a) Gaussiana com $\mu = 0$ e $\sigma = 1$; (b) Gaussiana com $\mu = 0$, $\sigma = 1$ e *outlier* em ± 25 ; (c) Lognormal com $\mu = 0$ e $\sigma = 0.5$.

3.1.1 LINSPEACE

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 10 ilustra o método *Linspace* para a distribuição Normal e a Figura 12 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 11 mostra as distribuições geradas, conforme é mostrado na Figura 9, discretizadas utilizando este método.

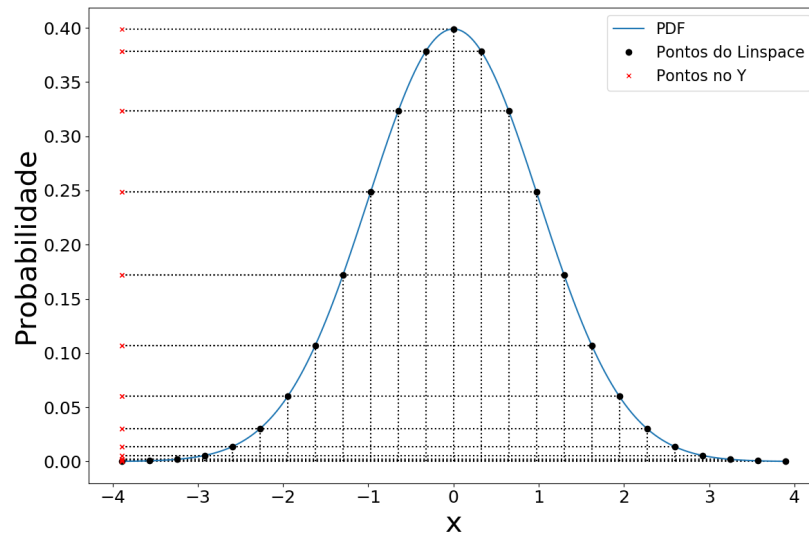


Figura 10: Ilustração do método Linspace aplicado à uma distribuição normal.

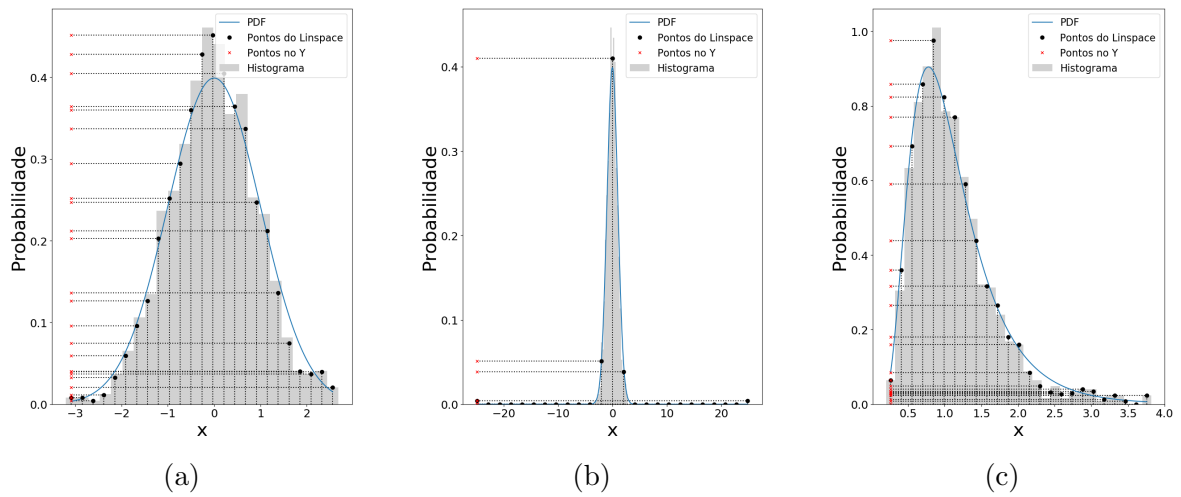


Figura 11: Histograma dos dados gerados utilizando a discretização pelo método *Linspace* sendo eles: (a) Gaussiana com $\mu = 0$ e $\sigma = 1$; (b) Gaussiana com $\mu = 0$, $\sigma = 1$ e *outlier* em ± 25 ; (c) Lognormal com $\mu = 0$ e $\sigma = 0.5$.

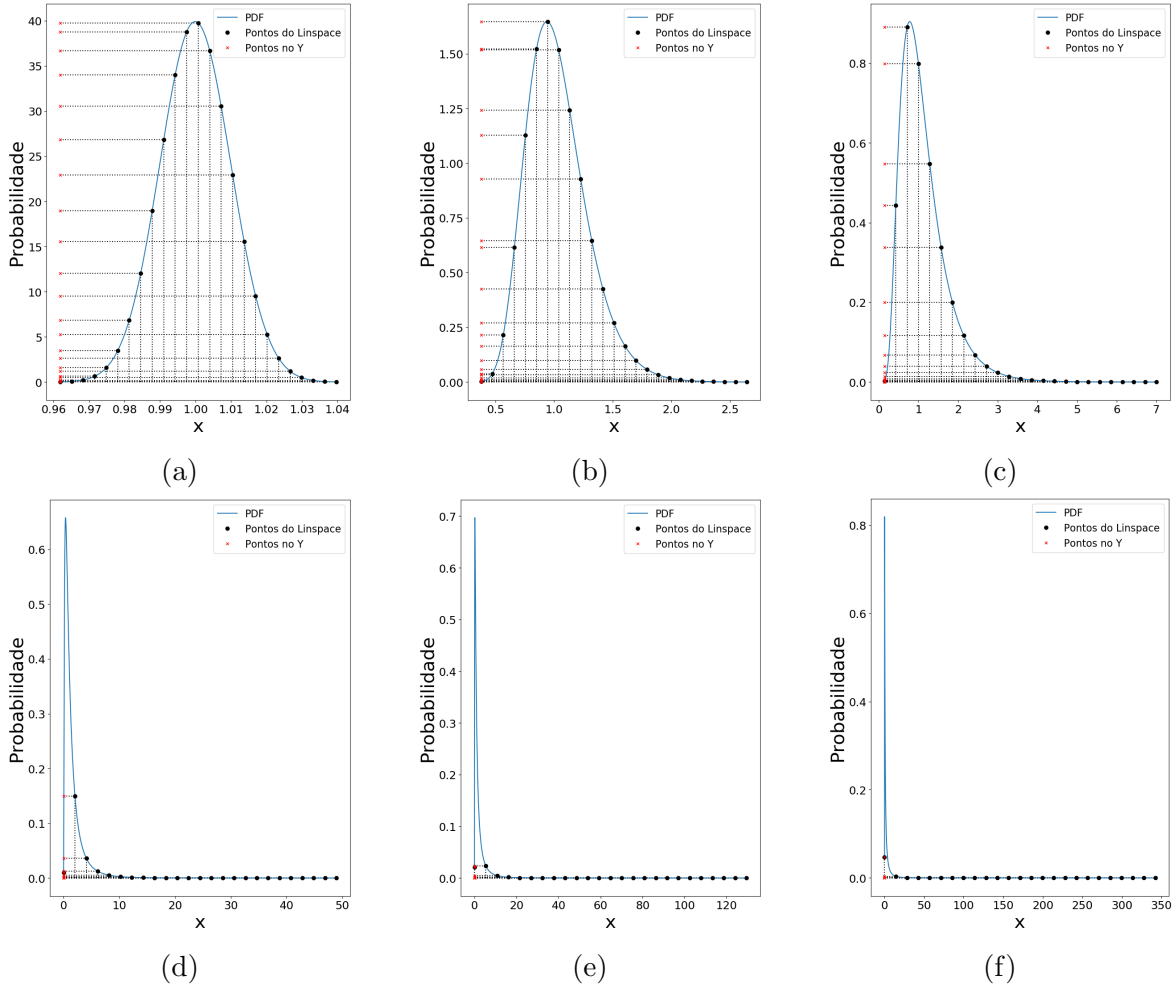


Figura 12: Ilustração do método Linspace aplicado à uma distribuição lognormal em que: (a) possui $\sigma = 0.01$; (b) possui $\sigma = 0.25$; (c) possui $\sigma = 0.5$; (d) possui $\sigma = 1$; (e) possui $\sigma = 1.25$; e (f) possui $\sigma = 1.5$

É possível perceber que este método atende de forma satisfatória distribuições que não possuem derivadas muito altas como é ilustrado na figura 10 e nas figuras 12a à 12c, embora nas Figuras 11a e 11c há um erro de estimação por devida à quantidade de eventos simulados. Nas figuras 12d à 12f e 11b o método em questão já não consegue descrever a curva, colocando um número insuficientes de pontos na região de alta probabilidade e um número maior de pontos na região de alta probabilidade.

3.1.2 CDFM

O método denominado nesse trabalho de CDFm representa a discretização baseada na 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 14 para a distribuição Normal, a Figura 13 para

a distribuição Lognormal e a Figura XXXXX para o *dataset*. Note que, quanto maior a probabilidade da PDF, maior o número de pontos na sua região e que os *outliers* não fazem mais tanto efeito, como é o caso do *Linspace*.

COLOCAR GRÁFICO 2X3

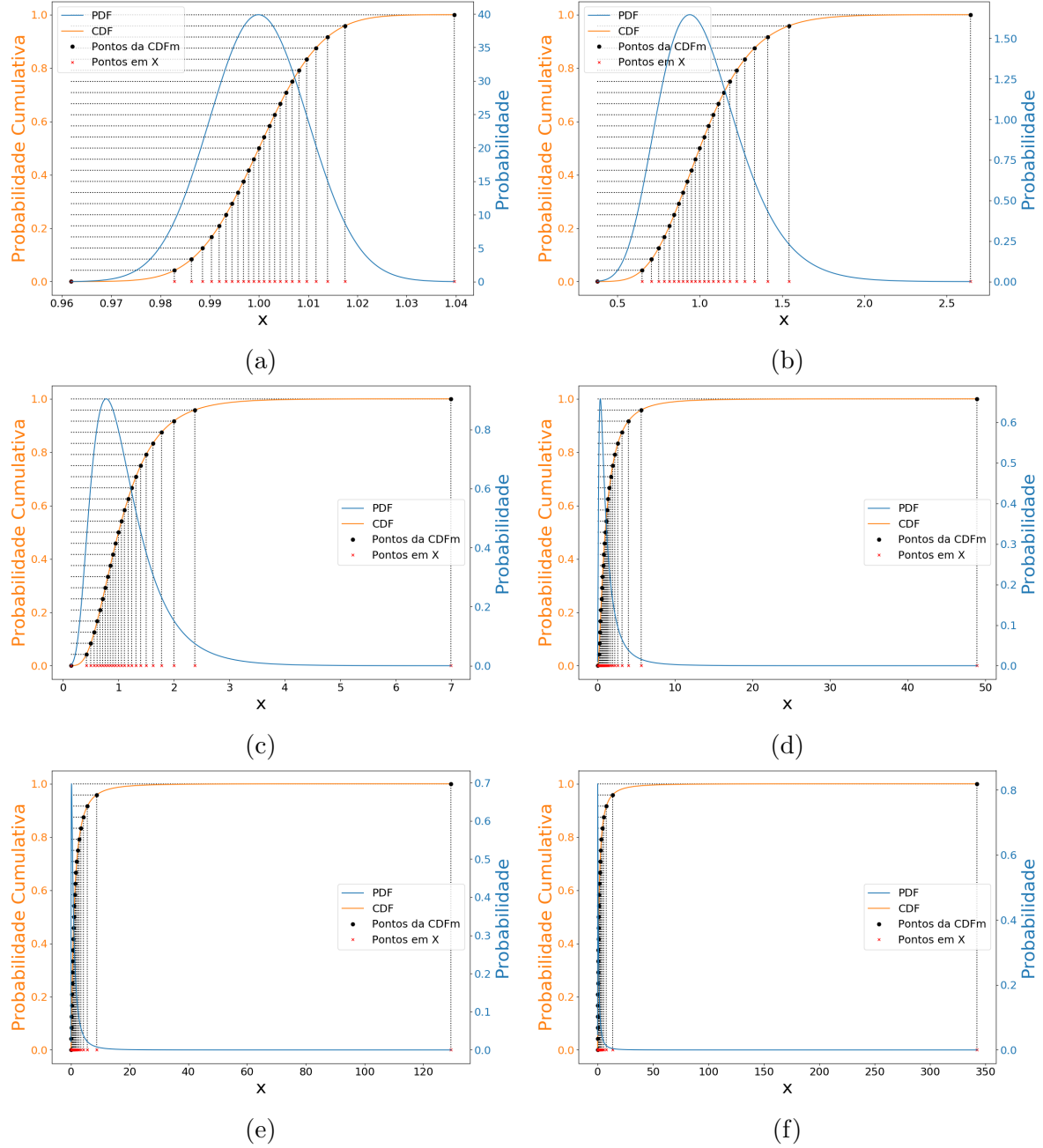


Figura 13: Ilustração do método *CDFm* aplicado à uma distribuição lognormal em que: (a) possui $\sigma = 0.01$; (b) possui $\sigma = 0.25$; (c) possui $\sigma = 0.5$; (d) possui $\sigma = 1$; (e) possui $\sigma = 1.25$; e (f) possui $\sigma = 1.5$

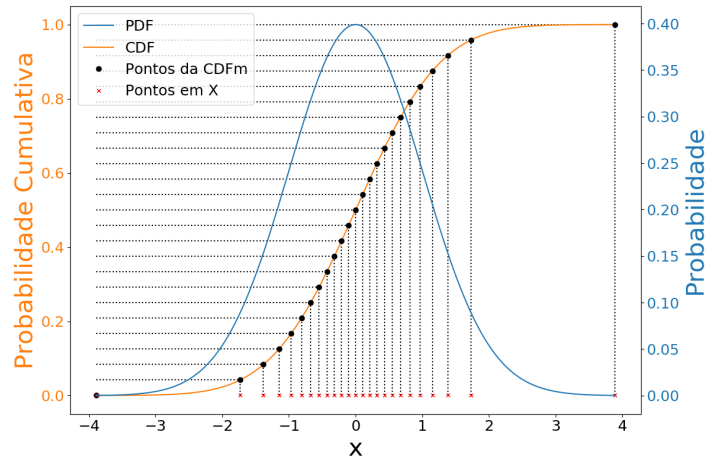


Figura 14: Ilustração da discretização da distribuição Gaussiana baseada em sua CDF.

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, denominado de PDFm 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 15 mostra como este método funciona para a distribuição Normal, a Figura 16 para a distribuição Lognormal e a Figura XXXX para dados gerados. Ela possui o efeito de incrementar o número de pontos onde a primeira derivada da PDF é maior, ou seja, quanto maior a inclinação da curva, mais pontos são colocados.

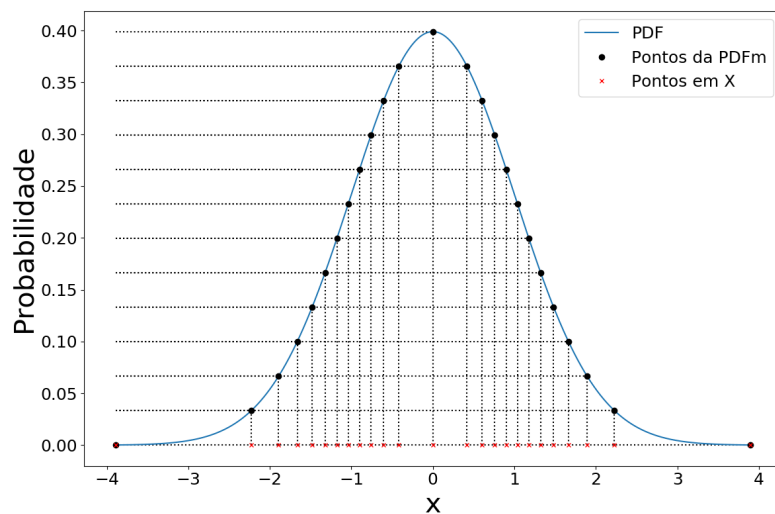


Figura 15: Ilustração da discretização da distribuição Gaussiana baseada em sua PDF.

COLOCAR GRÁFICO 2X3

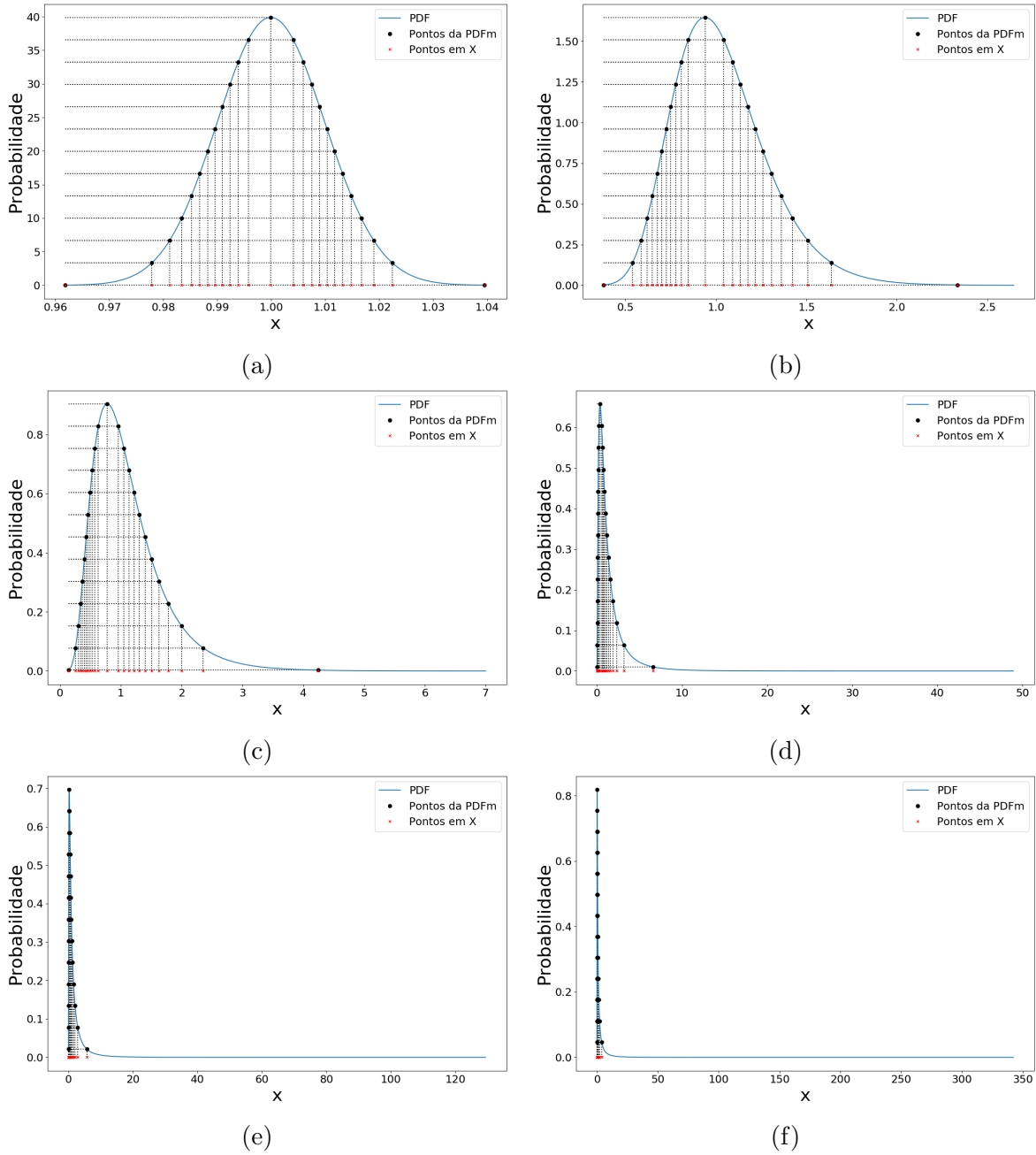


Figura 16: Ilustração do método *PDFm* aplicado à uma distribuição Lognormal em que: (a) possui $\sigma = 0.01$; (b) possui $\sigma = 0.25$; (c) possui $\sigma = 0.5$; (d) possui $\sigma = 1$; (e) possui $\sigma = 1.25$; e (f) possui $\sigma = 1.5$

COLOCAR GRÁFICO DOS DATASETS

3.1.4 *IPDF1*

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 17a e 17b **COLOCAR O GRÁFICO DA DERIVADA EM ANEXO E AQUI O**

GRÁFICO DA CDF PARA A LOGNORMAL E COM OS DATASETS

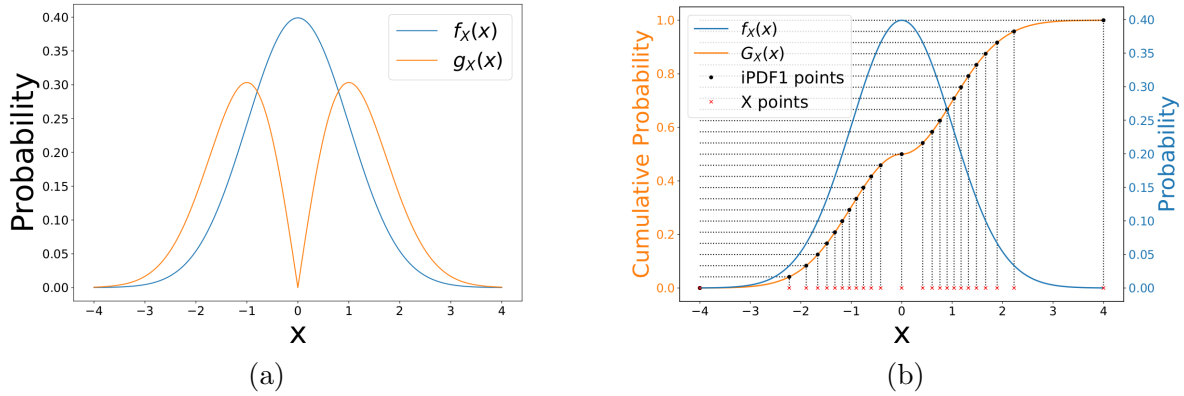


Figura 17: PDF Gaussiana e sua primeira derivada à esquerda. Ilustração da discretização da distribuição Gaussiana baseada na CDF da sua primeira derivada à direita.

As equações (3.3) e (2.5) descrevem este método matematicamente.

COLOCAR AQUI TB A EQUAÇÃO PRA LOGNORMAL

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

$$\int_{-\infty}^{\infty} \zeta(x) \cdot dx = c_1 \quad (3.3)$$

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

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 17b.

Já para a aplicação com dados gerados, ilustrado na Figura XXXX, precisamos utilizar a derivada discreta, que pode ser feita utilizando o método de diferenças finitas, que consiste em computar a inclinação de uma reta secante vizinha através dos pontos $(x, f(x))$ e $(x+h, f(x+h))$ (BURDEN; FAIRES, 2001) que pode ser vista na Figura XXXX. A inclinação dessa reta pode ser descrita pela Equação (3.4)

$$f'(x) = \frac{f(x+h) - f(x)}{h} \quad (3.4)$$

onde h é a distancia entre dois pontos vizinhos (*bins*) e x o valor do *bin*

BOTAR UMA FIGURA MANEIRA AQUI PRA EXPLICAR A DERIVADA DISCRETA

3.1.5 IPDF2

O método da Integral da distribuição da segunda derivada da PDF (iPDF2) é construído da mesma maneira da *IPDF1* mas usando a segunda derivada ao invés da primeira, como é mostrado na Figura 18a e 18b. 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

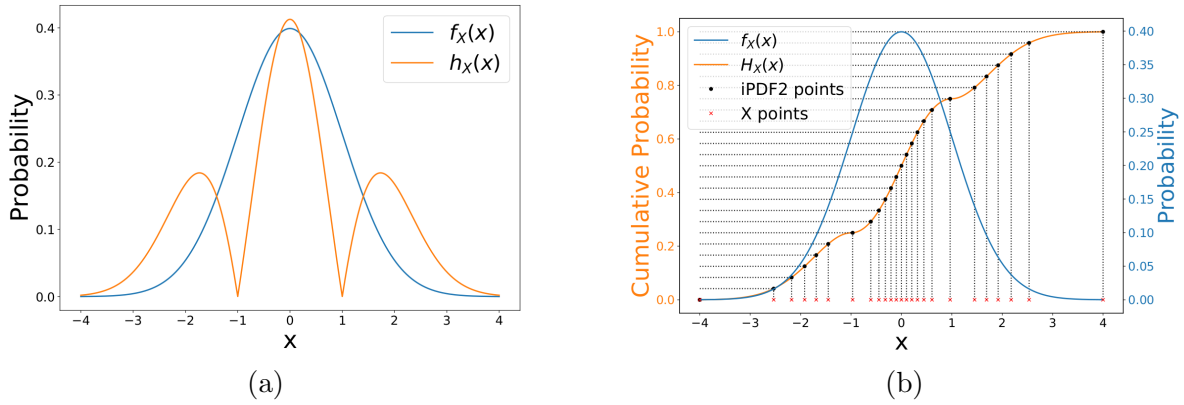


Figura 18: PDF Gaussiana e sua segunda derivada à esquerda. Ilustração da discretização da distribuição Gaussiana baseada na CDF da sua segunda derivada à direita.

COLOCAR AQUI TB A EQUAÇÃO PRA LOGNORMAL

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

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

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

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).

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

Para os dados gerados, o procedimento é o mesmo descrito na Seção 3.1.5 e a equação da segunda derivada discreta pode ser escrita pela Equação 3.7.

$$f''(x) = \frac{f'(x+h) - f'(x)}{h} \quad (3.7)$$

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.
- BURDEN, R.; FAIRES, J. *Numerical analysis 7th ed., brooks/cole, thomson learning*. [S.l.]: Inc, 2001.
- 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.