

**UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA**

David de Melo Souza

**Estatística não-paramétrica: estimação, classificação e uma nova abordagem de
seleção automática para largura de banda**

Juiz de Fora
2020

David de Melo Souza

Estatística não-paramétrica: estimação, classificação e uma nova abordagem de seleção automática para largura de banda

Tese apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Doutor em Engenharia Elétrica. Área de Concentração: Sistemas Eletrônicos.

Orientador: Rafael Antunes Nóbrega

JUIZ DE FORA

2020

Ficha catalográfica elaborada através do programa de geração automática da Biblioteca Universitária da UFJF, com os dados fornecidos pelo(a) autor(a)

Souza, David de Melo.

Estatística Não-Paramétrica: Estimação, classificação e uma nova abordagem de seleção automática para largura de banda / David de Melo Souza. -- 2020.

195 f.

Orientador: Rafael Antunes Nóbrega

Tese (doutorado) - Universidade Federal de Juiz de Fora, Faculdade de Engenharia. Programa de Pós-Graduação em Engenharia Elétrica, 2020.

1. Estimação não-paramétrica. 2. Largura de banda. 3. KDE. 4. Verossimilhança naive. 5. Classificação. I. Nóbrega, Rafael Antunes, orient. II. Título.

David de Melo Souza

Estatística não-paramétrica: estimação, classificação e uma nova abordagem de seleção automática para largura de banda

Tese apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Doutor em Engenharia Elétrica. Área de Concentração: Sistemas Eletrônicos.

Aprovada em 15 de Abril de 2020.

BANCA EXAMINADORA



Rafael Antunes Nóbrega - Orientador
Universidade Federal de Juiz de Fora



Prof. Augusto Santiago Cerqueira
Universidade Federal de Juiz de Fora



Prof. Luciano Manhães de Andrade Filho
Universidade Federal de Juiz de Fora



Luis Fernando Gomez Gonzalez
Universidade Estadual de Campinas



Prof. Marcello Luiz Rodrigues de Campos
Universidade Federal do Rio de Janeiro

AGRADECIMENTOS

Primeiro, gostaria de agradecer a Deus, razão suprema de todas as coisas.

Aos meus pais, Adélia e Carlos, pelo incentivo e apoio incondicional em todas as batalhas.

À minha esposa, Laila, pelo amor, companheirismo e paciência. Estive morto, agora posso voltar à vida e agradeço a Deus por ser ao meu lado.

Ao meu filho, Luan, pela força que exerce em mim. Prometo recompensar todo tempo longe de você.

Ao meu eterno orientador, Rafael Nóbrega. Me sinto agraciado por tê-lo como referência profissional e amigo.

Aos meus companheiros e amigos de Laboratório, Guilherme, Antônio, Tiago, Igor Abritta, Igor Pains, Mariana, Amaro, Felipe, Alan, Marcelo Paschoal, Luan, Carlos Amorim e Rafael Fusário. Vocês foram determinantes para que eu suportasse a estatística não-paramétrica e melhores ainda para dividir a loucura.

Finalmente, agradeço à CAPES(Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), à Universidade Federal de Juiz de Fora e à Faculdade de Engenharia por todo o suporte e pelas ferramentas necessárias ao desenvolvimento deste trabalho.

RESUMO

Esta tese teve como motivação conhecer o estado da arte em estimação não-paramétrica de densidade de probabilidade, avaliar as técnicas mais proeminentes encontradas em publicações científicas, compará-las em diversas situações e avaliar seu impacto em classificação utilizando verossimilhança. Para isto, foi realizado um estudo sobre a escolha automática da largura de banda, principal parâmetro utilizado pelos quatro estimadores não-paramétricos de densidade clássicos: Histograma, *Average Shifted Histogram* (ASH), Polígono de Frequência (PF) e *Kernel Density Estimation* (KDE). Em linhas gerais, o método KDE mostrou os melhores resultados em todas as distribuições testadas e devido a esse desempenho sua análise foi mais aprofundada, adentrando nas teorias do KDE com largura de banda variável. Ademais, foi percebido nos diversos testes realizados que os seletores baseados em validação-cruzada são mais resilientes do que os métodos de *Plug-In* (PI), levando a melhores resultados de estimação e classificação em realidades complexas. Por fim, este trabalho teve como desdobramento algumas contribuições para o estado da arte no assunto de investigação, cujas principais são elencadas a seguir: aumento do conhecimento sobre alguns dos principais estimadores não-paramétricos discutidos no mundo científico; desenvolvimento de uma técnica de avaliação de estimadores de densidade, nomeada de *Region of Interest Map* (RoIMap); proposta de uma técnica automática híbrida para ajustar o seletor de largura de banda variável, denominada *Region of Interest-based Kernel Density Estimation* (ROIKDE); e avaliação do impacto da estimação não-paramétrica em classificação de amostras.

Palavras-chave: Estimação não-paramétrica. Largura de banda. KDE. Verossimilhança *naive*. Classificação.

ABSTRACT

The thesis initial motivation was to know the state-of-the-art in non-parametric density estimation, compare different situations and assess their impact on the likelihood-based classification. Therefore, a study was carried out related to the automatic choice of bandwidth, the main parameter used by the four classic non-parametric estimators: Histogram, Average Shifted Histogram, Frequency Polygon and Kernel Density Estimation (KDE). In general, the KDE method showed the best results in all tested distributions and, due to this performance, its analysis was further developed, entering into the variable KDE theories with variable bandwidth. Furthermore, several tests shown that the selectors based on cross-validation are more resilient than the Plug-In methods, leading to better density estimation and classification results in complex problems. Finally, this thesis unfolded in some contributions to the state-of-the-art in the investigation subject, whose main ones are listed below: increased knowledge about some of the main non-parametric estimators discussed in the scientific world; development of a technique for evaluating density estimators called the Region of Interest Map (RoIMap); proposal for a hybrid automatic technique to adjust the variable bandwidth selector called Region of interest-based kernel density estimation (ROIKDE); and impact evaluation of the nonparametric estimation in classifying samples.

Keywords: Nonparametric estimation. Bandwidth. KDE. Likelihood naive. Classification.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplificação da estimação de densidade via Histograma.	34
Figura 2 – Distribuição Binomial com $p=0.5$	36
Figura 3 – Exemplificação da estimação de densidade via Polígono de Frequência.	42
Figura 4 – AMISE para Histograma e PF em uma distribuição Normal	42
Figura 5 – ASH e ASFP calculados para uma distribuição normal, com diversas variações de m	44
Figura 6 – Impacto da variação de h de banda fixa na estimação da PDF.	49
Figura 7 – Fluxograma das etapas de estimação e classificação.	60
Figura 8 – Distribuição D1a.	63
Figura 9 – Distribuição D1b.	63
Figura 10 – Distribuição D1c.	63
Figura 11 – Distribuição D2a.	64
Figura 12 – Distribuição D2b.	64
Figura 13 – Distribuição D2c.	64
Figura 14 – Distribuição D3a.	65
Figura 15 – Distribuição D3b.	65
Figura 16 – Distribuição D3c.	65
Figura 17 – Distribuição D4a.	66
Figura 18 – Distribuição D4b.	66
Figura 19 – Distribuição D4c.	66
Figura 20 – Área do erro da modelo discreto com o aumento do número de pontos de <i>grid</i>	67
Figura 21 – Área do erro com o aumento do número de pontos de estimação.	67
Figura 22 – Função erro E e o <i>bin</i> selecionado via BIC para o método LHM em uma distribuição D1a.	69
Figura 23 – Função de Custo $Z(h)$ do método de Shimazaki e Shinomoto.	70
Figura 24 – Função de Custo $Q(h)$ do método de Rudemo.	70
Figura 25 – Função de Custo $\log\phi(N_b)$ do método de Knuth.	70
Figura 26 – Área do erro de acordo com o número de <i>bins</i> para distribuição D1a.	71
Figura 27 – Área do erro de acordo com o número de <i>bins</i> para distribuição de D1c.	71
Figura 28 – Escolha do m para o método ASH na Distribuição D1a.	72
Figura 29 – Escolha do m para o método ASH na Distribuição D2a.	72

Figura 30 – Porcentagem total de mínimos locais para todas as iterações (1500) em todas as distribuições. *não convergiu no intervalo.	74
Figura 31 – Método MLCV.	74
Figura 32 – Método UCV.	74
Figura 33 – Método BCV1.	75
Figura 34 – Método BCV2.	75
Figura 35 – Método CCV.	75
Figura 36 – Método MCV.	75
Figura 37 – Método TCV.	75
Figura 38 – Método OSCV.	75
Figura 39 – Estimação de D1a via KDE fixo	78
Figura 40 – SiZeR <i>Map</i> para D1a	78
Figura 41 – Estimação de D3c via KDE fixo.	78
Figura 42 – SiZeR <i>Map</i> para D3c	78
Figura 43 – Eficiência da escolha do número de picos via PDF estimada.	79
Figura 44 – Eficiência da escolha do número de picos via SiZeR.	79
Figura 45 – Assimetria e Curtose das distribuições: Conjunto com 25 amostras.	79
Figura 46 – Assimetria e Curtose das distribuições: Conjunto com 500 amostras.	79
Figura 47 – Rol odernada através da variável aleatória	80
Figura 48 – Rol odernada através da probabilidade	80
Figura 49 – Rol odernada através da derivada	80
Figura 50 – Seletores ordenados pelo menor valor de área do erro de acordo com a probabilidade máxima.	81
Figura 51 – Seletores ordenados pelo menor valor de área do erro de acordo com a probabilidade mínima.	81
Figura 52 – Seletores ordenados pelo menor valor de área do erro de acordo com a derivada máxima.	81
Figura 53 – Seletores ordenados pelo menor valor de área do erro de acordo com a derivada mínima.	81
Figura 54 – Escolha da largura de banda via ROIKDE para Distribuição D1a.	83
Figura 55 – Escolha da largura de banda via ROIKDE para Distribuição D2a.	83
Figura 56 – Área do erro entre modelo discreto da distribuição D2a e sua estimacão via KDE com 25 amostras.	84
Figura 57 – Ferramenta RolMap utilizada na distribuição D2a.	85

Figura 58 – Variação de assimetria para uma Distribuição Log-Normal.	86
Figura 59 – Gráfico de assimetria pela variação do parâmetro σ	86
Figura 60 – Variação de curtose para uma Distribuição GGD.	87
Figura 61 – Gráfico de curtose pela variação do parâmetro ρ	87
Figura 62 – Dados simulados à partir das densidades vistas anteriormente.	88
Figura 63 – Densidades de um problema de física de partículas extraída de uma simulação realizada por Geant4.	88
Figura 64 – Performance do grupo G0 utilizando o estimador Histograma.	93
Figura 65 – Performance do grupo G0 utilizando o estimador PF.	93
Figura 66 – Performance do grupo G0 utilizando o estimador ASH.	93
Figura 67 – Performance do grupo G1 utilizando o estimador Histograma.	95
Figura 68 – Performance do grupo G1 utilizando o estimador PF.	95
Figura 69 – Performance do grupo G1 utilizando o estimador ASH.	95
Figura 70 – Performance do grupo G2 utilizando o estimador Histograma.	96
Figura 71 – Performance do grupo G2 utilizando o estimador PF.	96
Figura 72 – Performance do grupo G2 utilizando o estimador ASH.	96
Figura 73 – Performance do grupo G3 utilizando o estimador Histograma.	97
Figura 74 – Performance do grupo G3 utilizando o estimador PF.	97
Figura 75 – Performance do grupo G3 utilizando o estimador ASH.	97
Figura 76 – Performance do grupo G4 utilizando o estimador Histograma.	98
Figura 77 – Performance do grupo G4 utilizando o estimador PF.	98
Figura 78 – Performance do grupo G4 utilizando o estimador ASH.	98
Figura 79 – Variação dos <i>bins</i> para o Histograma com 1000 amostras.	99
Figura 80 – Área do erro para o Histograma com 1000 amostras.	99
Figura 81 – [G0]Performance do grupo contendo todas as Distribuições para o KDE.	101
Figura 82 – [G1]Performance do grupo contendo as Distribuições unimodais para o KDE.	102
Figura 83 – [G2]Performance do grupo contendo as Distribuições bimodais com deri- vada suave e trimodal pouco esparsa para o KDE.	102
Figura 84 – [G3]Performance do grupo contendo as Distribuições bastante esparsas para o KDE.	103
Figura 85 – [G4]Performance do grupo contendo as Distribuições com descontinui- dade para o KDE.	104
Figura 86 – Variação de h para o KDE com 25 amostras.	105

Figura 87 – Área do erro para o KDE com 25 amostras.	105
Figura 88 – Variação de h para o KDE com 1000 amostras.	106
Figura 89 – Área do erro para o KDE com 1000 amostras.	106
Figura 90 – Comparação entre os estimadores de banda fixa para Distribuição D1c.	107
Figura 91 – Comparação entre os estimadores de banda fixa para Distribuição D2c.	107
Figura 92 – Comparação entre os estimadores de banda fixa para Distribuição D3c.	107
Figura 93 – Comparação entre os estimadores de banda fixa para Distribuição D4c.	107
Figura 94 – Assimetria para o KDE de banda fixa.	108
Figura 95 – Curtose para o KDE de banda fixa.	109
Figura 96 – Ferramenta RolMap utilizada na distribuição D1b.	110
Figura 97 – Área do erro dos métodos variáveis para Distribuição D1b.	111
Figura 98 – Área do erro em relação ao ROIKDE para Distribuição D1b.	111
Figura 99 – Área do erro dos métodos variáveis para Distribuição D1a.	112
Figura 100–Área do erro dos métodos variáveis para Distribuição D1c.	112
Figura 101–Ferramenta RolMap utilizada na distribuição D2c.	113
Figura 102–Área do erro dos métodos variáveis para Distribuição D2c.	113
Figura 103–Área do erro em relação ao ROIKDE para Distribuição D2c.	113
Figura 104–Área do erro dos métodos variáveis para Distribuição D2a.	114
Figura 105–Área do erro dos métodos variáveis para Distribuição D2b.	114
Figura 106–Ferramenta RolMap utilizada na distribuição D3a.	114
Figura 107–Área do erro dos métodos variáveis para Distribuição D3a.	115
Figura 108–Área do erro em relação ao ROIKDE para Distribuição D3a.	115
Figura 109–Área do erro dos métodos variáveis para Distribuição D3b.	115
Figura 110–Área do erro dos métodos variáveis para Distribuição D3c.	115
Figura 111–Ferramenta RolMap utilizada na distribuição D4a.	116
Figura 112–Área do erro dos métodos variáveis para Distribuição D4a.	117
Figura 113–Área do erro em relação ao ROIKDE para Distribuição D4a.	117
Figura 114–Área do erro dos métodos variáveis para Distribuição D4b.	117
Figura 115–Área do erro dos métodos variáveis para Distribuição D4c.	117
Figura 116–Comparação geral entre os estimadores não-paramétricos de largura de banda variável.	118
Figura 117–Assimetria para o KDE de banda variável.	120
Figura 118–Curtose para o KDE de banda variável.	121
Figura 119–Caso representativo (1000 amostras) da ROC para os dados simulados	121

Figura 120–Performance dos algoritmos de classificação para os dados simulados.	122
Figura 121–Performance dos algoritmos de classificação para dados referentes a identificação de partículas.	123
Figura 122–Variável aleatória sem <i>outlier</i> na distribuição D1a.	139
Figura 123–Adição de <i>outliers</i> simétricos (Esquerda) - <i>Zoom</i> no histograma (direita)	139
Figura 124– <i>Outlier</i> em 100, adicionado na variável aleatória D1a.	140
Figura 125– <i>Outlier</i> em 100, adicionado na variável aleatória D2a.	140
Figura 126–Distribuição das distâncias D para algumas distribuições.	141
Figura 127–Detecção e degradação por <i>outlier</i> na distribuição D2a.	143
Figura 128–Detecção e degradação por <i>outlier</i> na distribuição D1b.	143
Figura 129– <i>Outliers</i> adicionados a distribuição tridimensional D2a.	143
Figura 130–Zoom na distribuição D2a.	143
Figura 131–Variável aleatória Gaussiana com descontinuidade em 0 e -1.	144
Figura 132–Variação da Binagem utilizando Histograma na distribuição D1a	145
Figura 133–Área do erro utilizando Histograma na distribuição D1a	145
Figura 134–Variação da Binagem utilizando PF na distribuição D1a	145
Figura 135–Área do erro utilizando PF na distribuição D1a	145
Figura 136–Variação da Binagem utilizando ASH na distribuição D1a	146
Figura 137–Área do erro utilizando ASH na distribuição D1a	146
Figura 138–Variação da Binagem utilizando Histograma na distribuição D1b	146
Figura 139–Área do erro utilizando Histograma na distribuição D1b	146
Figura 140–Variação da Binagem utilizando PF na distribuição D1b	146
Figura 141–Área do erro utilizando PF na distribuição D1b	146
Figura 142–Variação da Binagem utilizando ASH na distribuição D1b	147
Figura 143–Área do erro utilizando ASH na distribuição D1b	147
Figura 144–Variação da Binagem utilizando Histograma na distribuição D1c	147
Figura 145–Área do erro utilizando Histograma na distribuição D1c	147
Figura 146–Variação da Binagem utilizando PF na distribuição D1c	147
Figura 147–Área do erro utilizando PF na distribuição D1c	147
Figura 148–Variação da Binagem utilizando ASH na distribuição D1c	148
Figura 149–Área do erro utilizando ASH na distribuição D1c	148
Figura 150–Variação da Binagem utilizando Histograma na distribuição D2a	148
Figura 151–Área do erro utilizando Histograma na distribuição D2a	148
Figura 152–Variação da Binagem utilizando PF na distribuição D2a	148

Figura 153–Área do erro utilizando PF na distribuição D2a	148
Figura 154–Variação da Binagem utilizando ASH na distribuição D2a	149
Figura 155–Área do erro utilizando ASH na distribuição D2a	149
Figura 156–Variação da Binagem utilizando Histograma na distribuição D2b	149
Figura 157–Área do erro utilizando Histograma na distribuição D2b	149
Figura 158–Variação da Binagem utilizando PF na distribuição D2b	149
Figura 159–Área do erro utilizando PF na distribuição D2b	149
Figura 160–Variação da Binagem utilizando ASH na distribuição D2b	150
Figura 161–Área do erro utilizando ASH na distribuição D2b	150
Figura 162–Variação da Binagem utilizando Histograma na distribuição D2c	150
Figura 163–Área do erro utilizando Histograma na distribuição D2c	150
Figura 164–Variação da Binagem utilizando PF na distribuição D2c	150
Figura 165–Área do erro utilizando PF na distribuição D2c	150
Figura 166–Variação da Binagem utilizando ASH na distribuição D2c	151
Figura 167–Área do erro utilizando ASH na distribuição D2c	151
Figura 168–Variação da Binagem utilizando Histograma na distribuição D3a	151
Figura 169–Área do erro utilizando Histograma na distribuição D3a	151
Figura 170–Variação da Binagem utilizando PF na distribuição D3a	151
Figura 171–Área do erro utilizando PF na distribuição D3a	151
Figura 172–Variação da Binagem utilizando ASH na distribuição D3a	152
Figura 173–Área do erro utilizando ASH na distribuição D3a	152
Figura 174–Variação da Binagem utilizando Histograma na distribuição D3b	152
Figura 175–Área do erro utilizando Histograma na distribuição D3b	152
Figura 176–Variação da Binagem utilizando PF na distribuição D3b	152
Figura 177–Área do erro utilizando PF na distribuição D3b	152
Figura 178–Variação da Binagem utilizando ASH na distribuição D3b	153
Figura 179–Área do erro utilizando ASH na distribuição D3b	153
Figura 180–Variação da Binagem utilizando Histograma na distribuição D3c	153
Figura 181–Área do erro utilizando Histograma na distribuição D3c	153
Figura 182–Variação da Binagem utilizando PF na distribuição D3c	153
Figura 183–Área do erro utilizando PF na distribuição D3c	153
Figura 184–Variação da Binagem utilizando ASH na distribuição D3c	154
Figura 185–Área do erro utilizando ASH na distribuição D3c	154
Figura 186–Variação da Binagem utilizando Histograma na distribuição D4a	154

Figura 187–Área do erro utilizando Histograma na distribuição D4a	154
Figura 188–Variação da Binagem utilizando PF na distribuição D4a	154
Figura 189–Área do erro utilizando PF na distribuição D4a	154
Figura 190–Variação da Binagem utilizando ASH na distribuição D4a	155
Figura 191–Área do erro utilizando ASH na distribuição D4a	155
Figura 192–Variação da Binagem utilizando Histograma na distribuição D4b	155
Figura 193–Área do erro utilizando Histograma na distribuição D4b	155
Figura 194–Variação da Binagem utilizando PF na distribuição D4b	155
Figura 195–Área do erro utilizando PF na distribuição D4b	155
Figura 196–Variação da Binagem utilizando ASH na distribuição D4b	156
Figura 197–Área do erro utilizando ASH na distribuição D4b	156
Figura 198–Variação da Binagem utilizando Histograma na distribuição D4c	156
Figura 199–Área do erro utilizando Histograma na distribuição D4c	156
Figura 200–Variação da Binagem utilizando PF na distribuição D4c	156
Figura 201–Área do erro utilizando PF na distribuição D4c	156
Figura 202–Variação da Binagem utilizando ASH na distribuição D4c	157
Figura 203–Área do erro utilizando ASH na distribuição D4c	157
Figura 204–Variação de h utilizando KDE na distribuição D1a	157
Figura 205–Área do erro utilizando KDE na distribuição D1a	157
Figura 206–Variação de h utilizando KDE na distribuição D1b	157
Figura 207–Área do erro utilizando KDE na distribuição D1b	157
Figura 208–Variação de h utilizando KDE na distribuição D1c	158
Figura 209–Área do erro utilizando KDE na distribuição D1c	158
Figura 210–Variação de h utilizando KDE na distribuição D2a	158
Figura 211–Área do erro utilizando KDE na distribuição D2a	158
Figura 212–Variação de h utilizando KDE na distribuição D2b	158
Figura 213–Área do erro utilizando KDE na distribuição D2b	158
Figura 214–Variação de h utilizando KDE na distribuição D2c	159
Figura 215–Área do erro utilizando KDE na distribuição D2c	159
Figura 216–Variação de h utilizando KDE na distribuição D3a	159
Figura 217–Área do erro utilizando KDE na distribuição D3a	159
Figura 218–Variação de h utilizando KDE na distribuição D3b	159
Figura 219–Área do erro utilizando KDE na distribuição D3b	159
Figura 220–Variação de h utilizando KDE na distribuição D3c	160

Figura 221–Área do erro utilizando KDE na distribuição D3c	160
Figura 222–Variação de h utilizando KDE na distribuição D4a	160
Figura 223–Área do erro utilizando KDE na distribuição D4a	160
Figura 224–Variação de h utilizando KDE na distribuição D4b	160
Figura 225–Área do erro utilizando KDE na distribuição D4b	160
Figura 226–Variação de h utilizando KDE na distribuição D4c	161
Figura 227–Área do erro utilizando KDE na distribuição D4c	161
Figura 228–Variação de h para todas as distribuições com 50 amostras de treinamento no KDE	161
Figura 229–Área do erro para todas as distribuições com 50 amostras de treinamento no KDE	161
Figura 230–Variação de h para todas as distribuições com 100 amostras de treina- mento no KDE	162
Figura 231–Área do erro para todas as distribuições com 100 amostras de treinamento no KDE	162
Figura 232–Variação de h para todas as distribuições com 500 amostras de treina- mento no KDE	162
Figura 233–Área do erro para todas as distribuições com 500 amostras de treinamento no KDE	162
Figura 234–Variação dos <i>bins</i> para todas as distribuições com 25 amostras de treina- mento no Histograma	163
Figura 235–Área do erro para todas as distribuições com 25 amostras de treinamento no Histograma	163
Figura 236–Variação dos <i>bins</i> para todas as distribuições com 50 amostras de treina- mento no Histograma	163
Figura 237–Área do erro para todas as distribuições com 50 amostras de treinamento no Histograma	163
Figura 238–Variação dos <i>bins</i> para todas as distribuições com 100 amostras de treinamento no Histograma	164
Figura 239–Área do erro para todas as distribuições com 100 amostras de treinamento no Histograma	164
Figura 240–Variação dos <i>bins</i> para todas as distribuições com 500 amostras de treinamento no Histograma	164

Figura 241–Área do erro para todas as distribuições com 500 amostras de treinamento no Histograma	164
Figura 242–Variação dos <i>bins</i> para todas as distribuições com 25 amostras de treinamento no PF	165
Figura 243–Área do erro para todas as distribuições com 25 amostras de treinamento no PF	165
Figura 244–Variação dos <i>bins</i> para todas as distribuições com 50 amostras de treinamento no PF	165
Figura 245–Área do erro para todas as distribuições com 50 amostras de treinamento no PF	165
Figura 246–Variação dos <i>bins</i> para todas as distribuições com 100 amostras de treinamento no PF	166
Figura 247–Área do erro para todas as distribuições com 100 amostras de treinamento no PF	166
Figura 248–Variação dos <i>bins</i> para todas as distribuições com 500 amostras de treinamento no PF	166
Figura 249–Área do erro para todas as distribuições com 500 amostras de treinamento no PF	166
Figura 250–Variação dos <i>bins</i> para todas as distribuições com 1000 amostras de treinamento no PF	167
Figura 251–Área do erro para todas as distribuições com 1000 amostras de treinamento no PF	167
Figura 252–Variação dos <i>bins</i> para todas as distribuições com 25 amostras de treinamento no ASH	167
Figura 253–Área do erro para todas as distribuições com 25 amostras de treinamento no ASH	167
Figura 254–Variação dos <i>bins</i> para todas as distribuições com 50 amostras de treinamento no ASH	168
Figura 255–Área do erro para todas as distribuições com 50 amostras de treinamento no ASH	168
Figura 256–Variação dos <i>bins</i> para todas as distribuições com 100 amostras de treinamento no ASH	168
Figura 257–Área do erro para todas as distribuições com 100 amostras de treinamento no ASH	168

Figura 258–Variação dos <i>bins</i> para todas as distribuições com 500 amostras de treinamento no ASH	169
Figura 259–Área do erro para todas as distribuições com 500 amostras de treinamento no ASH	169
Figura 260–Variação dos <i>bins</i> para todas as distribuições com 1000 amostras de treinamento no ASH	169
Figura 261–Área do erro para todas as distribuições com 1000 amostras de treinamento no ASH	169
Figura 262–Comparação entre estimadores para Distribuição D1a.	170
Figura 263–Comparação entre estimadores para Distribuição D1b.	170
Figura 264–Comparação entre estimadores para Distribuição D2a.	170
Figura 265–Comparação entre estimadores para Distribuição D2b.	170
Figura 266–Comparação entre estimadores para Distribuição D3a.	171
Figura 267–Comparação entre estimadores para Distribuição D3b.	171
Figura 268–Comparação entre estimadores para Distribuição D4a.	171
Figura 269–Comparação entre estimadores para Distribuição D4b.	171
Figura 270–Ferramenta RolMap utilizada na distribuição D1a para 25 amostras. . .	172
Figura 271–Ferramenta RolMap utilizada na distribuição D1a para 1000 amostras. .	172
Figura 272–Ferramenta RolMap utilizada na distribuição D1b para 1000 amostras. .	173
Figura 273–Ferramenta RolMap utilizada na distribuição D1c para 25 amostras. . .	173
Figura 274–Ferramenta RolMap utilizada na distribuição D1c para 1000 amostras. .	174
Figura 275–Ferramenta RolMap utilizada na distribuição D2a para 25 amostras. . .	174
Figura 276–Ferramenta RolMap utilizada na distribuição D2a para 1000 amostras. .	175
Figura 277–Ferramenta RolMap utilizada na distribuição D2b para 1000 amostras. .	175
Figura 278–Ferramenta RolMap utilizada na distribuição D2c para 25 amostras. . .	176
Figura 279–Ferramenta RolMap utilizada na distribuição D2c para 1000 amostras. .	176
Figura 281–Ferramenta RolMap utilizada na distribuição D3b para 25 amostras. . .	177
Figura 280–Ferramenta RolMap utilizada na distribuição D3a para 1000 amostras. .	177
Figura 282–Ferramenta RolMap utilizada na distribuição D3b para 1000 amostras. .	178
Figura 283–Ferramenta RolMap utilizada na distribuição D3c para 25 amostras. . .	178
Figura 284–Ferramenta RolMap utilizada na distribuição D3c para 1000 amostras. .	179
Figura 285–Ferramenta RolMap utilizada na distribuição D4a para 1000 amostras. .	179
Figura 286–Ferramenta RolMap utilizada na distribuição D4b para 25 amostras. . .	180
Figura 287–Ferramenta RolMap utilizada na distribuição D4b para 1000 amostras. .	180

Figura 288–Ferramenta RoIMap utilizada na distribuição D4c para 25 amostras. . .	181
Figura 289–Ferramenta RoIMap utilizada na distribuição D4c para 1000 amostras. .	181
Figura 290–Área do erro em relação ao ROIKDE para D1a	182
Figura 291–Área do erro em relação ao ROIKDE para D1c	182
Figura 292–Área do erro em relação ao ROIKDE para D2a.	182
Figura 293–Área do erro em relação ao ROIKDE para D2b.	182
Figura 294–Área do erro em relação ao ROIKDE para D3b.	183
Figura 295–Área do erro em relação ao ROIKDE para D3c.	183
Figura 296–Área do erro em relação ao ROIKDE para D4b.	183
Figura 297–Área do erro em relação ao ROIKDE para D4c.	183

LISTA DE TABELAS

Tabela 1 – Algoritmos seletores de largura de banda utilizados no KDE. (*) implementados nessa tese. (**) nova abordagem desenvolvida nessa tese.	30
Tabela 2 – Definição da PDF D1a.	63
Tabela 3 – Definição da PDF D1b.	63
Tabela 4 – Definição da PDF D1c.	63
Tabela 5 – Definição da PDF D2a.	64
Tabela 6 – Definição da PDF D2b.	64
Tabela 7 – Definição da PDF D2c.	64
Tabela 8 – Definição da PDF D3a.	65
Tabela 9 – Definição da PDF D3b.	65
Tabela 10 – Definição da PDF D3c.	65
Tabela 11 – Definição da PDF D4a.	66
Tabela 12 – Definição da PDF D4b.	66
Tabela 13 – Definição da PDF D4c.	66
Tabela 14 – Tabela da área do erro referente a 10^4 pontos de estimação, 10^5 pontos de <i>grid</i> e o erro total residual de interpolação.	68
Tabela 15 – Parâmetros utilizados no cálculo da binagem ótima para o Histograma. (*O método LHM será utilizado com interpolação “nearest”)	69
Tabela 16 – Parâmetros utilizados no cálculo da binagem ótima para o PF. (*O método LHM será utilizado com interpolação “linear”)	71
Tabela 17 – Parâmetros utilizados no cálculo da binagem ótima para o ASH.	72
Tabela 18 – Parâmetros utilizados no cálculo da largura de banda fixa h para o KDE.	73
Tabela 19 – Divisão dos grupos de densidades.	91
Tabela 20 – Melhora geral na área do erro do método ROIKDE em relação aos demais métodos.	118
Tabela 21 – Desempenho do método ROIKDE em relação aos demais métodos.	119
Tabela 22 – Comparação relativa geral dos métodos de banda variável.	119
Tabela 23 – Melhor SP para os dados simulados com as densidades vistas anteriormente.	122
Tabela 24 – Melhor SP para os dados de identificação de partículas.	124
Tabela 25 – Área do erro para distribuição D1a.	184
Tabela 26 – Área do erro para distribuição D1b.	184

Tabela 27 – Área do erro para distribuição D1c.	185
Tabela 28 – Área do erro para distribuição D2a.	185
Tabela 29 – Área do erro para distribuição D2b.	186
Tabela 30 – Área do erro para distribuição D2c.	186
Tabela 31 – Área do erro para distribuição D3a.	187
Tabela 32 – Área do erro para distribuição D3b.	187
Tabela 33 – Área do erro para distribuição D3c.	188
Tabela 34 – Área do erro para distribuição D4a.	188
Tabela 35 – Área do erro para distribuição D4b.	189
Tabela 36 – Área do erro para distribuição D4c.	189
Tabela 37 – Área do erro para distribuição D1a.	189
Tabela 38 – Área do erro para distribuição D1b.	190
Tabela 39 – Área do erro para distribuição D1c.	190
Tabela 40 – Área do erro para distribuição D2a.	190
Tabela 41 – Área do erro para distribuição D2b.	190
Tabela 42 – Área do erro para distribuição D2c.	190
Tabela 43 – Área do erro para distribuição D3a.	191
Tabela 44 – Área do erro para distribuição D3b.	191
Tabela 45 – Área do erro para distribuição D3c.	191
Tabela 46 – Área do erro para distribuição D4a.	191
Tabela 47 – Área do erro para distribuição D4b.	191
Tabela 48 – Área do erro para distribuição D4c.	192
Tabela 49 – Desempenho do método AKDE em relação aos demais métodos.	192
Tabela 50 – Desempenho do método BKDE em relação aos demais métodos.	192
Tabela 51 – Desempenho do método VKDE em relação aos demais métodos.	193

LISTA DE ABREVIATURAS E SIGLAS

AISB	<i>Asymptotic Integrated Squared Bias</i>
AIV	<i>Asymptotic Integrated Variance</i>
AKDE	<i>Adaptive Kernel Density Estimation</i>
AMISE	<i>Asymptotic Mean Integrated Square Error</i>
ASFP	<i>Average Shifted Frequency Polygon</i>
ASH	<i>Average Shifted Histogram</i>
ATLAS	<i>A Toroidal LHC ApparatuS</i>
BCV1	<i>Biased Cross-Validation 1</i>
BCV2	<i>Biased Cross-Validation 2</i>
BCV	<i>Biased Cross-Validation</i>
BIC	<i>Bayesian Information Criterion</i>
BKDE	<i>Binned Kernel Density Estimator</i>
CCV	<i>Complete Cross-Validation</i>
CDF	<i>Cumulative Distribution Function</i>
CERN	<i>Conseil Européen pour la Recherche Nucléaire</i>
CV	<i>Cross-Validation</i>
ECDF	<i>Empirical Cumulative Distribution Function</i>
FD	<i>Freedman-Diaconis</i>
GEANT	<i>Geometry and Tracking</i>
GGD	<i>Generalized Gaussian Distribution</i>
IMSE	<i>Integrated Mean Square Error</i>
IQR	<i>Interquartile Range</i>
ISB	<i>Integrated Squared Bias</i>
ISE	<i>Integrated Square Error</i>
IV	<i>Integrated Variance</i>
KDE	<i>Kernel Density Estimator</i>
L1I	<i>L1 Improved</i>
LSCV	<i>Least Square Cross-Validation</i>
MAD	<i>Median Absolute Deviation</i>
MCV	<i>Modified Cross-Validation</i>
MISE	<i>Mean Integrated Square Error</i>
MKDE	<i>Multivariate Kernel Density Estimator</i>
MLCV	<i>Maximum Likelihood Cross-Validation</i>

OSCV	<i>One Sided Cross-Validation</i>
PDF	<i>Probability Density Function</i>
PF	Polígonos de Frequência
PI	<i>Plug-In</i>
ROC	<i>Receiver Operating Curve</i>
Roi	<i>Region of Interest</i>
ROIKDE	<i>Region of Interest-based Kernel Density Estimation</i>
RoiMap	<i>Region of Interest Map</i>
SC	<i>Scott</i>
SiZeR	<i>Significant Zero crossings of derivative</i>
SJ	<i>Sheather and Jones</i>
SP	Soma-Produto
SS	<i>Shimazaki-Shinomoto</i>
SV	<i>Silverman</i>
SVM1	Silverman Modificação 1 - Robusto
SVM2	Silverman Modificação 2 - Adaptativo
TCV	<i>Trimmed Cross-Validation</i>
TVM	Teorema do Valor Médio
TVMG	Teorema do Valor Médio Generalizado
UCV	<i>Unbiased Cross-Validation</i>
VKDE	<i>Variable Kernel Density Estimation</i>

LISTA DE SÍMBOLOS

α_σ	Fator do ROIKDE referente ao desvio padrão da Gaussiana
α	Parâmetro 1 utilizado no seletor L1 Improved
β	Inverso do parâmetro de escala
δ	Passo do <i>grid</i>
η	Parâmetro adaptativo do método de Silverman
Γ	Função gamma
γ	Largura de banda diferente de h
$\hat{\sigma}$	Estimador do desvio padrão
κ_1	Assimetria
λ	Constante de proporcionalidade
μ	Média
ϕ	Função de densidade qualquer
ρ	Parâmetro de forma
σ	Desvio padrão
σ^2	Variância
τ	Parâmetro 1 de ajuste de escala do método Sheather-Jones
θ	Parâmetro qualquer
ν	Parâmetro 2 de ajuste de escala do método Sheather-Jones
ε	Parâmetro 2 utilizado no seletor L1 Improved
φ_α	Parâmetro 1 positivo de forma da distribuição Beta
φ_β	Parâmetro 2 positivo de forma da distribuição Beta
\hat{f}	Estimador da densidade de probabilidade
a	Limitante inferior da distribuição Uniforme
B	<i>Bin</i>
b	Limitante superior da distribuição Uniforme
Bn	Distribuição Binomial
Bt	Distribuição Beta
C	Contagem no <i>bin</i>
c	Parâmetro associado a escala do seletor
d	Número de dimensões
D_{bg}	Probabilidade de detecção de ruído
D_{sg}	Probabilidade de detecção de sinal
dL	Discriminante

E	Função custo do seletor LHM
f^p	Fator de probabilidade
$f^{d'}$	Fator de primeira derivada
f_n	Densidade estimada utilizando Kernel de Epanechnikov
G	<i>Grid</i>
g	Parâmetro positivo qualquer
g_n	Densidade estimada utilizando Kernel L
h	Largura de banda
h^*	Largura de banda ótima
I	Função Indicadora
K	Função Kernel
K_e	Termo adicional da regra de Doane
L	Função Kernel diferente de K
L^2	Norma L^2
L_b	Verossimilhança de ruído
L_s	Verossimilhança de sinal
LN	Distribuição Log-Normal
m	Número de coleção de Histogramas ou Polígonos de Frequência
N	Distribuição Normal
n	Número de amostras
n_σ	Número de amostras descartáveis
N_b	Número de <i>bins</i>
p	Probabilidade de sucesso
P_b	Probabilidade de ruído
P_s	Probabilidade de sinal
Q	Função custo do seletor de Rudemo
r	Ordem da derivada
R	Rugosidade
s	Numero de sucessos em n amostras/tentativas
t	Centro do <i>bin</i>
U	Distribuição Uniforme
V	Volume da esfera unitária
w	Constante de suavização do método One Sided Cross-Validation
x	Amostra

X	Variável aleatória
Z	Função custo do seletor de Shimazaki-Shinomoto
$\bar{\theta}$	Função 1 utilizada pelo método <i>Complete Cross-Validation</i>
$\hat{\alpha}_2$	Função 2 utilizada no método Sheather-Jones
$\hat{\psi}_r$	Estimador de escala normal
\hat{S}_D	Função 1 utilizada no método Sheather-Jones
\hat{T}_D	Função 3 utilizada no método Sheather-Jones
ϕ_{MCV}	Função utilizada pelo método <i>Modified Cross-Validation</i>
ϕ_{TCV}	Função utilizada pelo método <i>Trimmed Cross-Validation</i>
ξ	Função 2 utilizada pelo método <i>Complete Cross-Validation</i>

SUMÁRIO

1	INTRODUÇÃO	27
1.1	MOTIVAÇÃO	28
1.2	OBJETIVO	28
1.3	O QUE FOI FEITO	29
1.4	ESTRUTURA DA TESE	31
2	REVISÃO BIBLIOGRÁFICA	32
2.1	HISTOGRAMA	33
2.1.1	Definição	34
2.1.2	Seletores de Largura de Banda	35
2.2	POLÍGONOS DE FREQUÊNCIA	41
2.3	ASH	43
2.4	KDE	44
2.4.1	Definição	47
2.4.2	KDE com Largura de Banda Fixa	49
2.4.2.1	<i>Seletores de largura de banda</i>	50
2.4.2.2	<i>KDE Multivariado</i>	56
2.4.3	KDE com Largura de Banda Variável	56
3	DESENVOLVIMENTO	59
3.1	DADOS DE ENTRADA	61
3.1.1	Distribuições	61
3.1.2	Erros de Interpolação	67
3.2	ESTIMAÇÃO NÃO-PARAMÉTRICA	68
3.2.1	Histograma	68
3.2.2	Polígonos de Frequência	70
3.2.3	Average Shifted Histogram	71
3.2.4	Kernel Density Estimator	72
3.2.4.1	<i>Largura de Banda Fixa</i>	72
3.2.4.2	<i>Largura de Banda Variável</i>	76
3.3	ANÁLISE DA ESTIMAÇÃO	83
3.3.1	Área do Erro	83
3.3.2	Region of Interest Map (RoIMap)	84

3.3.3	Teste de terceiro e quarto momento central	86
3.4	CLASSIFICAÇÃO	86
3.4.1	Conjunto de Dados	87
3.4.1.1	<i>Dados simulados neste trabalho</i>	87
3.4.1.2	<i>Identificação de partículas</i>	87
3.4.2	Classificador	89
3.4.3	Análise da classificação	90
4	RESULTADOS	91
4.1	ESTIMAÇÃO DE DENSIDADES	91
4.1.1	Largura de Banda Fixa	91
4.1.1.1	<i>Avaliação dos Seletores</i>	92
4.1.1.2	<i>Histograma, PF e ASH</i>	92
4.1.1.3	<i>KDE</i>	100
4.1.1.4	<i>Comparação entre os estimadores de banda fixa</i>	106
4.1.2	Teste de 3º e 4º momento central	107
4.1.3	Largura de Banda Variável	109
4.1.4	Avaliação dos Estimadores	109
4.1.5	Teste de 3º e 4º Momento Central	120
4.2	CLASSIFICAÇÃO	120
4.2.1	Dados Simulados	122
4.2.2	Identificação de Partículas	123
5	CONCLUSÃO	125
5.1	PRÓXIMOS PASSOS	127
	REFERÊNCIAS	129
	APÊNDICE A – CRITÉRIO DE ERRO PARA ESTIMAÇÃO DE DENSIDADES	134
A.1	CRITÉRIO L^2 APLICADO AO HISTOGRAMA	135
A.2	CRITÉRIO L^2 APLICADO AO KDE	137
	APÊNDICE B – PRÉ-PROCESSAMENTO	138
B.1	OUTLIERS	138
B.1.1	Degradação da Estimação	138
B.1.2	Detecção de <i>Outlier</i>	140
B.1.2.1	<i>Proposta de Algoritmo de Detecção</i>	140

B.1.2.2	<i>Z-Score</i>	141
B.1.2.3	<i>Z-scores Robusto</i>	142
B.1.3	Testando algoritmos de Detecção de <i>Outliers</i>	142
B.2	<i>SPIKES</i>	144
	APÊNDICE C – AVALIAÇÃO DOS MÉTODOS	145
C.1	BANDA FIXA	145
C.1.1	Histograma, PF e ASH	145
C.1.2	Kernel Density Estimation	157
C.1.3	Matriz geral de comparação de banda fixa	161
C.1.4	Comparação entre estimadores de banda fixa	170
C.2	BANDA VARIÁVEL	172
C.2.1	RoIMap	172
C.2.2	Comparação em relação ao ROIKDE	182
	APÊNDICE D – TABELAS	184
D.1	KDE DE BANDA FIXA	184
D.2	KDE DE BANDA VARIÁVEL	189

1 INTRODUÇÃO

Atualmente, a construção de modelos multivariados utilizados em inferências probabilísticas tem se tornado cada vez mais difícil, devido ao aumento da complexidade dos experimentos científicos. Neste contexto, técnicas de estimação não-paramétrica de densidades de probabilidade tem sido amplamente utilizadas em diversas áreas do conhecimento, como: arqueologia, física de partículas, economia, genética, engenharia, medicina, biomedicina, estatística, teoria da informação, matemática, ciência da computação, entre outras.

Apesar da grande flexibilidade desses métodos, a aspiração por seus modelos ótimos é vista de forma bastante subjetiva e questionável por parte da comunidade que trabalha com o tema, indicando que pode não haver um processo perfeito para construção do modelo não-paramétrico (HEIDENREICH; SCHINDLER; SPERLICH, 2013). Sabendo-se disso, é importante notar que a qualidade do modelo passa pela capacidade do analista de dados em avaliar o problema e aplicar a teoria adequada para determinada realidade, aumentando o grau de dificuldade na implementação de algoritmos com soluções automáticas. Nessa perspectiva, geralmente encontramos na literatura a otimização dos modelos em conjuntos específicos de dados, reservando a comparação genérica entre os modelos e técnicas de otimização a um número expressivamente menor de trabalhos.

Este trabalho apresentará o desenvolvimento histórico do tema, através das quatro principais técnicas de estimação não-paramétrica de densidade de probabilidade, enfatizando as alternativas desenvolvidas por diversos autores para a otimização da largura de banda responsável pela troca entre viés e variância da estimação. Ademais, o método *Kernel Density Estimator* (KDE) será visto com maiores detalhes, juntamente com um algoritmo proposto nesta tese, baseado no ajuste e escolha automática da largura de banda variável. Neste contexto, serão apresentadas técnicas capazes de mensurar a “qualidade” da estimação, para então comparar e avaliar as características dos diferentes seletores de largura de banda. Por fim, os estimadores serão utilizados na construção de um classificador de verossimilhança aplicados a um conjunto de dados simulados neste trabalho e um problema de identificação de partículas, simulado pelo *software Geometry and Tracking* (Geant) 4 (ALLISON *et al.*, 2006).

A abordagem desse trabalho tem um viés experimental, utilizando o conhecimento adquirido nas teorias, aqui expostas, em algoritmos capazes de estimar automaticamente diferentes tipos de distribuições e classificar suas respectivas amostras. Para uma melhor

adequação a alguns dos principais problemas encontrados na prática, as densidades de probabilidade desse estudo foram escolhidas por suas características representarem graus distintos de dificuldade para cada tipo de seletor de largura de banda, sendo possível avaliar a sua resiliência a cada um desses problemas. Todos os algoritmos utilizados nessa tese podem ser encontrados em: <https://github.com/davidmelosouza/nonparametric_analysis>.

1.1 MOTIVAÇÃO

No ano de 2015, na dissertação de mestrado (SOUZA, 2015), foi desenvolvido um estudo sobre a identificação de elétrons através do método de verossimilhança, utilizando dados fornecidos pelo detector *A Toroidal LHC ApparatuS* (ATLAS) do *Conseil Européen pour la Recherche Nucléaire* (CERN). Para a construção das densidades de probabilidade conjuntas, responsáveis por descrever as características dos sinais e ruídos de fundo do problema, foi utilizado a estimação não-paramétrica via KDE. Desde então, foi crescente o interesse em relação a essas técnicas e trabalhos vinculados ao tema, originando uma busca no estado da arte por respostas que transcendiam o objetivo do trabalho de mestrado citado acima. Através dessa motivação inicial e genérica foi possível pleitear objetivos mais específicos dentro da literatura, mediante sua revisão.

1.2 OBJETIVO

Ao avaliarmos a linha histórica evolutiva do tema, através do prisma da engenharia, identificamos a necessidade de aprofundar nas técnicas de estimação de densidade de probabilidade não-paramétrica e correlacionar esse tema com classificação baseada em verossimilhança. Em linhas gerais, este trabalho tem dois objetivos principais:

1. Compreender o impacto dos diferentes seletores de largura de banda aplicados a estimação não-paramétrica de densidade de probabilidade;
2. Utilizar o conhecimento adquirido no item anterior para construir um classificador baseado em verossimilhança.

Para alcançar os objetivos citados alguns pontos específicos foram desenvolvidos ao longo deste estudo, explicitando parâmetros e abordagens necessárias para amparar as afirmações que serão feitas a posteriori. Dentre esses pontos destacamos:

1. Minimização dos erros gerados pelos processos de discretização das densidades de probabilidade simuladas e cálculos relacionados;

2. Escolha do intervalo de avaliação da função custo da validação-cruzada: observando quase em sua totalidade a convergência dos métodos;
3. Definição de um método de referência resiliente para os seletores com vários mínimos locais;
4. Definição de um intervalo em comum para a estimação e modelo, que garanta área próxima a 1;
5. Escolha de distribuições com níveis de dificuldade distintos e contemplando descontinuidades;
6. Avaliação dos seletores em relação à variação do terceiro e quarto momento central das distribuições;
7. Avaliação de três classes de seletores aqui definidos como: proeminentes, clássicos e aqueles ainda não contemplados na literatura no contexto de comparação;
8. Utilização de diferentes quantidades de amostras com o objetivo de avaliar os seletores em diferentes condições;
9. Utilização do conhecimento adquirido pelo estudo bibliográfico na construção de um novo método automático para escolha da largura de banda;
10. Apresentação dos resultados utilizando técnicas de estatística robusta;
11. Avaliação dos seletores através de uma métrica intuitiva;
12. Correlação entre resultados quantitativos e qualitativos;
13. Análise conjunta entre estimação e classificação de amostras.

1.3 O QUE FOI FEITO

Todos os itens citados anteriormente foram contemplados nesta tese. Antes da construção dos algoritmos foi feita uma revisão bibliográfica dos principais estimadores não-paramétricos e seus respectivos seletores de largura de banda. Os métodos foram comparados através de técnicas clássicas e uma técnica desenvolvida neste trabalho, capaz de avaliar a estimação ao longo de toda variável aleatória. Por fim, os algoritmos de estimação foram utilizados na construção de um classificador baseado em verossimilhança, tendo sua performance avaliada em dois conjuntos de amostras distintos. Sobre o desenvolvimento

dos algoritmos utilizados nessa tese podemos separar as implementações em: estimadores não-paramétricos, seletores de largura de banda, avaliadores de estimação e classificação de amostras.

Neste trabalho os principais estimadores não-paramétricos, presentes na literatura, foram implementados e avaliados: Histograma, *Average Shifted Histogram* (ASH), Polígonos de Frequência (PF) e KDE. Maiores detalhes sobre cada um dos estimadores serão apresentados na Seção 3.2.

Após a construção dos estimadores faz-se necessário calcular a largura de banda, parâmetro responsável pelas características da estimação. Os seletores de largura de banda utilizados por Histograma, PF e ASH, que foram avaliados nessa tese, são: Regra de Sturges, Regra de Scott, Regra de Freedman e Diaconis, Regra de Doane, Método de Shimazaki-Shinomoto, Método de Rudemo, LHM, Knuth e Wand. Dentre esses seletores apenas o Wand não foi implementado neste trabalho. Todos esses seletores serão discutidos com maiores detalhes nas Seções 2.1, 2.2 e 2.3, respectivamente.

Para o método KDE existem vários trabalhos dedicados aos seletores de largura de banda. A Tabela 1 mostra os seletores de largura de banda fixa e variável, que serão apresentados com maiores detalhes na Seção 2.4. Vale destacar que dentre esses seletores se encontra uma nova abordagem desenvolvida nessa tese, denominada *Region of Interest-based Kernel Density Estimation* (ROIKDE). Maiores detalhes sobre esse método serão apresentados na Seção 3.2.4.2.

Tabela 1 – Algoritmos seletores de largura de banda utilizados no KDE. (*) implementados nessa tese. (**) nova abordagem desenvolvida nessa tese.

Seletores de largura de banda	
Banda Fixa	Banda Variável
*Maximum Likelihood	*Sample Point Estimator
* <i>Unbiased</i>	Shimazaki-Shinomoto Adaptativo
* <i>Biased</i>	*Binned Kernel Density Estimator
* <i>Complete</i>	**ROIKDE
* <i>Modified</i>	
* <i>Trimmed</i>	
* <i>One Sided</i>	
*Silverman	
*Silverman Robusto	
*Silverman Adaptativo	
*Scott	
Sheather and Jones	
*L1 <i>Improved</i>	

Fonte: Elaborada pelo autor (2020).

Os estimadores com diferentes seletores de largura de banda serão comparados através do método clássico de *BoxPlot*, que representará a “distância” média entre a estimação da densidade de probabilidade e o modelo. Além dessa comparação teremos um método desenvolvido neste trabalho, denominado *Region of Interest Map* (RoIMap), que é capaz de avaliar a estimação localmente em cada região da variável aleatória. Essa técnica será vista com maior detalhe na Seção 3.3.

Por fim, todo avanço percebido na estimação não-paramétrica será utilizado em um classificador de verossimilhança *naive* baseado nas estimações de densidade de probabilidade, realizadas via KDE, com seletores de largura de banda variável. Os resultados de cada um dos seletores serão comparados através de um critério que leva em consideração tanto a eficiência de detecção quanto o falso alarme destes métodos. A etapa de classificação será descrita na Seção 3.4.2.

As respectivas referências dos algoritmos/bibliotecas que não foram implementados nessa tese podem ser encontrados na Seção 3.

1.4 ESTRUTURA DA TESE

Esta tese está organizada da seguinte forma: No **capítulo 2** teremos a revisão bibliográfica sobre o tema principal, que permeia todo estudo: estimadores não-paramétricos. O **capítulo 3** define uma metodologia e implementa as teorias relativas aos seletores, através de algoritmos que serão testados em diversas realidades. Para aproximar as simulações da prática, as distribuições foram escolhidas com o intuito de abranger diversas áreas do conhecimento e, conjuntamente, avaliar a resiliência dos seletores a esses desafios. Além disso, nesse capítulo, classificação e estimação não-paramétrica serão relacionadas. No **capítulo 4** teremos os resultados e comparações desses seletores no ambiente proposto anteriormente. Testes relacionados a performance dos diversos seletores e estimadores serão analisados e os resultados da classificação discutidos. Por fim, teremos o **capítulo 5** onde diversas ponderações serão concluídas, de acordo com o trabalho desenvolvido, e algumas diretrizes serão discutidas.

2 REVISÃO BIBLIOGRÁFICA

A base teórica deste trabalho está alicerçada na estimação não-paramétrica de densidades de probabilidade. Por estimação não-paramétrica de densidades entende-se um conjunto de métodos capazes de representar a característica probabilística de uma determinada variável aleatória, geralmente utilizado quando essa característica não pode ser inferida a priori. Nesta seção, teremos uma revisão sobre os principais estimadores não-paramétricos de densidade, sendo apresentados alguns dos desafios enfrentados por analistas de dados em uma busca histórica por melhores seletores de largura de banda.

O erro de um dispositivo eletrônico medidor de glicemia, a vida útil de uma lâmpada e o número de neutrinos emitidos por um reator nuclear, são exemplos de variáveis aleatórias geradas em processos distintos. Embora exista um grande número de modelos paramétricos, ou *Probability Density Function* (PDF), capazes de descrever as características de determinadas variáveis aleatórias, na prática, um número relativamente pequeno desses modelos obteve destaque. Este fato ocorre devido a apenas algumas dessas funções possuírem características matemáticas desejáveis e/ou por se adequarem bem a uma parcela da realidade (FORBES *et al.*, 2011). Essa constatação evidencia uma demanda das pesquisas atuais, pois, com o aumento da complexidade dos experimentos científicos é natural que os modelos paramétricos se tornem cada vez mais complexos, dificultando a concepção e parametrização de funções matemáticas que representem as variáveis aleatórias em questão. Neste contexto, a utilização da estimação não-paramétrica vem sendo amplamente empregada em diversas áreas do conhecimento, devido à construção baseada nas observações tornar suas premissas menos restritivas e a natureza da estimação mais flexível e generalista.

David W. Scott, em seu livro (SCOTT, 2015), afirma que os estimadores não-paramétricos de densidades eliminam a necessidade da especificação de um modelo paramétrico a priori. Entretanto, esse benefício vem acompanhado da perda de eficiência na representação da densidade de probabilidade. Esse assunto é conhecido na literatura como “maldição da otimização”, colocando em pauta se essa perda pode ser justificada quando comparada ao risco de uma aplicação errônea de um modelo paramétrico. Neste mesmo livro, Scott elucida que estimadores paramétricos podem se tornar ineficientes com pequenas perturbações em seus parâmetros e aponta a ênfase moderna da estimação robusta, que através da estimação não-paramétrica sacrifica uma pequena porcentagem de eficiência com o intuito de alcançar uma maior insensibilidade ao erro de especificação do

modelo.

Em outras palavras, a estimação não-paramétrica possui uma abordagem diferente da estimação paramétrica. Dada uma função densidade de probabilidade $\phi(\cdot|\theta)$, bem conhecida na literatura, da família normal $N(\mu, \sigma^2)$, onde $\theta = (\mu, \sigma^2)$, a estimação paramétrica tem sua ênfase na obtenção do melhor estimador $\hat{\theta}$ de θ . Sob uma suposição errada do modelo seu viés não pode ser removido apenas com o aumento das observações e muitas vezes determinar se esse viés justifica a manutenção do modelo paramétrico pode ser complicado (FISHER, 1922). No caso não-paramétrico muda-se a perspectiva, visando uma boa estimação $\hat{\phi}(\cdot)$ de $\phi(\cdot)$. Neste trabalho, para estimar $\hat{\phi}(\cdot)$ abordaremos quatro estimadores não-paramétricos: Histograma (SILVERMAN, 1986), *Average Shifted Histogram* (ASH) (SCOTT, 1985a), Polígono de Frequência (PF) (SCOTT, 1985b) e KDE (SILVERMAN, 1986; SCOTT, 2015). Dentro destes tópicos existem duas propostas principais de seletores de largura de banda:

- *Plug-In* (PI): necessitam de um conhecimento a priori da PDF para sua construção. Esse conhecimento pode advir de uma presunção de Normalidade ou de alguma técnica de estimação.
- *Cross-Validation* (CV): utilizam uma estimação por validação-cruzada como valor esperado de alguns termos importantes de sua otimização ou para o cálculo de características da distribuição a ser estimada.

Geralmente os algoritmos relacionados ao Histograma, PF e ASH são explicitados em função do número de *bins* e os algoritmos referentes ao KDE estão em função da largura e banda. Vale notar que essas grandezas são relacionadas e a escolha do termo será em função do seletor estudado.

2.1 HISTOGRAMA

O Histograma é o estimador de densidade mais utilizado e mais antigo, geralmente não é exibido em escala de densidade, sendo muito representado por contagem. Essa diferença foi abordada por (EMERSON; HOAGLIN, 1983). Entretanto (SCOTT, 2015), julga essa distinção superficial, visto que, a informação visual de frequência transmitida pelo histograma possui a essência de uma função de densidade.

Discutir questões referentes a esse simples estimador será de grande valia no desenvolvimento posterior de teorias sobre estimadores mais complexos. Nesta etapa,

discutiremos sua definição, o estado da arte de algumas teorias relacionadas a busca pelo melhor estimador, além da aplicação da validação-cruzada nesse contexto.

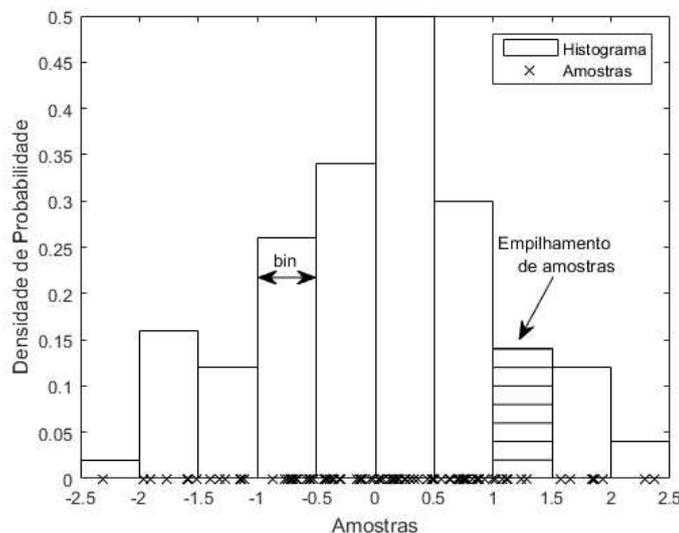
2.1.1 Definição

O histograma é completamente determinado pelas observações, ou amostras $\{x_1, \dots, x_n\}$ de uma função geradora $f(x)$ e a escolha do seu *grid*. Nesse contexto, *grid* representa o intervalo das amostras, subdividido em espaçamentos equidistantes, que são denominados *bins*. Suponhamos o *grid* $\{t_k, -\infty < k < \infty\}$, onde $B_k = [t_k, t_{k+1}]$ denota o *bin*. Suponha que $h = t_k - t_{k+1}$ para todo k . Então esse histograma é dito de banda fixa h . Um histograma de frequência é construído usando blocos de tamanho unitário e largura h “empilhados” nos seus respectivos *bins*. A integral da figura gerada é (nh) , onde n é o número de amostras. Para a construção de um histograma de densidade, os blocos devem ter altura de $1/(nh)$, para que cada bloco tenha área de $1/n$. Sendo I a função indicadora (KENNY *et al.*, 2003), o histograma é definido como:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n I_{[t_k, t_{k+1})}(x_i) \quad \text{Para todo } x \in B_k. \quad (2.1)$$

A Figura 1 traduz graficamente os conceitos mostrados anteriormente através de um histograma criado com 100 amostras normalizado pelo fator (nh) em densidade de probabilidade. Nessa figura é possível verificar o *bin*, cujo tamanho é h , além do empilhamento de cada amostra dentro de um determinado *bin*, que agora possui altura $1/(nh)$.

Figura 1 – Exemplificação da estimação de densidade via Histograma.



Fonte: Elaborada pelo autor (2020).

A estimação *naive* de densidade pelo histograma, como visto, é bastante simples. Entretanto, a análise feita de maneira mais profunda mostra dificuldades que serão vivenciadas, de forma similar, por estimadores de densidade mais complexos. Ou seja, encontrar a largura h^* ótima do *bin* pode ser complexo em determinadas situações como:

- **Função geradora não é Normal e nem outra função parametrizada conhecida:** devido à predominância na literatura de soluções para esses casos e a necessidade de alguns seletores em conhecer características da função geradora para sua otimização.
- **Quantidade de amostras insuficiente para descrever as derivadas e informações relevantes da função geradora:** dificulta o cálculo de parâmetros via validação-cruzada.
- **Outliers:** Valores de amostras muito discrepantes em relação a variável aleatória podem distorcer as variáveis de escala utilizadas em alguns seletores.

Com o intuito de responder a essa demanda, em relação à seleção do número de *bins*, existem algumas teorias na literatura trabalhadas ao longo da história. Na próxima seção serão apresentados seletores clássicos e abordagens mais recentes para o Histograma.

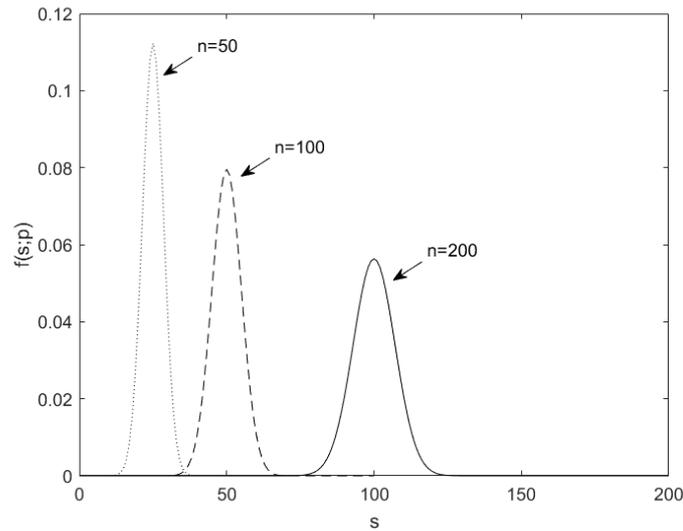
2.1.2 Seletores de Largura de Banda

Regra de Sturges: Tendo a densidade Normal como um ponto de referência, Sturges observou que a distribuição binomial $Bn(n, p = 0,5)$ poderia ser usada como modelo ideal de histograma. A seguir temos a Equação (2.2) da distribuição binomial:

$$f(s; n, p) = \binom{n}{s} p^s (1-p)^{n-s}, \quad (2.2)$$

onde s é o número de sucesso em n tentativas, num processo de Bernoulli com probabilidade de sucesso p . A Figura 2 mostra como a distribuição Binomial se aproxima de uma distribuição Normal quando corretamente dimensionada, com $p = 0.5$.

Para construir um histograma de frequência com N_b *bins*, centrado nos pontos $i = 1, \dots, N_b - 1$. Escolha a contagem do i ésimo *bin* como sendo um coeficiente binomial $\binom{N_b - 1}{i}$, o tamanho total da amostra é dado pela Equação (2.3):

Figura 2 – Distribuição Binomial com $p=0.5$.

Fonte: Elaborada pelo autor (2020).

$$n = \sum_{i=0}^{N_b-1} \binom{N_b-1}{i} = (1+1)^{N_b-1} = 2^{N_b-1} \quad (2.3)$$

Portanto, através da expansão binomial deriva a regra de Sturges (STURGES, 1926), mostrada na Equação (2.4):

$$N_b = 1 + \log_2 n, \quad (2.4)$$

onde n é o número de amostras e N_b é o número de *bins*. Essa regra foi idealizada para aplicar em histogramas de largura de banda h fixa.

Regra de Doane: Em 1976, Doane (DOANE, 1976) modifica a Regra de Sturges para adequá-la ao mundo real, onde predominam os casos em que as PDFs são assimétricas. Doane propõe adicionar um termo K_e à Regra de Sturges, inserindo o efeito da distorção do terceiro momento (Assimetria ou *Skewness*) κ_1 , que é definido pela Equação (2.5):

$$\kappa_1 = \frac{\sum (X-\bar{X})^3}{\left[\sum (X-\bar{X})^2\right]^{3/2}}, \quad (2.5)$$

onde X é uma variável aleatória e \bar{X} é a média dessa variável. O termo extra K_e adicionado a regra de Sturges é dado pela Equação (2.6):

$$K_e = \log_2 \left(1 + \frac{|\kappa_1|}{\sigma\kappa_1} \right) \quad (2.6)$$

Por fim, a Regra de Doane assume a forma mostrada na Equação (2.7):

$$N_b = 1 + \log_2(n) + \log_2 \left(1 + \frac{|\kappa_1|}{\sigma\kappa_1} \right) \quad (2.7)$$

onde N_b é o número de *bins*, n é o número de amostras, σ representa o desvio padrão, κ_1 é a assimetria e $\sigma\kappa_1 = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}$

Regra de Scott: Scott aponta em seu livro (SCOTT, 2015) um Teorema como base para sua regra, que obedece a minimização da norma L^2 . Esse teorema afirma que, se uma função ϕ possui uma derivada contínua, e integral de primeira derivada ao quadrado, então seu *Mean Integrated Square Error* (MISE) assintótico é dado pela Equação (2.8): (Mais informações sobre o cálculo do *Asymptotic Mean Integrated Square Error* (AMISE) podem ser encontradas no Apêndice A).

$$AMISE(h) = \frac{1}{nh} + \frac{1}{12}h^2R(\phi') \quad (2.8a)$$

$$h^* = \left(\frac{6}{R(\phi')} \right)^{1/3} n^{-1/3} \quad (2.8b)$$

$$AMISE^* = \left(\frac{3}{4} \right)^{2/3} R(\phi')^{1/3} n^{-2/3}, \quad (2.8c)$$

onde n é o número de amostras, h é a largura de banda ou do *bin* e $R(\phi)$ é a rugosidade da função densidade de probabilidade ϕ , sendo determinada pela Equação (2.9):

$$R(\phi) = \int \phi(x)^2 dx \quad (2.9)$$

Scott afirma que essa é apenas uma das medidas de rugosidade possíveis, baseada na norma L^2 , entretanto existem $R(\phi')$ e $R(\phi'')$. Essa definição é utilizada na perspectiva estatística levando em conta a primeira derivada da função, ou seja, as inclinações das funções. Como exemplo, essa definição descreveria a função Normal como pouco rugosa e a função Log-Normal como bastante rugosa, do ponto de vista estatístico.

A regra de Scott (SCOTT, 1979) é de simples aplicação, para o caso especial da distribuição Normal, pois utiliza apenas o desvio padrão σ das amostras e o número de amostras n , fazendo a substituição de $R(\phi') = 1/4\sqrt{\pi}\sigma^3$ em 2.8b temos a Equação (2.10):

$$h = (24\sqrt{\pi}\sigma^3/n)^{1/3} \approx 3.5\sigma n^{-1/3} \quad (2.10)$$

Regra de Freedman e Diaconis: Freedman-Diaconis (FD) (FREEDMAN; DIACONIS, 1981) propuseram uma regra mais robusta estatisticamente, alterando o parâmetro σ , que poderia sofrer com *outliers*, por um múltiplo do *Interquartile Range* (IQR) (WAN *et al.*, 2014). Sua fórmula é mostrada na Equação (2.11):

$$h = 2(IQR)n^{-1/3} \quad (2.11)$$

Método de Rudemo O método de Rudemo é calculado através da norma L^2 . A métrica *Integrated Square Error* (ISE) utilizada para minimização é mostrada na Equação (2.12):

$$ISE = \int (\hat{\phi}(x) - \phi(x))^2 dx \quad (2.12)$$

Ao expandir a equação (2.12) os termos de interesse se tornam explícitos, como visto na Equação (2.13):

$$ISE = \int \hat{\phi}(x)^2 dx - 2 \int \hat{\phi}(x)\phi(x) dx + \int \phi(x)^2 dx \quad (2.13)$$

O último termo, como não depende da estimação do número de *bins*, é desconsiderado da minimização. O segundo termo é chamado de “termo problemático” na literatura devido a presença de $\phi(x)$, e grande parte dos seletores baseados na norma L^2 visam estimar esse termo. O seletor de Rudemo estima o “termo problemático” segundo a Equação (2.14), como:

$$\int \hat{\phi}(x)^2 dx - \frac{2}{n} \sum_{i=1}^n \hat{\phi}_{-i}(x_i) \quad (2.14)$$

Para um histograma com contagem C_k em cada *bin* k a função custo $Q(h)$ de Rudemo (RUDEMO, 1982) é computado através da Equação (2.15):

$$Q(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)n^2h} \sum_k C_k^2 \quad (2.15)$$

Wand: Este método pode ser encontrado em (WAND, 1997), juntamente com uma proposta de implementação. Nessa proposta Wand propõe substituir a estimação do “termo problemático”, simbolizado por $\hat{\psi}_r(h)$, do MISE por uma aproximação. Dado $G_1 = \min(X_i) \leq G_2 \leq \dots \leq G_k = \max(X_i)$, para um intervalo espaçado igualmente $\delta = (G_k - G_1)/(M - 1)$ por todo o intervalo das amostras X . A aproximação linear para $\hat{\psi}_r(h)$ é dada pela equação 2.16:

$$\tilde{\psi}_r(h) = n^{-2} \sum_{j=1}^M \left(\sum_{j'=1}^M C_{j'} k_{j'-j}^{(r)} \right) C_j, \quad (2.16)$$

onde C_j é definido pela equação (2.17):

$$C_j = \sum_{i=1}^n (1 - |\delta^{-1} X_i - j|)_+ \quad (2.17)$$

$[x_+ = \max(0, x)]$ e C_j é a contagem no ponto G_j e:

$$k_j^{(r)} = h^{-r-1} L^{(r)}(\delta_j/h), \quad |j| = 0, \dots, M., \quad (2.18)$$

onde $L(x) = (2\pi)^{-1/2} e^{-x^2/2}$, $\hat{\sigma} = \min(\sigma, IQR/1.349)$ e $M = 400$.

Portanto, os dois estágios h_1 e h_2 propostos por Wand são definidos pelas Equações (2.17) e (2.20):

$$h_1 = \left\{ \frac{6}{-\tilde{\psi}_2(g_{11})n} \right\}^{1/3}, \quad (2.19)$$

onde $g_{11} = \{2/(3n)\}^{1/5} 2^{1/2} \hat{\sigma}$.

$$h_2 = \left\{ \frac{6}{-\tilde{\psi}_2(g_{21})n} \right\}^{1/3}, \quad (2.20)$$

onde $g_{21} = \left[2 / \left\{ (2\pi)^{1/2} \tilde{\psi}_4(g_{22})^{1/5} n \right\} \right]^{1/5} 2^{1/2} \hat{\sigma}$ e $g_{22} = \{2/(5n)\}^{1/7} 2^{1/2} \hat{\sigma}$.

Knuth: De acordo com Knuth (KNUTH, 2006), em problemas de otimização envolvendo densidade de probabilidade, geralmente é mais fácil maximizar a verossimilhança através do logaritmo. No caso deste método a equação a ser otimizada é definida pela Equação (2.21):

$$\begin{aligned} \log(\phi|x, I) = n \log(N_b) + \log \Gamma\left(\frac{N_b}{2}\right) \dots \\ - N_b \log \Gamma\left(\frac{1}{2}\right) - \log \Gamma\left(\frac{2n+N_b}{2}\right) + \sum_{k=1}^K \log \Gamma\left(C_k + \frac{1}{2}\right), \end{aligned} \quad (2.21)$$

onde N_b representa número de *bins* que maximiza a função ϕ , n é o número de amostras, e C_k é o número de contagem dentro do *bin* B_k .

Método de Shimazaki-Shinomoto: Shimazaki e Shinomoto (SHIMAZAKI; SHINOMOTO, 2007) propuseram um método para estimar a largura do *bin* de um histograma, denominado *Shimazaki-Shinomoto* (SS). Seu método utiliza a norma L^2 através do MISE e sua adaptação ao universo discreto desenvolve-se da seguinte forma:

1. Dividir o intervalo das variáveis aleatórias em N *bins* de tamanho h . Contar o número de eventos C_k que estão compreendidos dentro do $k^{\text{ésimo}}$ *bin*.
2. Calcular a média μ e variância σ^2 do número de amostras, através das Equações (2.22) e (2.23):

$$\mu \equiv \frac{1}{N} \sum_{i=1}^N C_k \quad (2.22a)$$

$$\sigma^2 \equiv \frac{1}{N} \sum_{i=1}^N (C_k - \mu)^2 \quad (2.22b)$$

3. A função custo Z será dada por:

$$Z(h) = \frac{2\mu - \sigma^2}{h^2} \quad (2.23)$$

Os autores afirmam que o método foi derivado do processo de Poisson e não através da validação-cruzada. Os resultados da largura de banda escolhidas pelos dois métodos, Poisson e CV, são diferentes.

LHM: De acordo com Lolla e Hoberock (LOLLA; HOBEROCK, 2011), resumidamente, o método LHM consiste em definir métricas de erro em quantidades observáveis ou computáveis de um conjunto de amostras X , para então selecionar o número de *bins* através do balanço entre erro e o custo computacional. A métrica de erro é definida pela Equação (2.24):

$$E = \sum_{i=1}^n ECDF(x_i) - CDF(x_i), \quad (2.24)$$

onde, *Empirical Cumulative Distribution Function* (ECDF) representa a função cumulativa empírica do conjunto de amostras X , e *Cumulative Distribution Function* (CDF) é a função cumulativa construída através da interpolação *nearest* ou linear, variando o número de *bins*. De acordo com o Teorema de Glivenko-Cantelli (STUTE; SCHUMANN, 1980), ambas as aproximações da CDF irão convergir para a ECDF quando o número de *bins* aumentar. Portanto, a curva de erro irá convergir para zero com o aumento do número de *bins*, onde existe um ponto na curva que representa o equilíbrio entre erro e custo computacional. O ponto é encontrado via *Bayesian Information Criterion* (BIC), descrito em (ZHAO; XU; FRÄNTI, 2008).

2.2 POLÍGONOS DE FREQUÊNCIA

Apesar de sua reconhecida utilidade, os histogramas possuem descontinuidades que inviabilizam alguns cálculos matemáticos, além de limitar a visualização gráfica no contexto multivariado. O PF é um estimador de densidade contínua que utiliza como base o histograma, entretanto a ligação entre os pontos estimados é feita por uma interpolação linear. Em 1985, Scott (SCOTT, 1985b) examinou a teoria dos polígonos de frequência univariado e multivariado e encontrou melhorias consideráveis em relação ao histograma.

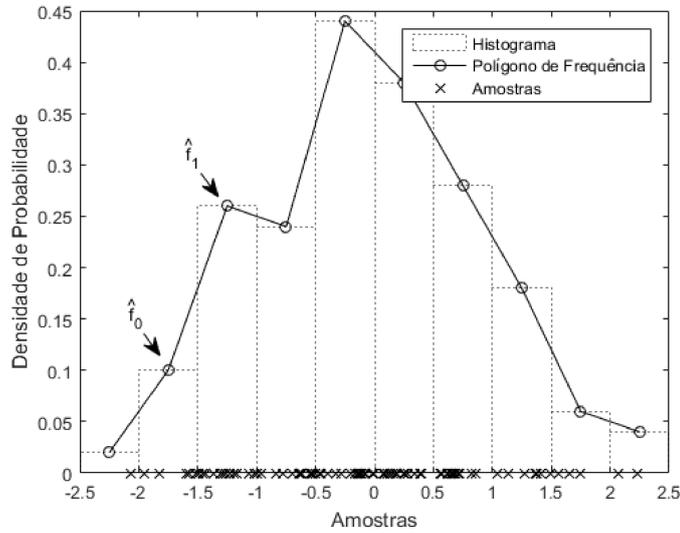
O Polígono de frequência conecta dois valores adjacentes de histogramas \hat{f}_0 e \hat{f}_1 entre os centros dos *bins*, como mostra a Figura 3. Sua definição é apresentada na Equação (2.25):

$$\hat{f}(x) = \left(\frac{1}{2} - \frac{x}{h}\right) \hat{f}_0 + \left(\frac{1}{2} + \frac{x}{h}\right) \hat{f}_1 \quad \text{Onde } -\frac{h}{2} \leq x < \frac{h}{2} \quad (2.25)$$

De acordo com Scott, o Polígono de Frequência consegue estimar melhor a PDF nas regiões onde a derivada é maior, se aproximando melhor de uma função contínua. Entretanto, se comporta pior nos picos, onde a segunda derivada e as magnitudes de densidade são maiores. Além disso, nas regiões onde existem descontinuidades o PF se comporta pior, sobrepondo esses pontos. Já o histograma é imune a esses efeitos na presença de descontinuidades. O PF encontra seu valor mínimo de AMISE utilizando *bins* de larguras maiores do que o Histograma, como pode ser visto na Figura 4.

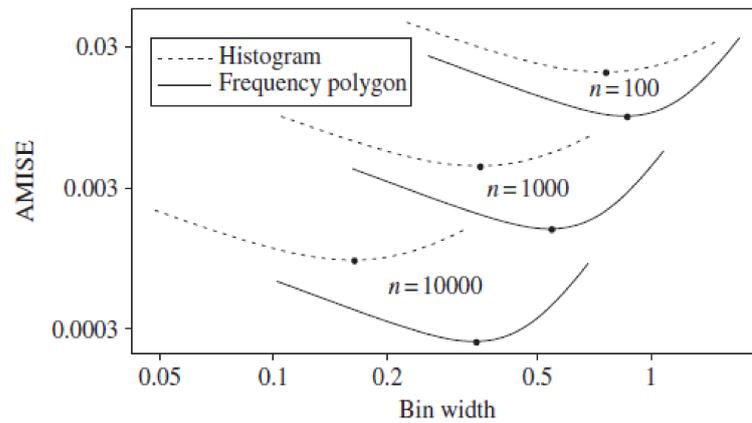
Portanto, o cálculo da largura do *bin* para o PF é refeito a partir de um Teorema encontrado em (SCOTT, 2015). Esse Teorema diz que sob uma suposição de ϕ'' absolutamente contínua e a rugosidade $R(\phi''') < \infty$, temos:

Figura 3 – Exemplificação da estimação de densidade via Polígono de Frequência.



Fonte: Elaborada pelo autor (2020).

Figura 4 – AMISE para Histograma e PF em uma distribuição Normal



Fonte: (SCOTT, 2015)

$$AMISE(h) = \frac{2}{3nh} + \frac{49}{2880}h^4R(\phi''') \tag{2.26a}$$

$$h^* = 2[15/(49R(\phi''))]^{1/5}n^{-1/5} \tag{2.26b}$$

$$h^* = (5/12)[49R(\phi'')/15]^{1/5}n^{-4/5} \tag{2.26c}$$

Supondo uma PDF Normal, $R(\phi'') = 3/(8\sqrt{\pi}\sigma^5)$, substituindo em (2.26b) temos:

$$h = 2.15\sigma n^{-1/5} \tag{2.27}$$

2.3 ASH

O histograma médio deslocado foi proposto como alternativa para solucionar um problema existente no histograma, causado pela variação do formato da densidade estimada, devido a escolha do início do *grid*. O método, basicamente, faz a média de vários PFs deslocados, gerando uma estimação final com aparência de polígono que pode ser denominada de *Average Shifted Frequency Polygon* (ASFP). Outro método mais simples é fazer o mesmo processo com o histograma, deslocando-os e realizando a média de suas densidades de probabilidade, o resultado final também será um histograma e pode ser chamado de ASH. Neste trabalho será utilizado somente a média de PFs e para facilitar a nomenclatura relativa a outros trabalhos, ambos casos serão chamados de ASH.

Para construir um estimador ASH considere a coleção de m histogramas, $\widehat{f}_1, \widehat{f}_2, \dots, \widehat{f}_m$ todos com largura de *bin* h e com a origem dada pela Equação (2.28):

$$t_0 = 0, \frac{h}{m}, \frac{2h}{m}, \dots, \frac{(m-1)h}{m} \quad (2.28)$$

Portanto, ASH será definido pela Equação (2.29):

$$\widehat{f}(\cdot) = \widehat{f}_{ASH}(\cdot) = \frac{1}{m} \sum_{i=1}^m f_i(\cdot) \quad (2.29)$$

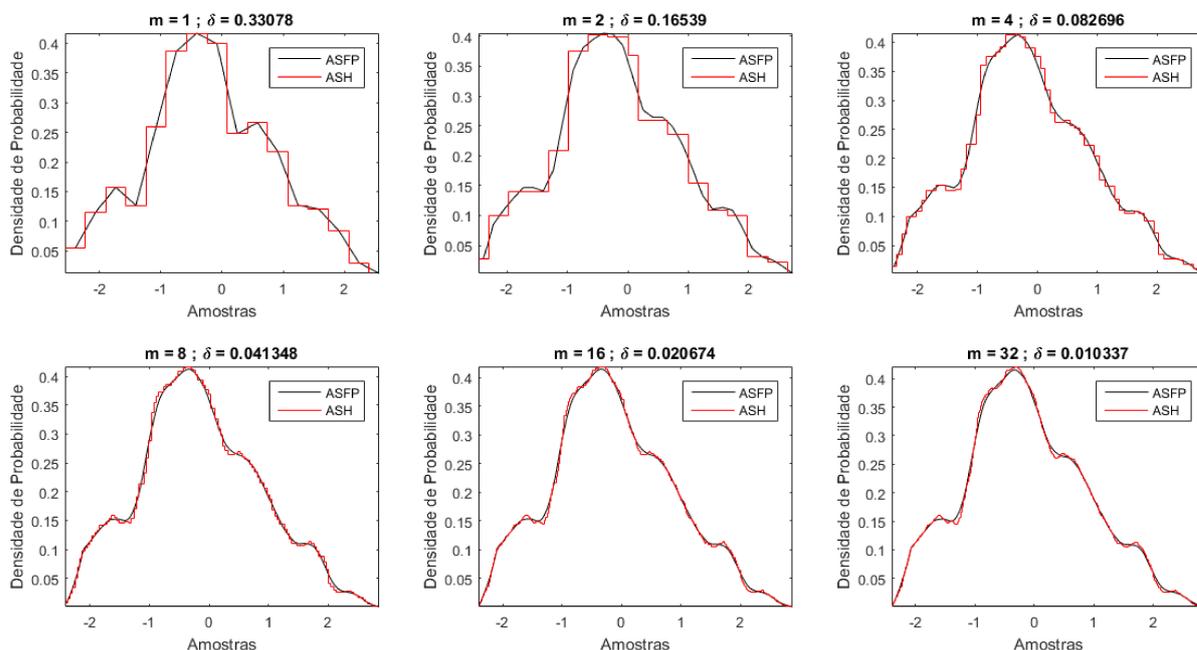
Agora, a largura final do *bin* será dada por $\delta = h/m$. A Figura 5 exemplifica o efeito do aumento de m . É interessante referir-se ao intervalo $[k\delta, (k+1)\delta]$ como intervalo do *bin* B_k , e dado C_k como a contagem no *bin* B_k , temos a expressão geral do ASH, normalizada por nh , como mostra a Equação (2.30):

$$\begin{aligned} \widehat{f}(x; m) &= \frac{1}{m} \sum_{i=1-m}^{m-1} \frac{(m-|i|)C_{k+i}}{nh} \\ &= \frac{1}{nh} \sum_{i=1-m}^{m-1} \left(1 - \frac{|i|}{m}\right) C_{k+i} \\ &= \frac{1}{nh} \sum_{|i| < m} (w_m(i)) C_{k+i} \end{aligned} \quad (2.30)$$

Portanto, para encontrar o h do método ASH Scott utiliza um Teorema (SCOTT, 2015) que permite o cálculo do AMISE com uma função de peso do triângulo isósceles. Essa expressão é dada pela Equação (2.31):

$$AMISE = \frac{2}{3nh} \left(1 + \frac{1}{2m^2}\right) + \frac{h^2}{12m^2} R(\phi') + \frac{h^4}{144} \left(1 - \frac{2}{m^2} + \frac{3}{5m^4}\right) R(\phi'') \quad (2.31)$$

Figura 5 – ASH e ASFP calculados para uma distribuição normal, com diversas variações de m .



Fonte: Elaborada pelo autor (2020).

Novamente, considerando a rugosidade da distribuição Normal, o h para o ASH é definido através da Equação (2.32):

$$h_{m=\infty} = [24/(nR(f''))]^{1/5} = 2.576\sigma n^{-1/5} \quad (2.32)$$

Pela Equação (2.31) é possível notar que os dois primeiros termos são referentes ao histograma, quando $m=1$. Por outro lado, quando $m \rightarrow \infty$ o segundo termo de viés do histograma desaparece e se assemelha ao PF. Geralmente, para $m \geq 10$, o termo do meio se torna insignificante comparado ao último termo. Em (SCOTT, 1985a), para um número moderados de amostras, foi dito que $5 \leq m \leq 10$ é suficiente para o cálculo do ASH. Em um teste com 100 amostras, com distribuição Normal, o aumento de eficiência na estimação para $m \geq 13$ se mostrou menor do que 1%, entretanto, o custo computacional aumentou consideravelmente com o aumento de m .

2.4 KDE

Atualmente, o estimador de densidades KDE tornou-se uma ferramenta comum para estudos empíricos em diversas áreas de pesquisas. Discussões e diversos estudos sobre esse tipo de estimador, com enfoque principalmente em otimizações da largura de banda h , tem sido muito presente na literatura. Apesar de grande parte desses estudos ter sua atenção voltada para regressão não-paramétrica, a escolha da largura de banda

em estimação de densidades é igualmente problemática. Essa dificuldade se traduz em afirmações como a de (HEIDENREICH; SCHINDLER; SPERLICH, 2013), de que pode não haver um processo perfeito para selecionar a largura de banda ótima. Entretanto, ainda assim, pode-se dizer qual seletor de largura de banda é considerado razoável para uma determinada aplicação.

A literatura fornece inúmeros trabalhos dedicados aos seletores de largura de banda para o método KDE, parâmetro responsável pelo ajuste entre viés e variância da estimação não-paramétrica. Entretanto, existe um número menor de trabalhos relativos a comparação entre esses seletores, que podem ser classificados principalmente em *Plug-in* (PI) e *Cross Validation* (CV). O primeiro faz algumas suposições a priori para definir parâmetros desconhecidos ou os estimam utilizando dados experimentais para, matematicamente, definir um valor de largura de banda. O último realiza essa tarefa otimizando uma função custo com base nos dados experimentais.

Entre os estudos comparativos considerando métodos baseados em PI e CV, em 1996, Jones, Marron e Sheather (JONES; MARRON; SHEATHER, 1996a; JONES; MARRON; SHEATHER, 1996b) avaliaram o desempenho de alguns métodos, indicando uma preferência pelo método *Sheather and Jones* (SJ) (SHEATHER; JONES, 1991) em relação aos chamados métodos de primeira geração, representados pelos métodos *Silverman* (SV) (SILVERMAN, 1986) e *Scott* (SC) (SCOTT, 1992) e os métodos clássicos de CV conhecidos como *Biased Cross-Validation* (BCV) e *Unbiased Cross-Validation* (UCV) (SCOTT; TERRELL, 1987). Os autores apontam que o UCV, também denominado *Least Square Cross-Validation* (LSCV), apresenta resultados insatisfatórios pois geralmente subestimam a largura de banda, fornecendo estimativas de alta variância, enquanto o BCV tende a superestimá-la, oferecendo estimativas com alto viés.

A conclusão sobre o UCV coincide com o trabalho de Cao, Cuevas e Manteiga (CAO; CUEVAS; MANTEIGA, 1994) de 1994. No entanto, neste trabalho, o desempenho do BCV foi comparável ao SJ. Este último trabalho reforça, como detectado em outros trabalhos (HALL; MARRON, 1987; HALL; MARRON *et al.*, 1987), o fraco desempenho do método *Maximum Likelihood Cross-Validation* (MLCV) (HABBEMA, 1974), principalmente ao lidar com distribuições que possuem caudas bastante longas.

Ainda em 1996, Chiu (CHIU, 1996), usando o critério de viés, mostrou que os métodos estabilizados de CV tem “melhor desempenho” quando comparados aos métodos padrões de CV, que apresentam maior variação e geralmente subestimam a largura de banda. Duas conclusões importantes puderam ser destacadas: (1) para densidades que

não são muito diferentes da densidade Normal o SJ obteve bom desempenho e mesmo para funções de densidades diferentes da Normal raramente fornece larguras de banda muito pequenas, e não superestima severamente a largura de banda; e (2) o BCV de Scott e Terrell não obteve bom desempenho com poucas amostras e também falhou em algumas misturas de densidades mesmo com muitas amostras.

Um extenso estudo foi desenvolvido em 1997 por Luc Devroye (DEVROYE *et al.*, 1997). Este trabalho avaliou o desempenho de 17 seletores de largura de banda em 28 funções de densidade. Para cada função, foram gerados 20 conjuntos de dados, cada um com 100 amostras. Nesse contexto, o autor afirmou que, em geral, os métodos PI têm uma maior probabilidade de superestimar o valor da largura de banda em comparação aos métodos CV. No entanto, os métodos PI alcançaram melhor desempenho para densidades unimodais suaves. Neste mesmo trabalho, o chamado método aprimorado *L1 Improved* (L1I), baseado em PI, obteve um erro de estimativa tão baixo quanto alguns dos melhores métodos de CV, para algumas densidades multimodais, alcançando o melhor desempenho geral.

Em 1999, o trabalho do Loader (LOADER *et al.*, 1999) representou uma mudança de paradigma em relação aos trabalhos anteriores, pois percebeu a maior variância das estimações baseadas em CV, em relação as estimações de PI, como algo positivo. Loader afirmou que os métodos PI tendem a superestimar a largura de banda e, portanto, podem perder informações ao não serem capazes de identificar certas variações que fazem parte de um determinado processo. Os métodos CV seriam então mais confiáveis nesse contexto. Loader também chamou a atenção para a suscetibilidade dos métodos PI à necessidade de suposições a priori para o cálculo da largura de banda. O trabalho de Loader testou cinco seletores de largura de banda em seis densidades diferentes usando 1000 conjuntos de dados, cada um com cerca de 100 a 200 amostras.

Em 2004, Sheather (SHEATHER, 2004) traz para o centro da discussão o problema dos métodos UCV e BCV com mais de um mínimo local. Esse problema foi solucionado em trabalhos anteriores. Hall e Marron estudaram o caso da UCV em 1991 (HALL; MARRON, 1991), mostrando que mínimos locais espúrios tendem a ocorrer no lado inferior da largura de banda ideal, e não no lado superior. Em 1996, Jones, Marron e Sheather (JONES; MARRON; SHEATHER, 1996a) aconselharam a usar o maior mínimo local para o método UCV e o menor para o BCV. Além disso, em (SHEATHER, 2004) Sheather lembra que, para densidades com baixa rugosidade, o método SJ tende a ser melhor que o UCV, enquanto, para aqueles com alta rugosidade, SJ tende a superestimar a largura de banda.

No entanto, o autor sugere a escolha do mínimo local do UCV mais próximo da largura de banda selecionada pelo SJ, com base nas informações de que esse método teria um bom desempenho geral. Por fim, Sheather recomenda, em qualquer caso, sempre verificar o gráfico produzido pelo processo de minimização do UCV, em vez de simplesmente confiar no resultado de uma rotina automática.

Em 2009, Liu, Yang, Webb e Boughton (LIU *et al.*, 2009) propuseram um estudo de comparação usando nove seletores de largura de banda para avaliar a relação entre os desempenhos de estimativa e classificação. Além de indicar a necessidade de mais estudos sobre o assunto, os autores evocaram uma declaração de Friedman de 1997 (FRIEDMAN, 1997) afirmando que maior precisão na estimativa de densidade não leva necessariamente a um melhor desempenho de classificação.

Em uma revisão de 2013, Heidenreich, Schindler e Sperlich (HEIDENREICH; SCHINDLER; SPERLICH, 2013) examinaram a evolução dos seletores de largura de banda atualizados e compararam o desempenho de oito métodos diferentes, com base em um conjunto de seis distribuições sem descontinuidades e caudas exponencialmente pesadas. Neste trabalho comparativo foram considerados conjuntos de 25 a 200 amostras. Além disso, os autores destacaram um método desenvolvido por Hart e Yi chamado *One Sided Cross-Validation* (OSCV) (HART; YI, 1998), afirmando que esse método deve ser utilizado quando não se souber nada sobre a densidade por trás do processo de geração do conjunto de dados. Heidenreich, Schindler e Sperlich também mostraram que uma mistura entre técnicas de PI e CV pode produzir seletores de largura de banda fixa estáveis.

Em 2017, Borrajo, Manteiga e Miranda (BORRAJO; GONZÁLEZ-MANTEIGA; MARTÍNEZ-MIRANDA, 2017) fizeram uma comparação entre cinco seletores de largura de banda usando seis distribuições diferentes, cobrindo uma ampla gama de densidades com diferentes complexidades, mostrando que os métodos CV obtiveram melhor desempenho para densidades muito complexas.

Para melhor compreender os seletores citados acima, bem como suas características e performance em diferentes tipos de densidades de probabilidade, veremos uma breve definição sobre o método KDE e logo após serão apresentadas as equações de cada um dos seletores que tiveram seus algoritmos implementados nesse estudo.

2.4.1 Definição

Suponha uma variável aleatória X com sua função geradora $f(x)$. Parzen–Rosenblatt (PARZEN, 1962) definiram a fórmula padrão do estimador $\hat{f}(x_i)$ chamado de KDE. Dedu-

ções mais detalhadas podem ser encontradas em (SILVERMAN, 1986) e (SCOTT, 2015). A fórmula geral do KDE é dada pela Equação (2.33):

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right), \quad (2.33)$$

onde, n é o número de amostras, h é a largura de banda e K é a função *Kernel*.

Função Kernel Uma função *kernel* $K(x) : \mathbb{R} \rightarrow \mathbb{R}$ é qualquer função que satisfaça as seguintes premissas:

- $\int_{-\infty}^{+\infty} K(x) dx = 1$;
- Não-negatividade: satisfaz $K(x) \geq 0$, para todo x . Neste caso, $K(x)$ é uma função densidade de probabilidade;
- Simetria: $K(x) = K(-x)$, para todo x ;

Sabendo que I representa a função indicadora (KENNY *et al.*, 2003), a seguir serão apresentados alguns exemplos de funções *kernel*:

- Triangular: $K(x) = (1 - |x|)I(|x| \leq 1)$
- Epanechnikov: $K(x) = \frac{3}{4}(1 - x)I(|x| \leq 1)$
- Quartic (Biweight): $K(x) = \frac{15}{16}(1 - x)I(|x| \leq 1)$
- Triweight: $K(x) = \frac{35}{32}(1 - x)I(|x| \leq 1)$
- Gaussiana: $K(x) = \frac{1}{\sqrt{2\pi}}e^{(-\frac{1}{2}x)}$

Se considerarmos $u = (x - X_i)$ e $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$ temos o estimador definido pela Equação (2.34):

$$\widehat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i), \quad (2.34)$$

onde, n é o número de amostras, x é a posição na variável aleatória que desejamos estimar a probabilidade, X representa a variável aleatória e K_h é a função *kernel* com alguns termos implícitos. Nessa tese todas as funções *kernel* utilizadas na estimação de densidade serão Gaussianas, exceto para o método *L1 improved* (L1I), que será Epanechnikov. Mais detalhes serão explicitados na Seção 2.4.2.1.

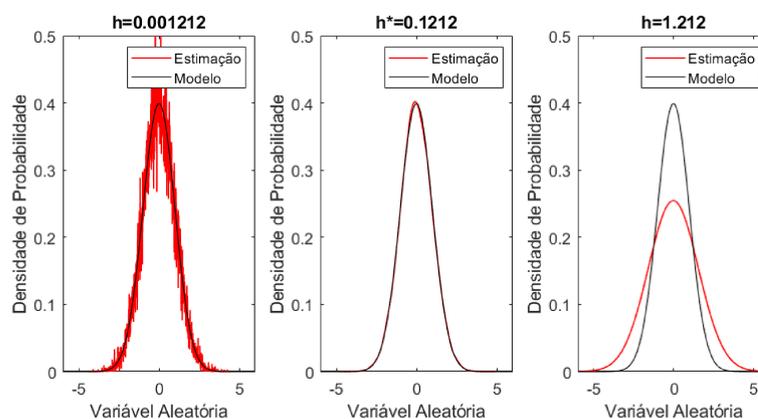
A busca pela otimização da largura de banda h é responsável pela maior parte dos esforços dos estatísticos em estimação de densidades não-paramétricas. A largura de banda h pode ser fixa ou variável, e controla a suavização da estimativa de probabilidade, sendo sua escolha de fundamental importância no desempenho de uma boa estimação.

2.4.2 KDE com Largura de Banda Fixa

Nesse tipo de método será escolhida uma largura banda, para toda a estimação da densidade de probabilidade, baseada nas amostras ou em alguma suposição a priori. Ou seja, independente da posição em que a estimação for feita, em relação a variável aleatória, o h será mantido o mesmo.

A Figura 6 mostra a variação da largura de banda h em três valores distintos: a figura da esquerda mostra uma estimação com alta variância e valor pequeno de h ; a figura central mostra a largura de banda ótima h^* , calculada pela Regra de Silverman (SILVERMAN, 1986), como uma estimação mais “próxima” do modelo analítico e com uma “boa” compensação entre viés e variância; por fim, a figura da direita mostra uma largura de banda h muito grande, e com isso a estimação do KDE não consegue descrever as maiores derivadas, se mostrando bastante suave e com um viés elevado.

Figura 6 – Impacto da variação de h de banda fixa na estimação da PDF.



Fonte: Elaborada pelo autor (2020).

Geralmente os minimizadores utilizados pelos seletores de largura de banda estão em função de h . Os parâmetros do ISE são encontrados para $h > 0$, e geralmente é utilizado nos seletores de CV. O MISE é uma função determinística de h , geralmente utilizado em seletores PI. Porém, existem exceções nas minimizações dos dois casos. A crítica geral pode ser resumida da seguinte forma: seletores baseados em CV levam a uma baixa suavização e são conhecidos por dificilmente estabilizar para grandes conjuntos de amostras; seletores PI possuem as melhores propriedades assintóticas comparada à

validação-cruzada (HALL; MARRON, 1987). Entretanto, tende a superestimar a largura de banda em distribuições com rugosidade maior do que a Normal.

Nos tópicos a seguir serão apresentados diversos seletores baseados em CV e PI. Dentre eles, alguns possuem a capacidade de calcular a largura de banda tanto para a densidade de probabilidade quanto para suas derivadas.

2.4.2.1 Seletores de largura de banda

Maximum Likelihood Cross-Validation (MLCV): O seletor proposto por Habbema (HABBEMA, 1974) é conhecido como MLCV e tem como objetivo escolher h para que a *pseudo-likelihood* $\prod_{i=1}^n \hat{f}_h(X_i)$ seja maximizada. O método de validação-cruzada utilizado para calcular $\hat{f}_{h,i}(X_i)$ será o *leave-one-out* (EFRON, 1982), que substituirá $\hat{f}_h(x)$, dada pela Equação (2.35):

$$\hat{f}_{h,i}(X_i) = \frac{1}{(n-1)h} \sum_{\substack{j=1 \\ j \neq i}}^n K\left(\frac{X_j - X_i}{h}\right) \quad (2.35)$$

Portanto, deve-se escolher um h para aproximar do máximo finito de $h_{mlcv} = \arg \max_{h>0} MLCV(h)$. A função a ser maximizada, do seletor MLCV, é definida pela Equação (2.36):

$$MLCV(h) = \left(n^{-1} \sum_{i=1}^n \log \left[\sum_{j \neq i} K\left(\frac{X_j - X_i}{h}\right) \right] - \log [(n-1)h] \right) \quad (2.36)$$

onde, n representa o número de amostras, K é a função *kernel*, X é a variável aleatória e h a largura de banda.

Unbiased Cross-Validation (UCV): O seletor proposto por Rudemo e Bowman (RUDEMO, 1982; BOWMAN, 1984) é provavelmente o método mais popular e estudado nesse contexto. Uma adaptação estendendo a formulação para derivadas de r^{th} ordem foi proposta por (HARDLE; MARRON; WAND, 1990). Basicamente, o objetivo principal do método é minimizar o $ISE(h)$. O critério de minimização é $h_{mlcv} = \arg \min_{h>0} UCV(h, r)$, onde UCV é definido pela Equação (2.37):

$$UCV(h, r) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{(-1)^r}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \left(K^{(r)} * K^{(r)} - 2K^{(2r)} \right) \left(\frac{X_j - X_i}{h} \right) \quad (2.37)$$

onde, R é a rugosidade, r é a ordem da derivada, n representa o número de amostras, K é a função *kernel*, X é a variável aleatória e h a largura de banda.

Biased Cross-Validation (BCV): Este seletor foi proposto por Scott e Terrell (SCOTT; TERRELL, 1987) e tem como objetivo minimizar o AMISE. Neste método foi proposto estimar $R(\phi^{(r+2)})$ como mostra a Equação (2.38):

$$R(\phi^{(r+2)}) = R(\phi_h^{(r+2)}) - \frac{R(K^{(r+2)})}{nh^{2r+5}} \quad (2.38)$$

Existem duas versões do BCV, que se distinguem através de duas formas de estimar $R(\phi^{(r+2)})$. Pode ser usada a teoria proposta por Scott (SCOTT; TERRELL, 1987), como apresentado na Equação (2.39):

$$R(\phi^{(r+2)}) = \frac{(-1)^{r+2}}{n(n-1)h^{2r+5}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \left(K^{(r+2)} * K^{(r+2)} \right) \left(\frac{X_j - X_i}{h} \right) \quad (2.39)$$

Ou, $R(\phi^{(r+2)})$ pode ser estimado de acordo com Jones (JONES; KAPPENMAN, 1992), de acordo com a Equação (2.40):

$$R(\phi^{(r+2)}) = \frac{(-1)^{r+2}}{n(n-1)h^{2r+5}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \left(K^{(2r+4)} \right) \left(\frac{X_j - X_i}{h} \right) \quad (2.40)$$

Portanto, é possível definir o critério de minimização, de *Biased Cross-Validation 1* (BCV1) e *Biased Cross-Validation 2* (BCV2), para $R(\phi^{(r+2)})$ conforme as Equações (2.41) e (2.42):

$$BCV_1(h, r) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{\mu_2^2(K)}{4} \frac{(-1)^r}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \left(K^{(r+2)} * K^{(r+2)} \right) \left(\frac{X_j - X_i}{h} \right) \quad (2.41)$$

$$BCV_2(h, r) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{\mu_2^2(K)}{4} \frac{(-1)^r}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \left(K^{(2r+4)} \right) \left(\frac{X_j - X_i}{h} \right) \quad (2.42)$$

onde, ϕ representa a função de uma densidade qualquer, R é a rugosidade, r é a ordem da derivada, μ_2^2 é o segundo momento ao quadrado, n representa o número de amostras, K é a função *kernel*, X é a variável aleatória e h representa a largura de banda.

Complete Cross-Validation (CCV): Jones e Kappenman (JONES; KAPPENMAN, 1992) propuseram o método de CCV para a largura de banda em estimações via KDE. Este seletor pode ser estendido para encontrar o h responsável por estimar as derivadas, de

acordo com o trabalho de Hall (HALL; MARRON *et al.*, 1987). O critério de minimização é definido pela Equação (2.43):

$$CCV(h, r) = R(\phi_h^{(r)}) - \bar{\theta}_r(h) + \frac{1}{2}\mu_2(K)h^2\bar{\theta}_{r+1}(h) + \frac{1}{24}(6\mu_2^2(K) - \xi(K))h^4\bar{\theta}_{r+2}(h) \quad (2.43)$$

onde, ϕ representa a função de uma densidade qualquer, R é a rugosidade, r é a ordem da derivada, μ_2^2 é o segundo momento ao quadrado, n representa o número de amostras, K é a função *kernel*, X é a variável aleatória e h representa a largura de banda. Além disso, os parâmetros $R(\phi_h^{(r)})$, $\bar{\theta}_r(h)$ e $\delta(K)$ são apresentados nas Equações (2.44), (2.45) e (2.46), respectivamente.

$$R(\phi_h^{(r)}) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{(-1)^r}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \left(K^{(r)} * K^{(r)} \right) \left(\frac{X_j - X_i}{h} \right) \quad (2.44)$$

$$\bar{\theta}_r(h) = \frac{(-1)^r}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \left(K^{(2r)} \right) \left(\frac{X_j - X_i}{h} \right) \quad (2.45)$$

$$\xi(K) = \int_{\mathfrak{R}} x^4 K(x) dx \quad (2.46)$$

Modified Cross-Validation (MCV): O seletor proposto por Stute (STUTE, 1992) visa aproximar o “termo problemático” através da projeção de Hajek (HÁJEK *et al.*, 1968). O critério de minimização é definido pela Equação (2.47):

$$MCV(h, r) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{(-1)^r}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \left(\varphi_{MCV}^{(r)} \right) \left(\frac{X_j - X_i}{h} \right) \quad (2.47)$$

onde, R é a rugosidade, r é a ordem da derivada, n representa o número de amostras, K é a função *kernel*, X é a variável aleatória e h a largura de banda. A função $\varphi^{(r)}(c)$ é definida como:

$$\varphi_{MCV}^{(r)}(c) = \left(K^{(r)} * K^{(r)} - K^{(2r)} - \frac{\mu_2(K)}{2} K^{(2r+2)} \right) (c) \quad (2.48)$$

e c representa $\left(\frac{X_j - X_i}{h} \right)$.

Trimmed Cross-Validation (TCV): (FELUCH; KORONACKI, 1992) propôs uma modificação no seletor UCV, através de um “corte” na versão *unbiased*. O critério de minimização é dado pela Equação (2.49):

$$TCV(h, r) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{(-1)^r}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \left(\varphi_{TCV}^{(r)} \right) \left(\frac{X_j - X_i}{h} \right) \quad (2.49)$$

onde, R é a rugosidade, r é a ordem da derivada, n representa o número de amostras, K é a função *kernel*, X é a variável aleatória e h a largura de banda. A função $\varphi^{(r)}(c)$ é definida como:

$$\varphi_{TCV}^{(r)}(c) = \left[K^{(r)} * K^{(r)} - K^{(2r)} - 1 \left(|c| > \frac{c_n}{h^{2r+1}} \right) \right] (c) \quad (2.50)$$

Ademais, c representa $\left(\frac{X_j - X_i}{h} \right)$ e $c_n = 1/n$.

One Sided Cross-Validation (OSCV): Esse seletor baseado em CV proposto por Savchuk (SAVCHUK, 2017) usa a definição tradicional desenvolvida por Rudemo (RUDEMO, 1982) e modifica a função custo utilizada por Martinez (MARTINEZ-MIRANDA; NIELSEN; SPERLICH, 2009). O método OSCV se baseia em um *kernel* L e pode ser definido como mostra a Equação (2.51).

$$OSCV_L(\gamma) = R(\hat{f}_{\gamma,L}) - \frac{2}{n} \sum_{i=1}^n \hat{f}_{\gamma,L}^{-i}(X_i), \quad (2.51)$$

onde, R representa a rugosidade, $\hat{f}_{\gamma,L}$ é a estimaco de densidade usando a largura de banda γ e o *kernel* L . $\hat{f}_{\gamma,L}^{-i}$ é uma modificaco atravs da CV “*leave-one-out*” computada para todas as amostras, exceto X_i . A constante de suavizaco da largura de banda utilizada, presente no mtodo, foi escolhida para o caso de densidades pouco suaves. ($w = 0, 5730$)

Silverman (SV): Silverman, em (SILVERMAN, 1986), parte do pressuposto que a abordagem natural é usar uma distribuico da famlia Normal para atribuir valores ao termo $\int f''(x)^2 dx$ para o clculo do tamanho ideal da largura de banda. Considerando a distribuico Normal ϕ com varincia σ^2 e o nmero de amostras n , temos a Equaco (2.52):

$$\begin{aligned} \int f''(x) dx &= \sigma^{-5} \int \phi''(x)^2 dx \\ &= \frac{3}{8} \pi^{-1/2} \sigma^{-5} \approx 0.212 \sigma^{-5} \end{aligned} \quad (2.52)$$

No caso de utilização do *Kernel* Gaussiano, a largura de banda será dada pela Equação (2.53):

$$\begin{aligned} h &= (4\pi)^{-1/10} \frac{3}{8} \pi^{-1/2} \sigma n^{-1/5} \\ &= \left(\frac{4}{3}\right)^{-1/5} \sigma n^{-1/5} = 1.06 \sigma n^{-1/5} \end{aligned} \quad (2.53)$$

Silverman Modificação 1 - Robusto (SVM1): A primeira modificação de Silverman pretende aumentar a robustez do método, amenizando o efeito do espalhamento de escala das distribuições e dos *outliers*. As fórmulas foram reescritas em termo do IQR da distribuição em questão e o número de amostras n . A nova largura de banda foi definida pela Equação (2.54):

$$h = 0.79 IQR(x) n^{-1/5} \quad (2.54)$$

Quando utilizado em distribuições com caudas pesadas o seletor obtém uma vantagem de performance, entretanto para distribuições bimodais o resultado se torna pior, devido a tendência de *oversmoothing* do seletor.

Silverman Modificação 2 - Adaptativo (SVM2): A segunda modificação tenta equilibrar o melhor das duas realidades, utilizando uma estimação adaptativa para o espalhamento da distribuição $\eta = \min(\sigma, IQR/1.34)$, como definido pela Equação (2.55):

$$h = 0.9 \eta n^{-1/5} \quad (2.55)$$

Sheather and Jones (SJ): O seletor de Sheather and Jones é bastante utilizado e bem referenciado na literatura em relação aos métodos de PI para cálculo da largura de banda. Esse seletor pode ser encontrado em (SHEATHER; JONES, 1991) e utiliza a mesma idéia do seletor de Park-Marron (PARK; MARRON, 1990) com uma alteração no estimador de $\widehat{\|f''\|_2^2}$, corrigindo problemas que podem ocasionar resultados com valores negativos. A largura de banda do método SJ é definida pela Equação (2.56):

$$h = \left[\frac{R(K)}{\sigma_K^4 \hat{S}_D(\hat{\alpha}_2(h))} \right]^{1/5} n^{-1/5}, \quad (2.56)$$

onde,

$$\hat{S}_D(\gamma) = \{n(n-1)\}^{-1} \gamma^{-5} \sum_{i=1}^n \sum_{j=1}^n \phi^{iv} \{\gamma^{-1}(X_i - X_j)\}, \quad (2.57)$$

e,

$$\hat{\alpha}_2(h) = 1.357 \left\{ \hat{S}_D(\tau) / \hat{T}_D(v) \right\}^{1/7} h^{5/7} \quad (2.58)$$

$$\hat{T}_D(v) = -\{n(n-1)\}^{-1} v^{-7} \sum_{i=1}^n \sum_{j=1}^n \phi^{vi} \{v^{-1}(X_i - X_j)\} \quad (2.59)$$

Temos que, $\tau = 0.920IQR(x)n^{-1/7}$, $v = 0.912IQR(x)n^{-1/9}$, γ é uma outra largura de banda diferente de h e ϕ é uma configuração diferente do *kernel* K .

Scott (SC): A largura de banda h pela regra de Scott é proporcional ao fator $n^{(-1/(d+4))}$ e usa o fator de escala $\hat{\sigma} = IQR/1.348$ para aumentar sua robustez. A regra de Scott é definida para o caso multivariado pela Equação (2.60):

$$h_i = \hat{\sigma}_i n^{-1/(d+4)} \quad (2.60)$$

onde, i se refere a uma respectiva dimensão, n é o tamanho da amostra e d representa o número de dimensões.

L1 Improved (L1I): Este seletor pode ser encontrado em (DEVROYE *et al.*, 1997) e tem como particularidade, dentre os outros seletores, ser construído utilizando a função *kernel* Epanechnikov. Este seletor é definido pela Equação (2.61):

$$h_{pi,L1I} = \min \left\{ \left(\frac{\sqrt{15/(2\pi)}\alpha}{\varepsilon} \right) n^{-1/5}, h_{ms,L1} \right\} \quad (2.61)$$

onde,

$$h' \leftarrow h_{ref,L1}(h_{dk}) \quad (2.62)$$

$$\alpha = \int \sqrt{f_n h'} \quad (2.63)$$

$$R = \alpha \sqrt{\int (K - L)^2} / \sqrt{nh'} \int |f_n h' - g_n h'| \quad (2.64)$$

$$h'' = h' \max(1, (10R)^{2/5}) \quad (2.65)$$

$$\varepsilon = 2 \int |f_n h'' - g_n h''| / h''^2 \int x^2 K \quad (2.66)$$

2.4.2.2 KDE Multivariado

As pesquisas científicas atuais, geralmente, possuem variáveis aleatórias a serem analisadas mutuamente. Em algumas dessas análises existe a necessidade de calcular a probabilidade conjunta entre essas variáveis. O raciocínio da estimação via KDE pode ser expandido, criando uma generalização multidimensional, que será denominada *Multivariate Kernel Density Estimation* (MKDE). Essa adaptação é dada pela Equação (2.67):

$$\begin{aligned} \hat{f}_h^{d_1, 2, \dots, n}(x^{d_1, 2, \dots, n}) = \\ \frac{1}{n} \sum_{i=1}^N \frac{1}{h^{d_1}} \frac{1}{h^{d_2}} \frac{1}{h^{d_n}} \sum_{i=1}^n K\left(\frac{(x^{d_1} - X^{d_1}_i)}{h^{d_1}}\right) K\left(\frac{(x^{d_2} - X^{d_2}_i)}{h^{d_2}}\right) K\left(\frac{(x^{d_n} - X^{d_n}_i)}{h^{d_n}}\right) \end{aligned} \quad (2.67)$$

Onde d_1, d_2, \dots, d_n denota o valor em cada dimensão que está sendo analisada. Geralmente, devido a preocupação com o custo computacional, essas equações são feitas matricialmente. De forma análoga a (Equação 2.68):

$$\begin{aligned} \hat{f}_H(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\det(H)} K\{H^{-1}(x - X_i)\} = \\ \frac{1}{n} \sum_{k=1}^n K_H(x - X_i) \end{aligned} \quad (2.68)$$

com $H = \text{diag}(h_1, h_2, n)$.

2.4.3 KDE com Largura de Banda Variável

Como alternativa aos seletores de banda fixa existem os seletores de banda variável, que são capazes de alterar o valor de h de acordo com alguma informação a priori. Em Terrel e Scott (TERRELL; SCOTT, 1992) observamos um estudo voltado apenas para KDE com este tipo de largura de banda, apresentando algumas alternativas. A primeira proposta

foi desenvolvida por Loftsgaarden (LOFTSGAARDEN; QUESENBERRY *et al.*, 1965), sendo chamada de “ K_{th} nearest neighbor” e pode ser definida pela Equação (2.69):

$$\widehat{f}_{h_i}(x) = \frac{K_{h_i}}{nV_d h_i x^d} \quad (2.69)$$

onde $h_i(x)$ é a distância de x até a $n_{ésima}$ amostra mais próxima, V_d é o volume da esfera unitária S_d em \mathbb{R}^d , onde d representa a dimensão da amostra. O estimador K_{th} nearest neighbor pode ser escrito como um estimador de kernel se $K(u)$ for escolhido para ser uma densidade Uniforme na esfera unitária S_d , e pode ser encontrado na Equação (2.70):

$$\widehat{f}_{h_i}(x) = \frac{1}{nh_i V_d x^d} \frac{1}{n} \sum_{i=1}^n K_{h_i}(x - X_i) \quad (2.70)$$

A segunda proposta foi feita por Breiman (BREIMAN; MEISEL; PURCELL, 1977) e sua definição é mostrada na Equação (2.71):

$$\widehat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_i^d} K_{h_i}(x - X_i) \quad (2.71)$$

onde h_k é a distância Euclideana de x_k ao $k_{ésimo}$ vizinho mais próximo da amostra. Esse método é assintoticamente equivalente a escolher $h_k \propto f(x_k)^{-1/d}$. Abramson (ABRAMSON, 1982) propôs usar $h_k \propto f(x_k)^{-1/2}$ para todas as dimensões.

A duas propostas vistas nas Equações (2.70) e (2.71) sugerem o estudo de dois métodos simples de variação da largura de banda. Primeiro a largura de banda do kernel pode depender apenas de x_k , o ponto onde é feita a estimação, e será representada por h_k . O segundo método depende das amostras X_i e a largura de banda será representada por h_i .

Baloon Estimator: O primeiro método é denominado *baloon estimator*, que é uma generalização de (TUKEY; TUKEY, 1981), sendo definido pela Equação (2.72):

$$\widehat{f}_{h_k}(x) = \frac{1}{nh_k(x)^d} \sum_{i=1}^n K_{h_k}(x - X_i) \quad (2.72)$$

onde $h_k = h(x_k)$ é uma largura de banda que varia de acordo com o ponto de estimação x_k .

Sample Smoothing Estimator: O segundo método é chamado de *sample smoothing estimator*. Nessa proposta existe a variação da largura de banda para cada amostra e sua definição é dada pela Equação (2.73):

$$\widehat{f}_h^i(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_i^d} K_{h_i}(x - X_i) \quad (2.73)$$

onde $h_i = h(x_i)$ é uma largura de banda que varia de acordo com a amostra X_i . Existem algumas propostas mostradas anteriormente que utilizam a distância Euclidiana para definir h_i e h_k , Abramson (ABRAMSON, 1982) sugere o método apresentado pela Equação (2.74):

$$h(x_i) = h \left[\frac{\lambda}{f_p(x_i)} \right]^{\frac{1}{2}} \quad (2.74)$$

Silverman (SILVERMAN, 1986) propõe outro parâmetro com o objetivo de aumentar a robustez do KDE através da inserção de λ , chamado de constante de proporcionalidade. Essa constante é definida através da Equação (2.75):

$$\lambda = e^{-n^{-1} \sum_{i=1}^n \log(f_p(x_i))} \quad (2.75)$$

Binned Kernel Density Estimator: Um método alternativo de *sample estimator* foi desenvolvido por Silverman e Scott (SILVERMAN, 1982; SCOTT; SHEATHER, 1985), sendo conhecido como *Binned Kernel Density Estimator* (BKDE). Ao invés de trabalhar diretamente com todas as amostras da variável aleatória X , os autores separaram as amostras em intervalos igualmente espaçados da variável aleatória, como são representados os *bins*. Para resumir o conceito do método considere a largura do *bin* δ , t_j representando o centro do bin B_j , o número de amostras dentro do intervalo de B_j denotado por C_j , $\delta = t_{j+1} - t_j$ para todo j e $\sum C_j = n$. Agora, BKDE pode ser definido pela Equação (2.76):

$$\widehat{f}(x) = \frac{1}{nh} \sum_{j=-\infty}^{\infty} C_j K \left(\frac{x-t_j}{h} \right) \quad (2.76)$$

3 DESENVOLVIMENTO

Os algoritmos de estimação não-paramétrica sofrem influência direta das características do conjunto de amostras, ou seja, em conjuntos de dados experimentais a presença de *outliers* e outros tipos de degradações podem alterar o tamanho do intervalo de estimação, adicionar erros de interpolação e prejudicar a performance de alguns seletores de largura de banda. Geralmente, uma das estratégias adotadas por analistas de dados é aplicar um pré-processamento no conjunto de amostras com o intuito de mitigar esses efeitos antes da estimação. Desta forma, com as densidades estimadas em mãos, será possível combinar essas informações em um classificador baseado em verossimilhança, que terá sua performance influenciada pela resiliência dos seletores de largura de banda (e suas respectivas estimações) aos graus de dificuldades distintos encontrados em análises práticas. Portanto, uma abordagem bastante comum neste tipo de estudo segue a seguinte metodologia:

1. Dados de Entrada;
2. Pré-processamento; (Apêndice B)
3. Estimação não-paramétrica;
4. Análise da estimação;
5. Construção da verossimilhança;
6. Análise da classificação.

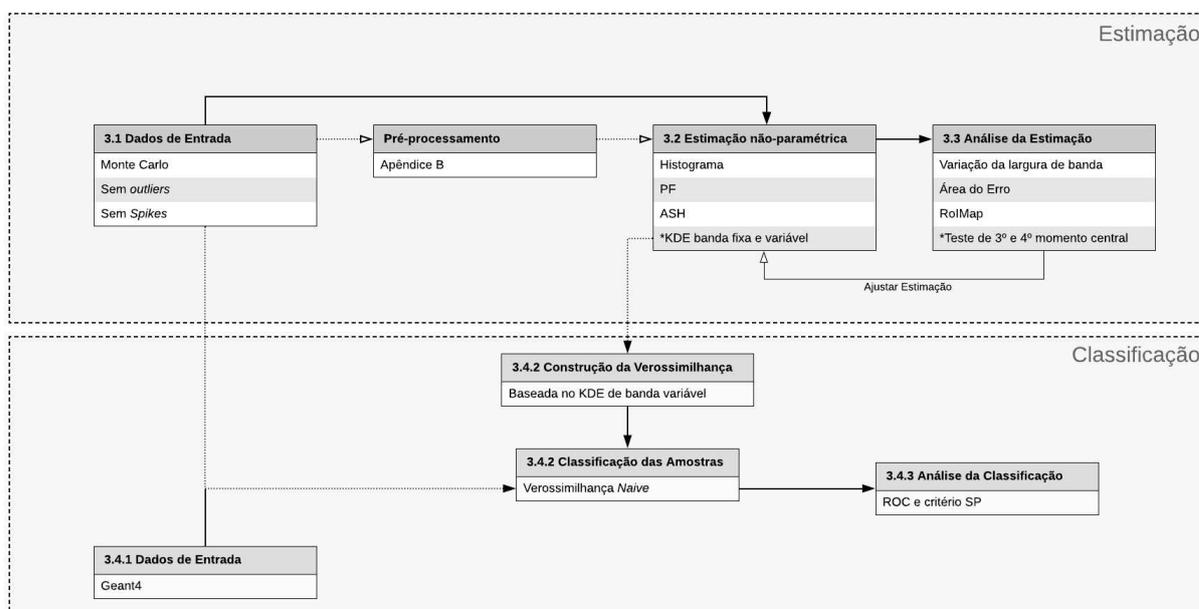
A Figura 7 apresenta o fluxograma das etapas listadas anteriormente e pode ser dividido em dois temas principais:

1. **Estimação:** Basicamente, na etapa **3.1 Dados de entrada** temos o conjunto de amostras de entrada do sistema e essas amostras podem conter características indesejáveis que serão corrigidas através da teoria adequada na etapa **Pré-processamento** (Apêndice B). Após essa “correção”, o conjunto de amostras estará “ajustado” para a estimação não-paramétrica de densidade ser realizada na etapa **3.2 Estimação não-paramétrica**, utilizando seletores de largura de banda e escolhendo o método de estimação adequado. Após a realização da estimação existe a necessidade de avaliar sua “qualidade”, que será executada na etapa **3.3 Análise da estimação e**

toda informação obtida será utilizada em propostas de melhorias dos seletores de largura de banda, reajustando parâmetros na etapa **3.2 Estimação não-Paramétrica**.

2. **Classificação:** Após construir e avaliar as características dos seletores e estimadores não-paramétricos é possível combinar as respectivas densidades de um determinado conjunto de amostras, que podem ser encontrados na etapa **3.4.1 Dados de entrada**, em um classificador baseado em verossimilhança, como mostra a etapa **3.4.2 Construção da verossimilhança e Classificação das Amostras**. Por fim, esses dados serão avaliados através de medidas conhecidas de performance na etapa **3.4.3 Análise da classificação**.

Figura 7 – Fluxograma das etapas de estimação e classificação.



Fonte: Elaborada pelo autor (2020).

No desenvolvimento deste trabalho serão abordados os itens: **3.1 Dados de entrada**, **3.2 Estimação não-paramétrica**, **3.3 Análise da estimação**, **3.4.1 Dados de entrada**, **3.4.2 Construção da verossimilhança e Classificação das Amostras** e **3.4.3 Análise da classificação**. O item **Pré-processamento** será discutido brevemente no Apêndice B, devido aos conjuntos de amostras contidos nesse estudo não apresentarem *outliers* e *spikes* ocasionadas por um único valor.

3.1 DADOS DE ENTRADA

3.1.1 Distribuições

Os dados de entrada foram gerados através do método de Monte Carlo, sendo possível obter as variáveis aleatórias e suas funções geradoras. Essa abordagem possibilita a análise de conjuntos de amostras com características distintas e com diferentes quantidades de eventos. Além disso, será possível comparar a estimação não-paramétrica via diferentes seletores com o seu respectivo modelo ideal.

Em estudos práticos geralmente encontramos desde distribuições bastante conhecidas, e bem contempladas na literatura, até distribuições altamente complexas. Dentro dessa perspectiva, a escolha das distribuições deve-se a necessidade de avaliar algumas características julgadas importantes no âmbito desse trabalho e representar a diversidade encontrada em estudos práticos.

As distribuições podem ser divididas em 4 grupos principais: D1, que representa distribuições unimodais; D2, distribuições formadas pela mistura de outras duas distribuições; D3, distribuições formadas pela mistura de 3 distribuições e D4, distribuições contendo descontinuidades. Os subgrupos “a, b e c” representam alterações na característica de escala, esparsidade das distribuições mistas, curtose e assimetria dos grupos principais. Abaixo serão listadas as motivações para a escolha de cada uma das doze distribuições, suas respectivas figuras e tabelas de características/parâmetros. Todas as distribuições foram criadas à partir das distribuições: Log-Normal, *Generalized Gaussian Distribution* (GGD), Beta e Uniforme.

1. **D1a: (Figura 8) (Tabela 2)**

- Distribuição Gaussiana, padrão em análises estatísticas, com curtose igual a 3 e assimetria igual a 0.

2. **D1b: (Figura 9) (Tabela 3)**

- Distribuição com alta assimetria e cauda longa.

3. **D1c: (Figura 10) (Tabela 4)**

- Distribuição com alta curtose e rápida transição de pico.

4. **D2a: (Figura 11) (Tabela 5)**

- Duas Gaussianas deslocadas (uma próxima a outra) com diferentes variâncias.

5. **D2b: (Figura 12) (Tabela 6)**

- Duas distribuições assimétricas com pico e vale com transição suave e alta derivada nas caudas.

6. **D2c: (Figura 13) (Tabela 7)**

- Duas distribuições distantes com alta curtose.

7. **D3a: (Figura 14) (Tabela 8)**

- Duas distribuições (com alta e baixa curtose) somada a uma distribuição Log-Normal.

8. **D3b: (Figura 15) (Tabela 9)**

- Três distribuições para explorar transições de pico suaves e rápidas e transição rápida na cauda.

9. **D3c: (Figura 16) (Tabela 10)**

- Três distribuições separadas com alta curtose formando uma distribuição trimodal.

10. **D4a: (Figura 17) (Tabela 11)**

- Duas distribuições uniformes com alto desvio padrão média e quatro descontinuidades.

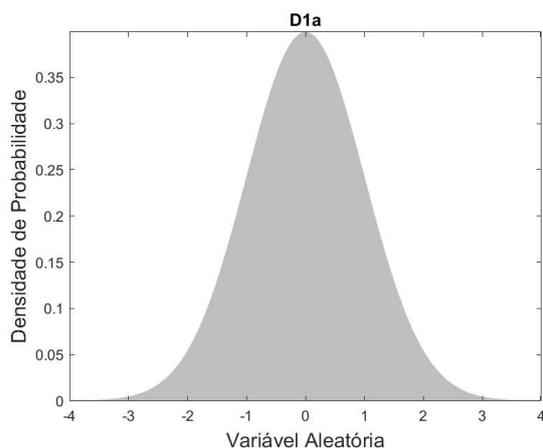
11. **D4b: (Figura 18) (Tabela 12)**

- Duas distribuições bem separadas com caudas longas e duas descontinuidades.

12. **D4c: (Figura 19) (Tabela 13)**

- Três distribuições uniformes bem separadas formando uma distribuição com seis descontinuidades.

Figura 8 – Distribuição D1a.



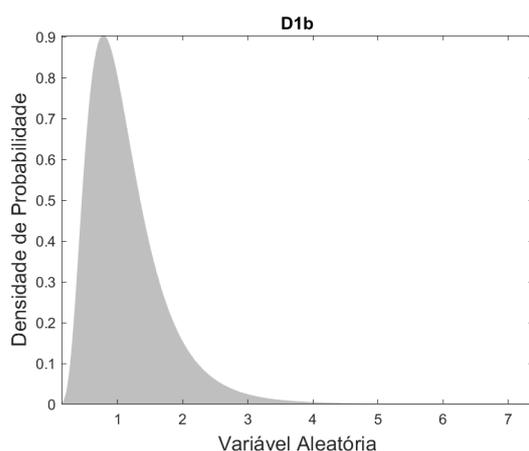
Fonte: Elaborada pelo autor (2020).

Tabela 2 – Definição da PDF D1a.

Equação
$GGD(x; \mu, \beta, \rho) = \frac{\beta^{1/2}}{2\Gamma(1+1/\rho)} e^{(-\beta\rho/2 x-\mu ^\rho)}$
Parâmetros
$\mu = 0, \beta = 1, \rho = 2$

Fonte: Elaborada pelo autor (2020).

Figura 9 – Distribuição D1b.



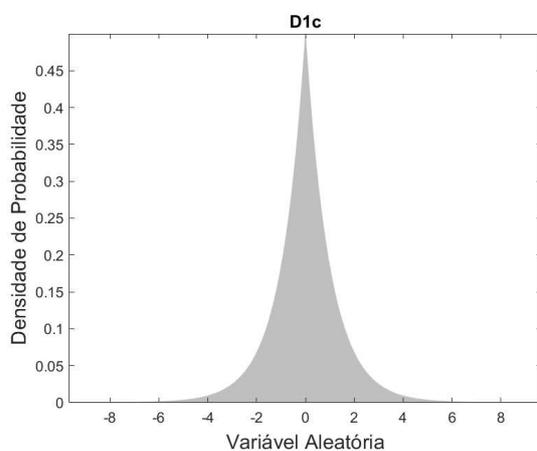
Fonte: Elaborada pelo autor (2020).

Tabela 3 – Definição da PDF D1b.

Equação
$LN(x \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$
Parâmetros
$\mu = 0, \sigma = 0.5$

Fonte: Elaborada pelo autor (2020).

Figura 10 – Distribuição D1c.



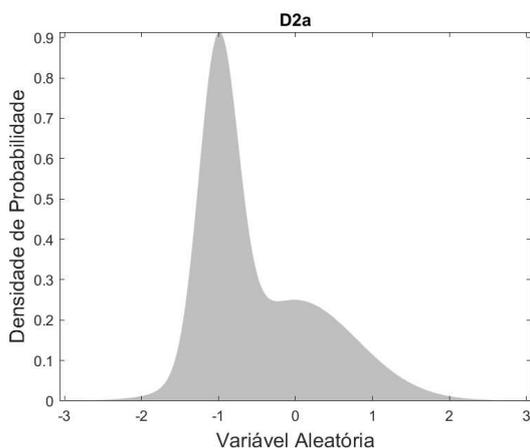
Fonte: Elaborada pelo autor (2020).

Tabela 4 – Definição da PDF D1c.

Equação
$GGD(x; \mu, \beta, \rho) = \frac{\beta^{1/2}}{2\Gamma(1+1/\rho)} e^{(-\beta\rho/2 x-\mu ^\rho)}$
Parâmetros
$\mu = 0, \beta = 1, \rho = 1$

Fonte: Elaborada pelo autor (2020).

Figura 11 – Distribuição D2a.



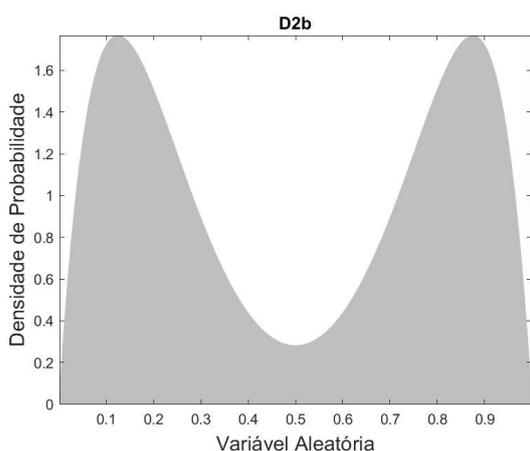
Fonte: Elaborada pelo autor (2020).

Tabela 5 – Definição da PDF D2a.

Equação
$GGD(x; \mu_1, \beta_1, \rho_1) + GGD(x; \mu_2, \beta_2, \rho_2) = \frac{\left(\frac{\beta_1^{1/2}}{2\Gamma(1+1/\rho_1)} e^{(-\beta_1 \rho_1/2 x - \mu_1 ^{\rho_1})} \dots \right)}{2}$ $\left(\dots + \frac{\beta_2^{1/2}}{2\Gamma(1+1/\rho_2)} e^{(-\beta_2 \rho_2/2 x - \mu_2 ^{\rho_2})} \right)$
Parâmetros
$\mu_1 = 0, \beta_1 = 1.25, \rho_1 = 2$ $\mu_2 = -1, \beta_2 = 4, \rho_2 = 2$

Fonte: Elaborada pelo autor (2020).

Figura 12 – Distribuição D2b.



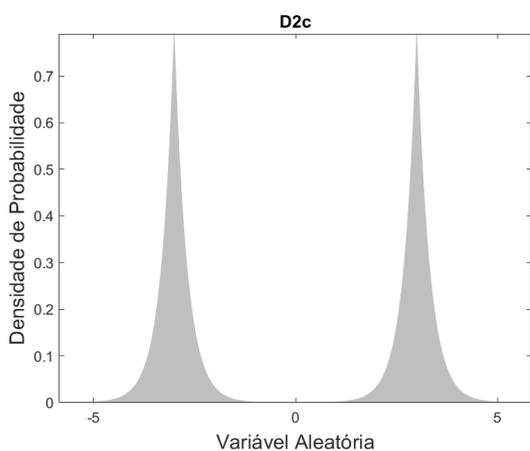
Fonte: Elaborada pelo autor (2020).

Tabela 6 – Definição da PDF D2b.

Equação
$Bt(x; \varphi_{\alpha_1}, \varphi_{\beta_1}) + Bt(x; \varphi_{\alpha_2}, \varphi_{\beta_2}) = \frac{\left(\frac{\Gamma(\varphi_{\alpha_1} + \varphi_{\beta_1})}{\Gamma(\varphi_{\alpha_1})\Gamma(\varphi_{\beta_1})} x^{\varphi_{\alpha_1}-1} (1-x)^{\varphi_{\beta_1}-1} + \frac{\Gamma(\varphi_{\alpha_2} + \varphi_{\beta_2})}{\Gamma(\varphi_{\alpha_2})\Gamma(\varphi_{\beta_2})} x^{\varphi_{\alpha_2}-1} (1-x)^{\varphi_{\beta_2}-1} \right)}{2}$
Parâmetros
$\alpha_1 = 8, \beta_1 = 2$ $\alpha_2 = 2, \beta_2 = 8$

Fonte: Elaborada pelo autor (2020).

Figura 13 – Distribuição D2c.



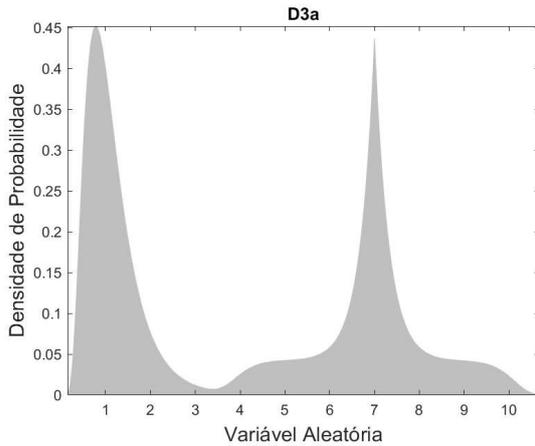
Fonte: Elaborada pelo autor (2020).

Tabela 7 – Definição da PDF D2c.

Equação
$GGD(x; \mu_1, \beta_1, \rho_1) + GGD(x; \mu_2, \beta_2, \rho_2) = \frac{\left(\frac{\beta_1^{1/2}}{2\Gamma(1+1/\rho_1)} e^{(-\beta_1 \rho_1/2 x - \mu_1 ^{\rho_1})} \dots \right)}{2}$ $\left(\dots + \frac{\beta_2^{1/2}}{2\Gamma(1+1/\rho_2)} e^{(-\beta_2 \rho_2/2 x - \mu_2 ^{\rho_2})} \right)$
Parâmetros
$\mu_1 = -3, \beta_1 = 10, \rho_1 = 1$ $\mu_2 = 3, \beta_2 = 10, \rho_2 = 1$

Fonte: Elaborada pelo autor (2020).

Figura 14 – Distribuição D3a.



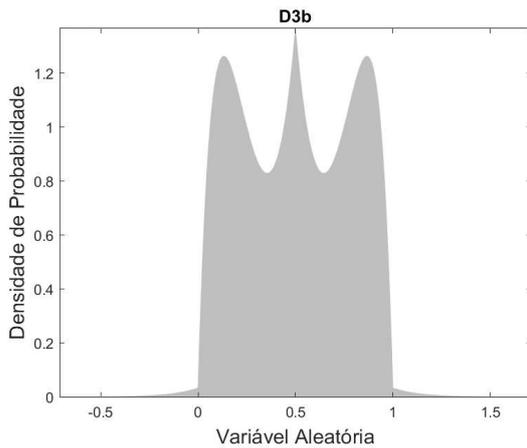
Fonte: Elaborada pelo autor (2020).

Tabela 8 – Definição da PDF D3a.

Equação
$GGD(x; \mu_1, \beta_1, \rho_1) + GGD(x; \mu_2, \beta_2, \rho_2) \cdots$ $\cdots + LN(x; \mu_3, \sigma) =$ $\left(\frac{\beta_1^{1/2}}{2\Gamma(1+1/\rho_1)} e^{(-\beta_1 \rho_1/2 x-\mu_1 ^{\rho_1})} \cdots \right)$ $\left(\cdots + \frac{\beta_2^{1/2}}{2\Gamma(1+1/\rho_2)} e^{(-\beta_2 \rho_2/2 x-\mu_2 ^{\rho_2})} \cdots \right)$ $\left(\cdots + \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu_3)^2}{2\sigma^2}} \right)$
Parâmetros
$\mu_1 = 7, \beta_1 = 0.1, \rho_1 = 10$ $\mu_2 = 7, \beta_2 = 10, \rho_2 = 1$ $\mu_3 = 0, \sigma = 0.5$

Fonte: Elaborada pelo autor (2020).

Figura 15 – Distribuição D3b.



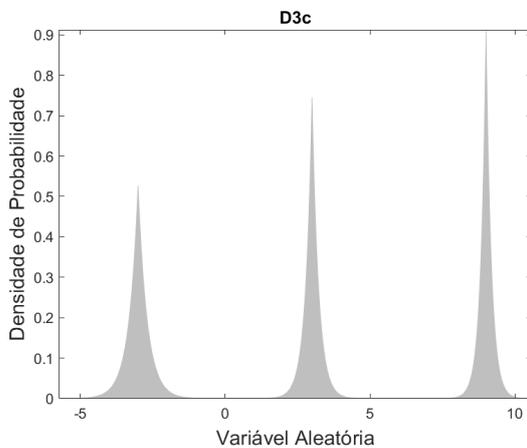
Fonte: Elaborada pelo autor (2020).

Tabela 9 – Definição da PDF D3b.

Equação
$Bt(x; \varphi_{\alpha_1}, \varphi_{\beta_1}) + Bt(x; \varphi_{\alpha_2}, \varphi_{\beta_2}) \cdots$ $\cdots + GGD(x; \mu_3, \beta_3, \rho_3) =$ $\left(\frac{\Gamma(\varphi_{\alpha_1} + \varphi_{\beta_1})}{\Gamma(\varphi_{\alpha_1})\Gamma(\varphi_{\beta_1})} x^{\varphi_{\alpha_1}-1} (1-x)^{\varphi_{\beta_1}-1} \cdots \right)$ $\left(\cdots + \frac{\Gamma(\varphi_{\alpha_2} + \varphi_{\beta_2})}{\Gamma(\varphi_{\alpha_2})\Gamma(\varphi_{\beta_2})} x^{\varphi_{\alpha_2}-1} (1-x)^{\varphi_{\beta_2}-1} \cdots \right)$ $\left(\cdots + \frac{\beta_3^{1/2}}{2\Gamma(1+1/\rho_3)} e^{(-\beta_3 \rho_3/2 x-\mu_3 ^{\rho_3})} \right)$
Parâmetros
$\alpha_1 = 8, \beta_1 = 2$ $\alpha_2 = 2, \beta_2 = 8$ $\mu_3 = 0, \beta_3 = 50, \rho_3 = 1$

Fonte: Elaborada pelo autor (2020).

Figura 16 – Distribuição D3c.



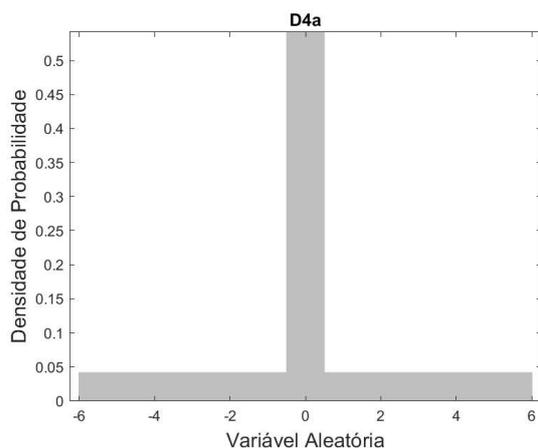
Fonte: Elaborada pelo autor (2020).

Tabela 10 – Definição da PDF D3c.

Equação
$GGD(x; \mu_1, \beta_1, \rho_1) + GGD(x; \mu_2, \beta_2, \rho_2) \cdots$ $\cdots + GGD(x; \mu_3, \beta_3, \rho_3) =$ $\left(\frac{\beta_1^{1/2}}{2\Gamma(1+1/\rho_1)} e^{(-\beta_1 \rho_1/2 x-\mu_1 ^{\rho_1})} \cdots \right)$ $\left(\cdots + \frac{\beta_2^{1/2}}{2\Gamma(1+1/\rho_2)} e^{(-\beta_2 \rho_2/2 x-\mu_2 ^{\rho_2})} \cdots \right)$ $\left(\cdots + \frac{\beta_3^{1/2}}{2\Gamma(1+1/\rho_3)} e^{(-\beta_3 \rho_3/2 x-\mu_3 ^{\rho_3})} \right)$
Parâmetros
$\mu_1 = -3, \beta_1 = 10, \rho_1 = 1$ $\mu_2 = 3, \beta_2 = 20, \rho_2 = 1$ $\mu_3 = 9, \beta_3 = 30, \rho_3 = 1$

Fonte: Elaborada pelo autor (2020).

Figura 17 – Distribuição D4a.



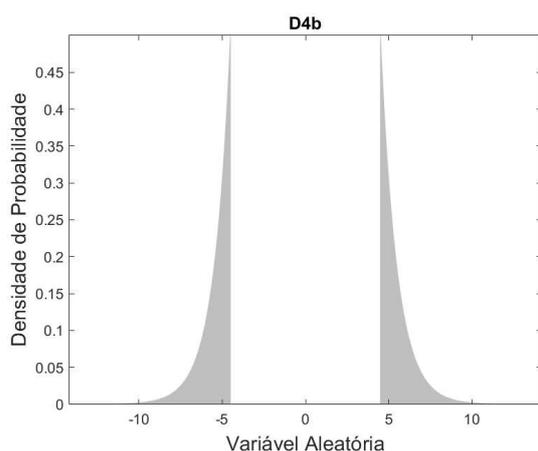
Fonte: Elaborada pelo autor (2020).

Tabela 11 – Definição da PDF D4a.

Equação
$U(x; a, b) = \begin{cases} \frac{1}{b-a} & , a \leq x \leq b \\ 0 & , x < a \text{ or } x > b \end{cases}$
$f(x; a_1, b_1, a_2, b_2) = \frac{U(x; a_1, b_1) + U(x; a_2, b_2)}{2}$
Parâmetros
$a_1 = -6, b_1 = 6$
$a_2 = -0.5, b_2 = 0.5$

Fonte: Elaborada pelo autor (2020).

Figura 18 – Distribuição D4b.



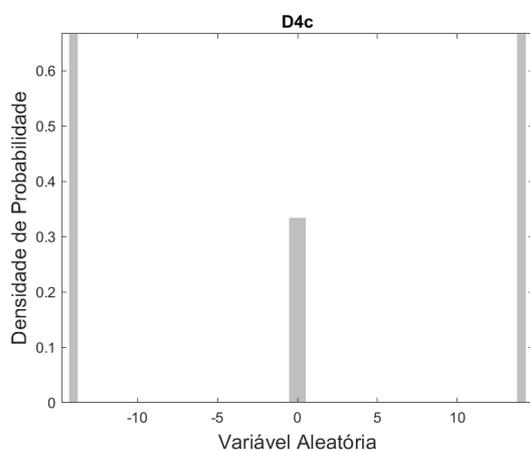
Fonte: Elaborada pelo autor (2020).

Tabela 12 – Definição da PDF D4b.

Equação
$GGD(x; \mu, \beta, \rho) = \begin{cases} \frac{\beta^{1/2}}{2\Gamma(1+1/\rho)} e^{(-\beta\rho/2 x-\mu ^\rho)} & , x \leq -4.5 \\ \frac{\beta^{1/2}}{2\Gamma(1+1/\rho)} e^{(-\beta\rho/2 x-\mu ^\rho)} & , x \geq 4.5 \\ 0 & , -4.5 \leq x \leq 4.5 \end{cases}$
Parâmetros
$\mu = 0, \beta = 1, \rho = 1$

Fonte: Elaborada pelo autor (2020).

Figura 19 – Distribuição D4c.



Fonte: Elaborada pelo autor (2020).

Tabela 13 – Definição da PDF D4c.

Equação
$U(x; a, b) = \begin{cases} \frac{1}{b-a} & , a \leq x \leq b \\ 0 & , x < a \text{ or } x > b \end{cases}$
$f(x; a_1, b_1, a_2, b_2, a_3, b_3) = \frac{U(x; a_1, b_1) + U(x; a_2, b_2) + U(x; a_3, b_3)}{3}$
Parâmetros
$a_1 = -14.25, b_1 = -13.75$
$a_2 = -0.5, b_2 = 0.5$
$a_3 = 13.756, b_3 = 14.25$

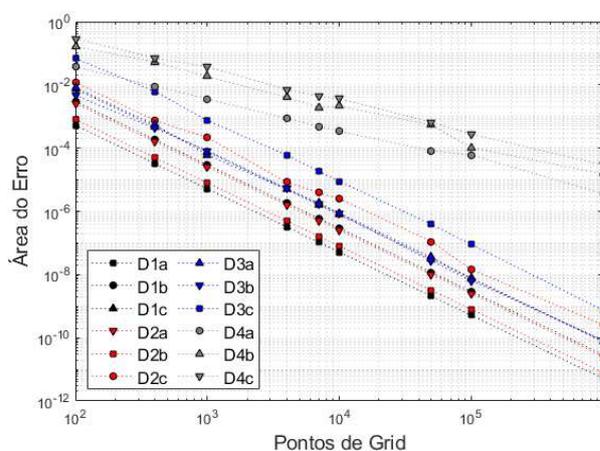
Fonte: Elaborada pelo autor (2020).

3.1.2 Erros de Interpolação

O viés prático deste estudo naturalmente leva a abordagem para o universo discreto, com o intuito de facilitar algumas operações via integral de Riemman, como: cálculo da área do erro, divisão da estimação em regiões de interesse, por exemplo. Ao discretizarmos a análise surge uma componente denominada erro de interporlação, que deve ser desprezível em relação ao erro de área entre estimação não-paramétrica e modelo ideal. A interpolação linear foi escolhida para este trabalho devido a sua facilidade de implementação, boa performance e um maior controle nas regiões de extrapolação.

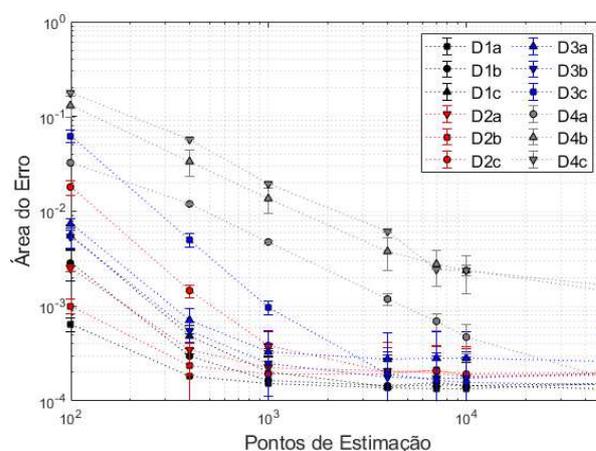
Desta etapa em diante definiremos o modelo ideal como um modelo discreto, utilizando n_{grid} número de pontos. Portanto, o número de pontos de *grid* é relativo aos pontos discretos da PDF, e devem ser escolhidos de modo a ser desprezível na parcela do erro final. Outra parcela importante é a estimação não-paramétrica, e para a escolha de seu número de pontos n_{est} deve-se ter o mesmo cuidado. A Figura 20 mostra o erro entre a integral analítica e a integral de Riemann de acordo com o aumento do n_{grid} para uma área referente a 4σ da distribuição Gaussiana, que corresponde a área de ≈ 0.9999 . Essa parcela nos mostra qual o erro do nosso modelo discreto de acordo com n_{grid} . A Figura 21 representa o erro de área entre a melhor estimação possível com o modelo analítico pelo aumento do n_{est} , levando em consideração 100 iterações com diferentes intervalos da variável aleatória em questão. Essa parcela nos mostra qual o erro intrínseco da nossa estimação de acordo com n_{est} .

Figura 20 – Área do erro da modelo discreto com o aumento do número de pontos de *grid*.



Fonte: Elaborada pelo autor (2020).

Figura 21 – Área do erro com o aumento do número de pontos de estimação.



Fonte: Elaborada pelo autor (2020).

Foi definido o número de 10^5 pontos para o modelo discreto (portanto $n_{grid} = 10^5$) e 10^4 pontos para a estimação não-paramétrica (portanto $n_{est} = 10^4$). Essa escolha justifica-

Tabela 14 – Tabela da área do erro referente a 10^4 pontos de estimação, 10^5 pontos de *grid* e o erro total residual de interpolação.

Distribuição	Erro(Estimação) $\pm\sigma$	Erro(Grid)	Erro Total Residual $\pm\sigma$
D1a	$1,62.10^{-4} \pm 1,64.10^{-4}$	$5,16.10^{-10}$	$2,29.10^{-4} \pm 1,64.10^{-4}$
D1b	$1,31.10^{-4} \pm 1,13.10^{-4}$	$2,93.10^{-9}$	$1,86.10^{-4} \pm 1,13.10^{-4}$
D1c	$1,57.10^{-4} \pm 1,23.10^{-4}$	$7,78.10^{-9}$	$2,21.10^{-4} \pm 1,23.10^{-4}$
D2a	$2,05.10^{-4} \pm 1,95.10^{-4}$	$2,53.10^{-9}$	$2,90.10^{-4} \pm 1,95.10^{-4}$
D2b	$1,96.10^{-4} \pm 1,54.10^{-4}$	$7,97.10^{-10}$	$2,78.10^{-4} \pm 1,54.10^{-4}$
D2c	$1,89.10^{-4} \pm 1,77.10^{-4}$	$1,46.10^{-8}$	$2,68.10^{-4} \pm 1,77.10^{-4}$
D3a	$2,62.10^{-4} \pm 2,12.10^{-4}$	$7,45.10^{-9}$	$3,70.10^{-4} \pm 2,12.10^{-4}$
D3b	$1,95.10^{-4} \pm 1,64.10^{-4}$	$6,41.10^{-9}$	$2,75.10^{-4} \pm 1,64.10^{-4}$
D3c	$1,45.10^{-4} \pm 1,21.10^{-4}$	$8,98.10^{-8}$	$2,05.10^{-4} \pm 1,21.10^{-4}$
D4a	$4,68.10^{-4} \pm 1,45.10^{-4}$	$6,05.10^{-5}$	$6,62.10^{-4} \pm 1,45.10^{-4}$
D4b	$2,42.10^{-3} \pm 9,69.10^{-4}$	$1,03.10^{-4}$	$3,42.10^{-3} \pm 9,69.10^{-4}$
D4c	$2,36.10^{-3} \pm 3,05.10^{-4}$	$2,79.10^{-4}$	$3,33.10^{-3} \pm 3,05.10^{-4}$

Fonte: Elaborada pelo autor (2020).

se pelo baixo valor de erro neste ponto não influenciar nas avaliações posteriores, pela velocidade dos algoritmos ser aceitável e por nenhuma distribuição ter entrado na região de erro numérico. As diferenças entre os erros ocorre predominantemente pelas diferenças de derivadas das distribuições, além de processos intrínsecos deste cálculo de área do erro. A tabela 14 mostra o erro com 10^4 pontos de estimação, 10^5 pontos de *grid* e o erro total residual da análise. É importante salientar que, estes valores são resíduos causados pela discretização do modelo ideal e da estimação e estarão presentes durante todo estudo. Os maiores erros, presentes nas distribuições de derivadas infinitas D4b e D4c alcançam desprezíveis 0.3% da área total da distribuição.

3.2 ESTIMAÇÃO NÃO-PARAMÉTRICA

Nesta etapa serão mostrados os detalhes construtivos dos algoritmos responsáveis pelo cálculo da largura de banda dos métodos clássicos de estimação não-paramétrica: Histograma, PF, ASH e KDE. Na seção dedicada ao KDE de banda variável será apresentado o método desenvolvido no âmbito deste trabalho, denominado ROIKDE.

3.2.1 Histograma

Os seletores utilizados no cálculo do número de *bins* foram vistos na Seção 2.1 e as principais características utilizadas na construção dos algoritmos serão apresentadas na Tabela 15.

Onde n representa o número de eventos (ou amostras), x a variável aleatória, κ_1 é a medida de assimetria, R é o intervalo da variável aleatória x , σ o desvio padrão, μ a média,

Tabela 15 – Parâmetros utilizados no cálculo da binagem ótima para o Histograma. (*O método LHM será utilizado com interpolação “nearest”)

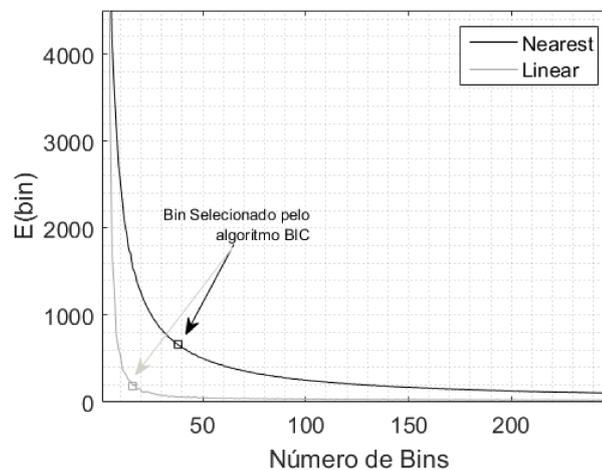
Método	Número de Bins	Parâmetros
Sturges	$N_b = 1 + \log_2(n)$	n
Doane	$N_b = 1 + \log_2(n) + \log_2\left(1 + \frac{\sqrt{\kappa_1}}{\sigma\sqrt{\kappa_1}}\right)$	n e κ_1
FD	$N_b = R/(2(IQR(x)))n^{-1/5}$	R, n e x
Scott	$N_b = R/(3.5\sigma n^{-1/3})$	R, n e σ
SS	$N_b = bin \rightarrow \min(Z(h) = 2\mu - \sigma/h^2)$	μ, σ e h
Rudemo	$N_b = bin \rightarrow \min(Q(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)n^2h} \sum_k n_k^2)$	n, h e n_k
LHM*	$N_b = bin \rightarrow \text{knee}(E(h) = \sum_{i=1}^n ECDF(x_i) - CDF(x_i))$	x
Knuth	$\log(\phi x, I) = n \log(N_b) + \log \Gamma\left(\frac{N_b}{2}\right) \dots$ $-N_b \log \Gamma\left(\frac{1}{2}\right) - \log \Gamma\left(\frac{2n+N_b}{2}\right) + \sum_{k=1}^{N_b} \log \Gamma\left(C_k + \frac{1}{2}\right)$	n, N_b e C_k
Wand	$N_b = R / \left\{ \frac{6}{-\tilde{\psi}_2(g_{21})n} \right\}^{-1/3}$	$n, \tilde{\psi}_2$ e g_{21}

Fonte: Elaborada pelo autor (2020).

h a largura do *bin*, e C_k representa a contagem em cada *bin* k . N_b representa o número ótimo de *bins* que maximiza a função F. Os parâmetros $\tilde{\psi}_2$ e g_{21} são calculados na Seção 2.1.

Dos métodos citados na tabela 15, quatro são iterativos. O método LHM, calcula a função de erro E à partir da diferença entre a ECDF e a CDF (construída com interpolação “nearest” ou linear). A Figura 22 mostra a função de erro e o ponto escolhido pelo algoritmo BIC, desenvolvido em (ZHAO; XU; FRÄNTI, 2008). No método do Histograma, a interpolação utilizada será *nearest*, devido as suas características construtivas; para o método PF e ASH a interpolação será linear.

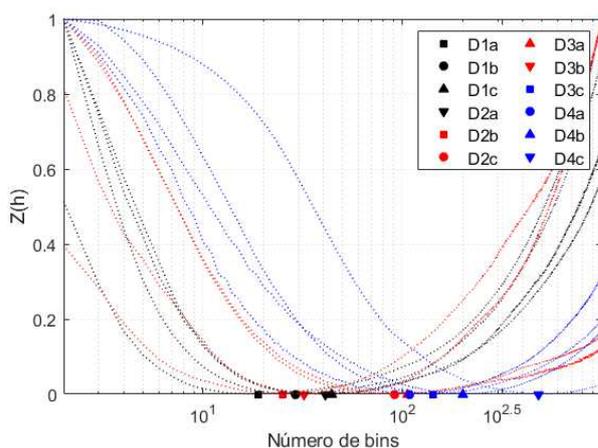
Figura 22 – Função erro E e o *bin* selecionado via BIC para o método LHM em uma distribuição D1a.



Fonte: Elaborada pelo autor (2020).

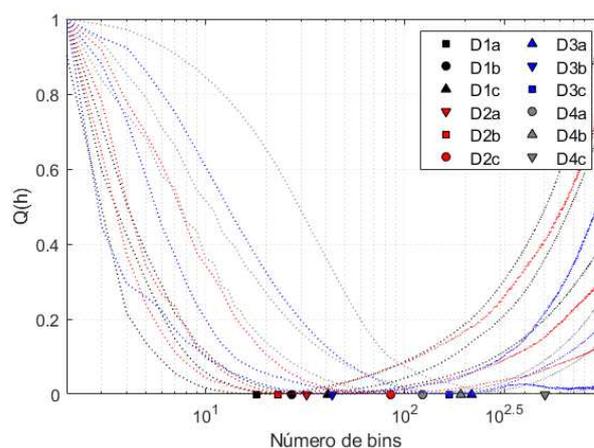
A Figura 23 mostra o número de *bins* selecionado pelo método de SS através da função custo $Z(h)$, já na Figura 24 temos o método de Rudemo e sua respectiva função custo $Q(h)$, ligeiramente diferente do método SS. Por fim, a Figura 25 mostra a ruidosa função custo do método de Knuth $\log\phi(N_b)$, sendo uma das causas da grande variância do método, principalmente com poucas amostras.

Figura 23 – Função de Custo $Z(h)$ do método de Shimazaki e Shinomoto.



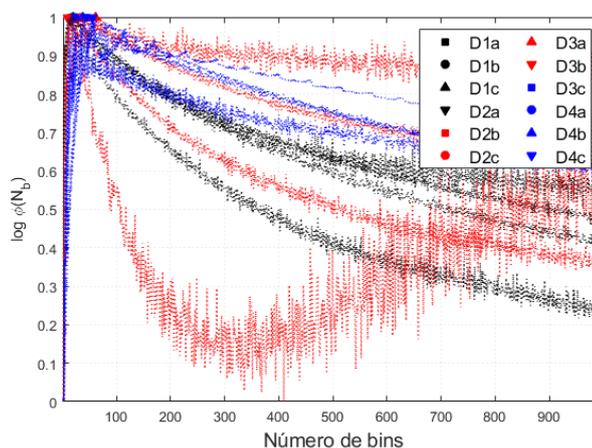
Fonte: Elaborada pelo autor (2020).

Figura 24 – Função de Custo $Q(h)$ do método de Rudemo.



Fonte: Elaborada pelo autor (2020).

Figura 25 – Função de Custo $\log\phi(N_b)$ do método de Knuth.



Fonte: Elaborada pelo autor (2020).

3.2.2 Polígonos de Frequência

De acordo com a Seção 2.2, existe uma teoria prevista para o seletor de Scott quando o estimador utilizado for o PF. O método LHM sofrerá uma alteração na interpolação de sua CDF para linear, aproximando sua característica de uma estimação por polígonos

de frequência. Essas alterações serão resumidas na Tabela 16 e os demais métodos serão utilizados de forma similar ao Histograma, como visto anteriormente.

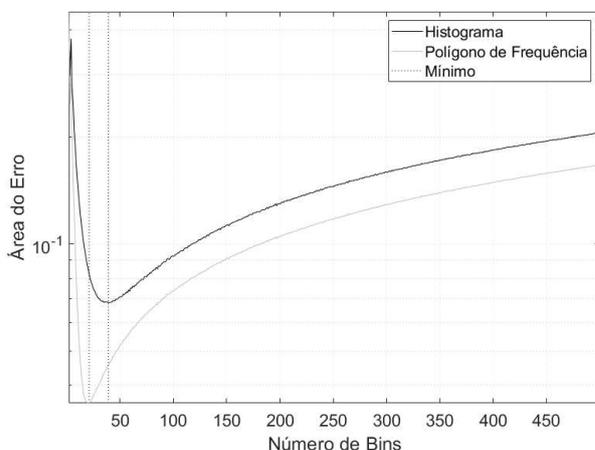
Tabela 16 – Parâmetros utilizados no cálculo da binagem ótima para o PF. (*O método LHM será utilizado com interpolação “linear”)

Métodos	Número de Bins	Parâmetros
Scott	$N_b = R / (2.15\sigma n^{-1/5})$	R, n e σ
LHM*	$N_b = bin \rightarrow \text{knee}(E(h) = \sum_{i=1}^n ECDF(x_i) - CDF(x_i))$	x

Fonte: Elaborada pelo autor (2020).

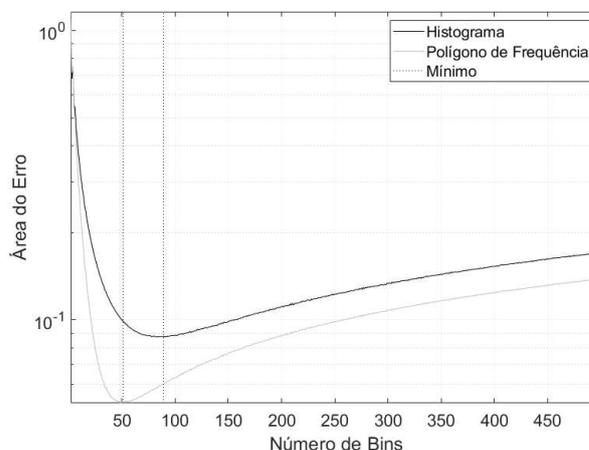
De acordo com Scott, a largura do *bin* referente ao Polígono de Frequência deve ser maior do que a largura do *bin* do Histograma, analogamente o número de *bins* deve ser menor, como mostra a Figura 4. As Figuras 26 e 27 mostram a área do erro entre o modelo discreto e o Histograma (PF) para as distribuições D1a e D1c, respectivamente, onde é possível observar que PF necessita de um menor número de *bins* para atingir a sua menor área do erro. O cálculo da área do erro será discutido com maiores detalhes na Seção 3.3.

Figura 26 – Área do erro de acordo com o número de *bins* para distribuição D1a.



Fonte: Elaborada pelo autor (2020).

Figura 27 – Área do erro de acordo com o número de *bins* para distribuição de D1c.



Fonte: Elaborada pelo autor (2020).

3.2.3 Average Shifted Histogram

O método ASH utilizará os seletores vistos anteriormente para o PF, exceto o seletor de Scott, que utiliza uma alteração prevista em teoria. O resumo dessa alteração será mostrado na Tabela 17. Nesse tipo de estimador, existe a necessidade de definir o termo m , responsável pelo número de subdivisões do *bin*. De acordo com a Seção 2.3 o valor de m deve ser ≤ 10 . As Figuras 28 e 29 mostram o impacto da variação de m em relação ao erro de área entre o estimador via ASH e o modelo discreto, utilizando a interpolação *nearest*

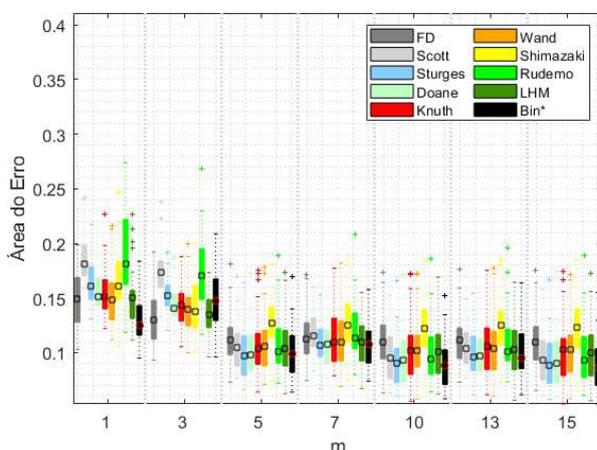
(Figura 28) e linear (Figura 29) para a distribuição D1a. Neste trabalho, como indicado por Scott e confirmado no teste, o benefício de performance em relação ao custo computacional deixa de ser vantajoso acima de $m = 10$, sendo esse o valor escolhido.

Tabela 17 – Parâmetros utilizados no cálculo da binagem ótima para o ASH.

Métodos	Número de Bins	Parâmetros
Scott	$N_b = R/(2.576\sigma n^{-1/5})$	R, n e σ

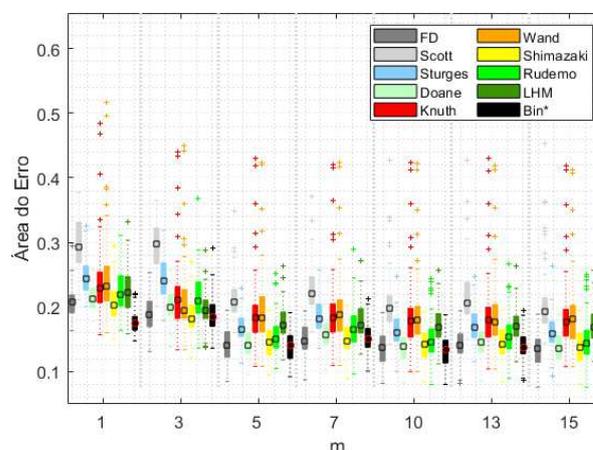
Fonte: Elaborada pelo autor (2020).

Figura 28 – Escolha do m para o método ASH na Distribuição D1a.



Fonte: Elaborada pelo autor (2020).

Figura 29 – Escolha do m para o método ASH na Distribuição D2a.



Fonte: Elaborada pelo autor (2020).

3.2.4 Kernel Density Estimator

A etapa de desenvolvimento do KDE foi dividida em largura de banda fixa e largura de banda variável. Os seletores de banda fixa foram organizados em PI e CV, como visto na Seção 2.4.2. Além disso, foram implementados os métodos de largura de banda variável baseados nas teorias vistas na Seção 2.4.3, que são: *Binned Kernel Density Estimation* (BKDE), *Variable Kernel Density Estimation* (VKDE), *Adaptive Kernel Density Estimation* (AKDE) encontrado em (SHIMAZAKI; SHINOMOTO, 2010), e uma nova abordagem de escolha automática da largura de banda variável h_i , denominada ROIKDE.

3.2.4.1 Largura de Banda Fixa

Plug-In: Dos seletores de PI, vistos anteriormente, apenas o SJ e L11 não são baseados principalmente em variáveis de escala, utilizando uma estimativa da rugosidade $R(f'')$ alternativa para o cálculo de sua largura de banda. Os seletores SV, SVM1, SVM2 e SC utilizam a suposição de distribuição Gaussiana para o conjunto de amostras, diferindo em

alterações pontuais descritas na Seção 2.4.2. O resumo dos seletores pode ser encontrado na Tabela 18.

Tabela 18 – Parâmetros utilizados no cálculo da largura de banda fixa h para o KDE.

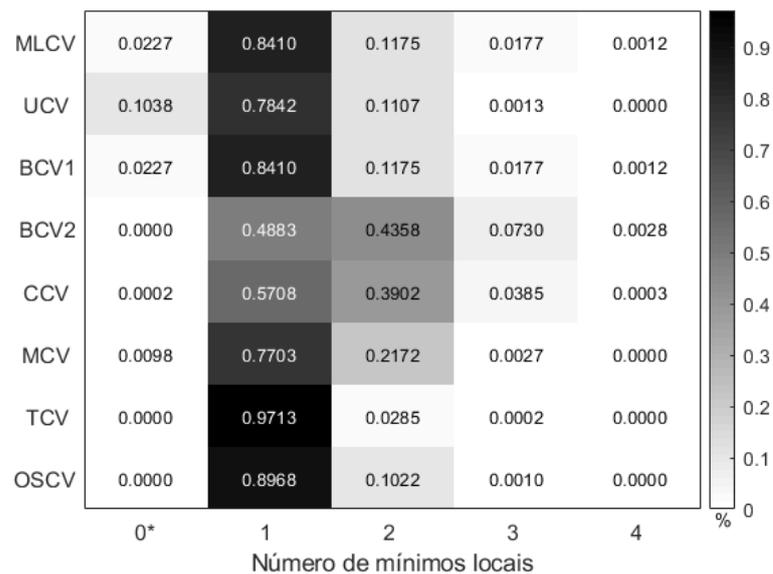
Método	Largura de Banda	Parâmetros
SV	$h = 1.06\sigma n^{-1}$	n e σ
SVM1	$h = 0.79IQR(x)n^{-1/5}$	x , n e IQR
SVM2	$h = 0.9\eta n^{-1/5}$	η e n
SJ	$h = \left[\frac{R(K)}{\sigma_K^4 \hat{S}_D(\hat{\alpha}_2(h))} \right]^{1/5} n^{-1/5}$	n , $R(K)$, σ_K^4 , \hat{S}_D e $\hat{\alpha}_2(h)$
SC	$h = \hat{\sigma}_i n^{-1/(d+4)}$	$\hat{\sigma}_i$, n and d
L1I	$h = \min \left\{ \left(\frac{\sqrt{15/(2\pi)}\alpha}{\varepsilon} \right) n^{-1/5}, h_{ms,L1} \right\}$	n , α , ε , $h_{ms,L1}$.

Fonte: Elaborada pelo autor (2020).

Onde n representa o número de amostras, x a variável aleatória, σ o desvio padrão de x , $IQR(x)$ o interquartil de x , $\eta = \min(\sigma, IQR/1.34)$, $R(K^r)$ é a rugosidade da função kernel de $r^{ésima}$ ordem, d representa o número de dimensões, $\hat{\sigma}_i$ é o mesmo que $IQR/1.348$, σ_K^4 representa o segundo momento do Kernel elevado ao quadrado, \hat{S}_D , $\hat{\alpha}_2(h)$, podem ser encontrados em 2.4.2.1, α e ε podem ser encontrados em 2.4.2.1 e $h_{ms,L1} = 2.71042\sigma n^{-1/5}$.

Cross-Validation: Como intervalo de h necessário para encontrar os pontos de mínimo (máximo) da função custo (verossimilhança) foi definido h entre 10^{-3} até $10^{1.2}$, com 100 subdivisões. Como visto anteriormente, é possível encontrar mais de um mínimo (máximo) local nos seletores de CV e como mostra a Figura 30 os seletores TCV e OSCV parecem os mais resilientes nesse contexto. Além disso, o seletor BCV2 parece encontrar mais dificuldades de convergência do que os demais para um único mínimo e o seletor UCV teve maior dificuldade em convergir dentro do intervalo estipulado. Portanto, a abordagem adotada nessa tese, que será utilizada nos resultados posteriores, utiliza o seletor TCV como referência do mínimo da função custo dos métodos CV, ou seja, quando o seletor encontrar mais de um valor de convergência, será adotado o valor mais próximo do TCV e nos poucos casos onde TCV apresentar mais de um mínimo, será utilizado o seletor SJ.

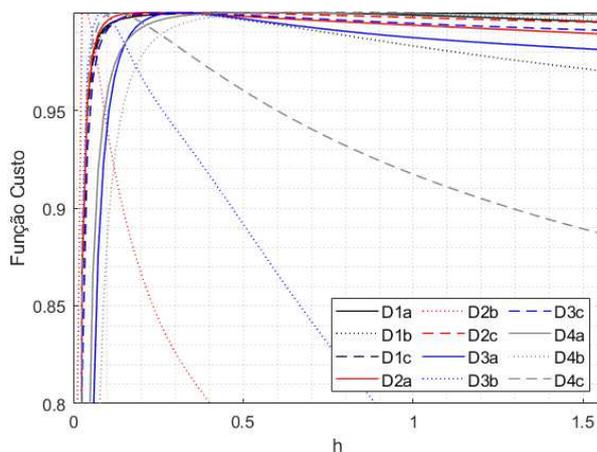
Figura 30 – Porcentagem total de mínimos locais para todas as iterações (1500) em todas as distribuições. *não convergiu no intervalo.



Fonte: Elaborada pelo autor (2020).

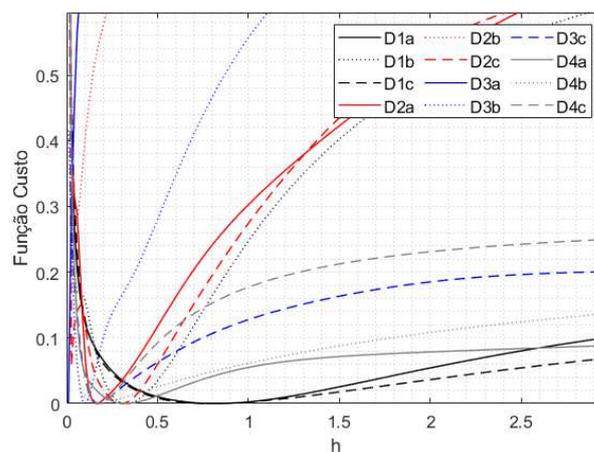
As Figuras 31 a 38 mostram as funções custo normalizadas dos seletores de CV avaliados, para diferentes distribuições. A Figura 31 mostra o seletor MLCV diferindo dos outros ao buscar a máxima verossimilhança para a melhor largura de banda, todos os outros seletores utilizam o valor mínimo da sua respectiva função custo.

Figura 31 – Método MLCV.



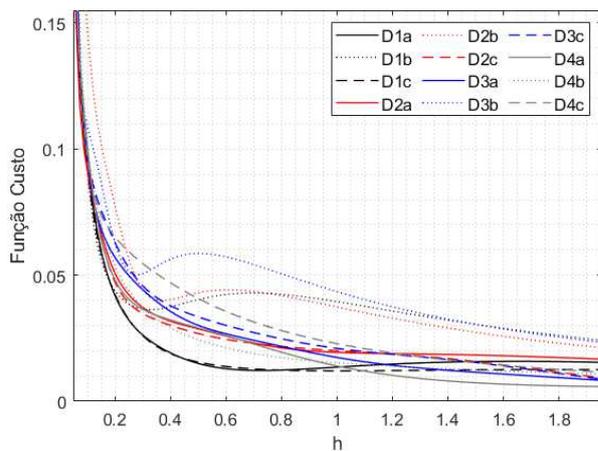
Fonte: Elaborada pelo autor (2020).

Figura 32 – Método UCV.



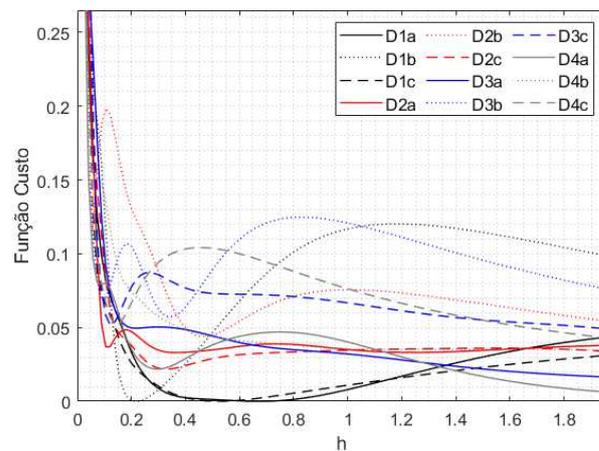
Fonte: Elaborada pelo autor (2020).

Figura 33 – Método BCV1.



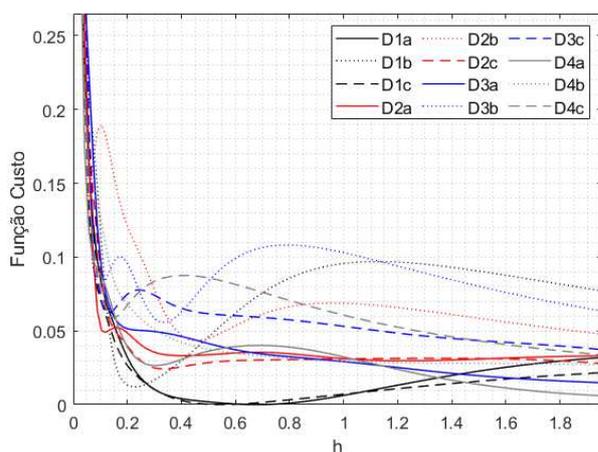
Fonte: Elaborada pelo autor (2020).

Figura 34 – Método BCV2.



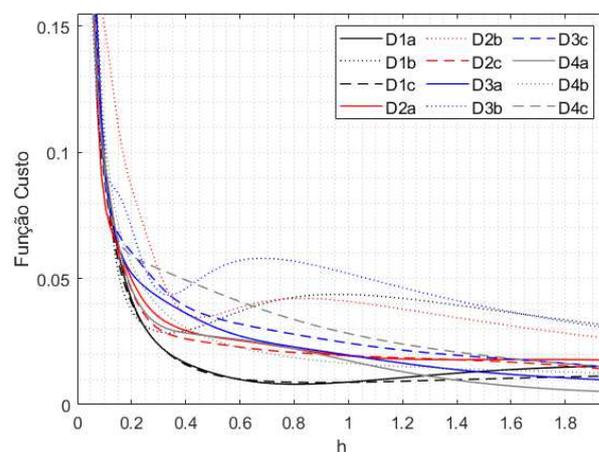
Fonte: Elaborada pelo autor (2020).

Figura 35 – Método CCV.



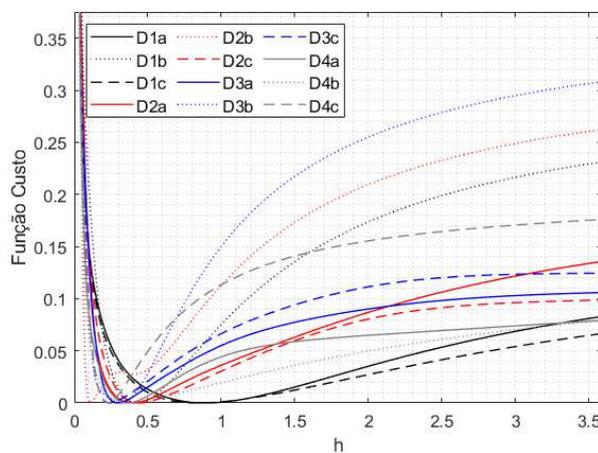
Fonte: Elaborada pelo autor (2020).

Figura 36 – Método MCV.



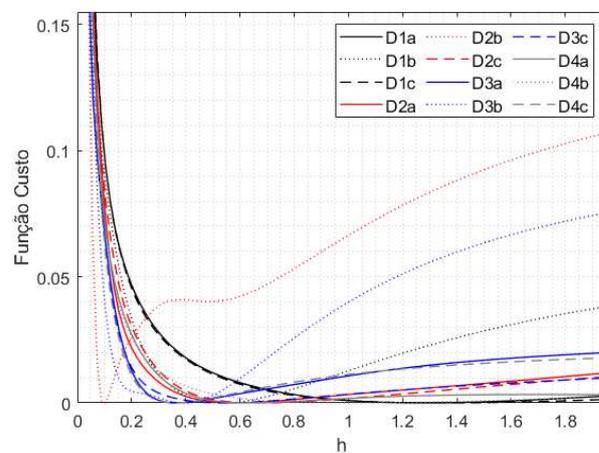
Fonte: Elaborada pelo autor (2020).

Figura 37 – Método TCV.



Fonte: Elaborada pelo autor (2020).

Figura 38 – Método OSCV.



Fonte: Elaborada pelo autor (2020).

3.2.4.2 Largura de Banda Variável

Para o seletor de largura variável teremos três representantes do *Sample Smoothing Estimator*: VKDE, BKDE e ROIKDE. Para representar do método *Ballon Estimator* teremos o AKDE. Neste trabalho os métodos BKDE e AKDE não serão alterados, respeitando suas concepções originais e utilizando os algoritmos de seus respectivos artigos. O método clássico VKDE será avaliado com mais profundidade, buscando a melhor otimização possível do parâmetro h . Por fim, as etapas necessárias para construção do método ROIKDE serão detalhadas.

AKDE: O algoritmo do AKDE foi extraído do trabalho (SHIMAZAKI; SHINOMOTO, 2010). Esse método tem como característica a técnica de *ballon estimator* e utiliza o processo de Poisson na otimização do termo problemático do MISE, para o cálculo da banda fixa. Logo após é feito um processo adaptativo da largura banda variável através do cálculo local de MISE.

BKDE: O algoritmo utilizado como BKDE foi extraído do trabalho (WOLTERS, 2009) e utilizará como largura de banda h o método OSCV (devido à complexidade das densidades estudadas aqui), embora no artigo referido o método SJ tenha sido utilizado. A função de suavização responsável pela ponderação de sua largura de banda baseia-se na lógica de um histograma, como mostrado na Seção 2.4.3.

VKDE: Esse método de *Sample Smoothing Estimator* foi implementado incorporando as melhorias verificadas na literatura ao longo do tempo. A partir dessa implementação foi possível alterar os parâmetros responsáveis por melhorar a performance do algoritmo. A fórmula base para o cálculo da largura de banda variável, que tem sido amplamente utilizada, pode ser vista na Equação (3.1).

$$h(x_i) = h \left[\frac{\lambda}{f_p(x_i)} \right]^{\frac{1}{2}} \quad (3.1)$$

Onde $h_i = h(x_i)$ é a largura de banda variável, que varia em função de cada amostra i , h é a largura de banda fixa, λ é uma constante de proporcionalidade que auxilia na robustez da estimação e seu cálculo pode ser visto na seção 2.4.3, e $f_p(x_i)$ será denominada função adaptativa de suavização.

Dos parâmetros citados acima, h e $f_p(x_i)$ serão avaliados no Capítulo 4, buscando uma estimação de densidade automática robusta.

ROIKDE: O seletor automático desenvolvido neste trabalho é dividido em quatro estágios: (1) Robustez Relativa; (2) Ajuste de Banda Fixa; (3) Ajuste de Banda Variável; (4) Estimacão via VKDE.

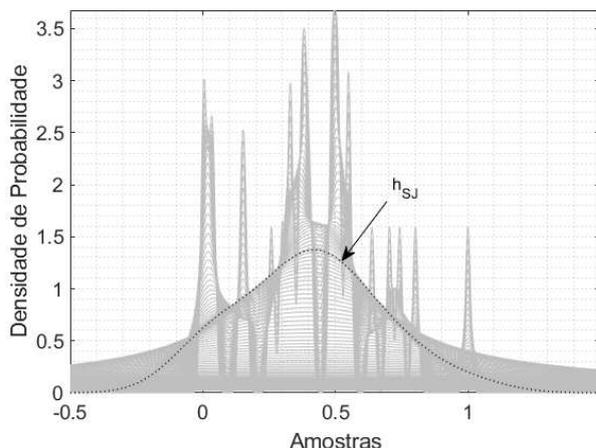
(1) Robustez Relativa: Nesta etapa o algoritmo utiliza a distância Euclideana, dada pela Equacão (B.1) do Apêndice B, para eliminar as amostras relativamente mais distantes que seriam posteriormente utilizadas no cálculo da largura de banda fixa. Ou seja, primeiramente selecionamos quantas amostras n_σ descartar pela Equacão (3.2):

$$n_\sigma = n(1 - \alpha_\sigma), \quad (3.2)$$

onde, n é o número total de amostras, α_σ é um fator referente a área de 3σ da distribuição Gaussiana ($\alpha_\sigma = 0.9973$). (2) Ajuste de Banda Fixa: Esse estágio propõe uma escolha automática para a largura de banda fixa, baseada em características que foram percebidas durante este estudo e serão explicitadas posteriormente na Seção 4. Os seletores baseados em PI tendem a superestimar a largura de banda em distribuições mais complexas e os seletores baseados em CV tendem a ter maiores variâncias em distribuições unimodais com poucas amostras. Portanto, a primeira estratégia do algoritmo é verificar se a distribuição é unimodal ou multimodal através do método *Significant Zero crossings of derivative* (SiZeR) (CHAUDHURI; MARRON, 1999). A Figura 39 mostra a distribuição D1a sendo estimada com 25 amostras pelo KDE de banda fixa em um intervalo de h de 10^{-2} a $10^{1.2}$, e sendo destacada a estimacão feita pela escolha de SJ. A Figura 40 mostra o método SiZeR *Map*, que como padrão mostra a cor azul para derivadas positivas, a cor vermelha para derivadas negativas, a cor roxa para transições desprezíveis de derivada e a cor cinza para regiões onde existe considerável flutuação estatística (regiões com menos de 5 amostras). Basicamente, o método utiliza as informações das derivadas de cada estimacão da Figura 39 para mostrar visualmente o cruzamento da derivada em zero, deixando de ser positiva (azul) e se tornando negativa (vermelha).

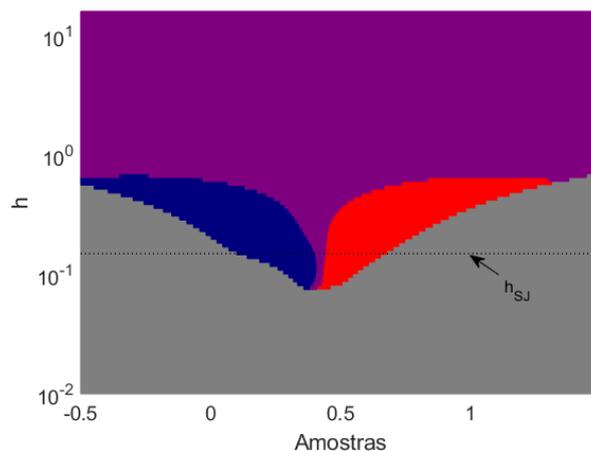
Para melhor fixar o método, as Figuras 41 e 42 mostram os resultados para a distribuição D3c. A Figura 41 mostra a estimacão para o mesmo intervalo de h anterior e a Figura 42 mostra graficamente as transições no cruzamento da derivada em zero. Adicionalmente a esse algoritmo foi implementado uma contagem automática do número de transições de pico, baseado nas transições encontradas no seletor SJ. A escolha do SJ nesse algoritmo se deu pela robustez desse método em diversas distribuições obtendo pouca variância (quesito importante para esta etapa) e nas distribuições onde o método

Figura 39 – Estimação de D1a via KDE fixo



Fonte: Elaborada pelo autor (2020).

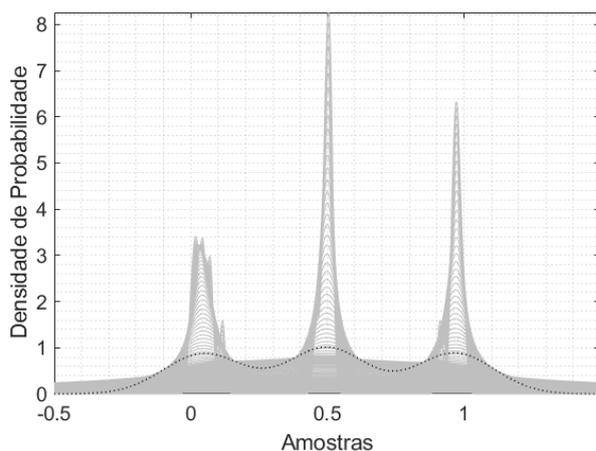
Figura 40 – SiZeR Map para D1a



Fonte: Elaborada pelo autor (2020).

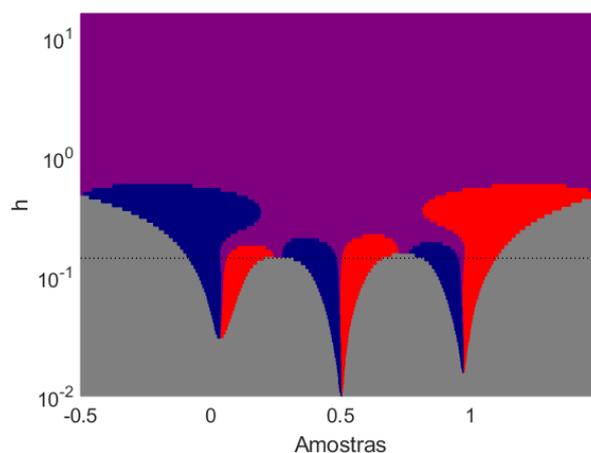
superestima h , ainda é possível discernir os picos.

Figura 41 – Estimação de D3c via KDE fixo.



Fonte: Elaborada pelo autor (2020).

Figura 42 – SiZeR Map para D3c

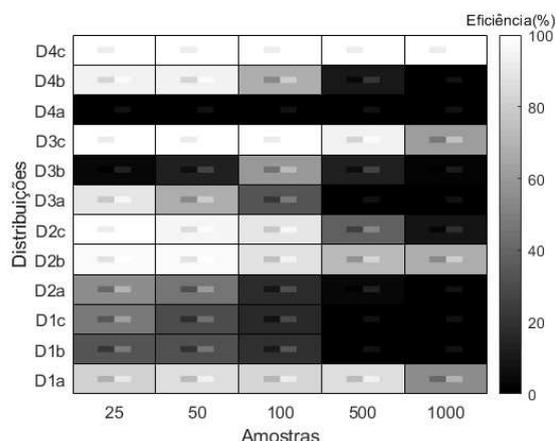


Fonte: Elaborada pelo autor (2020).

O método SiZeR apresentou resultados extremamente resilientes na escolha do número de picos em conjuntos acima de 500 amostras, a Figura 43 mostra a eficiência na identificação do número de picos feita através de um algoritmo que extrai os picos diretamente da PDF, cada quadrante representa o valor da eficiência para as 12 distribuições e para diferentes conjuntos de amostras, no centro de cada quadrante é possível observar um subquadrante com os valores máximos/mínimos que o mesmo pode assumir de acordo com o desvio padrão robusto. A Figura 44 é obtida através do método SiZeR, sendo possível avaliar que em uma análise automática, com picos dificilmente discerníveis, essa etapa deve ser utilizada apenas em conjuntos acima de 500 amostras.

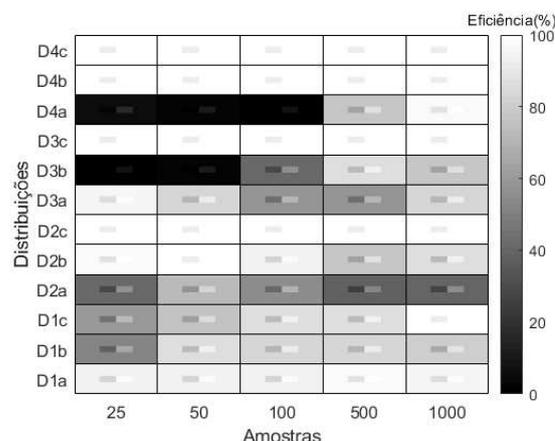
Caso o número de transições seja único, a distribuição teoricamente é considerada unimodal iniciando outra etapa, que será responsável por descobrir se a distribuição é

Figura 43 – Eficiência da escolha do número de picos via PDF estimada.



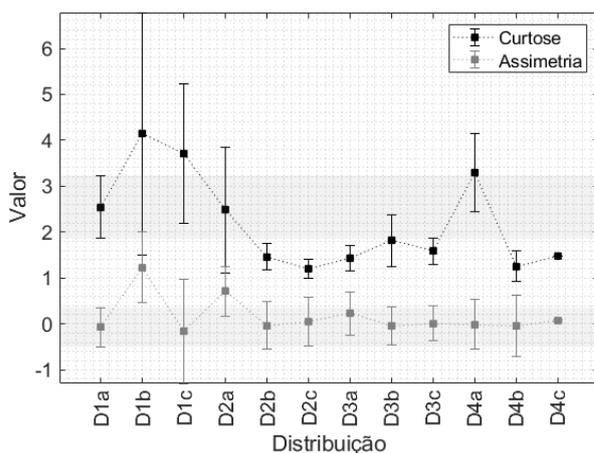
Fonte: Elaborada pelo autor (2020).

Figura 44 – Eficiência da escolha do número de picos via SiZeR.



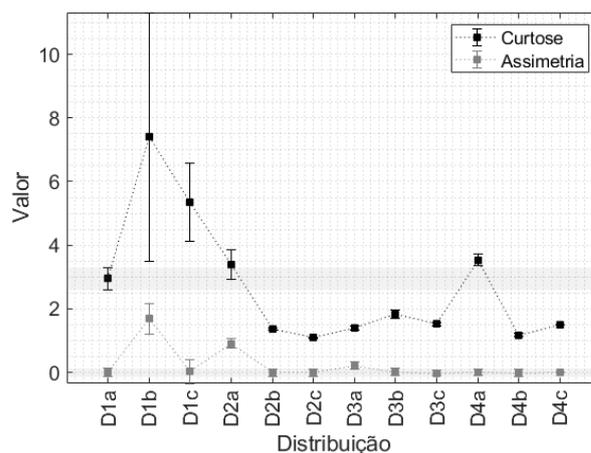
Fonte: Elaborada pelo autor (2020).

Figura 45 – Assimetria e Curtose das distribuições: Conjunto com 25 amostras.



Fonte: Elaborada pelo autor (2020).

Figura 46 – Assimetria e Curtose das distribuições: Conjunto com 500 amostras.



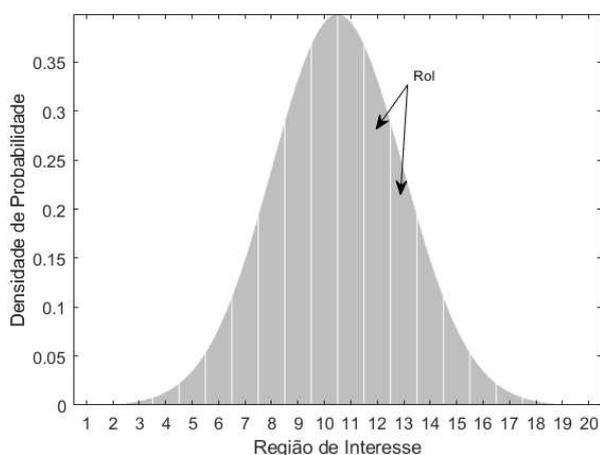
Fonte: Elaborada pelo autor (2020).

Gaussiana ou não. Portanto, é feito um teste de curtose e assimetria nas amostras para descobrir suas características de Normalidade. A Figura 45 mostra o teste feito para todas as distribuições com conjuntos de 25 amostras, sendo possível perceber que existe uma zona de interseção que classifica D2a e D3b como Normal (assimetria ≈ 0 e curtose ≈ 3). A Figura 46 mostra que 500 amostras são suficientes para diferenciar bem essas distribuições, sendo indicado utilizar esta etapa apenas em conjuntos acima de 500 amostras. Para casos inferiores a 500 amostras aconselha-se utilizar o algoritmo OSCV, devido a sua robustez em diferentes situações.

(3) Ajuste de Banda Variável: Esta etapa utiliza as informações contidas na área entre as segmentações imaginárias do intervalo das amostras, denominadas regiões de interesse, para decidir automaticamente a melhor largura de banda fixa h utilizada no cálculo

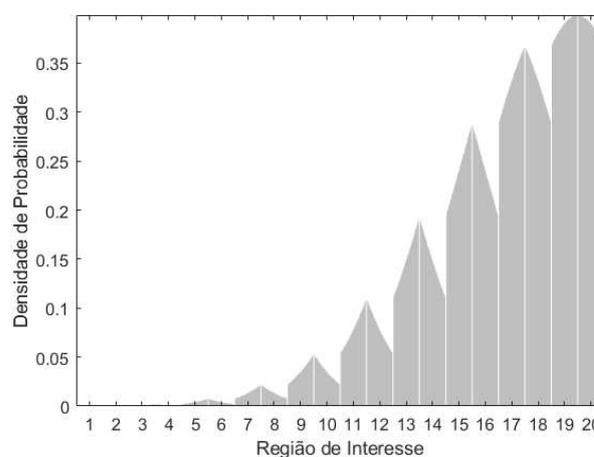
da largura de banda variável h_i . A Figura 47, mostra a distribuição D1a dividida em 20 *Region of Interest* (ROI)s, organizadas em ordem crescente em relação a variável aleatória. A Figura 48 representa a mesma distribuição D1a com as Rols ordenadas em ordem crescente com a probabilidade e a Figura 49 mostra as Rols ordenadas de acordo com a primeira derivada. É importante notar que é feita uma média de cada parâmetro (variável aleatória, probabilidade ou primeira derivada) dentro da respectiva Rol e posteriormente essa média determina a ordem da Rol.

Figura 47 – Rol odernada através da variável aleatória



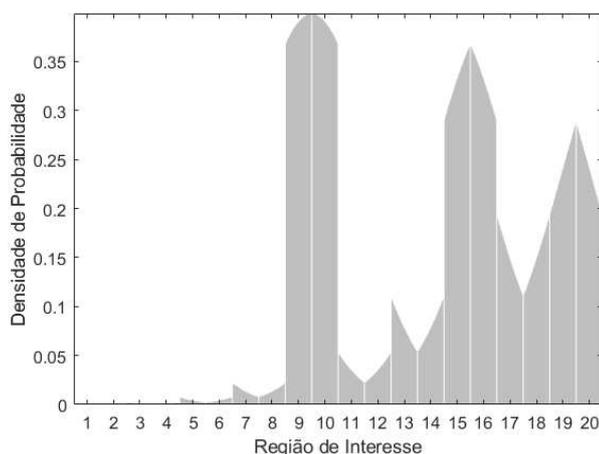
Fonte: Elaborada pelo autor (2020).

Figura 48 – Rol odernada através da probabilidade



Fonte: Elaborada pelo autor (2020).

Figura 49 – Rol odernada através da derivada



Fonte: Elaborada pelo autor (2020).

Na seção 4 será visto que existe dependência entre probabilidade/derivada com o erro de estimação dos seletores. Por exemplo, alguns seletores são capazes de descrever derivadas suaves com maior precisão, outros obtém melhor performance em derivadas rápidas, com isso será possível ponderar de maneira consciente qual largura de banda utilizar em cada situação. Suponhamos um conjunto com as distribuições mais suaves (D1a,

capaz de lidar e classificar essas características está fora do escopo desse trabalho. Portanto, foram escolhidos os seletores com as melhores performances médias capazes de lidar de forma satisfatória com as diferentes complexidades das distribuições. São eles:

- Probabilidade: Baixa = SJ; Alta = OSCV.
- Derivada: Baixa = SJ; Alta = BCV2.

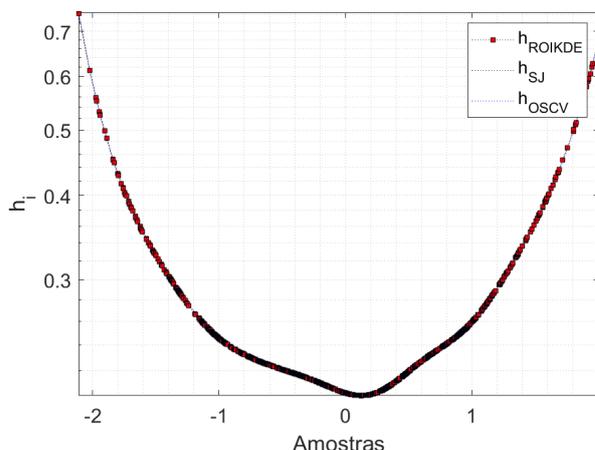
A Equação (3.3) define as características da largura de banda variável calculada pelo método ROIKDE:

$$h_i^{ROIKDE} = \frac{1}{2} \left\{ \begin{array}{l} [h_i^{OSCV} \cdot f_i^p + h_i^{SJ} \cdot (1 - f_i^p)] + \\ [h_i^{BCV2} \cdot f_i^{d'} + h_i^{SJ} \cdot (1 - f_i^{d'})] \end{array} \right\} \quad (3.3)$$

Onde h_i^{OSCV} representa o h variável de cada amostra i gerada pelo seletor OSCV, logo h_i^{SJ} e h_i^{BCV2} representam o h de cada amostra para os seletores SJ e BCV2, respectivamente. A quantidade f_i^p representa um fator relativo a estimação da probabilidade da amostra i , ou seja, é feita uma estimação prévia utilizando o KDE de banda fixa com o h do estágio anterior (2-Ajuste de Banda Fixa), e essa estimação é normalizada pelo maior valor de probabilidade. Logo o fator $f_i^{d'}$ é calculado com o auxílio do módulo da derivada da mesma estimação do KDE de banda fixa, e sua normalização é feita através de seu valor máximo. O efeito desse fator de probabilidade/derivada é fazer o h fixo encontrado na etapa (2-Ajuste de Banda Fixa) tender aos valores desejados de acordo com a ponderação da probabilidade/derivada.

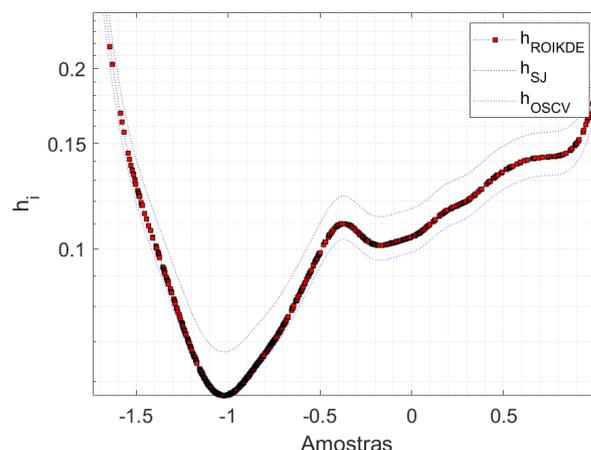
A Figura 54 mostra o funcionamento do algoritmo ROIKDE para a distribuição D1a. Nesse caso percebemos que os seletores SJ e OSCV selecionaram valores de h_i bem próximos, com isso a ponderação de f_i^p e $f_i^{d'}$ mantiveram o h_i do ROIKDE próximo aos dois seletores. Na Figura 55 é possível observar OSCV com larguras de banda menores do que o SJ, com isso o ROIKDE tende a se aproximar do seletor OSCV nas regiões de maiores primeiras derivadas e alta probabilidade, aproveitando as melhores características de cada seletor.

Figura 54 – Escolha da largura de banda via ROIKDE para Distribuição D1a.



Fonte: Elaborada pelo autor (2020).

Figura 55 – Escolha da largura de banda via ROIKDE para Distribuição D2a.



Fonte: Elaborada pelo autor (2020).

(4) Estimação via VKDE: Após a escolha automática da largura de banda fixa e o ajuste da banda variável, a estimação será desenvolvida de acordo com a teoria exposta na Seção 3.2.4.2 baseada no método VKDE.

3.3 ANÁLISE DA ESTIMAÇÃO

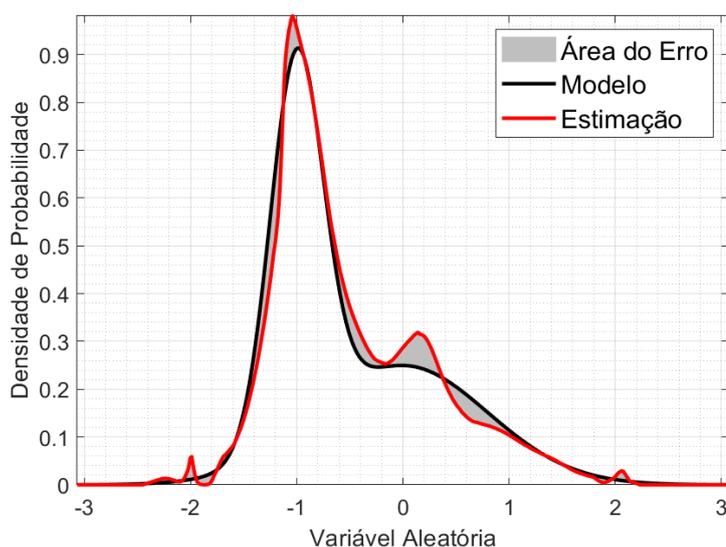
Para avaliar a estimação dos métodos, vistos anteriormente, foram escolhidas três abordagens: Área do Erro, RoIMap e um teste de terceiro e quarto momento central.

3.3.1 Área do Erro

Uma questão bastante relevante em estudos de estimação de densidade é a escolha da métrica/similaridade ideal para avaliar a performance dos estimadores. Em (SOUZA; COSTA; NÓBREGA, 2017) diversas métricas foram avaliadas no contexto de estimação, sendo possível constatar que a maioria das métricas baseadas em L_1 geralmente tendem a avaliar todas as regiões ao longo da variável aleatória com o mesmo peso. Já as métricas baseadas em L_2 tendem a dar maior peso para as regiões com maior probabilidade. Portanto, foi decidido utilizar como medida de performance a área do erro, que é uma medida da família L_1 . Essa escolha foi embasada em três fatores: peso equânime ao longo de toda variável aleatória, simplicidade de implementação e referencial intuitivo. Ou seja, como esse trabalho destina-se a avaliar estimatórias de PDFs, que possuem como uma de suas características a integral de sua área ser unitária, a escolha da área promove um senso de proporcionalidade mais forte do que em medidas como MISE, por exemplo, devido ao seu limite inferior possível ser 0 e seu limite superior possível ser 2.

A Figura 56 mostra como é feito o cálculo da área entre modelo e estimação. A região sombreada representa a área onde a estimaco obteve valores diferentes do modelo. Por fim, a melhor estimaco é a que possui menor valor de área total de erro, que ocorrerá quando a estimaco estiver o mais próximo possível do modelo. O cálculo da área do erro será feito através da Integral de Riemann, utilizando 10^5 pontos de *grid*, como definido anteriormente.

Figura 56 – Área do erro entre modelo discreto da distribuico D2a e sua estimaco via KDE com 25 amostras.



Fonte: Elaborada pelo autor (2020).

3.3.2 *Region of Interest Map (RoIMap)*

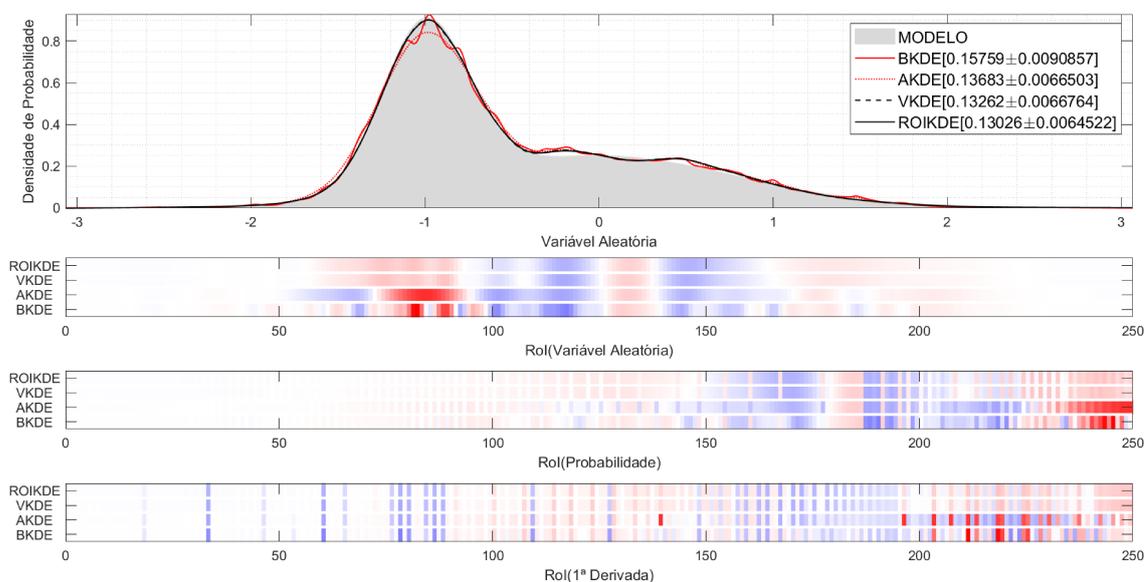
Essa abordagem foi desenvolvida, neste trabalho, com o intuito de auxiliar nas análises comparativas entre os estimadores e possui dois objetivos principais: (1) avaliar o valor da área do erro nas Rols e (2) avaliar a dependência entre a probabilidade (ou derivada) com a área do erro nas Rols.

Definições: O RoIMap tem como representaco gráfica duas cores, azul quando o valor da estimaco é maior do que o modelo e vermelho quando a estimaco é menor do que o modelo. As cores estão normalizadas pelo maior valor absoluto de erro, portanto, a intensidade da cor reflete o valor da área do erro naquela Rol em relaco a Rol com maior área do erro, logo, quanto mais clara a cor, menor será a área do erro.

As Figuras 47, 48 e 49 mostram que as Rols são áreas compreendidas entre as subdivises imaginárias equidistantes no eixo da variável aleatória. Com isso, é possível avaliar separadamente o que ocorre dentro de cada regio e perceber a sua sensibilidade

quanto a variação da probabilidade ou primeira derivada. A Figura 57, superior, mostra a distribuição D2a juntamente com a estimação feita por quatro KDEs de banda variável utilizando 500 amostras. Nessa figura é possível perceber a dificuldade de se avaliar visualmente os detalhes das estimações e proporções entre os erros. Abaixo da distribuição temos a avaliação através do RoIMap, e suas representações ordenando as Rols pela variável aleatória (análogo a Figura 47), probabilidade (análogo a Figura 48) e primeira derivada (análogo a Figura 49). Através dessa representação é possível perceber que os métodos ROIKDE e VKDE possuem estimações bem próximas, o método AKDE teve mais dificuldade em representar o pico próximo ao valor -1 na variável aleatória e o método BKDE parece flutuar mais do que os demais métodos. Essa nova perspectiva nos dá uma visão mais detalhada ao longo de toda variável aleatória, enquanto os valores de média e desvio padrão contidos na legenda apresentam uma visão geral da estimação. Outro fator relevante é a capacidade de avaliar a dependência do erro de área em relação a probabilidade e primeira derivada, neste caso parece existir uma maior dependência em relação a probabilidade do que pela derivada, ou seja, o efeito de Poisson teve maior influência na estimação do que as primeiras derivadas, que neste caso são relativamente suaves.

Figura 57 – Ferramenta RoIMap utilizada na distribuição D2a.



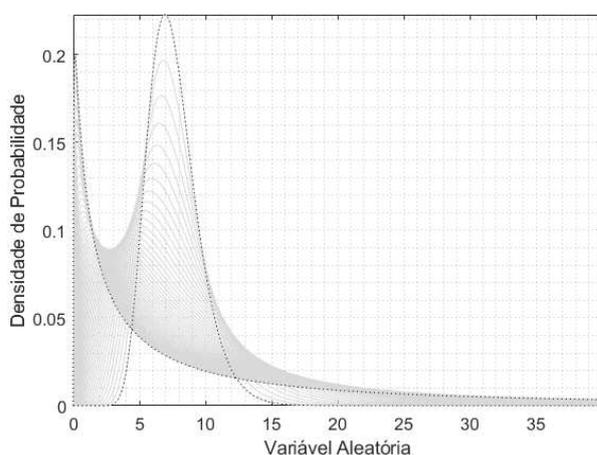
Fonte: Elaborada pelo autor (2020).

Vale ressaltar que a análise e entendimento mais aprofundado das estimações feitas com o RoIMap possibilitaram a criação do algoritmo ROIKDE.

3.3.3 Teste de terceiro e quarto momento central

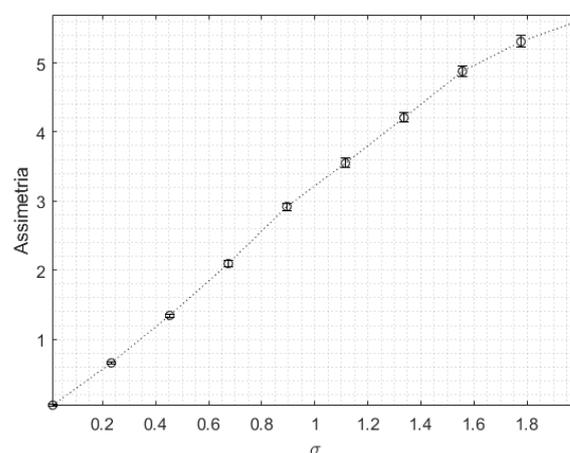
Os estimadores serão avaliados através da sua sensibilidade a variação do 3º (assimetria) e 4º (curtose) momento central. Ou seja, serão avaliadas diversas distribuições dentro de um determinado intervalo de curtose e assimetria. Para o teste do 3º momento será utilizada a distribuição Log-Normal. A Figura 58 mostra a diferença de formato que pode ser verificada através da variação de σ na equação da Log-Normal, quanto menor o valor de σ mais próxima de uma Gaussiana a distribuição se torna, caso contrário a distribuição tenderá a uma exponencial. A Figura 59 mostra os valores de σ utilizados no teste e seus respectivos valores de assimetria.

Figura 58 – Variação de assimetria para uma Distribuição Log-Normal.



Fonte: Elaborada pelo autor (2020).

Figura 59 – Gráfico de assimetria pela variação do parâmetro σ .



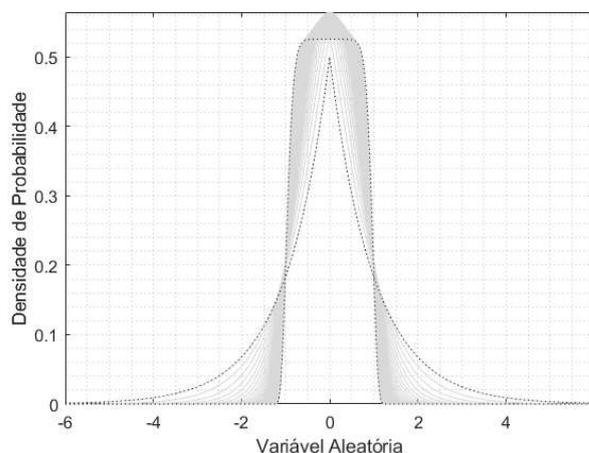
Fonte: Elaborada pelo autor (2020).

No teste de 4º momento foi utilizado a distribuição GGD, devido a sua capacidade de variação de curtose. A Figura 60 mostra o efeito na distribuição ocasionado pela variação do parâmetro de forma ρ , quanto menor o valor de ρ a distribuição mais a distribuição GGD se aproxima de uma distribuição de Laplace, e quanto maior o valor de ρ a distribuição GGD tenderá a uma distribuição uniforme. Na Figura 61 temos os valores de ρ utilizados no teste do 4º momento e seus respectivos valores de curtose.

3.4 CLASSIFICAÇÃO

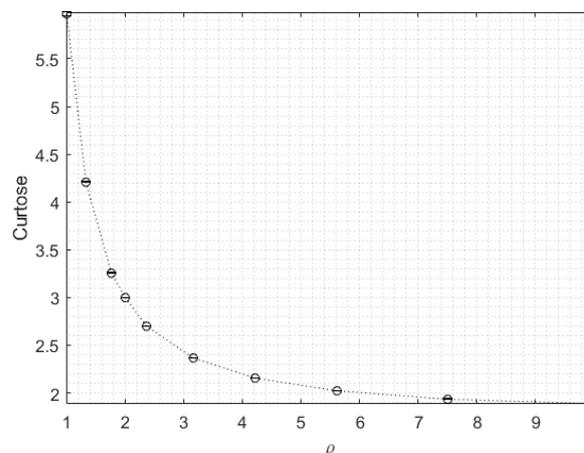
Nesta seção serão vistos os conceitos utilizados para classificação de dois casos representativos: utilizando amostras simuladas neste trabalho e amostras extraídas de uma realidade de identificação de partículas. Os dados simulados aqui serão construídos utilizando como função geradora as distribuições vistas anteriormente, já o problema de

Figura 60 – Variação de curtose para uma Distribuição GGD.



Fonte: Elaborada pelo autor (2020).

Figura 61 – Gráfico de curtose pela variação do parâmetro ρ .



Fonte: Elaborada pelo autor (2020).

identificação de partículas utilizará os dados gerados pelo *software* Geant4. Como classificador utilizaremos a verossimilhança *naive* (ZHU; NANDI, 2015) e as suas densidades de probabilidade serão construídas utilizando o KDE com largura e banda variável.

3.4.1 Conjunto de Dados

Os conjuntos de dados serão divididos em duas classes: sinal e ruído. As definições sobre cada conjunto de dados serão mostradas a seguir:

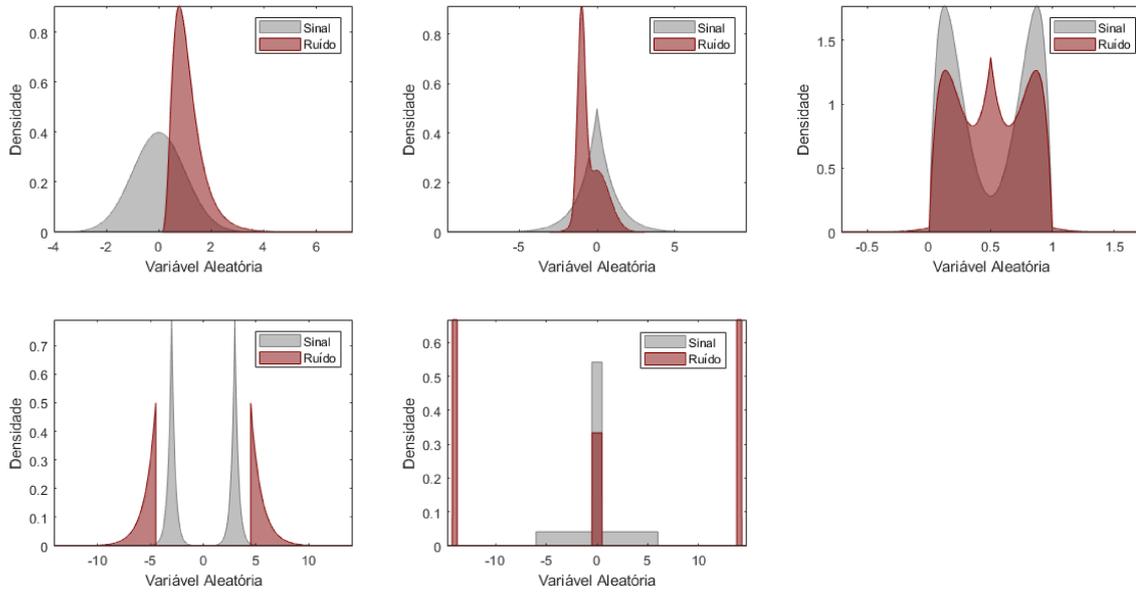
3.4.1.1 Dados simulados neste trabalho

Será construída uma realidade multivariada com cinco dimensões, utilizando as distribuições apresentadas na Seção 3.1. Para a classe de sinais, teremos as distribuições D1a, D1c, D2b, D2c e D4a. Para a classe de ruído, teremos as distribuições D1b, D2a, D3b, D4b e D4c. As respectivas densidades, a área de interseção entre sinal e ruído e o grau de dificuldade de cada distribuição serão apresentados na Figura 62. Vale destacar que todas as variáveis aleatórias de cada dimensão são independentes, sendo ideal para o método de verossimilhança *naive*.

3.4.1.2 Identificação de partículas

Será observada uma realidade com 13 variáveis discriminantes oriundas de um problema de identificação de partículas, gerados pelo *software* Geant4. Uma estimativa relativamente suave dessas variáveis é apresentada na Figura 63 e mais detalhes sobre esse conjunto de dados podem ser encontrados em (SOUZA, 2015). É importante salientar

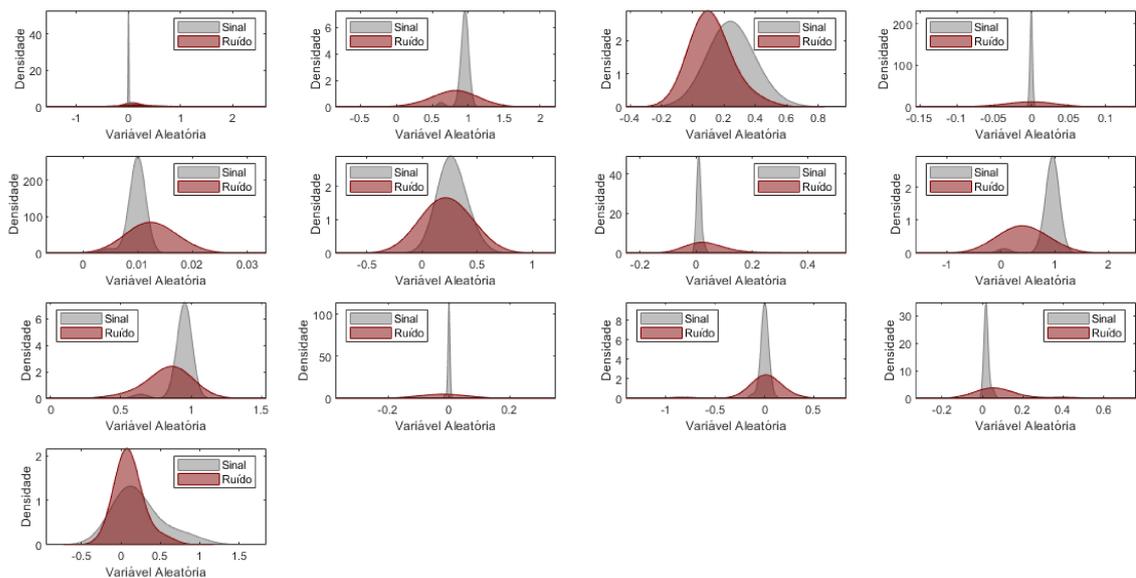
Figura 62 – Dados simulados à partir das densidades vistas anteriormente.



Fonte: Elaborada pelo autor (2020).

que nessa realidade existe um certo grau de dependência entre as variáveis aleatórias, que será desconsiderada na verossimilhança *naïve*. No mesmo trabalho citado acima existem alternativas para diminuir o impacto dessa dependência. Entretanto, o objetivo dessa tese é melhorar a densidade conjunta das classes através dos métodos disponíveis na literatura e a nova abordagem proposta aqui, o ROIKDE.

Figura 63 – Densidades de um problema de física de partículas extraída de uma simulação realizada por Geant4.



Fonte: Elaborada pelo autor (2020).

3.4.2 Classificador

As estimações vistas anteriormente podem ser combinadas em um classificador de verossimilhança. Neste trabalho utilizaremos o conceito de verossimilhança *naive*, desconsiderando as dependências entre as variáveis aleatórias e construindo a probabilidade conjunta através da multiplicação das densidades marginais.

O classificador baseado em verossimilhança é uma abordagem bastante comum em análise multivariada e permite a avaliação simultânea de todas os atributos antes de rotular determinada classe. Este método utiliza a informação da densidade conjunta das classes, combinadas em uma variável discriminante, para definir qual a probabilidade de determinada amostra pertencer a uma das classes. Portanto, a máxima verossimilhança é a melhor representação possível da PDF conjunta dado o parâmetro θ . Sua fórmula geral é conhecida pela Equação 3.4:

$$L_s(\theta) = P_s(x|\theta) \quad e \quad L_b(\theta) = P_b(x|\theta), \quad (3.4)$$

onde, L_s e P_s representam a verossimilhança do sinal e probabilidade conjunta de sinal, respectivamente, e L_b e P_b representam a verossimilhança do ruído e probabilidade conjunta do ruído, respectivamente. Neste contexto, as amostras em x são fixas e θ pode ser ajustado, buscando a máxima verossimilhança.

Se o problema for multivariado, e as variáveis aleatórias forem independentes, é possível fazer uma simplificação na formulação da verossimilhança através da multiplicação das densidades de probabilidade de cada dimensão, como pode ser visto nas Equações 3.5 e 3.6. Após a definição das quantidades L_s e L_b será possível combinar suas representações em uma variável discriminante, mostrada na Equação 3.7.

$$L_s(x) = \prod_{i=1}^n P_{s,i}(x_i) \quad (3.5)$$

$$L_b(x) = \prod_{i=1}^n P_{b,i}(x_i) \quad (3.6)$$

$$dL = \frac{L_s}{L_s + L_b}, \quad (3.7)$$

onde $P_{s,i}(x_i)$ e $P_{b,i}(x_i)$ são as probabilidades associadas a cada uma das n amostras da variável aleatória de sinal e ruído, respectivamente. L_s e L_b são os valores discretos da multiplicação das probabilidades de cada variável aleatória e dL o discriminante responsável por rotular as amostras.

3.4.3 Análise da classificação

Para avaliar a classificação realizada pelo método de verossimilhança, utilizaremos duas abordagens distintas: *Receiver Operating Curve* (ROC) e critério Soma-Produto (SP).

- **ROC:** Utilizada quando deseja-se compreender a relação entre eficiência de detecção e falso alarme (ou rejeição de falso positivo) em todos os pontos de operação do algoritmo avaliado. Mais detalhes sobre esse clássico avaliador pode ser encontrado em (FAWCETT, 2006).
- **SP:** este critério, também utilizado na discriminação binária, representa o equilíbrio entre a probabilidade de detecção de sinal e a probabilidade de detecção de ruído. Sua fórmula é dada pela equação 3.8:

$$SP = 100\% \sqrt{\sqrt{D_{sg} D_{bg}} \left(\frac{D_{sg} + D_{bg}}{2} \right)} \quad (3.8)$$

onde, D_{sg} é a probabilidade de detecção de sinal e D_{bg} é a probabilidade de detecção de ruído.

4 RESULTADOS

Após a construção dos algoritmos seletores de largura de banda, vistos na Seção 3.2, definição dos métodos para avaliação da estimação, propostos na Seção 3.3 e a breve explanação sobre classificação mostrada na Seção 3.4, serão apresentados os resultados deste trabalho. Os resultados serão divididos em dois assuntos: (1) Estimação de Densidades e (2) Classificação.

4.1 ESTIMAÇÃO DE DENSIDADES

Essa etapa se dividirá entre estimadores com largura de banda fixa e estimadores com largura banda variável. Além disso, devido ao volume de resultados, foi escolhido agrupar as densidades para facilitar a análise. Esses grupos estão divididos de acordo com características em comum, que impactam de formas distintas os seletores. Os grupos se encontram na Tabela 19.

Tabela 19 – Divisão dos grupos de densidades.

Grupo	Densidades	Característica
G0	Todas	Distribuições unimodais, bimodais, trimodais, derivadas rápidas e lentas, e descontinuidades.
G1	D1a, D1b e D1c	Distribuições unimodais.
G2	D2a, D2b e D3b	Distribuições bimodais suaves e trimodal pouco esparsa.
G3	D2c, D3a e D3c	Distribuições bastante esparsas.
G4	D4a, D4b e D4c	Distribuições com descontinuidades.

Fonte: Elaborada pelo autor (2020).

4.1.1 Largura de Banda Fixa

Nesta primeira etapa utilizaremos os métodos de Histograma, PF, ASH e KDE na estimação das densidades através dos seus respectivos seletores de largura banda fixa. Os testes realizados podem ser divididos em:

1. **Avaliação dos seletores:** Nesta etapa será avaliada a área do erro de cada grupo, através da ferramenta gráfica *boxplot*.
2. **Comparação entre os estimadores:** Será feita uma comparação direta entre os melhores seletores de largura de banda para cada estimador, através da mediana e desvio padrão robusto dos estimadores.
3. **Testes de 3º e 4º momento:** Será avaliada a performance dos seletores utilizados no KDE, variando a curtose e assimetria das duas distribuições citadas na Seção 3.3.3.

4.1.1.1 Avaliação dos Seletores

Os estimadores Histograma, PF e ASH serão analisados conjuntamente na Seção 4.1.1.2, devido ao fato de compartilharem dos mesmos seletores de largura de banda, salvo algumas alterações nos seletores Scott e LHM, como visto anteriormente na Seção 3.2. Logo após, na Seção 4.1.1.3 o método de KDE será avaliado, utilizando seus respectivos seletores de largura de banda.

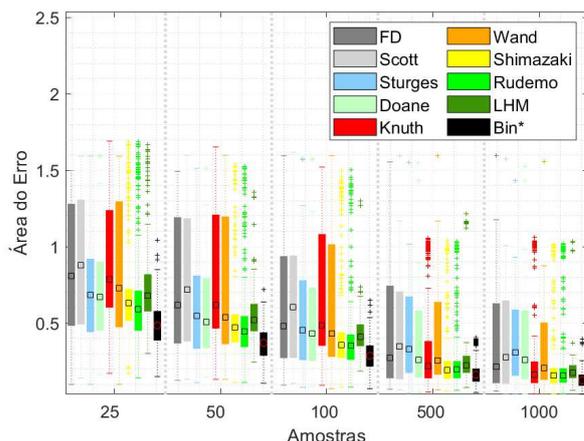
4.1.1.2 Histograma, PF e ASH

Nesta seção, para evitar a redundância nas explicações, serão abordadas os cinco grupos: G0, contendo todas as distribuições; G1, que possui distribuições unimodais; G2, possui distribuições bimodais com derivadas suaves e uma distribuição Trimodal pouco esparsa; G3, possui distribuições com derivadas rápidas e bastante esparsas e G4, com distribuições que possuem descontinuidades. Qualquer particularidade das distribuições individuais será frisada, e suas respectivas figuras poderão ser encontradas no Apêndice C.

As Figuras 64, 65 e 66 mostram diferentes estimadores não-paramétricos e seus respectivos seletores do número de *bins* sendo expostos a todas as distribuições vistas na Seção 3.1. Portanto, esse teste visa avaliar a resiliência/adaptabilidade dos seletores a essas distribuições e seus graus de dificuldade. Vale ressaltar que o método *bin** representa o *bin* ótimo, ou seja, o melhor valor possível do número de *bins* encontrado com auxílio da função geradora. Logo na Figura 64, para o Histograma, observamos os seletores de Shimazaki e Rudemo com as melhores performances gerais. Além disso é percebido que os seletores FD, Scott, Sturges, Doane, Knuth e Wand possuem grande variância, embora à partir de 500 amostras Knuth melhore consideravelmente sua dispersão. O seletor LHM obteve um desempenho intermediário, convergindo mais lentamente para performances próximas ao Shimazaki e Rudemo com aproximadamente 1000 amostras. A Figura 65, para o PF, mostra o seletor LHM, agora com interpolação linear, com a estimativa mais próxima aos dois melhores seletores, Rudemo e Shimazaki. Na Figura 66, com o estimador ASH, o seletor LHM melhora ainda mais sua performance, chegando a ser o melhor seletor até 50 amostras. De forma geral, os seletores iterativos Rudemo, LHM e Shimazaki obtiveram os melhores resultados. Os seletores baseados em variáveis de escala, Scott e FD, obtiveram os piores resultados e com grande variância, devido ao fato de distribuições muito diferentes da Normal e multimodais esparsas atrapalharem a estimação desses parâmetros. Os seletores oriundos da Binomial, Sturges e Doane, foram projetados

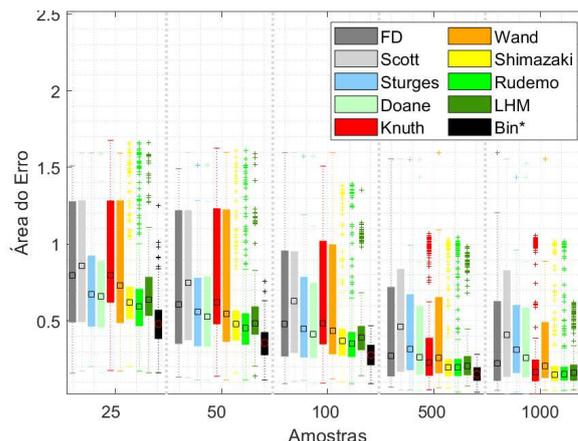
com presunção de Normalidade, portanto não se adaptam bem a derivadas rápidas e distribuições multimodais. A aproximação do “termo problemático” feita por Wand encontrou dificuldade em distribuições multimodais. O seletor iterativo de Knuth encontrou dificuldade de convergência abaixo de 500 amostras, até mesmo em distribuições unimodais.

Figura 64 – Performance do grupo G0 utilizando o estimador Histograma.



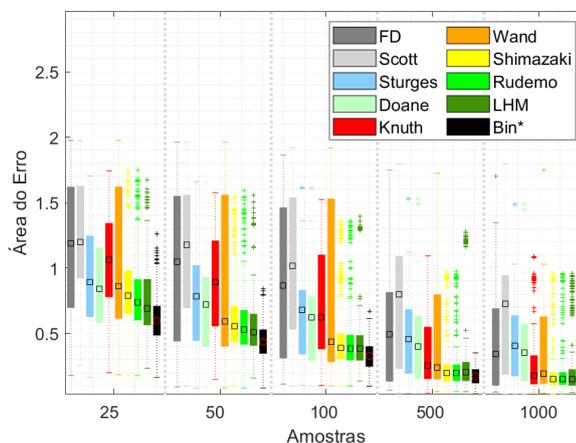
Fonte: Elaborada pelo autor (2020).

Figura 65 – Performance do grupo G0 utilizando o estimador PF.



Fonte: Elaborada pelo autor (2020).

Figura 66 – Performance do grupo G0 utilizando o estimador ASH.



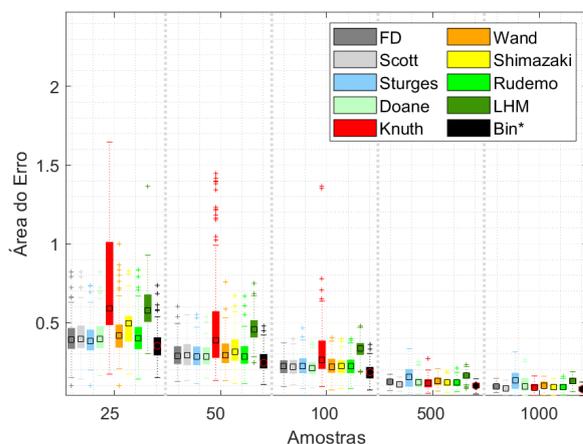
Fonte: Elaborada pelo autor (2020).

O Grupo G1 busca avaliar os seletores em uma realidade unimodal, agrupando os resultados das estimações efetuadas nas densidades D1a, D1b e D1c na mesma análise. As Figuras 67, 68 e 69, diferentemente do grupo G0, mostram uma performance mais homogênea entre os seletores. Na Figura 67 temos os seletores Knuth e LHM com dificuldade de convergência até 100 amostras e os demais seletores com performances bem parecidas. O seletor LHM superestimou o número de bins nas três distribuições e o Knuth subestimou. O seletor Knuth, através de sua função custo extremamente ruidosa, encontra bastante dificuldade em lidar com poucas amostras. Na distribuição D1a, com baixa

rugosidade, todos os seletores convergiram para estimações bem próximas com o aumento das amostras, exceto o seletor LHM. Para a distribuição D1b, assimétrica e com derivadas rápidas e lentas, os seletores Sturges e Doane subestimaram e LHM superestimou os *bins*, piorando a performance dos seletores nessa realidade; A alteração feita no Doane, considerando a assimetria, colocou o seletor dentre os melhores em D1b. Já para D1c, distribuição simétrica e com rápida transição de pico, Sturges e Doane obtiveram as piores performances com o aumento das amostras, devido a subestimarem o número de *bins*. Shimazaki e Rudemo, apesar de serem seletores iterativos que dependem das amostras, apresentaram as melhores performances gerais, embora Shimazaki tenha mostrado uma convergência mais lenta do que Rudemo. A Figura 68, para o PF, se difere no seguinte quesito, o seletor LHM melhorou sua performance relativa, devido aos seletores (exceto LHM e Scott) serem projetados para o Histograma, piorando relativamente a performance geral. De forma análoga temos a Figura 69. Até 100 amostras existe uma queda na performance geral dos seletores, com destaque negativo para o seletor Scott que teve dificuldade em lidar com a assimetria da distribuição D1b, devido a utilização do desvio padrão como variável de escala.

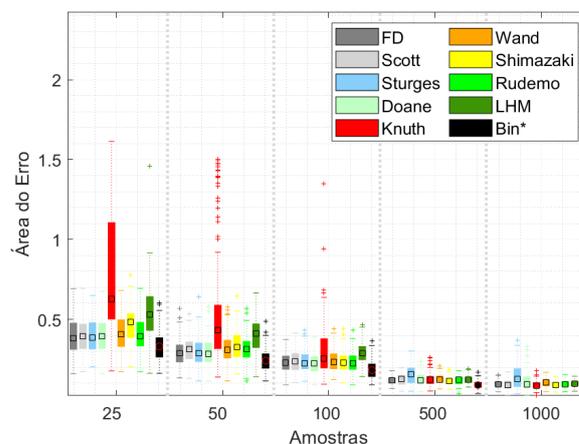
O Grupo G2 avalia as distribuições D2a, D2b e D3b, que são multimodais pouco esparsas, embora contemplem derivadas rápidas, lentas e transição de pico rápida. As Figuras 70, 71 e 72 mostram a performance geral dos seletores pior do que o G1, afinal, o grau de dificuldade dessa estimacão é maior. Na Figura 70 é mostrada a grande dispersão do seletor Knuth e sua dificuldade em convergir, principalmente na distribuição D2b. Nas demais distribuições do G2 Knuth converge com 500 amostras; LHM superestima o número de *bins* e possui desempenho intermediário; Shimazaki encontrou mais dificuldade do que Rudemo com poucas amostras, principalmente na distribuição D3b; FD e Scott encontraram maior dificuldade em D2b baseando-se em variáveis de escala; Rudemo se mostrou mais resiliente nas três distribuições e em todos os conjuntos de amostras, embora em D2a, com derivadas suaves, Scott e FD tivessem performances ligeiramente melhores. A Figura 71 mostra a convergência de Knuth ainda mais lenta e LHM melhorando sua performance em D2a à partir de 500 amostras; Scott piora sua performance em relação a FD em D2b. A Figura 72 mostra a dispersão de FD, Scott, Sturges e Doane aumentando consideravelmente, principalmente devido aos seletores subestimarem o número de *bins* em D2b, fazendo com que os seletores tivessem perda de performance; Em D2b e D3b, a dificuldade de Shimazaki com poucas amostras parece maior em relação ao Rudemo com o estimador ASH; o seletor Wand figurou dentre os melhores nas distibuições D2b e D3b até 500 amostras. Novamente

Figura 67 – Performance do grupo G1 utilizando o estimador Histograma.



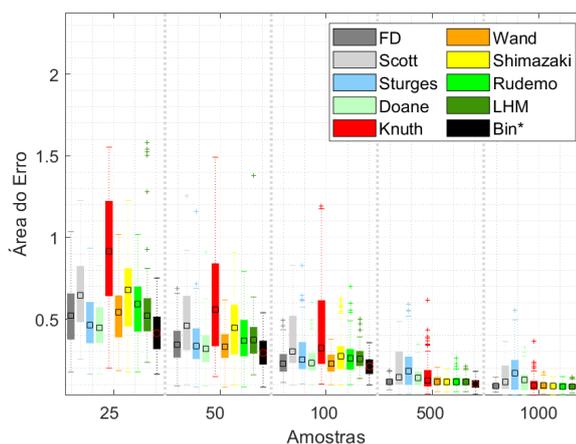
Fonte: Elaborada pelo autor (2020).

Figura 68 – Performance do grupo G1 utilizando o estimador PF.



Fonte: Elaborada pelo autor (2020).

Figura 69 – Performance do grupo G1 utilizando o estimador ASH.

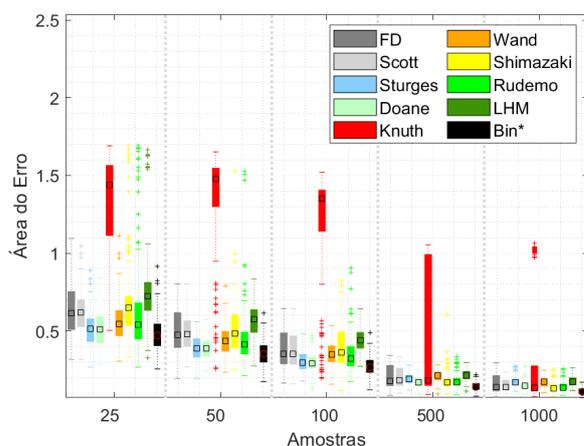


Fonte: Elaborada pelo autor (2020).

o seletor Rudemo se mostrou mais resiliente do que os demais nessa nova realidade, com destaque para o Wand em até 500 amostras.

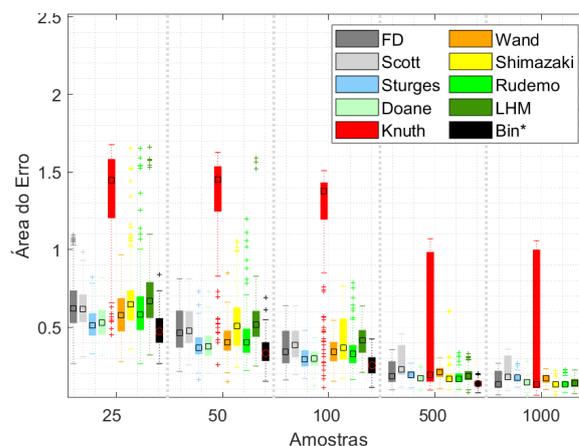
O Grupo G3 considera as distribuições D2c, D3c, e D3a, que são as distribuições com derivadas finitas mais esparsas. Esse teste é interessante pois distorce consideravelmente as variáveis de escala em relação à densidade Normal, afetando drasticamente a performance dos seletores que utilizam essas informações. As Figuras 73, 74 e 75 mostram uma performance geral pior do que o grupo G2, juntamente com o aumento da dispersão dos seletores. Na Figura 73 temos os seletores FD e Scott com as piores performances, influenciada principalmente pelo péssimo desempenho em D2c e D3c. De maneira análoga temos o seletor Wand, embora apresente uma convergência melhor do que os dois anteriores; Sturges e Doane são seletores que conhecidamente tendem a subestimar em distribuições com rugosidade maior do que a Normal, sendo natural o seu fraco desempe-

Figura 70 – Performance do grupo G2 utilizando o estimador Histograma.



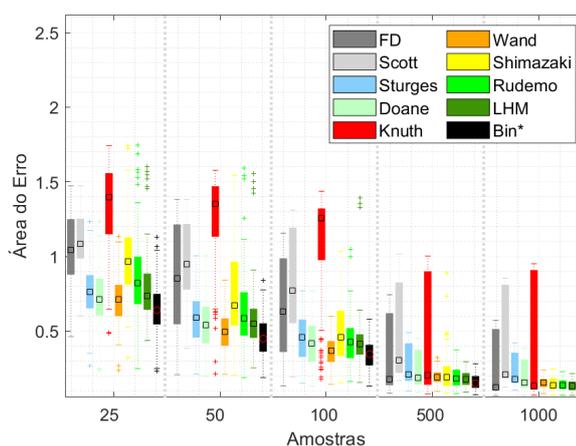
Fonte: Elaborada pelo autor (2020).

Figura 71 – Performance do grupo G2 utilizando o estimador PF.



Fonte: Elaborada pelo autor (2020).

Figura 72 – Performance do grupo G2 utilizando o estimador ASH.

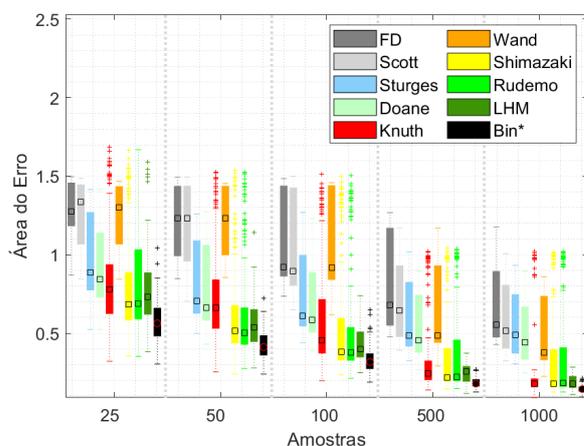


Fonte: Elaborada pelo autor (2020).

nho em distribuições esparsas, neste caso os piores resultados são promovidos por D3c. Shimazaki e Rudemo obtiveram resultados impressionantes em D2c e D3c, entretanto, não conseguem lidar com esparsidade, assimetria, derivadas rápidas e lentas ao mesmo tempo, promovendo resultados com grande dispersão em D3a. Knuth conseguiu convergir para bons resultados em D2c e D3c com 500 amostras; como destaque positivo temos LHM, sua função custo baseada na CDF parece ter conseguido contemplar as diferentes características das distribuições. A Figura 74 mostra o seletor de Scott piorando sua performance relativa e os seletores de Shimazaki e Rudemo diminuindo sua dispersão, com a mesma tendência sendo vista também na Figura 75.

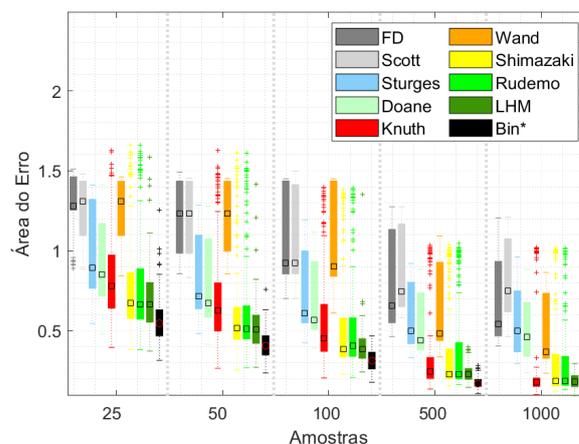
O Grupo G4 contempla as distribuições com derivada infinita, ou seja, com descontinuidades. Além disso será verificado como os seletores lidam com derivadas infinitas em realidades esparsas, pouco esparsa e com derivadas finitas. As Figuras 76, 77 e 78

Figura 73 – Performance do grupo G3 utilizando o estimador Histograma.



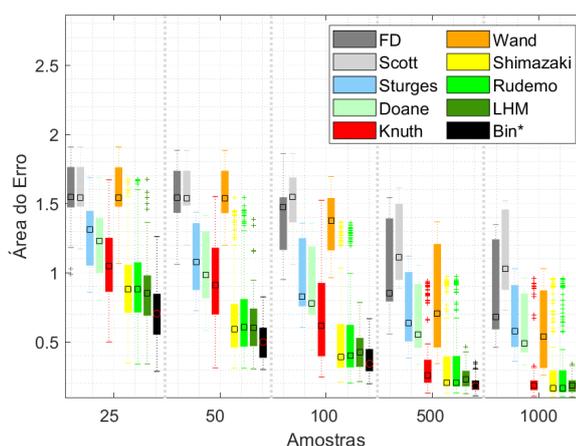
Fonte: Elaborada pelo autor (2020).

Figura 74 – Performance do grupo G3 utilizando o estimador PF.



Fonte: Elaborada pelo autor (2020).

Figura 75 – Performance do grupo G3 utilizando o estimador ASH.

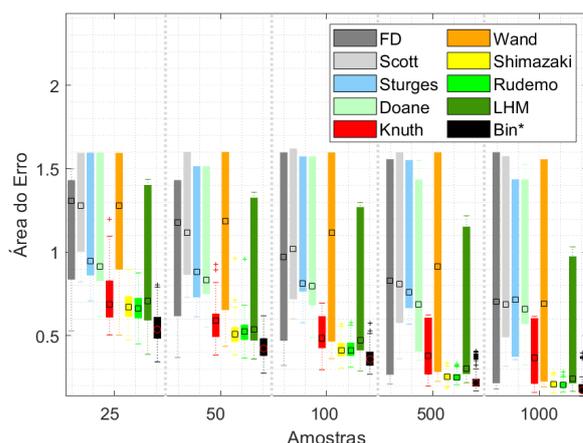


Fonte: Elaborada pelo autor (2020).

mostram uma performance geral pior do que o G3 e com maior dispersão, exceto para os seletores Rudemo e Shimazaki. Na Figura 76 observamos apenas os seletores Shimazaki e Rudemo com uma dispersão relativamente menor, com o seletor de Rudemo ligeiramente melhor. Os demais seletores sofrem demasiadamente com a distribuição D4c, chegando ao ponto de FD, Scott e Wand não convergirem com o aumento de amostras. Apenas os seletores iterativos, Rudemo e Shimazaki (baseados em validação-cruzada e Poisson) chegaram a resultados próximos a binagem ótima. Já na distribuição D4b, os seletores Scott, FD e Wand começaram em um patamar de erro maior do que em D4c, entretanto, conseguiram convergir; Os seletores Knuth e LHM obtiveram performances intermediárias e tanto Rudemo quanto Shimazaki tiveram erro de estimação próximos ao *bin* ótimo. Na distribuição D2a, Scott foi o pior seletor, Knuth, Sturges e Doane foram intermediários, embora Knuth tenha sido melhor até 100 amostras; LHM, Wand, Rudemo e Shimazaki obtiveram os

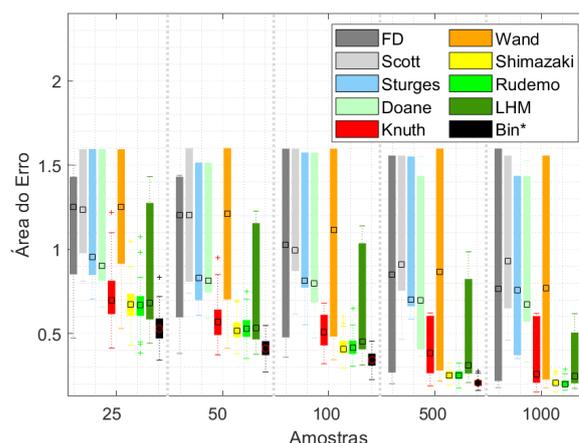
melhores resultados, embora Wand tenha convergido apenas com 50 amostras. A Figura 77 mostra LHM diminuindo a sua dispersão e o seletor Scott piorando relativamente. Já a Figura 78, mostra a piora de Scott sendo acentuada, Wand piorando sua medida de posição e a dispersão de Knuth aumentando.

Figura 76 – Performance do grupo G4 utilizando o estimador Histograma.



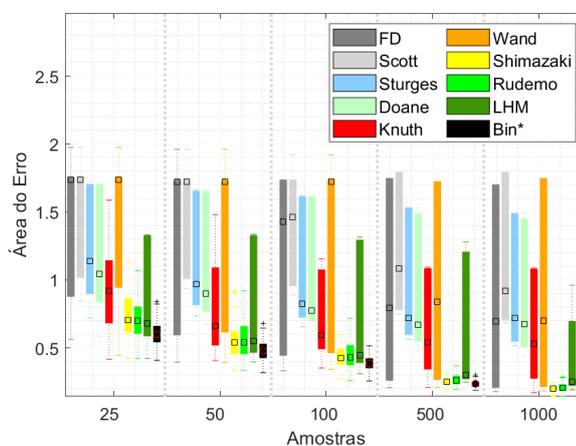
Fonte: Elaborada pelo autor (2020).

Figura 77 – Performance do grupo G4 utilizando o estimador PF.



Fonte: Elaborada pelo autor (2020).

Figura 78 – Performance do grupo G4 utilizando o estimador ASH.

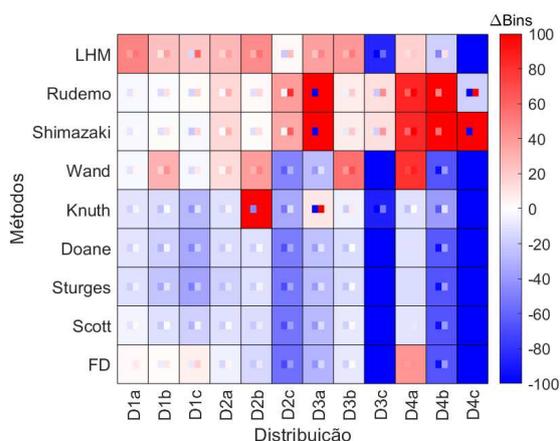


Fonte: Elaborada pelo autor (2020).

Aa Figuras 79 e 80 sumarizam todas as distribuições por todos os seletores para estimações realizadas pelo Histograma com 1000 amostras. Dentro de cada quadrante, que representa a mediana (2º quartil) existem sub-quadrantes que representam o 1º quartil e o 3º quartil, possibilitando avaliar a dispersão dos valores. A Figura 79 mostra a variação do número de *bins* em relação ao *bin* ótimo, valores positivos representam que o seletor superestimou os *bins*, levando a estimação a ter um menor viés e maior variância; valores negativos representam que o seletor subestimou os *bins*, tornando a estimação mais suave e com grande viés. Valores foram saturados em $|100|$ para melhorar a resolução do resultado.

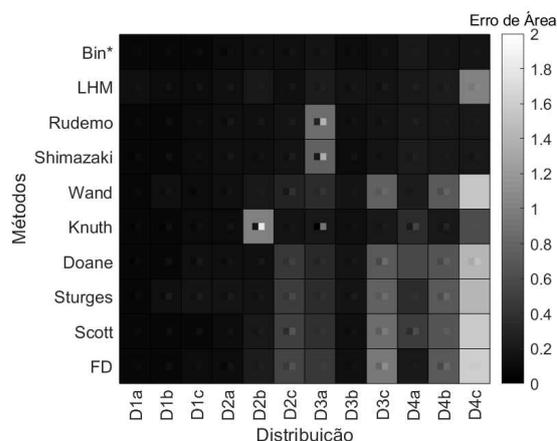
A Figura 80 apresenta a área do erro das estimações, sendo possível notar que os seletores possuem maiores dificuldades em distribuições esparsas e/ou com descontinuidade, como as densidades D2b, D2c, D3a, D3c, D4a, D4b, D4c; para densidades unimodais, pouco esparsas e com derivadas suaves os seletores obtiveram boas performances. As demais figuras, referentes aos seletores PF e ASH com diferentes conjuntos de amostras se encontram no Apêndice C.1.3.

Figura 79 – Variação dos bins para o Histograma com 1000 amostras.



Fonte: Elaborada pelo autor (2020).

Figura 80 – Área do erro para o Histograma com 1000 amostras.



Fonte: Elaborada pelo autor (2020).

De forma geral percebemos que os seletores iterativos de Rudemo e Shimazaki parecem ser os mais resilientes, e devem ser utilizados quando não se sabe nada a priori sobre a distribuição. Ademais, o seletor de Rudemo se saiu ligeiramente melhor, principalmente com poucas amostras. A única distribuição em que os dois seletores encontraram maior dificuldade foi D3a, onde havia a soma das características de uma distribuição esparsa, multimodal e assimétrica. O seletor de Knuth se comportou bem em distribuições com grau de dificuldade elevado, para o número de amostras acima de 500, deixando a desejar em distribuições relativamente simples, abaixo de 500 amostras. Wand parece ter dificuldade com distribuições esparsas e os seletores baseados em variáveis de escala encontram muita dificuldade em distribuições com alta rugosidade e multimodais. Os seletores Doane e Sturges tendem a subestimar o número de bins, levando a estimações mais suaves, que os torna relativamente resilientes em distribuições pouco esparsas e com poucas amostras.

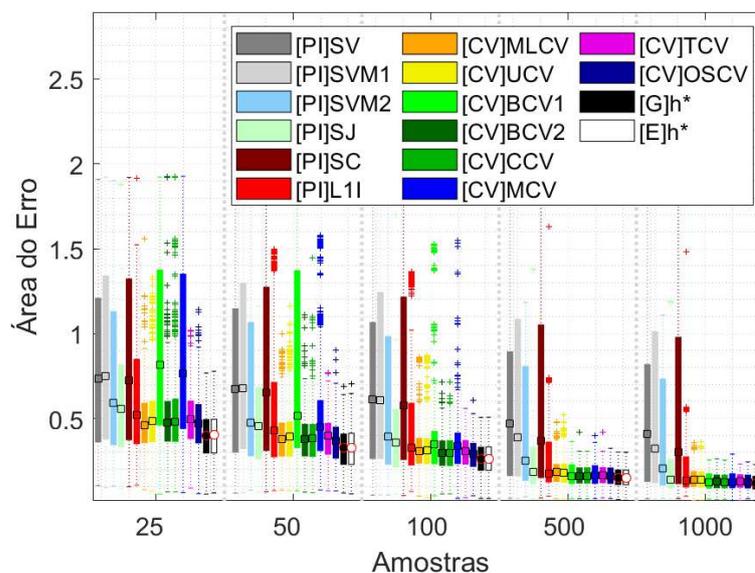
4.1.1.3 KDE

O estimador não-paramétrico KDE de banda fixa será avaliado na mesma perspectiva vista anteriormente. Todas as figuras relativas às distribuições individuais podem ser encontrados no Apêndice C.1.2. O método KDE utilizará o *kernel* Gaussiano para todos os seletores, com exceção para o seletor L1I que utilizará o *kernel* Epanechnikov. Ademais, encontraremos nas figuras seguintes uma largura de banda ótima para o KDE associada ao *kernel* Gaussiano $[G]h_*$ e ao *kernel* Epanechnikov $[E]h_*$, que foram encontradas com o auxílio da função geradora de cada distribuição estudada.

A Figura 81 representa o grupo G0, com todas as distribuições, e evidencia que os seletores baseados em validação-cruzada tendem a ser mais resilientes do que os seletores de PI com o aumento do número de amostras. Este fato ratifica a afirmação de Luc Devroyne (DEVROYE *et al.*, 1997) de que os seletores de PI tendem a superestimar mais as larguras de banda dos que os seletores de CV. Ou seja, os seletores de PI geralmente levam a estimações mais suaves e com maior viés do que os seletores de CV, principalmente em densidades com rugosidade maior do que a distribuição Normal e com variáveis de escala mais esparsas em relação a Normal. Em contrapartida, os seletores de CV realizam estimações com maior variância e baixo viés, se adaptando melhor a densidades mais complexas, como visto em (BORRAJO; GONZÁLEZ-MANTEIGA; MARTÍNEZ-MIRANDA, 2017). Além disso, é possível perceber que SJ e L1I são mais resilientes do que os demais seletores de PI e que BCV1 encontra dificuldades de convergência abaixo de 100 amostras. Os seletores, MLCV (embora com maior dispersão e *outliers*), UCV, BCV2, CCV, MCV e TCV apresentaram resultados bem próximos. Já o seletor OSCV, foi ligeiramente melhor, coincidindo com as afirmações feitas por Heidenreich (HEIDENREICH; SCHINDLER; SPERLICH, 2013), sobre a preferência na escolha de OSCV quando não se souber nada a priori sobre a densidade.

O grupo G1, composto por densidades unimodais, é apresentado na Figura 82 e mostra todos os seletores com performances relativamente próximas com o aumento das amostras, exceto os seletores SV, L1I e MLCV. Além disso, é possível perceber que os seletores de CV e L1I encontraram mais dificuldade em convergir para melhores performances com poucas amostras. Na distribuição D1a os seletores de PI se saíram melhor, afinal a maioria deles foi baseada na presunção de Normalidade e L1I teve performance intermediária; os seletores de CV tiveram mais dificuldade do que os demais em convergir até 100 amostras, embora OSCV tenha figurado dentre os melhores; com o aumento das amostras os seletores de PI, exceto L1I, continuaram melhores e com menor variância,

Figura 81 – [G0]Performance do grupo contendo todas as Distribuições para o KDE.

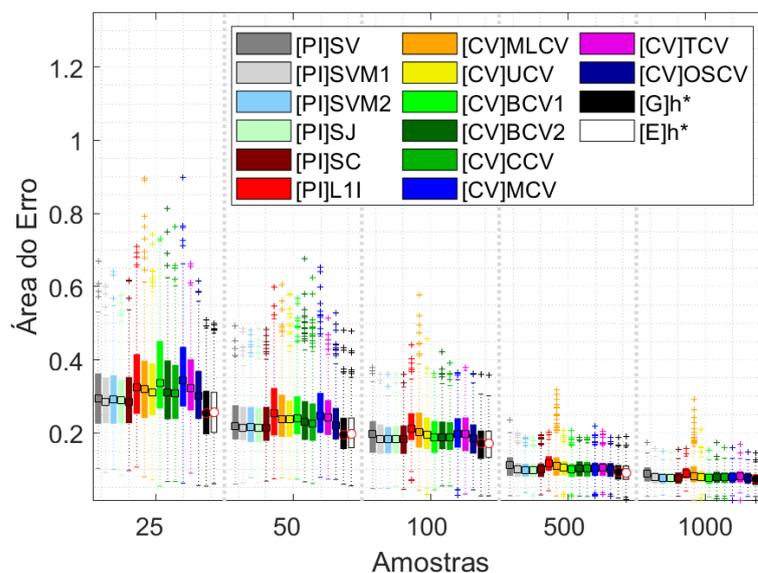


Fonte: Elaborada pelo autor (2020).

tendo SV (baseado em desvio padrão) com as melhores performances. Em D1b, distribuição assimétrica, os seletores baseados em interquartil, SC, SVM1 e SVM2 obtiveram boa performance, SV perdeu eficiência devido a presunção de Normalidade, superestimando o valor de h e novamente L1I teve dificuldades de convergência, junto com MLCV. Na densidade D1c, com rápida transição de pico, MLCV aumentou sua variância, os seletores baseados em interquartil se saíram bem, SV continuou superestimando h . Vale notar que, na densidade com menor rugosidade o seletor SV teve os melhores resultados, bem próximos aos valores ótimos de h^* ; com o aumento da rugosidade os seletores baseados em interquartil se saíram bem, juntamente com os seletores de CV, com exceção de MLCV e BCV1 com poucas amostras.

A Figura 83 apresenta o grupo G2, com distribuições multimodais relativamente pouco esparsas, sendo possível notar que até 100 amostras os seletores de CV encontram dificuldade em convergir, exceto BCV2, CCV e OSCV. Os seletores de PI encontram dificuldade em interpretar essas distribuições, salvo SJ e L1I, que apresentam ótimos resultados até mesmo com poucas amostras. Na distribuição D2a, BCV1 e MCV encontram dificuldade em convergir com poucas amostras, além disso, seletores baseados em escala SV, SVM1, SVM2 tem convergência lenta com o aumento da amostras e o seletor UCV apresenta grande variância; L1I e SJ continuam sendo destaques dos PIs; os demais seletores de CV estabilizam em 500 amostras. Em D2b, com a distribuição um pouco mais esparsa e com derivadas rápidas e lentas, os seletores baseados em interquartil e desvio padrão sofrem ainda mais, obtendo performances muito abaixo dos demais seletores

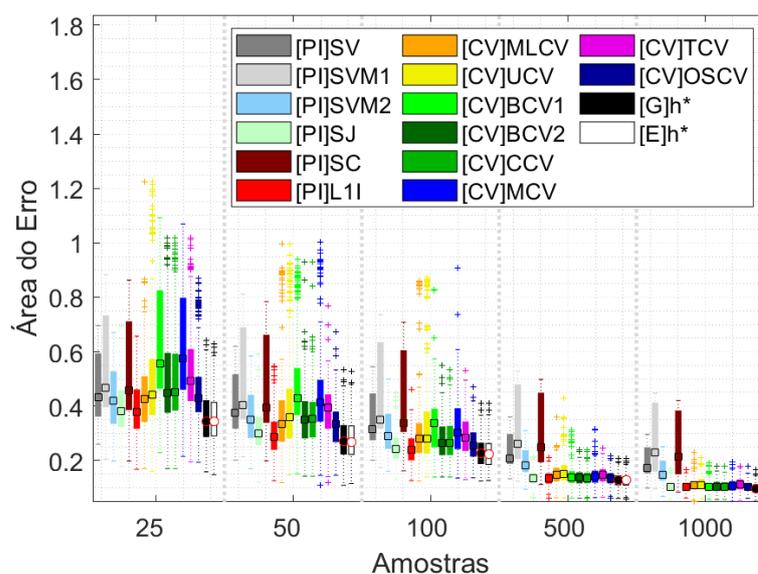
Figura 82 – [G1]Performance do grupo contendo as Distribuições unimodais para o KDE.



Fonte: Elaborada pelo autor (2020).

e aumentando drasticamente sua variância no grupo G2. Já em D3b, MCV piorou sua performance relativa, convergindo mais lentamente junto ao BCV1, mostrando que esses seletores possuem dificuldade em lidar com derivadas rápidas e lentas em realidades com poucas amostras, mesmo em densidades pouco esparsas. De forma geral esse teste mostra a falta de robustez de seletores baseados em escala, e que seletores de PI como L1I e SJ conseguem bons resultados em densidades pouco esparsas.

Figura 83 – [G2]Performance do grupo contendo as Distribuições bimodais com derivada suave e trimodal pouco esparsa para o KDE.

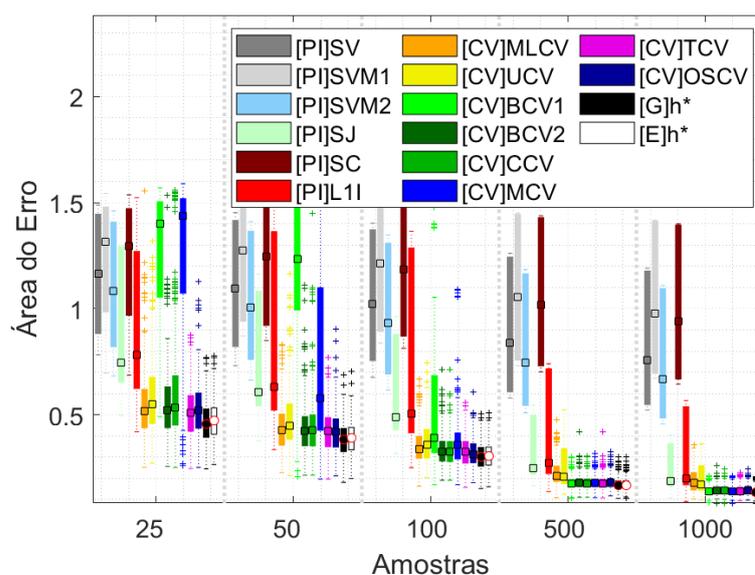


Fonte: Elaborada pelo autor (2020).

A Figura 84 mostra o grupo G3, que contempla densidades relativamente esparsas.

Nessa figura é possível perceber que os seletores de PI baseados em escala possuem grande dispersão e baixa performance; L1I e SJ convergem com aproximadamente 500 amostras e com grande variância; os seletores de CV parecem resilientes a essa realidade, embora BCV1 e MCV tenham dificuldade em convergir com poucas amostras. Analisando a densidade D2c, os seletores SV, SVM1, SVM2 e SC encontram uma dificuldade sabida, devido a distorção das variáveis de escala nessa realidade; L1I e seu método adaptativo e SJ, com a rugosidade calculada de forma alternativa, possuem desempenho intermediário; os seletores de CV convergem bem com poucas amostras, exceto BCV1 e MCV; o seletor UCV apresenta a maior dispersão dentre os seletores de CV. Em D3a L1I tem performance próxima aos melhores seletores de CV; os seletores baseados em variáveis de escala continuam com performances ruins; MLCV encontra dificuldade com o aumento das amostras e UCV, construída através do “termo problemático” (teorizado por Rudemo) encontra dificuldade em convergir, assim como o seletor de Rudemo (visto anteriormente para Histograma). Resumidamente, os seletores baseados em escala perdem muita performance, os seletores SJ e L1I tiveram performances intermediárias, entretanto, quanto mais esparsa a densidade pior será seu desempenho; os seletores de CV foram resilientes com o aumento das amostras nessa realidade, embora o seletor UCV tenha dificuldade em lidar com derivadas rápidas e lentas juntamente com a esparsidade.

Figura 84 – [G3]Performance do grupo contendo as Distribuições bastante esparsas para o KDE.

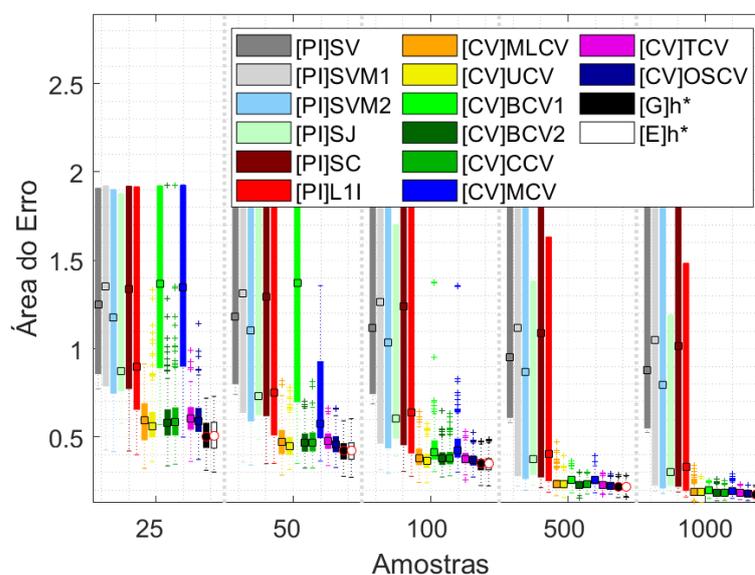


Fonte: Elaborada pelo autor (2020).

O grupo G4 é formado pelas distribuições com descontinuidades (ou derivadas infinitas). A Figura 85 mostra todos os seletores de PI com grande dispersão, os seletores

BCV1 e MCV convergindo apenas com 500 amostras e os seletores de CV resilientes, mesmo com poucas amostras. De forma geral, os seletores L1I e SJ parecem ter perdido relativamente mais performance do que no grupo anterior G3, principalmente devido a sua performance na densidade D4c. Avaliando D4a, percebemos que os seletores tiveram performances bem próximas, os seletores SVM1, SVM2 e SC (baseados em interquartil) tiveram boas performances. Entretanto, o seletor baseado em desvio padrão foi o que teve pior performance, devido a distorção no desvio padrão gerar um h muito maior do que o necessário, levando a estimação a ter um alto viés; novamente BCV1 convergiu apenas com 500 amostras. Em D4b, distribuição esparsa com derivadas finitas e infinitas, os seletores baseados em escala tiveram as piores performances, o seletor SJ se saiu melhor do que L1I evidenciando novamente que o SJ parece ser mais imune a esparsidade do que L1I; Na última distribuição D4c, bastante esparsa e somente com derivadas infinitas, os seletores de escala tiveram suas piores performances, próximas ao erro máximo ≈ 2 em algumas iterações. Ou seja, a esparsidade faz com que esses seletores estimem variáveis de escala bem maiores do que o necessário, superestimando h e levando a estimatórias muito suaves, porém existem derivadas infinitas que precisariam de h (tendendo a zero) para serem minimamente descritas; L1I e SJ encontraram dificuldade nessa realidade; Os seletores de CV se mantiveram resilientes, principalmente com o aumento das amostras; BCV1 e MCV tiveram dificuldade de convergência, estabilizando em uma performance inferior aos demais seletores de CV.

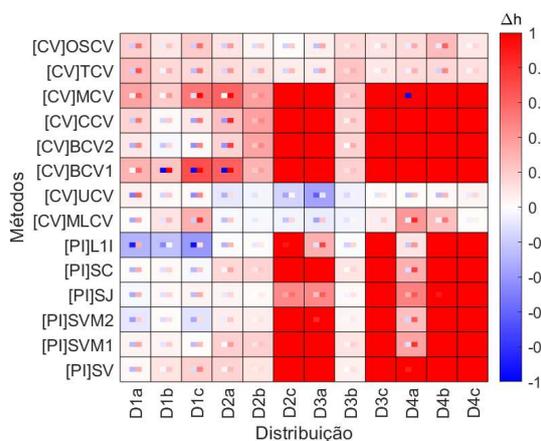
Figura 85 – [G4]Performance do grupo contendo as Distribuições com descontinuidade para o KDE.



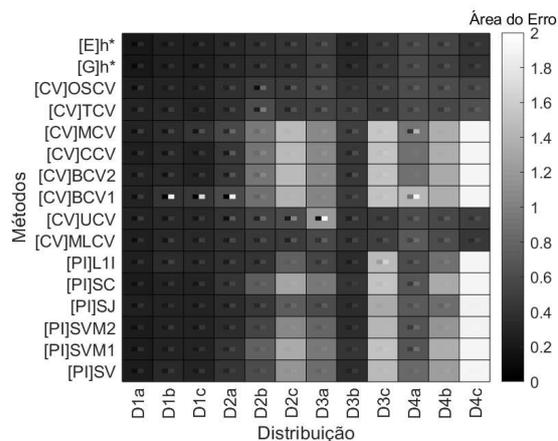
Fonte: Elaborada pelo autor (2020).

A Figura 87 sintetiza o desempenho de todos os seletores em todas as distribuições, em uma realidade com 25 amostras de treinamento. No centro de cada quadrante é possível observar um subquadrante com os valores máximos/mínimos que o mesmo pode assumir de acordo com o interquartil. A Figura 86 apresenta a variação da largura de banda h em relação ao h^* ótimo, saturada em $|1|$ para melhorar a resolução. Se Δh for positivo o seletor superestimou h levando a estimações mais suaves com maior viés, caso Δh seja negativo o seletor subestimou a largura de banda levando a uma estimacão com maior variância e menor viés. É importante notar que a maioria dos seletores tende a superestimar h , e quanto maior a rugosidade e esparsidade mais os seletores tendem a superestimar h com poucas amostras. Ao relacionarmos Δh com a Figura 87 percebemos que superestimar a largura de banda h degrada mais a performance dos seletores do que quando a largura de banda é subestimada. Além disso, percebemos os seletores TCV, OSCV e MLCV bastante resilientes em todas as distribuições, mesmo com poucas amostras.

Figura 86 – Variação de h para o KDE com 25 amostras. Figura 87 – Área do erro para o KDE com 25 amostras.



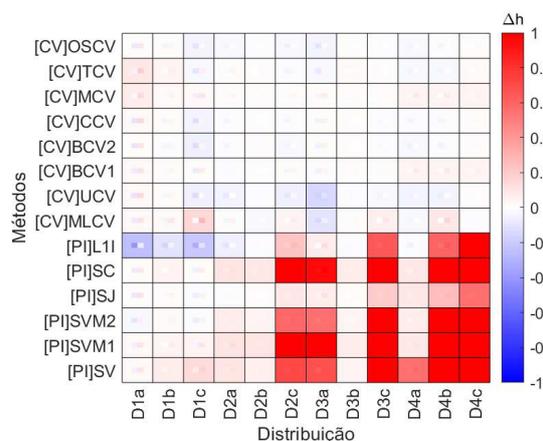
Fonte: Elaborada pelo autor (2020).



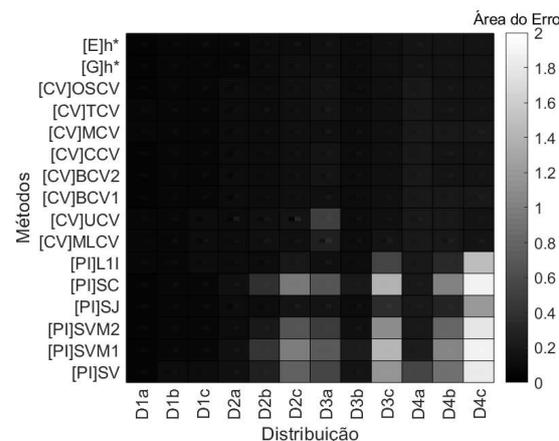
Fonte: Elaborada pelo autor (2020).

As Figuras 88 e 89 mostram a mesma realidade anterior para estimacões feitas com 1000 amostras. Logo na Figura 88 é possível perceber que os seletores de PI realmente superestimam com maior intensidade h em densidades com rugosidade maior do que a Normal; os seletores CV convergem bem em todas as densidades, exceto os seletores UCV e MLCV que encontram dificuldade em estimar D3a. Na Figura 89, percebemos novamente que o viés causado por uma estimacão com h maior do que o necessário degrada a performance com maior intensidade; Ademais, percebemos que a maioria dos seletores de CV conseguem resultados próximos a h^* em densidades complexas; L11 e SJ tem performances intermediárias, porém perdem performance em densidades esparsas. As demais figuras referentes a essa análise se encontram no Apêndice C.1.3.

Figura 88 – Variação de h para o KDE com 1000 amostras. Figura 89 – Área do erro para o KDE com 1000 amostras.



Fonte: Elaborada pelo autor (2020).



Fonte: Elaborada pelo autor (2020).

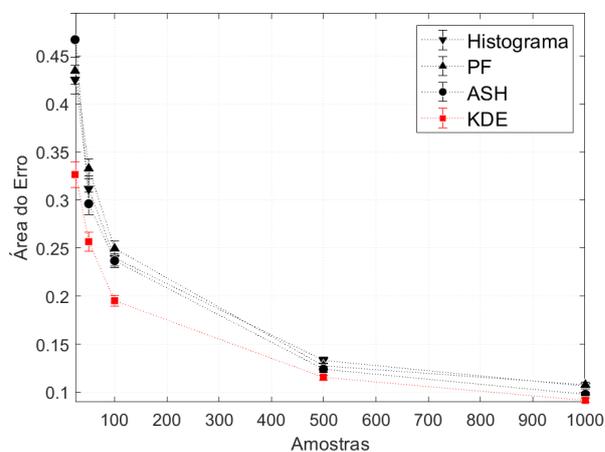
4.1.1.4 Comparação entre os estimadores de banda fixa

Após a avaliação das características dos seletores do Histograma, PF, ASH e KDE, será realizada uma comparação de desempenho dos seletores considerados mais resilientes nas análises anteriores. Essa comparação é importante devido à necessidade de encontrar uma função adaptativa de suavização f_{pi} robusta, que será utilizada nos métodos de banda variável VKDE e ROIKDE. Para os estimadores Histograma, PF, ASH utilizaremos o seletor de Rudemo e para o KDE o seletor OSCV.

A seguir veremos a comparação através de quatro casos representativos de densidades D1c (Figura 90), D2c (Figura 91), D3c (Figura 92) e D4c (Figura 93), apontando o KDE com as menores áreas de erro entre estimação e modelo. As figuras complementares podem ser encontradas no Apêndice C.1.4 e apontam na mesma direção, em todas as distribuições o método KDE apresenta as melhores performances. Ademais, é percebido que os outros métodos (Histograma, PF, ASH) intercalam suas performances, devido à existência de diversos componentes conflitantes que alteram a performance desses métodos; ou seja o Histograma tem dificuldade em descrever derivadas rápidas, entretanto todos os respectivos seletores foram desenvolvidos para essa realidade, salvo algumas exceções; o PF descreve melhor as derivadas do que Histograma, entretanto, tem dificuldade em representar picos e os vales entre distribuições multimodais com poucas amostras e por necessitar de menos *bins* do que o Histograma (para chegar a uma melhor estimação) pode se beneficiar fortuitamente da tendência dos seletores em subestimar o número de *bins* em densidades com alta rugosidade. O ASH necessita de menos *bins* do que o PF, entretanto a média feita com m valores buscando minimizar o impacto da primeira amostra

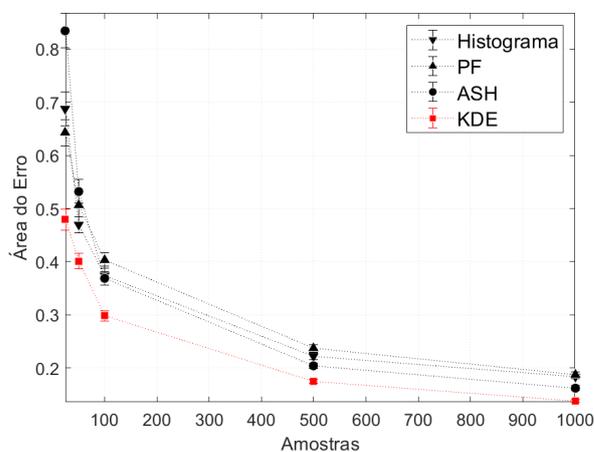
na estimação pode atenuar derivadas rápidas nas bordas das densidades.

Figura 90 – Comparação entre os estimadores de banda fixa para Distribuição D1c.



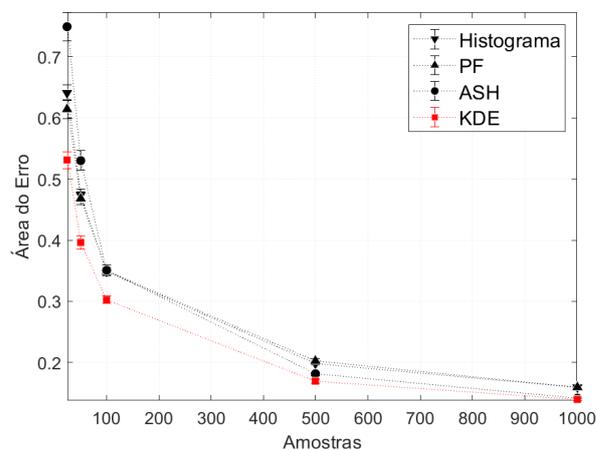
Fonte: Elaborada pelo autor (2020).

Figura 91 – Comparação entre os estimadores de banda fixa para Distribuição D2c.



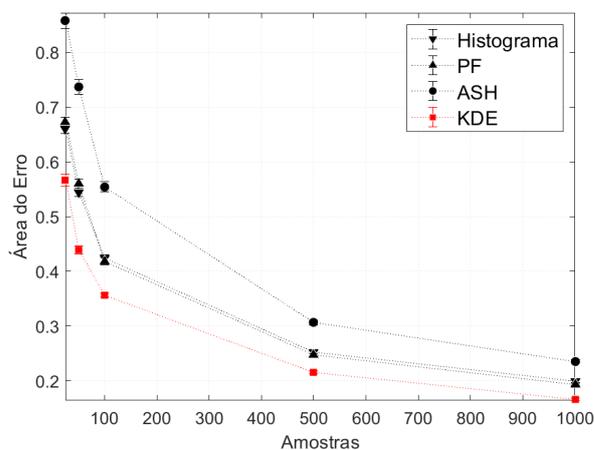
Fonte: Elaborada pelo autor (2020).

Figura 92 – Comparação entre os estimadores de banda fixa para Distribuição D3c.



Fonte: Elaborada pelo autor (2020).

Figura 93 – Comparação entre os estimadores de banda fixa para Distribuição D4c.



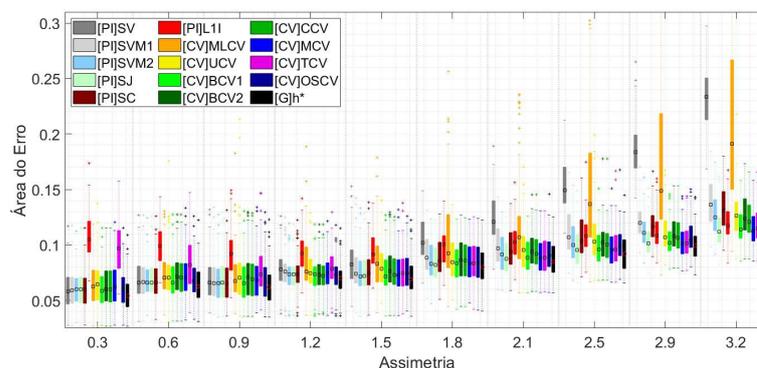
Fonte: Elaborada pelo autor (2020).

4.1.2 Teste de 3º e 4º momento central

O teste de 3º momento, ou assimetria, foi descrito na Seção 3.3.3. Basicamente, o valor do parâmetro σ da distribuição Log-Normal será alterado, com isso a distribuição irá alterar sua simetria. A Figura 94 mostra o erro de área de cada seletor para o KDE de banda fixa, de acordo com a assimetria do conjunto de amostras. É importante salientar que, para manter a área total da distribuição em 0.9999 o intervalo da distribuição precisa ser alterado, adicionando ao teste um erro de interpolação, que foi quantificado como desprezível nessa análise. O valor de assimetria no gráfico representa a mediana do cálculo de assimetria

feito em cada conjunto de 1000 amostras, escolhido por ser um quantidade de amostras em que a característica dos seletores é mais clara em relação ao erro de estimação com 25 iterações. Portanto, percebemos que, para valores de assimetria próximos de zero, ou seja, para distribuição tendendo a Gaussiana, com pouca assimetria, todos os seletores obtiveram os melhores resultados, com exceção L1I e TCV, que sofrem com a redução das amostras ocasionada por cortes internos dos algoritmos, L1I sofre cortes pela função *Kernel L* e de Epanechnikov, e TCV corta amostras consideradas proporcionalmente distantes do intervalo de h selecionado. Com o aumento da assimetria os seletores SV e MLCV são os que mais perdem performance. Esse fato é ocasionado pela tendência desses seletores de encontrar valores maiores para h nessa realidade, dificultando a descrição de derivadas mais altas, que tendem a crescer com o aumento de assimetria. Outra observação importante é o aumento de resiliência das modificação em SVM1, SVM2 e SC que utilizam o interquartil. Os outros seletores de CV mantêm a performance próxima ao valor de erro encontrado pelo h^* .

Figura 94 – Assimetria para o KDE de banda fixa.

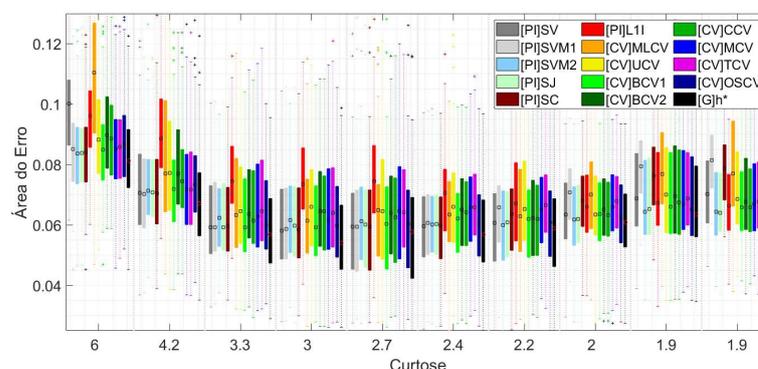


Fonte: Elaborada pelo autor (2020).

Para o teste de 4º momento foi utilizada a distribuição GGD, como visto anteriormente. De modo análogo ao teste anterior o parâmetro de forma ρ foi alterado com o intuito de variar a curtose do conjunto de amostras. Como referência temos o valor de *curtose* = 3, que representa a Gaussiana. Valores baixos de curtose significam que a distribuição tende a ter formato mais parecido a uma distribuição Uniforme, e para altos valores de curtose a distribuição tende a distribuição Laplace. A Figura 95 mostra que para valores próximos a 3 de curtose os seletores tendem a apresentar menor área de erro. Quando a distribuição tende a Laplace os seletores perdem mais eficiência do que quando a distribuição tende ao formato Uniforme. No caso de alta curtose os seletores SV, L1I e MLCV foram os que apresentaram piores resultados. Em baixos valores de curtose o seletor SVM1, alternando com o seletor SVM1, baseado somente no interquartil, e MLCV foram os mais degradados.

O mesmo efeito não é percebido em SVM2 devido ao parâmetro de escala ser adaptativo. Os outros seletores de CV mantêm a performance próxima ao valor de erro encontrado pelo h^* , exceto MLCV.

Figura 95 – Curtose para o KDE de banda fixa.



Fonte: Elaborada pelo autor (2020).

4.1.3 Largura de Banda Variável

Nesta etapa apenas o estimador KDE será avaliado, os demais métodos necessitam alterar o espaçamento entre os pontos de *grid* para emular um seletor de banda variável, assunto que não será contemplado neste trabalho. Portanto, avaliaremos as estimações dos seletores apresentados na Seção 3.2.4.2, que são os métodos AKDE, BKDE, VKDE e ROIKDE.

A largura de banda variável possibilita ao KDE uma estimaco mais generalista e adiciona um grau de liberdade em um contexto onde as estimaces diferiam apenas na troca entre viés e variância. Foi mostrado anteriormente que do ponto de vista de resiliência da estimaco é preferível um seletor que subestime a largura de banda h fixa em detrimento de um que superestime.

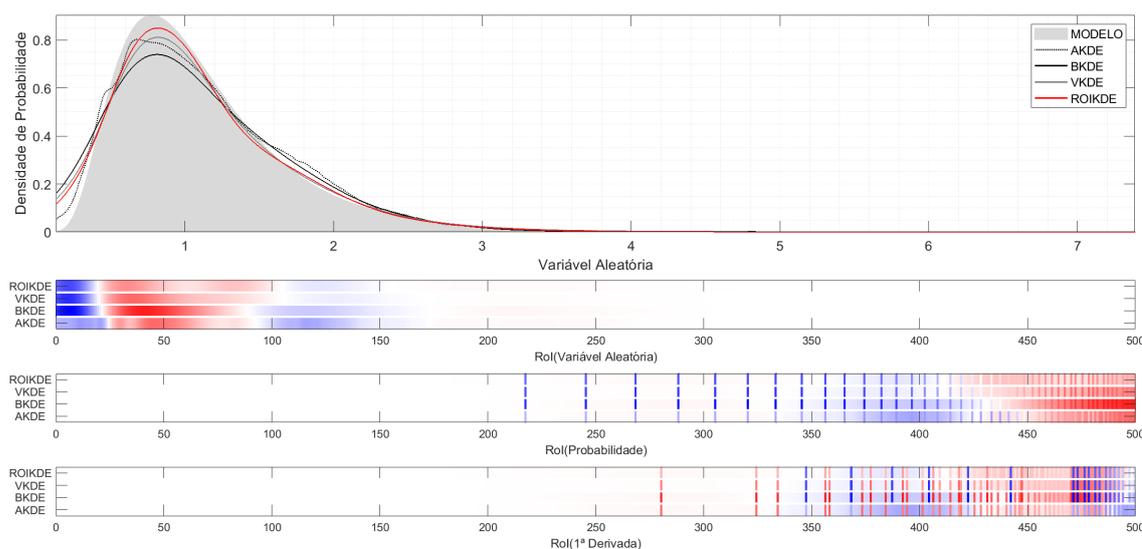
No decorrer dessa seo serão apresentadas estimaces construdas com 25 amostras em um caso representativo de cada grupo. As demais densidades avaliadas através do RoIMap, construdas com estimaces de 25 e 1000 amostras, encontram-se no Apêndice C.2.

4.1.4 Avaliaco dos Estimadores

Para avaliar as características das estimaces ao longo da variável aleatória utilizaremos o método RoIMap para discernir as diferenças/similaridades entre os seletores. Na Figura 96 temos a distribuo D1b em uma realidade com 25 amostras, sendo possível

observar o método AKDE com uma estimação pouco suave, BKDE sendo a mais suave de todas e as estimações VKDE e ROIKDE com características parecidas. Ademais, é possível perceber que a estimação de ROIKDE na inclinação da esquerda consegue ser mais rápida do que o VKDE, chegando ao pico próximo a RoI 50 (da variável aleatória) com o menor erro dentre as demais estimações. Já na descida, à direita do pico, ROIKDE consegue ajustar sua derivada para um erro próximo à melhor estimação nessa faixa, próximo a RoI 100 (da variável aleatória), e segue estimando a cauda bem próximo a VKDE. Também é possível perceber no RoIMap a dependência da área do erro da estimação em relação a probabilidade e a primeira derivada. Devido a este fator o método ROIKDE utiliza a informação a priori de probabilidade e derivada, construída através de uma estimação de KDE fixo, para corrigir a largura de banda variável nas regiões de alta probabilidade e derivada.

Figura 96 – Ferramenta RoIMap utilizada na distribuição D1b.

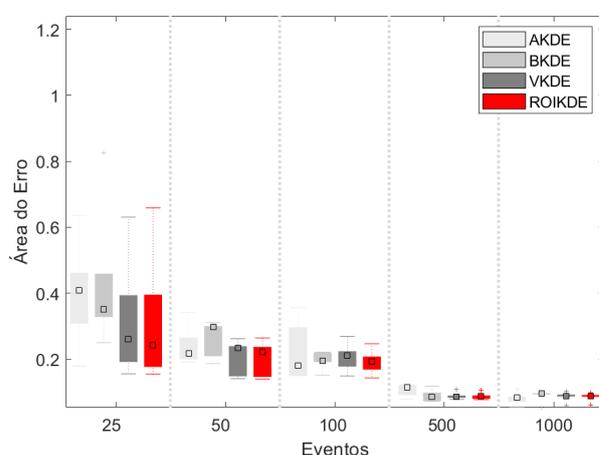


Fonte: Elaborada pelo autor (2020).

Naturalmente, ao avaliarmos o somatório dos erros ao longo da variável aleatória na Figura 97, em 25 amostras, o método ROIKDE obtém a melhor performance média, bem próximo a VKDE. A diferença entre ROIKDE e VKDE tende a diminuir com o aumento das amostras, pois as larguras de banda fixas utilizadas como limitante superior e inferior da largura de banda variável do ROIKDE tendem a se aproximar em distribuições unimodais. Embora seja comum na literatura a representação dos resultados através da mediana e interquartil, sendo possível avaliar qual a medida de posição e dispersão da área do erro das estimações dos métodos, comparar os métodos entre si através de uma representatividade estatística se torna mais complexo. Portanto, a Figura 98 mostra a diferença da área do erro

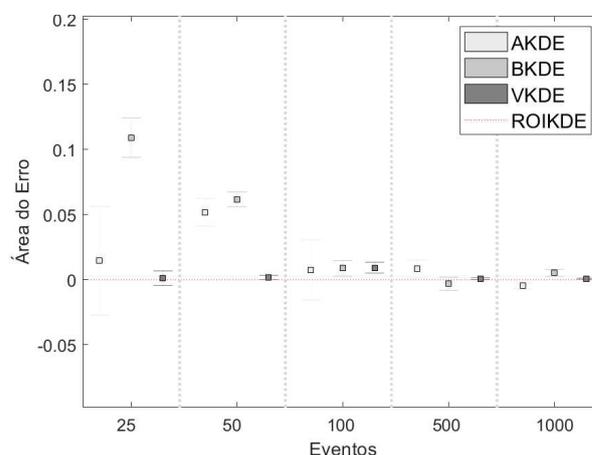
entre os métodos desenvolvidos na literatura (AKDE, BKDE e VKDE) e o método proposto nessa tese (ROIKDE), ou seja, a cada iteração dos algoritmos utilizando o mesmo conjunto de amostras é observada a diferença da área do erro entre todos os métodos em relação ao ROIKDE. Caso a diferença seja positiva, os métodos obtiveram um erro maior do que o ROIKDE, caso a diferença seja negativa, o ROIKDE obteve pior performance do que o respectivo método. Neste caso é possível perceber que o ROIKDE teve performance estatisticamente igual ao VKDE em quase todos os conjuntos de amostras, melhor do que AKDE e BKDE até 100 amostras e pior do que o AKDE com 1000 amostras, embora com performance bem parecida.

Figura 97 – Área do erro dos métodos variáveis para Distribuição D1b.



Fonte: Elaborada pelo autor (2020).

Figura 98 – Área do erro em relação ao ROIKDE para Distribuição D1b.

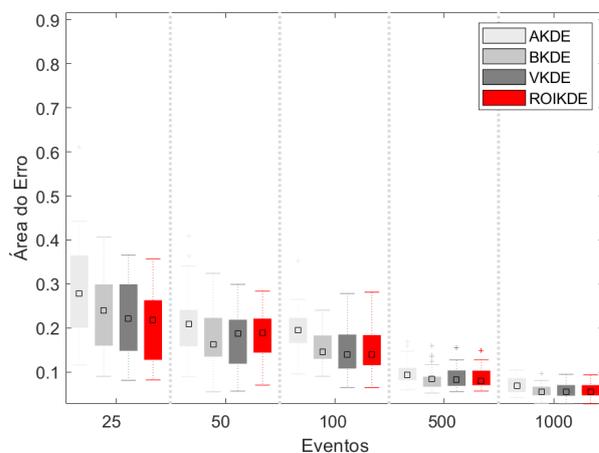


Fonte: Elaborada pelo autor (2020).

Ao avaliarmos as outras distribuições unimodais através da Figura 99 observamos a mesma característica vista em D2b, ROIKDE começa ligeiramente melhor e se aproxima de VKDE com o aumento das amostras e AKDE, com sua estimação de maior variância, tem a pior performance na distribuição de rugosidade mais baixa. Na Figura 100, com rápida transição de pico, temos o AKDE, realizando sua estimação ruidosa, com dificuldade em descrever as derivadas mais suaves de D1c, e BKDE, com uma estimação com alto viés, estimando mal a transição rápida de pico. O método ROIKDE parece ligeiramente melhor do que VKDE, principalmente com poucas amostras. As demais figuras, referentes a todas as análises da representatividade estatística, da comparação dos métodos em relação ao ROIKDE podem ser encontradas em C.2.2.

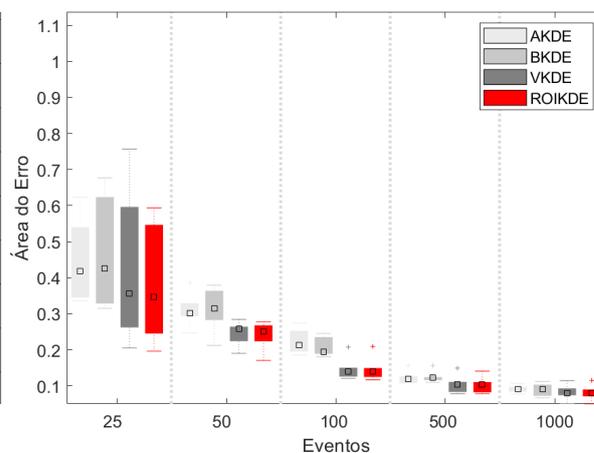
A Figura 101 mostra a distribuição D2c, esparsa e com rápida transição de pico. Avaliando o RoIMap, próximo aos valores -3 e 3 da variável aleatória, percebemos o método AKDE com as melhores estimações, entretanto, a medida que se afasta das

Figura 99 – Área do erro dos métodos variáveis para Distribuição D1a.



Fonte: Elaborada pelo autor (2020).

Figura 100 – Área do erro dos métodos variáveis para Distribuição D1c.



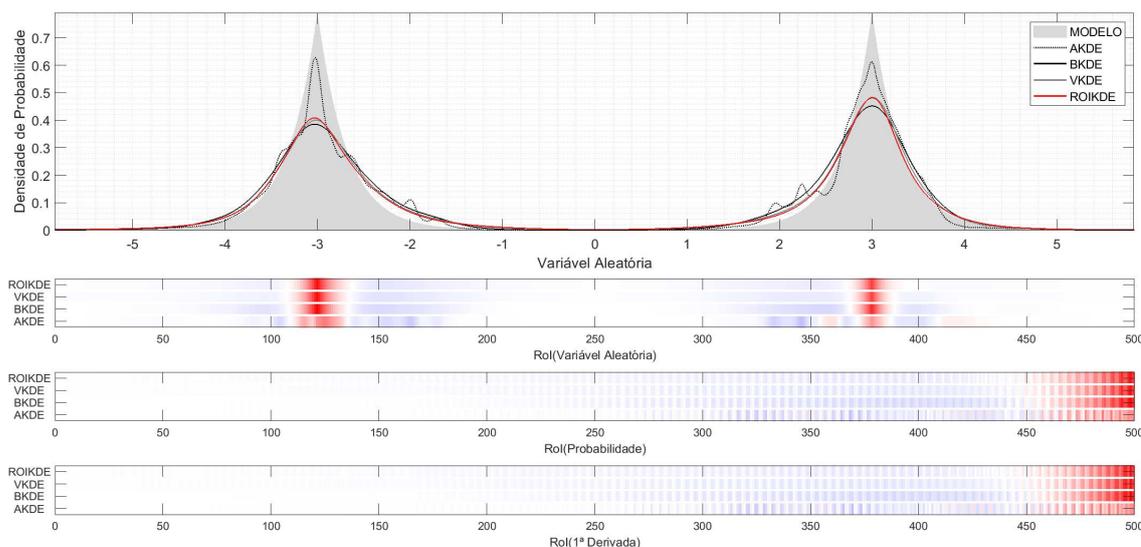
Fonte: Elaborada pelo autor (2020).

regiões de pico a estimação possui grande variância. Em seguida temos o BKDE como o outro extremo, a estimação com maior viés. De fato o BKDE teve uma característica bem próxima ao VKDE e ROIKDE, no pico a estimação foi ligeiramente pior e a medida que a derivada da densidade decresce BKDE superestima a densidade mais do que VKDE e ROIKDE. O método ROIKDE consegue atingir uma derivada maior do que VKDE, chegando a maiores valores de probabilidade próximo ao pico em -3 da variável aleatória. No vale formado entre os dois picos ROIKDE consegue estimar melhor do que AKDE mantendo uma estimação suave, ou seja, após atingir uma inclinação maior na subida ROIKDE conseguiu estimar melhor do que AKDE uma derivada suave. No pico à direita, em 3 na variável aleatória, VKDE e ROIKDE tem valores próximos, porém, ROIKDE consegue estimar melhor a descida à direita deste pico. Novamente existe uma dependência entre a área do erro e probabilidade/derivada.

A Figura 102 mostra o método ROIKDE novamente com uma performance melhor do que os demais métodos e se aproximando de VKDE com o aumento das amostras. AKDE sofre com poucas amostras, levando a uma estimação com alta variância e BKDE tem performance intermediária. Ao avaliarmos a Figura 103 percebemos que o ROIKDE realmente obteve melhores performances do que AKDE e BKDE em todos os conjuntos de amostras, e foi melhor do que VKDE com 50 e 500 amostras.

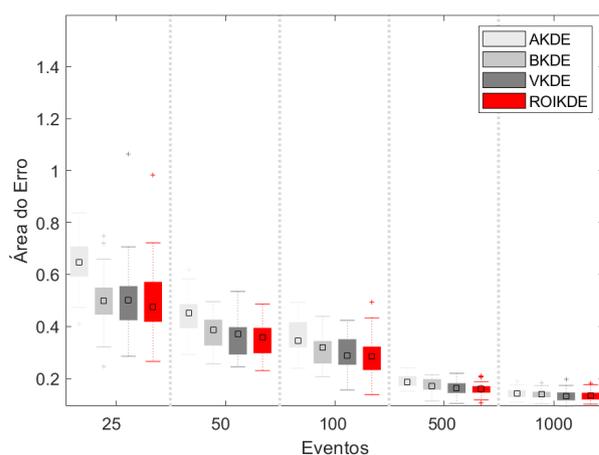
A Figura 104 mostra a estimação em uma densidade Bimodal com derivadas relativamente suaves. Nesse cenário ROIKDE é ligeiramente melhor com poucas amostras, ROIKDE e VKDE possuem as melhores performances. BKDE, com estimções suaves, tem boa performance devido à baixa rugosidade da densidade. Na Figura 105, embora a

Figura 101 – Ferramenta RoIMap utilizada na distribuição D2c.



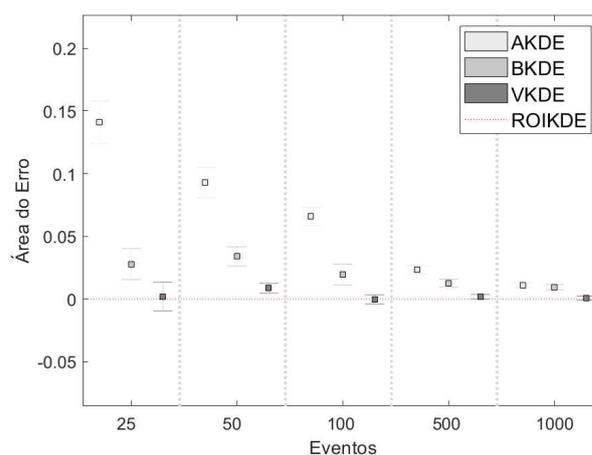
Fonte: Elaborada pelo autor (2020).

Figura 102 – Área do erro dos métodos variáveis para Distribuição D2c.



Fonte: Elaborada pelo autor (2020).

Figura 103 – Área do erro em relação ao ROIKDE para Distribuição D2c.

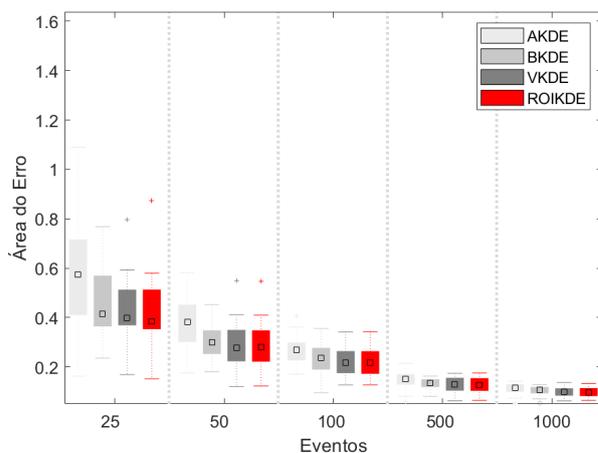


Fonte: Elaborada pelo autor (2020).

densidade possui derivadas relativamente rápidas temos uma transição de pico lenta e um vale com derivadas lentas, beneficiando BKDE que obtém estimações mais suaves, chegando a melhor performance dentre os demais métodos. Os métodos ROIKDE e VKDE tiveram estimações bem próximas a BKDE, sendo que o método AKDE (geralmente com maior variância) tem dificuldade em estimar distribuições com derivadas suaves.

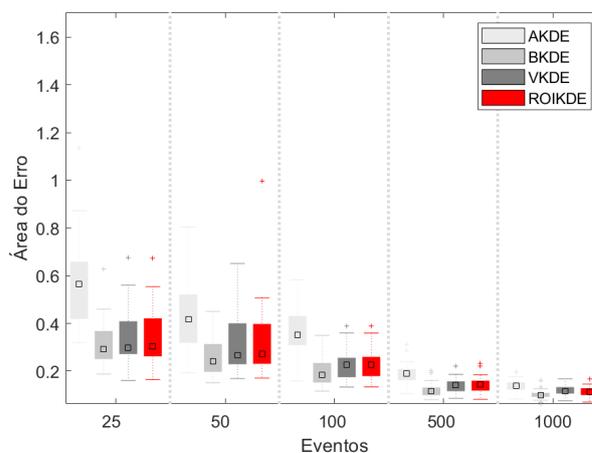
A distribuição D3a, mostrada na Figura 106, tem como características ser esparsa, assimétrica, com rápida transição de pico, derivadas rápidas e lentas. Avaliando o RoIMap percebemos a grande flutuação de AKDE ao longo de toda a variável aleatória. BKDE superestima a largura de banda e tem uma estimação muito suave, VKDE e ROIKDE tem estimações e performances bem próximas. Entretanto, se avaliarmos o pico próximo a 0.7

Figura 104 – Área do erro dos métodos variáveis para Distribuição D2a.



Fonte: Elaborada pelo autor (2020).

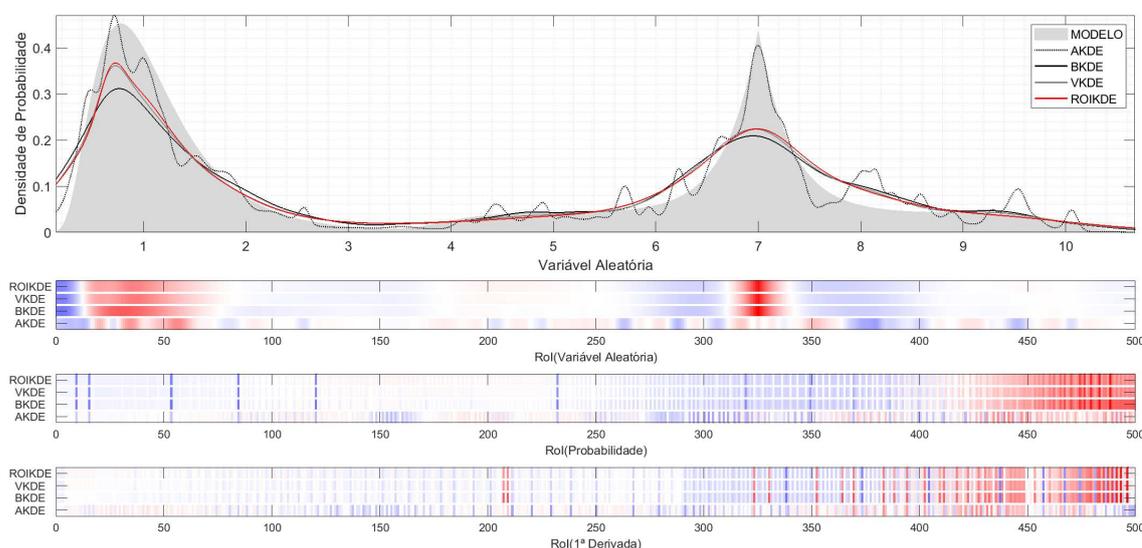
Figura 105 – Área do erro dos métodos variáveis para Distribuição D2b.



Fonte: Elaborada pelo autor (2020).

da variável aleatória percebemos ROIKDE com valores maiores de probabilidade, além disso, na inclinação à esquerda desse pico ROIKDE atingiu derivadas maiores que VKDE. No segundo pico, próximo ao valor 7 da v.a. temos uma defasagem entre as estimações de VKDE e ROIKDE, na inclinação da esquerda ROIKDE estima melhor e na direita VKDE é melhor.

Figura 106 – Ferramenta RoIMap utilizada na distribuição D3a.

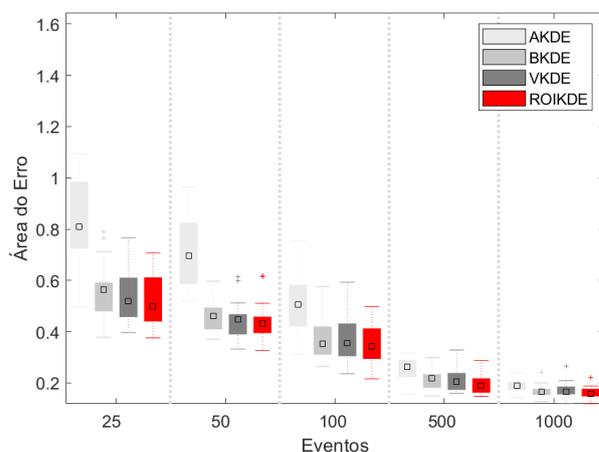


Fonte: Elaborada pelo autor (2020).

Analisando a Figura 107 observamos ROIKDE com a melhor performance e bem próxima a VKDE. O método AKDE tem dificuldade em convergir com poucas amostras, levando a estimações com grande variância, embora consiga convergir com 500 amostras. A Figura 108 mostra que de fato ROIKDE foi melhor do que os demais métodos em todos os conjuntos de amostras, embora o método BKDE com 100 amostras tenha performance

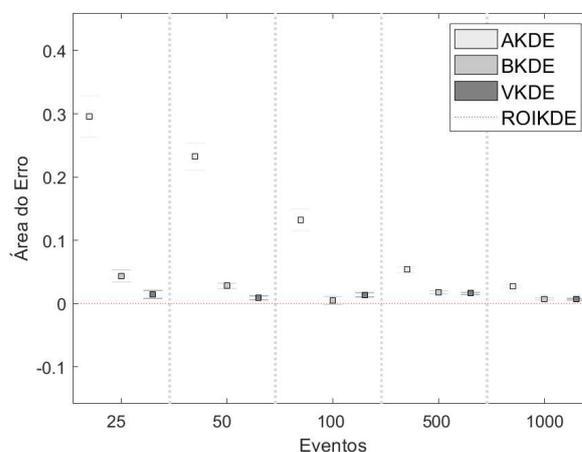
estatisticamente igual ao ROIKDE.

Figura 107 – Área do erro dos métodos variáveis para Distribuição D3a.



Fonte: Elaborada pelo autor (2020).

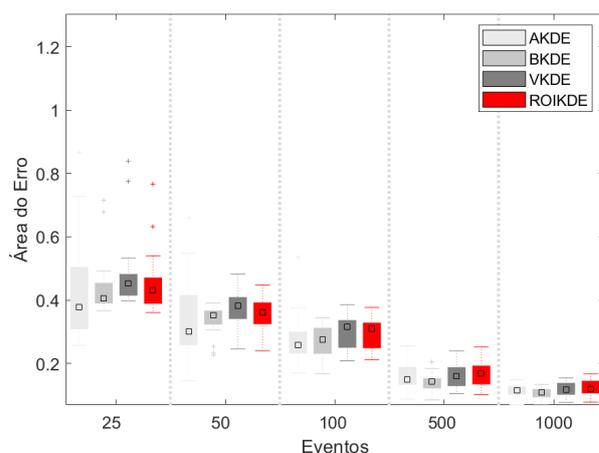
Figura 108 – Área do erro em relação ao ROIKDE para Distribuição D3a.



Fonte: Elaborada pelo autor (2020).

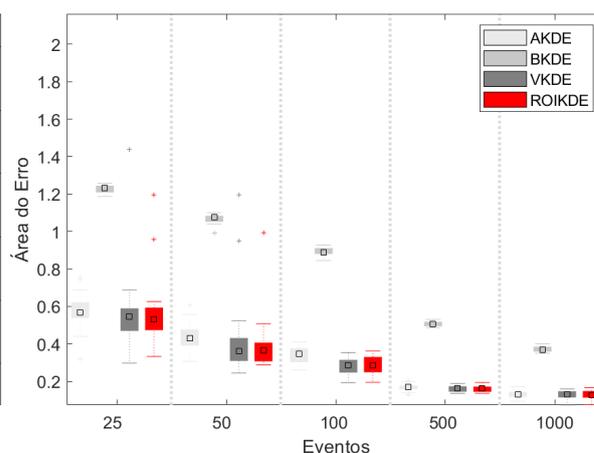
A Figura 109 contempla uma realidade trimodal pouco esparsa. Nessa densidade novamente ROIKDE obteve os melhores resultados, seguido por VKDE. Na Figura 110 os métodos são submetidos a uma distribuição bastante esparsa, onde o método BKDE, que leva a estimações mais suaves, obteve as piores estimações. Em contrapartida, o método AKDE conseguiu se adaptar bem a esse grau de dificuldade, obtendo a melhor performance. Os métodos ROIKDE e VKDE se aproximaram da performance de BKDE à partir de 500 amostras.

Figura 109 – Área do erro dos métodos variáveis para Distribuição D3b.



Fonte: Elaborada pelo autor (2020).

Figura 110 – Área do erro dos métodos variáveis para Distribuição D3c.

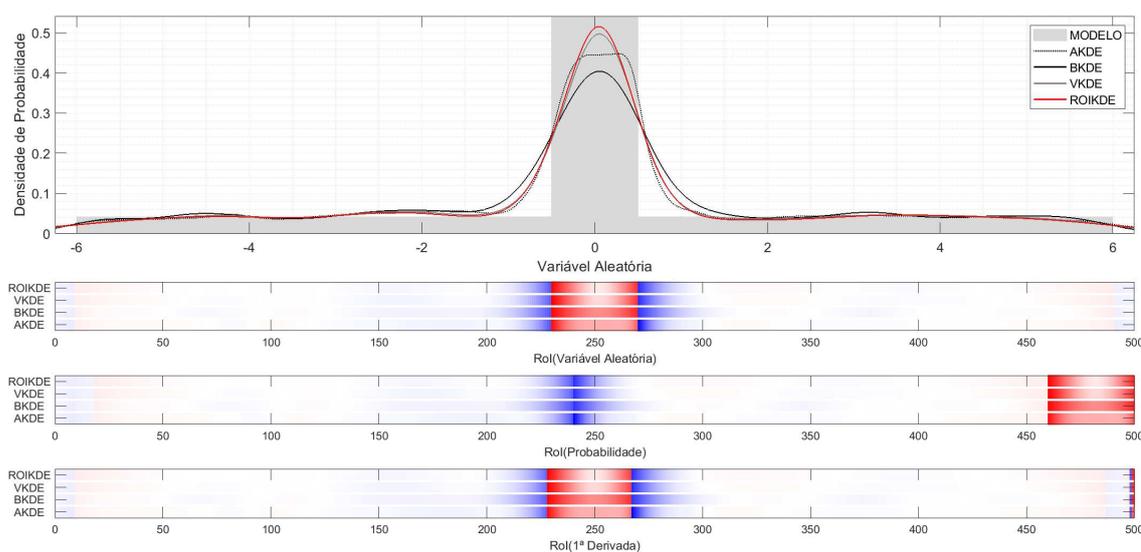


Fonte: Elaborada pelo autor (2020).

A Figura 111 mostra uma realidade em que a função geradora possui derivada infinita, sendo possível observar que o método AKDE consegue atingir as maiores derivadas com 25 amostras de treinamento, além disso, consegue apresentar uma característica

uniforme no pico de sua estimação. Entretanto, VKDE consegue atingir valores de pico mais próximo do modelo, perdendo apenas para ROIKDE. O método BKDE obteve a estimação mais suave dentre as demais. Vale ressaltar que avaliar a dependência em relação a derivada em métodos com derivadas infinitas se torna complexo.

Figura 111 – Ferramenta RoIMap utilizada na distribuição D4a.



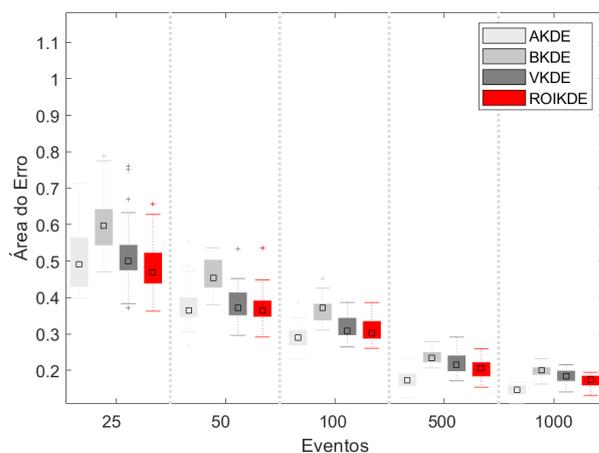
Fonte: Elaborada pelo autor (2020).

Na Figura 112, observamos o ROIKDE com a melhor estimação com 25 amostras, porém o método AKDE tende a alcançar a melhor performance com o aumento das amostras. O método BKDE consolidou a sua dificuldade em estimar derivadas rápidas obtendo a pior performance dentre os métodos. O ROIKDE, apesar de não obter a melhor performance nessa realidade, conseguiu ser melhor do que o VKDE em todas os conjuntos de amostras. A Figura 113 mostra que o método AKDE à partir de 100 amostras obtêm melhores resultados do que o ROIKDE, embora o ROIKDE tenha conseguido se sair melhor do que o VKDE em todos os conjuntos de amostras.

Por fim, a Figura 114 representa D4b, distribuição esparsa que possui derivadas finitas, rápida e lentas, e infinitas. Nessa realidade as estimações obtiveram performances mais próximas. Já na distribuição 115, esparsa e com derivadas infinitas, BKDE realizou estimações extremamente suaves, superestimando a largura de banda variável nas regiões com amostras. Os métodos ROIKDE e VKDE tiveram as melhores performances e bem próximas.

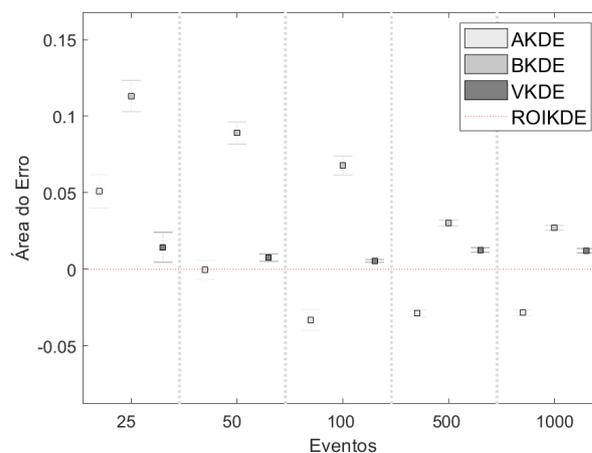
Para sintetizar os resultados anteriores e realizar a comparação entre o ROIKDE e todos os demais métodos, com representatividade estatística, foram reunidas as estimações realizadas pelos métodos em todas as distribuições e comparadas em relação a estimação

Figura 112 – Área do erro dos métodos variáveis para Distribuição D4a.



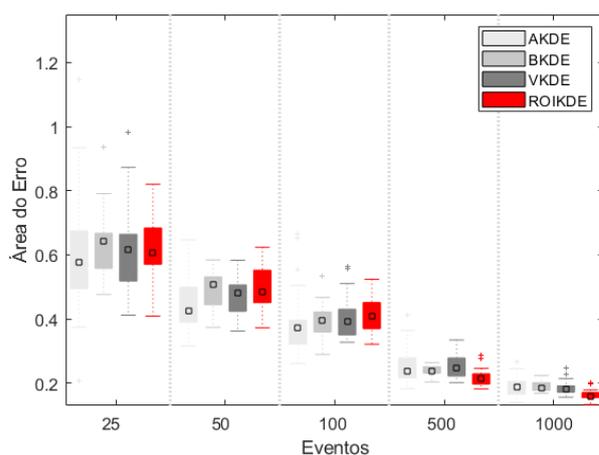
Fonte: Elaborada pelo autor (2020).

Figura 113 – Área do erro em relação ao ROIKDE para Distribuição D4a.



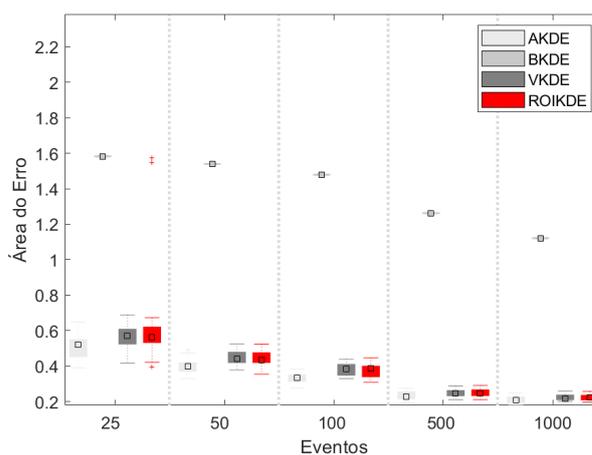
Fonte: Elaborada pelo autor (2020).

Figura 114 – Área do erro dos métodos variáveis para Distribuição D4b.



Fonte: Elaborada pelo autor (2020).

Figura 115 – Área do erro dos métodos variáveis para Distribuição D4c.

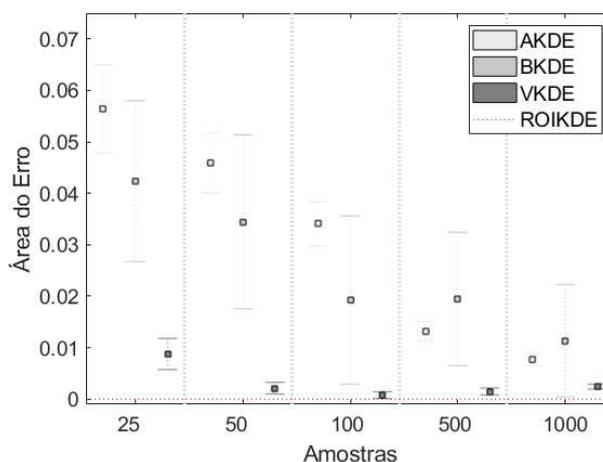


Fonte: Elaborada pelo autor (2020).

do método ROIKDE. A Figura 116 mostra que os estimadores VKDE e ROIKDE apresentam os resultados mais próximos, embora VKDE tenha apresentado maiores erros em todos os conjuntos de amostras. O método BKDE obteve resultados com maior variância em relação ao ROIKDE, embora AKDE tenha obtido a pior performance média. Resumidamente, é possível perceber que o método ROIKDE obteve a melhor performance média dentre todos os métodos e essa melhora foi maior com poucas amostras. Além disso, a partir de 100 amostras o método ROIKDE tende a aumentar a diferença em relação ao VKDE, fato que pode estar relacionado a melhora na estimação da derivada com o aumento do número das amostras.

A Tabela 20 quantifica as afirmações anteriores, mostrando que o ganho médio máximo do ROIKDE ocorre em relação ao AKDE, chegando a $\approx 5\%$. Em relação ao VKDE

Figura 116 – Comparação geral entre os estimadores não-paramétricos de largura de banda variável.



Fonte: Elaborada pelo autor (2020).

o ganho é relativamente pequeno, chegando $\approx 1\%$ com 25 amostras e com 100 amostras o ganho é praticamente desprezível. Portanto, é possível perceber que existe de fato um ganho médio do ROIKDE em relação aos demais métodos.

Tabela 20 – Melhora geral na área do erro do método ROIKDE em relação aos demais métodos.

Amostras	Área do Erro		
	AKDE	BKDE	VKDE
25	0,0564 ± 0,0086	0,0424 ± 0,0156	0,0088 ± 0,0030
50	0,0459 ± 0,0058	0,0345 ± 0,0169	0,0021 ± 0,0011
100	0,0341 ± 0,0043	0,0193 ± 0,0163	0,0008 ± 0,0007
500	0,0132 ± 0,0018	0,0195 ± 0,0130	0,0015 ± 0,0007
1000	0,0077 ± 0,0012	0,0114 ± 0,0109	0,0025 ± 0,0005

Fonte: Elaborada pelo autor (2020).

Ao longo deste trabalho ficou claro que alguns seletores PI tendem a realizar estimações mais suaves do que os seletores CV, para a banda fixa. Por meios distintos, o método baseado em banda variável BKDE tende a ser mais suave do que o AKDE, e os métodos VKDE e ROIKDE parecem se adaptar melhor aos diferentes níveis de dificuldade explorados aqui. Essa tendência natural de alguns estimadores, que privilegia alguns tipos de densidades, dificulta a tarefa de encontrar métodos automáticos que consigam lidar com os dois extremos de viés e variância. Para avaliar a adequação dos métodos em diferentes realidades, uma comparação bem simples pode ser vista na Tabela 21, onde percebemos que o método ROIKDE pode ser melhor (M), igual (I) ou pior (P) estatisticamente do que no máximo 3 métodos (AKDE, BKDE e VKDE) em cada realidade. É possível observar a maior probabilidade do ROIKDE em ser melhor do que pelo menos um dos métodos em quase todas as distribuições analisadas neste trabalho, e ser o pior dos 3 métodos apenas

Tabela 21 – Desempenho do método ROIKDE em relação aos demais métodos.

Distribuição	Amostras														
	25			50			100			500			1000		
	M	I	P	M	I	P	M	I	P	M	I	P	M	I	P
D1a	3	0	0	1	2	0	2	0	1	1	1	1	2	0	1
D1b	1	2	0	3	0	0	2	1	0	2	1	0	2	0	1
D1c	2	1	0	3	0	0	3	0	0	2	1	0	3	0	0
D2a	3	0	0	2	1	0	2	1	0	2	1	0	2	1	0
D2b	1	1	1	1	1	1	1	1	1	1	1	1	2	0	1
D2c	2	1	0	3	0	0	2	1	0	3	0	0	2	1	0
D3a	3	0	0	3	0	0	2	1	0	3	0	0	3	0	0
D3b	1	1	1	1	0	2	1	0	2	0	0	3	0	0	3
D3c	2	1	0	2	1	0	2	0	1	2	1	0	2	1	0
D4a	3	0	0	2	1	0	2	0	1	2	0	1	2	0	1
D4b	0	2	1	1	0	2	0	1	2	3	0	0	3	0	0
D4c	1	1	1	2	0	1	1	1	1	1	0	2	2	0	1

Fonte: Elaborada pelo autor (2020).

na distribuição D3b com 500 e 1000 amostras. O mesmo teste para os demais métodos pode ser encontrado no Apêndice D.2.

A Tabela 22 apresenta o resultado percentual final do teste visto na Tabela 21 para todos os métodos. O valor encontrado na tabela representa a probabilidade do método de referência ser melhor, igual ou pior do que ≥ 1 dos demais métodos. A Tabela 22 mostra o método BKDE como a pior capacidade adaptativa dos demais métodos até 100 amostras, de 500 amostras em diante o método AKDE foi um dos piores. O método VKDE se saiu melhor do que os dois métodos anteriores chegando próximo ao ROIKDE com 100 amostras, embora tenha figurado $\approx 33\%$ das vezes entre os (≥ 1) piores métodos, contra $\approx 25\%$ do ROIKDE. Para os outros conjuntos de amostras (25, 50, 500 e 1000) o ROIKDE obteve no mínimo $\approx 13\%$ a mais do que o VKDE em relação aos resultados de quando o método é melhor (≥ 1) do que os demais.

Tabela 22 – Comparação relativa geral dos métodos de banda variável.

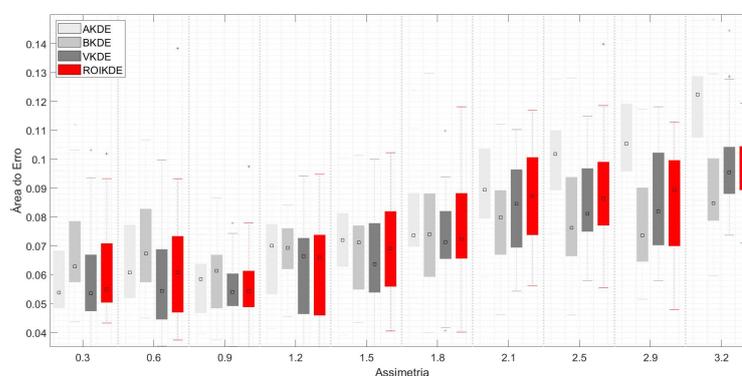
Método de Referência	Comparação	Amostras				
		25	50	100	500	1000
AKDE	Melhor	0,3333	0,3611	0,3611	0,2500	0,3333
	Igual	0,0833	0,0556	0,0833	0,1111	0,1111
	Pior	0,5833	0,5833	0,5556	0,6389	0,5556
BKDE	Melhor	0,2222	0,2500	0,3056	0,3611	0,3611
	Igual	0,1111	0,0556	0,1111	0,2222	0,0833
	Pior	0,6667	0,6944	0,5833	0,4167	0,5556
VKDE	Melhor	0,4722	0,5278	0,5000	0,4167	0,4167
	Igual	0,2500	0,1111	0,1667	0,2222	0,1111
	Pior	0,2778	0,3611	0,3333	0,3611	0,4722
ROIKDE	Melhor	0,6111	0,6667	0,5556	0,6111	0,6944
	Igual	0,2778	0,1667	0,1944	0,1667	0,0833
	Pior	0,1111	0,1667	0,2500	0,2222	0,2222

Fonte: Elaborada pelo autor (2020).

4.1.5 Teste de 3º e 4º Momento Central

Primeiramente será feito o teste de variação da assimetria para avaliar o comportamento dos métodos que utilizam largura de banda variável. A Figura 117 mostra o método AKDE com a pior performance com o aumento da assimetria, esse fator ocorre devido ao método descrever as derivadas lenta da cauda da Log-Normal com grande variância. O método BKDE tende a ser mais resiliente do que os demais métodos com o aumento da assimetria e os métodos VKDE e ROIKDE tem performances bem próximas, figurando dentre os melhores métodos com baixa assimetria.

Figura 117 – Assimetria para o KDE de banda variável.



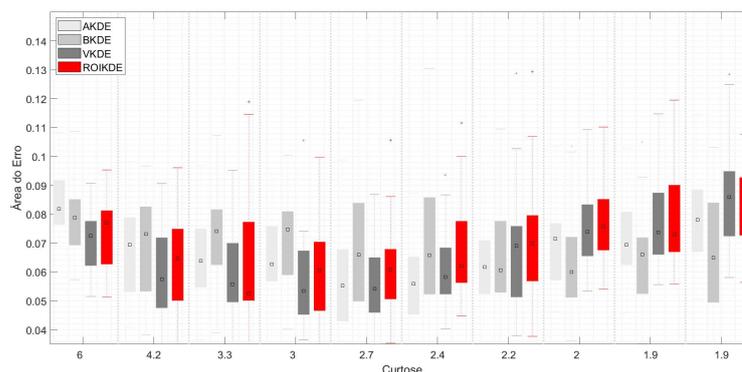
Fonte: Elaborada pelo autor (2020).

No teste de curtose VKDE e ROIKDE continuam com comportamento bem próximo, embora com padrão diferente dos métodos de banda fixa, que tendem a encontrar mais dificuldade em curtoses altas e quando a distribuição tende ao formato de Laplace. O método AKDE encontrou dificuldade em descrever a cauda da distribuição de Laplace, devido a sua estimação bastante ruidosa. Com alta curtose BKDE tem dificuldade em estimar a transição rápida de pico, devido à sua estimação suave. Os métodos VKDE e ROIKDE têm dificuldade em estimar densidades que tendem ao formato uniforme, levando os métodos aos piores resultados com baixa Curtose. Este fato pode ser observado em uma realidade parecida na estimação da distribuição D4a.

4.2 CLASSIFICAÇÃO

A análise de classificação será divididas em duas etapas: Dados Simulados e Identificação de partículas (Geant4). As respectivas distribuições de sinal e ruído de cada conjunto de amostras foram apresentadas na Seção 3.4.1. É importante salientar que todos os algoritmos de estimação de densidade utilizados na verossimilhança *naive*

Figura 118 – Curtose para o KDE de banda variável.

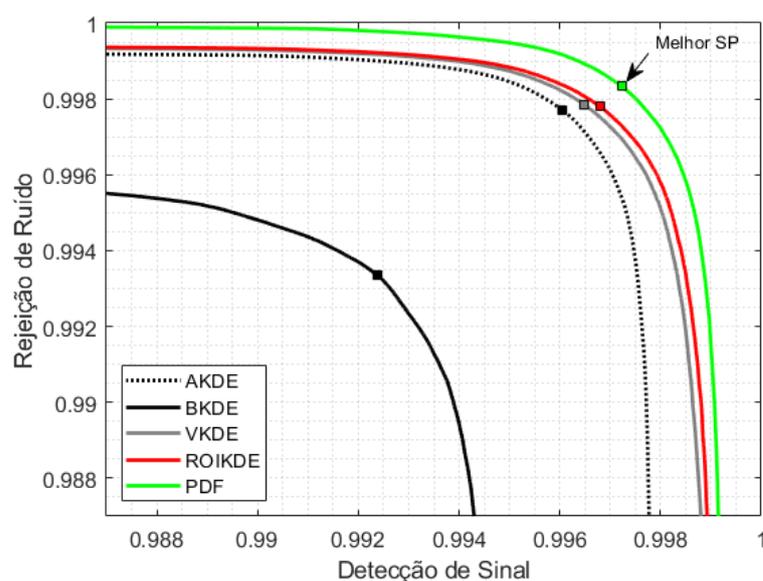


Fonte: Elaborada pelo autor (2020).

são automáticos. Além disso, apenas os seletores de banda variável foram utilizados nessa seção, devido a sua melhor performance geral dentre os demais estimadores não-paramétricos.

Para simplificar a visualização e comparação dos resultados, com diferentes conjuntos de treinamento, foi utilizado o parâmetro SP. A Figura 119 apresenta a ROC para o conjunto de Dados Simulados com 1000 amostras de treinamento, sendo possível observar em todas as curvas o ponto de melhor SP. Vale destacar que o método ROIKDE obteve os melhores valores de SP em todos os testes feitos, propagando o ganho de performance na estimação da densidade conjunta, devido a sua adaptabilidade, para a classificação das amostras.

Figura 119 – Caso representativo (1000 amostras) da ROC para os dados simulados

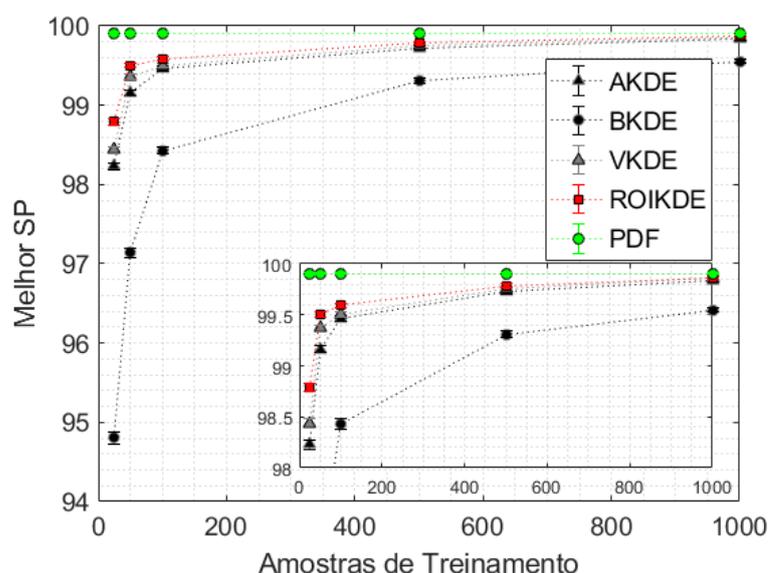


Fonte: Elaborada pelo autor (2020).

4.2.1 Dados Simulados

A Figura 120 apresenta o melhor SP de acordo com o aumento das amostras de treinamento. Foram realizados 50 treinamentos distintos e para cada treinamento foram testados 100 conjuntos de amostras diferentes. Nesse caso é possível perceber o método ROIKDE com os maiores valores de SP, seguido do método VKDE. Como nesse conjunto de amostras existiam distribuições relativamente complexas, com derivadas rápidas e esparsas, o método AKDE se saiu melhor do que o método BKDE, que tende realizar estimações mais suaves. Além disso, o método BKDE teve maior dificuldade em convergir para performance próxima a PDF dentro do intervalo de amostras avaliado.

Figura 120 – Performance dos algoritmos de classificação para os dados simulados.



Fonte: Elaborada pelo autor (2020).

A Tabela 23 mostra a relação numérica dos diferentes métodos, sendo possível constatar a proximidade estatística de VKDE e ROIKDE acima de 500 amostras. O método AKDE converge para performance próxima ao VKDE e ROIKDE com 1000 amostras. Ademais, é possível observar que o método ROIKDE obteve as melhores performances em todos os conjuntos de amostras.

Tabela 23 – Melhor SP para os dados simulados com as densidades vistas anteriormente.

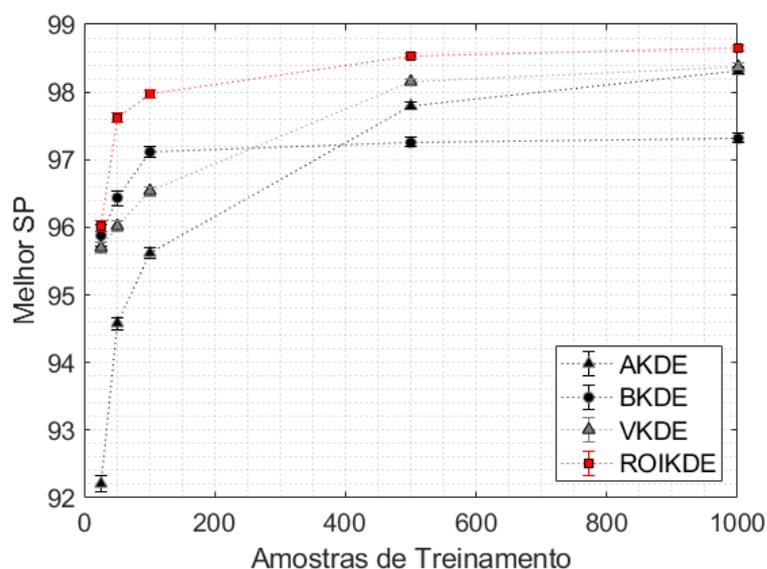
Amostras	AKDE	BKDE	VKDE	ROIKDE	PDF
25	98.22 ± 0.04	94.80 ± 0.08	98.43 ± 0.04	98.78 ± 0.03	99.90 ± 0.01
50	99.16 ± 0.03	97.14 ± 0.06	99.36 ± 0.03	99.50 ± 0.02	99.91 ± 0.01
100	99.46 ± 0.02	98.42 ± 0.04	99.50 ± 0.02	99.59 ± 0.02	99.91 ± 0.01
500	99.72 ± 0.02	99.30 ± 0.03	99.76 ± 0.01	99.78 ± 0.01	99.90 ± 0.01
1000	99.83 ± 0.01	99.54 ± 0.02	99.85 ± 0.01	99.86 ± 0.01	99.91 ± 0.01

Fonte: Elaborada pelo autor (2020).

4.2.2 Identificação de Partículas

Nesta etapa, dados referentes ao problema de identificação de partículas foram classificados. A Figura 121 mostra o algoritmo ROIKDE novamente com a melhor performance em todos os conjuntos de amostras. O método BKDE obteve bons resultados com poucas amostras (até 100), devido à sua estimação suave o tornar mais generalista do que os demais nesta realidade sem derivadas infinitas e com derivadas relativamente menores do que o conjunto de dados anterior. O método AKDE, devido a sua estimação mais ruidosa em densidades relativamente suaves e com poucas amostras, obteve a pior performance até 100 amostras. O método VKDE teve dificuldades em classificar os eventos com poucas amostras devido ao seletor OSCV, por se adaptar bem em diversas situações, tender a fazer estimações menos suaves (do que o seletor SJ utilizado no ROIKDE) em conjuntos com poucas amostras.

Figura 121 – Performance dos algoritmos de classificação para dados referentes a identificação de partículas.



Fonte: Elaborada pelo autor (2020).

A Tabela 24 mostra o método ROIKDE com performance estatisticamente melhor do que os demais métodos. O método AKDE começou com a pior performance de SP e convergiu para um resultado próximo ao VKDE com 1000 amostras. O método BKDE teve uma boa performance com 25 amostras, entretanto, teve a convergência mais lentas do que os demais métodos. O método VKDE conseguiu atingir a segunda melhor performance apenas com 500 amostras.

Tabela 24 – Melhor SP para os dados de identificação de partículas.

Amostras	AKDE	BKDE	VKDE	ROIKDE
25	92.21 ± 0.12	95.87 ± 0.15	95.69 ± 0.08	96.01 ± 0.08
50	94.57 ± 0.09	96.42 ± 0.11	96.02 ± 0.07	97.62 ± 0.07
100	95.62 ± 0.07	97.11 ± 0.07	96.53 ± 0.07	97.98 ± 0.06
500	97.80 ± 0.05	97.26 ± 0.06	98.14 ± 0.05	98.53 ± 0.04
1000	98.31 ± 0.04	97.32 ± 0.06	98.37 ± 0.04	98.65 ± 0.04

Fonte: Elaborada pelo autor (2020).

5 CONCLUSÃO

Neste trabalho foram abordados dois assuntos principais: (1) análise dos seletores de largura de banda desenvolvidos na literatura e (2) classificação de amostras utilizando as estimações realizadas por esses seletores.

Durante o desenvolvimento desta primeira etapa quatro estimadores não-paramétricos de densidade, Histograma, PF, ASH e KDE, foram estudados. Na análise de banda fixa concluiu-se que os seletores de Rudemo e SS obtiveram os resultados mais resilientes, para os três primeiros estimadores, e para o método KDE o seletor OSCV obteve as melhores performances gerais. Ademais, foi possível observar algumas afirmações feitas em trabalhos anteriores sobre os seletores de banda fixa do KDE, como: o seletor SJ obtém bom desempenho para distribuições que não são muito diferentes da Normal (CHIU, 1996); os seletores de PI alcançam os melhores resultados com distribuições unimodais suaves e tem maior probabilidade de subestimar a largura de banda do que os seletores de CV (DEVROYE *et al.*, 1997); de fato existe a probabilidade de encontrar diversos mínimos locais nos seletores BCV1 e UCV; para densidades com baixa rugosidade, SJ tende a ser melhor do que o UCV, e com alta rugosidade SJ tende a superestimar h (SHEATHER, 2004); de fato OSCV obteve as melhores performances médias (HEIDENREICH; SCHINDLER; SPERLICH, 2013); e de forma genérica os seletores de CV alcançaram melhores resultados em densidades muito complexas (BORRAJO; GONZÁLEZ-MANTEIGA; MARTÍNEZ-MIRANDA, 2017).

Além de ratificar algumas afirmações esse trabalho conseguiu avaliar nuances que não foram exploradas em trabalhos à época, como: o seletor UCV apesar de tender a subestimar a largura de banda com muitas amostras, superestimou h em densidades como D1a, D1b, D1c, D3a, D3b e D4a com poucas amostras; da mesma forma que BCV1 como padrão superestima h (JONES; MARRON; SHEATHER, 1996a; JONES; MARRON; SHEATHER, 1996b; CAO; CUEVAS; MANTEIGA, 1994), BCV2 subestimou a largura de banda na densidade D2a com 25 amostras; com poucas amostras BCV1 tem desempenho ruim e BCV2 tende a ser melhor, porém, para realidades com muitas amostras BCV2 tende a figurar dentre as melhores eficiências, enquanto BCV1 tem desempenho intermediário; BCV1 e BCV2 estimam relativamente bem até mesmo em densidades com descontinuidade, com o aumento das amostras; o desempenho de BCV1 de fato é comparável ao SJ (CAO; CUEVAS; MANTEIGA, 1994) em alguns casos, entretanto, para distribuições com baixa rugosidade e poucas amostras, SJ tende a ser melhor que BCV1; em distribuições com alta

rugosidade BCV1 tende a ser melhor, já com descontinuidade BCV2 é bem superior e BCV1 tende a sofrer mais com poucas amostras; o seletor MLCV embora apontado com fraco desempenho em distribuições de caudas pesadas (CAO; CUEVAS; MANTEIGA, 1994), apresentou desempenho geral intermediário em realidades com distribuições complexas; o L1I possui resultados interessantes para um seletor de PI, estimando bem a largura de banda h até mesmo pra distribuições difíceis, alcançando em alguns casos o melhor desempenho (DEVROYE *et al.*, 1997); em distribuições como: D2c, D3a, D3b, D3c, D4a e D4b, o seletor L1I figurou dentre as melhores performances, até mesmo com poucas amostras; entretanto, L1I apresentou em D1c, D1d, D2d convergência lenta, e superestimou h em densidades complexas como: D3d, D4c, D4d (que são multimodais esparsas e com transições rápidas de pico); outro ponto é a sugestão de escolher o mínimo local mais próximo ao SJ (SHEATHER, 2004), que funciona bem para distribuições pouco esparsas. Entretanto, como visto, em distribuições com a variável de escala muito distorcida em relação a Normal, como distribuições multimodais esparsas, o seletor SJ tende a superestimar bastante o h . Portanto, sugerimos utilizar o TCV como referência para escolha do mínimo, e nos raríssimos casos onde TCV possuir mais de um mínimo sugerimos utilizar o SJ.

Ainda sobre estimação de densidades, agora com o KDE de banda variável, mostramos que o método AKDE tende a gerar estimações com maior variância com poucas amostras e o método BKDE tende a apresentar maior viés, principalmente em densidades esparsas. O método VKDE obteve bons resultados utilizando o seletor OSCV como largura de banda, tendo performance inferior apenas ao método desenvolvido nesta tese, o ROIKDE.

Na segunda etapa, de classificação, foi observada a capacidade do algoritmo ROIKDE em selecionar de forma automática a melhor largura de banda de acordo com a complexidade, curtose/assimetria, quantidade de amostras, valor de probabilidade e valor de derivada. Essa versatilidade rendeu ao ROIKDE as melhores performances de classificação nas duas realidades propostas nesse trabalho, indicando que essa abordagem pode ser interessante em algoritmos de classificação, baseados em verossimilhança, principalmente em realidades com poucas amostras onde seja interessante utilizar algoritmos de seleção de largura de banda automática. Vale frisar que este trabalho também fornece uma base sólida para casos onde o analista possa avaliar as características da densidade e escolher o seletor de largura de banda adequado.

Através das análises realizadas e aprofundamento de afirmações do estado da arte, alguns avanços foram mostrados no âmbito deste trabalho. Foi realizada uma comparação

utilizando os seletores mais clássicos e proeminentes da literatura, alguns seletores nunca comparados entre si, principalmente em relação aos seletores de CV. Além disso, com base nos resultados, foi proposta uma alternativa de escolha do método mais resiliente de referência para os seletores CV, justificado através da proporção de mínimos locais em cada seletor de CV e performance desses respectivos seletores. Foram definidos alguns parâmetros essenciais para garantir representatividade dos resultados apresentados, como número de pontos de estimação, *grid*, garantia de convergência dos métodos de CV e escolha compatível do intervalo da estimação e do modelo. Avaliamos os resultados de forma separada e conjunta, possibilitando-nos discernir entre características de cada caso e performance geral dos seletores. Ademais, os seletores foram avaliados em uma abordagem nova relativa a variação de terceiro e quarto momento central, mostrando os seletores L1I e TCV com dificuldades em baixa curtose em uma realidade com a variável de escala bem pequena. Foi criado um método denominado RoIMap, capaz de avaliar a estimação ao longo de toda variável aleatória e perceber a dependência da estimação em relação a derivada e probabilidade. E por fim, o estudo aprofundado dos seletores de largura de banda, juntamente com o RoIMap, possibilitou a criação de um algoritmo híbrido automático de escolha e adaptação da largura de banda variável, denominado ROIKDE, que obteve melhores resultados nos testes de estimação e classificação realizados no âmbito desse trabalho.

5.1 PRÓXIMOS PASSOS

Essa tese mostrou um estudo sobre estimadores não-paramétricos de densidades e classificação baseada em verossimilhança. Algumas diretrizes relevantes e ideias futuras foram separadas, e serão listadas a seguir:

1. Estender para uma análise de resiliência dos algoritmos, adicionando *outliers* nas amostras e verificando o comportamento dos seletores de largura de banda fixa e variável, nessa nova realidade.
2. Avaliar a possibilidade de alterar o parâmetro λ e α da largura de banda variável e também avaliar os valores padrões utilizados, se são as melhores escolhas para diferentes tipos de distribuições
3. Estender o estudo para distribuições multivariadas;

4. Aprofundar a análise do ROIKDE delimitando melhor suas características, etapas e possíveis melhorias;
5. Avaliar o impacto dos seletores de *Bootstrap* nos estimadores não-paramétricos vistos aqui;
6. Avaliar dependências de ordens superiores na estimação não-paramétrica multivariada;
7. Avaliar o impacto da melhora da estimação em classificação de amostras de acordo com o grau de dependência.

REFERÊNCIAS

- ABRAMSON, I. S. On bandwidth variation in kernel estimates-a square root law. **The annals of Statistics**, JSTOR, p. 1217–1223, 1982.
- ALLISON, J.; AMAKO, K.; APOSTOLAKIS, J.; ARAUJO, H.; DUBOIS, P. A.; ASAI, M.; BARRAND, G.; CAPRA, R.; CHAUVIE, S.; CHYTRACEK, R. *et al.* Geant4 developments and applications. **IEEE Transactions on nuclear science**, IEEE, v. 53, n. 1, p. 270–278, 2006.
- BORRAJO, M. I.; GONZÁLEZ-MANTEIGA, W.; MARTÍNEZ-MIRANDA, M. D. Bandwidth selection for kernel density estimation with length-biased data. **Journal of Nonparametric Statistics**, Taylor & Francis, v. 29, n. 3, p. 636–668, 2017.
- BOWMAN, A. W. An alternative method of cross-validation for the smoothing of density estimates. **Biometrika**, Oxford University Press, v. 71, n. 2, p. 353–360, 1984.
- BREIMAN, L.; MEISEL, W.; PURCELL, E. Variable kernel estimates of multivariate densities. **Technometrics**, Taylor & Francis Group, v. 19, n. 2, p. 135–144, 1977.
- CAO, R.; CUEVAS, A.; MANTEIGA, W. G. A comparative study of several smoothing methods in density estimation. **Computational Statistics & Data Analysis**, North-Holland, v. 17, n. 2, p. 153–176, 1994.
- CHAUDHURI, P.; MARRON, J. S. Sizer for exploration of structures in curves. **Journal of the American Statistical Association**, Taylor & Francis, v. 94, n. 447, p. 807–823, 1999.
- CHIU, S.-T. A comparative review of bandwidth selection for kernel density estimation. **Statistica Sinica**, JSTOR, p. 129–145, 1996.
- DEVROYE, D.; BEIRLANT, J.; CAO, R.; FRAIMAN, R.; HALL, P.; JONES, M.; LUGOSI, G.; MAMMEN, E.; MARRON, J.; SÁNCHEZ-SELLERO, C. *et al.* Universal smoothing factor selection in density estimation: theory and practice. **Test**, Springer, v. 6, n. 2, p. 223–320, 1997.
- DOANE, D. P. Aesthetic frequency classifications. **The American Statistician**, Taylor & Francis, v. 30, n. 4, p. 181–183, 1976.
- EFRON, B. **The jackknife, the bootstrap, and other resampling plans**. [S.l.]: Siam, 1982. v. 38.
- EMERSON, J. D.; HOAGLIN, D. C. **Understanding robust and exploratory data analysis**. [S.l.: s.n.], 1983.
- FAWCETT, T. An introduction to roc analysis. **Pattern recognition letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006.
- FELUCH, W.; KORONACKI, J. A note on modified cross-validation in density estimation. **Computational statistics & data analysis**, Elsevier, v. 13, n. 2, p. 143–151, 1992.
- FISHER, R. **On the mathematical foundations of theoretical statistics. Reprinted in Contributions to Mathematical Statistics (by R. A. Fisher)(1950), J.** [S.l.]: Wiley & Sons, New York, 1922.

FORBES, C.; EVANS, M.; HASTINGS, N.; PEACOCK, B. **Statistical distributions**. [S.l.]: John Wiley & Sons, 2011.

FREEDMAN, D.; DIACONIS, P. On the histogram as a density estimator: L² theory. **Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete**, Springer, v. 57, n. 4, p. 453–476, 1981.

FRIEDMAN, J. H. On bias, variance, 0/1-loss, and the curse-of-dimensionality. **Data Mining and Knowledge Discovery**, JSTOR, p. 55–77, 1997. Disponível em: <<https://doi.org/10.1023/A:1009778005914>>.

HABBEMA, J. A stepwise discriminant analysis program using density estimation. In: PHYSICA-VERLAG. **Compstat**. [S.l.], 1974. p. 101–110.

HÁJEK, J. *et al.* Asymptotic normality of simple linear rank statistics under alternatives. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 39, n. 2, p. 325–346, 1968.

HALL, P.; MARRON, J. S. Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. **Probability Theory and Related Fields**, Springer, v. 74, n. 4, p. 567–581, 1987.

HALL, P.; MARRON, J. S. Local minima in cross-validation functions. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, v. 53, n. 1, p. 245–252, 1991.

HALL, P.; MARRON, J. S. *et al.* Estimation of integrated squared density derivatives. **Stat. Prob. Lett.**, v. 6, n. 2, p. 109–115, 1987.

HARDLE, W.; MARRON, J.; WAND, M. Bandwidth choice for density derivatives. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 223–232, 1990.

HART, J. D.; YI, S. One-sided cross-validation. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 93, n. 442, p. 620–631, 1998.

HEIDENREICH, N.-B.; SCHINDLER, A.; SPERLICH, S. Bandwidth selection for kernel density estimation: a review of fully automatic selectors. **AStA Advances in Statistical Analysis**, Springer, v. 97, n. 4, p. 403–433, 2013.

JONES, M.; KAPPENMAN, R. On a class of kernel density estimate bandwidth selectors. **Scandinavian Journal of Statistics**, JSTOR, p. 337–349, 1992.

JONES, M. C.; MARRON, J. S.; SHEATHER, S. J. A brief survey of bandwidth selection for density estimation. **Journal of the American Statistical Association**, Taylor & Francis, v. 91, n. 433, p. 401–407, 1996.

JONES, M. C.; MARRON, J. S.; SHEATHER, S. J. Progress in data-based bandwidth selection for kernel density estimation. **Computational Statistics**, Springer Verlag, v. 11, n. 3, p. 337–381, 1996.

KENNY, Q. Y. *et al.* Indicator function and its application in two-level factorial designs. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 31, n. 3, p. 984–994, 2003.

KNUTH, K. H. Optimal data-based binning for histograms. **arXiv preprint physics/0605197**, 2006.

LIU, B.; YANG, Y.; WEBB, G. I.; BOUGHTON, J. A comparative study of bandwidth choice in kernel density estimation for naive bayesian classification. In: SPRINGER. **Pacific-Asia Conference on Knowledge Discovery and Data Mining**. [S.l.], 2009. p. 302–313.

LOADER, C. R. *et al.* Bandwidth selection: classical or plug-in? **The Annals of Statistics**, Institute of Mathematical Statistics, v. 27, n. 2, p. 415–438, 1999.

LOFTSGAARDEN, D. O.; QUESENBERRY, C. P. *et al.* A nonparametric estimate of a multivariate density function. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 36, n. 3, p. 1049–1051, 1965.

LOLLA, S. V. G.; HOBEROCK, L. L. On selecting the number of bins for a histogram. In: CITESEER. **Proceedings of the International Conference on Data Mining (DMIN)**. [S.l.], 2011. p. 1.

MARTINEZ-MIRANDA, M.; NIELSEN, J.; SPERLICH, S. **One sided Cross Validation in density estimation**, in “**Operational Risk Towards Basel III: Best Practices and Issues in Modeling, Management and Regulation**”, ed. GN Gregoriou. [S.l.]: John Wiley and Sons, Hoboken, New Jersey, 2009.

PARK, B. U.; MARRON, J. S. Comparison of data-driven bandwidth selectors. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 85, n. 409, p. 66–72, 1990.

PARZEN, E. On estimation of a probability density function and mode. **The annals of mathematical statistics**, JSTOR, v. 33, n. 3, p. 1065–1076, 1962.

RUDEMO, M. Empirical choice of histograms and kernel density estimators. **Scandinavian Journal of Statistics**, JSTOR, p. 65–78, 1982.

SAVCHUK, O. Y. One-sided cross-validation for nonsmooth density functions. **arXiv preprint arXiv:1703.05157**, 2017.

SCOTT, D. **Multivariate density estimation. 1992**. [S.l.]: Wiley, New York, 1992.

SCOTT, D. W. On optimal and data-based histograms. **Biometrika**, Oxford University Press, v. 66, n. 3, p. 605–610, 1979.

SCOTT, D. W. Averaged shifted histograms: effective nonparametric density estimators in several dimensions. **The Annals of Statistics**, JSTOR, p. 1024–1040, 1985.

SCOTT, D. W. Frequency polygons: theory and application. **Journal of the American Statistical Association**, Taylor & Francis, v. 80, n. 390, p. 348–354, 1985.

SCOTT, D. W. **Multivariate density estimation: theory, practice, and visualization**. [S.l.]: John Wiley & Sons, 2015.

SCOTT, D. W.; SHEATHER, S. J. Kernel density estimation with binned data. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 14, n. 6, p. 1353–1359, 1985.

SCOTT, D. W.; TERRELL, G. R. Biased and unbiased cross-validation in density estimation. **Journal of the American Statistical Association**, Taylor & Francis, v. 82, n. 400, p. 1131–1146, 1987.

- SHEATHER, S. J. Density estimation. **Statistical science**, JSTOR, p. 588–597, 2004.
- SHEATHER, S. J.; JONES, M. C. A reliable data-based bandwidth selection method for kernel density estimation. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 683–690, 1991.
- SHIMAZAKI, H.; SHINOMOTO, S. A method for selecting the bin size of a time histogram. **Neural computation**, MIT Press, v. 19, n. 6, p. 1503–1527, 2007.
- SHIMAZAKI, H.; SHINOMOTO, S. Kernel bandwidth optimization in spike rate estimation. **Journal of computational neuroscience**, Springer, v. 29, n. 1-2, p. 171–182, 2010.
- SILVERMAN, B. W. Algorithm as 176: Kernel density estimation using the fast fourier transform. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, JSTOR, v. 31, n. 1, p. 93–99, 1982.
- SILVERMAN, B. W. **Density estimation for statistics and data analysis**. [S.l.]: Routledge, 1986.
- SOUZA, D. de M. Estimaco de densidades multivariadas para a filtragem de eventos baseado em um detector de altas energias com fina segmentaco. 2015.
- SOUZA, D. M.; COSTA, I. A.; NBREGA, R. A. A study of distance/similarity measurements in the context of signal processing (density estimation). In: IEEE. **Instrumentation Systems, Circuits and Transducers (INSCIT), 2017 2nd International Symposium on**. [S.l.], 2017. p. 1–6.
- STURGES, H. A. The choice of a class interval. **Journal of the american statistical association**, New York, v. 21, n. 153, p. 65–66, 1926.
- STUTE, W. Modified cross-validation in density estimation. **Journal of Statistical Planning and Inference**, Elsevier, v. 30, n. 3, p. 293–305, 1992.
- STUTE, W.; SCHUMANN, G. A general glivenko-cantelli theorem for stationary sequences of random observations. **Scandinavian Journal of Statistics**, JSTOR, p. 102–104, 1980.
- TERRELL, G. R.; SCOTT, D. W. Variable kernel density estimation. **The Annals of Statistics**, JSTOR, p. 1236–1265, 1992.
- TUKEY, P.; TUKEY, J. W. Data driven view selection, agglomeration, and sharpening. **Interpreting multivariate data**, Chichester: Wiley, p. 215–243, 1981.
- WAN, X.; WANG, W.; LIU, J.; TONG, T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. **BMC medical research methodology**, BioMed Central, v. 14, n. 1, p. 135, 2014.
- WAND, M. Data-based choice of histogram bin width. **The American Statistician**, Taylor & Francis, v. 51, n. 1, p. 59–64, 1997.
- WOLTERS, M. A. A greedy algorithm for unimodal kernel density estimation by data sharpening. 2009.
- ZHAO, Q.; XU, M.; FRANTI, P. Knee point detection on bayesian information criterion. In: IEEE. **Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on**. [S.l.], 2008. v. 2, p. 431–438.

ZHU, Z.; NANDI, A. K. **Automatic modulation classification: principles, algorithms and applications.** [S.l.]: John Wiley & Sons, 2015.

APÊNDICE A – CRITÉRIO DE ERRO PARA ESTIMAÇÃO DE DENSIDADES

A comparação entre diferentes tipos de estimadores pressupõe a especificação de um critério que possa diferenciá-los, e conseqüentemente ser otimizado. O conceito de otimalidade não é absoluto, porém está intimamente ligado a escolha de um critério. A escolha do critério é subjetiva podendo variar desde argumentos teóricos até a aplicação em questão. Na estimação paramétrica, respeitadas as premissas corretas, o estimador ótimo será ideal para qualquer propósito relacionado. Já em estimação não-paramétrica, um estimador pode ser “ótimo” em uma aplicação e sub-ótimo para outra.

Geralmente quando calculamos parâmetros com estimadores enviesados, a variância é substituída pelo *Mean Square Error* (MSE), que pode ser representada pela soma da variância e o viés ao quadrado. Para estimação pontual de uma função de densidade pelo estimador $\hat{f}(x)$, temos:

$$\begin{aligned} MSE\{\hat{f}(x)\} &= E[\hat{f}(x) - f(x)]^2 \\ &= Var\{\hat{f}(x)\} + Bias^2\{\hat{f}(x)\} \end{aligned} \quad (A.1)$$

Onde $Bias\{\hat{f}(x)\} = E[\hat{f}(x)] - f(x)$. Essa equação trata o problema de estimativa de densidade não-paramétrica como um problema de estimação pontual, com parâmetro desconhecido $\theta = f(x)$. Embora tal abordagem seja interessante o objetivo principal desta tese é a estimação de toda superfície da densidade. Geralmente, devido a sua fácil manipulação, predomina na literatura o critério L^2 nesse contexto (SCOTT, 2015), sendo referido como ISE:

$$ISE = \int [\hat{f}(x) - f(x)]^2 dx \quad (A.2)$$

Apesar da fácil manipulação o ISE é uma variável aleatória complexa, devido a dependência de uma PDF desconhecida, de um estimador particular e do tamanho da amostra. O ISE é relativo a realizações particulares de n pontos e para a maioria das finalidades será necessário calcular a média do ISE sobre essas realizações. Sendo denominado MISE.

$$\begin{aligned} MISE &= E[ISE] = E \int [\hat{f}(x) - f(x)]^2 dx \\ &= \int E [\hat{f}(x) - f(x)]^2 dx = \int MSE \{\hat{f}(x)\} dx \\ &\equiv IMSE \end{aligned} \quad (A.3)$$

A quantidade *Integrated Mean Square Error* (IMSE) evidencia que o MISE tem duas interpretações equivalentes: É a medida do erro global médio e o erro pontual acumulado.

A.1 CRITÉRIO L^2 APLICADO AO HISTOGRAMA

O histograma, com *bin* B_k de tamanho h , é definido por:

$$\hat{f}(x) = \frac{v_k}{nh}, \quad x \in B_k \quad (\text{A.4})$$

A análise da variável aleatória do histograma, $\hat{f}(x)$ é simples, visto que a contagem dos *bins* v_k é similar a uma variável aleatória binomial.

$$v_k \sim B(n, p_k), \quad p_k = \int_{B_k} f(t) dt \quad (\text{A.5})$$

Considere o MSE de $\hat{f}(x)$ para $x \in B_k$. Agora, $E[v_k] = np_k$ e $Var[v_k] = np_k(1 - p_k)$. Portanto,

$$Var(\hat{f}(x)) = \frac{Var v_k}{(nh)^2} = \frac{p_k(1-p_k)}{nh^2} \quad (\text{A.6})$$

$$Bias \hat{f}(x) = E[\hat{f}(x)] - f(x) = \frac{1}{nh} E v_k - f(x) = \frac{p_k}{h} - f(x) \quad (\text{A.7})$$

Suponha que $f(x)$ seja uma função Lipschitz contínua, então $|f(x) - f(y)| < \gamma_k |x - y|$. Pelo Teorema do Valor Médio (TVM) temos que:

$$p_k = \int_{B_k} f(t) dt = hf(\xi), \quad \xi \in B_k \quad (\text{A.8})$$

Portanto, a variância e o viés são definidos por:

$$Var(\hat{f}(x)) \leq \frac{p_k}{nh^2} = \frac{f(\xi)}{nh} \quad (\text{A.9})$$

$$|Bias \hat{f}(x)| = |f(\xi) - f(x)| \leq \gamma_k |\xi - x| \leq \gamma_k h \quad (\text{A.10})$$

Sob a suposição de A.1:

$$MSE \hat{f}(x) = \frac{f(\xi)}{nh} + \gamma^2 h^2 \quad (\text{A.11})$$

No contexto prático é necessário computar o MISE através da soma do MSE, de cada *bin*, ao longo de toda superfície de densidade. A IV é definida por:

$$IV = \int_{-\infty}^{\infty} Var \hat{f}(x) dx = \sum_{-\infty}^{\infty} \int_{B_k} Var \hat{f}(x) dx \quad (A.12)$$

Da equação A.6 a integral sobre B_k é $p_k(1 - p_k)/nh$. Agora $\sum p_k = \int f(x) dx = 1$. Lembrando que $\sum \phi(\xi)h = \int \phi(x) dx + o(1)$ de acordo com a aproximação da Integral de Riemann. Portanto, utilizando a teoria TVM para p_k , $\sum p_k = \sum f(\xi)^2 h^2 = h \sum f(\xi)^2 h = h[\int f(x)^2 dx + o(1)]$. Temos:

$$IV = \frac{1}{nh} - \frac{R(f)}{n} + o(n^{-1}) \quad (A.13)$$

Onde $R(f)$, como visto em refeq:06, representa a rugosidade da função f .

Para o cálculo do viés, utilizando um *bin* $B_0 = [0, h)$. A probabilidade do *bin* pode ser aproximada por:

$$\begin{aligned} p_0 &= \int_0^h f(t) dt = \int_0^h [f(x) + (t-x)f'(x) + \frac{1}{2}(t-x)^2 + \dots] dt \\ &= hf(x) + h\left(\frac{h}{2} - x\right) f'(x) + o(h^3) \end{aligned} \quad (A.14)$$

Portanto, da equação A.7 temos:

$$Bias \hat{f}(x) = \frac{p_0}{h} - f(x) = \left(\frac{h}{2} - x\right) f'(x) + o(h^2) \quad (A.15)$$

Utilizando o Teorema do Valor Médio Generalizado (TVMG), o principal termo do *Integrated Squared Bias* (ISB) para este *bin* é:

$$\int_{B_0} \left(\frac{h}{2} - x\right)^2 f'(x)^2 dx = f'(\eta_0)^2 \int_{-\infty}^{\infty} \left(\frac{h}{2} - x\right)^2 dx = \frac{h^3}{12} f'(\eta_0)^2, \eta_0 \in B_0 \quad (A.16)$$

Considerando a generalização para todos os bins B_k temos o ISB total:

$$ISB = \frac{h^2}{12} \sum_{k=-\infty}^{\infty} f'(\eta_0)^2 \times h = \frac{h^2}{12} \int_{-\infty}^{\infty} f'(x)^2 dx + o(h^2) \quad (A.17)$$

Removendo a componente referente a Integral de Riemann é possível calcular o *Asymptotic Integrated Squared Bias* (AISB), versão assintótica do viés $ISB = AISB + o(h^2)$.

$$AISB = \frac{h^2}{12} \int_{-\infty}^{\infty} f'(x)^2 dx = \frac{1}{12} h^2 R(f') \quad (\text{A.18})$$

De forma similar, a *Asymptotic Integrated Variance* (AIV) e o AMISE se referem aos principais termos nas aproximações de *Integrated Variance* (IV) e MISE, respectivamente.

A.2 CRITÉRO L^2 APLICADO AO KDE

A análise estatística do estimador por *kernel* é mais simples do que por histograma, como o estimador por *kernel* é média aritmética de n variáveis (i.i.d.):

$$K_h(x, X_i) = \frac{1}{h} K\left(\frac{x-X_i}{h}\right) \quad (\text{A.19})$$

Assim sendo,

$$E\{\hat{f}(x)\} = E K_h(x, X), \quad Var\{\hat{f}(x)\} = \frac{1}{n} Var K_h(x, X), \quad (\text{A.20})$$

O valor esperado pode ser calculado como:

$$\begin{aligned} E K_h(x, X) &= \int \frac{1}{h} K\left(\frac{x-t}{h}\right) f(t) dt = \int K(w) f(x-hw) dw \\ &= f(x) \int K(w) - h f'(x) dw \int w K(w) + \frac{1}{2} h^2 f''(x) \int w^2 K(w) + \dots, \end{aligned} \quad (\text{A.21})$$

A variância é dada por:

$$Var K_h(x, X) = E\left[\frac{1}{h} K\left(\frac{x-X_i}{h}\right)\right]^2 - \left[E\frac{1}{h} K\left(\frac{x-X_i}{h}\right)\right]^2 \quad (\text{A.22})$$

O segundo termo de A.22 foi calculado em A.21. Enquanto o primeiro termo pode ser aproximado por:

$$\int \frac{1}{h^2} K\left(\frac{x-t}{h}\right)^2 f(t) dt = \int \frac{1}{h} K(w)^2 f(x-hw) dw \approx \frac{f(x)R(K)}{h} \quad (\text{A.23})$$

Se o kernel K satisfaz,

$$\int K(w) = 1, \quad \int w K(w) = 0, \quad \int w^2 K(w) \equiv \sigma_K^2 > 0 \quad (\text{A.24})$$

O valor esperado de $\hat{f}(x)$ será igual a $f(x)$ para ordem de $o(h^2)$. Portanto,

$$\begin{aligned} Bias(x) &= \frac{1}{2} \sigma_K^2 h^2 f''(x) + o(h^4) \Rightarrow \\ ISB &= \frac{1}{4} \sigma_K^4 h^4 R(f'') + o(h^6) \end{aligned} \quad (\text{A.25})$$

Similarmente em A.22, A.23 e A.21,

$$\begin{aligned} Var(x) &= \frac{f(x)R(K)}{nh} - \frac{f(x)^2}{n} + o\left(\frac{h}{n}\right) \Rightarrow \\ IV &= \frac{R(K)}{nh} - \frac{R(f)^2}{n} + \dots \end{aligned} \quad (\text{A.26})$$

APÊNDICE B – PRÉ-PROCESSAMENTO

Esta seção mostra a necessidade de um Pré-Processamento no conjunto de amostras para a melhora da performance dos estimadores não-paramétricos. Esse pré-processamento utiliza técnicas de estatística robusta para amenizar efeitos indesejáveis no conjunto amostras.

Em estudos de estimação não-paramétrica os métodos dependem principalmente do conjunto de amostras para sua construção. Através desses dados seus momentos e variáveis de escala são extraídos e utilizados em cálculos de largura de banda ou tamanho do *bin*, por exemplo. Além disso, o próprio *range* de estimação pode sofrer distorções devido a algumas características dos dados. Nesta etapas dois problemas principais podem ser citados: *Outliers* e *Spikes*.

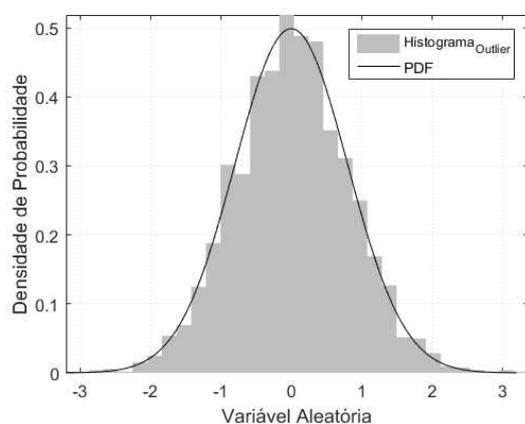
B.1 OUTLIERS

Os *outliers* distorcem a estimação não-paramétrica de densidade, degradando as inferências feitas pelos estimadores baseados em escala. Nesta seção será mostrado exemplos de como ocorre essa degradação e alguns métodos para detecção de *outliers*.

B.1.1 Degradação da Estimação

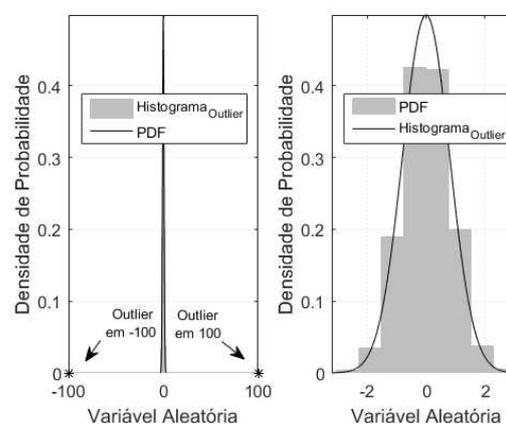
A Figura 122 mostra a variável aleatória D1a sem *outliers*, representada pelo histograma e sua PDF. Para o cálculo do número de *bins* escolhemos o método do Scott, que para um conjunto de 2500 amostras escolheu 31 *bins*. Na Figura 123 foram adicionados dois *outliers* simétricos em 100 e -100 e claramente a área de erro entre histograma e PDF aumentou, o cálculo dos *bins* foi 264, porém o aumento expressivo do número de *bins* não foi suficiente para obter bons resultados no novo intervalo. Esse efeito justifica-se pelo método de Scott, como visto em 2.1, utilizar somente o desvio padrão σ e o número de eventos n como informação. Neste caso, a média permaneceu a mesma devido aos *outliers* simétricos, já o desvio padrão variou de $\sigma = 0.8177$ sem *outlier* para $\sigma = 2.9436$ com *outlier*.

Figura 122 – Variável aleatória sem *outlier* na distribuição D1a.



Fonte: Elaborada pelo autor (2020).

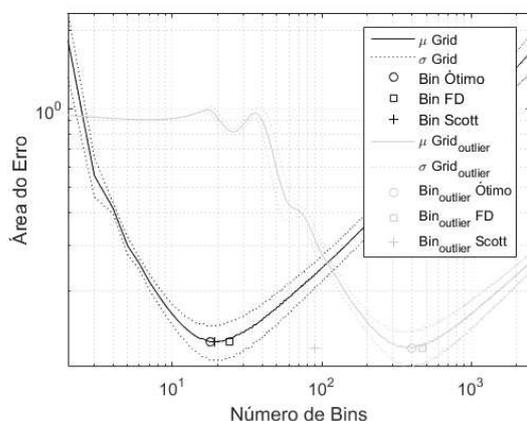
Figura 123 – Adição de *outliers* simétricos (Esquerda) - *Zoom* no histograma (direita)



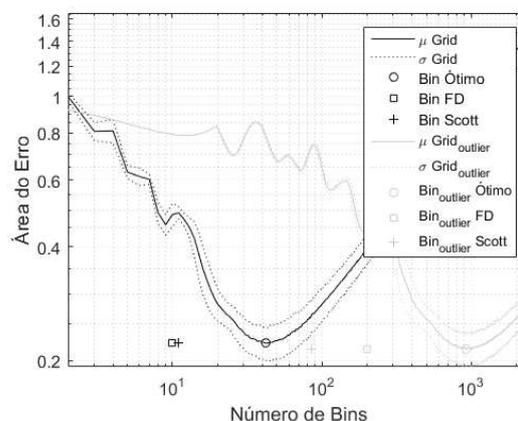
Fonte: Elaborada pelo autor (2020).

Extrapolando o pensamento, na Figura 124 e 125, podemos analisar o efeito do *outlier* em duas distribuições, a D1a e a D2a. Dado um conjunto com 1000 amostras é possível avaliar o efeito de um *outlier* de valor 100 na estimação de um histograma, ao variarmos o número de *bins*, em relação a PDF. A Figura 124 utiliza a variável aleatória D1a e mostra duas realidades. Em preto temos a análise sem *outlier* em cinza a análise com *outlier* = 100, a curva contínua representa o a área média do erro entre histograma e PDF, a linha pontilhada representa o desvio padrão dessa área do erro. Nesse panorama é possível avaliar os estimadores de *bin* Scott e FD, nas duas realidades, sem e com *outlier*. O número de *bins* calculado pelo Scott, com uma variância de $\sigma = 1.3$, alcança o mesmo mínimo do Bin ótimo, como esperado nessa realidade sem *outlier*. Porém, o Método FD superestima o número de *bins*. Com *outlier*, como visto anteriormente, o Scott estima mal o número de *bins*. Entretanto, o FD estima próximo ao número ótimo de *bins*. Lembrando que o FD é uma alternativa robusta ao Scott. A Figura 125, mostra uma realidade onde nenhum dos métodos, FD e Scott, são ótimos, mesmo na realidade sem *outlier*.

Figura 124 – *Outlier* em 100, adicionado na Figura 125 – *Outlier* em 100, adicionado na variável aleatória D1a. Figura 125 – *Outlier* em 100, adicionado na variável aleatória D2a.



Fonte: Elaborada pelo autor (2020).



Fonte: Elaborada pelo autor (2020).

B.1.2 Detecção de *Outlier*

Algumas métodos utilizam o conhecimento *à priori* das características do conjunto de amostras, como os métodos baseados em classificação e os métodos estatísticos paramétricos. Outros métodos, baseados em estatística, utilizam estimadores não-paramétricos como KDE e Histograma. No contexto de estimação não-paramétrica de densidades, devido a sua natureza, seria natural utilizar métodos que dispensassem informações prévias sobre o modelo dos dados e métodos que não utilizassem estimação não-paramétrica de densidade na detecção de *outlier*, evitando a natureza cíclica do problema. Outro fator é que a escolha do método depende das características dos dados, inviabilizando uma conclusão geral. Nesta seção será apresentada uma proposta capaz de solucionar problemas multivariados, com o intuito de demonstrar o conceito de detecção de *outlier*.

B.1.2.1 Proposta de Algoritmo de Detecção

Como ponto de partida utilizaremos a redução dimensional através da distância euclidiana, logo após a abordagem estatística acima citada será desenvolvida, utilizando um método robusto e outro não robusto. A construção do algoritmo será não-supervisionada, sendo separada em duas etapas. Uma para calcular a redução dimensional e outra para identificar quais amostras são *outliers*.

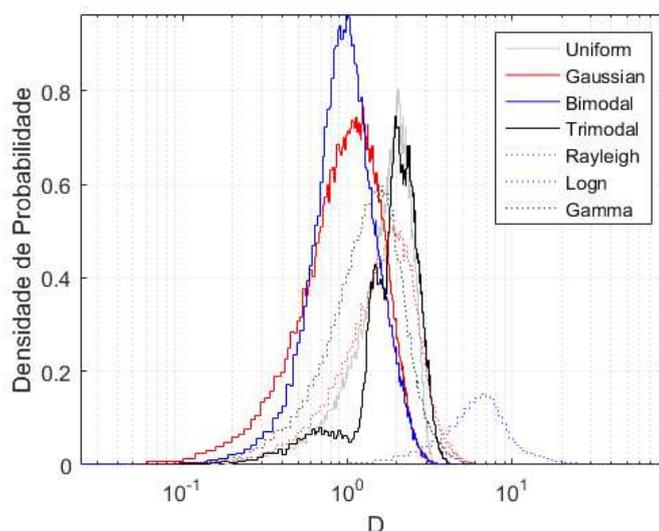
1. **Redução dimensional:** Nessa etapa a redução dimensional foi computadas via distância Euclidiana, mostrada na equação B.1, concatenando as N -dimensões em

uma dimensão que representa a distância D . A equação abaixo representa a distância entre os ponto p e q , em um sistema de coordenadas cartesianas n -dimensional.

$$D(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (\text{B.1})$$

A Figura 126 mostra o histograma normalizado da distância D removendo os valores referentes aos *outliers*, onde é possível perceber que nenhuma das distribuições possuem a diferença entre suas amostras a distribuição Gaussiana. Entretanto, como não é o foco deste trabalho, suposições de distribuição Gaussiana serão feitas nesta etapa, sendo utilizados dois métodos conhecidos da abordagem estatística denominados Z-Score e Z-Score Robusto.

Figura 126 – Distribuição das distâncias D para algumas distribuições.



Fonte: Elaborada pelo autor (2020).

2. **Identificação:** Após calculada a Distância D o algoritmo precisa identificar os *outliers*. Portanto, testes foram feitos utilizando dois seletores. Uma versão não robusta e outra robusta.

B.1.2.2 Z-Score

Outro método para identificar *outliers* é o chamado *Z-scores*. Para obter o *z-score* absoluto os elementos das variáveis são padronizados, extraíndo de cada elemento a média da variável e dividindo por seu desvio padrão correspondente:

$$z = \frac{|x - \mu(x)|}{\sigma(x)} \quad (\text{B.2})$$

Então, cada elemento com *z-score* maior que 2,5 ou 3 pode ser identificado como *outlier*. Esse número é justificado pela suposição de distribuição normal do *z-scores*, ou seja, espera-se que 99,40% ou 99,73% dos dados esteja dentro do intervalo de 2.5 e 3 vezes o desvio padrão do conjunto de amostras. Entretanto, neste trabalho, utilizamos 4 vezes o desvio padrão, que representa 99,99%. Embora não seja uma característica desejável, a suposição de distribuição normal será feita quando a falta de eventos justificar essa premissa. Decidir essa utilização não é uma tarefa trivial, sendo diretamente relacionada ao conceito de "maldição da otimização".

B.1.2.3 *Z-scores Robusto*

Alterando os parâmetros pelos seus já conhecidos respectivos robustos, o *z-scores* robusto é dado por:

$$z = \frac{|x - \tilde{\mu}(x)|}{\sigma_{Q_n}(x)} \quad (\text{B.3})$$

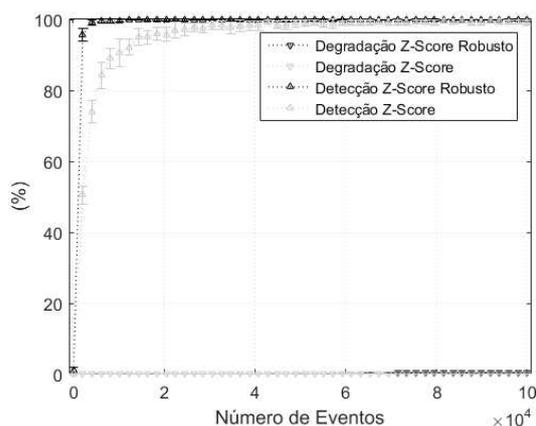
B.1.3 **Testando algoritmos de Detecção de *Outliers***

Para avaliar a abordagem citada acima foi criado um conjunto de amostras tridimensionais utilizando a distribuição D2a e D1b, apresentadas na seção 3.1. Então, 100 amostras de *outliers* foram adicionadas a esse conjunto, com uma distribuição normal de média $\mu = 50\sigma_d$, onde σ_d representa o desvio padrão *Median Absolute Deviation* (MAD) de cada dimensão.

As Figuras 127 e 128 mostram a porcentagem de degradação e detecção dos algoritmos de detecção de *outliers* em relação ao aumento do número de amostras. Por degradação entende-se quantas amostras "boas" foram identificadas como *outlier*. Já detecção significa quantos eventos *outliers* foram identificados corretamente. A figura 127 mostra o algoritmo operando na distribuição D2a. Neste caso, percebemos que o algoritmo robusto obtêm melhores índices de detecção mesmo com poucas amostras, entretanto é mais degradado com mais amostras. Este fato justifica-se pelos *outliers* influenciarem cada vez menos o algoritmo Z-score não robusto, no cálculo dos momentos, com o aumento do número de eventos, que aumenta sua capacidade de detecção. A degradação do algoritmo

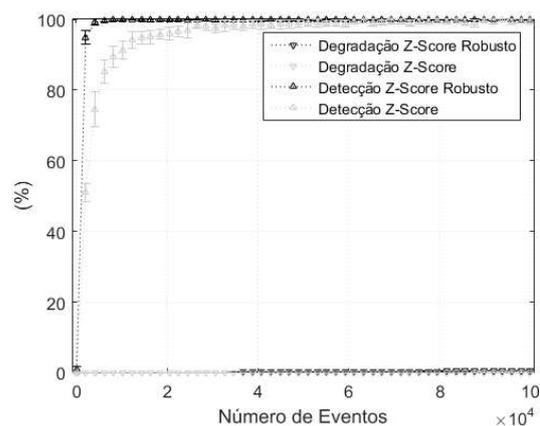
robusto aumenta devido o desvio padrão ser menor neste caso, por evitar o efeito dos *outliers*, do que o algoritmo Z-score não robusto. A Figura 128 representa o funcionamento com a distribuição D1b, e o raciocínio é análogo a D2a.

Figura 127 – Detecção e degradação por *outlier* na distribuição D2a.



Fonte: Elaborada pelo autor (2020).

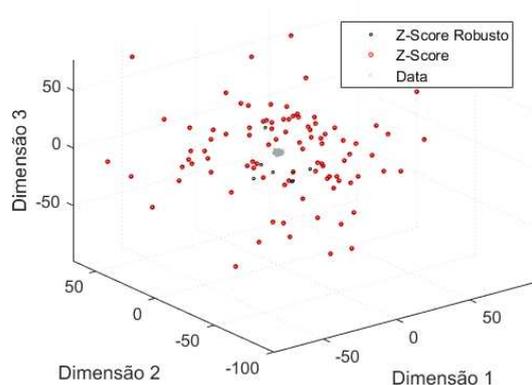
Figura 128 – Detecção e degradação por *outlier* na distribuição D1b.



Fonte: Elaborada pelo autor (2020).

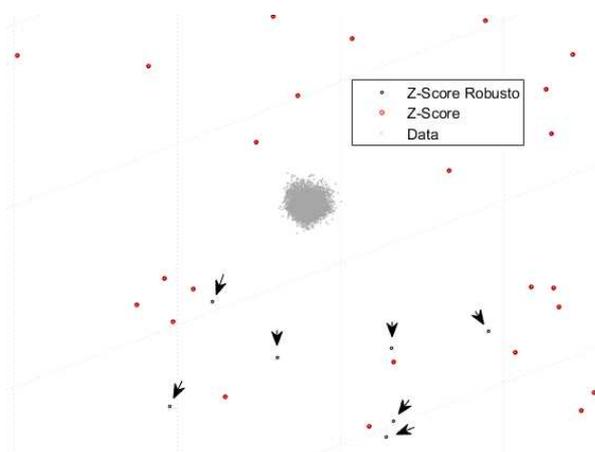
Analisando as Figuras 129 e 130 tridimensionais, é possível perceber funcionamento dos algoritmos com 10000 amostras. Em cinza, na região central temos a distribuição D2a e seus *outliers*, as amostras consideradas *outlier* pelos algoritmos foram circuladas de preto pelo Z-Score Robusto e vermelho pelo Z-Score. É possível perceber, nesta realidade, a tendência do algoritmo robusto de detectar *outliers* mais próximos da distribuição, que em contrapartida, com o aumento do número de eventos, resulta na maior degradação do conjunto de amostras.

Figura 129 – *Outliers* adicionados a distribuição tridimensional D2a.



Fonte: Elaborada pelo autor (2020).

Figura 130 – Zoom na distribuição D2a.



Fonte: Elaborada pelo autor (2020).

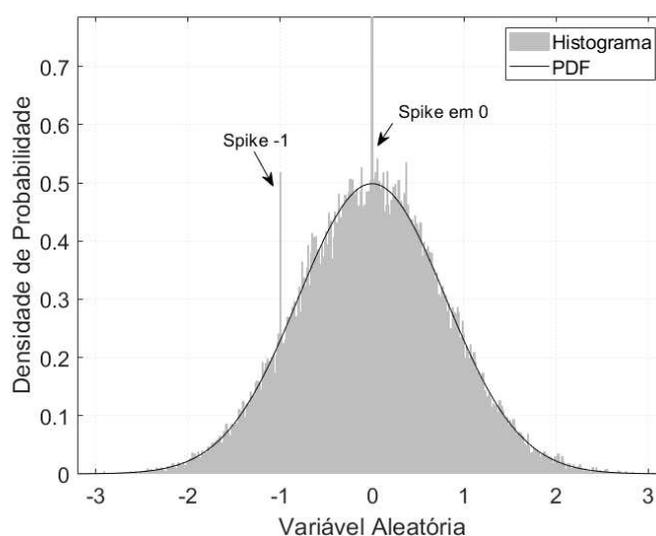
B.2 SPIKES

Spikes são características de alguns conjuntos de dados causadas por diferentes fatores:

1. Operações matemáticas: Em algumas variáveis podemos ter operações como divisão por zero, $\log(0)$, números imaginários, que podem levar a um acúmulo de valores em um determinado *bin* escolhido para representá-los.
2. Resolução: Alguns tipos de variáveis podem levar a valores que necessitam ser arredondados, causando acúmulo em determinado *bin*.
3. Proposital: Alguns erros nas variáveis podem ser representados por um valor discrepante, como -9999, por exemplo. Neste caso o *spike* pode se confundir com *outlier*, sendo necessária a avaliação do analista dos dados.

A Figura 131 mostra um exemplo de *spike* na variável aleatória Gaussiana, em -1 e 0. A utilização desses valores indiscriminadamente, sem análise prévia, distorce tanto a média quanto o desvio padrão do conjunto de dados, diminuindo a performance dos algoritmos de estimação. Tanto o significado desses valores, quanto a sua remoção devem estar a critério do analista. Uma alternativa seria sua remoção antes da estimação, e sua adição posterior, normalizada pela área correspondente, quando há necessidade de representá-los.

Figura 131 – Variável aleatória Gaussiana com descontinuidade em 0 e -1.



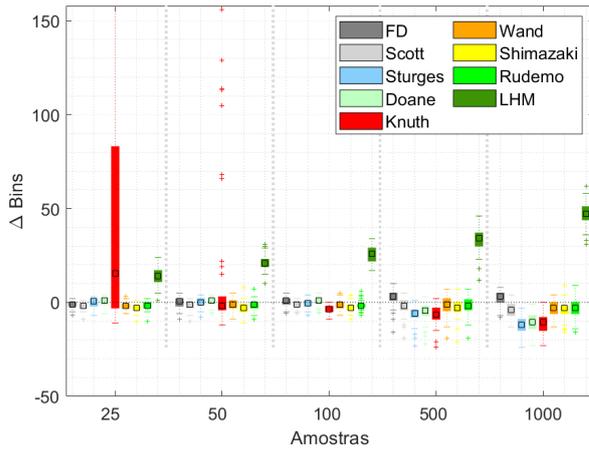
Fonte: Elaborada pelo autor (2020).

APÊNDICE C – AVALIAÇÃO DOS MÉTODOS

C.1 BANDA FIXA

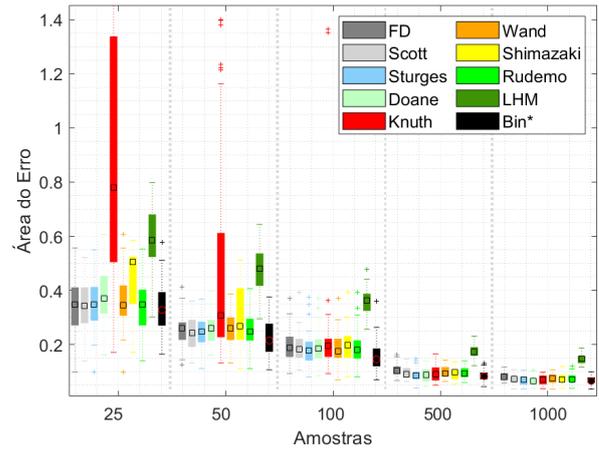
C.1.1 Histograma, PF e ASH

Figura 132 – Variação da Binagem utilizando Histograma na distribuição D1a



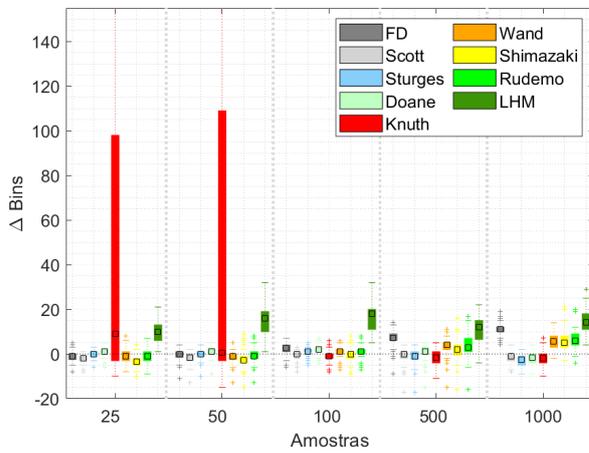
Fonte: Elaborada pelo autor (2020).

Figura 133 – Área do erro utilizando Histograma na distribuição D1a



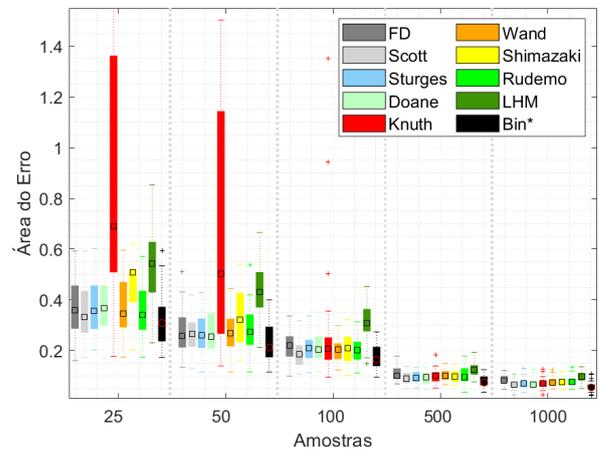
Fonte: Elaborada pelo autor (2020).

Figura 134 – Variação da Binagem utilizando PF na distribuição D1a



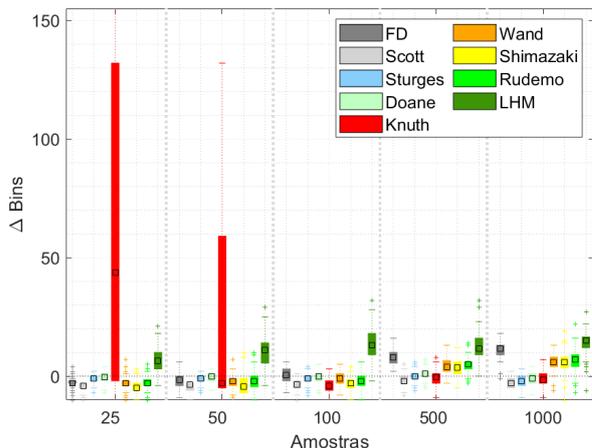
Fonte: Elaborada pelo autor (2020).

Figura 135 – Área do erro utilizando PF na distribuição D1a



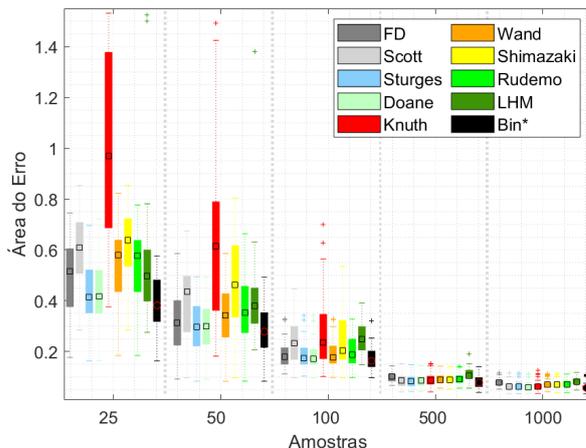
Fonte: Elaborada pelo autor (2020).

Figura 136 – Variação da Binagem utilizando ASH na distribuição D1a



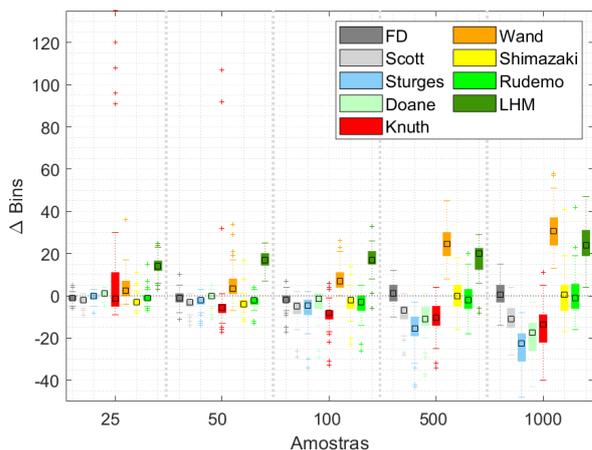
Fonte: Elaborada pelo autor (2020).

Figura 137 – Área do erro utilizando ASH na distribuição D1a



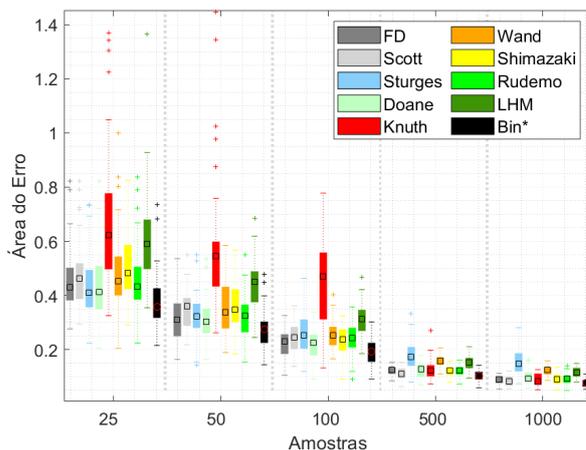
Fonte: Elaborada pelo autor (2020).

Figura 138 – Variação da Binagem utilizando Histograma na distribuição D1b



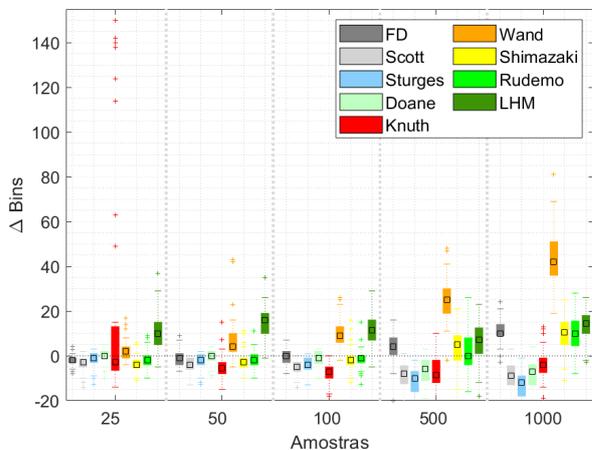
Fonte: Elaborada pelo autor (2020).

Figura 139 – Área do erro utilizando Histograma na distribuição D1b



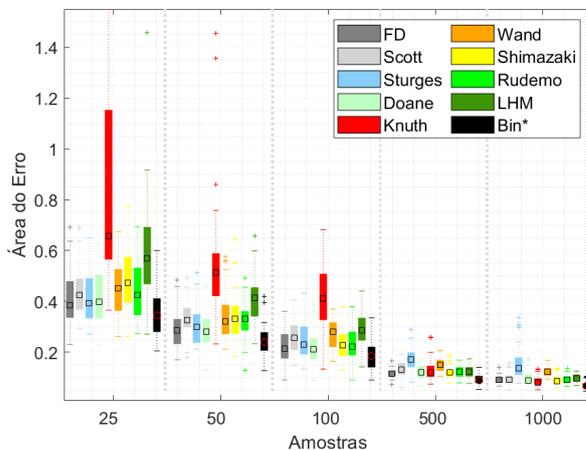
Fonte: Elaborada pelo autor (2020).

Figura 140 – Variação da Binagem utilizando PF na distribuição D1b



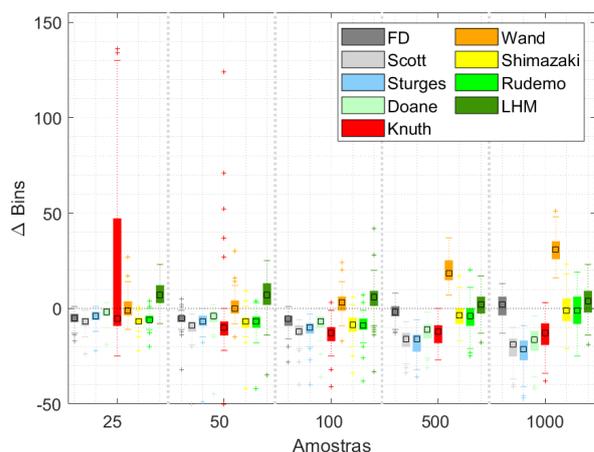
Fonte: Elaborada pelo autor (2020).

Figura 141 – Área do erro utilizando PF na distribuição D1b



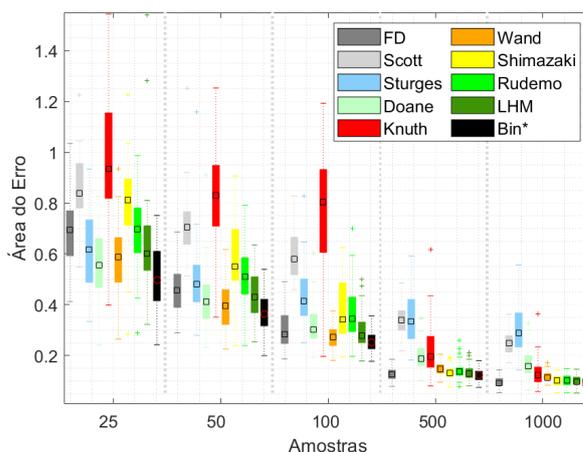
Fonte: Elaborada pelo autor (2020).

Figura 142 – Variação da Binagem utilizando ASH na distribuição D1b



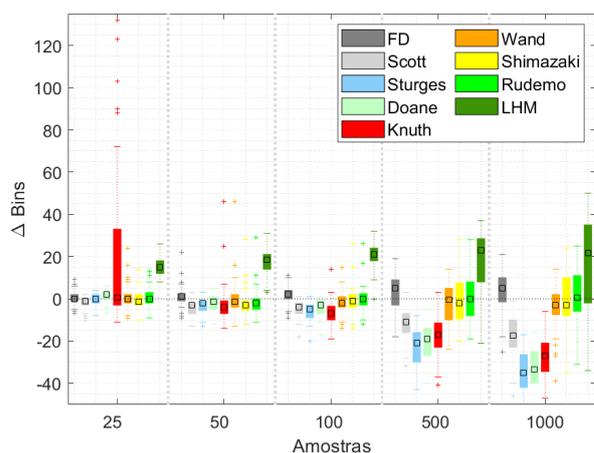
Fonte: Elaborada pelo autor (2020).

Figura 143 – Área do erro utilizando ASH na distribuição D1b



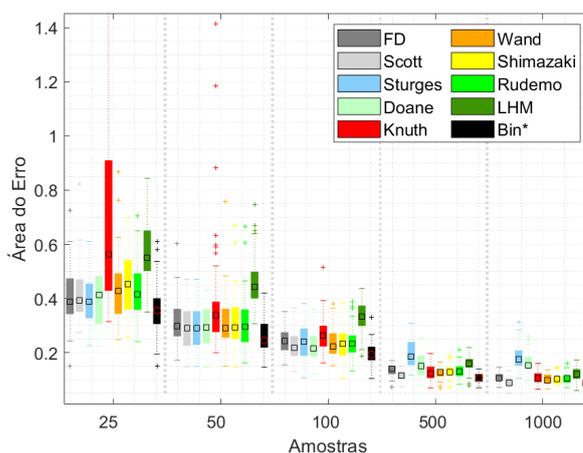
Fonte: Elaborada pelo autor (2020).

Figura 144 – Variação da Binagem utilizando Histograma na distribuição D1c



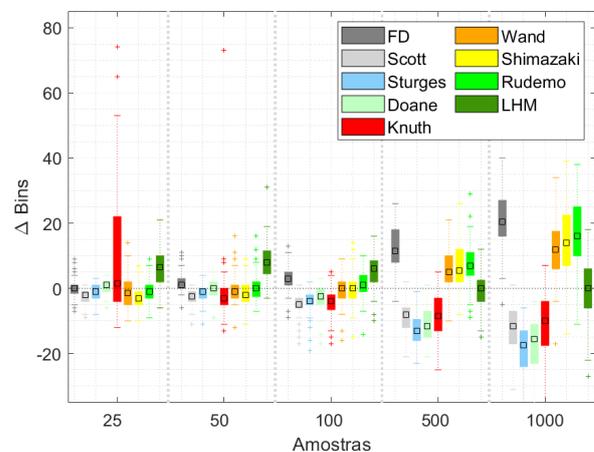
Fonte: Elaborada pelo autor (2020).

Figura 145 – Área do erro utilizando Histograma na distribuição D1c



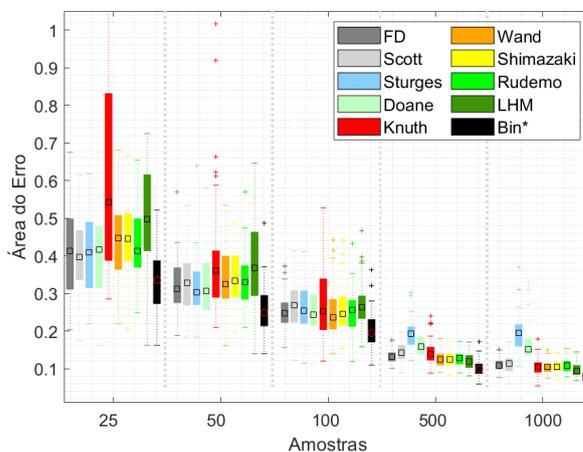
Fonte: Elaborada pelo autor (2020).

Figura 146 – Variação da Binagem utilizando PF na distribuição D1c



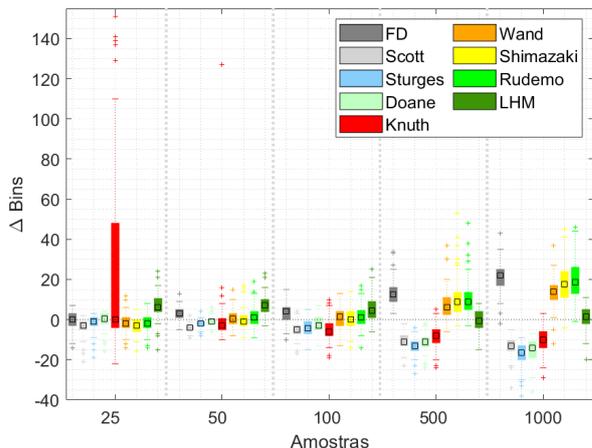
Fonte: Elaborada pelo autor (2020).

Figura 147 – Área do erro utilizando PF na distribuição D1c



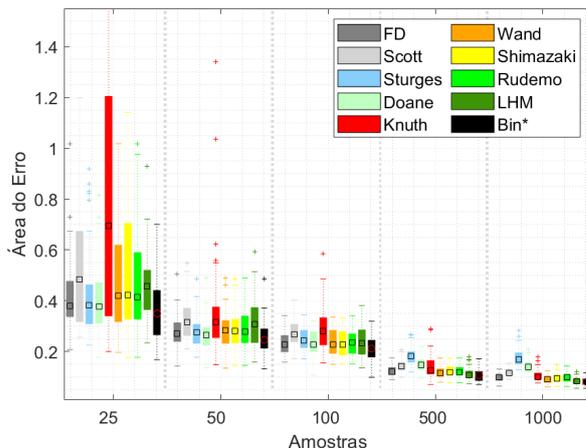
Fonte: Elaborada pelo autor (2020).

Figura 148 – Variação da Binagem utilizando ASH na distribuição D1c



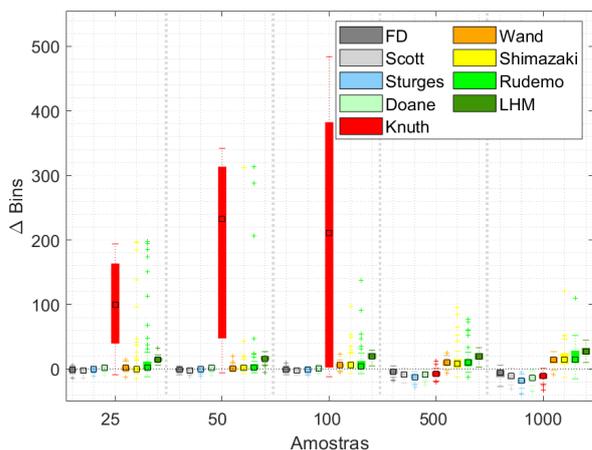
Fonte: Elaborada pelo autor (2020).

Figura 149 – Área do erro utilizando ASH na distribuição D1c



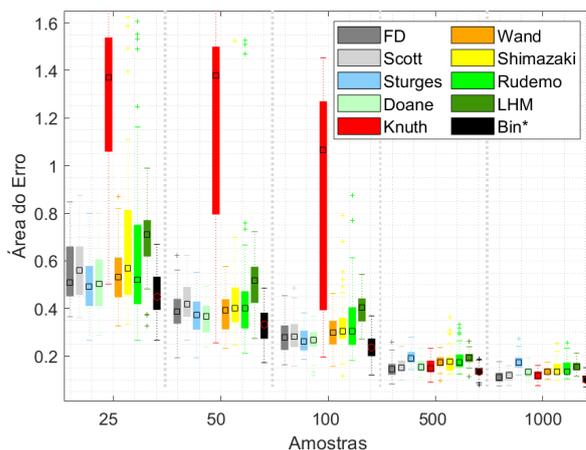
Fonte: Elaborada pelo autor (2020).

Figura 150 – Variação da Binagem utilizando Histograma na distribuição D2a



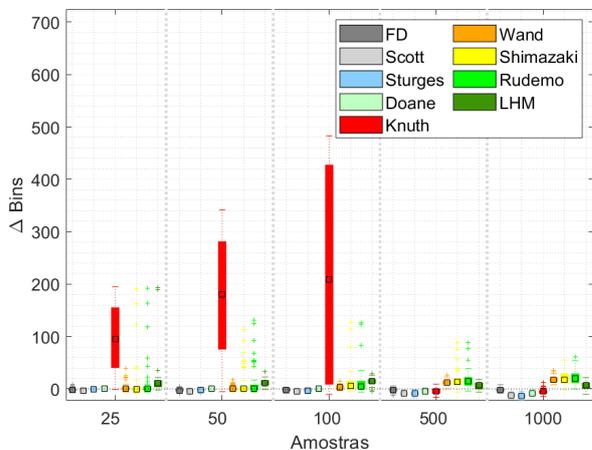
Fonte: Elaborada pelo autor (2020).

Figura 151 – Área do erro utilizando Histograma na distribuição D2a



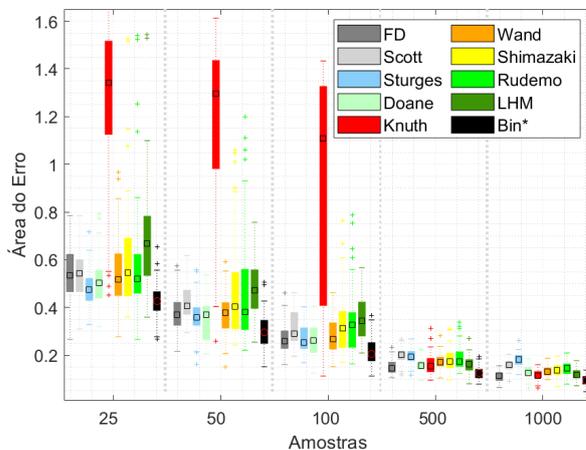
Fonte: Elaborada pelo autor (2020).

Figura 152 – Variação da Binagem utilizando PF na distribuição D2a



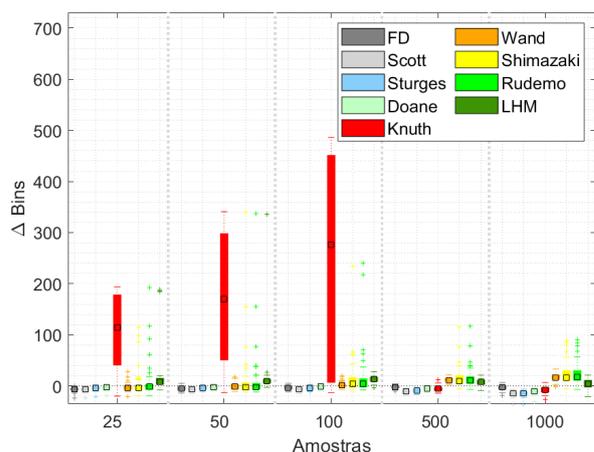
Fonte: Elaborada pelo autor (2020).

Figura 153 – Área do erro utilizando PF na distribuição D2a



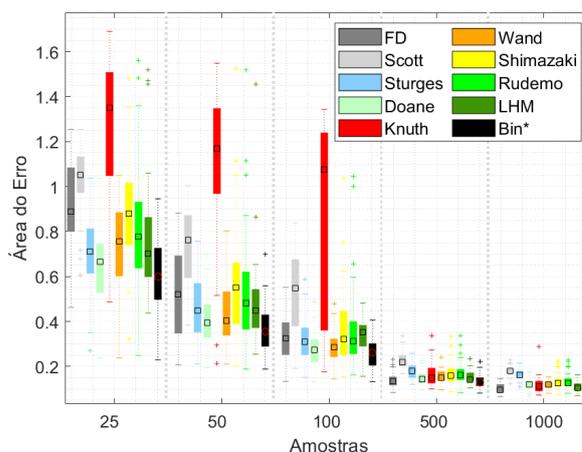
Fonte: Elaborada pelo autor (2020).

Figura 154 – Variação da Binagem utilizando ASH na distribuição D2a



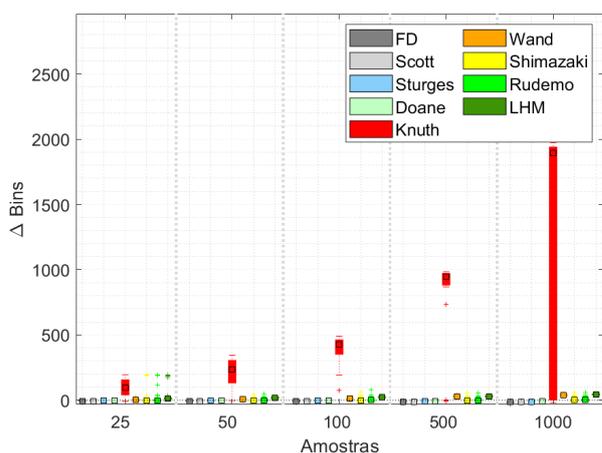
Fonte: Elaborada pelo autor (2020).

Figura 155 – Área do erro utilizando ASH na distribuição D2a



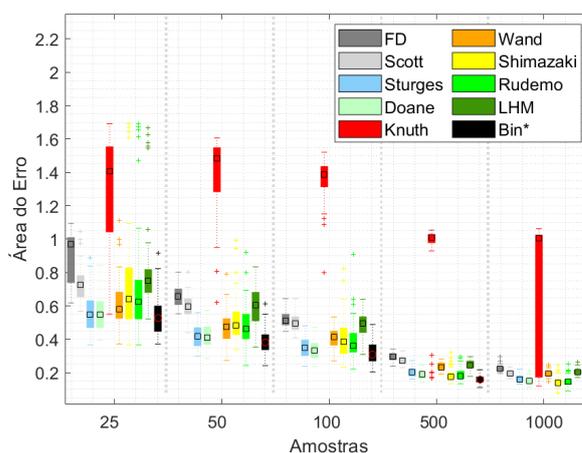
Fonte: Elaborada pelo autor (2020).

Figura 156 – Variação da Binagem utilizando Histograma na distribuição D2b



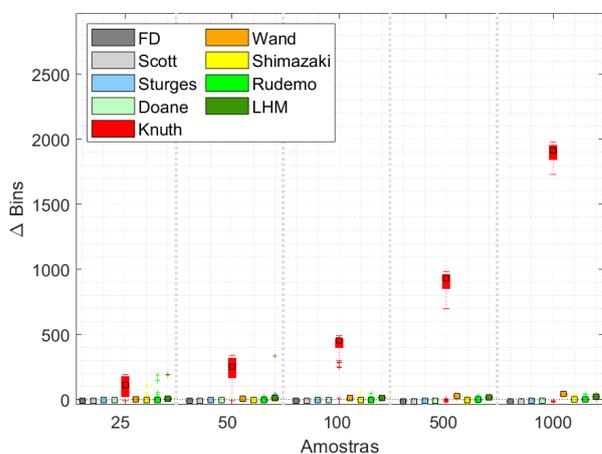
Fonte: Elaborada pelo autor (2020).

Figura 157 – Área do erro utilizando Histograma na distribuição D2b



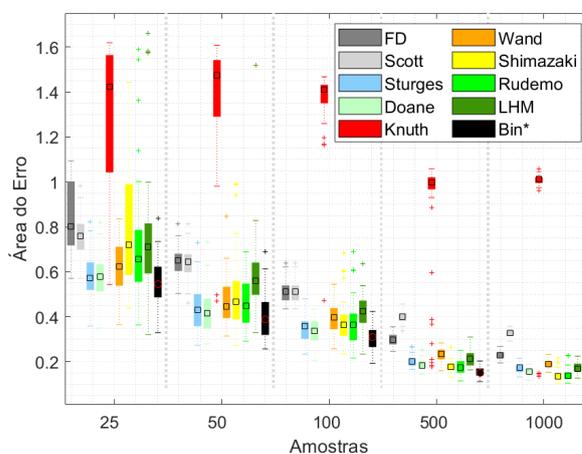
Fonte: Elaborada pelo autor (2020).

Figura 158 – Variação da Binagem utilizando PF na distribuição D2b



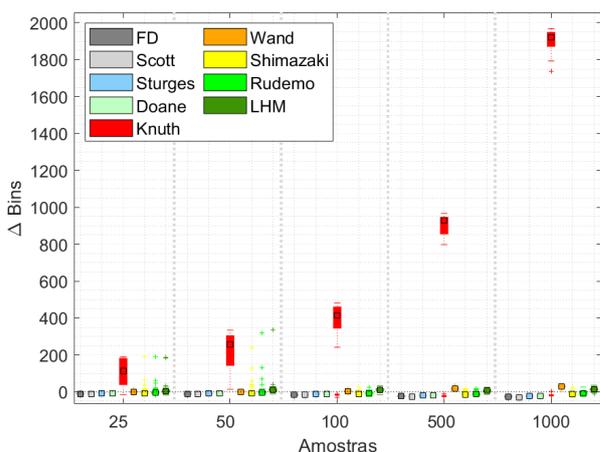
Fonte: Elaborada pelo autor (2020).

Figura 159 – Área do erro utilizando PF na distribuição D2b



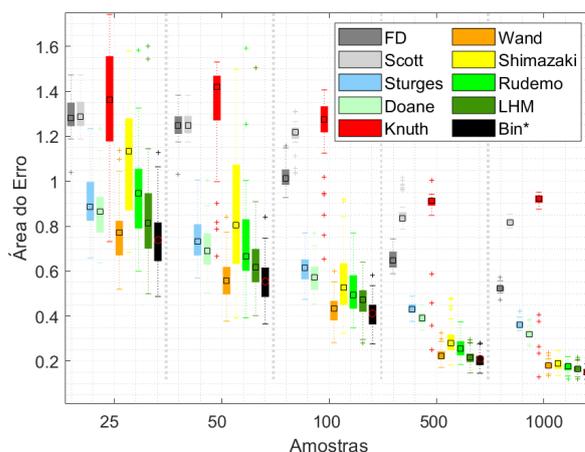
Fonte: Elaborada pelo autor (2020).

Figura 160 – Variação da Binagem utilizando ASH na distribuição D2b



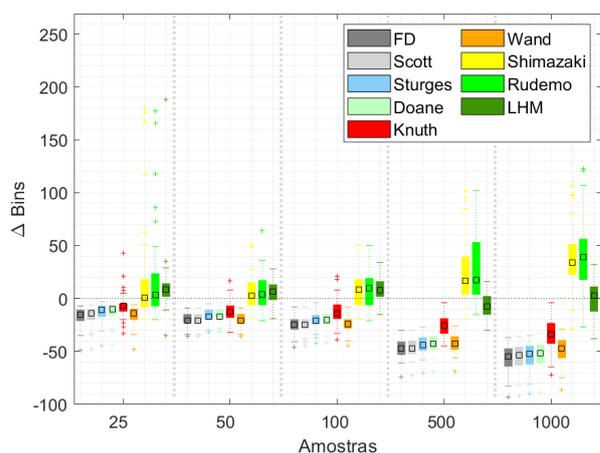
Fonte: Elaborada pelo autor (2020).

Figura 161 – Área do erro utilizando ASH na distribuição D2b



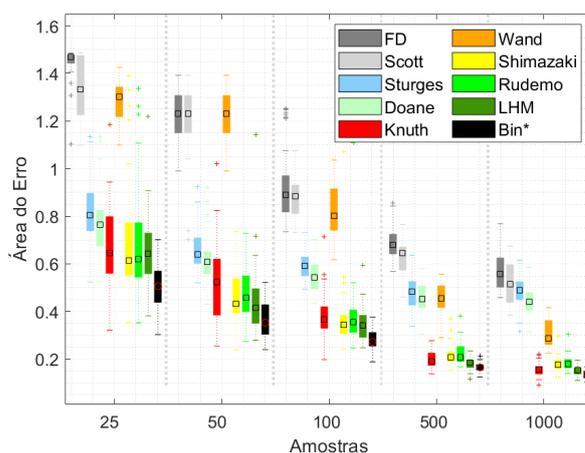
Fonte: Elaborada pelo autor (2020).

Figura 162 – Variação da Binagem utilizando Histograma na distribuição D2c



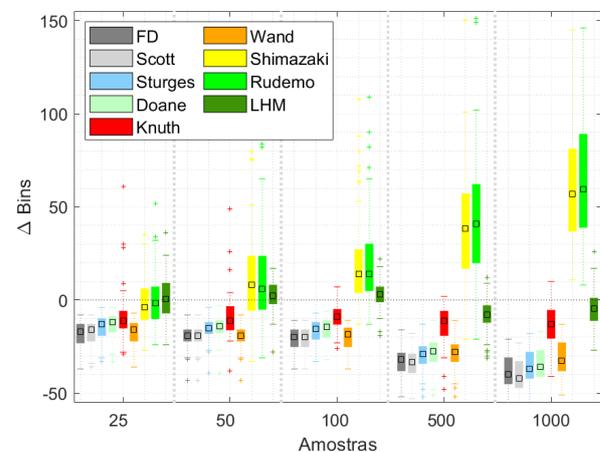
Fonte: Elaborada pelo autor (2020).

Figura 163 – Área do erro utilizando Histograma na distribuição D2c



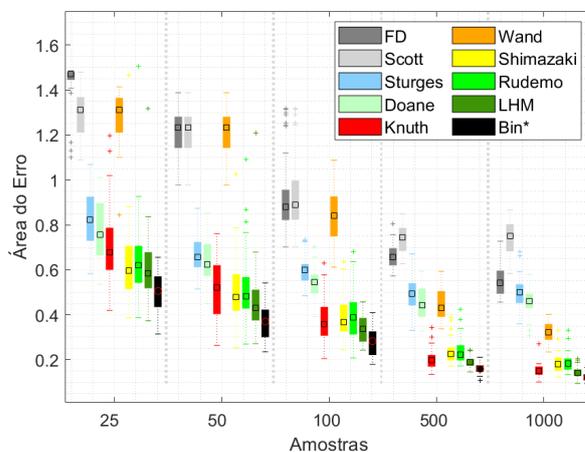
Fonte: Elaborada pelo autor (2020).

Figura 164 – Variação da Binagem utilizando PF na distribuição D2c



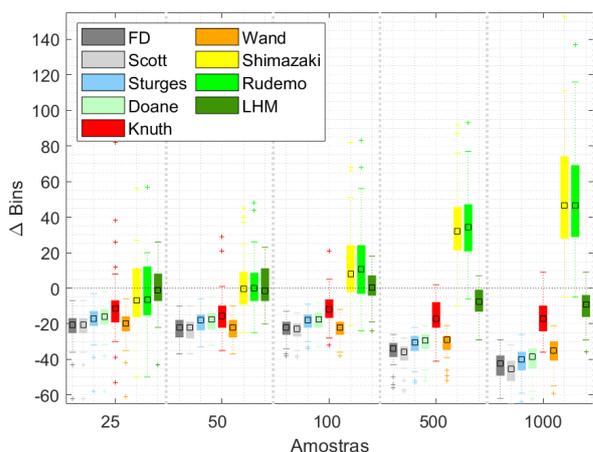
Fonte: Elaborada pelo autor (2020).

Figura 165 – Área do erro utilizando PF na distribuição D2c



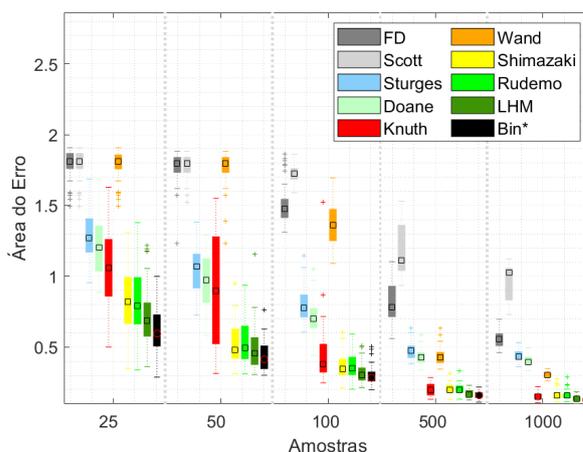
Fonte: Elaborada pelo autor (2020).

Figura 166 – Variação da Binagem utilizando ASH na distribuição D2c



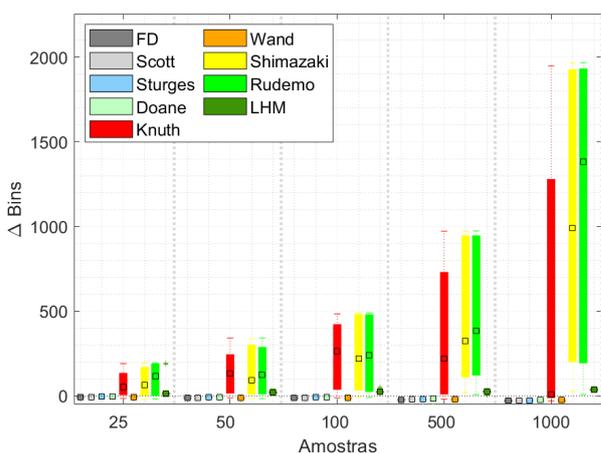
Fonte: Elaborada pelo autor (2020).

Figura 167 – Área do erro utilizando ASH na distribuição D2c



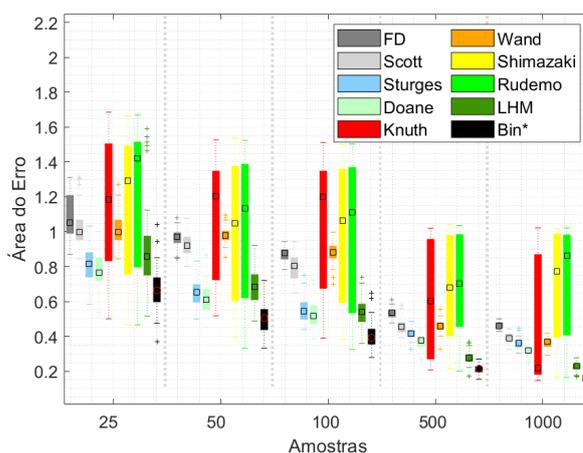
Fonte: Elaborada pelo autor (2020).

Figura 168 – Variação da Binagem utilizando Histograma na distribuição D3a



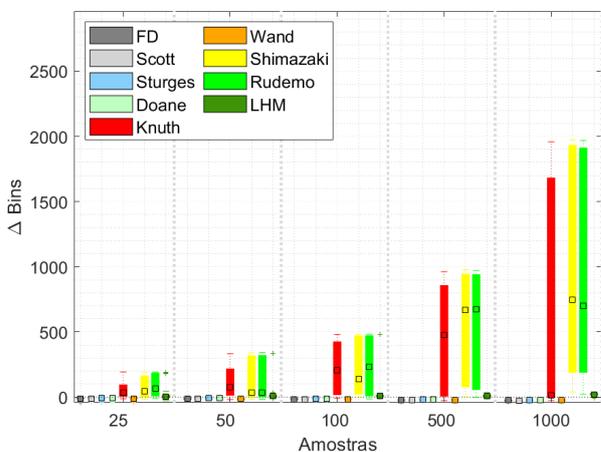
Fonte: Elaborada pelo autor (2020).

Figura 169 – Área do erro utilizando Histograma na distribuição D3a



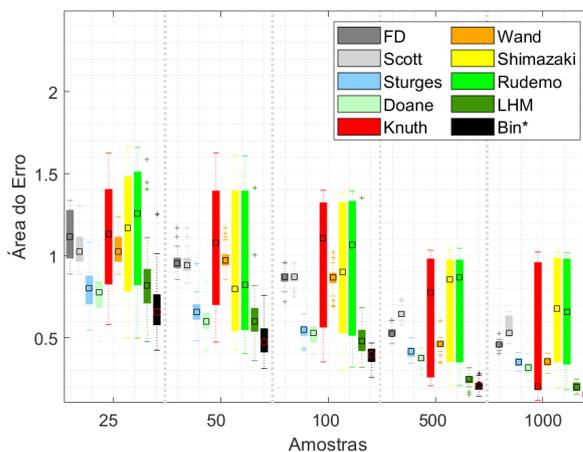
Fonte: Elaborada pelo autor (2020).

Figura 170 – Variação da Binagem utilizando PF na distribuição D3a



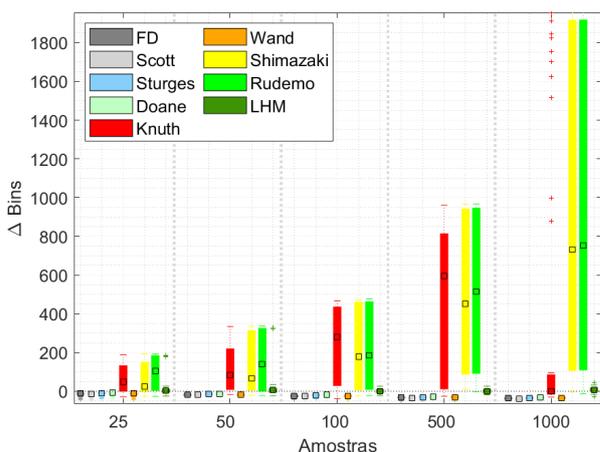
Fonte: Elaborada pelo autor (2020).

Figura 171 – Área do erro utilizando PF na distribuição D3a



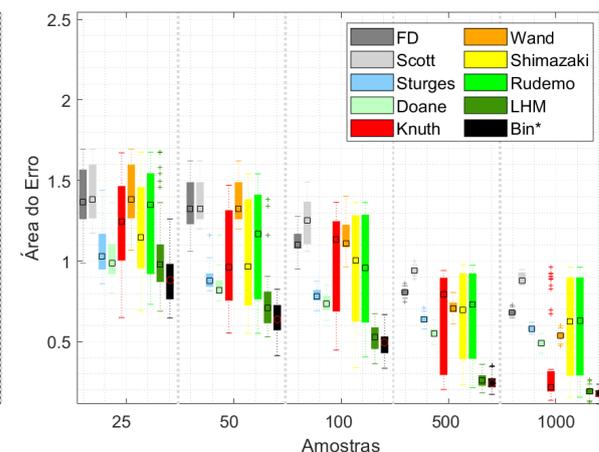
Fonte: Elaborada pelo autor (2020).

Figura 172 – Variação da Binagem utilizando ASH na distribuição D3a



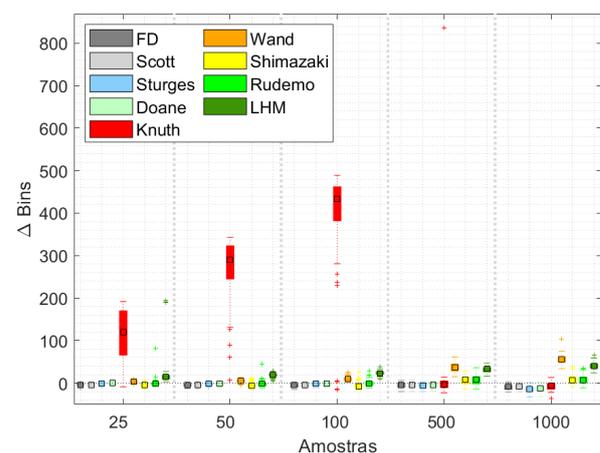
Fonte: Elaborada pelo autor (2020).

Figura 173 – Área do erro utilizando ASH na distribuição D3a



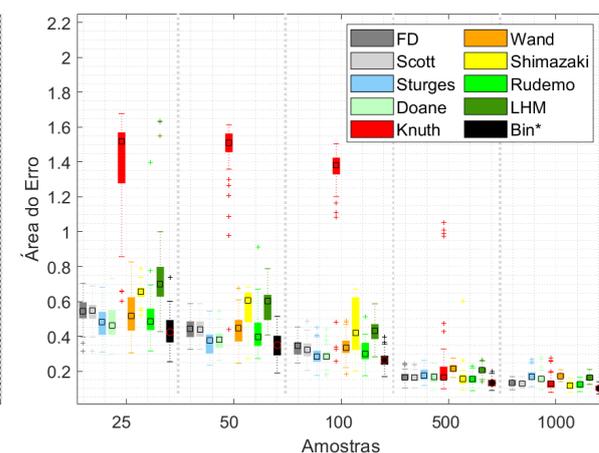
Fonte: Elaborada pelo autor (2020).

Figura 174 – Variação da Binagem utilizando Histograma na distribuição D3b



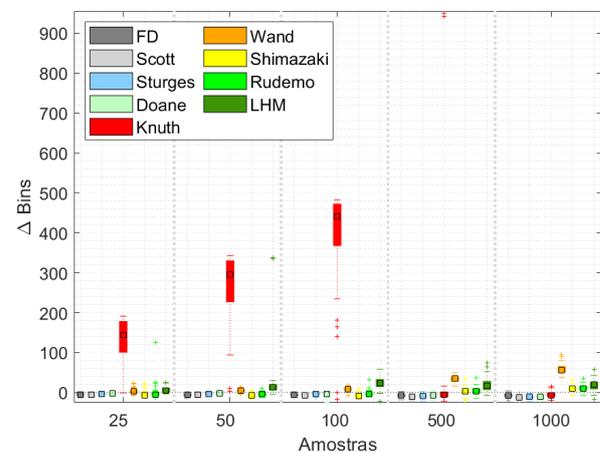
Fonte: Elaborada pelo autor (2020).

Figura 175 – Área do erro utilizando Histograma na distribuição D3b



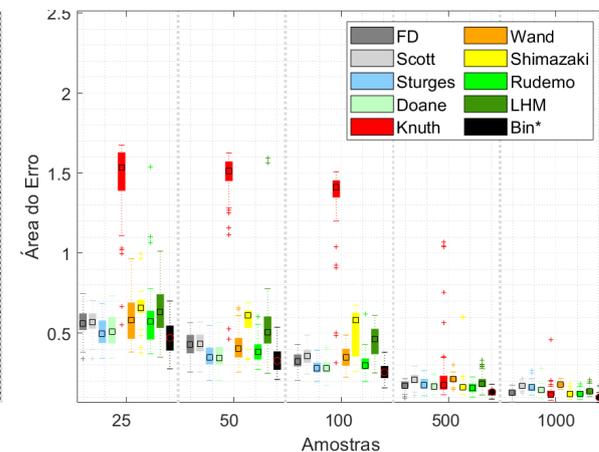
Fonte: Elaborada pelo autor (2020).

Figura 176 – Variação da Binagem utilizando PF na distribuição D3b



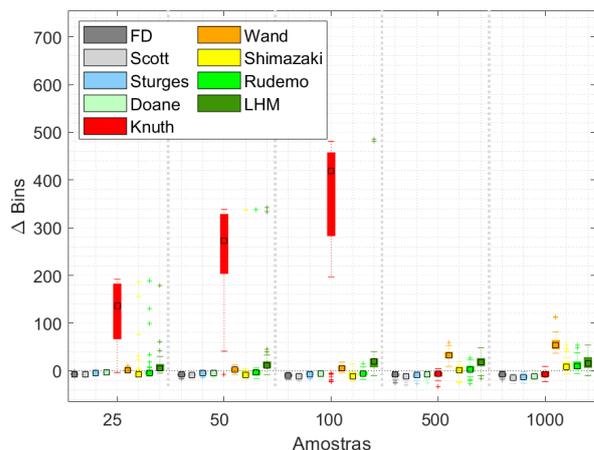
Fonte: Elaborada pelo autor (2020).

Figura 177 – Área do erro utilizando PF na distribuição D3b



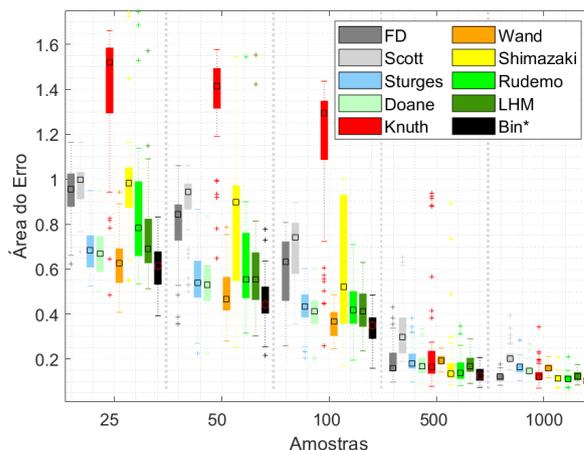
Fonte: Elaborada pelo autor (2020).

Figura 178 – Variação da Binagem utilizando ASH na distribuição D3b



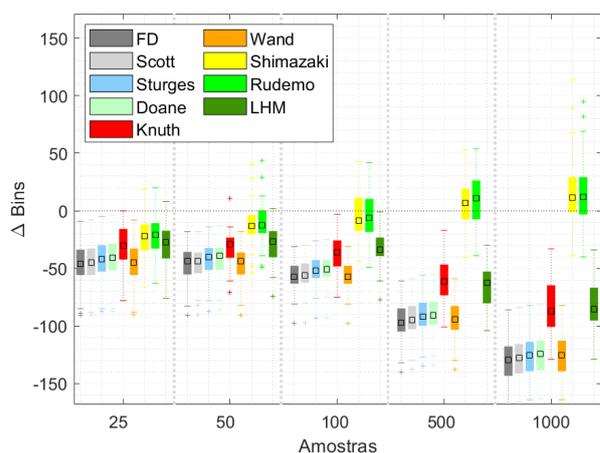
Fonte: Elaborada pelo autor (2020).

Figura 179 – Área do erro utilizando ASH na distribuição D3b



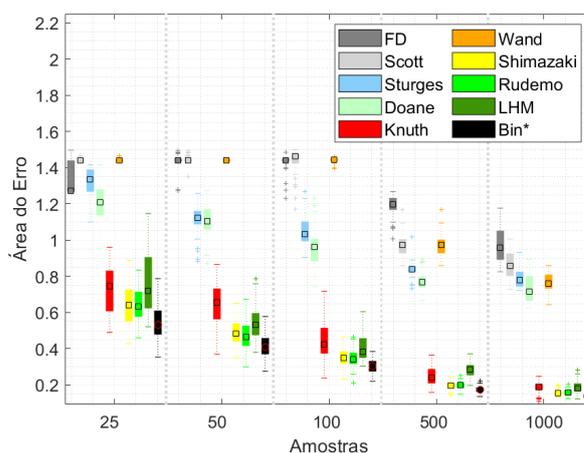
Fonte: Elaborada pelo autor (2020).

Figura 180 – Variação da Binagem utilizando Histograma na distribuição D3c



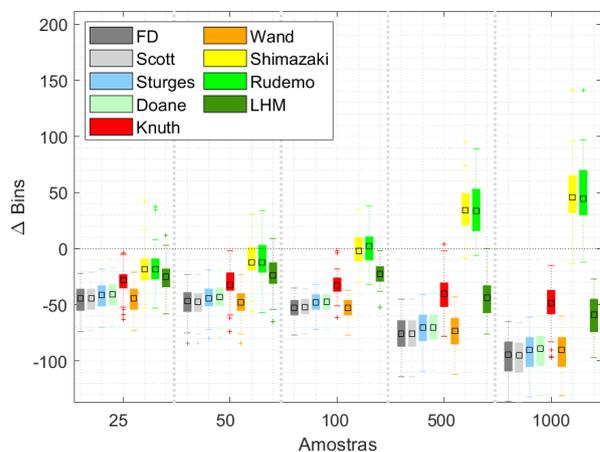
Fonte: Elaborada pelo autor (2020).

Figura 181 – Área do erro utilizando Histograma na distribuição D3c



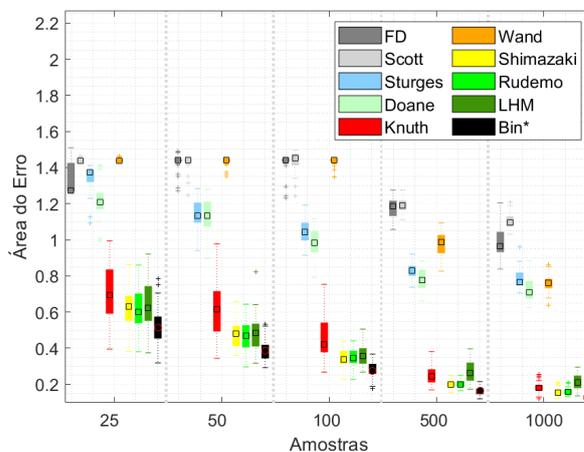
Fonte: Elaborada pelo autor (2020).

Figura 182 – Variação da Binagem utilizando PF na distribuição D3c



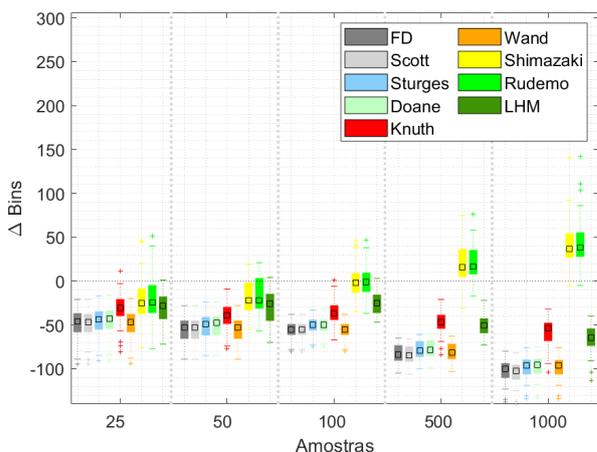
Fonte: Elaborada pelo autor (2020).

Figura 183 – Área do erro utilizando PF na distribuição D3c



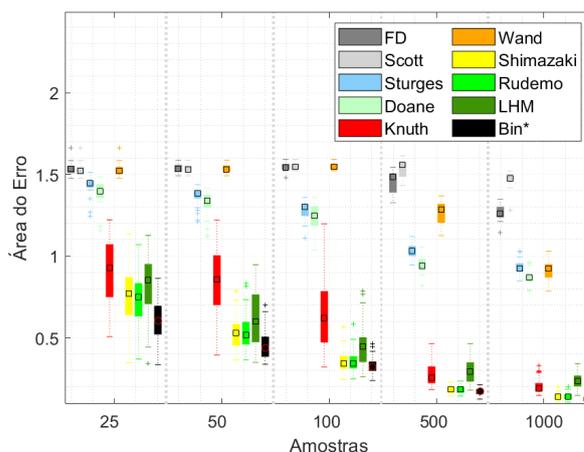
Fonte: Elaborada pelo autor (2020).

Figura 184 – Variação da Binagem utilizando ASH na distribuição D3c



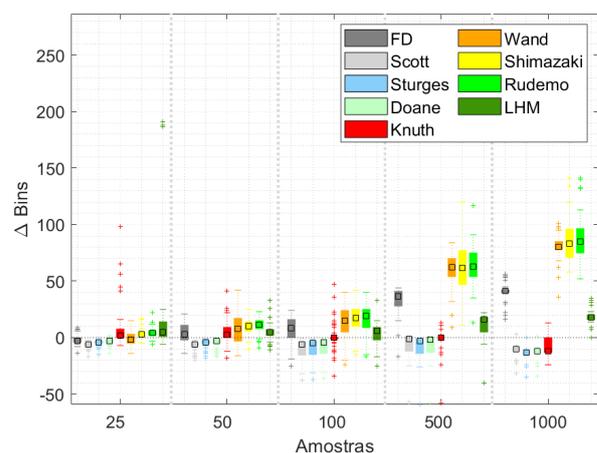
Fonte: Elaborada pelo autor (2020).

Figura 185 – Área do erro utilizando ASH na distribuição D3c



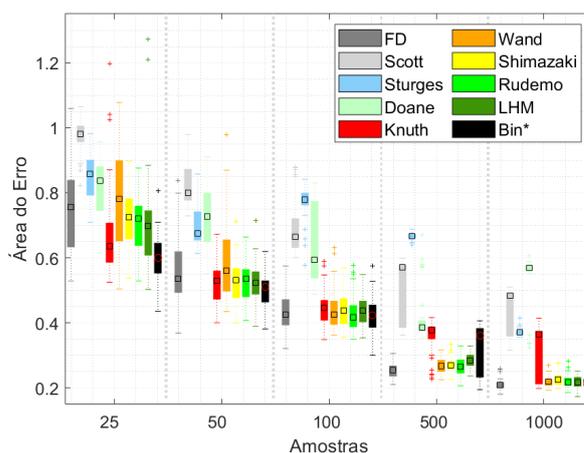
Fonte: Elaborada pelo autor (2020).

Figura 186 – Variação da Binagem utilizando Histograma na distribuição D4a



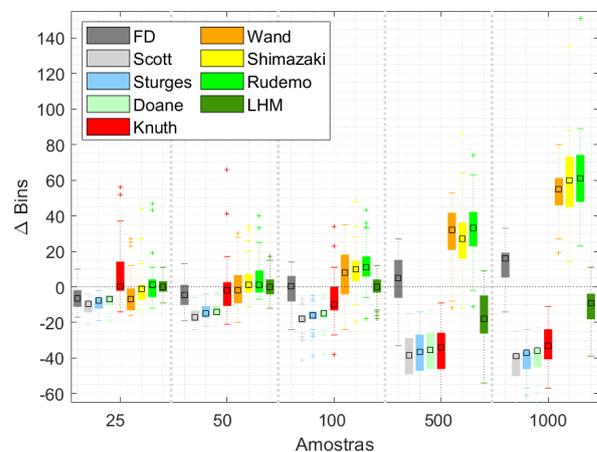
Fonte: Elaborada pelo autor (2020).

Figura 187 – Área do erro utilizando Histograma na distribuição D4a



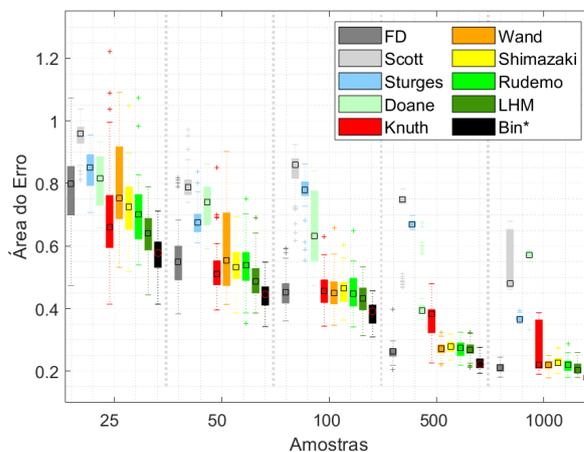
Fonte: Elaborada pelo autor (2020).

Figura 188 – Variação da Binagem utilizando PF na distribuição D4a



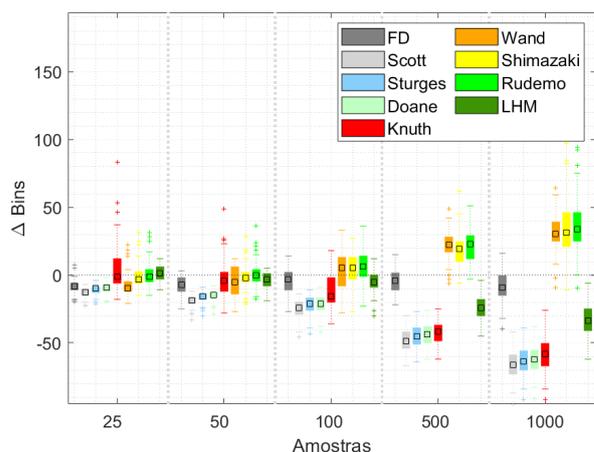
Fonte: Elaborada pelo autor (2020).

Figura 189 – Área do erro utilizando PF na distribuição D4a



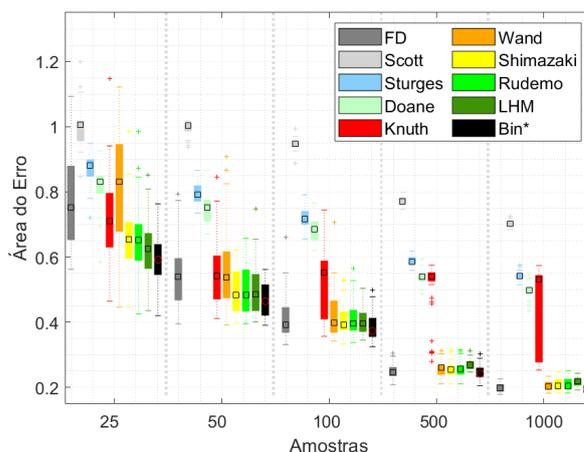
Fonte: Elaborada pelo autor (2020).

Figura 190 – Variação da Binagem utilizando ASH na distribuição D4a



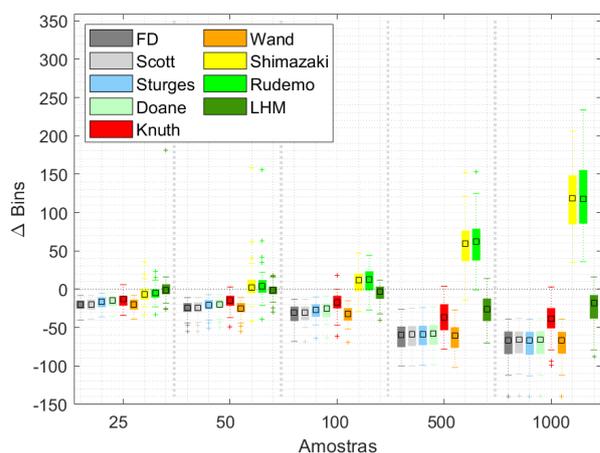
Fonte: Elaborada pelo autor (2020).

Figura 191 – Área do erro utilizando ASH na distribuição D4a



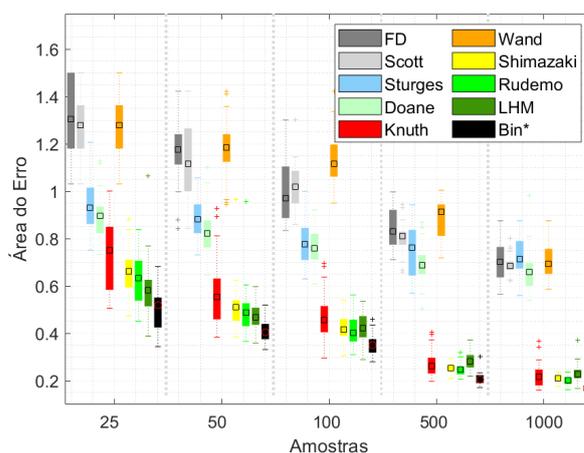
Fonte: Elaborada pelo autor (2020).

Figura 192 – Variação da Binagem utilizando Histograma na distribuição D4b



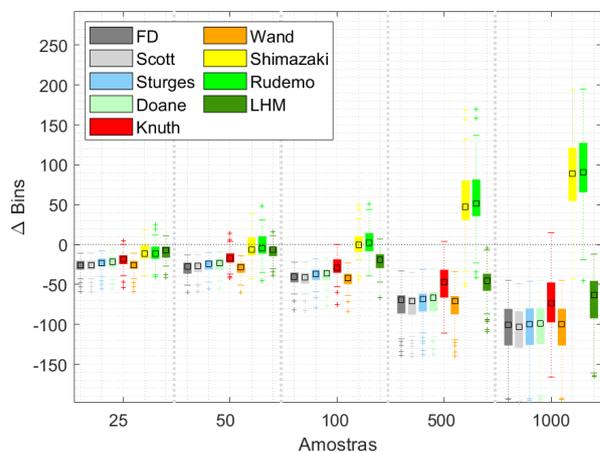
Fonte: Elaborada pelo autor (2020).

Figura 193 – Área do erro utilizando Histograma na distribuição D4b



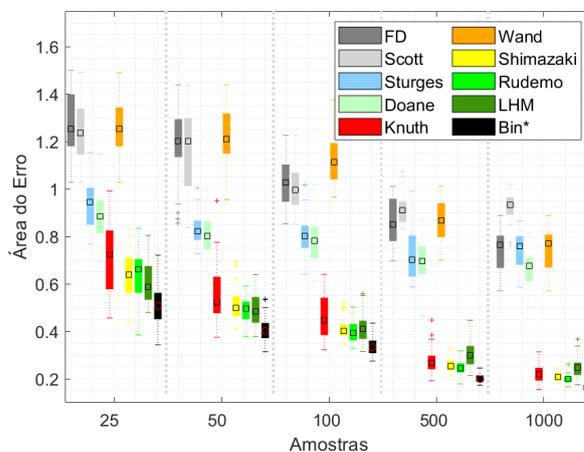
Fonte: Elaborada pelo autor (2020).

Figura 194 – Variação da Binagem utilizando PF na distribuição D4b



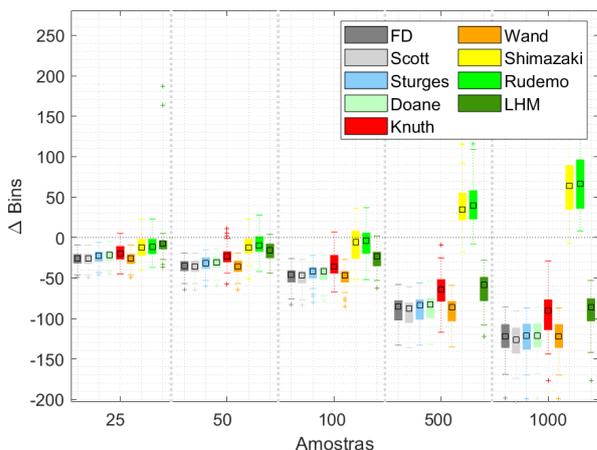
Fonte: Elaborada pelo autor (2020).

Figura 195 – Área do erro utilizando PF na distribuição D4b



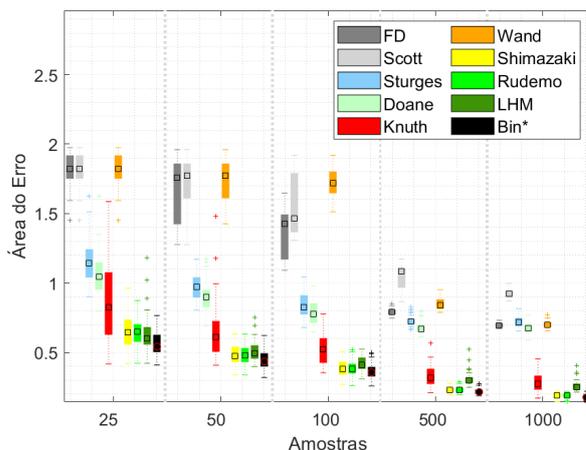
Fonte: Elaborada pelo autor (2020).

Figura 196 – Variação da Binagem utilizando ASH na distribuição D4b



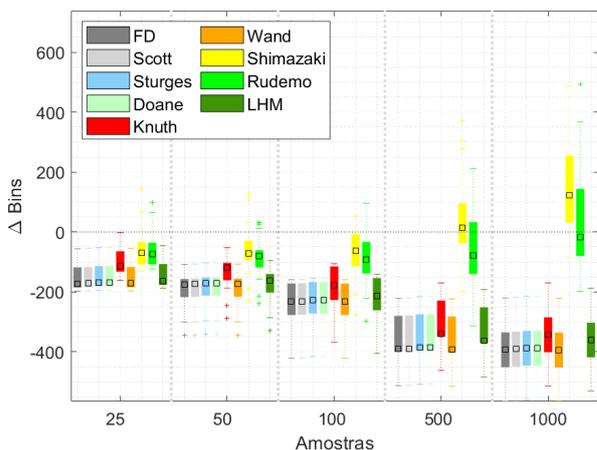
Fonte: Elaborada pelo autor (2020).

Figura 197 – Área do erro utilizando ASH na distribuição D4b



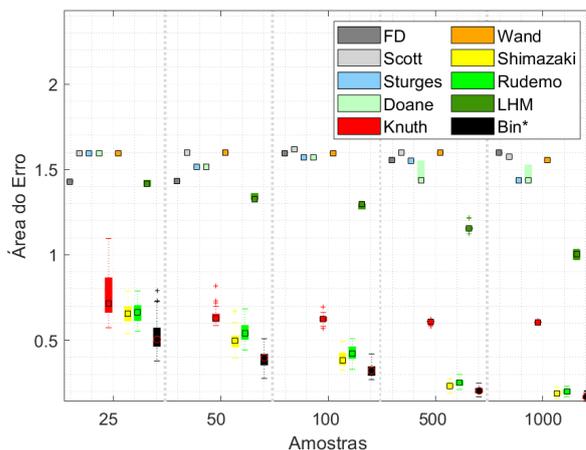
Fonte: Elaborada pelo autor (2020).

Figura 198 – Variação da Binagem utilizando Histograma na distribuição D4c



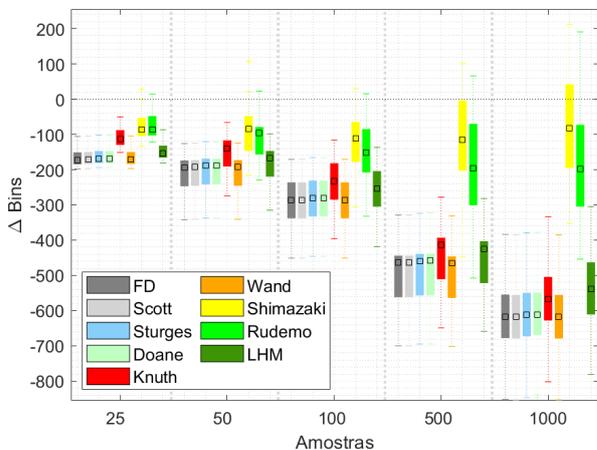
Fonte: Elaborada pelo autor (2020).

Figura 199 – Área do erro utilizando Histograma na distribuição D4c



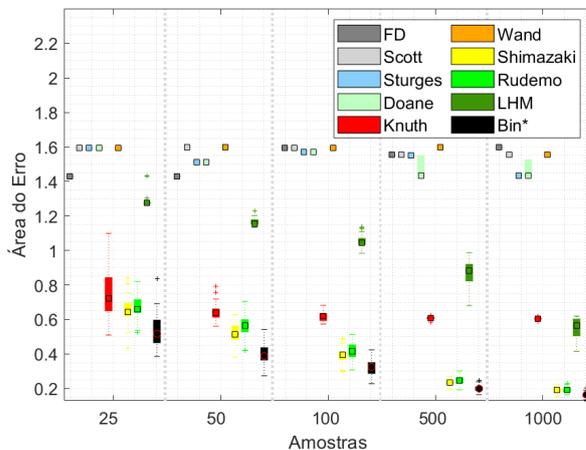
Fonte: Elaborada pelo autor (2020).

Figura 200 – Variação da Binagem utilizando PF na distribuição D4c



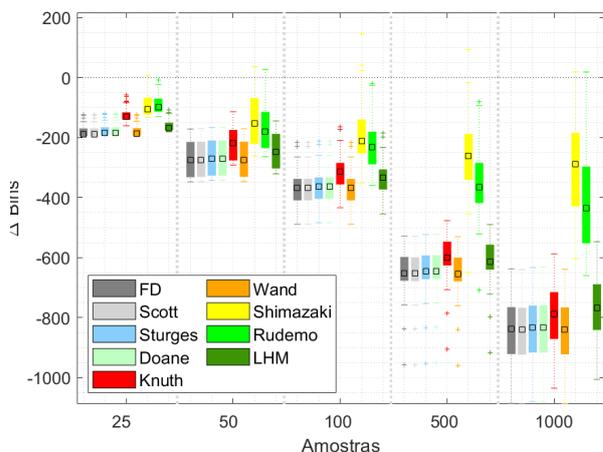
Fonte: Elaborada pelo autor (2020).

Figura 201 – Área do erro utilizando PF na distribuição D4c



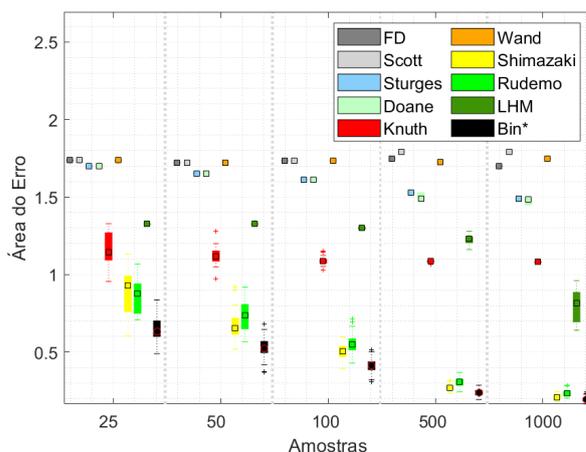
Fonte: Elaborada pelo autor (2020).

Figura 202 – Variação da Binagem utilizando ASH na distribuição D4c



Fonte: Elaborada pelo autor (2020).

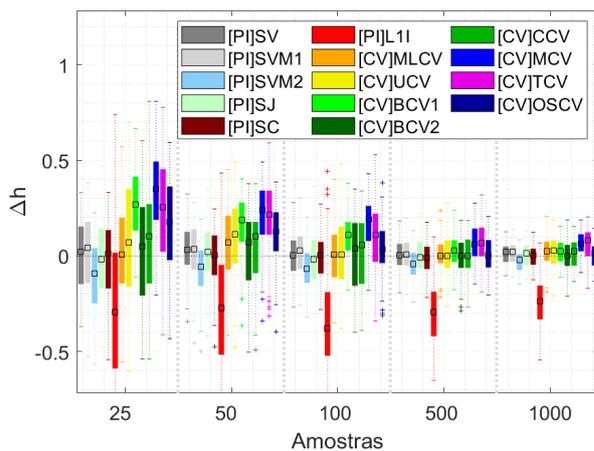
Figura 203 – Área do erro utilizando ASH na distribuição D4c



Fonte: Elaborada pelo autor (2020).

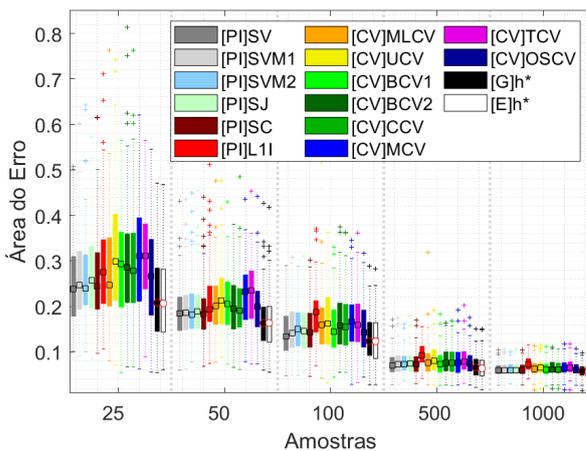
C.1.2 Kernel Density Estimation

Figura 204 – Variação de h utilizando KDE na distribuição D1a



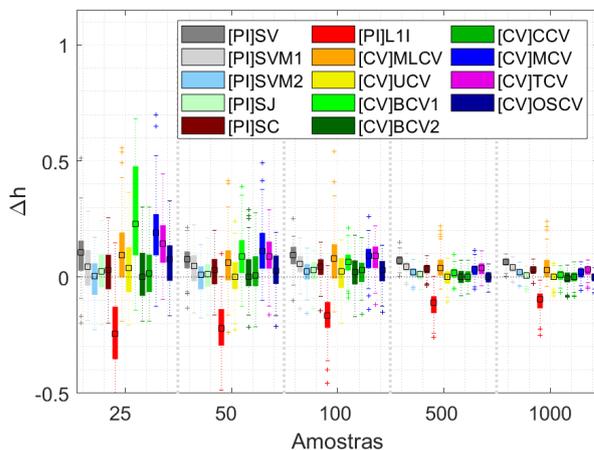
Fonte: Elaborada pelo autor (2020).

Figura 205 – Área do erro utilizando KDE na distribuição D1a



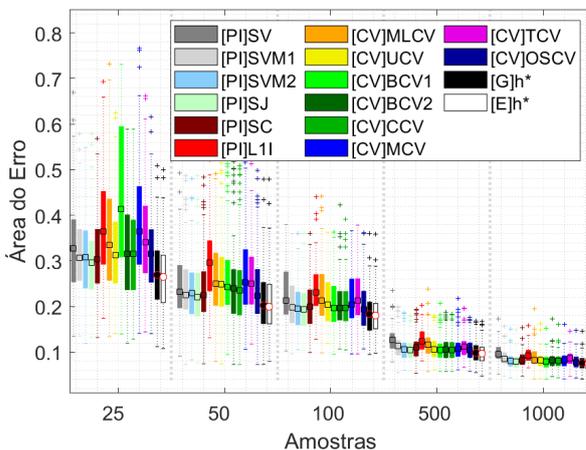
Fonte: Elaborada pelo autor (2020).

Figura 206 – Variação de h utilizando KDE na distribuição D1b



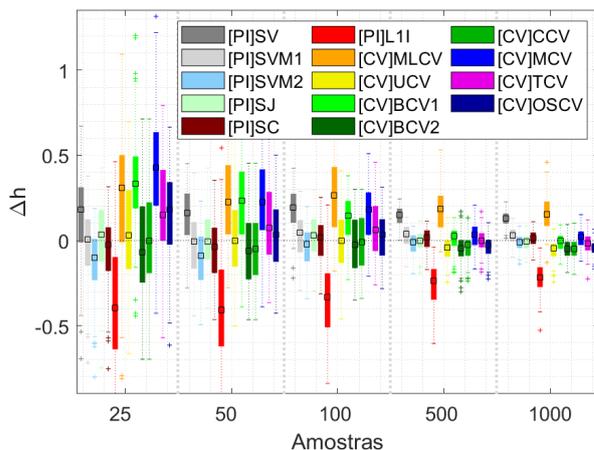
Fonte: Elaborada pelo autor (2020).

Figura 207 – Área do erro utilizando KDE na distribuição D1b



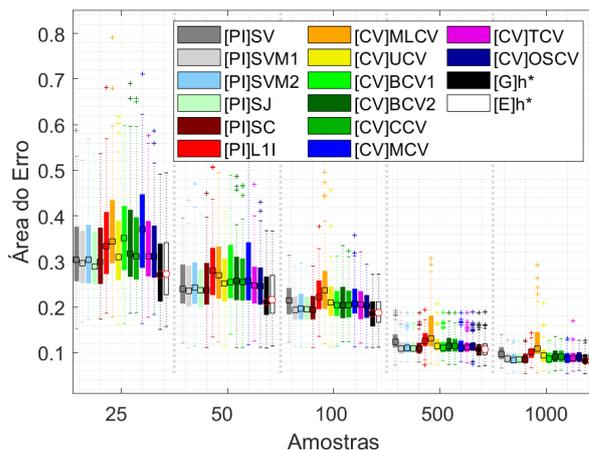
Fonte: Elaborada pelo autor (2020).

Figura 208 – Variação de h utilizando KDE na distribuição D1c



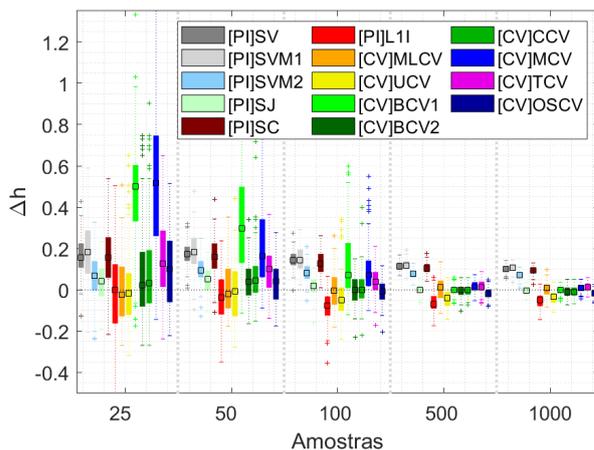
Fonte: Elaborada pelo autor (2020).

Figura 209 – Área do erro utilizando KDE na distribuição D1c



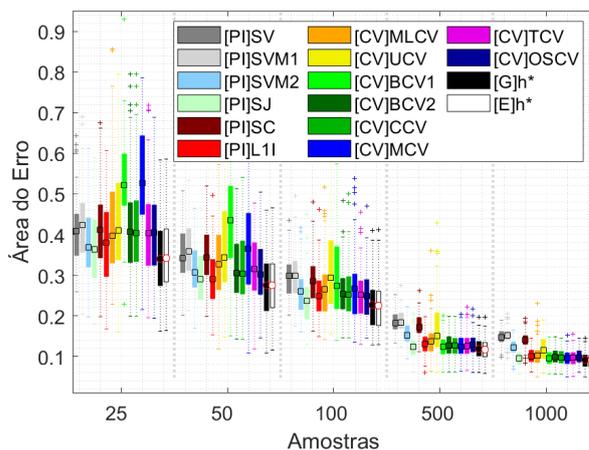
Fonte: Elaborada pelo autor (2020).

Figura 210 – Variação de h utilizando KDE na distribuição D2a



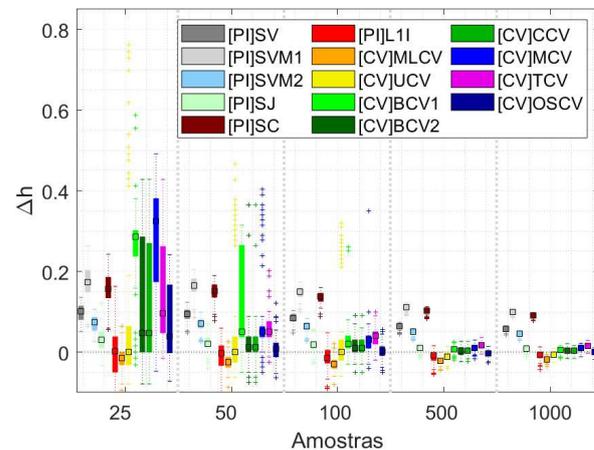
Fonte: Elaborada pelo autor (2020).

Figura 211 – Área do erro utilizando KDE na distribuição D2a



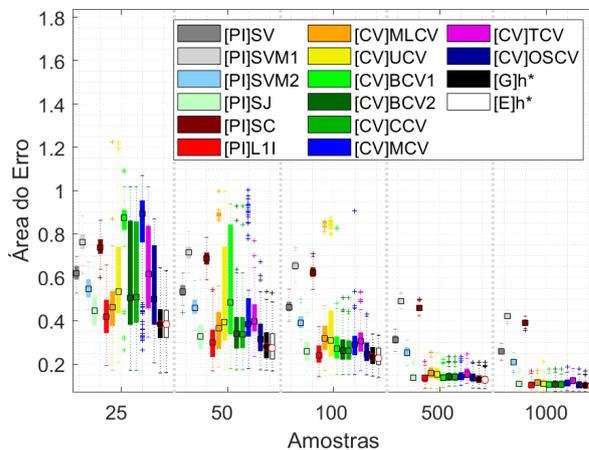
Fonte: Elaborada pelo autor (2020).

Figura 212 – Variação de h utilizando KDE na distribuição D2b



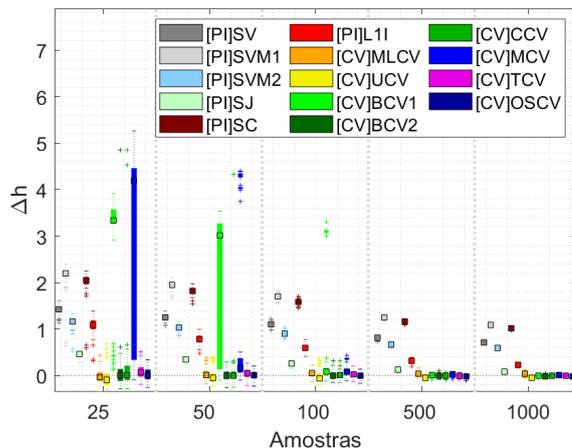
Fonte: Elaborada pelo autor (2020).

Figura 213 – Área do erro utilizando KDE na distribuição D2b



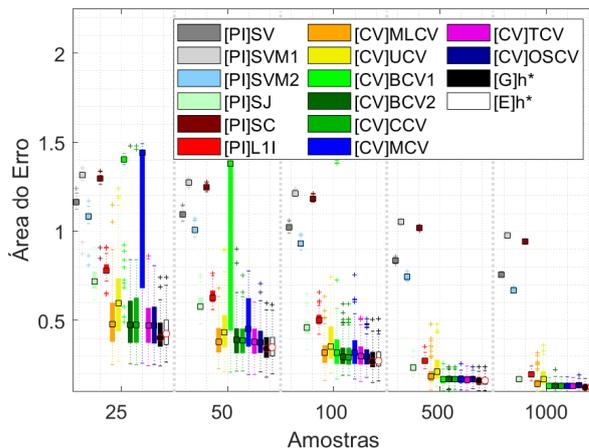
Fonte: Elaborada pelo autor (2020).

Figura 214 – Variação de h utilizando KDE na distribuição D2c



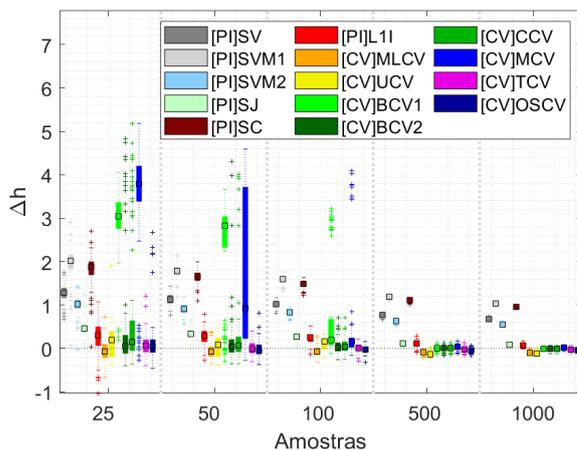
Fonte: Elaborada pelo autor (2020).

Figura 215 – Área do erro utilizando KDE na distribuição D2c



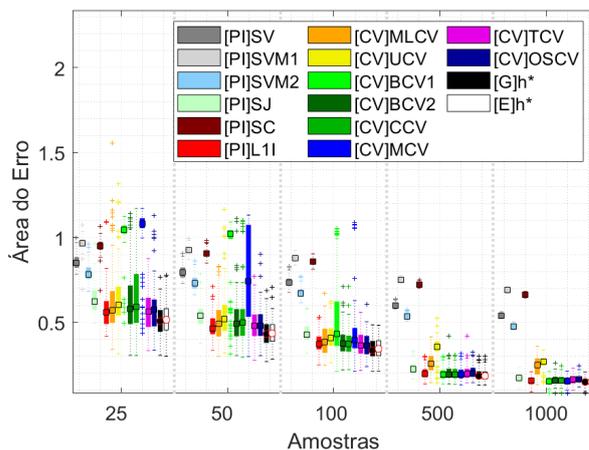
Fonte: Elaborada pelo autor (2020).

Figura 216 – Variação de h utilizando KDE na distribuição D3a



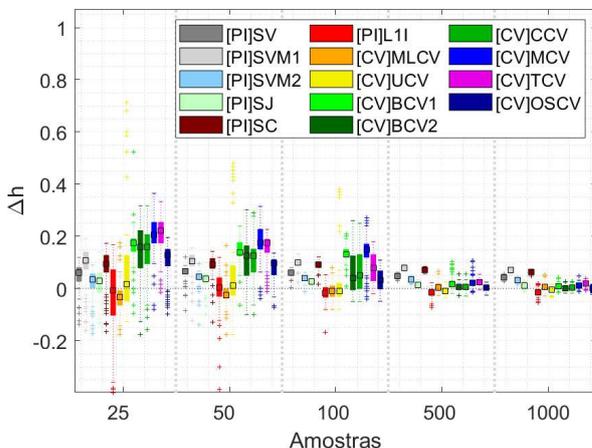
Fonte: Elaborada pelo autor (2020).

Figura 217 – Área do erro utilizando KDE na distribuição D3a



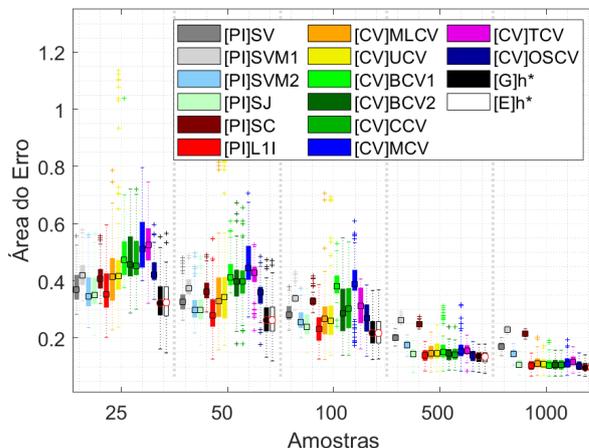
Fonte: Elaborada pelo autor (2020).

Figura 218 – Variação de h utilizando KDE na distribuição D3b



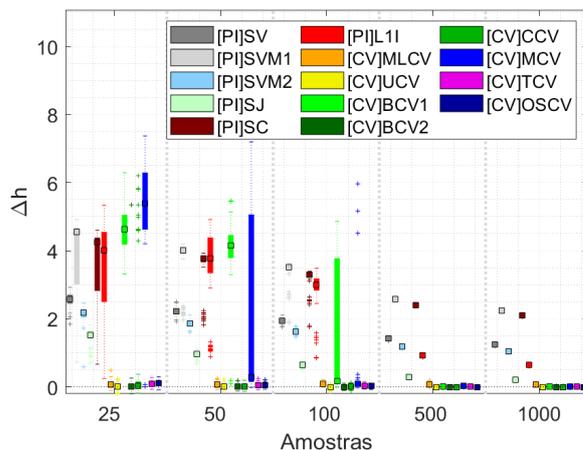
Fonte: Elaborada pelo autor (2020).

Figura 219 – Área do erro utilizando KDE na distribuição D3b



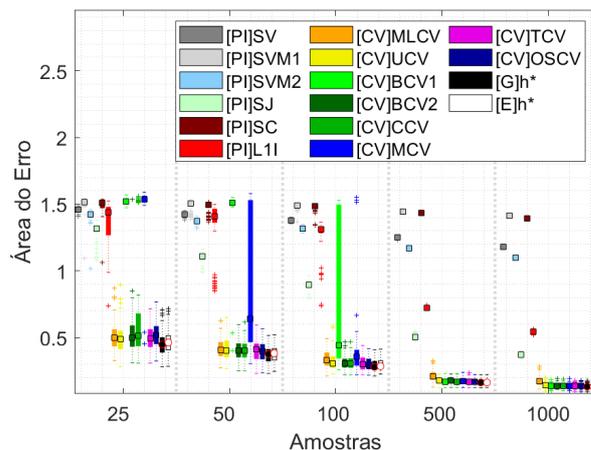
Fonte: Elaborada pelo autor (2020).

Figura 220 – Variação de h utilizando KDE na distribuição D3c



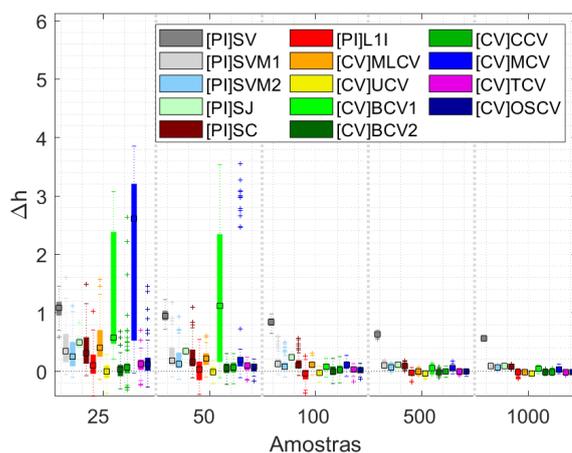
Fonte: Elaborada pelo autor (2020).

Figura 221 – Área do erro utilizando KDE na distribuição D3c



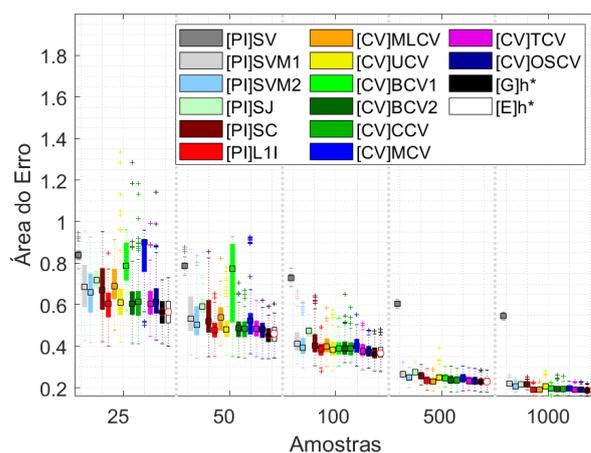
Fonte: Elaborada pelo autor (2020).

Figura 222 – Variação de h utilizando KDE na distribuição D4a



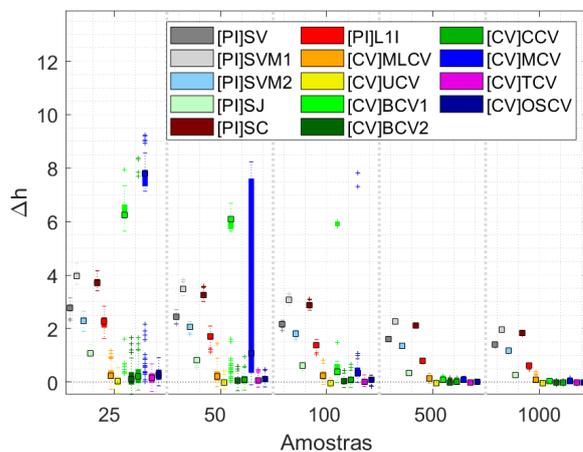
Fonte: Elaborada pelo autor (2020).

Figura 223 – Área do erro utilizando KDE na distribuição D4a



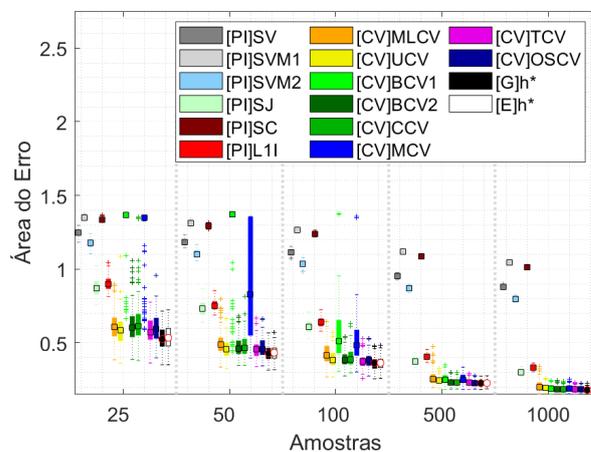
Fonte: Elaborada pelo autor (2020).

Figura 224 – Variação de h utilizando KDE na distribuição D4b



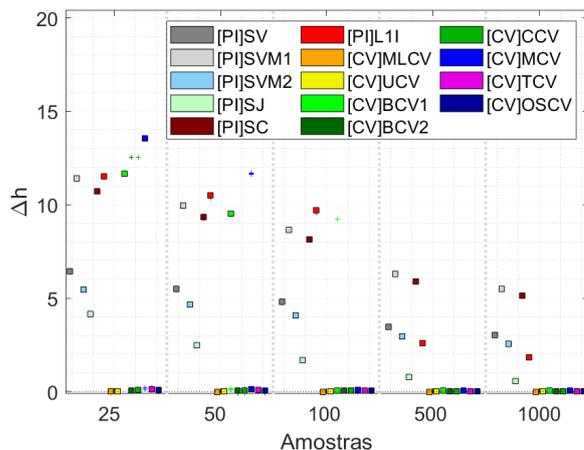
Fonte: Elaborada pelo autor (2020).

Figura 225 – Área do erro utilizando KDE na distribuição D4b



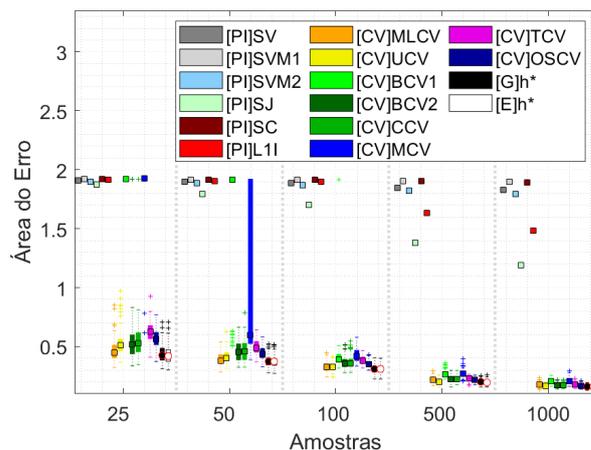
Fonte: Elaborada pelo autor (2020).

Figura 226 – Variação de h utilizando KDE na distribuição D4c



Fonte: Elaborada pelo autor (2020).

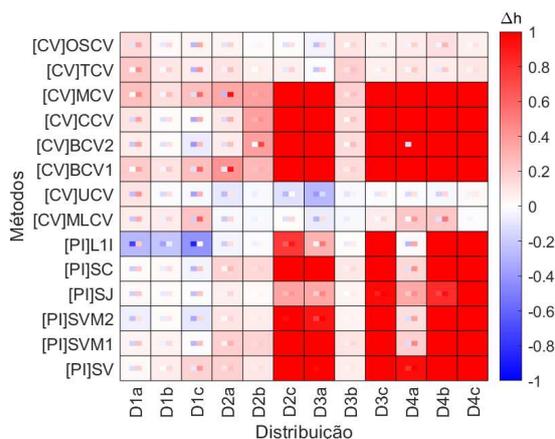
Figura 227 – Área do erro utilizando KDE na distribuição D4c



Fonte: Elaborada pelo autor (2020).

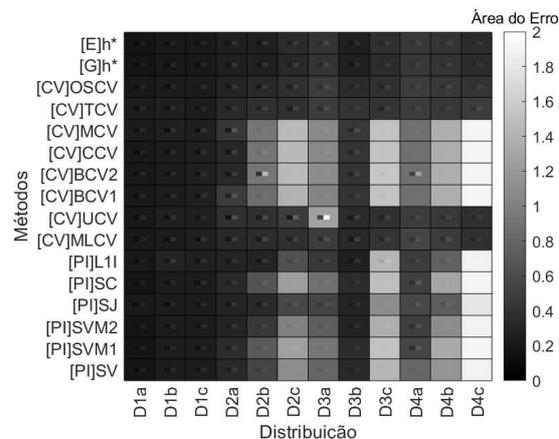
C.1.3 Matriz geral de comparação de banda fixa

Figura 228 – Variação de h para todas as distribuições com 50 amostras de treinamento no KDE



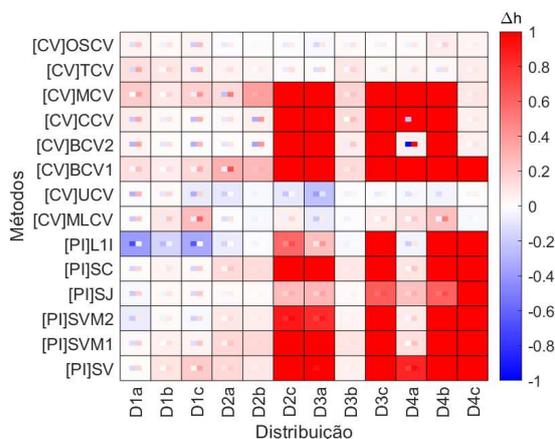
Fonte: Elaborada pelo autor (2020).

Figura 229 – Área do erro para todas as distribuições com 50 amostras de treinamento no KDE



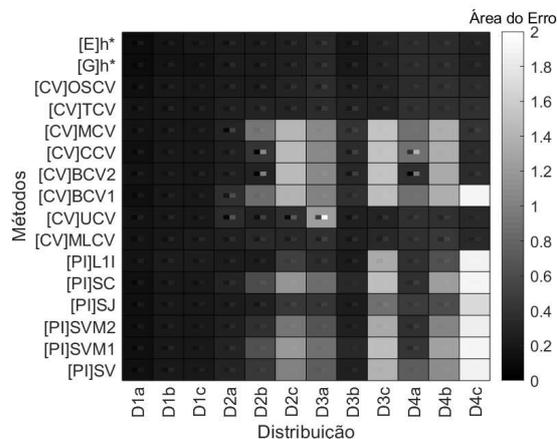
Fonte: Elaborada pelo autor (2020).

Figura 230 – Variação de h para todas as distribuições com 100 amostras de treinamento no KDE



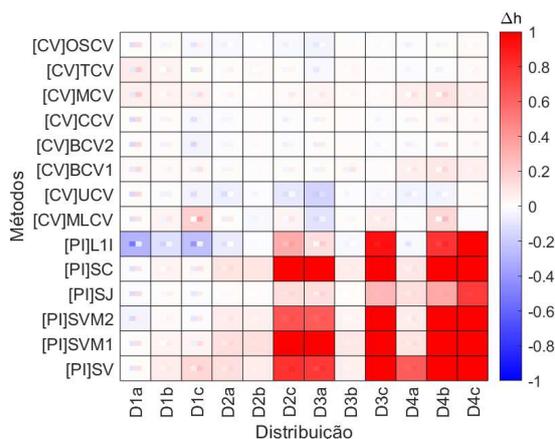
Fonte: Elaborada pelo autor (2020).

Figura 231 – Área do erro para todas as distribuições com 100 amostras de treinamento no KDE



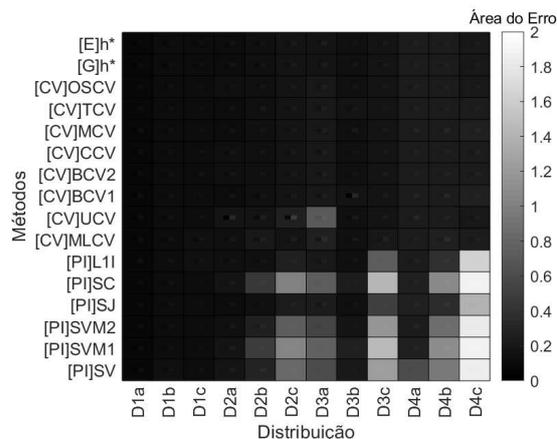
Fonte: Elaborada pelo autor (2020).

Figura 232 – Variação de h para todas as distribuições com 500 amostras de treinamento no KDE



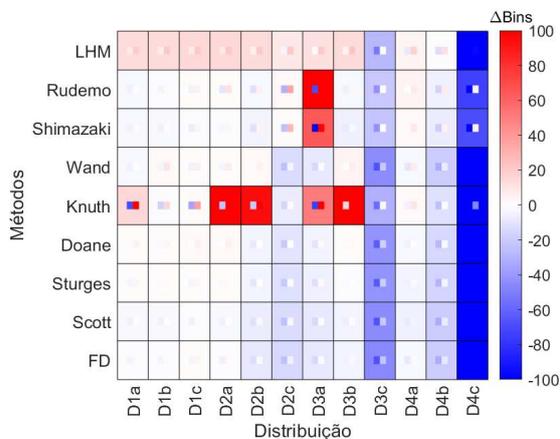
Fonte: Elaborada pelo autor (2020).

Figura 233 – Área do erro para todas as distribuições com 500 amostras de treinamento no KDE



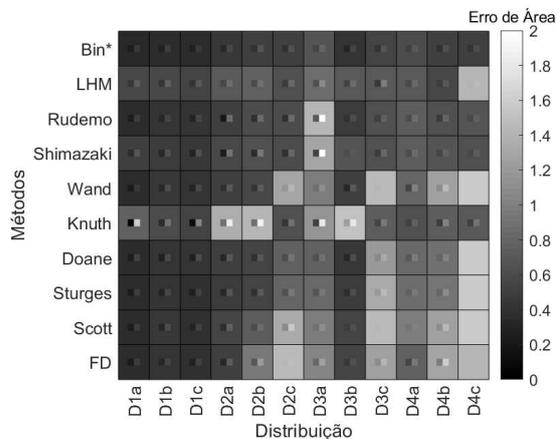
Fonte: Elaborada pelo autor (2020).

Figura 234 – Variação dos *bins* para todas as distribuições com 25 amostras de treinamento no Histograma



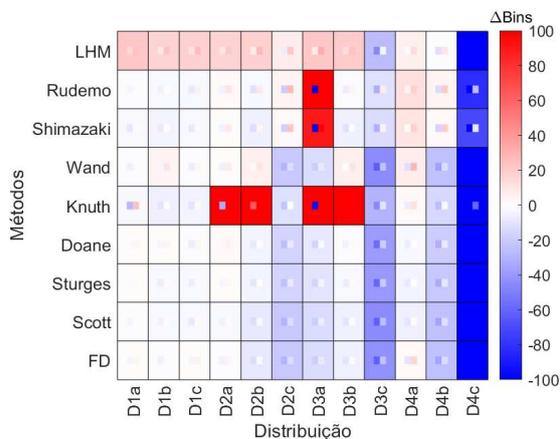
Fonte: Elaborada pelo autor (2020).

Figura 235 – Área do erro para todas as distribuições com 25 amostras de treinamento no Histograma



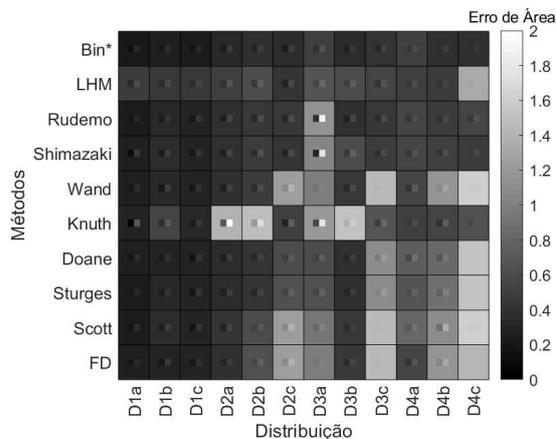
Fonte: Elaborada pelo autor (2020).

Figura 236 – Variação dos *bins* para todas as distribuições com 50 amostras de treinamento no Histograma



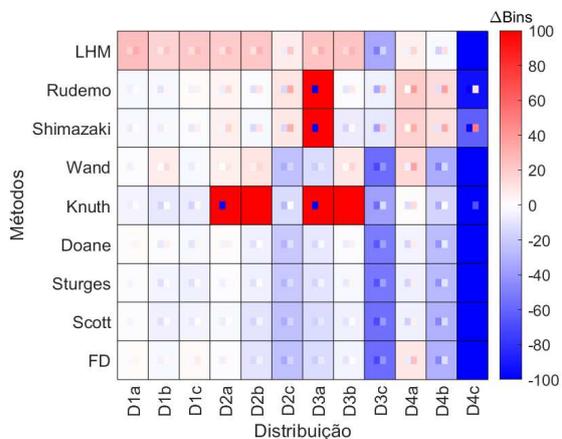
Fonte: Elaborada pelo autor (2020).

Figura 237 – Área do erro para todas as distribuições com 50 amostras de treinamento no Histograma



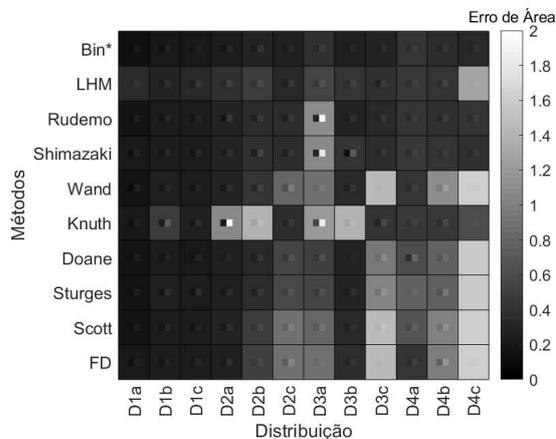
Fonte: Elaborada pelo autor (2020).

Figura 238 – Variação dos *bins* para todas as distribuições com 100 amostras de treinamento no Histograma



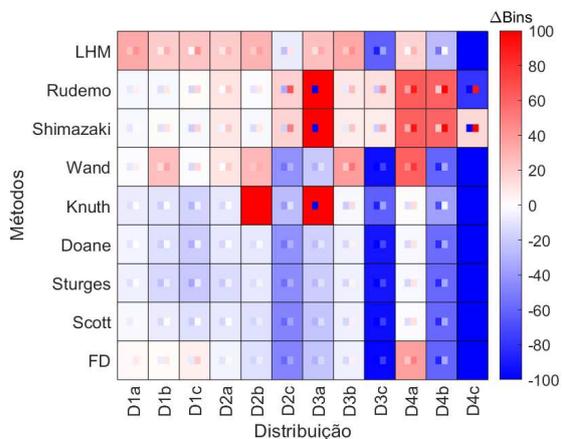
Fonte: Elaborada pelo autor (2020).

Figura 239 – Área do erro para todas as distribuições com 100 amostras de treinamento no Histograma



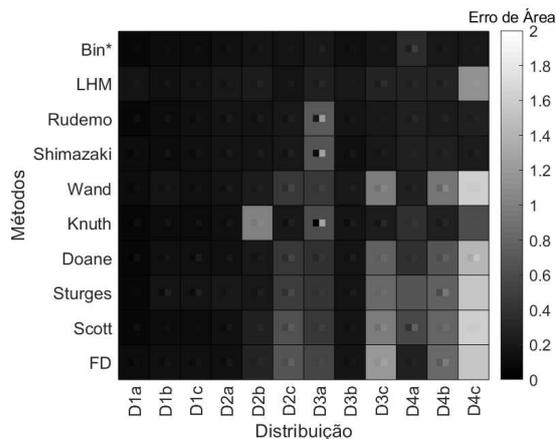
Fonte: Elaborada pelo autor (2020).

Figura 240 – Variação dos *bins* para todas as distribuições com 500 amostras de treinamento no Histograma



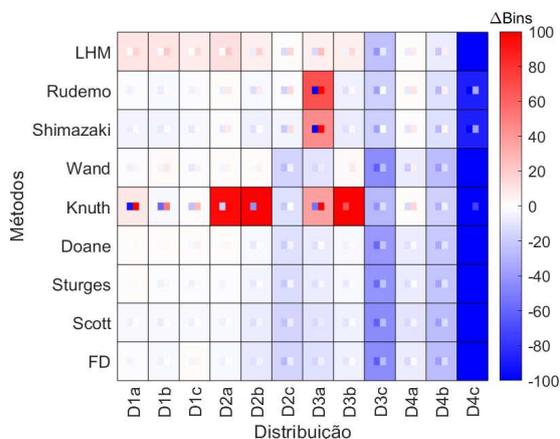
Fonte: Elaborada pelo autor (2020).

Figura 241 – Área do erro para todas as distribuições com 500 amostras de treinamento no Histograma



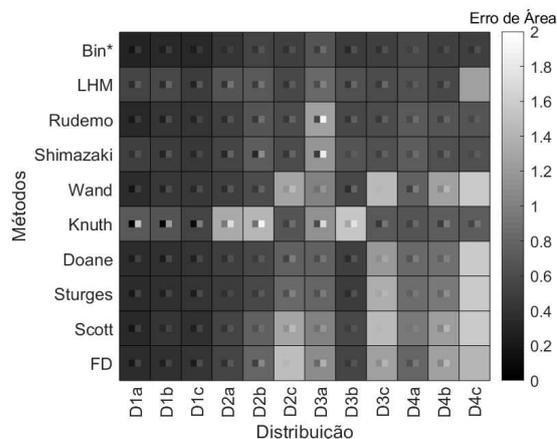
Fonte: Elaborada pelo autor (2020).

Figura 242 – Variação dos *bins* para todas as distribuições com 25 amostras de treinamento no PF



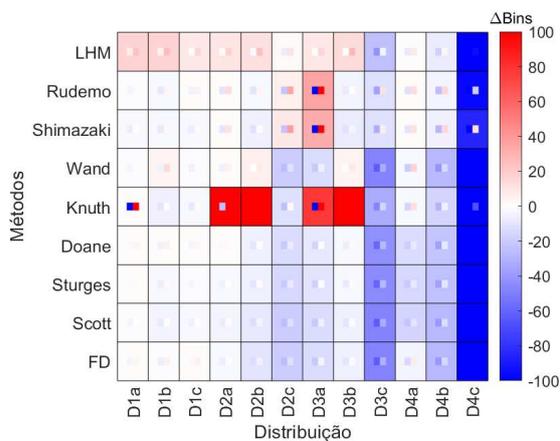
Fonte: Elaborada pelo autor (2020).

Figura 243 – Área do erro para todas as distribuições com 25 amostras de treinamento no PF



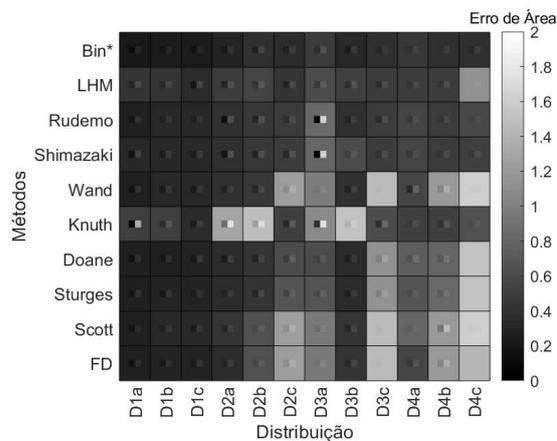
Fonte: Elaborada pelo autor (2020).

Figura 244 – Variação dos *bins* para todas as distribuições com 50 amostras de treinamento no PF



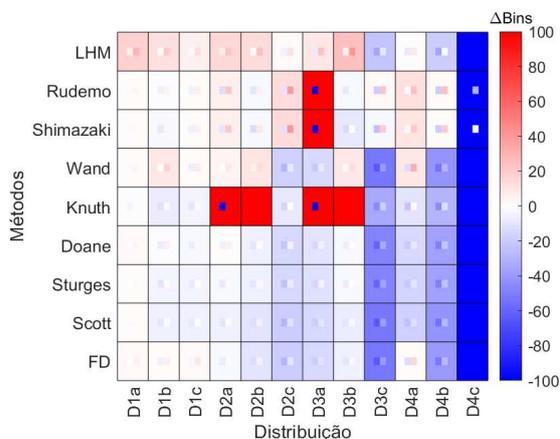
Fonte: Elaborada pelo autor (2020).

Figura 245 – Área do erro para todas as distribuições com 50 amostras de treinamento no PF



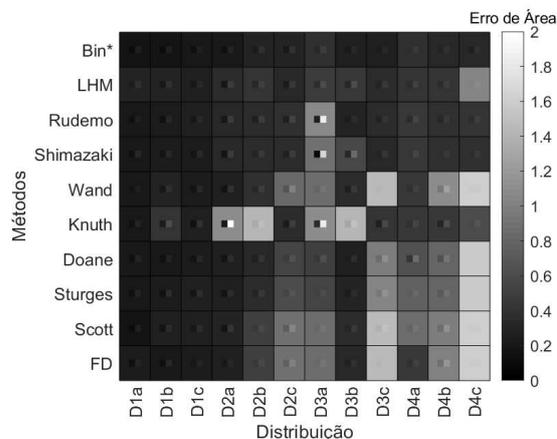
Fonte: Elaborada pelo autor (2020).

Figura 246 – Variação dos *bins* para todas as distribuições com 100 amostras de treinamento no PF



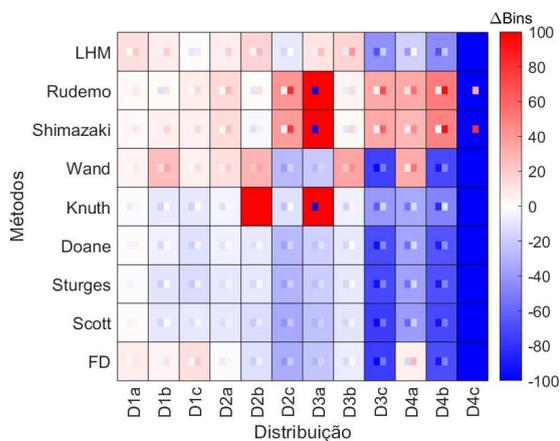
Fonte: Elaborada pelo autor (2020).

Figura 247 – Área do erro para todas as distribuições com 100 amostras de treinamento no PF



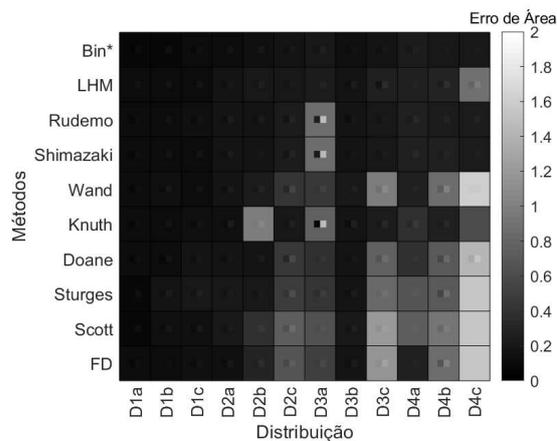
Fonte: Elaborada pelo autor (2020).

Figura 248 – Variação dos *bins* para todas as distribuições com 500 amostras de treinamento no PF



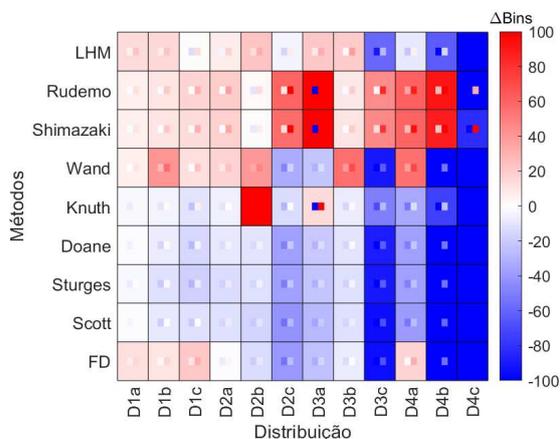
Fonte: Elaborada pelo autor (2020).

Figura 249 – Área do erro para todas as distribuições com 500 amostras de treinamento no PF



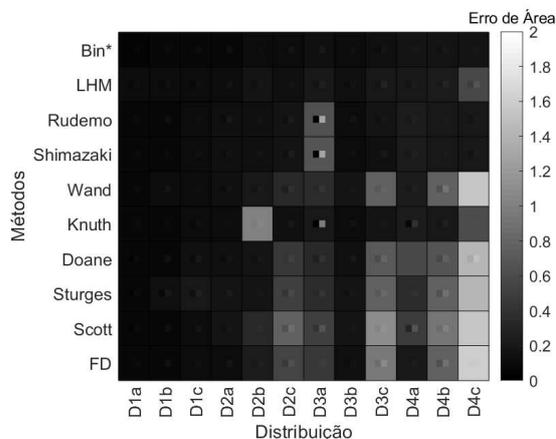
Fonte: Elaborada pelo autor (2020).

Figura 250 – Variação dos bins para todas as distribuições com 1000 amostras de treinamento no PF



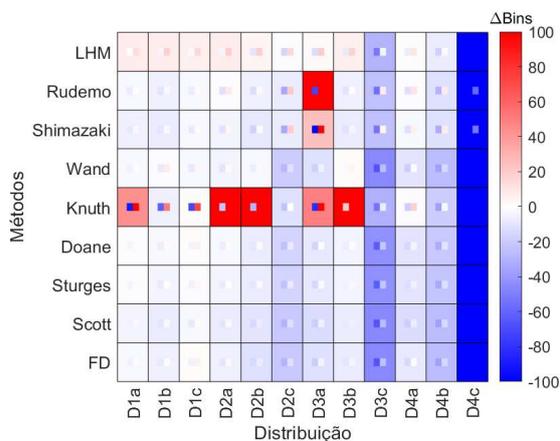
Fonte: Elaborada pelo autor (2020).

Figura 251 – Área do erro para todas as distribuições com 1000 amostras de treinamento no PF



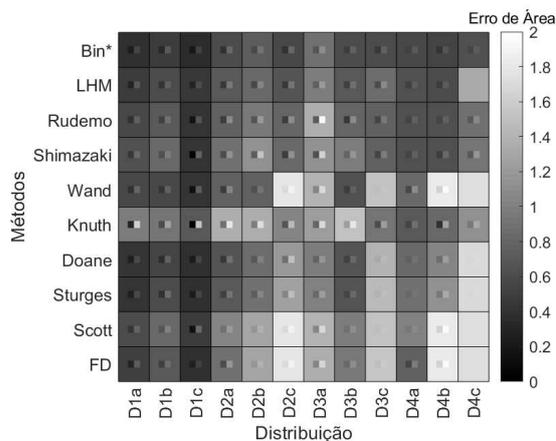
Fonte: Elaborada pelo autor (2020).

Figura 252 – Variação dos bins para todas as distribuições com 25 amostras de treinamento no ASH



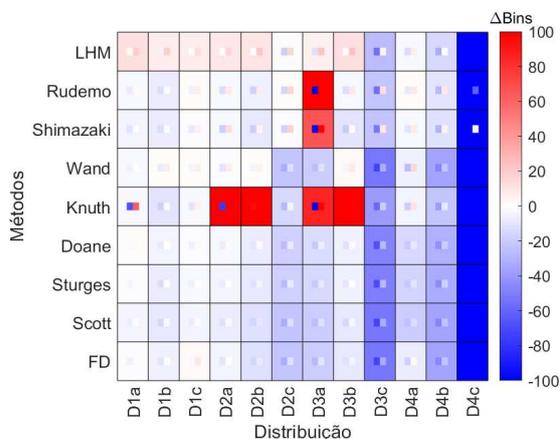
Fonte: Elaborada pelo autor (2020).

Figura 253 – Área do erro para todas as distribuições com 25 amostras de treinamento no ASH



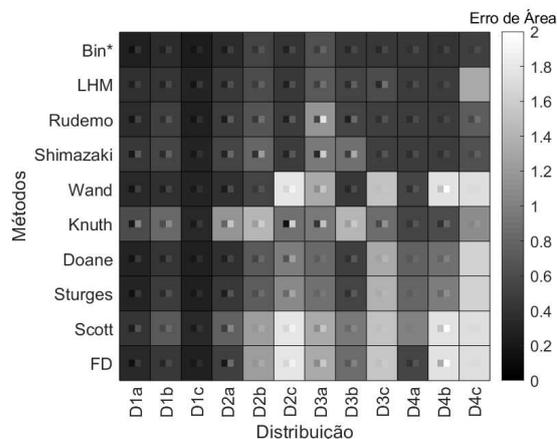
Fonte: Elaborada pelo autor (2020).

Figura 254 – Variação dos bins para todas as distribuições com 50 amostras de treinamento no ASH



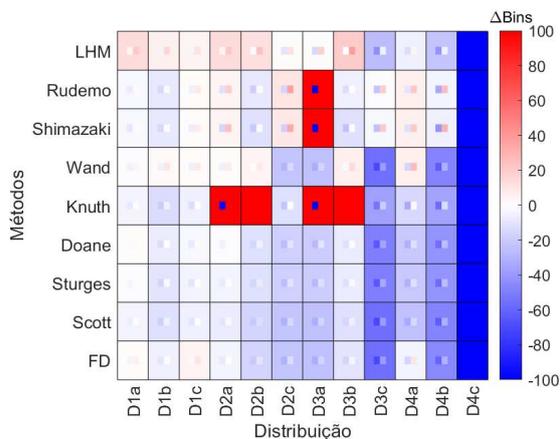
Fonte: Elaborada pelo autor (2020).

Figura 255 – Área do erro para todas as distribuições com 50 amostras de treinamento no ASH



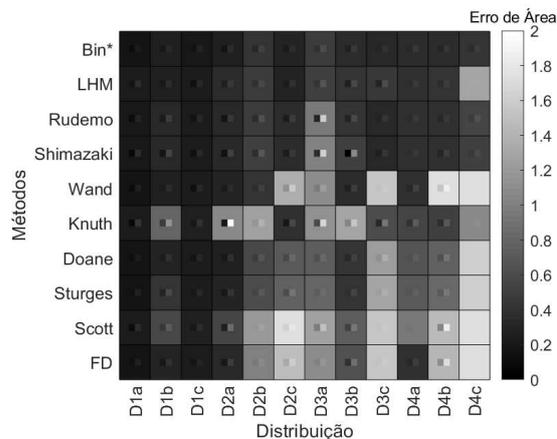
Fonte: Elaborada pelo autor (2020).

Figura 256 – Variação dos bins para todas as distribuições com 100 amostras de treinamento no ASH



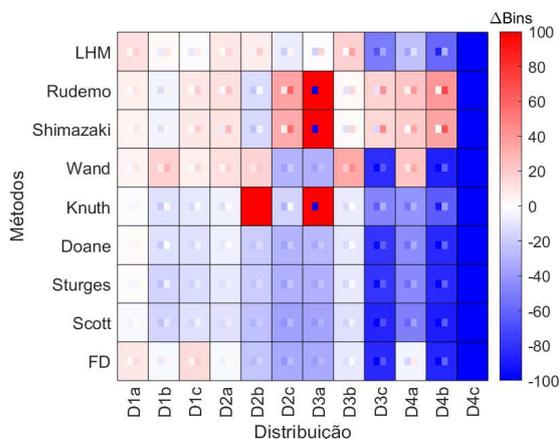
Fonte: Elaborada pelo autor (2020).

Figura 257 – Área do erro para todas as distribuições com 100 amostras de treinamento no ASH



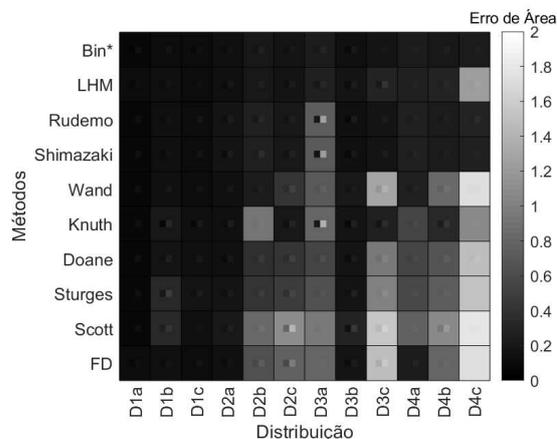
Fonte: Elaborada pelo autor (2020).

Figura 258 – Variação dos bins para todas as distribuições com 500 amostras de treinamento no ASH



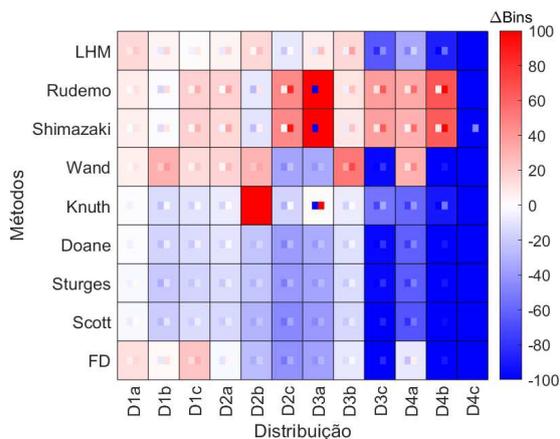
Fonte: Elaborada pelo autor (2020).

Figura 259 – Área do erro para todas as distribuições com 500 amostras de treinamento no ASH



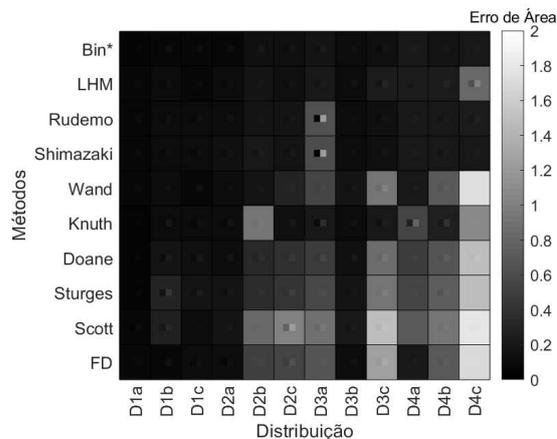
Fonte: Elaborada pelo autor (2020).

Figura 260 – Variação dos bins para todas as distribuições com 1000 amostras de treinamento no ASH



Fonte: Elaborada pelo autor (2020).

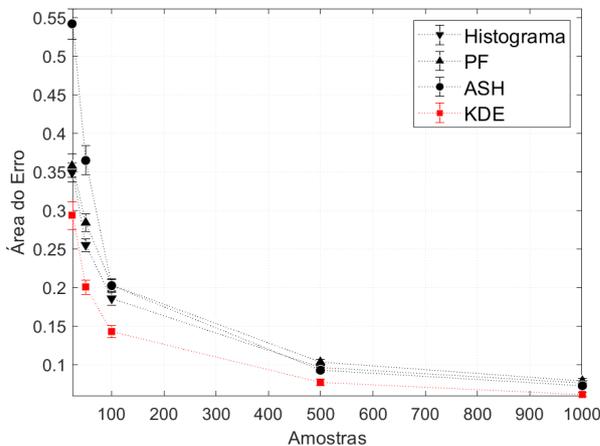
Figura 261 – Área do erro para todas as distribuições com 1000 amostras de treinamento no ASH



Fonte: Elaborada pelo autor (2020).

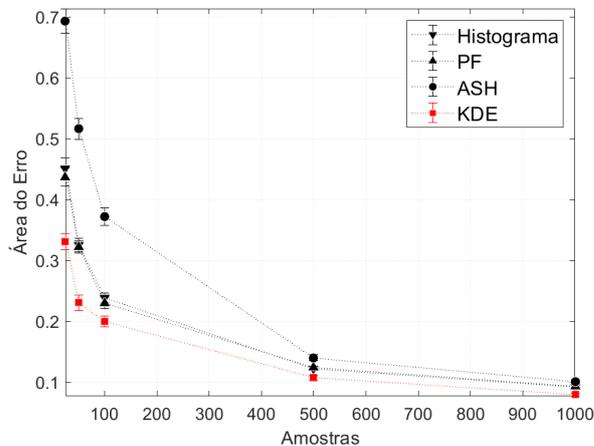
C.1.4 Comparação entre estimadores de banda fixa

Figura 262 – Comparação entre estimadores para Distribuição D1a.



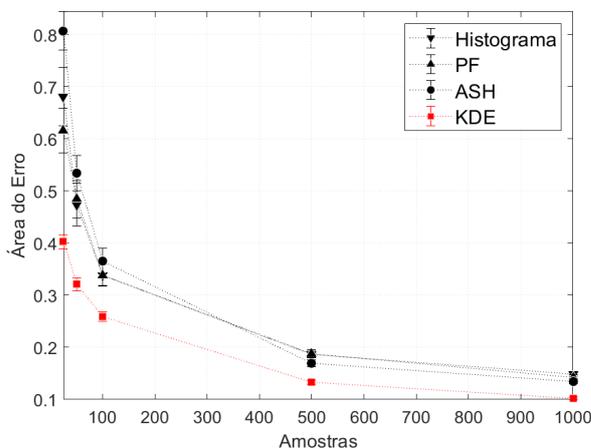
Fonte: Elaborada pelo autor (2020).

Figura 263 – Comparação entre estimadores para Distribuição D1b.



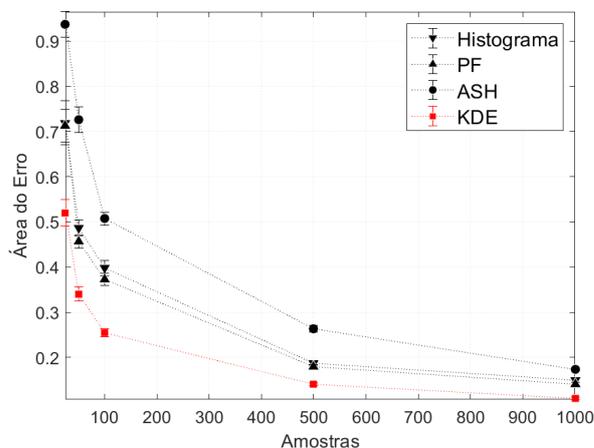
Fonte: Elaborada pelo autor (2020).

Figura 264 – Comparação entre estimadores para Distribuição D2a.



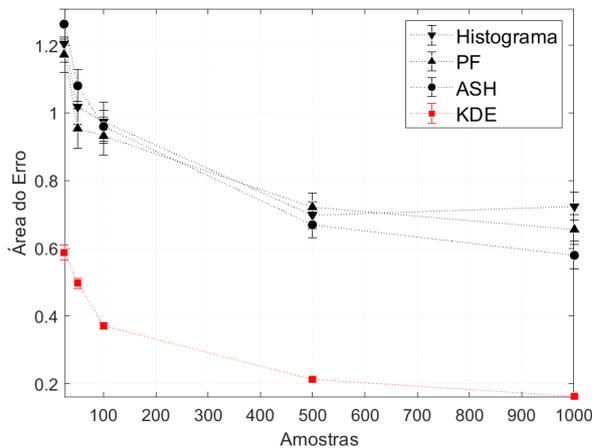
Fonte: Elaborada pelo autor (2020).

Figura 265 – Comparação entre estimadores para Distribuição D2b.



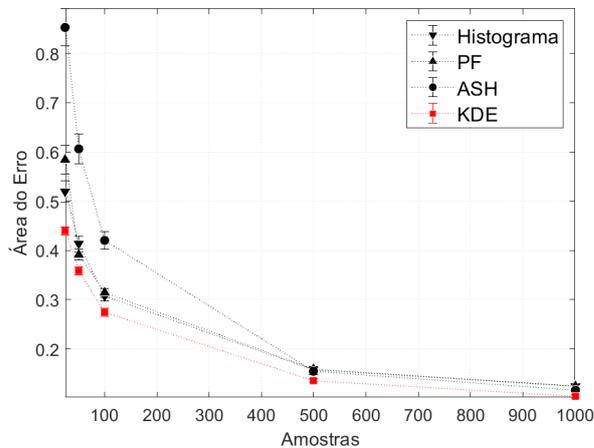
Fonte: Elaborada pelo autor (2020).

Figura 266 – Comparação entre estimadores para Distribuição D3a.



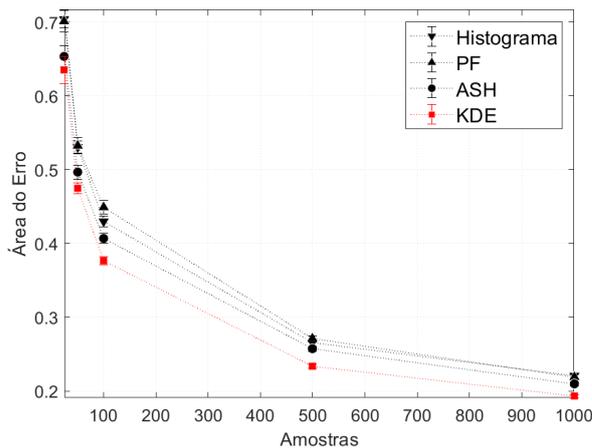
Fonte: Elaborada pelo autor (2020).

Figura 267 – Comparação entre estimadores para Distribuição D3b.



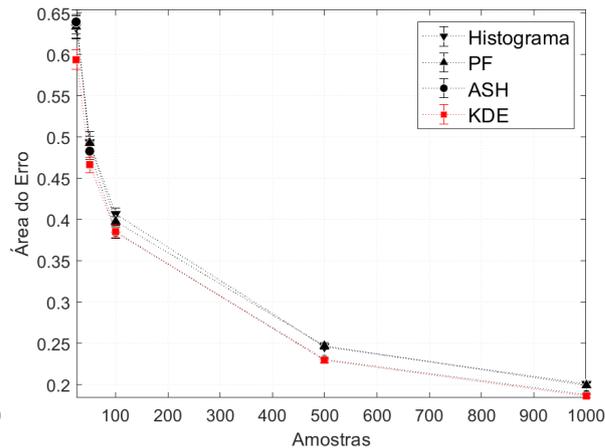
Fonte: Elaborada pelo autor (2020).

Figura 268 – Comparação entre estimadores para Distribuição D4a.



Fonte: Elaborada pelo autor (2020).

Figura 269 – Comparação entre estimadores para Distribuição D4b.

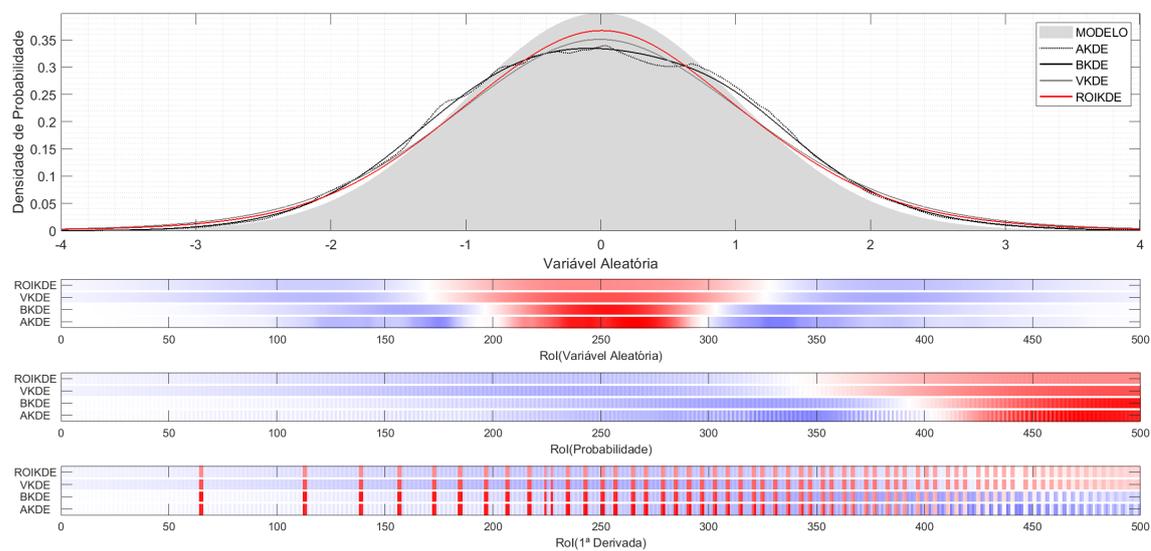


Fonte: Elaborada pelo autor (2020).

C.2 BANDA VARIÁVEL

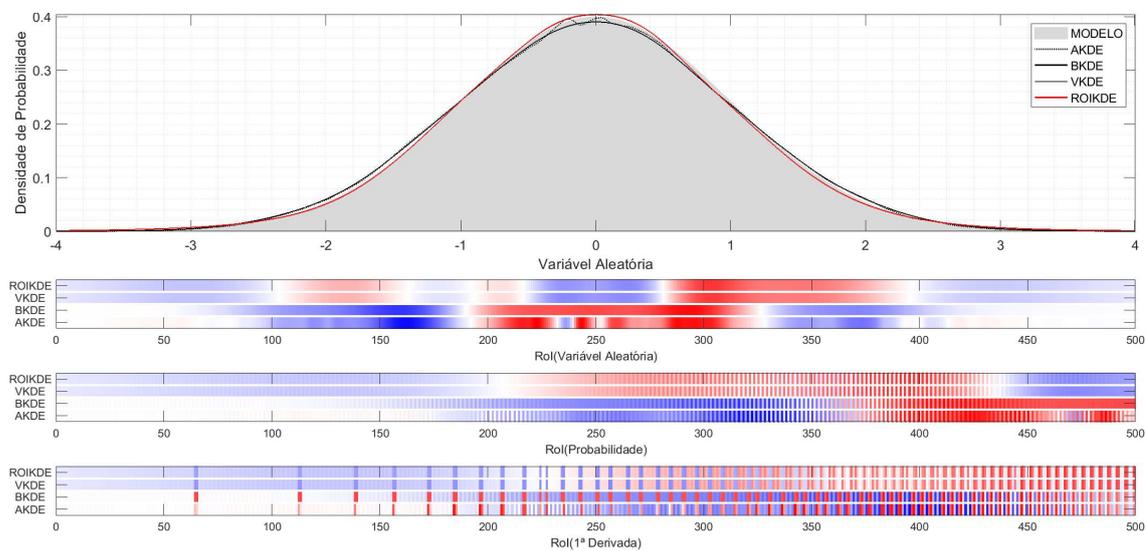
C.2.1 RoIMap

Figura 270 – Ferramenta RoIMap utilizada na distribuição D1a para 25 amostras.



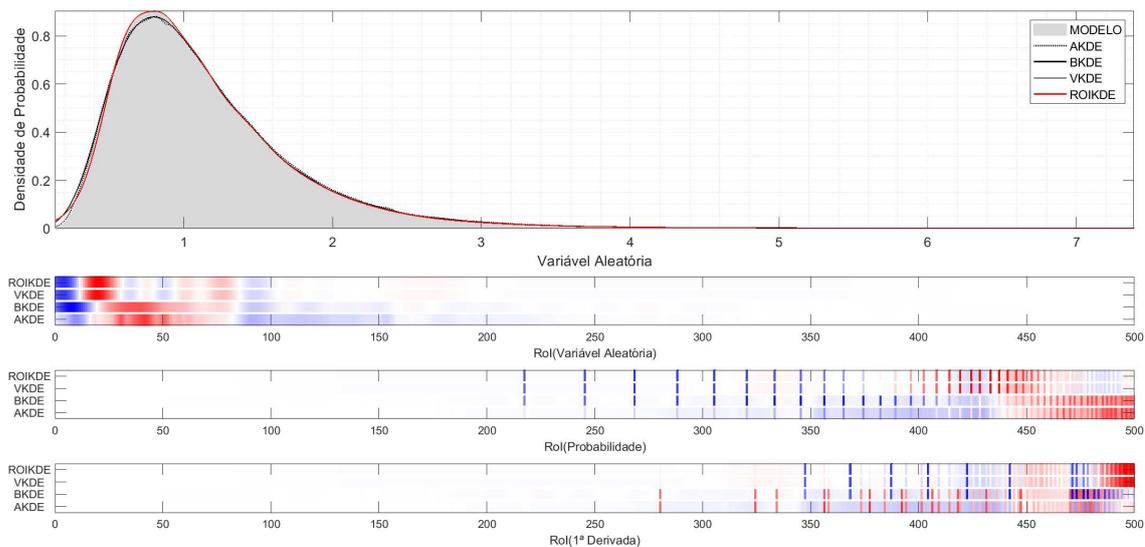
Fonte: Elaborada pelo autor (2020).

Figura 271 – Ferramenta RoIMap utilizada na distribuição D1a para 1000 amostras.



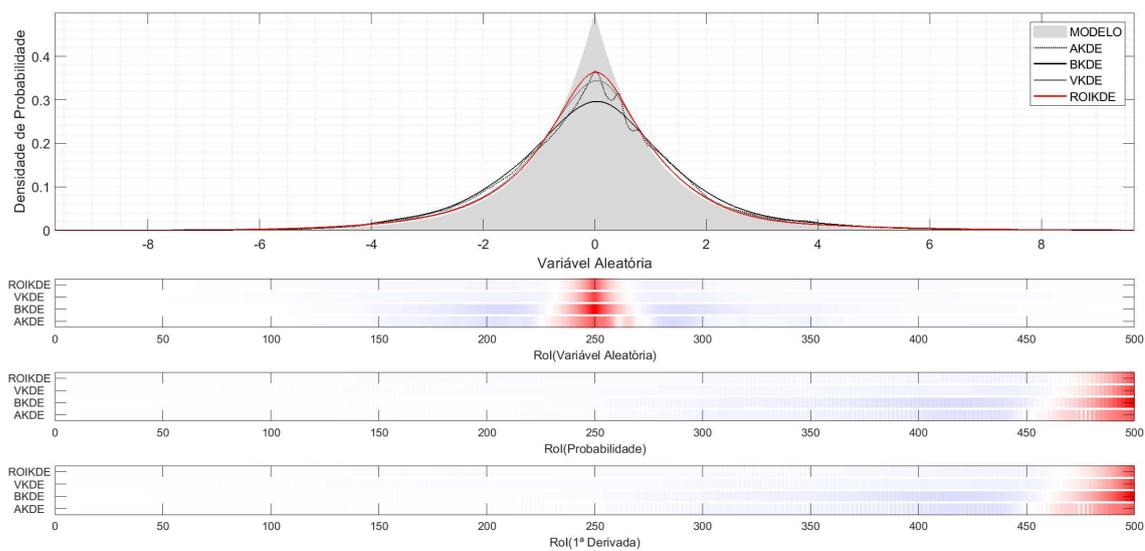
Fonte: Elaborada pelo autor (2020).

Figura 272 – Ferramenta RoIMap utilizada na distribuição D1b para 1000 amostras.



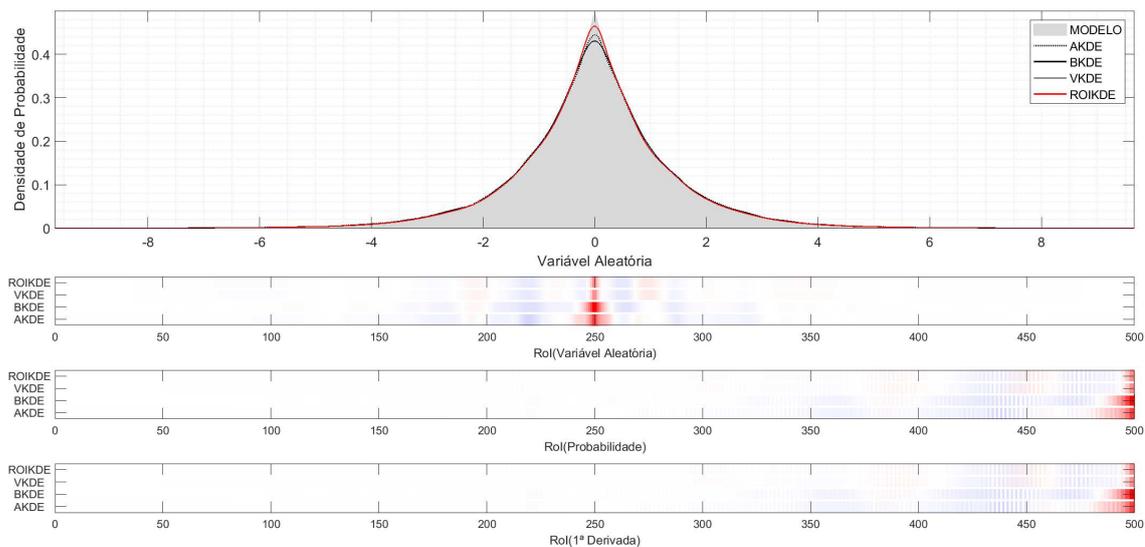
Fonte: Elaborada pelo autor (2020).

Figura 273 – Ferramenta RoIMap utilizada na distribuição D1c para 25 amostras.



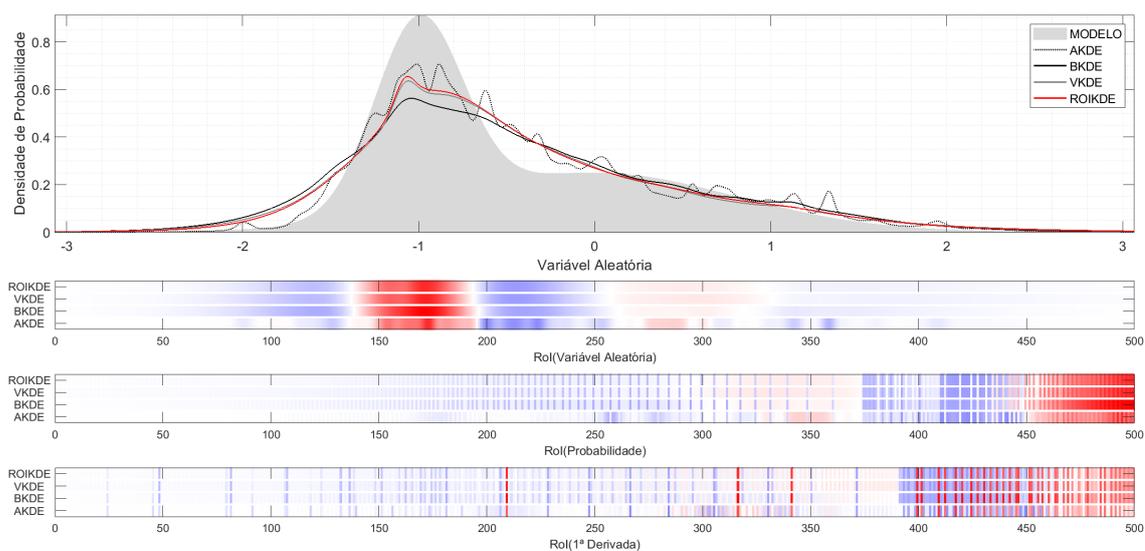
Fonte: Elaborada pelo autor (2020).

Figura 274 – Ferramenta RolMap utilizada na distribuição D1c para 1000 amostras.



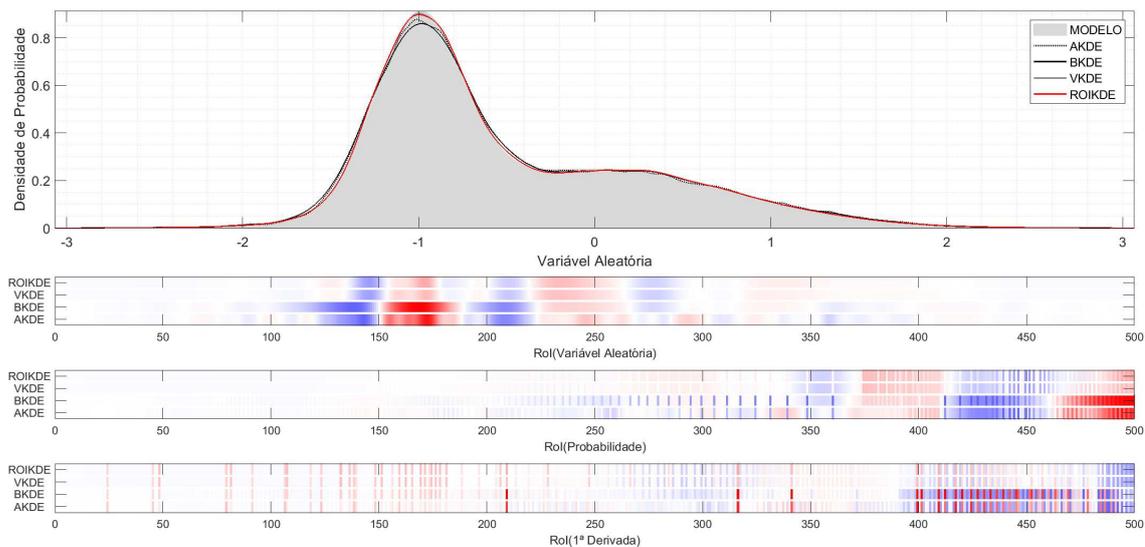
Fonte: Elaborada pelo autor (2020).

Figura 275 – Ferramenta RolMap utilizada na distribuição D2a para 25 amostras.



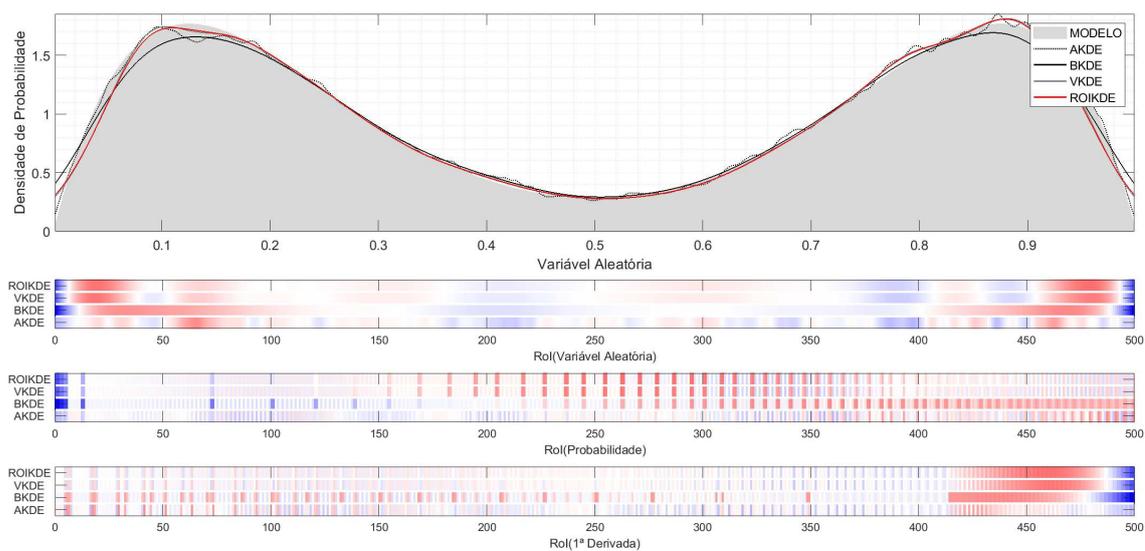
Fonte: Elaborada pelo autor (2020).

Figura 276 – Ferramenta RoIMap utilizada na distribuição D2a para 1000 amostras.



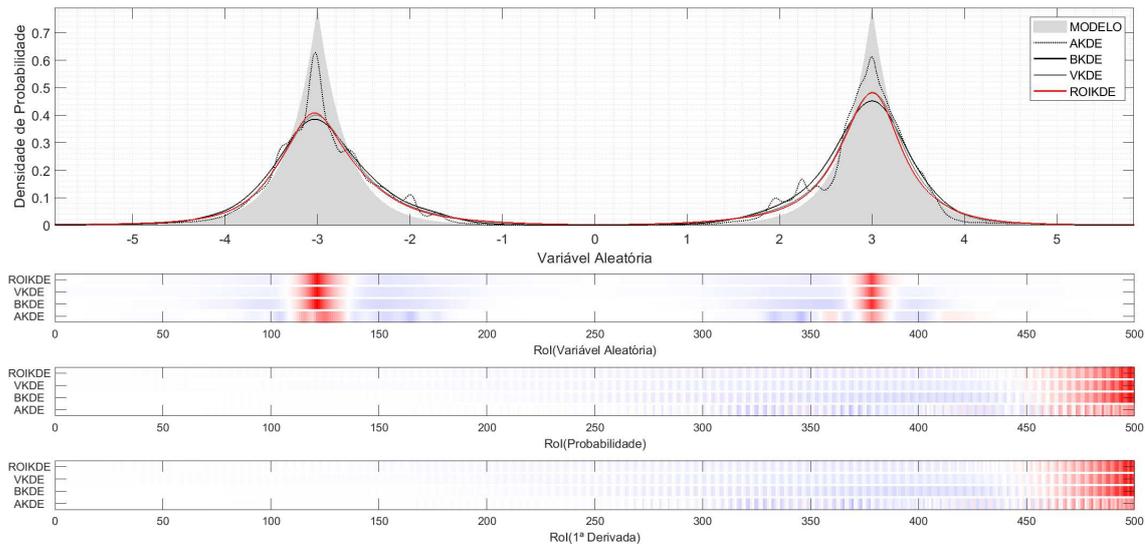
Fonte: Elaborada pelo autor (2020).

Figura 277 – Ferramenta RoIMap utilizada na distribuição D2b para 1000 amostras.



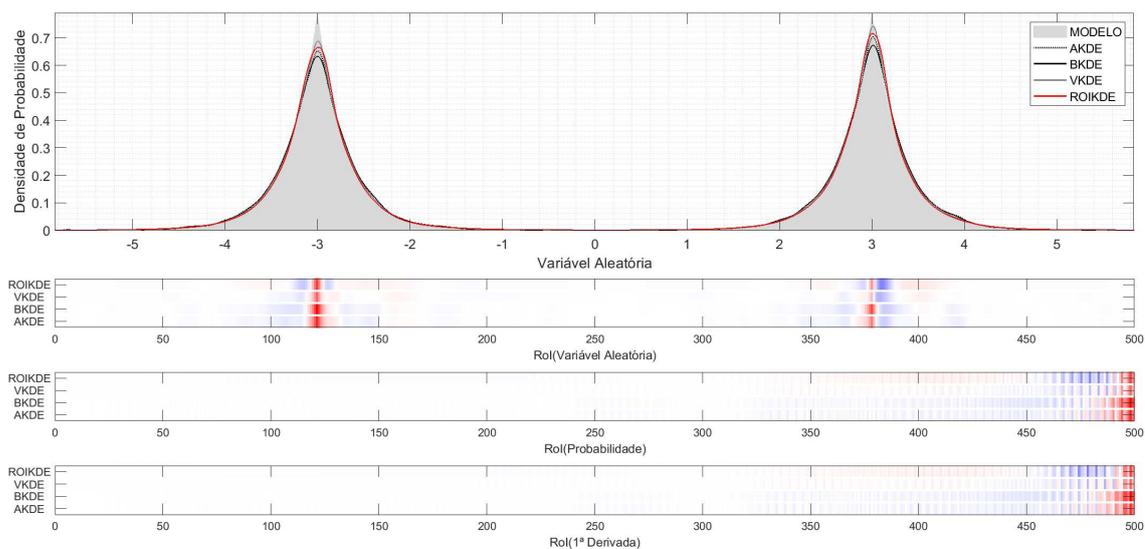
Fonte: Elaborada pelo autor (2020).

Figura 278 – Ferramenta RoIMap utilizada na distribuição D2c para 25 amostras.



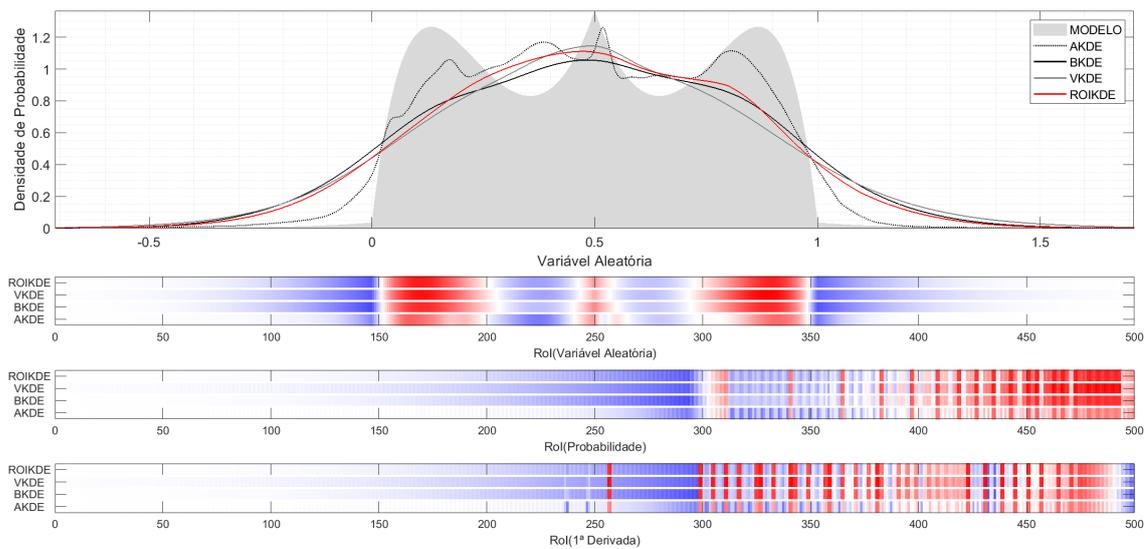
Fonte: Elaborada pelo autor (2020).

Figura 279 – Ferramenta RoIMap utilizada na distribuição D2c para 1000 amostras.



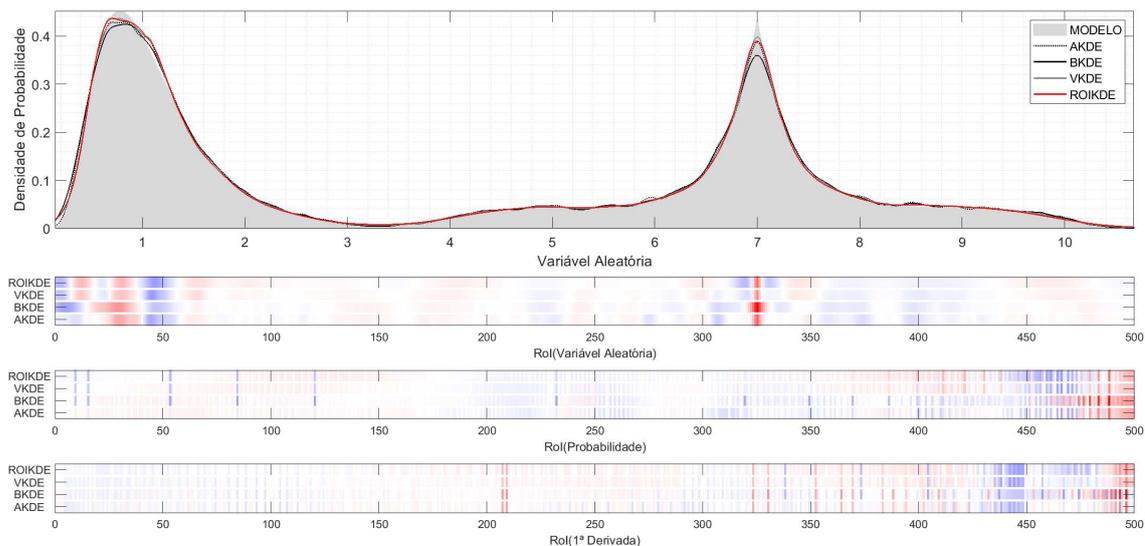
Fonte: Elaborada pelo autor (2020).

Figura 281 – Ferramenta RoIMap utilizada na distribuição D3b para 25 amostras.



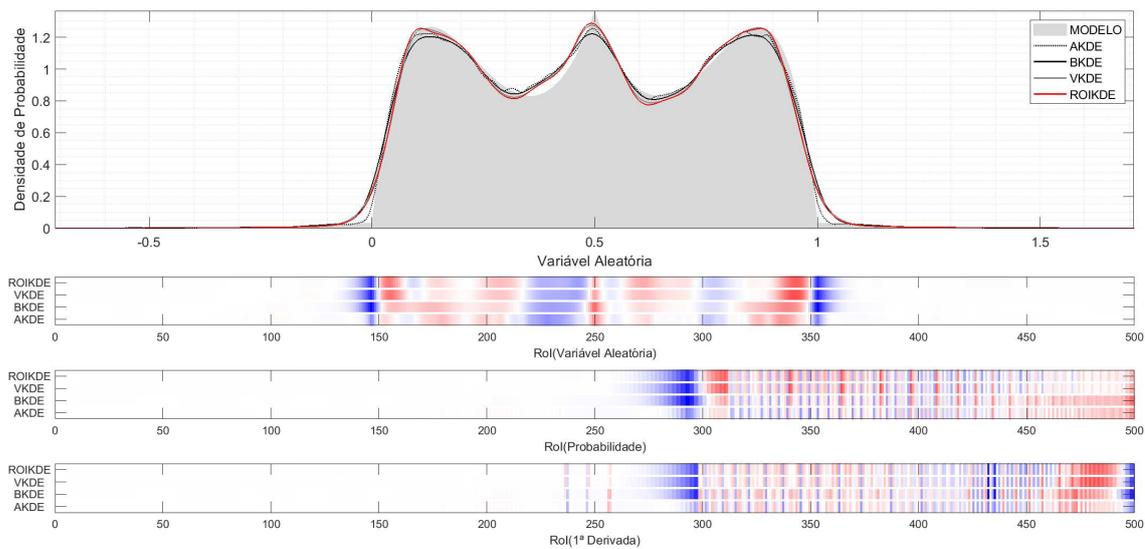
Fonte: Elaborada pelo autor (2020).

Figura 280 – Ferramenta RoIMap utilizada na distribuição D3a para 1000 amostras.



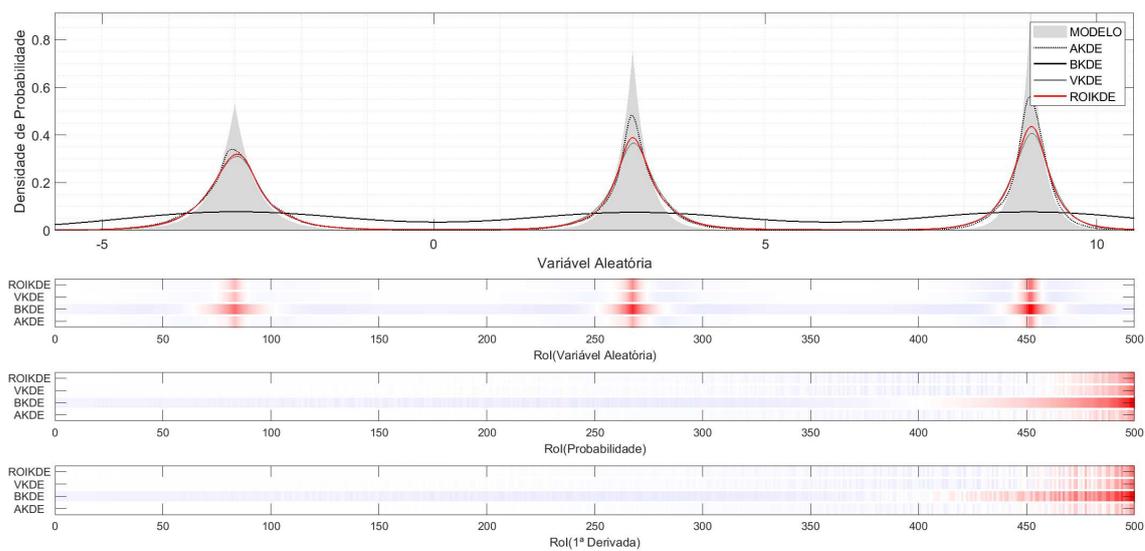
Fonte: Elaborada pelo autor (2020).

Figura 282 – Ferramenta RoIMap utilizada na distribuição D3b para 1000 amostras.



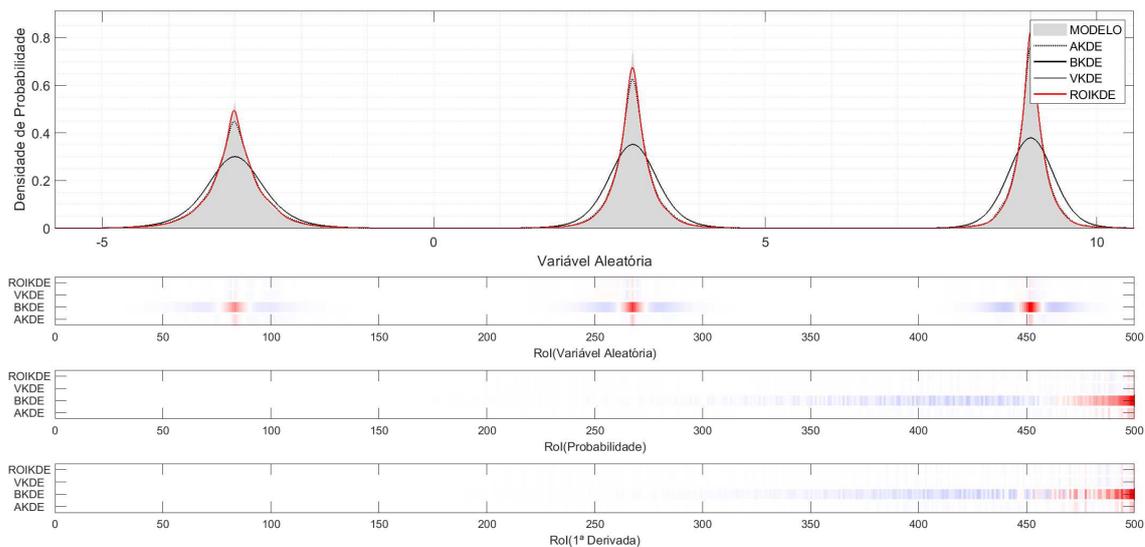
Fonte: Elaborada pelo autor (2020).

Figura 283 – Ferramenta RoIMap utilizada na distribuição D3c para 25 amostras.



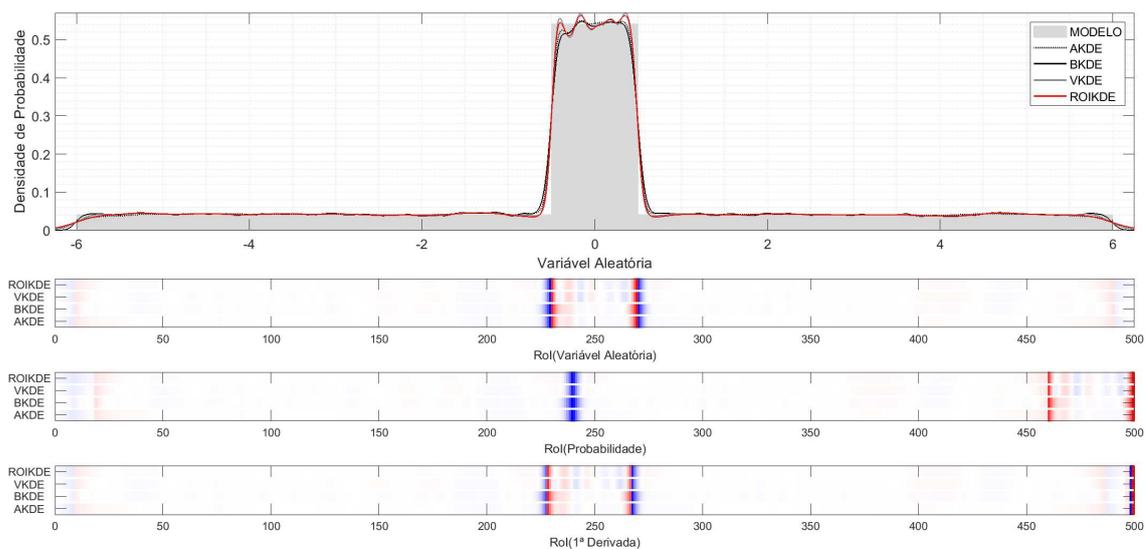
Fonte: Elaborada pelo autor (2020).

Figura 284 – Ferramenta RoIMap utilizada na distribuição D3c para 1000 amostras.



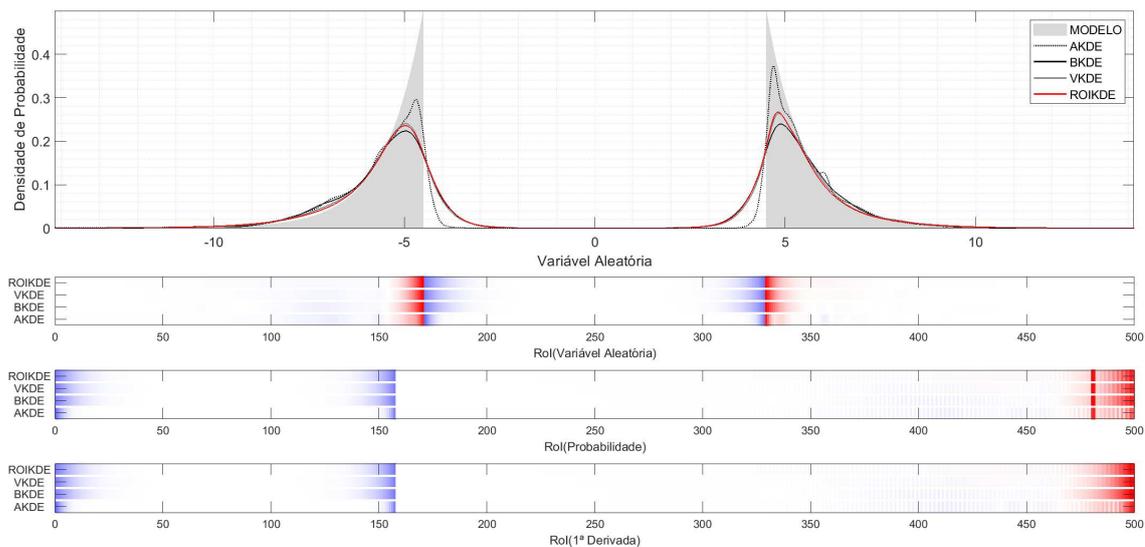
Fonte: Elaborada pelo autor (2020).

Figura 285 – Ferramenta RoIMap utilizada na distribuição D4a para 1000 amostras.



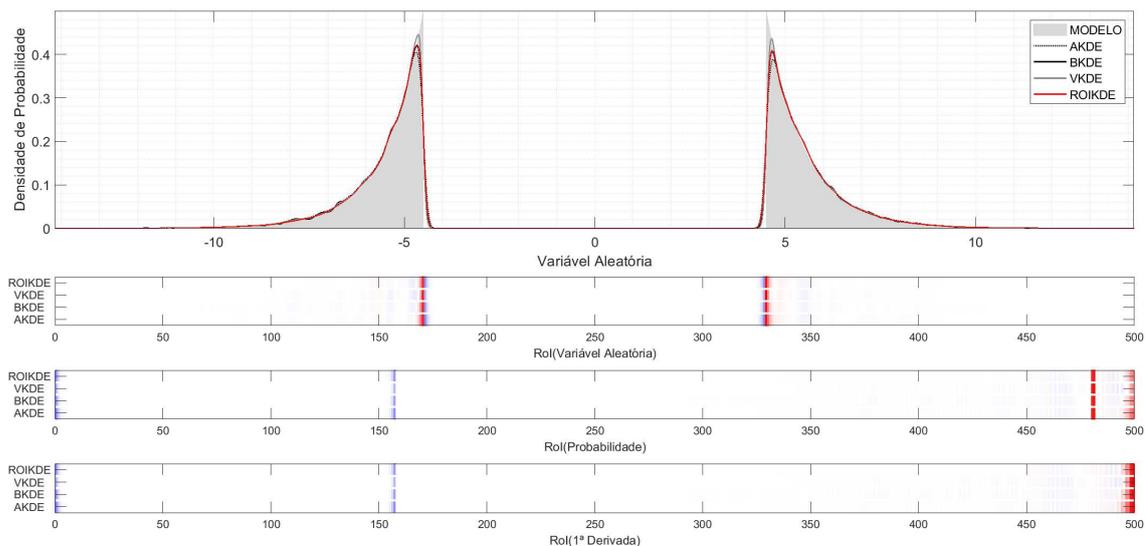
Fonte: Elaborada pelo autor (2020).

Figura 286 – Ferramenta RoIMap utilizada na distribuição D4b para 25 amostras.



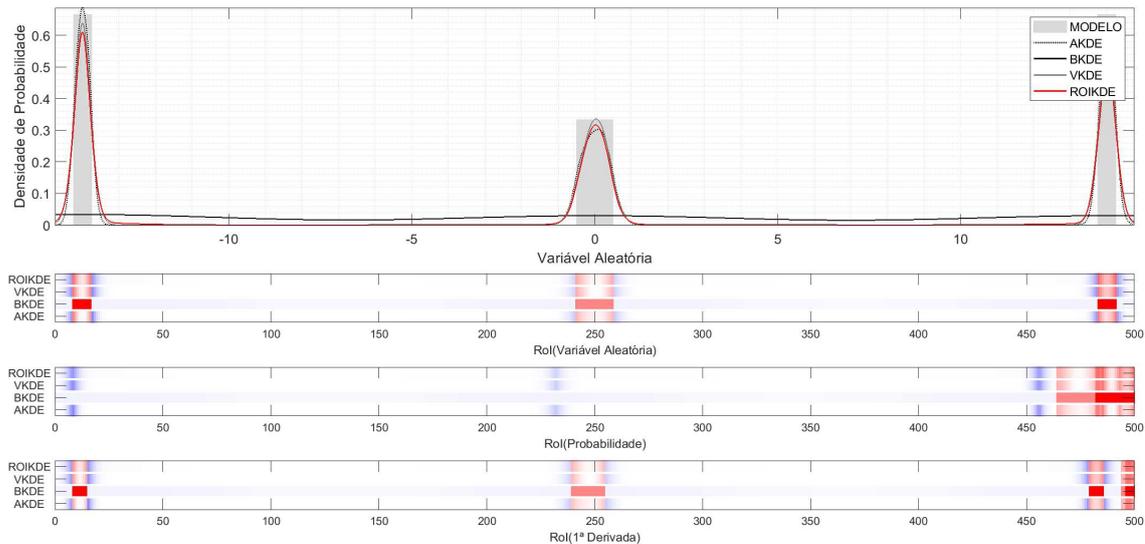
Fonte: Elaborada pelo autor (2020).

Figura 287 – Ferramenta RoIMap utilizada na distribuição D4b para 1000 amostras.



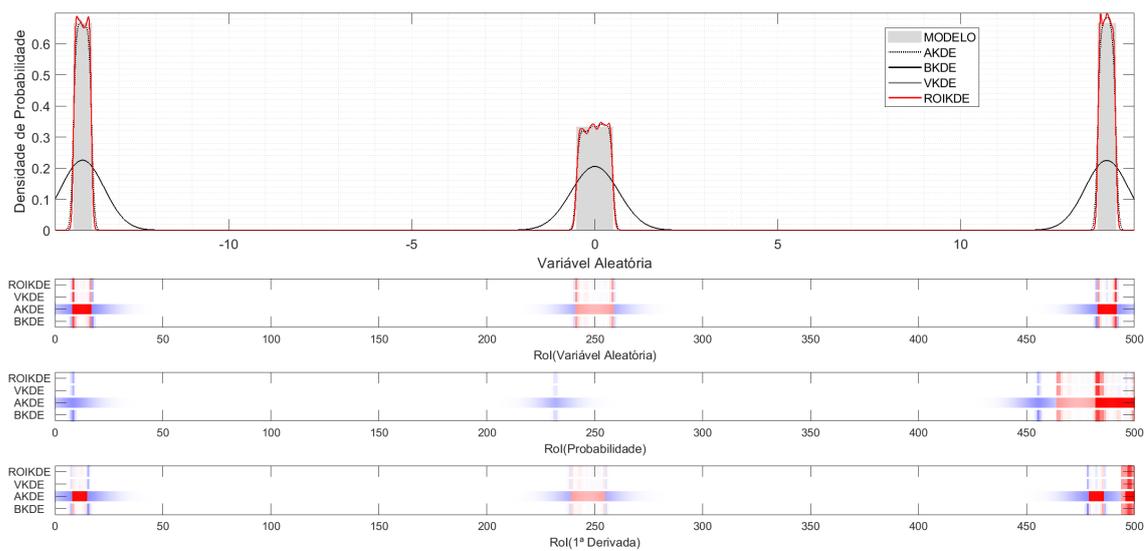
Fonte: Elaborada pelo autor (2020).

Figura 288 – Ferramenta RoIMap utilizada na distribuição D4c para 25 amostras.



Fonte: Elaborada pelo autor (2020).

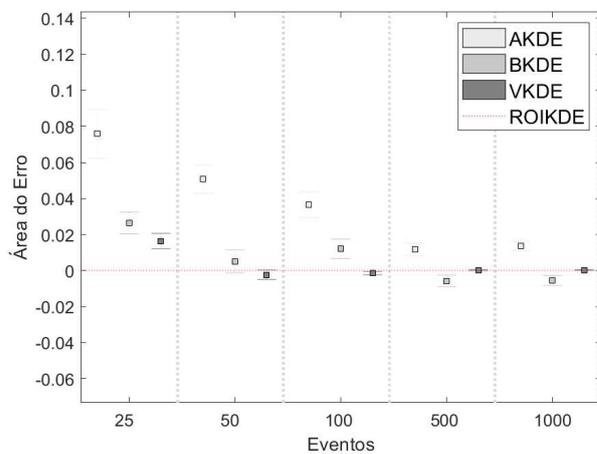
Figura 289 – Ferramenta RoIMap utilizada na distribuição D4c para 1000 amostras.



Fonte: Elaborada pelo autor (2020).

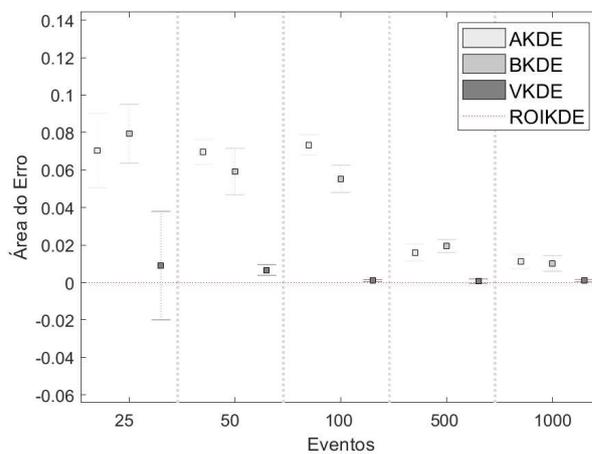
C.2.2 Comparação em relação ao ROIKDE

Figura 290 – Área do erro em relação ao ROIKDE para D1a



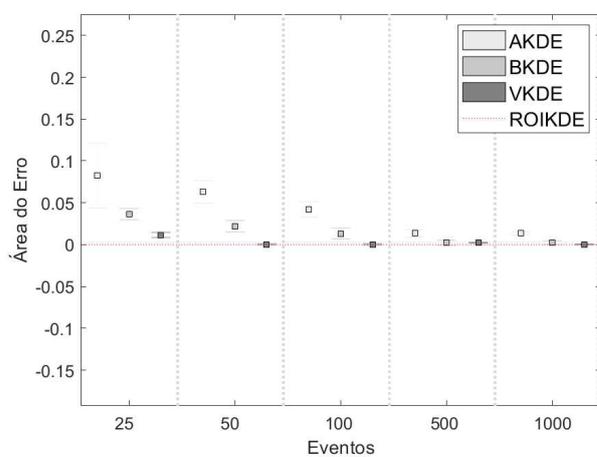
Fonte: Elaborada pelo autor (2020).

Figura 291 – Área do erro em relação ao ROIKDE para D1c



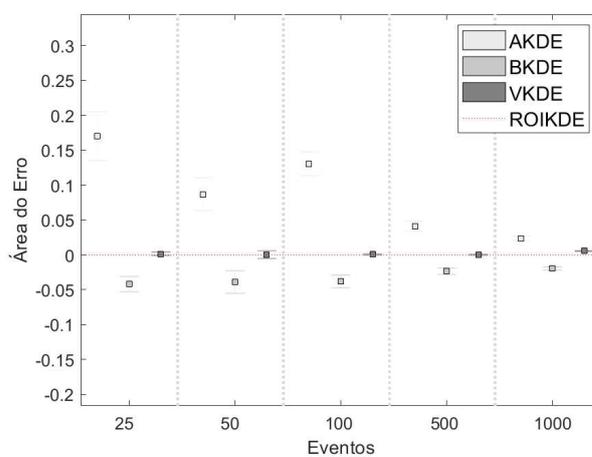
Fonte: Elaborada pelo autor (2020).

Figura 292 – Área do erro em relação ao ROIKDE para D2a.



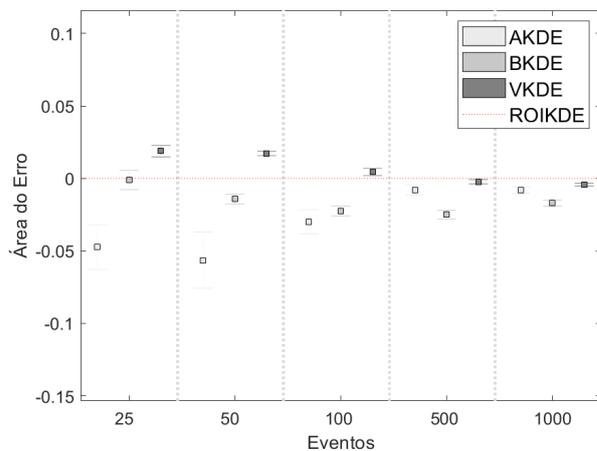
Fonte: Elaborada pelo autor (2020).

Figura 293 – Área do erro em relação ao ROIKDE para D2b.



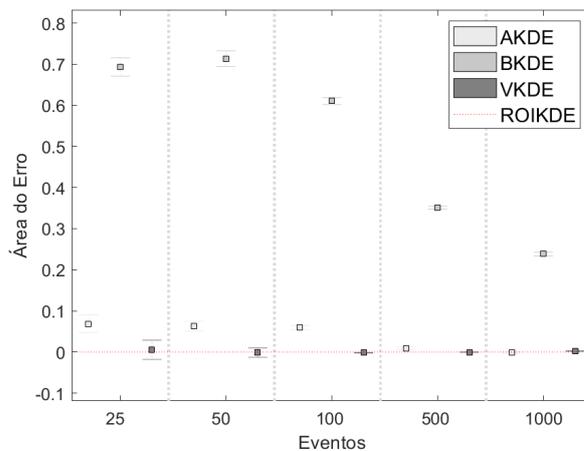
Fonte: Elaborada pelo autor (2020).

Figura 294 – Área do erro em relação ao ROIKDE para D3b.



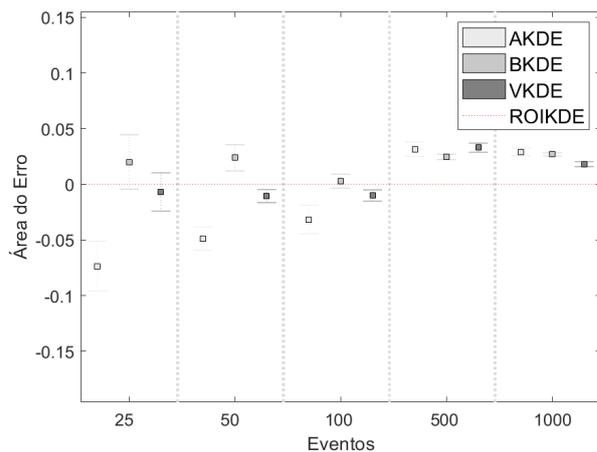
Fonte: Elaborada pelo autor (2020).

Figura 295 – Área do erro em relação ao ROIKDE para D3c.



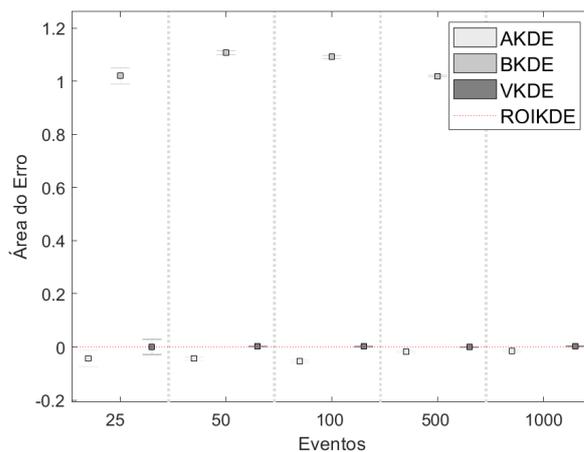
Fonte: Elaborada pelo autor (2020).

Figura 296 – Área do erro em relação ao ROIKDE para D4b.



Fonte: Elaborada pelo autor (2020).

Figura 297 – Área do erro em relação ao ROIKDE para D4c.



Fonte: Elaborada pelo autor (2020).

APÊNDICE D – TABELAS

D.1 KDE DE BANDA FIXA

Tabela 25 – Área do erro para distribuição D1a.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
SV	0,24[0,13]	0,18[0,07]	0,13[0,07]	0,07[0,03]	0,06[0,02]
SVM1	0,25[0,13]	0,19[0,08]	0,14[0,08]	0,07[0,03]	0,06[0,02]
SVM2	0,24[0,11]	0,18[0,08]	0,15[0,07]	0,07[0,03]	0,06[0,02]
SJ	0,26[0,14]	0,19[0,07]	0,15[0,08]	0,07[0,03]	0,06[0,02]
SC	0,24[0,12]	0,18[0,07]	0,14[0,07]	0,07[0,03]	0,06[0,02]
L1I	0,28[0,14]	0,19[0,08]	0,19[0,07]	0,09[0,03]	0,07[0,02]
MLCV	0,25[0,15]	0,2[0,08]	0,16[0,08]	0,08[0,04]	0,06[0,02]
UCV	0,3[0,19]	0,21[0,1]	0,16[0,1]	0,08[0,04]	0,07[0,02]
BCV1	0,29[0,16]	0,21[0,09]	0,15[0,08]	0,07[0,04]	0,06[0,02]
BCV2	0,29[0,15]	0,19[0,1]	0,16[0,09]	0,08[0,04]	0,06[0,02]
CCV	0,28[0,15]	0,19[0,08]	0,16[0,09]	0,08[0,04]	0,06[0,02]
MCV	0,31[0,18]	0,23[0,1]	0,17[0,08]	0,08[0,04]	0,06[0,02]
TCV	0,31[0,14]	0,24[0,1]	0,16[0,08]	0,08[0,04]	0,07[0,02]
OSCV	0,27[0,17]	0,2[0,08]	0,14[0,08]	0,07[0,03]	0,06[0,02]
TRG	0,21[0,14]	0,16[0,07]	0,12[0,07]	0,07[0,03]	0,06[0,02]
TRE	0,21[0,14]	0,16[0,08]	0,12[0,08]	0,06[0,03]	0,06[0,02]

Fonte: Elaborada pelo autor (2020).

Tabela 26 – Área do erro para distribuição D1b.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
SV	0,25[0,1]	0,22[0,08]	0,17[0,05]	0,11[0,02]	0,09[0,02]
SVM1	0,3[0,1]	0,24[0,06]	0,21[0,04]	0,13[0,02]	0,11[0,02]
SVM2	0,25[0,11]	0,21[0,08]	0,16[0,05]	0,1[0,02]	0,08[0,01]
SJ	0,26[0,11]	0,22[0,08]	0,17[0,05]	0,1[0,02]	0,08[0,01]
SC	0,29[0,11]	0,23[0,07]	0,2[0,05]	0,12[0,02]	0,1[0,02]
L1I	0,27[0,12]	0,22[0,09]	0,16[0,06]	0,1[0,03]	0,07[0,02]
MLCV	0,27[0,15]	0,24[0,11]	0,17[0,07]	0,11[0,03]	0,08[0,02]
UCV	0,32[0,09]	0,25[0,08]	0,18[0,06]	0,1[0,02]	0,08[0,02]
BCV1	0,34[0,15]	0,29[0,09]	0,24[0,07]	0,11[0,03]	0,08[0,01]
BCV2	0,34[0,13]	0,24[0,08]	0,18[0,06]	0,1[0,02]	0,08[0,02]
CCV	0,34[0,13]	0,24[0,08]	0,19[0,06]	0,1[0,03]	0,08[0,01]
MCV	0,39[0,18]	0,33[0,13]	0,23[0,08]	0,11[0,03]	0,08[0,02]
TCV	0,35[0,1]	0,27[0,05]	0,21[0,04]	0,11[0,02]	0,08[0,02]
OSCV	0,31[0,09]	0,24[0,06]	0,18[0,05]	0,1[0,02]	0,08[0,02]
TRG	0,24[0,1]	0,2[0,08]	0,15[0,05]	0,09[0,03]	0,07[0,01]
TRE	0,23[0,08]	0,2[0,08]	0,15[0,05]	0,09[0,03]	0,07[0,02]

Fonte: Elaborada pelo autor (2020).

Tabela 27 – Área do erro para distribuição D1c.

Seletores	Amostras				
	25 Mediana[IQR]	50 Mediana[IQR]	100 Mediana[IQR]	500 Mediana[IQR]	1000 Mediana[IQR]
SV	0,33[0,14]	0,23[0,09]	0,21[0,09]	0,13[0,03]	0,1[0,02]
SVM1	0,31[0,11]	0,22[0,09]	0,2[0,08]	0,11[0,03]	0,09[0,02]
SVM2	0,31[0,13]	0,23[0,09]	0,2[0,07]	0,11[0,03]	0,08[0,02]
SJ	0,3[0,1]	0,22[0,09]	0,19[0,07]	0,1[0,03]	0,08[0,02]
SC	0,3[0,12]	0,22[0,09]	0,2[0,07]	0,11[0,03]	0,08[0,02]
L11	0,37[0,16]	0,3[0,11]	0,23[0,07]	0,12[0,04]	0,09[0,02]
MLCV	0,34[0,18]	0,25[0,11]	0,21[0,09]	0,12[0,03]	0,08[0,03]
UCV	0,31[0,13]	0,25[0,11]	0,2[0,08]	0,11[0,03]	0,08[0,02]
BCV1	0,41[0,29]	0,24[0,1]	0,2[0,08]	0,1[0,03]	0,08[0,02]
BCV2	0,32[0,16]	0,24[0,1]	0,2[0,06]	0,11[0,03]	0,08[0,02]
CCV	0,32[0,15]	0,24[0,1]	0,2[0,06]	0,1[0,03]	0,08[0,02]
MCV	0,37[0,17]	0,25[0,12]	0,2[0,08]	0,11[0,03]	0,08[0,02]
TCV	0,34[0,15]	0,25[0,1]	0,21[0,08]	0,11[0,03]	0,09[0,02]
OSCV	0,32[0,12]	0,22[0,1]	0,19[0,07]	0,1[0,03]	0,08[0,02]
TRG	0,27[0,11]	0,2[0,09]	0,18[0,06]	0,1[0,03]	0,08[0,02]
TRE	0,26[0,1]	0,2[0,09]	0,18[0,06]	0,1[0,03]	0,07[0,02]

Fonte: Elaborada pelo autor (2020).

Tabela 28 – Área do erro para distribuição D2a.

Seletores	Amostras				
	25 Mediana[IQR]	50 Mediana[IQR]	100 Mediana[IQR]	500 Mediana[IQR]	1000 Mediana[IQR]
SV	0,3[0,12]	0,24[0,08]	0,21[0,06]	0,12[0,03]	0,1[0,02]
SVM1	0,3[0,11]	0,24[0,09]	0,19[0,05]	0,11[0,02]	0,09[0,02]
SVM2	0,3[0,13]	0,24[0,09]	0,2[0,05]	0,11[0,02]	0,08[0,02]
SJ	0,29[0,11]	0,24[0,08]	0,2[0,05]	0,11[0,02]	0,09[0,02]
SC	0,3[0,12]	0,24[0,09]	0,19[0,05]	0,11[0,02]	0,09[0,02]
L11	0,33[0,13]	0,28[0,1]	0,22[0,06]	0,13[0,03]	0,1[0,02]
MLCV	0,34[0,14]	0,27[0,11]	0,24[0,08]	0,13[0,07]	0,11[0,05]
UCV	0,31[0,13]	0,25[0,09]	0,21[0,06]	0,11[0,02]	0,09[0,02]
BCV1	0,35[0,14]	0,25[0,12]	0,2[0,05]	0,11[0,02]	0,09[0,02]
BCV2	0,32[0,15]	0,26[0,09]	0,2[0,07]	0,12[0,03]	0,09[0,02]
CCV	0,31[0,14]	0,25[0,09]	0,2[0,06]	0,11[0,03]	0,09[0,02]
MCV	0,37[0,16]	0,26[0,13]	0,21[0,05]	0,11[0,02]	0,09[0,02]
TCV	0,31[0,12]	0,25[0,08]	0,21[0,06]	0,11[0,02]	0,09[0,02]
OSCV	0,31[0,11]	0,25[0,08]	0,2[0,05]	0,11[0,02]	0,09[0,02]
TRG	0,27[0,13]	0,21[0,08]	0,19[0,05]	0,11[0,02]	0,08[0,02]
TRE	0,27[0,11]	0,22[0,08]	0,19[0,05]	0,11[0,02]	0,08[0,02]

Fonte: Elaborada pelo autor (2020).

Tabela 29 – Área do erro para distribuição D2b.

Seletores	Amostras				
	25 Mediana[IQR]	50 Mediana[IQR]	100 Mediana[IQR]	500 Mediana[IQR]	1000 Mediana[IQR]
SV	0,41[0,1]	0,34[0,1]	0,3[0,07]	0,18[0,04]	0,15[0,02]
SVM1	0,42[0,12]	0,36[0,1]	0,3[0,08]	0,18[0,04]	0,15[0,02]
SVM2	0,37[0,12]	0,31[0,09]	0,26[0,07]	0,15[0,04]	0,12[0,02]
SJ	0,36[0,14]	0,29[0,11]	0,24[0,08]	0,12[0,03]	0,1[0,03]
SC	0,41[0,13]	0,34[0,1]	0,28[0,08]	0,17[0,04]	0,14[0,02]
L11	0,38[0,16]	0,29[0,1]	0,25[0,07]	0,13[0,03]	0,1[0,03]
MLCV	0,4[0,17]	0,33[0,13]	0,27[0,09]	0,14[0,04]	0,1[0,03]
UCV	0,41[0,19]	0,34[0,17]	0,29[0,15]	0,15[0,08]	0,11[0,05]
BCV1	0,52[0,13]	0,43[0,18]	0,27[0,15]	0,12[0,03]	0,1[0,03]
BCV2	0,41[0,15]	0,31[0,12]	0,25[0,07]	0,13[0,04]	0,1[0,03]
CCV	0,4[0,15]	0,3[0,13]	0,25[0,08]	0,13[0,04]	0,1[0,03]
MCV	0,53[0,19]	0,36[0,17]	0,27[0,1]	0,12[0,04]	0,1[0,03]
TCV	0,4[0,15]	0,32[0,12]	0,25[0,08]	0,12[0,04]	0,1[0,03]
OSCV	0,4[0,15]	0,3[0,11]	0,25[0,09]	0,13[0,03]	0,1[0,03]
TRG	0,34[0,13]	0,28[0,12]	0,23[0,09]	0,12[0,04]	0,09[0,03]
TRE	0,34[0,13]	0,28[0,11]	0,23[0,08]	0,12[0,03]	0,09[0,03]

Fonte: Elaborada pelo autor (2020).

Tabela 30 – Área do erro para distribuição D2c.

Seletores	Amostras				
	25 Mediana[IQR]	50 Mediana[IQR]	100 Mediana[IQR]	500 Mediana[IQR]	1000 Mediana[IQR]
SV	0,44[0,08]	0,39[0,06]	0,34[0,05]	0,24[0,03]	0,2[0,02]
SVM1	0,44[0,13]	0,36[0,1]	0,29[0,07]	0,17[0,03]	0,14[0,03]
SVM2	0,44[0,15]	0,35[0,09]	0,28[0,07]	0,16[0,03]	0,13[0,03]
SJ	0,43[0,11]	0,35[0,08]	0,28[0,06]	0,16[0,03]	0,12[0,03]
SC	0,44[0,14]	0,36[0,1]	0,29[0,07]	0,17[0,03]	0,13[0,03]
L11	0,47[0,15]	0,38[0,09]	0,3[0,08]	0,17[0,03]	0,13[0,03]
MLCV	0,45[0,11]	0,37[0,12]	0,29[0,08]	0,18[0,06]	0,14[0,04]
UCV	0,49[0,2]	0,41[0,19]	0,32[0,13]	0,2[0,12]	0,16[0,07]
BCV1	0,5[0,14]	0,5[0,11]	0,33[0,19]	0,16[0,03]	0,13[0,03]
BCV2	0,49[0,17]	0,37[0,12]	0,29[0,09]	0,17[0,03]	0,13[0,03]
CCV	0,49[0,15]	0,37[0,14]	0,29[0,09]	0,16[0,03]	0,13[0,03]
MCV	0,55[0,16]	0,47[0,21]	0,31[0,12]	0,16[0,04]	0,12[0,03]
TCV	0,44[0,12]	0,36[0,09]	0,28[0,08]	0,16[0,03]	0,13[0,03]
OSCV	0,48[0,14]	0,37[0,11]	0,29[0,08]	0,16[0,03]	0,13[0,03]
TRG	0,41[0,1]	0,32[0,09]	0,27[0,06]	0,15[0,03]	0,12[0,03]
TRE	0,41[0,11]	0,33[0,08]	0,27[0,06]	0,15[0,03]	0,12[0,03]

Fonte: Elaborada pelo autor (2020).

Tabela 31 – Área do erro para distribuição D3a.

Seletores	Amostras				
	25 Mediana[IQR]	50 Mediana[IQR]	100 Mediana[IQR]	500 Mediana[IQR]	1000 Mediana[IQR]
SV	0,62[0,06]	0,53[0,05]	0,46[0,04]	0,31[0,03]	0,26[0,02]
SVM1	0,76[0,07]	0,71[0,05]	0,65[0,04]	0,49[0,02]	0,42[0,02]
SVM2	0,55[0,08]	0,46[0,06]	0,39[0,05]	0,25[0,04]	0,21[0,02]
SJ	0,44[0,13]	0,33[0,11]	0,26[0,08]	0,14[0,03]	0,11[0,03]
SC	0,74[0,06]	0,69[0,05]	0,62[0,04]	0,46[0,02]	0,39[0,02]
L11	0,42[0,15]	0,3[0,12]	0,24[0,08]	0,13[0,03]	0,1[0,03]
MLCV	0,46[0,16]	0,36[0,17]	0,32[0,11]	0,16[0,05]	0,11[0,03]
UCV	0,53[0,3]	0,39[0,43]	0,31[0,21]	0,15[0,05]	0,11[0,03]
BCV1	0,88[0,09]	0,48[0,5]	0,27[0,1]	0,14[0,03]	0,1[0,03]
BCV2	0,51[0,48]	0,34[0,14]	0,26[0,09]	0,14[0,03]	0,11[0,03]
CCV	0,51[0,46]	0,34[0,14]	0,26[0,09]	0,14[0,03]	0,1[0,03]
MCV	0,89[0,19]	0,38[0,17]	0,29[0,09]	0,14[0,03]	0,11[0,03]
TCV	0,61[0,38]	0,4[0,12]	0,3[0,09]	0,15[0,04]	0,12[0,03]
OSCV	0,5[0,36]	0,32[0,12]	0,25[0,08]	0,14[0,03]	0,1[0,03]
TRG	0,38[0,13]	0,28[0,12]	0,23[0,08]	0,13[0,03]	0,1[0,02]
TRE	0,38[0,13]	0,27[0,12]	0,23[0,08]	0,13[0,03]	0,1[0,02]

Fonte: Elaborada pelo autor (2020).

Tabela 32 – Área do erro para distribuição D3b.

Seletores	Amostras				
	25 Mediana[IQR]	50 Mediana[IQR]	100 Mediana[IQR]	500 Mediana[IQR]	1000 Mediana[IQR]
SV	1,16[0,03]	1,09[0,02]	1,02[0,02]	0,84[0,01]	0,76[0,01]
SVM1	1,32[0,02]	1,27[0,01]	1,21[0,01]	1,05[0,01]	0,98[0,01]
SVM2	1,08[0,03]	1,01[0,03]	0,93[0,02]	0,75[0,01]	0,67[0,01]
SJ	0,72[0,09]	0,58[0,07]	0,46[0,07]	0,24[0,03]	0,17[0,03]
SC	1,3[0,02]	1,25[0,01]	1,18[0,01]	1,02[0,01]	0,94[0,01]
L11	0,78[0,05]	0,63[0,06]	0,5[0,05]	0,27[0,03]	0,2[0,03]
MLCV	0,48[0,22]	0,38[0,13]	0,32[0,1]	0,19[0,05]	0,14[0,04]
UCV	0,6[0,29]	0,43[0,18]	0,35[0,18]	0,21[0,1]	0,17[0,07]
BCV1	1,4[0,02]	1,38[0,95]	0,32[0,13]	0,17[0,04]	0,13[0,03]
BCV2	0,48[0,23]	0,39[0,14]	0,3[0,09]	0,17[0,03]	0,13[0,03]
CCV	0,48[0,25]	0,39[0,13]	0,29[0,09]	0,17[0,03]	0,13[0,03]
MCV	1,44[0,77]	0,45[0,27]	0,32[0,13]	0,17[0,04]	0,13[0,03]
TCV	0,47[0,19]	0,38[0,14]	0,3[0,1]	0,17[0,04]	0,13[0,03]
OSCV	0,47[0,2]	0,38[0,12]	0,3[0,08]	0,17[0,03]	0,14[0,02]
TRG	0,4[0,13]	0,34[0,11]	0,27[0,08]	0,16[0,04]	0,13[0,03]
TRE	0,42[0,14]	0,35[0,1]	0,27[0,08]	0,16[0,03]	0,13[0,03]

Fonte: Elaborada pelo autor (2020).

Tabela 33 – Área do erro para distribuição D3c.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
SV	0,35[0,08]	0,29[0,05]	0,26[0,04]	0,19[0,02]	0,17[0,02]
SVM1	0,39[0,08]	0,34[0,05]	0,3[0,03]	0,23[0,02]	0,2[0,01]
SVM2	0,32[0,07]	0,27[0,05]	0,24[0,04]	0,17[0,02]	0,15[0,02]
SJ	0,33[0,08]	0,27[0,07]	0,23[0,05]	0,14[0,02]	0,11[0,02]
SC	0,38[0,07]	0,32[0,04]	0,29[0,03]	0,22[0,02]	0,19[0,01]
L11	0,31[0,07]	0,26[0,06]	0,21[0,05]	0,13[0,03]	0,1[0,02]
MLCV	0,34[0,09]	0,26[0,09]	0,23[0,06]	0,14[0,04]	0,11[0,05]
UCV	0,38[0,13]	0,28[0,08]	0,23[0,06]	0,13[0,03]	0,1[0,02]
BCV1	0,46[0,1]	0,39[0,06]	0,36[0,05]	0,16[0,04]	0,12[0,03]
BCV2	0,38[0,16]	0,3[0,1]	0,23[0,07]	0,13[0,03]	0,1[0,02]
CCV	0,41[0,15]	0,31[0,12]	0,23[0,06]	0,13[0,03]	0,1[0,02]
MCV	0,52[0,11]	0,43[0,08]	0,38[0,12]	0,16[0,04]	0,11[0,03]
TCV	0,38[0,13]	0,3[0,1]	0,23[0,07]	0,13[0,03]	0,1[0,02]
OSCV	0,41[0,09]	0,33[0,06]	0,24[0,07]	0,13[0,03]	0,1[0,02]
TRG	0,28[0,07]	0,24[0,06]	0,19[0,05]	0,12[0,03]	0,1[0,02]
TRE	0,28[0,06]	0,24[0,06]	0,19[0,05]	0,12[0,03]	0,1[0,02]

Fonte: Elaborada pelo autor (2020).

Tabela 34 – Área do erro para distribuição D4a.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
SV	0,85[0,06]	0,79[0,06]	0,73[0,04]	0,6[0,02]	0,54[0,01]
SVM1	0,97[0,04]	0,93[0,03]	0,88[0,02]	0,75[0,01]	0,69[0,01]
SVM2	0,78[0,06]	0,73[0,06]	0,67[0,04]	0,53[0,01]	0,48[0,02]
SJ	0,62[0,11]	0,54[0,09]	0,43[0,07]	0,22[0,05]	0,17[0,04]
SC	0,95[0,04]	0,91[0,03]	0,86[0,03]	0,72[0,01]	0,66[0,01]
L11	0,56[0,13]	0,46[0,09]	0,37[0,07]	0,2[0,04]	0,16[0,03]
MLCV	0,57[0,18]	0,49[0,15]	0,39[0,14]	0,26[0,08]	0,25[0,09]
UCV	0,6[0,17]	0,52[0,15]	0,41[0,1]	0,36[0,05]	0,27[0,04]
BCV1	1,05[0,04]	1,02[0,05]	0,43[0,26]	0,19[0,04]	0,15[0,03]
BCV2	0,58[0,22]	0,49[0,16]	0,38[0,09]	0,19[0,05]	0,16[0,04]
CCV	0,59[0,28]	0,5[0,15]	0,37[0,09]	0,19[0,05]	0,16[0,04]
MCV	1,08[0,05]	0,74[0,54]	0,4[0,12]	0,19[0,05]	0,15[0,04]
TCV	0,56[0,17]	0,48[0,12]	0,37[0,11]	0,2[0,05]	0,17[0,04]
OSCV	0,58[0,15]	0,48[0,12]	0,36[0,1]	0,2[0,05]	0,17[0,03]
TRG	0,51[0,12]	0,43[0,11]	0,34[0,09]	0,19[0,04]	0,15[0,03]
TRE	0,51[0,13]	0,44[0,1]	0,34[0,09]	0,19[0,04]	0,15[0,03]

Fonte: Elaborada pelo autor (2020).

Tabela 35 – Área do erro para distribuição D4b.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
SV	0,37[0,08]	0,33[0,05]	0,28[0,04]	0,2[0,02]	0,17[0,02]
SVM1	0,42[0,07]	0,37[0,05]	0,34[0,02]	0,26[0,01]	0,23[0,01]
SVM2	0,35[0,1]	0,3[0,06]	0,25[0,05]	0,17[0,02]	0,15[0,02]
SJ	0,35[0,09]	0,3[0,07]	0,24[0,07]	0,15[0,03]	0,11[0,03]
SC	0,41[0,07]	0,36[0,05]	0,33[0,02]	0,25[0,01]	0,22[0,01]
L1I	0,35[0,12]	0,28[0,09]	0,23[0,08]	0,14[0,03]	0,1[0,03]
MLCV	0,41[0,15]	0,33[0,14]	0,27[0,1]	0,15[0,04]	0,11[0,03]
UCV	0,42[0,1]	0,34[0,14]	0,26[0,1]	0,15[0,05]	0,11[0,03]
BCV1	0,47[0,11]	0,41[0,07]	0,38[0,06]	0,15[0,04]	0,1[0,03]
BCV2	0,46[0,14]	0,4[0,09]	0,29[0,15]	0,15[0,04]	0,11[0,03]
CCV	0,45[0,12]	0,4[0,08]	0,3[0,14]	0,14[0,03]	0,11[0,03]
MCV	0,51[0,16]	0,45[0,12]	0,39[0,07]	0,15[0,04]	0,11[0,03]
TCV	0,53[0,11]	0,43[0,05]	0,31[0,12]	0,16[0,03]	0,12[0,03]
OSCV	0,42[0,06]	0,36[0,06]	0,27[0,09]	0,14[0,03]	0,1[0,03]
TRG	0,32[0,1]	0,26[0,08]	0,22[0,08]	0,14[0,03]	0,1[0,02]
TRE	0,32[0,1]	0,26[0,08]	0,22[0,08]	0,14[0,03]	0,1[0,02]

Fonte: Elaborada pelo autor (2020).

Tabela 36 – Área do erro para distribuição D4c.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
SV	1,46[0,02]	1,42[0,01]	1,38[0,02]	1,25[0,01]	1,18[0,01]
SVM1	1,51[0,04]	1,5[0,01]	1,49[0,01]	1,45[0,01]	1,41[0,01]
SVM2	1,42[0,03]	1,37[0,02]	1,32[0,02]	1,17[0,01]	1,1[0,01]
SJ	1,32[0,04]	1,11[0,04]	0,9[0,04]	0,51[0,02]	0,37[0,02]
SC	1,51[0,05]	1,5[0,01]	1,48[0,01]	1,43[0,01]	1,4[0,01]
L1I	1,44[0,21]	1,41[0,1]	1,31[0,05]	0,72[0,01]	0,54[0,01]
MLCV	0,5[0,12]	0,41[0,1]	0,33[0,09]	0,21[0,05]	0,17[0,04]
UCV	0,49[0,13]	0,4[0,12]	0,31[0,08]	0,18[0,03]	0,14[0,02]
BCV1	1,52[0,03]	1,51[0,03]	0,44[1,15]	0,17[0,03]	0,14[0,02]
BCV2	0,5[0,15]	0,4[0,1]	0,3[0,06]	0,18[0,03]	0,14[0,02]
CCV	0,51[0,24]	0,4[0,1]	0,3[0,07]	0,17[0,03]	0,14[0,02]
MCV	1,53[0,04]	0,64[1,06]	0,36[0,11]	0,17[0,03]	0,14[0,02]
TCV	0,5[0,13]	0,41[0,11]	0,3[0,08]	0,17[0,03]	0,14[0,01]
OSCV	0,52[0,13]	0,4[0,11]	0,3[0,07]	0,17[0,03]	0,14[0,02]
TRG	0,45[0,11]	0,38[0,09]	0,28[0,06]	0,16[0,04]	0,13[0,01]
TRE	0,46[0,11]	0,38[0,09]	0,29[0,06]	0,16[0,03]	0,13[0,02]

Fonte: Elaborada pelo autor (2020).

D.2 KDE DE BANDA VARIÁVEL

Tabela 37 – Área do erro para distribuição D1a.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
AKDE	0,28[0,17]	0,21[0,08]	0,20[0,06]	0,09[0,03]	0,07[0,03]
BKDE	0,24[0,14]	0,16[0,09]	0,15[0,05]	0,08[0,03]	0,06[0,02]
VKDE	0,22[0,15]	0,19[0,10]	0,14[0,08]	0,08[0,03]	0,05[0,02]
ROIKDE	0,22[0,14]	0,19[0,09]	0,14[0,07]	0,08[0,03]	0,05[0,02]

Fonte: Elaborada pelo autor (2020).

Tabela 38 – Área do erro para distribuição D1b.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
AKDE	0,41[0,23]	0,22[0,09]	0,18[0,16]	0,11[0,03]	0,08[0,02]
BKDE	0,35[0,24]	0,3[0,1]	0,2[0,04]	0,09[0,03]	0,1[0,01]
VKDE	0,26[0,27]	0,23[0,1]	0,21[0,06]	0,09[0,01]	0,09[0,01]
ROIKDE	0,24[0,29]	0,22[0,1]	0,19[0,05]	0,09[0,01]	0,09[0,01]

Fonte: Elaborada pelo autor (2020).

Tabela 39 – Área do erro para distribuição D1c.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
AKDE	0,34[0,18]	0,26[0,14]	0,22[0,07]	0,12[0,02]	0,09[0,02]
BKDE	0,34[0,19]	0,23[0,12]	0,19[0,05]	0,12[0,03]	0,09[0,02]
VKDE	0,26[0,19]	0,2[0,11]	0,17[0,05]	0,11[0,04]	0,08[0,02]
ROIKDE	0,26[0,19]	0,2[0,11]	0,17[0,05]	0,12[0,03]	0,08[0,02]

Fonte: Elaborada pelo autor (2020).

Tabela 40 – Área do erro para distribuição D2a.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
AKDE	0,58[0,33]	0,38[0,16]	0,27[0,07]	0,15[0,04]	0,12[0,03]
BKDE	0,41[0,21]	0,3[0,09]	0,24[0,09]	0,14[0,03]	0,11[0,03]
VKDE	0,4[0,14]	0,28[0,13]	0,22[0,09]	0,13[0,05]	0,1[0,03]
ROIKDE	0,38[0,16]	0,28[0,13]	0,22[0,1]	0,13[0,05]	0,1[0,03]

Fonte: Elaborada pelo autor (2020).

Tabela 41 – Área do erro para distribuição D2b.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
AKDE	0,57[0,24]	0,42[0,21]	0,35[0,12]	0,19[0,05]	0,14[0,03]
BKDE	0,29[0,12]	0,24[0,12]	0,18[0,08]	0,12[0,03]	0,1[0,02]
VKDE	0,3[0,14]	0,27[0,17]	0,23[0,09]	0,14[0,04]	0,12[0,03]
ROIKDE	0,3[0,17]	0,27[0,17]	0,23[0,09]	0,15[0,04]	0,11[0,03]

Fonte: Elaborada pelo autor (2020).

Tabela 42 – Área do erro para distribuição D2c.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
AKDE	0,65[0,12]	0,45[0,09]	0,35[0,1]	0,19[0,04]	0,14[0,02]
BKDE	0,5[0,1]	0,39[0,1]	0,32[0,09]	0,17[0,04]	0,14[0,02]
VKDE	0,5[0,14]	0,37[0,1]	0,29[0,1]	0,16[0,04]	0,13[0,03]
ROIKDE	0,47[0,16]	0,36[0,1]	0,29[0,1]	0,16[0,03]	0,13[0,02]

Fonte: Elaborada pelo autor (2020).

Tabela 43 – Área do erro para distribuição D3a.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
AKDE	0,81[0,27]	0,7[0,24]	0,5[0,16]	0,26[0,07]	0,19[0,03]
BKDE	0,56[0,12]	0,46[0,08]	0,35[0,11]	0,22[0,05]	0,17[0,02]
VKDE	0,52[0,16]	0,45[0,08]	0,36[0,13]	0,2[0,07]	0,17[0,03]
ROIKDE	0,5[0,17]	0,43[0,07]	0,34[0,12]	0,19[0,06]	0,16[0,03]

Fonte: Elaborada pelo autor (2020).

Tabela 44 – Área do erro para distribuição D3b.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
AKDE	0,38[0,2]	0,3[0,16]	0,26[0,07]	0,15[0,06]	0,11[0,03]
BKDE	0,41[0,07]	0,35[0,04]	0,28[0,08]	0,14[0,03]	0,11[0,02]
VKDE	0,45[0,07]	0,38[0,07]	0,32[0,09]	0,16[0,06]	0,12[0,04]
ROIKDE	0,43[0,08]	0,36[0,07]	0,31[0,08]	0,17[0,06]	0,12[0,04]

Fonte: Elaborada pelo autor (2020).

Tabela 45 – Área do erro para distribuição D3c.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
AKDE	0,57[0,09]	0,43[0,09]	0,35[0,07]	0,17[0,02]	0,13[0,02]
BKDE	1,23[0,03]	1,08[0,03]	0,89[0,03]	0,5[0,02]	0,37[0,02]
VKDE	0,54[0,12]	0,36[0,12]	0,29[0,07]	0,16[0,03]	0,13[0,03]
ROIKDE	0,53[0,12]	0,36[0,1]	0,29[0,08]	0,16[0,03]	0,13[0,03]

Fonte: Elaborada pelo autor (2020).

Tabela 46 – Área do erro para distribuição D4a.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
AKDE	0,49[0,14]	0,36[0,06]	0,29[0,05]	0,17[0,04]	0,15[0,02]
BKDE	0,6[0,1]	0,45[0,07]	0,37[0,05]	0,23[0,03]	0,2[0,02]
VKDE	0,5[0,08]	0,37[0,06]	0,31[0,05]	0,21[0,05]	0,18[0,03]
ROIKDE	0,47[0,09]	0,37[0,04]	0,3[0,05]	0,21[0,04]	0,17[0,02]

Fonte: Elaborada pelo autor (2020).

Tabela 47 – Área do erro para distribuição D4b.

Seletores	Amostras				
	25	50	100	500	1000
	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]	Mediana[IQR]
AKDE	0,58[0,18]	0,42[0,12]	0,37[0,09]	0,24[0,06]	0,19[0,04]
BKDE	0,64[0,12]	0,51[0,09]	0,39[0,07]	0,24[0,02]	0,19[0,02]
VKDE	0,62[0,17]	0,48[0,08]	0,39[0,08]	0,25[0,06]	0,18[0,02]
ROIKDE	0,61[0,12]	0,48[0,1]	0,4[0,08]	0,22[0,03]	0,16[0,02]

Fonte: Elaborada pelo autor (2020).

Tabela 48 – Área do erro para distribuição D4c.

Seletores	Amostras				
	25 Mediana[IQR]	50 Mediana[IQR]	100 Mediana[IQR]	500 Mediana[IQR]	1000 Mediana[IQR]
AKDE	0,52[0,1]	0,4[0,05]	0,34[0,04]	0,23[0,04]	0,21[0,03]
BKDE	1,58[0,03]	1,54[0,02]	1,48[0,03]	1,26[0,03]	1,12[0,03]
VKDE	0,57[0,08]	0,44[0,06]	0,39[0,06]	0,25[0,03]	0,22[0,03]
ROIKDE	0,56[0,09]	0,43[0,06]	0,39[0,06]	0,25[0,03]	0,22[0,03]

Fonte: Elaborada pelo autor (2020).

Tabela 49 – Desempenho do método AKDE em relação aos demais métodos.

Distribuição	Amostras														
	25			50			100			500			1000		
	M	I	P	M	I	P	M	I	P	M	I	P	M	I	P
D1a	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3
D1b	1	2	0	1	0	2	0	3	0	0	1	2	3	0	0
D1c	0	1	2	0	1	2	0	0	3	0	1	2	0	1	2
D2a	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3
D2b	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3
D2c	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3
D3a	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3
D3b	3	0	0	3	0	0	3	0	0	2	0	1	2	0	1
D3c	1	0	2	1	0	2	1	0	2	1	0	2	1	2	0
D4a	1	0	2	2	1	0	3	0	0	3	0	0	3	0	0
D4b	3	0	0	3	0	0	3	0	0	0	2	1	0	1	2
D4c	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0

Fonte: Elaborada pelo autor (2020).

Tabela 50 – Desempenho do método BKDE em relação aos demais métodos.

Distribuição	Amostras														
	25			50			100			500			1000		
	M	I	P	M	I	P	M	I	P	M	I	P	M	I	P
D1a	1	1	1	1	1	1	1	0	2	3	0	0	3	0	0
D1b	0	0	3	0	0	3	0	2	1	0	3	0	0	0	3
D1c	0	1	2	0	1	2	1	0	2	0	1	2	0	1	2
D2a	1	0	2	1	0	2	1	0	2	1	2	0	1	1	1
D2b	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
D2c	1	0	2	1	0	2	1	0	2	1	0	2	1	0	2
D3a	1	0	2	1	0	2	2	1	0	1	1	1	2	0	1
D3b	1	1	1	2	0	1	2	0	1	3	0	0	3	0	0
D3c	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3
D4a	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3
D4b	0	1	2	0	0	3	0	1	2	1	1	1	0	1	2
D4c	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3

Fonte: Elaborada pelo autor (2020).

Tabela 51 – Desempenho do método VKDE em relação aos demais métodos.

Distribuição	Amostras														
	25			50			100			500			1000		
	M	I	P	M	I	P	M	I	P	M	I	P	M	I	P
D1a	1	1	1	2	1	0	3	0	0	1	1	1	1	0	2
D1b	1	2	0	2	0	1	0	2	1	1	1	1	1	0	2
D1c	2	1	0	2	0	1	2	0	1	2	1	0	2	0	1
D2a	2	0	1	2	1	0	2	1	0	1	1	1	1	2	0
D2b	1	1	1	1	1	1	1	1	1	1	1	1	1	0	2
D2c	2	1	0	2	0	1	2	1	0	2	0	1	2	1	0
D3a	2	0	1	2	0	1	1	0	2	1	1	1	1	0	2
D3b	0	0	3	0	0	3	0	0	3	1	0	2	1	0	2
D3c	2	1	0	2	1	0	3	0	0	2	1	0	1	1	1
D4a	2	0	1	1	0	2	1	0	2	1	0	2	1	0	2
D4b	1	1	1	2	0	1	2	0	1	0	1	2	2	0	1
D4c	1	1	1	1	0	2	1	1	1	2	0	1	1	0	2

Fonte: Elaborada pelo autor (2020).