Luciano Valiensi Lima

Métodos Clássicos e Bayesianos de Estimação da Janela Ótima em Núcleo-Estimadores

Belo Horizonte – MG

07 / 05 / 2007

Luciano Valiensi Lima

Métodos Clássicos e Bayesianos de Estimação da Janela Ótima em Núcleo-Estimadores

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito à obtenção do título de Mestre em Estatística.

Orientadora: Prof. Dra. Sueli Aparecida Mingoti

Co-Orientador: Prof. Dr. Gregorio Saraiva Atuncar

Universidade Federal Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Belo Horizonte – MG

26 / 04 / 2007

Resumo

Neste trabalho é discutido o problema de estimação da função densidade de probabilidade através do núcleo-estimador, no caso univariado. São descritas algumas abordagens de núcleo-estimadores e suas propriedades, sob o enfoque de janela fixa e de janela variável. Consequentemente, também foram implementados alguns métodos, clássicos e bayesianos, de estimação da janela ótima. No contexto de janela fixa são abordados apenas métodos clássicos de estimação, mais especificamente metodologias *Plug-in* (CHIU, 1991; SHEATHER; JONES, 1991), uma vez que não foi encontrado nenhum método bayesiano na literatura de núcleo-estimadores para estimação da janela ótima neste contexto. Sob a óptica de janela variável são apresentados tanto métodos clássicos como métodos bayesianos (BREWER, 2000; GANGOPADHYAY; CHEUNG, 2002). O principal objetivo deste trabalho foi comparar o desempenho de algumas metodologias de estimação da janela ótima presentes na literatura em diferentes abordagens de núcleo-estimadores. Para as comparações foram realizadas diversas simulações para diferentes cenários, observando assim identificar vantagens, desvantagens ou características peculiares de cada metodologia, em termos erro de estimação. Também são apresentados exemplos de aplicação, com o intuito de demonstrar o uso das metodologias no contexto prático.

Palavras Chave: Núcleo-estimador, Janela váriavel, Métodos Clássicos e Bayesianos.

Abstract

In this dissertation the estimation of univariate density functions through kernel methodology is discussed considering fixed and variable kernel estimators. The main purpose was to compare the performance of some methodologies of data-based bandwidth selection considering different approach of kernel estimation. Some methods for bandwidth selection under classics and bayesian approach were implemented. For fixed bandwidth only Plug-in (CHIU, 1991; SHEATHER; JONES, 1991) methodology was considered. For variable bandwidth classical and bayesian (BREWER, 2000; GANGOPADHYAY; CHEUNG, 2002) methods were evaluated. The comparison was performed by using Monte Carlo simulation. Different scenarios were simulated with the purpose to identify the advantages and peculiar characteristics of each methodology in terms of error criteria. Some examples of application were presented to show how the methodologies can be used in practical context.

Keywords: Kernel density estimation, variable bandwidth, Classic and bayesian methods.

Sumário

Lista de Figuras p. vii

Lista de Tabelas p. xii

1 Introdução p. 1

1.1 Objetivos e Contribuições . . . . . p. 2

1.2 Organização . . . . . p. 3

2 Revisão da Literatura p. 4

2.1 Núcleo-Estimador Fixo . . . . . p. 4

2.2 Núcleo-Estimador Variável . . . . . p. 5

3 Núcleo-Estimador Univariado p. 8

3.1 Núcleo-Estimador . . . . . p. 8

3.2 Propriedades do Núcleo-Estimador Fixo . . . . . p. 8

3.2.1 Avaliação da Acurácia . . . . . p. 8

3.2.2 Escolha da Função Núcleo . . . . . p. 10

3.2.3 Escolha do Parâmetro de Suavidade . . . . . p. 11

3.3 Núcleo-Estimador Variável . . . . . p. 12

3.3.1 Núcleo-Estimador Local . . . . . p. 12

3.3.2 Núcleo-Estimador por Pontos Amostrais . . . . . p. 14

4 Métodos para a Estimação da Janela Ótima p. 17

4.1 Método Plug-in . . . . . p. 17

4.1.1 Abordagem de Chiu (1991) . . . . . p. 18

4.1.1.1 Definições e Resultados . . . . . p. 18

4.1.1.2 Metodologia Proposta . . . . . p. 19

4.1.2 Abordagem de Sheather e Jones (1991) . . . . . p. 20

4.2 Abordagem de Sain e Scott (1996) . . . . . p. 22

4.3 Metodologias Bayesianas . . . . . p. 25

4.3.1 Abordagem de Brewer (2000) . . . . . p. 25

4.3.2 Abordagem de Gangopadhyay e Cheung (2002) . . . . . p. 29

5 Simulações p. 33

5.1 Implementação Computacional . . . . . p. 33

5.1.1 Ox . . . . . p. 33

5.2 Modelos Simulados . . . . . p. 33

6 Resultados p. 39

6.1 $X \sim \text{Normal}(0,1)$ . . . . . p. 40

6.2 $X \sim \text{Gama}(4,2)$ . . . . . p. 44

6.3 $X \sim \text{Weibull}(1,6)$ . . . . . p. 48

6.4 $X \sim$ Qui-Quadrado(7) . . . . . p. 52

6.5 $X \sim \frac{1}{5}\text{Normal}(0,1) + \frac{1}{5}\text{Normal}\left(\frac{1}{4}, \frac{4}{9}\right) + \frac{3}{5}\text{Normal}\left(\frac{13}{12}, \frac{25}{81}\right)$ . . . . . p. 56

6.6 $X \sim \frac{1}{2}\text{Normal}(2,1) + \frac{1}{2}\text{Normal}(5,1)$ . . . . . p. 60

6.7 $X \sim \frac{1}{2}\text{Normal}(2,1) + \frac{1}{2}\text{Normal}(6,1)$ . . . . . p. 65

6.8 $X \sim \frac{1}{2}\text{Normal}(2,1) + \frac{1}{2}\text{Normal}(8,1)$ . . . . . p. 69

6.9 $X \sim \frac{1}{2}\text{Normal}(2,1) + \frac{1}{2}\text{Normal}(6,3)$ . . . . . p. 73

6.10 $X \sim \frac{1}{2}\text{Normal}(2,1) + \frac{1}{2}\text{Normal}(8,3)$ . . . . . p. 77

6.11 $X \sim \frac{4}{5}\text{Normal}(0,1) + \frac{1}{5}\text{Normal}\left(2, \frac{1}{25}\right)$ . . . . . p. 81

6.12 $X \sim \frac{1}{4}\text{Normal}(3,1) + \frac{1}{2}\text{Normal}\left(6, \frac{1}{2}\right) + \frac{1}{4}\text{Normal}(9,1)$ . . . . . p. 85

6.13 $X \sim \frac{2}{3}\text{Normal}(0,1) + \frac{1}{3}\text{Normal}\left(0, \frac{1}{100}\right)$ . . . . . p. 89

6.14 $X \sim \frac{4}{5}\text{Weibull}\left(\frac{1}{100}, 6\right) + \frac{1}{5}\text{Weibull}(1,6)$ . . . . . p. 93

6.15 Comentários . . . . . p. 97

7 Aplicações p. 98

7.1 Aplicação 1 . . . . . p. 99

7.2 Aplicação 2 . . . . . p. 103

8 Conclusões p. 106

8.1 Propostas futuras . . . . . p. 107

Apêndice A – Resultados da Simulação p. 108

A.1 $X \sim \text{Normal}(0,1)$ . . . . . p. 108

A.2 $X \sim \text{Gama}(4,2)$ . . . . . p. 111

A.3 $X \sim \text{Weibull}(1,6)$ . . . . . p. 113

A.4 $X \sim \text{Qui-Quadrado}(7)$ . . . . . p. 115

A.5 $X \sim Mistura(1)$ . . . . . p. 117

A.6 $X \sim Mistura(2)$ . . . . . p. 119

A.7 $X \sim Mistura(3)$ . . . . . p. 121

A.8 $X \sim Mistura(4)$ . . . . . p. 123

A.9 $X \sim Mistura(5)$ . . . . . p. 125

A.10 $X \sim Mistura(6)$ . . . . . p. 127

A.11 $X \sim Mistura(7)$ . . . . . p. 129

A.12 $X \sim Mistura(8)$ . . . . . p. 131

A.13 $X \sim Mistura(9)$ . . . . . p. 133

A.14 $X \sim Mistura(10)$ . . . . . p. 135

Referências p. 137

Lista de Figuras

1.1 Comportamento do núcleo-estimador para diferentes valores da janela h. p. 2

4.1 Comportamento da distribuição Gama-Invertida(d_1, d_2) para diferentes valores de d_1. . . . . p. 29

5.1 Estimativa da distribuição Normal(0,1) para diferentes distribuições *a priori* não informativas, $n = 100$. . . . . p. 38

5.2 Estimativa da distribuição Mistura(4) para diferentes distribuições *a priori* não informativas, $n = 100$. . . . . p. 38

6.1 Boxplot das estimativas da janela ótima para a distribuição Normal(0,1). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. p. 40

6.2 Boxplot do Erros Integrados ao estimar a distribuição Normal(0,1). . . p. 43

6.3 Boxplot das estimativas da janela ótima para a distribuição Gama(4,2). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. p. 44

6.4 Boxplot do Erros Integrados ao estimar a distribuição Gama(4,2) . . . p. 47

6.5 Boxplot das estimativas da janela ótima para a distribuição Weibull(1,6). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. p. 48

6.6 Boxplot do Erros Integrados ao estimar a distribuição Weibull(1,6) . . . p. 51

6.7 Boxplot das estimativas da janela ótima para a distribuição Qui-Quadrado(7). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. p. 52

6.8 Boxplot do Erros Integrados ao estimar a distribuição Qui-Quadrado(7) p. 55

6.9 Boxplot das estimativas da janela ótima para a Mistura(1). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. . . p. 56

6.10 Boxplot do Erros Integrados ao estimar a Mistura(1) . . . . . p. 59

6.11 Boxplot das estimativas da janela ótima para a Mistura(2). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. . . p. 60

6.12 Função característica empírica pra diferentes tamanhos amostrais. A linha contínua representa o ponto de corte $\frac{c}{n}$, em que $c = 3$. . . . . p. 61

6.13 Boxplot do Erros Integrados ao estimar a Mistura(2) . . . . . p. 64

6.14 Boxplot das estimativas da janela ótima para a Mistura(3). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. . . p. 65

6.15 Boxplot do Erros Integrados ao estimar a Mistura(3) . . . . . p. 68

6.16 Boxplot das estimativas da janela ótima para a Mistura(4). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. . . p. 69

6.17 Boxplot do Erros Integrados ao estimar a Mistura(4). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. . . . . p. 72

6.18 Boxplot das estimativas da janela ótima para a Mistura(5). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. . . p. 73

6.19 Boxplot do Erros Integrados ao estimar a Mistura(5) . . . . . p. 76

6.20 Boxplot das estimativas da janela ótima para a Mistura(6). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. . . p. 77

6.21 Boxplot do Erros Integrados ao estimar a Mistura(6) . . . . . p. 80

6.22 Boxplot das estimativas da janela ótima para a Mistura(7). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. . . p. 81

6.23 Boxplot do Erros Integrados ao estimar a Mistura(7) . . . . . p. 84

6.24 Boxplot das estimativas da janela ótima para a Mistura(8). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. . . p. 85

6.25 Boxplot do Erros Integrados ao estimar a Mistura(8) . . . . . p. 88

6.26 Boxplot das estimativas da janela ótima para a Mistura(9). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. . . p. 89

6.27 Boxplot do Erros Integrados ao estimar a Mistura(9) . . . . . p. 92

6.28 Boxplot das estimativas da janela ótima para a Mistura(10). A linha contínua representa a janela ótima teórica que minimiza o EQMIA. . . p. 93

6.29 Boxplot do Erros Integrados ao estimar a Mistura(10) . . . . . p. 96

7.1 Histograma para o tempo (em minutos) sucessivo entre erupções. . . . p. 100

7.2 Núcleo-estimador para o tempo (em minutos) sucessivo entre erupções. p. 102

7.3 Núcleo-estimador para o tempo (em minutos) sucessivo entre erupções. p. 102

7.4 Histograma para as despesas. . . . . p. 103

7.5 Núcleo-estimador para as despesas anuais de armazéns (em 1000 doláres). p. 105

7.6 Núcleo-estimador para as despesas anuais de armazéns (em 1000 doláres). p. 105

A.1 Estimativa da distribuição Normal(0,1) com $n = 30$. . . . . p. 108

A.2 Estimativa da distribuição Normal(0,1) com $n = 50$. . . . . p. 109

A.3 Estimativa da distribuição Normal(0,1) com $n = 100$. . . . . p. 109

A.4 Estimativa da distribuição Normal(0,1) com $n = 200$. . . . . p. 110

A.5 Estimativa da distribuição Gama(4,2) com $n = 30$. . . . . p. 111

A.6 Estimativa da distribuição Gama(4,2) com $n = 50$. . . . . p. 111

A.7 Estimativa da distribuição Gama(4,2) com $n = 100$. . . . . p. 112

A.8 Estimativa da distribuição Gama(4,2) com $n = 200$. . . . . p. 112

A.9 Estimativa da distribuição Weibull(1,6) com $n = 30$. . . . . p. 113

A.10 Estimativa da distribuição Weibull(1,6) com $n = 50$. . . . . p. 113

A.11 Estimativa da distribuição Weibull(1,6) com $n = 100$. . . . . p. 114

A.12 Estimativa da distribuição Weibull(1,6) com $n = 200$. . . . . p. 114

A.13 Estimativa da distribuição Qui-Quadrado(7) com $n = 30$. . . . . p. 115

A.14 Estimativa da distribuição Qui-Quadrado(7) com $n = 50$. . . . . p. 115

A.15 Estimativa da distribuição Qui-Quadrado(7) com $n = 100$. . . . . p. 116

A.16 Estimativa da distribuição Qui-Quadrado(7) com $n = 200$. . . . . p. 116

A.17 Estimativa da distribuição Mistura(1) com $n = 30$. . . . . p. 117

A.18 Estimativa da distribuição Mistura(1) com $n = 50$. . . . . p. 117

A.19 Estimativa da distribuição Mistura(1) com $n = 100$. . . . . p. 118

A.20 Estimativa da distribuição Mistura(1) com $n = 200$. . . . . p. 118

A.21 Estimativa da distribuição Mistura(2) com $n = 30$. . . . . p. 119

A.22 Estimativa da distribuição Mistura(2) com $n = 50$. . . . . p. 119

A.23 Estimativa da distribuição Mistura(2) com $n = 100$. . . . . p. 120

A.24 Estimativa da distribuição Mistura(2) com $n = 200$. . . . . p. 120

A.25 Estimativa da distribuição Mistura(3) com $n = 30$. . . . . p. 121

A.26 Estimativa da distribuição Mistura(3) com $n = 50$. . . . . p. 121

A.27 Estimativa da distribuição Mistura(3) com $n = 100$. . . . . p. 122

A.28 Estimativa da distribuição Mistura(3) com $n = 200$. . . . . p. 122

A.29 Estimativa da distribuição Mistura(4) com $n = 30$. . . . . p. 123

A.30 Estimativa da distribuição Mistura(4) com $n = 50$. . . . . p. 123

A.31 Estimativa da distribuição Mistura(4) com $n = 100$. . . . . p. 124

A.32 Estimativa da distribuição Mistura(4) com $n = 200$. . . . . p. 124

A.33 Estimativa da distribuição Mistura(5) com $n = 30$. . . . . p. 125

A.34 Estimativa da distribuição Mistura(5) com $n = 50$. . . . . p. 125

A.35 Estimativa da distribuição Mistura(5) com $n = 100$. . . . . p. 126

A.36 Estimativa da distribuição Mistura(5) com $n = 200$. . . . . p. 126

A.37 Estimativa da distribuição Mistura(6) com $n = 30$. . . . . p. 127

A.38 Estimativa da distribuição Mistura(6) com $n = 50$. . . . . p. 127

A.39 Estimativa da distribuição Mistura(6) com $n = 100$. . . . . p. 128

A.40 Estimativa da distribuição Mistura(6) com $n = 200$. . . . . p. 128

A.41 Estimativa da distribuição Mistura(7) com $n = 30$. . . . . p. 129

A.42 Estimativa da distribuição Mistura(7) com $n = 50$. . . . . p. 129

A.43 Estimativa da distribuição Mistura(7) com $n = 100$. . . . . p. 130

A.44 Estimativa da distribuição Mistura(7) com $n = 200$. . . . . p. 130

A.45 Estimativa da distribuição Mistura(8) com $n = 30$. . . . . p. 131

A.46 Estimativa da distribuição Mistura(8) com $n = 50$. . . . . p. 131

A.47 Estimativa da distribuição Mistura(8) com $n = 100$. . . . . p. 132

A.48 Estimativa da distribuição Mistura(8) com $n = 200$. . . . . p. 132

A.49 Estimativa da distribuição Mistura(9) com $n = 30$. . . . . p. 133

A.50 Estimativa da distribuição Mistura(9) com $n = 50$. . . . . p. 133

A.51 Estimativa da distribuição Mistura(9) com $n = 100$. . . . . p. 134

A.52 Estimativa da distribuição Mistura(9) com $n = 200$. . . . . p. 134

A.53 Estimativa da distribuição Mistura(10) com $n = 30$. . . . . p. 135

A.54 Estimativa da distribuição Mistura(10) com $n = 50$. . . . . p. 135

A.55 Estimativa da distribuição Mistura(10) com $n = 100$. . . . . p. 136

A.56 Estimativa da distribuição Mistura(10) com $n = 200$. . . . . p. 136

Lista de Tabelas

4.1 Possíveis regras para determinação de d_1 . . . . . p. 28

6.1 Estatísticas Descritivas para as estimativas da janela ótima para a distribuição Normal(0,1). . . . . p. 40

6.2 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 41

6.3 Estatísticas Descritivas para as estimativas da janela ótima para a distribuição Gama(4,2) . . . . . p. 44

6.4 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 45

6.5 Estatísticas Descritivas para as estimativas da janela ótima para a distribuição Weibull(1,6) . . . . . p. 48

6.6 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 49

6.7 Estatísticas Descritivas para as estimativas da janela ótima para a distribuição Qui-Quadrado(7) . . . . . p. 52

6.8 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 53

6.9 Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(1) . . . . . p. 56

6.10 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 57

6.11 Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(2) . . . . . p. 60

6.12 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 62

6.13 Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(3) . . . . . p. 65

6.14 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 66

6.15 Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(4) . . . . . p. 69

6.16 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 70

6.17 Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(5) . . . . . p. 73

6.18 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 74

6.19 Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(6) . . . . . p. 77

6.20 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 78

6.21 Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(7) . . . . . p. 82

6.22 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 82

6.23 Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(8) . . . . . p. 85

6.24 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 86

6.25 Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(9) . . . . . p. 89

6.26 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 90

6.27 Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(10) . . . . . p. 93

6.28 Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200). . . . . p. 94

7.1 Estatísticas Descritivas para o tempo (em minutos) entre erupções sucessivas. . . . . p. 99

7.2 Informações sobre as estimativas da janela ótima . . . . . p. 101

7.3 Estatísticas Descritivas para as despesas anuais com alimento ($em1000$). p. 103

7.4 Informações sobre as estimativas da janela ótima . . . . . p. 104

1 Introdução

Um problema comum nas diversas áreas do conhecimento reside na necessidade de inferir sobre um fenômeno de interesse de forma eficaz. Esta eficácia está associada ao conhecimento da estrutura probabilística geradora dos dados que, em geral, é desconhecida. Uma forma particular de se determinar tal estrutura é através da função densidade de probabilidade (fdp).

Existem distintas formas de se estimar a função densidade e uma possibilidade está na abordagem paramétrica. Nesta abordagem, assumimos que a densidade desconhecida pertence a uma família paramétrica de modelos, que satisfazem certas condições.

A robustez desta abordagem depende da escolha de um modelo paramétrico. Se o modelo designado for o verdadeiro ou próximo deste, as inferências a respeito da distribuição geradora dos dados serão plausíveis. No entanto, se o modelo for especificado incorretamente, as inferências poderão ser inadequadas, levando a interpretações errôneas do fenômeno em questão.

Uma forma alternativa de tratar o problema está na abordagem não-paramétrica. Nesta abordagem, as suposições que são feitas sobre a estrutura probabilística geradora dos dados são fracas ou inexistentes, ou seja, "os dados falam por si só".

Dentre as várias metodologias não-paramétricas, um importante método para a estimação de funções densidade de probabilidade é a do Núcleo-Estimador, também conhecido na literatura como suavização pelo método do núcleo.

O desempenho do núcleo-estimador depende essencialmente da escolha do parâmetro de suavidade ou janela. Este parâmetro, geralmente denotado por h, determina o grau de suavização a ser feita. Isto é, ele é responsável por quão rapidamente a função oscila.

A Figura 1 mostra o comportamento do núcleo-estimador para diferentes valores da janela h. A linha contínua representa a densidade de uma distribuição normal padrão, enquanto que a linha tracejada representa o núcleo-estimador baseado numa amostra

aleatória de tamanho 200 desta distribuição.

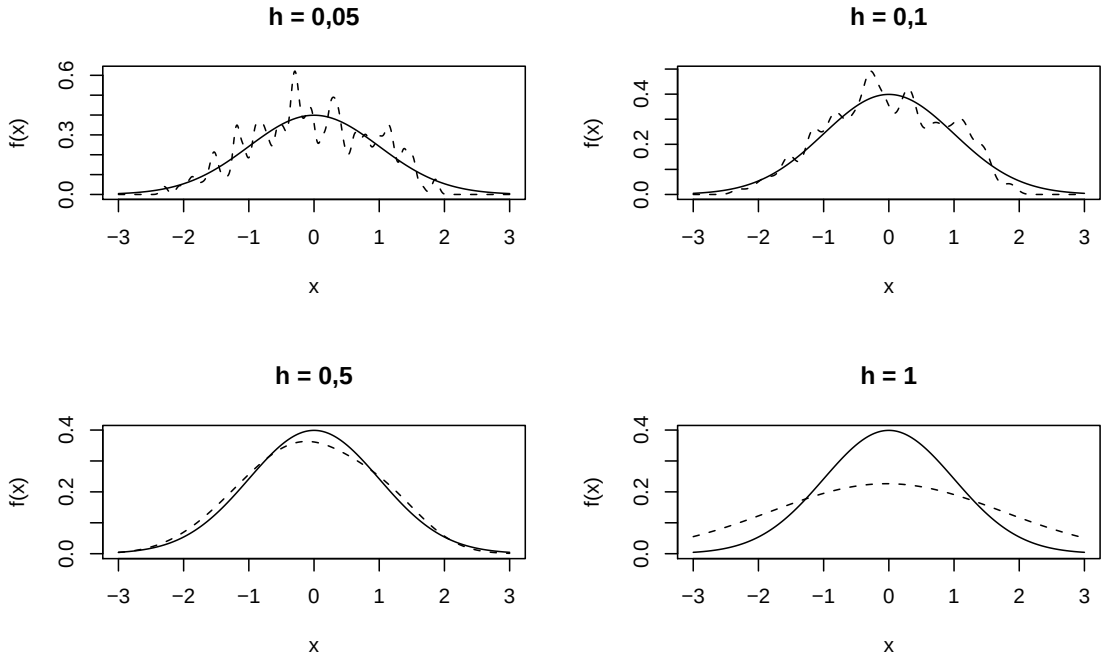

Figura 1.1: Comportamento do núcleo-estimador para diferentes valores da janela h.

Observe que, se h for muito pequeno, a estimativa da densidade apresentará muita oscilação, e se h for muito grande, a estimativa será super-suavizada, ocultando importantes características da função. Sendo assim, é necessário algum critério para encontrar h de tal modo a minimizar estes problemas, ou seja, encontrar uma janela ótima h_{opt}. Uma forma idealizada é através da minimização da discrepância (erro) entre a estimativa suavizada e a verdadeira densidade.

Nos últimos anos, o núcleo-estimador vem tendo um grande desenvolvimento, devido ao seu bom desempenho e ao número reduzido de parâmetros a ser ajustado em sua utilização. Como conseqüência, tem sido amplamente aplicado em diversos campos do conhecimento.

1.1 Objetivos e Contribuições

Na literatura podem ser encontradas diversas propostas de estimação da janela ótima, bem como aperfeiçoamentos do Núcleo-Estimador. Tendo em vista esta quantidade de trabalhos, esta dissertação tem como objetivo a implementação e a comparação de diferentes metodologias, clássicas e bayesianas, de estimação da janela ótima em diferentes

abordagens de núcleo-estimadores, visando assim identificar vantagens, desvantagens ou características peculiares de algumas metodologias presentes na literatura.

Vale salientar que o intuito desta dissertação é comparar as metodologias da forma em que elas foram propostas na literatura, fugindo do escopo estudos mais aprofundados e propostas de modificação de cada metodologia.

1.2 Organização

A presente dissertação está disposta do seguinte modo: no Capítulo 2, há uma breve revisão da literatura relativa a núcleo-estimadores para função densidade de probabilidade. No Capítulo 3, são apresentadas diferentes abordagens de núcleo-estimadores para o caso univariado, assim como suas respectivas propriedades. No Capítulo 4, estão descritos os métodos, clássicos e bayesianos, de estimação da janela ótima que serão abordados nesta dissertação. O Capítulo 5 contém a descrição dos procedimentos de simulação e alguns detalhes da implementação computacional. No Capítulo 6, são expostos os resultados obtidos em simulação. No Capítulo 7, são apresentados alguns exemplos de aplicação das metodologias estudadas, e o Capítulo 8 contém as conclusões e considerações finais do trabalho.

2 Revisão da Literatura

2.1 Núcleo-Estimador Fixo

Fix e Hodges (1951) foram os primeiros autores na literatura que propuseram as idéias básicas de núcleo-estimador, utilizando uma função núcleo $U(-1,1)$. Rosenblatt (1956) e Parzen (1962) estudaram a classe geral do núcleo-estimador univariado, também conhecido na literatura por núcleo-estimador fixo. A terminologia fixa foi adotada devido a janela h ser constante para todo $x \in \mathbb{R}$.

Rosenblatt (1956) mostrou que não existe estimador não-viciado para função densidade de probabilidade, o que motivou a procura de estimadores assintoticamente não-viciados. Parzen (1962) examinou as propriedades assintóticas do núcleo-estimador em termos do erro quadrático médio (EQM) e mostrou que, sob algumas condições, o núcleo-estimador é assintoticamente não-viciado e assintoticamente normal. Outros resultados sobre propriedades do núcleo-estimador podem ser encontrados em Bertrand-Retali (1978) e Devroye e Györfi (1985).

Epanechnikov (1969) fez um estudo da eficiência, em termos assintóticos, de diferentes funções núcleo, a fim de encontrar uma função núcleo ótima. O mesmo problema foi considerado por Bartlett (1963). Contudo, Silverman (1986) e Scott (1992) mostraram que o desempenho do núcleo-estimador, em termos de erro, é dominado essencialmente pela escolha do parâmetro de suavidade.

Um dos primeiros métodos automáticos de estimação da janela ótima foi o método *"plug-in"*, proposto por Woodroofe (1970) e Nadaraya (1974). No entanto, este método não foi implementado em termos práticos. Habbema *et al.* (1974) e Duin (1976) propuseram validação cruzada por verossimilhança para a escolha da janela ótima, baseados na minimização da estimativa da medida de divergência de Kullback-Leibler. Entretanto, este método pode produzir estimativas inconsistentes, como foi observado no trabalho de Schuster e Gregory (1981).

Rudemo (1982) e Bowman (1984), independentemente, desenvolveram o método de validação cruzada por mínimos quadrados (VCMQ). No entanto, o método possui alta variabilidade, o que compromete seu desempenho. Na tentativa de melhorar a metodologia proposta por Rudemo (1982) e Bowman (1984), Scott e Terrell (1987) propuseram o método de validação cruzada viciada (VCV), que mostrou ser mais estável que o VCMQ no sentido de possuir uma menor variância assintótica. Contudo, essa redução da variância implica num aumento do vício. Uma comparação teórica entre ambas as metodologias pode ser encontrada em Jones e Kappenman (1992).

Baseados nas idéias do método *"plug-in"*, Chiu (1991) e Sheather e Jones (1991) desenvolveram diferentes metodologias para a escolha da janela ótima e mostraram que, sob certas condições, suas propostas apresentam ótimos resultados.

Outras metodologias de estimação da janela ótima, além de outros estudos sobre núcleo-estimador fixo, podem ser encontradas na literatura. Para uma descrição mais completa sobre núcleo-estimador e suas propriedades, bem como outras informações, veja Silverman (1986), Scott (1992), Wand e Jones (1995), Jones *et al.* (1996), Simonoff (1996) e Bowman e Azzalini (1997).

2.2 Núcleo-Estimador Variável

Em geral, o núcleo-estimador fixo possui um desempenho insatisfatório em estimar densidades que exibem mudanças abruptas, multi-modalidade, forte assimetria , etc. (CACOULLOS, 1966; MINNOTTE, 1992). Wand *et al.* (1991) propuseram a utilização de transformações nos dados na tentativa de solucionar estes problemas. A estimativa da densidade é obtida utilizando transformação inversa.

Outras propostas na literatura para contornar estes problemas, em geral, são baseadas na idéia de variar a janela h ou a forma funcional do núcleo em diferentes partes da função. Essas propostas são conhecidas na literatura como núcleo-estimador com janela variável ou núcleo-estimador variável, respectivamente.

Loftsgaarden e Quesenberry (1965) introduziram o núcleo-estimador local, como uma forma particular do método dos k vizinhos mais próximos para a estimação de densidades. Nesta abordagem, a janela varia de acordo com o ponto em que a densidade será estimada. Uma desvantagem deste método é que a estimativa da densidade gerada pelo núcleo nem sempre integra 1. Uma discussão sobre o núcleo-estimador local pode ser encontrada em Schucany (1989) e Jones (1990). Terrell e Scott (1992) fizeram um estudo detalhado sobre

as propriedades gerais do núcleo-estimador local.

Hazelton (1999) interessado apenas na estimação pontual, utilizou o método *"bootstrap"* para a estimação da janela ótima do núcleo-estimador local. Com o enfoque na estimação global, Gangopadhyay e Cheung (2002) apresentaram uma metodologia bayesiana para a escolha da janela ótima.

Uma outra variação de núcleo-estimador com janela variável foi introduzida por Breiman *et al.* (1977). A proposta dos autores foi variar a janela de acordo com os pontos amostrais. Essa abordagem é conhecida na literatura estatística como núcleo-estimador adaptativo. Abramson (1982) propôs a regra da raiz quadrada (*"square root law"*) para a estimação das janelas e mostrou que a sua proposta reduz o vício pontual do estimador. Silverman (1986) propôs com detalhes uma forma de implementação para o estimador de Abramson (1982).

Terrell e Scott (1992) fizeram um estudo sobre algumas propriedades do núcleo-estimador adaptativo e propuseram uma explicação para o fato da abordagem apresentar o problema de não-localidade.

Sain e Scott (1996) estudaram alguns aspectos práticos e teóricos sobre o núcleo-estimador adaptativo através do núcleo-estimador adaptativo por blocos (SILVERMAN, 1982; SCOTT; SHEATHER, 1985). Além disso, Sain e Scott (1996) apresentaram um procedimento para a estimação das janelas ótimas, obtendo bons resultados.

Brewer (2000) propôs o primeiro procedimento bayesiano para a estimação das janelas ótimas, através de modelos gráficos bayesianos e validação cruzada por verossimilhança. Em seu trabalho, Brewer (2000) argumenta que a utilização de uma distribuição *"a priori "* apropriada e de uma função núcleo com suporte infinito contorna o problema de inconsistência da metodologia de validação cruzada por verossimilhança. No trabalho de Brewer (2000) é mostrado que o método apresentou bons resultados tanto em situações teóricas como em situações práticas.

Para maiores detalhes sobre núcleo-estimador variável, veja Silverman (1986), Terrell e Scott (1992), Jones (1990) e Sain (1994).

A extensão do núcleo-estimador fixo para o caso multivariado foi inicialmente considerado por Cacoullos (1966), que utilizou uma matriz diagonal de janelas, $H = hI$ para a estimação, em que h é o parâmetro de suavidade, constante para todas as variáveis, e I é uma matriz identidade $d \times d$. Devroye e Györfi (1985) examinaram algumas propriedades dessa proposta. Epanechnikov (1969) investigou a utilização de uma matriz diagonal com

diferentes janelas para cada variável, isto é, $H = diag(h_1, \ldots, h_d)$.

Fukunaga (1972) propôs utilizar transformações nos dados, de modo que não seja necessário trabalhar com a matriz completa de janelas; a mesma idéia foi abordada por Silverman (1986). Deheuvels (1977) foi o primeiro a discutir o caso em que a matriz de janelas é completa. Scott (1992) e Terrell e Scott (1992) fizeram um estudo detalhado sobre implementação e visualização de estimadores não-paramétricos de densidades multivariadas. Wand e Jones (1994) examinaram o método *"plug-in"* para dados multivariados.

Outras referências sobre núcleo-estimador multivariado e suas propriedades Sain (1994), Wand e Jones (1995), Simonoff (1996), Cavalcante (2004) e Duong (2004). Para mais informações sobre estimação das janelas, consulte Duong e Hazelton (2003), Duong e Hazelton (2005) e Zhang *et al.* (2006).

3 Núcleo-Estimador Univariado

3.1 Núcleo-Estimador

Dada uma amostra aleatória $X_1, X_2, \ldots, X_n$, de uma distribuição univariada contínua, com função densidade de probabilidade f, o núcleo-estimador de f avaliado no ponto x é definido por:

$$\begin{aligned} \hat{f}(x) &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - X_i\right), \end{aligned} \tag{3.1}$$

em que K é uma função, denominada núcleo. Note que $K_h(u) = h^{-1}K(u/h)$.

Em geral, a função núcleo é uma função densidade de probabilidade. Isto garante que $\hat{f}(x)$ também seja uma função densidade. Além disso, a forma aditiva de (3.1) assegura que $\hat{f}$ contenha as mesmas propriedades de continuidade e diferenciabilidade de K.

O parâmetro h ($h > 0$) é chamado de janela ou parâmetro de suavidade, pois controla a quantidade de suavização a ser feita na estimativa de f. O núcleo-estimador definido por (3.1) também é conhecido na literatura como núcleo-estimador fixo ou núcleo-estimador global, por considerar a janela h constante para todo ponto x.

3.2 Propriedades do Núcleo-Estimador Fixo

3.2.1 Avaliação da Acurácia

Existem diversas medidas para avaliar a acurácia de um estimador. Uma medida pontual comumente utilizada é o erro quadrático médio (EQM), que no caso do núcleo-estimador é definido por:

$$\begin{aligned}
\text{EQM}\left[\hat{f}(x)\right] &= \text{E}\left\{\left[f(x)-\hat{f}(x)\right]^2\right\} \\
&= \text{E}\left\{\hat{f}(x)-\text{E}\left[\hat{f}(x)\right]\right\}^2+\left\{\text{E}\left[\hat{f}(x)\right]-f(x)\right\}^2 \\
&= \text{Var}\left[\hat{f}(x)\right]+\left\{\text{Vício}\left[\hat{f}(x)\right]\right\}^2.
\end{aligned}$$

Outra medida muito utilizada para avaliar núcleo-estimador é o erro quadrático médio integrado (EQMI),

$$\begin{aligned}
\text{EQMI}(\hat{f}) &= \int \text{E}\left\{\left[f(x)-\hat{f}(x)\right]^2\right\}dx \\
&= \int \text{Var}\left[\hat{f}(x)\right]dx+\int\left\{\text{Vício}\left[\hat{f}(x)\right]\right\}^2dx,
\end{aligned}$$

que é uma medida global de discrepância de $\hat{f}$ com relação a f.

Em ambas as medidas existe a compensação entre variância e vício, ou seja, o aumento da variância implica na redução do vício, e vice-versa. Esse efeito é conhecido na literatura estatística como "balanceamento" (*trade-off*) entre variância e vício.

O erro quadrático integrado (EQI) também é amplamente utilizado para a avaliação do núcleo-estimador e é dado por:

$$\text{EQI}(\hat{f}) = \int\left[f(x)-\hat{f}(x)\right]^2dx.$$

Na prática, o EQM e o EQMI não podem ser calculados exatamente para o caso geral do núcleo-estimador. No entanto, sob algumas suposições, é possível obter aproximações de ambas as medidas.

Suponha que:

1. A função núcleo K seja simétrica em torno de zero;
2. $\int tK(t)dt = 0$;
3. $\int t^2K(t)dt = \sigma_K^2 > 0$;
4. A função desconhecida f tenha as duas primeiras derivadas contínuas;

5. $\int [f''(x)]^2 \, dx = R(f'') < \infty$.

Se $n \to \infty$, $h \to 0$ e $nh \to \infty$ para todo x, então pela expansão em série de Taylor, temos que

$$\text{Var}\left[\hat{f}(x)\right] = \frac{f(x)R(K)}{nh} + O(n^{-1}), \tag{3.2}$$

e

$$\text{Vício}\left[\hat{f}(x)\right] = \frac{h^2\sigma_K^2 f''(x)}{2} + O(h^4)$$

em que

$$R(K) = \int [K(u)]^2 \, du.$$

Combinando a variância e o vício ao quadrado, temos o EQM dado por:

$$\text{EQM}\left[\hat{f}(x)\right] = \frac{f(x)R(K)}{nh} + \frac{h^4\sigma_K^4 \left[f''(x)\right]^2}{4} + O(n^{-1}) + O(h^6). \tag{3.3}$$

Então, uma aproximação para (3.3) é dada por:

$$\text{EQMA}\left[\hat{f}(x)\right] = \frac{f(x)R(K)}{nh} + \frac{h^4\sigma_K^4 \left[f''(x)\right]^2}{4}. \tag{3.4}$$

Integrando (3.4) sobre toda a reta, temos uma aproximação para o erro quadrático médio integrado dado por:

$$\text{EQMIA}(\hat{f}) = \frac{R(K)}{nh} + \frac{h^4\sigma_K^4 R(f'')}{4}, \tag{3.5}$$

em que $R(f'')$ é dada pela expressão

$$R(f'') = \int [f''(u)]^2 \, du.$$

3.2.2 Escolha da Função Núcleo

Um aspecto-chave do núcleo-estimador está associado ao fato do seu desempenho, em termos do erro, não depender diretamente da forma funcional do núcleo. Diversos trabalhos na literatura demonstram o fato de que a qualidade do núcleo-estimador depende

essencialmente da escolha da janela h (SILVERMAN, 1986; SCOTT, 1992).

Alguns núcleos são apresentados a seguir.

$$\begin{aligned}
\text{Gaussiano:} \quad & \mathrm{K(u)} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right), -\infty < u < +\infty \\
\text{Epanechnikov:} \quad & \mathrm{K(u)} = \frac{3}{4}(1-u^2), -1 \leq u \leq 1 \\
\text{Biponderado:} \quad & \mathrm{K(u)} = \frac{15}{16}(1-u^2)^2, -1 \leq u \leq 1 \\
\text{Triponderado:} \quad & \mathrm{K(u)} = \frac{35}{32}(1-u^2)^3, -1 \leq u \leq 1 \\
\text{Uniforme:} \quad & \mathrm{K(u)} = \frac{1}{2}, -1 \leq u \leq 1
\end{aligned}$$

Nesta dissertação, a escolha da função núcleo estará restrita ao núcleo gaussiano. Para maiores informações sobre função núcleo, veja Silverman (1986) e Scott (1992).

3.2.3 Escolha do Parâmetro de Suavidade

Em geral, a escolha da janela h está relacionada à especificação (otimização) de alguma medida de desempenho, isto é, a escolha de h está vinculada à minimização de alguma medida de acurácia. Tomando como medida o EQMIA, a janela ótima será dada pela minimização da expressão (3.5) em relação a h.

Dessa forma, a janela que minimiza (3.5) é dada por:

$$h_{opt} = \left[\frac{R(K)}{\sigma_K^4 R(f'')}\right]^{1/5} n^{-1/5}, \tag{3.6}$$

que é denominada janela ótima.

Note que h_{opt} depende do tamanho da amostra, n, da função núcleo, K, e da densidade desconhecida, f, através da expressão $R(f'')$.

Substituindo (3.6) em (3.5), temos que

$$\inf_{h>0} \mathrm{EQMIA}(\hat{f}) = \frac{5}{4}\left\{\sigma_K^4 \left[R(K)\right]^4 R(f'')\right\}^{1/5} n^{-4/5},$$

é o menor EQMIA possível ao estimar f.

Na prática, a expressão (3.6) não pode ser calculada diretamente, pois depende da densidade f, que é desconhecida. Entretanto, existem alguns métodos para a estimação da janela ótima h_{opt}. Estes métodos serão abordados no Capítulo 4.

Para mais detalhes a respeito do núcleo-estimador, veja Silverman (1986), Scott (1992), Wand e Jones (1995) e Simonoff (1996).

3.3 Núcleo-Estimador Variável

Em muitas situações o núcleo-estimador fixo apresenta bons resultados. Contudo, em determinados problemas, tais como distribuições com caudas pesadas, distribuições assimétricas e distribuições multi-modais, o método tende a apresentar um desempenho insatisfatório.

No intuito de contornar estes problemas, pontualmente e globalmente, algumas modificações do núcleo-estimador foram desenvolvidas no decorrer dos anos. Essas modificações são baseadas na idéia de variar a janela h ou a forma funcional do núcleo em diferentes partes da função.

Nesta dissertação, serão abordadas as formulações em que existe variação da janela h em partes distintas da função. Essas abordagens do núcleo-estimador são conhecidas na literatura como núcleo-estimador com janela variável ou núcleo-estimador variável.

Nas próximas seções serão melhor detalhadas as abordagens utilizadas.

3.3.1 Núcleo-Estimador Local

O núcleo-estimador, $\hat{f}(x)$, definido em (3.1) depende de uma única janela h. Todavia, esta suposição pode não ser razoável em algumas situações em que a quantidade ótima de suavização a ser feita pode variar em todo o suporte da distribuição. Desse modo, uma extensão intuitiva de (3.1) seria dispor diferentes janelas, $h(x)$, para cada ponto x em que f está sendo estimada.

Sendo assim, temos o Núcleo-Estimador Local de f, avaliado no ponto x definido por:

$$\begin{aligned} \hat{f}_L(x) &= \frac{1}{nh(x)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h(x)}\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} K_{h(x)}\left(x - X_i\right), \end{aligned} \tag{3.7}$$

em que $K_{h(y)}(u) = [h(y)]^{-1} K(u/h(y))$. Note que, neste caso, a janela é uma função do ponto, x, em que f está sendo estimada.

Esta primeira variação do núcleo-estimador, conhecida como *"Baloon estimator"*, foi introduzida por Loftsgaarden e Quesenberry (1965) como uma forma particular do método dos k vizinhos mais próximos para a estimação de densidades.

Uma consequência do fato da janela depender de x é que, em geral, $\int \hat{f}_L(x)dx \neq 1$, ou seja, $\hat{f}_L$ não é uma função densidade de probabilidade.

Terrell e Scott (1992) provaram que, sob algumas suposições, é possível obter aproximações da variância e do vício ao quadrado. Assumindo que a segunda derivada f'' seja contínua e diferente de 0 no ponto x, e $\int |f''(x)|dx < \infty$, então

$$\text{Var}\left[\hat{f}_L(x)\right] \approx \frac{f(x)R(K)}{nh(x)} \tag{3.8}$$

e

$$\left\{\text{Vício}\left[\hat{f}_L(x)\right]\right\}^2 \approx \frac{1}{4}\left\{[h(x)]^2 f''(x)\right\}^2,$$

quando $nh(x) \to \infty$ e $h(x) \to 0$.

Nesta abordagem, a seleção da janela ótima é feita para cada ponto x, a fim de otimizar a quantidade de suavização. Minimizando o EQMA com respeito a $h(x)$, Terrell e Scott (1992) mostraram que a janela ótima no ponto x é dada por:

$$h_{opt}(x) = \left\{\frac{R(K)f(x)}{\sigma_K^4 \left[f''(x)\right]^2}\right\}^{1/5} n^{-1/5},$$

exceto nos pontos de inflexão (ROSENBLATT, 1956; TERRELL; SCOTT, 1992).

Outros resultados que Terrell e Scott (1992) obtiveram em seus trabalhos foram

$$\inf_{h(x)>0} \text{EQMIA}(\hat{f}_L) = \frac{5}{4}\left\{\sigma_K^4 \left[R(K)\right]^4 R(f^2 f'')\right\}^{1/5} n^{-4/5}$$

e

$$\left[R(f^2 f'')\right]^{1/5} \leq \left[R(f'')\right]^{1/5},$$

para todo f. Observe que a taxa de convergência de $\hat{f}_L$, com respeito a n, é idêntica à taxa encontrada em (3.7), $n^{-4/5}$. Ou seja, o núcleo-estimador local possui a mesma taxa

de convergência que o núcleo-estimador fixo, em relação ao tamanho da amostra.

Na prática, é necessário especificar a função h antes da suavização ser feita. Para maiores informações sobre esta abordagem, veja Jones (1990), Terrell e Scott (1992), Hazelton (1999).

3.3.2 Núcleo-Estimador por Pontos Amostrais

A segunda abordagem de núcleo-estimador variável segue um raciocínio similar ao do núcleo-estimador local, porém a variação da janela ocorre de acordo com os pontos amostrais. Conseqüentemente, cada observação da amostra possui uma janela respectiva. Esta versão de núcleo-estimador foi introduzida na literatura por Breiman *et al.* (1977) e é dada por:

$$\begin{aligned} \hat{f}_{PA}(x) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i}K\left(\frac{x-X_i}{h_i}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}K_{h_i}(x-X_i), \end{aligned} \tag{3.9}$$

em que h_i é uma função de X_i, ou seja $h_i = w(X_i)$.

A vantagem desta formulação em comparação com (3.7), descrita anteriormente, é que se a função núcleo for uma função densidade, então (3.9) também será uma densidade.

Assim como no núcleo-estimador local, é necessário escolher previamente a função w antes da suavização ser feita. Abramson (1982) sugeriu que $h_i \propto \tilde{f}(x_i)^{-1/2}$, em que $\tilde{f}$ é uma estimativa piloto gerada pelo núcleo-estimador fixo; dessa forma o vício pontual do estimador passa da ordem $O(h^2)$ para $O(h^4)$ (SILVERMAN, 1986). Abramson (1982) intitulou sua proposta como regra da raiz quadrada (*"square-root law"*).

Abramson (1982) limitou seu trabalho à estimação pontual, e mostrou que sua sugestão produz melhores resultados do que o núcleo-estimador fixo. Contudo, a regra da raiz quadrada foi muito utilizada na prática como um estimador global, apresentando ótimos resultados para amostras pequenas ($n < 200$) (TERRELL; SCOTT, 1992).

Silverman (1986) propôs em detalhes a implementação das idéias de Abramson (1982), que ele atribuiu o nome de núcleo-estimador adaptativo. A proposta de Silverman (1986) é definida como:

$$\begin{aligned}\hat{f}_S(x) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda\delta_i}K\left(\frac{x-X_i}{\lambda\delta_i}\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}K_{\lambda\delta_i}(x-X_i),\end{aligned} \tag{3.10}$$

em que λ é uma janela fixa e o parâmetro δ_i é um fator de localidade (*"local bandwidth factor"*), que tem a incumbência de tornar $\lambda\delta_i$ pequeno na região das modas e grande nas caudas.

O fator de localidade, δ_i, é dado por:

$$\delta_i = \left[\frac{\tilde{f}(X_i)}{b}\right]^{-\alpha}, \tag{3.11}$$

em que $\tilde{f}$ é uma estimativa piloto gerada pelo núcleo-estimador fixo, (3.1), utilizando a *"Silverman's rule-of-thumb"* (SILVERMAN, 1986) para estimação da janela, isto é,

$$h = 0,9An^{-1/5},$$

em que

$$A = \frac{\min\{\text{desvio-padrão , amplitude do intervalo interquartílico}\}}{1,34},$$

o termo b é a média geométrica de $\tilde{f}(X_i)$, isto é,

$$\log b = \frac{1}{n}\sum_{i=1}^{n}\log\tilde{f}(X_i),$$

e o parâmetro α, $0 \leq \alpha \leq 1$, é denominado parâmetro de sensibilidade (*"sensitivity parameter"*). Este parâmetro é responsável pela influência de δ_i na estimativa final de λ. Note que, se $\alpha = 0$, esta abordagem equivale ao núcleo-estimador fixo. Silverman (1986) optou por $\alpha = 1/2$ pelo mesmo motivo que Abramson (1982), ou seja, reduzir o vício pontual da ordem de $O(h^2)$ para $O(h^4)$.

Ainda em seus trabalhos, Silverman (1986) notou que utilizar a mesma janela na estimação piloto e na estimação final produz bons resultados, ou seja, aproveitar a janela utilizada na estimação de $\tilde{f}$, na expressão (3.11), como a janela fixa λ na estimação de $\hat{f}_S$, (3.10).

Terrell e Scott (1992) provaram que, sob algumas suposições, é possível obter aproximações da variância e do vício. Assumindo que $n \to \infty$, $w(x) \to 0$ e $nw(x) \to \infty$ para todo x, então, se f e w têm as duas primeiras derivadas contínuas,

$$\text{Var}\left[\hat{f}_{PA}(x)\right] \approx \frac{f(x)R(K)}{nw(x)}$$

e

$$\left\{\text{Vício}\left[\hat{f}_{PA}(x)\right]\right\}^2 \approx \frac{1}{4}\left\{\left\{[w(x)]^2 f(x)\right\}''\right\}^2.$$

Observe que a aproximação da variância é similar aos casos de janela fixa (3.2) e de janela local (3.8). Entretanto, na expressão do vício ao quadrado, a janela foi movida para dentro do diferencial.

Estudos mais detalhados sobre as propriedades gerais do núcleo-estimador por pontos amostrais não podem ser feitos, pois é impossível encontrar expressões para o cálculo do EQMI ou para aproximações do EQMI, sem o conhecimento da forma funcional da janela $h_i = w(x_i)$.

Para mais detalhes sobre esta abordagem, veja Silverman (1986), Jones (1990) e Terrell e Scott (1992).

4 *Métodos para a Estimação da Janela Ótima*

A implementação prática do núcleo-estimador requer a especificação do parâmetro de suavidade h. Como foi explicitado anteriormente na Seção 3.2.3, esta escolha visa otimizar alguma medida de acurácia, tal como o erro quadrático integrado ou erro quadrático médio integrado .

Ao longo dos anos, vários estudos foram feitos em busca de metodologias que estimassem "automaticamente" a janela ótima através da minimização de alguma medida de acurácia. Esses métodos, nomeados na literatura como métodos automáticos de seleção de janela, são baseados na idéia de que a quantidade ótima de suavização a ser feita deve depender unicamente dos dados.

A seguir, serão apresentadas algumas metodologias para a estimação da janela ótima, tanto no contexto de núcleo-estimador fixo quanto no contexto de núcleo-estimador variável.

4.1 Método Plug-in

O método *"Plug-in"* está baseado na idéia de substituir a quantidade desconhecida $R(f'')$, que aparece na expressão (3.6), por uma estimativa $\widehat{R(f'')}$ ou $R(\widehat{f''})$, ou seja, dado que a janela ótima para o Núcleo-Estimador fixo foi definida em (3.6) por

$$h_{opt} = \left[\frac{R(K)}{\sigma_K^4 R(f'')}\right]^{1/5} n^{-1/5}.$$

Substituindo $R(f'')$ por uma estimativa apropriada, $\hat{G}$, temos

$$\hat{h}_{opt} = \left[\frac{R(K)}{\sigma_K^4 \hat{G}}\right]^{1/5} n^{-1/5},$$

que é um estimador da janela ótima.

A seguir serão apresentadas duas abordagens distintas do método *plug-in*.

4.1.1 Abordagem de Chiu (1991)

4.1.1.1 Definições e Resultados

Inicialmente, serão apresentadas algumas definições e resultados que foram utilizados por Chiu (1991) para o desenvolvimento de sua proposta de estimação da janela ótima.

Definição 4.1.1 (Função Característica) *Seja X uma variável aleatória. A função característica de X é a função $\varphi : \mathbb{R} \to C$ definida por*

$$\varphi(\lambda) = \varphi_X(\lambda) = E\left[e^{i\lambda X}\right],$$

em que $\lambda \in \mathbb{R}$ e $i = \sqrt{-1}$.

Definição 4.1.2 (Função Característica Empírica) *Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória de uma distribuição univariada. Então, a função característica empírica, $\hat{\varphi}$, avaliada em λ, é definida por*

$$\hat{\varphi}(\lambda) = \frac{1}{n}\sum_{j=1}^{n} e^{i\lambda X_j}.$$

Teorema 4.1.1 (Fórmula de Inversão) *Se φ é a função característica de uma função de distribuição limitada F, e $F(a,b] = F(b) - F(a)$, então*

$$F(a,b] = \lim_{c\to\infty} \frac{1}{2\pi}\int_{-c}^{c} \frac{e^{-i\lambda a} - e^{-i\lambda b}}{i\lambda}\varphi(\lambda)d\lambda$$

para todos os pontos a,b $(a < b)$ em que F é contínua. Se φ é Lebesgue integrável em $(-\infty, \infty)$, então a função densidade f é dada por

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\lambda x}\varphi(\lambda)d\lambda$$

e é função densidade de F, isto é, f não-negativa e $F(x) = \int_{-\infty}^{x} f(u)du$ para todo x.

Teorema 4.1.2 (Identidade de Parseval) *Seja φ a função característica de f. Então,*

$$\int_{-\infty}^{\infty} [f(x)]^2\, dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\ \varphi(\lambda)\ |^2\, d\lambda.$$

4.1.1.2 Metodologia Proposta

Chiu (1991) propôs um estimador para $R(f'')$ fundamentado no seguinte fato: pela fórmula de inversão, temos que

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\lambda x}\varphi(\lambda)d\lambda, \tag{4.1}$$

e derivando duas vezes a igualdade (4.1), pode ser mostrado que a relação a seguir é válida,

$$f''(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} -\lambda^2 e^{-i\lambda x}\varphi(\lambda)d\lambda. \tag{4.2}$$

Então, utilizando a Identidade de Parseval em (4.2), segue que

$$\begin{aligned} R(f'') &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \lambda^4|\varphi(\lambda)|^2 d\lambda \\ &= \frac{1}{\pi}\int_{0}^{\infty} \lambda^4|\varphi(\lambda)|^2 d\lambda. \end{aligned} \tag{4.3}$$

Baseado em (4.3), Chiu (1991) propõe o estimador

$$\hat{G} = \pi^{-1}\int_0^{\Lambda} \lambda^4 \left[|\hat{\varphi}(\lambda)|^2 - \frac{1}{n}\right] d\lambda, \tag{4.4}$$

em que $\Lambda = \min\left\{\lambda : |\hat{\varphi}(\lambda)|^2 \leq \frac{c}{n}\right\}$ para alguma constante $c > 1$ e $\hat{\varphi}$ é a função característica empírica, dada pela definição 4.1.2. Chiu (1991) argumenta que, para altos valores de λ, $|\hat{\varphi}(\lambda)|^2$ não contém muita informação relativa a f, optando assim pela condição $\min\left\{\lambda : |\hat{\varphi}(\lambda)|^2 \leq \frac{c}{n}\right\}$ para escolha de Λ.

Chiu (1991) mostra que a escolha de c não é importante quando f é suficientemente suave e sugere que para propósitos práticos, $-\ln 0,15 \leq c \leq -\ln 0,05$, produz bons resultados.

Portanto, substituindo (4.4) em (3.6), temos que

$$\hat{h}_{chiu} = \left[\frac{R(K)}{\sigma_K^4 \hat{G}}\right]^{1/5} n^{-1/5},$$

é o estimador da janela ótima.

No entanto, $\hat{h}_{chiu}$ não é um estimador $\sqrt{n}$ consistente de h_{opt}. Chiu (1991), ainda em seu trabalho, sugere uma correção para $\hat{h}_{chiu}$ baseado no trabalho de Hall *et al.* (1991) para contornar esse inconveniente. Deste modo, o estimador ajustado da janela ótima $\hat{h}_{adj}$, é dado por

$$\hat{h}_{adj} = \left[\hat{h}_{chiu} + \frac{\tilde{R}'_n\left(\hat{h}_{chiu}\right)}{\tilde{A}''_n\left(\hat{h}_{chiu}\right)}\right] n^{-1/5},$$

em que

$$\begin{aligned}\tilde{R}_n(h) &= (24\pi)^{-1} n^{-2/5} h^6 \int_0^{\Lambda} \lambda^6 \left[|\hat{\varphi}(\lambda)|^2 - \frac{1}{n}\right] d\lambda \\ &\times \int x^2 K(x)dx \int x^4 K(x)dx\end{aligned}$$

e

$$\tilde{A}_n(h) = 4^{-1} h^4 \sigma_K^4 \hat{G} + h^{-1} R(K) - \tilde{R}_n(h).$$

4.1.2 Abordagem de Sheather e Jones (1991)

Sheather e Jones (1991) propuseram uma outra abordagem para o método *"plug-in"*, denominado *"plug-in"* multi-estágio. Considere que a densidade desconhecida f seja suficientemente suave, então

$$\begin{aligned}R(f^{(s)}) &= \int \left[f^{(s)}(x)\right]^2 dx \\ &= (-1)^s \int f^{(2s)}(x) f(x) dx,\end{aligned}$$

em que $f^{(s)}$ é a s-ésima derivada de f. Logo, a expressão (3.6) pode ser reescrita como

$$R(f'') = \int f^{(4)}(x) f(x) dx.$$

Seja

$$\begin{aligned} \psi_r &= \int f^{(r)}(x) f(x) dx \\ &= \mathrm{E}\left[f^{(r)}(X)\right], \end{aligned} \tag{4.5}$$

para todo r. Nesse caso, um estimador "natural" para ψ_r é dado por

$$\hat{\psi}_r = \frac{1}{n} \sum_{i=1}^{n} \widehat{f^{(r)}}(X_i),$$

o que motivou o construção do estimador

$$\begin{aligned} \hat{\psi}_r &= \frac{1}{n} \sum_{i=1}^{n} \widehat{f^{(r)}}(X_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{n g^{r+1}} \sum_{j=1}^{n} K^{(r)} \left(\frac{X_i - X_j}{g} \right) \right] \\ &= \frac{1}{n^2 g^{r+1}} \sum_{i=1}^{n} \sum_{j=1}^{n} K^{(r)} \left(\frac{X_i - X_j}{g} \right), \end{aligned}$$

(HALL; MARRON, 1987b; JONES; SHEATHER, 1991) em que K e g são, respectivamente, a função núcleo e o parâmetro de suavidade.

Reescrevendo (3.6) através de (4.5), obtemos

$$h_{opt} = \left[\frac{R(K)}{\sigma_K^4 \psi_4} \right]^{1/5} n^{-1/5}.$$

Diante disso, temos o estimador da janela ótima enunciado por

$$\hat{h}_{sj} = \left[\frac{R(K)}{\sigma_K^4 \hat{\psi}_4} \right]^{1/5} n^{-1/5}.$$

Note que, como $\hat{h}_{sj}$ depende de $\hat{\psi}_4$, conseqüentemente dependerá da janela g.

Sheather e Jones (1991) mostraram que a janela ótima , $g_{r_{opt}}$, que minimiza uma aproximação do erro quadrático médio,

$$\mathrm{EQM}\left(\hat{\psi}_r\right) = \mathrm{E}\left[\left(\hat{\psi}_r - \psi_r\right)^2\right],$$

é dada por

$$g_{r_{opt}} = \left[\frac{2K^{(r)}(0)}{-\sigma_k^2 \psi_{r+2}}\right]^{1/(r+3)} n^{-1/(r+3)}.$$

Portanto, a janela ótima para estimar ψ_4 é

$$g_{4_{opt}} = \left[\frac{2K^{(4)}(0)}{-\sigma_K^2 \psi_6}\right]^{1/7} n^{-1/7}.$$

Note que $g_{4_{opt}}$ depende do funcional ψ_6, que é desconhecido, ou seja, precisa ser estimado da mesma forma que ψ_4. Por conseguinte, a janela ótima, $g_{6_{opt}}$, é dada por

$$g_{6_{opt}} = \left[\frac{2K^{(6)}(0)}{-\sigma_K^2 \psi_8}\right]^{1/9} n^{-1/9}.$$

Observe que $g_{6_{opt}}$ depende de ψ_8, que precisará ser estimado através de $\hat{\psi}_8$, que por consequência dependerá de ψ_{10}, e assim por diante, tornando o procedimento insolúvel.

A proposta dos autores para resolução deste problema foi adotar uma distribuição de referência em determinado estágio de estimação de ψ_r, isto é, supondo que são realizadas m estimações sucessivas (estágios) dos funcionais do núcleo (ψ_r), então, na $(m+1)$-ésima estimação suponha que a função densidade f é dada por alguma distribuição paramétrica. Dessa forma, o funcional pode ser calculado diretamente sem a necessidade de estimação de ψ_r.

Wand e Jones (1995) demostraram que, ao assumir uma distribuição de referência normal com variância σ^2, então para todo r par

$$\psi_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1}(r/2)!\pi^{1/2}}.$$

Para maiores informações sobre o método *"plug-in"*, veja Woodroofe (1970), Chiu (1991), Sheather e Jones (1991), Wand e Jones (1995), Polansky e Baker (2000) e Damasceno (2000). Uma aplicação em controle de qualidade é apresentada em Glória (2006) para processos multivariados.

4.2 Abordagem de Sain e Scott (1996)

Sain e Scott (1996) propuseram uma metodologia para a estimação das janelas ótimas no caso do Núcleo-Estimador por Pontos Amostrais via Núcleo-Estimador por Blocos-

Pontos Amostrais (*"binned sample points estimator"*).

Silverman (1982) e Scott e Sheather (1985) descreveram a noção de núcleo-estimador por blocos como uma abordagem prática do núcleo-estimador fixo. Considere a reta dividida em blocos de tamanho θ, e seja t_j o centro do j-ésimo bloco, B_j. O número de observações contidas no j-ésimo bloco, B_j, é denotado por n_j. Note que $t_j - t_{j-1} = \theta$ para todo j e $n = \sum_j n_j$.

Dessa forma, o núcleo-estimador por blocos é dado por

$$\begin{aligned} \hat{f}_b(x) &= \frac{1}{nh} \sum_{j=-\infty}^{\infty} n_j K\left(\frac{x - t_j}{h}\right) \\ &= \frac{1}{n} \sum_{j=-\infty}^{\infty} n_j K_h\left(x - t_j\right). \end{aligned} \tag{4.6}$$

Na prática, somente um número finito de blocos contém observações, logo (4.6) pode ser reduzida a

$$\hat{f}_{bt}(x) = \frac{1}{n} \sum_{j=1}^{m} n_j K_h\left(x - t_j\right), \tag{4.7}$$

em que m é o número de blocos contendo pelo menos uma observação.

Hall (1982) deduziu o EQM para ambos os estimadores, (4.6) e (4.7). Scott e Sheather (1985) expandiram os resultados de Hall (1982) e encontraram uma expressão para o EQMI aproximado para o estimador (4.7), que é dada por:

$$\text{EQMIA}(\hat{f}_{bt}) = \frac{R(K)}{nh} + \frac{h^4 \sigma_K^4 R(f'')}{4} \left(1 + \frac{\theta^2}{12 h^2 \sigma_K^2}\right).$$

Perceba que dividir os dados em blocos aumenta apenas o termo relativo ao vício, quando comparado com o EQMIA proporcionado pelo núcleo-estimador fixo, dado em (3.5).

A versão adaptativa para (4.7) é dada por:

$$\hat{f}_{bpa}(x) = \frac{1}{n} \sum_{j=1}^{m} n_j K_{h_j}\left(x - t_j\right). \tag{4.8}$$

Sain e Scott (1996), utilizando uma função núcleo gaussiana, deduziram uma expressão fechada para o erro quadrático médio integrado de (4.8), que é dado por

$$\begin{aligned}\text{EQMI}\left(\hat{f}_{bpa}\right) &= \frac{1}{2n\sqrt{\pi}}\sum_j \frac{p_j(1-p_j)+np_j^2}{h_j} \\ &\quad + \frac{n-1}{n}\sum_{i\neq j} p_i p_j K_{\sqrt{h_i^2+h_j^2}}(t_i - t_j) \\ &\quad - \frac{2}{n}\sum_j p_j \int K_{h_j}(x-t_j)f(x)dx + R(f),\end{aligned}$$

em que $p_j = \int_{B_j} f(x)dx$ (WAND; JONES, 1995; SAIN; SCOTT, 1996).

A proposta apresentada por Sain e Scott (1996) consiste em estimar as janelas ótimas através de (4.8), posteriormente associar estas janelas aos pontos amostrais que estão contidos em cada bloco, e então, estimar a função densidade utilizando o núcleo-estimador por pontos amostrais (3.9).

A obtenção das janelas ótimas é feita através do método de validação cruzada por mínimos quadrados (VCMQ) (RUDEMO, 1982; BOWMAN, 1984). O método VCMQ visa minimizar o erro quadrático integrado, que neste caso é dado por

$$\begin{aligned}\text{EQI}\left(\hat{f}_{bpa}\right) &= \int \left[\hat{f}_{bpa}(x) - f(x)\right]^2 dx \\ &= R(\hat{f}_{bpa}) - 2\int \hat{f}_{bpa}(x)f(x)dx + R(f).\end{aligned}$$

Observe que $R(f)$ não depende das janelas e pode ser ignorado no procedimento de minimização.

Utilizando o núcleo gaussiano, Sain e Scott (1996) mostraram que $R(\hat{f}_{bpa})$ é dado por:

$$R(\hat{f}_{bsp}) = \frac{1}{2n^2\sqrt{\pi}}\sum_{j=1}^{m}\frac{n_j^2}{h_j} + \frac{1}{n^2}\sum_{i\neq j} n_i n_j K_{\sqrt{h_i^2+h_j^2}}(t_i - t_j). \tag{4.9}$$

O produto cruzado, $\int \hat{f}_{bpa}(x)f(x)dx = \text{E}\left[\hat{f}_{bpa}(X)\right]$, pode ser estimado de forma não-viciada pelo *"leave-one-out estimator"*

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{f}_{-i}(x_i), \tag{4.10}$$

em que

$$\tilde{f}_{-i}(x_i) = \frac{1}{n-1} \sum_{j=1}^{m} n_{ij}^{*} K\left(\frac{x_i - t_j}{h_j}\right)$$

e

$$n_{ij}^{*} = \begin{cases} n_j - 1, & x_i \in B_j \\ n_j, & \text{c.c.} \end{cases}$$

Combinando (4.9) e (4.10), a função VCMQ é dada por:

$$\text{VCMQ}\,(h_1, \ldots, h_m) = R(\hat{f}_{bpa}) - \frac{2}{n} \sum_{i=1}^{n} \tilde{f}_{-i}(x_i). \tag{4.11}$$

As estimativas das janelas ótimas são obtidas através da minimização de (4.11) sob m e $\{h_j, j = 1, \ldots, m\}$. Na prática, é necessário utilizar algum método de otimização vetorial para encontrar o mínimo de (4.11).

Sain e Scott (1996) mostraram que esta metodologia é pouco sensível à localização dos blocos e apresenta melhor desempenho que o núcleo-estimador fixo e o método de Abramson (1982) para o núcleo-estimador por pontos amostrais. No entanto, Sain e Scott (1996) argumentam que o método é instável para pequenas amostras, ou seja, quando se tem poucas observações nos blocos.

4.3 Metodologias Bayesianas

Os principais resultados para a estimação da janela ótima são quase que exclusivamente não-bayesianos, contudo com a prosperidade da estatística bayesiana nos últimos anos, existe a necessidade de desenvolver, bem como avaliar a capacidade dos métodos com enfoque bayesiano como alternativa aos com enfoque clássico.

Nas próximas seções serão apresentadas duas metodologias bayesianas para estimação das janelas ótimas.

4.3.1 Abordagem de Brewer (2000)

Brewer (2000) propôs um procedimento para a estimação das janelas ótimas do núcleo estimador adaptativo, (3.10), sob a óptica bayesiana. A idéia de Brewer (2000) consiste em utilizar validação cruzada por verossimilhança para tratar o fator de localidade, δ_i,

que aparece na expressão (3.10) como um parâmetro de escala, e assim, proceder uma análise bayesiana.

Considere, inicialmente, uma amostra aleatória, $X_1, X_2, \ldots, X_n$, de uma distribuição univariada contínua, com função densidade de probabilidade f. Uma aproximação para a log-verossimilhança é dada por

$$\begin{aligned} \log L(h|x_1, x_2, \ldots, x_n) &= \sum_{j=1}^{n} \log \hat{f}(x_j) \\ &= \sum_{j=1}^{n} \log \left[\frac{1}{n} \sum_{i=1}^{n} K_h (x_j - x_i) \right]. \end{aligned} \tag{4.12}$$

Uma maneira intuitiva de estimar a janela ótima seria maximizar (4.12). Entretanto, a maximização direta de L, com respeito a h, resulta em $h = 0$. Uma forma de evitar este problema é mediante o princípio de validação cruzada, isto é, estimar $\hat{f}(x_j)$ baseado no subconjunto $\{x_i, i \neq j\}$. Logo,

$$\hat{f}(x_j) = \frac{1}{n-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} K_h (x_j - x_i).$$

Sendo assim, (4.12) pode ser reescrita como

$$\log L(h|x_1, x_2, \ldots, x_n) = \sum_{j=1}^{n} \log \left[\frac{1}{n-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} K_h (x_j - x_i) \right]. \tag{4.13}$$

Deste modo, uma estimativa de h é obtida através da maximização de (4.13). Este método de estimação da janela ótima foi proposto independentemente por Habbema *et al.* (1974) e Duin (1976), e é denominado método de validação cruzada por verossimilhança.

Baseado nesta metodologia e no núcleo-estimador adaptativo (3.10), Brewer (2000) expôs a seguinte aproximação da função de verossimilhança:

$$L(\delta_1, \delta_2, \ldots, \delta_n | x_1, x_2, \ldots, x_n) = \prod_{j=1}^{n} \left[\frac{1}{n-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{1}{h\delta_j} K\left(\frac{x_j - x_i}{h\delta_j} \right) \right] \quad (4.14)$$
$$= \prod_{j=1}^{n} f(x_j | \{x_{-j}\}, \delta_j, h),$$

em que K é o núcleo normal, h é uma janela piloto, que é fixa e conhecida, e $\{x_{-j}\}$ equivale a exclusão da observação x_j da amostra. Observe que δ_j aparece como parâmetro de escala na expressão (4.14).

Com o intuito de explorar a conjugação entre a verossimilhança Normal e a família Gama, Brewer (2000) adotou uma distribuição *a priori* Gama-Invertida com parâmetros d_1 e d_2 para δ_j. Além disso, Brewer (2000) impôs uma parametrização na *priori* de tal forma que

$$\mathrm{E}[\delta_j] = 1,$$

implicando na seguinte relação entre os hiperparâmetros da distribuição *a priori*:

$$d_2 = \left[\frac{\Gamma(d_1)}{\Gamma\left(d_1 + \frac{1}{2}\right)} \right]^2, \quad \text{para } d_1 > \tfrac{1}{2}$$

Dessa forma, a distribuição *a posteriori* de δ_j é dada por:

$$f(\delta_j | \{x_{-j}\}, h, d_1, d_2) = \frac{d_2^{d_1}}{\Gamma(d_1)} \left(\frac{1}{\delta_j^2} \right)^{d_1 - 1} \exp\left(-\frac{d_2}{\delta_j^2} \right) \times \frac{1}{n-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} K_{h\delta_j}(x_j - x_i).$$

Utilizando a média *a posteriori* como estimativa para os fatores de localidade δ_j, Brewer (2000) mostrou que

$$\mathrm{E}[\delta_j | \{x_{-j}\}, h, d_1, d_2] = \frac{\Gamma(d_1) \sum_{\substack{i=1 \\ i \neq j}}^{n} \left[d_2 + \frac{(x_j - x_i)^2}{2h^2} \right]^{-d_1}}{\Gamma\left(d_1 + \frac{1}{2}\right) \sum_{\substack{i=1 \\ i \neq j}}^{n} \left[d_2 + \frac{(x_j - x_i)^2}{2h^2} \right]^{-(d_1 + \frac{1}{2})}}.$$

Portanto, as estimativas das janelas ótimas são dadas pelo produto,

$$\hat{h}_j = h\hat{\delta}_j \, , \forall j,$$

em que $\hat{\delta}_j$ é a média *a posteriori*, ou seja,

$$\hat{\delta}_j = \frac{\Gamma\left(d_1\right) \sum\limits_{\substack{i=1 \\ i \neq j}}^{n} \left[d_2 + \frac{(x_j - x_i)^2}{2h^2} \right]^{-d_1}}{\Gamma\left(d_1 + \frac{1}{2}\right) \sum\limits_{\substack{i=1 \\ i \neq j}}^{n} \left[d_2 + \frac{(x_j - x_i)^2}{2h^2} \right]^{-(d_1 + \frac{1}{2})}}.$$

Desta maneira, o método depende de duas entradas para ser utilizado: a janela piloto h, que pode ser estimada por metodologias *plug-in*; e o hiperparâmetro d_1 da distribuição *a priori* Gama-Invertida.

Conforme estudos de Brewer (2000), a escolha de d_1 pode ser sumariada pela Tabela 4.1.

Tabela 4.1: Possíveis regras para determinação de d_1

Tamanho amostral	d_1
$n \leq 30$	2,0
$30 < n \leq 75$	1,2
$75 < n \leq 200$	1,1
$n > 200$	1,0

Uma explicação possível para a Tabela 4.1, pode ser encontrada ao observamos o comportamento da distribuição *a priori* Gama-Invertida para diferentes valores de d_1, através da Figura 4.1.

Note que, quanto maior o valor de d_1, mais concentrada é a distribuição a priori, ou seja, para pequenas amostras, Brewer (2000) propõe utilizar uma distribuição mais informativa. Enquanto que, para grandes amostras, faz uso de uma *priori* menos informativa. Este fato demonstra, possivelmente, uma falta de robustez da metodologia em relação à escolha dos hiperparâmetros. No entanto, para afirmação de tal fato seria necessário um estudo mais detalhado sobre a sensibilidade do modelo, algo que foge ao escopo desta dissertação.

Brewer (2000) estendeu o seu modelo introduzindo uma dependência de vizinhança entre os fatores de localidade δ_j. A dependência é dada da seguinte forma:

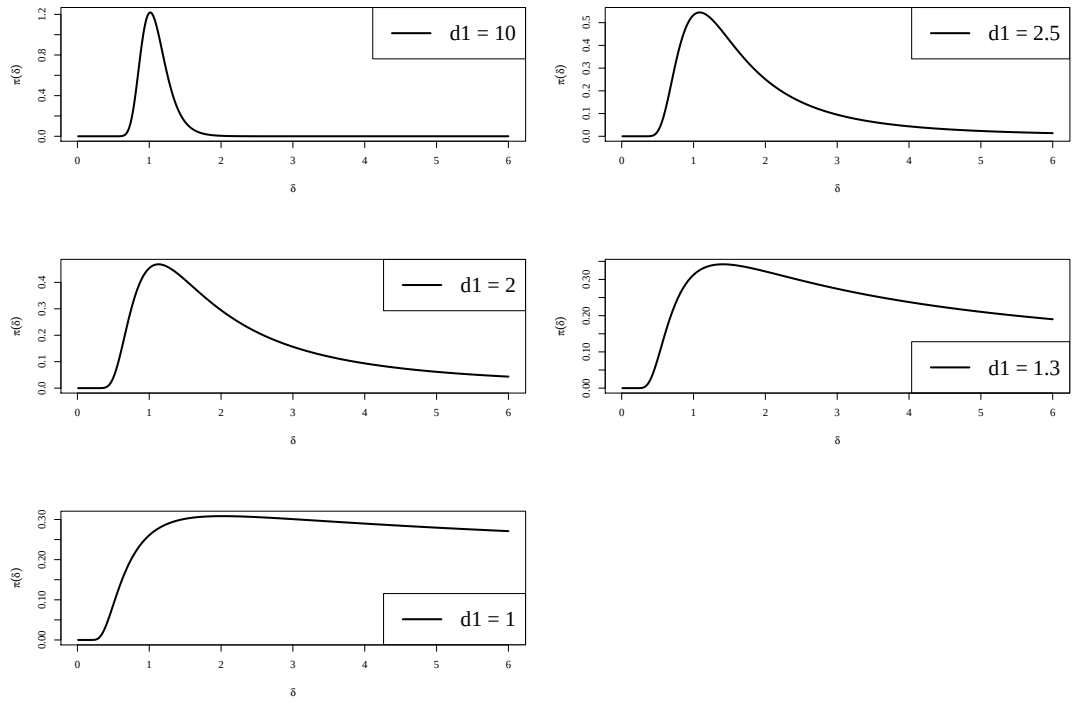

Figura 4.1: Comportamento da distribuição Gama-Invertida(d_1, d_2) para diferentes valores de d_1.

$$\delta_j \sim GI\left(\frac{1}{5}\left[d_1 + d_2 \times (\delta_{j-2} + \delta_{j-1} + \delta_{j+1} + \delta_{j+2})\right], d_2\right),$$

em que $j \in \{1, 2, \ldots, n-1, n-2\}$. Note que as observações precisam ser ordenadas para assegurar a funcionalidade do método.

Brewer (2000) argumenta que utilizar os quatro vizinhos mais próximos, ou seja, dois vizinhos de cada lado de δ_j, é a opção mais plausível, pois apenas um vizinho não produz melhoras significativas, e mais que dois vizinhos aumenta a complexidade do modelo. Para inferências sobre δ_j, é necessária a utilização de Métodos de Monte Carlo via Cadeias de Markov (MCMC), devido à dificuldade de encontrar uma expressão exata para a distribuição conjunta do modelo.

De acordo com Brewer (2000), os modelos propostos possuem um bom desempenho, tanto teoricamente como na prática.

4.3.2 Abordagem de Gangopadhyay e Cheung (2002)

Gangopadhyay e Cheung (2002) propuseram uma abordagem bayesiana para a estimação da janela ótima para o núcleo-estimador local. A sugestão de Gangopadhyay e

Cheung (2002) é tratar a janela h como um parâmetro verdadeiro, atribuindo-lhe uma distribuição *a priori* em cada ponto em que a densidade está sendo estimada. Entretanto, uma dificuldade óbvia é que h não é um parâmetro verdadeiro, ou seja, a janela h não indexa uma família paramétrica de modelos. Conseqüentemente, não é possível calcular a distribuição *a posteriori* diretamente.

A fim de superar esta dificuldade, Gangopadhyay e Cheung (2002) reformularam o problema da seguinte maneira. Se o interesse é estimar f pontualmente em x, apenas a vizinhança de x é de vital importância na estimação. Dessa forma, considere a versão truncada de $f(x)$ dada pela convolução de f e K_h, isto é,

$$\begin{aligned} f_h(x) = f * K_h(x) &= \int f(u)K_h(x-u)du \\ &= \int \frac{f(u)}{h} K\left(\frac{x-u}{h}\right) du, \end{aligned} \tag{4.15}$$

em que a função núcleo K é densidade de probabilidade. Observe que f_h pode ser vista como a densidade da variável aleatória $Z = X + \epsilon$, em que X é uma variável aleatória com densidade f e ϵ é uma contaminação aleatória com média zero e densidade dada por K_h.

Na expressão (4.15), a janela h aparece como um parâmetro de escala. Desse modo, seja $\pi(h)$ a distribuição sobre todos os possíveis valores de h, então a distribuição *a posteriori* de h dado $X = x$ é dada por

$$\pi\left(h|x\right) = \frac{f_h(x)\pi(h)}{\int f_h(x)\pi(h)dh}. \tag{4.16}$$

No entanto, $\pi\left(h|x\right)$ não pode ser calculada diretamente, pois $f_h(x)$ é desconhecida. Gangopadhyay e Cheung (2002) propuseram utilizar um *"plug-in"* na expressão (4.16).

Perceba que (4.15) pode ser reescrita como

$$f_h(x) = \mathrm{E}_f\left[K_h(X-x)\right],$$

Assim, baseado numa amostra aleatória $X_1, \ldots, X_n$, uma estimativa plausível para $f_h(x)$ é dada pela média amostral, ou seja,

$$\begin{aligned}\hat{f}_h(x) &= \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) \\ &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \end{aligned} \tag{4.17}$$

que é simplesmente o núcleo-estimador de $f(x)$. Substituindo (4.17) em (4.16), temos uma estimativa da distribuição *a posteriori* dada por

$$\hat{\pi}\left(h|X_1, \ldots, X_n, x\right) = \frac{\hat{f}_h(x)\pi(h)}{\int \hat{f}_h(x)\pi(h)dh}. \tag{4.18}$$

Uma possível escolha para a janela local $h = h(x)$ é a média *a posteriori*, que é dada por

$$h^* = h^*(x) = \int h\hat{\pi}\left(h|X_1, \ldots, X_n, x\right) dh. \tag{4.19}$$

Em muitas situações, as expressões (4.18) e (4.19) não possuem formas fechadas, sendo necessários procedimentos computacionais intensivos para a determinação das estimativas.

Gangopadhyay e Cheung (2002) abordaram uma situação em que é possível encontrar uma expressão fechada para (4.19). Suponha que K seja a função núcleo guassiana, isto é,

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right), \quad -\infty < u < \infty.$$

Neste caso, a janela h aparece como desvio-padrão na densidade normal $K_h(X_i - x)$. A distribuição conjugada *a priori* de h é dada por

$$\tau(h) = \frac{2}{\Gamma(\alpha)\beta^{\alpha}} \frac{1}{h^{2\alpha+1}} \exp\left(-\frac{1}{\beta h^2}\right), \quad h > 0, \tag{4.20}$$

que representa a raiz quadrada de uma distribuição Gama-Invertida(α, β).

Neste caso, a esperança e variância da distribuição *a priori* são dadas por:

$$\mathrm{E}\left[h\right] = \frac{\Gamma\left(\alpha - \frac{1}{2}\right)}{\Gamma\left(\alpha\right)} \beta^{-\frac{1}{2}}$$

e

$$\text{Var}\left[h\right] = \frac{1}{\left(\alpha - 1\right)\beta} - \left[\frac{\Gamma\left(\alpha - \frac{1}{2}\right)}{\Gamma\left(\alpha\right)}\beta^{-\frac{1}{2}}\right]^2.$$

Dessa maneira, Gangopadhyay e Cheung (2002) mostraram que

$$\hat{\pi}\left(h|X_1, \ldots, X_n, x\right) = \frac{\sum_{i=1}^{n}\left(1/h^{2\alpha+2}\right)\exp\left\{-(1/h^2)((X_i - x)^2/2 + 1/\beta)\right\}}{\sum_{i=1}^{n}\left(\Gamma(\alpha + 1/2)/2\right)\left\{((X_i - x)^2/2 + 1/\beta)\right\}^{-\alpha-1/2}}$$

e, conseqüentemente

$$h_{gc}(x) = \frac{\Gamma(\alpha)}{\sqrt{2\beta}\Gamma(\alpha + 1/2)}\frac{\sum_{i=1}^{n}\left\{1/\left(\beta(X_i - x)^2 + 2\right)\right\}^{\alpha}}{\sum_{i=1}^{n}\left\{1/\left(\beta(X_i - x)^2 + 2\right)\right\}^{\alpha+1/2}},$$

que é a estimativa da janela ótima.

Conforme estudos, Gangopadhyay e Cheung (2002) salientam que o estimador $\hat{f}_{h_{gc}}(x)$ é relativamente robusto quanto à escolha dos hiperparâmetros, e que a metodologia proposta apresenta desempenho superior quando comparada aos métodos de validação cruzada por mínimos quadrados e validação cruzada viciada (RUDEMO, 1982; BOWMAN, 1984; SCOTT; TERRELL, 1987).

5 Simulações

Neste Capítulo serão apresentados os detalhes das simulações, que foram realizadas com o objetivo de comparar as diferentes abordagens de núcleo-estimadores, assim como os diversos métodos de estimação da janela ótima.

5.1 Implementação Computacional

A implementação computacional dos núcleo-estimadores, apresentados no Capítulo 3, e das metodologias de estimação da janela ótima, mencionadas no Capítulo 4, foram feitas em linguagem Ox versão 4.04 (DOORNIK, 2005).

5.1.1 Ox

Ox é uma linguagem de programação, matricial, orientada a objetos e dotada de uma abrangente biblioteca de funções matemáticas e estatísticas. A linguagem Ox possui uma sintaxe muito similar às linguagens C e C++. Sua principal característica é a velocidade, que a torna uma boa opção quando a resolução de problemas necessitam de procedimentos computacionais intensivos.

Ox está disponível para Windows, Linux e várias plataformas Unix. Para mais detalhes veja Doornik (2005).

5.2 Modelos Simulados

Nesta seção serão apresentados os modelos probabilísticos que foram utilizados nas simulações. Estes modelos foram escolhidos visando avaliar e comparar a capacidade das diferentes metodologias presentadas no Capítulo 4 em captar características peculiares da função densidade de probabilidade, tais como assimetria e multi-modalidade. As simulações foram conduzidas através da reconstrução de densidades conhecidas.

Inicialmente, serão apresentadas as parametrizações dos modelos probabilísticos que foram empregados nas simulações. No Apêndice A são apresentados alguns resultados conjuntamente com os gráficos das funções densidade de probabilidade dos modelos que foram utilizados nas simulações.

Definição 5.2.1 (Modelo Normal) *A variável aleatória X tem distribuição normal com parâmetros μ e σ^2, denotada por $X \sim N(\mu, \sigma^2)$, se a função densidade de probabilidade de X é dada por*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, x \in \Re$$

com média $\mu \in \Re$ e variância $\sigma^2 > 0$.

Definição 5.2.2 (Modelo Gama) *A variável aleatória X tem distribuição gama com parâmetros α e β, denotada por $X \sim Gama(\alpha, \beta)$, se a função densidade de probabilidade de X é dada por*

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\} I_{[0,\infty)}(x),$$

em que $\alpha > 0$, $\beta > 0$ e $\Gamma(\cdot)$ é a função gama. Neste caso, a esperança e a variância de X são dadas respectivamente por:

$$E[X] = \frac{\alpha}{\beta} \text{ e } Var[X] = \frac{\alpha}{\beta^2}.$$

Definição 5.2.3 (Modelo Weibull) *A variável aleatória X tem distribuição Weibull com parâmetros α e β, denotada por $X \sim Weibull(\alpha, \beta)$, se a função densidade de probabilidade de X é dada por*

$$f_X(x) = \alpha\beta x^{\beta-1} \exp\left\{-\alpha x^\beta\right\} I_{(0,\infty)}(x),$$

em que $\alpha > 0$ e $\beta > 0$, sendo a esperança e a variância de X dadas respectivamente por:

$$E[X] = \left(\frac{1}{\alpha}\right)^{1/\beta} \Gamma(1+\beta^{-1}) \text{ e } Var[X] = \left(\frac{1}{\alpha}\right)^{2/\beta} \left\{\Gamma(1+2\beta^{-1}) - \left[\Gamma(1+\beta^{-1})\right]^2\right\}.$$

Definição 5.2.4 (Modelo Qui-Quadrado) *A variável aleatória X tem distribuição qui-quadrado com k graus de liberdade, denotada por $X \sim Q_k$, se a função densidade de probabilidade de X é dada por*

$$f_X(x) = \frac{1}{\Gamma(k/2)} \left(\frac{1}{2}\right)^{k/2} x^{(k/2)-1} \exp\left\{-\frac{1}{2}x\right\} I_{(0,\infty)}(x),$$

em que $k > 0$ *e* $\Gamma(\cdot)$ *é a função gama. Neste caso,* k *é a média de* X *e* $2k$ *é a variância de* X*.*

As seguintes funções densidade de probabilidade foram consideradas nas simulações:

1. Estimação da função densidade a partir de amostras aleatórias da distribuição Normal(0,1);

2. Estimação da função densidade a partir de amostras aleatórias da distribuição Gama(4,2);

3. Estimação da função densidade a partir de amostras aleatórias da distribuição Weibull(1,6);

4. Estimação da função densidade a partir de amostras aleatórias da distribuição Qui-Quadrado(7);

5. Estimação da função densidade a partir de amostras aleatórias de uma mistura de normais, $\frac{1}{5}N(0,1)+\frac{1}{5}N\left(\frac{1}{4},\frac{4}{9}\right)+\frac{3}{5}N\left(\frac{13}{12},\frac{25}{81}\right)$, que será denominada Mistura(1);

6. Estimação da função densidade a partir de amostras aleatórias de uma mistura de normais, $\frac{1}{2}N(2,1)+\frac{1}{2}N(5,1)$, que será denominada Mistura(2);

7. Estimação da função densidade a partir de amostras aleatórias de uma mistura de normais, $\frac{1}{2}N(2,1)+\frac{1}{2}N(6,1)$, que será denominada Mistura(3);

8. Estimação da função densidade a partir de amostras aleatórias de uma mistura de normais, $\frac{1}{2}N(2,1)+\frac{1}{2}N(8,1)$, que será denominada Mistura(4);

9. Estimação da função densidade a partir de amostras aleatórias de uma mistura de normais, $\frac{1}{2}N(2,1)+\frac{1}{2}N(6,3)$, que será denominada Mistura(5);

10. Estimação da função densidade a partir de amostras aleatórias de uma mistura de normais, $\frac{1}{2}N(2,1)+\frac{1}{2}N(8,3)$, que será denominada Mistura(6);

11. Estimação da função densidade a partir de amostras aleatórias de uma mistura de normais, $\frac{4}{5}N(0,1)+\frac{1}{5}N\left(2,\frac{1}{25}\right)$, que será denominada Mistura(7);

12. Estimação da função densidade a partir de amostras aleatórias de uma mistura de normais, $\frac{1}{4}N(3,1)+\frac{1}{2}N\left(6,\frac{1}{2}\right)+\frac{1}{4}N(9,1)$, que será denominada Mistura(8);

13. Estimação da função densidade a partir de amostras aleatórias de uma mistura de normais, $\frac{2}{3}N(0,1) + \frac{1}{3}N\left(0, \frac{1}{100}\right)$, que será denominada Mistura(9);

14. Estimação da função densidade a partir de amostras aleatórias de uma mistura de weibull, $\frac{4}{5}Weibull\left(\frac{1}{100}, 6\right) + \frac{1}{5}Weibull(1,6)$, que será denominada Mistura(10);

Em cada modelo, foram geradas 1000 amostras de tamanho 30, 50, 100 e 200 para avaliação dos ajustes e estimação das funções densidade de probabilidade. As medidas de acurácia, que foram empregadas nas avaliações dos ajustes e na comparação dos estimadores, são:

- Erro Quadrático Integrado (EQI),

$$\text{EQI}(\hat{f}) = \int \left[f(x) - \hat{f}(x)\right]^2 dx.$$

- Erro Absoluto Integrado (EAI),

$$\text{EAI}(\hat{f}) = \int \mid f(x) - \hat{f}(x) \mid dx.$$

- Erro Integrado (EI),

$$\text{EI}(\hat{f}) = \int \left[\hat{f}(x) - f(x)\right] dx.$$

- Erro Quadrático Médio Integrado (EQMI),

$$\text{EQMI}(\hat{f}) = \int \text{E}\left\{\left[f(x) - \hat{f}(x)\right]^2\right\} dx.$$

Todavia, essas medidas não podem ser calculadas exatamente, sendo necessária a recorrência de aproximações por métodos numéricos. A metodologia de integração numérica utilizada para o cálculo foi 3/8 de Simpson Composto com 300 pontos. Como para cada amostra é calculado o EQI, uma estimativa do EQMI será dada pela média aritmética dos 1000 EQI calculados em cada modelo simulado.

A seguir, serão detalhadas algumas particularidades de cada metodologia de estimação da janela ótima:

- Sheather e Jones (1991):

  Em praticamente todos os casos foram utilizados 5 estágios para a estimação da janela ótima, excluindo unicamente o modelo normal, em que foram utilizados 3

estágios. Em Glória (2006), é mostrado que, com 5 estágios obtém-se boas estimativas.

- Sain e Scott (1996):

  Este estimador será utilizado apenas nas aplicações, pois é um método dispendioso computacionalmente. A explicação de tal fato está na necessidade de aplicação de métodos de otimização vetorial para estimação da janela ótima.

- Brewer (2000):

  Foram consideradas ambas as propostas *plug-in*, Chiu (1991) e Sheather e Jones (1991), como estimativas para a janela piloto, h.

  A escolha dos hiperparâmetros foi feita conforme a Tabela 4.1 na página 28.

- Gangopadhyay e Cheung (2002):

  A escolha das distribuições *a priori* para h foi feita de duas formas: *priori* informativa e *priori* não informativa. Para a escolha da *priori* não informativa, usou-se a *priori* de Jeffreys's (PAULINO *et al.*, 2003), que neste caso é dada por

  $$\tau(h) \propto h^{-1} I_{(0,+\infty)}(h). \tag{5.1}$$

  Perceba que ao tomarmos $\alpha \to 0$ e $\beta \to \infty$ em (4.20) temos uma distribuição gama-invertida pouco informativa, que se aproxima de uma distribuição de Jeffreys em (5.1).

  No entanto, o método apresentou resultados insatisfatórios devido a esta escolha, como mostram as Figuras 5.1 e 5.2, que representam, respectivamente, o caso da distribuição Normal$(0, 1)$ e o caso da mistura(4).

  Sendo assim, para a determinação da *priori* informativa, escolheu-se α e β de modo que a esperança da distribuição *a priori* estivesse "próxima" da janela ótima teórica com um coeficiente de variação de aproximadamente 15%. Vale ressaltar que neste caso, estamos considerando "próximo" sendo aproximadamente três vezes maior ou aproximadamente três vezes menor, ou seja, a esperança da distribuição *a priori* estará centrada quase em $h_{opt}/3$ ou $3h_{opt}$.

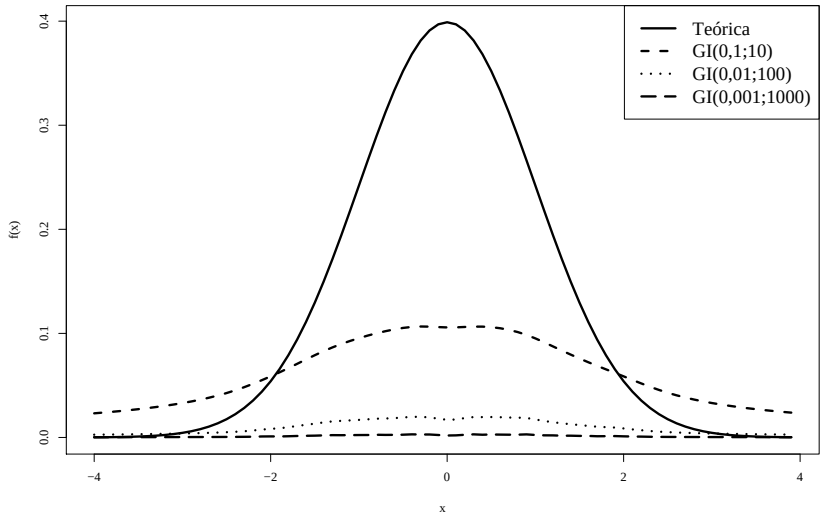

Figura 5.1: Estimativa da distribuição Normal(0,1) para diferentes distribuições *a priori* não informativas, $n = 100$.

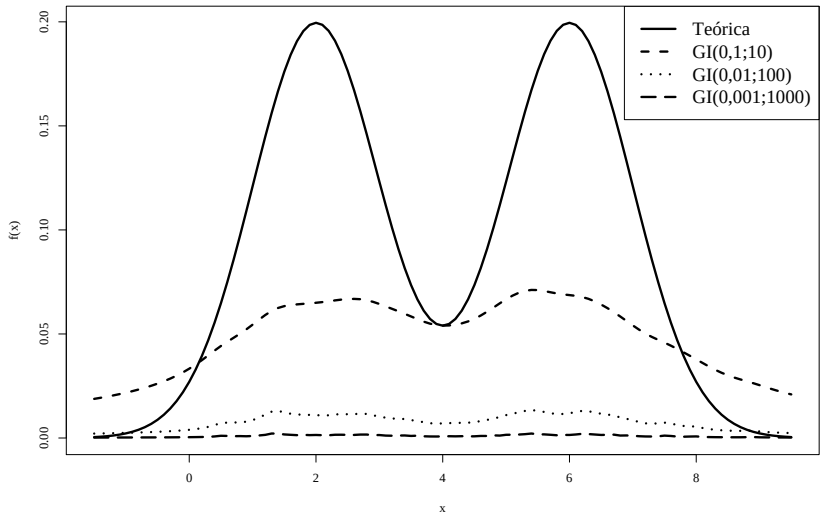

Figura 5.2: Estimativa da distribuição Mistura(4) para diferentes distribuições *a priori* não informativas, $n = 100$.

6 Resultados

Na análise dos resultados, serão mostrados alguns gráficos de *boxplot* para as estimativas da janela ótima, h_{opt}, e para os erros de estimação da função densidade em cada modelo simulado. Para melhor visualização dos gráficos, foram feitas as seguintes abreviações:

- O método *plug-in* de Sheather e Jones (1991) foi nomeado como SJ;
- O método de Brewer (2000) com a janela fixa estimada por Chiu (1991) foi nomeado como BR1 e, no caso em que a estimativa é auferida através de Sheather e Jones (1991), como BR2;
- O método Gangopadhyay e Cheung (2002) foi nomeado como GC1 e GC2, em que cada caso representa uma escolha diferente da distribuição *a priori*. Para GC1, a escolha foi feita de tal forma que a esperança da distribuição *a priori* fosse aproximadamente três vezes menor que a janela ótima teórica ($h_{opt}/3$). No caso de GC2, a escolha foi feita foi feita de tal forma que a esperança da *priori* fosse aproximadamente três vezes maior que a janela ótima teórica ($3h_{opt}$). Em ambos os casos foi utilizado um coeficiente de variação de aproximadamente 15%, como explicitado no Capítulo 5.

A análise dos resultados de simulação seguiram o seguinte padrão:

1. Comparação entre as estimativas da janela ótima obtidas pelo método *plug-in*, para os diferentes tamanhos amostrais;
2. Comparação dos erros das estimativas da função densidade obtidas pelos métodos utilizados em simulação, para os diferentes tamanhos amostrais;
3. Conclusões.

6.1 $X \sim \text{Normal}(0, 1)$

Neste caso, espera-se que todos os métodos apresentem desempenho similar, devido à unimodalidade e à suavidade da densidade normal.

Os resultados relativos a esta metodologia *plug-in* são apresentados a seguir.

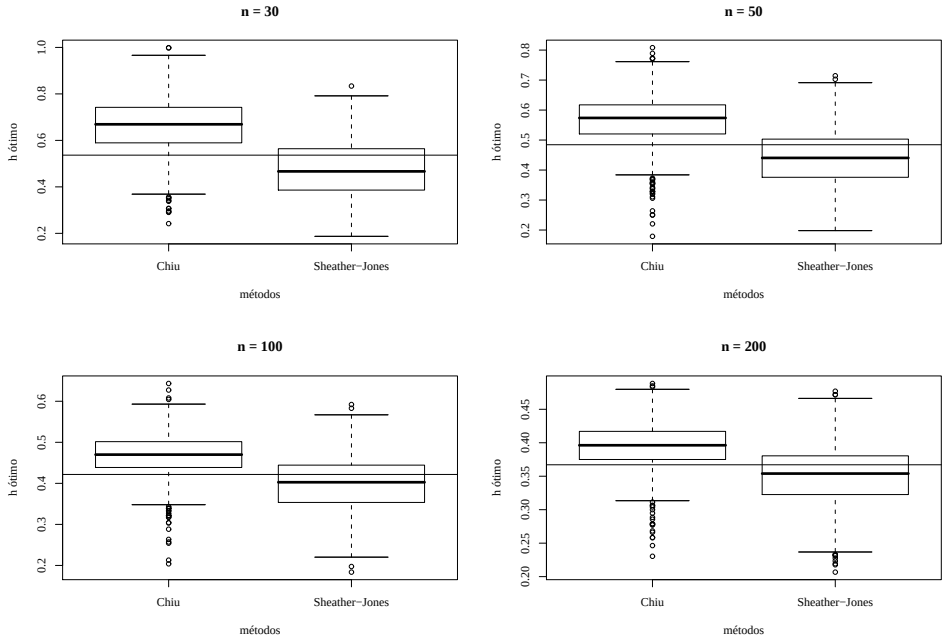

Figura 6.1: Boxplot das estimativas da janela ótima para a distribuição Normal(0,1). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

Tabela 6.1: Estatísticas Descritivas para as estimativas da janela ótima para a distribuição Normal(0,1).

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	0,5364	Chiu	0,6617	0,1143	0,1253
		SJ	0,4736	0,1201	0,0628
50	0,4843	Chiu	0,5673	0,0832	0,0830
		SJ	0,4416	0,0912	0,0427
100	0,4216	Chiu	0,4671	0,0517	0,0455
		SJ	0,3967	0,0652	0,0249
200	0,3670	Chiu	0,3946	0,0339	0,0276
		SJ	0,3505	0,0435	0,0165

A Figura 6.1 mostra que, neste caso, o estimador proposto por Chiu (1991) tende, em

média, a superestimar a janela ótima teórica, enquanto que o estimador de Sheather e Jones (1991) tende, em média, à subestimação. Essas conclusões são reforçadas ao analisarmos as medidas descritivas para as estimativas da janela ótima, que estão apresentadas na Tabela 6.1.

Em relação à variabilidade, o estimador de Chiu (1991) apresentou menor dispersão. Esse comportamento também foi observado em Glória (2006). Todavia, o estimador apresentou um grande número de estimativas discrepantes com respeito à média, quando comparado com o estimador Sheather e Jones (1991).

Uma comparação global entre os métodos implementados pode ser feita através da Tabela 6.2, que apresenta o EQMI, e da Figura 6.2, que apresenta os gráficos *boxplot* dos erros integrados.

Tabela 6.2: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0076	0,0054	0,0030	0,0018
Sheather-Jones	0,0092	0,0062	0,0033	0,0019
Brewer (Chiu) (BR1)	0,0073	0,0055	0,0031	0,0019
Brewer (Sheather-Jones) (BR2)	0,0099	0,0073	0,0039	0,0022
Gangopadhyay-Cheung (GI(99; 0, 15)) (GC1)	0,0134	0,0085	0,0040	0,0021
Gangopadhyay-Cheung (GI(99; 0, 04)) (GC2)	0,0063	0,0045	0,0028	0,0020

A Tabela 6.2 mostra que, para amostras de tamanho 100 e 200, todas as metodologias tiveram um desempenho semelhante, em termos de EQMI. No entanto, para amostras de tamanho 30 e 50, o método de Sheather e Jones (1991) apresentou um desempenho inferior.

O desempenho do estimador de Brewer (2000), quando utilizado com Chiu (1991), acompanhou o desempenho desse método. No entanto, quando utilizado com Sheather e Jones (1991), o estimador apresentou um desempenho inferior para amostras de tamanho 30, 50 e 100.

O estimador de Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 15), também

teve um desempenho inferior aos demais para amostras de tamanho 30, 50 e 100. No entanto, este comportamento pode ser arrogado à sensibilidade do método em relação à especificação dos hiperparâmetros da distribuição *a priori*, dado que utilizando uma *priori* GI$(99; 0, 04)$ a metodologia mostrou-se superior aos demais para amostras de tamanho 30, 50 e 100. Todas as conclusões são reforçadas através da análise da Figura 6.2.

De acordo com os resultados apresentados, verificou-se que os diversos estimadores são equivalentes para amostras maiores que 100, em termos de erro. Além disso, constatou-se que, sob as condições simuladas, as metodologias de janela variável não apresentaram ampla melhora em relação aos métodos de janela fixa. Este fato, possivelmente, pode ser explicado pela suavidade da distribuição normal, isto é, a quantidade de suavização a ser feita é similar em todas as partes da função.

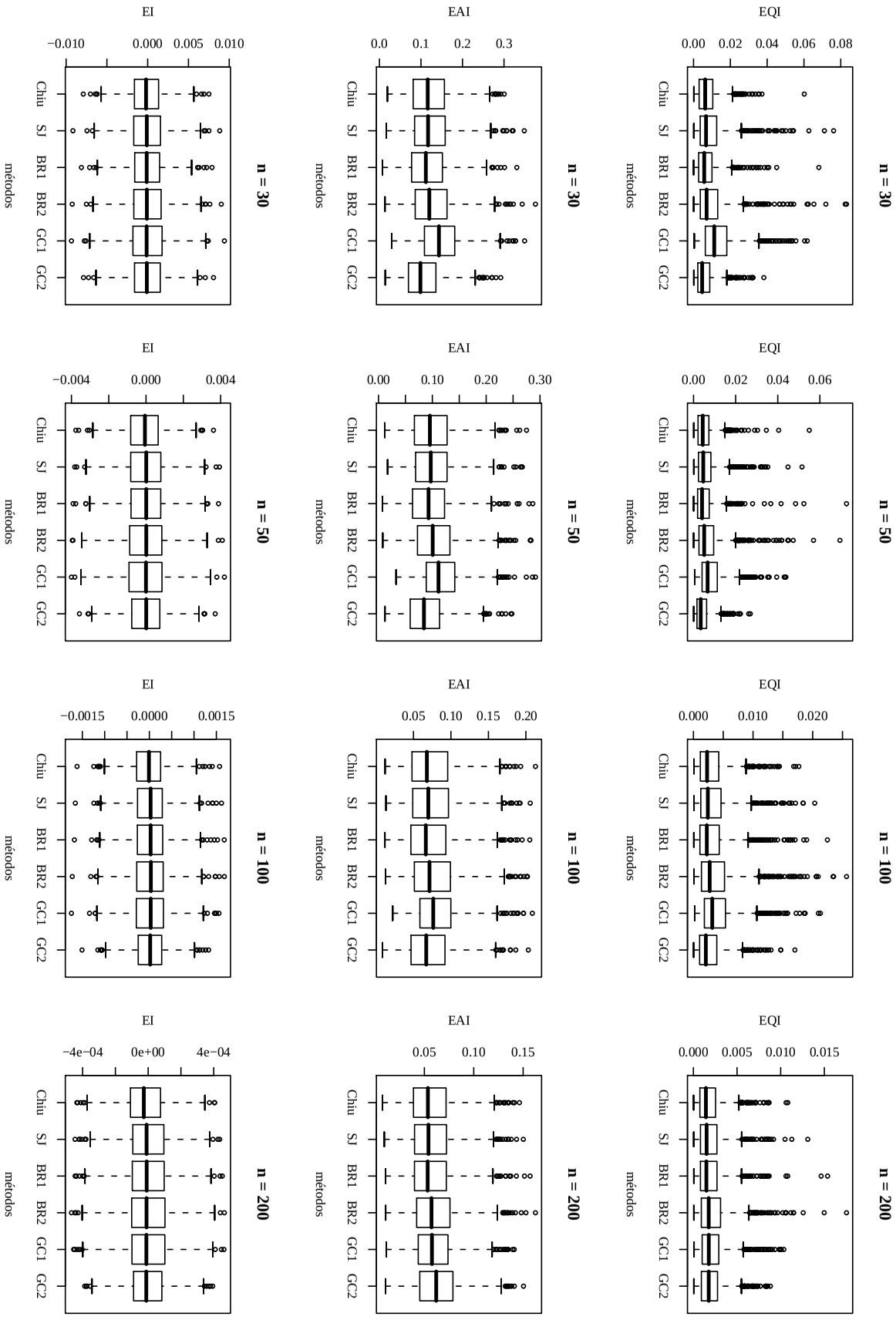

Figura 6.2: Boxplot do Erros Integrados ao estimar a distribuição Normal(0,1).

6.2 $X \sim$ Gama(4, 2)

Neste caso espera-se que os métodos de janela variável apresentem melhor desempenho, devido à assimetria da densidade gama. Os resultados relativos à metodologia *plug-in* são apresentados a seguir.

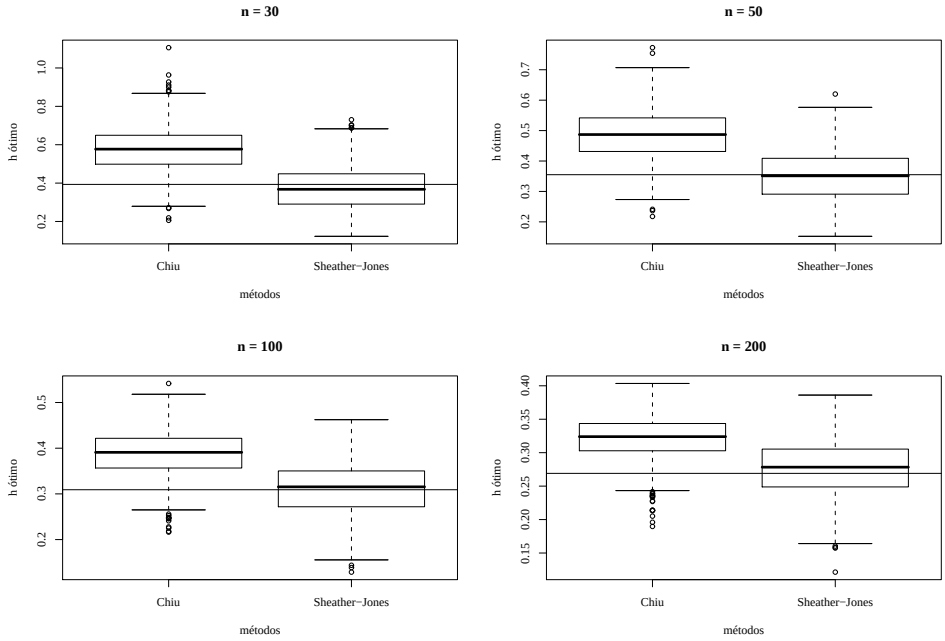

Figura 6.3: Boxplot das estimativas da janela ótima para a distribuição Gama(4,2). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

Tabela 6.3: Estatísticas Descritivas para as estimativas da janela ótima para a distribuição Gama(4,2)

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	0,3932	Chiu	0,5757	0,1153	0,1825
		SJ	0,3748	0,1077	0,0184
50	0,3550	Chiu	0,4842	0,0799	0,1292
		SJ	0,3520	0,0804	0,0030
100	0,3090	Chiu	0,3881	0,0498	0,0791
		SJ	0,3112	0,0560	0,0022
200	0,2690	Chiu	0,3218	0,0321	0,0528
		SJ	0,2755	0,0391	0,0065

A Figura 6.3 e a Tabela 6.3 mostram que, assim como no caso da distribuição normal

padrão, o estimador proposto por Chiu (1991) tende, em média, a superestimar a janela ótima. Entretanto, neste caso o estimador de Sheather e Jones (1991) tende, em média, para o ótimo teórico baseado no EQMIA. Em relação a dispersão, o estimador de Sheather e Jones (1991) apresentou maior variabilidade para amostras de tamanho 100 e 200.

Uma comparação global entre as metodologias implementadas pode ser feita através da Tabela 6.4, que apresenta o EQMIA, e da Figura 6.4, que apresenta os gráficos *boxplot* dos erros integrados.

Tabela 6.4: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0198	0,0137	0,0081	0,0048
Sheather-Jones	0,0252	0,0157	0,0091	0,0051
Brewer (Chiu) (BR1)	0,0195	0,0138	0,0086	0,0051
Brewer (Sheather-Jones) (BR2)	0,0275	0,0186	0,0108	0,0060
Gangopadhyay-Cheung (GI(99; 0, 28)) (GC1)	0,0392	0,0235	0,0120	0,0059
Gangopadhyay-Cheung (GI(99; 0, 07)) (GC2)	0,0179	0,0118	0,0072	0,0050

A Tabela 6.4 e a Figura 6.4 mostram que, para amostras de tamanho 30 e 50, o estimador de Chiu (1991), o método de Brewer (2000) com janela piloto estimada através de Chiu (1991) e o estimador de Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 07) tiveram os melhores desempenhos.

Vale notar que, apesar do estimador Sheather e Jones (1991) estar muito próximo do ótimo teórico, a metodologia não apresentou bons resultados perante aos outros estimadores para amostras de tamanho 30 e 50. Uma explicação mais detalhada ou conclusiva para tal fato não foi encontrada durante o trabalho, havendo assim necessidade de estudos futuros para melhor entendimento deste fenômemo.

Ainda pela Tabela 6.4 podemos notar que, se utilizarmos uma boa estimativa da janela piloto, o método de Brewer (2000) tende a apresentar bons resultados, enquanto que se a estimativa não for boa, o método tende a ter resultados piores.

O estimador de Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 28), também teve

um desempenho inferior. No entanto, novamente, este comportamento pode ser arrogado à sensibilidade do método em relação à especificação dos hiperparâmetros da distribuição *a priori*, dado que, utilizando uma *priori* GI$(99; 0, 07)$, a metodologia mostrou-se equivalente ou superior as demais.

De acordo com os resultados apresentados, verificou-se que os diversos estimadores são equivalentes para amostras de tamanho 200, em termos de erro. Além disso, constatou-se que, sob as condições simuladas, para amostras menores que 200, a metodologia de janela variável de Gangopadhyay e Cheung (2002) com *priori* GI$(99; 0, 28)$ apresentou uma melhora significativa em relação aos métodos de janela fixa. Este fato, possivelmente, pode ser explicado pela assimetria da distribuição gama, isto é, a quantidade de suavização a ser feita varia para diferentes partes da função.

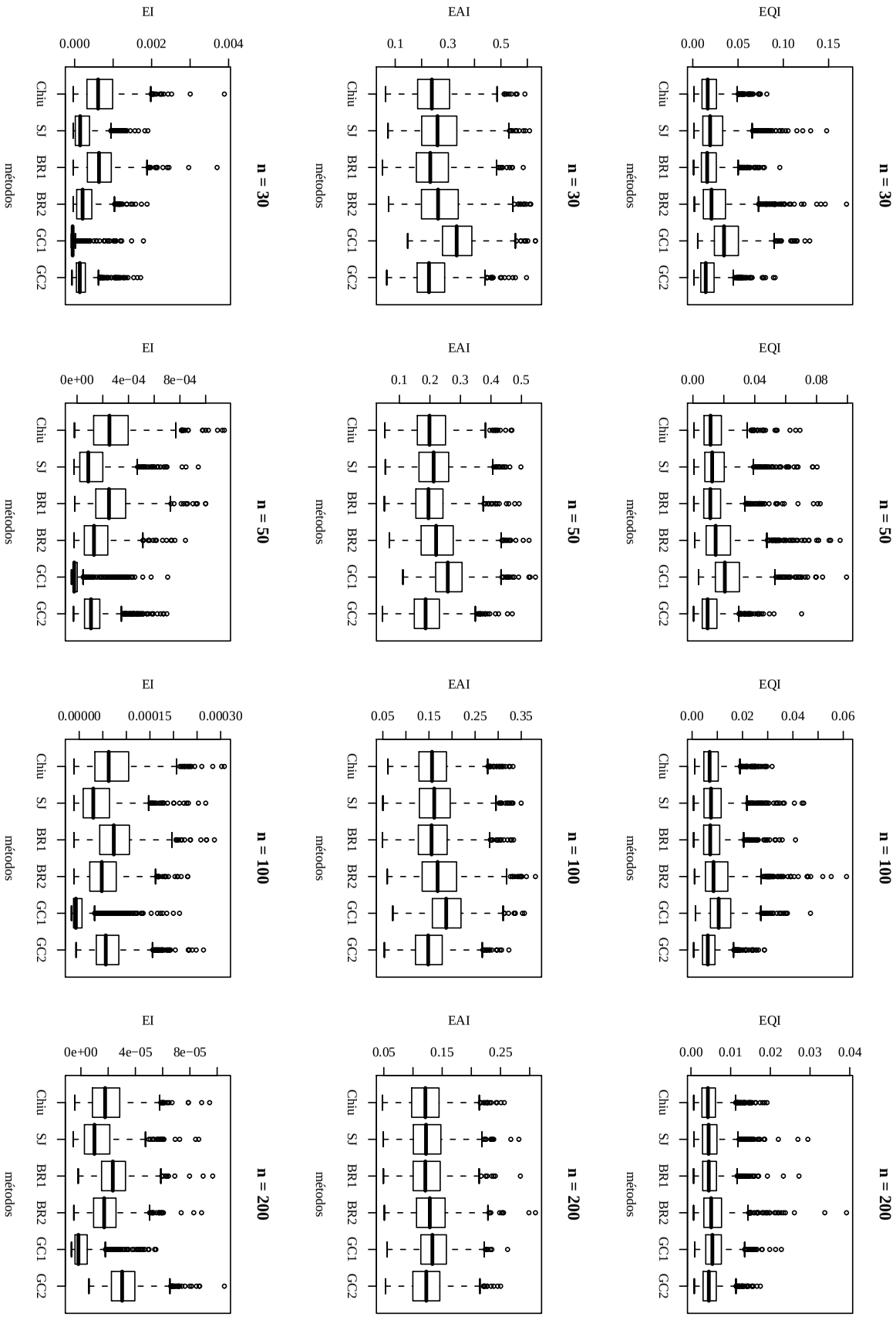

Figura 6.4: Boxplot do Erros Integrados ao estimar a distribuição Gama(4,2)

6.3 $X \sim$ Weibull$(1, 6)$

Assim como no caso da distribuição Gama, neste caso espera-se que os métodos de janela variável apresentem melhor desempenho, devido à assimetria da densidade Weibull. Os resultados relativos à metodologia *plug-in* são apresentados a seguir.

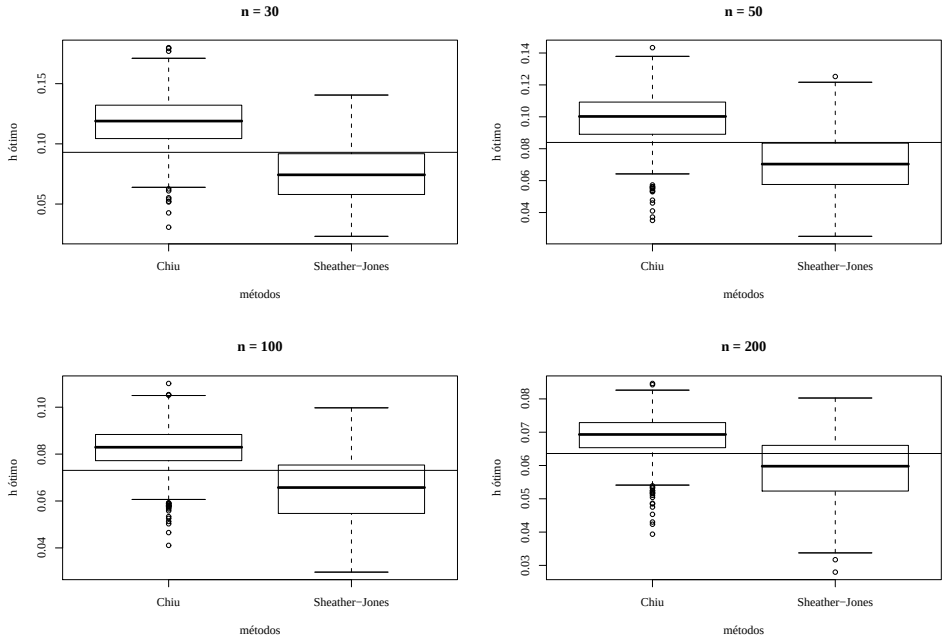

Figura 6.5: Boxplot das estimativas da janela ótima para a distribuição Weibull(1,6). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

Tabela 6.5: Estatísticas Descritivas para as estimativas da janela ótima para a distribuição Weibull(1,6)

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	0,0930	Chiu	0,1176	0,0213	0,0246
		SJ	0,0755	0,0230	0,0175
50	0,0839	Chiu	0,0987	0,0150	0,0148
		SJ	0,0708	0,0181	0,0131
100	0,0731	Chiu	0,0823	0,0091	0,0092
		SJ	0,0650	0,0135	0,0081
200	0,0636	Chiu	0,0687	0,0059	0,0051
		SJ	0,0586	0,0094	0,0050

A Figura 6.5 e a Tabela 6.5 mostram que, neste caso, o estimador proposto por Chiu

(1991) tende, em média, a superestimar a janela ótima teórica, enquanto que o estimador de Sheather e Jones (1991) tende, em média, à subestimação da janela ótima.

Em relação a variabilidade, ambos estimadores apresentaram desempenho similar para amostra de tamanho 30. Para os demais tamanhos amostrais, o estimador Sheather e Jones (1991) apresentou maior dispersão, assim como foi observado nos estudos de Glória (2006).

Assim como nos modelos Normal e Gama, podemos observar que o estimador de Chiu (1991) apresenta maior vício que o estimador de Sheather e Jones (1991), porém esse acréscimo de vício do método de Chiu (1991) é compensado por uma menor variabilidade.

Uma comparação global entre as metodologias implementadas pode ser feita através da Tabela 6.6, que apresenta o EQMI, e da Figura 6.6, que apresenta os gráficos *boxplot* dos erros integrados.

Tabela 6.6: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0833	0,0547	0,0334	0,0204
Sheather-Jones	0,1218	0,0754	0,0416	0,0234
Brewer (Chiu) (BR1)	0,0821	0,0596	0,0365	0,0224
Brewer (Sheather-Jones) (BR2)	0,1339	0,0920	0,0509	0,0279
Gangopadhyay-Cheung (GI(99; 5)) (GC1)	0,1564	0,0937	0,0478	0,0248
Gangopadhyay-Cheung (GI(99; 1, 25)) (GC2)	0,0704	0,0448	0,0309	0,0222

A Tabela 6.6 e a Figura 6.6 mostram que, para amostras de tamanho 200, as metodologias tiveram um desempenho semelhante, em termos de EQMI. Para amostras de tamanho 30, 50 e 100, o estimador de Chiu (1991) e o estimador de Gangopadhyay e Cheung (2002) com *priori* GI(99; 1, 25) tiveram os melhores desempenhos, enquanto que o estimador de Sheather e Jones (1991) e Gangopadhyay e Cheung (2002) com *priori* GI(99; 5) apresentaram desempenho inferior.

O desempenho da metodologia de Brewer (2000) novamente seguiu o seguinte padrão: se utilizarmos uma boa estimativa da janela piloto, o método de tende a apresentar bons

resultados, enquanto que se a estimativa não for boa o método tende a ter resultados piores.

De acordo com os resultados apresentados, verificou-se que os diversos estimadores são equivalentes para amostras de tamanho 200, em termos de erro. Constatou-se que sob as condições simuladas para amostras menores que 200, a metodologia de janela variável de Gangopadhyay e Cheung (2002) com *priori* GI$(99; 1, 25)$ apresentou uma melhora significativa em relação aos métodos de janela fixa. Este fato, possivelmente, pode ser explicado pela assimetria da distribuição Weibull.

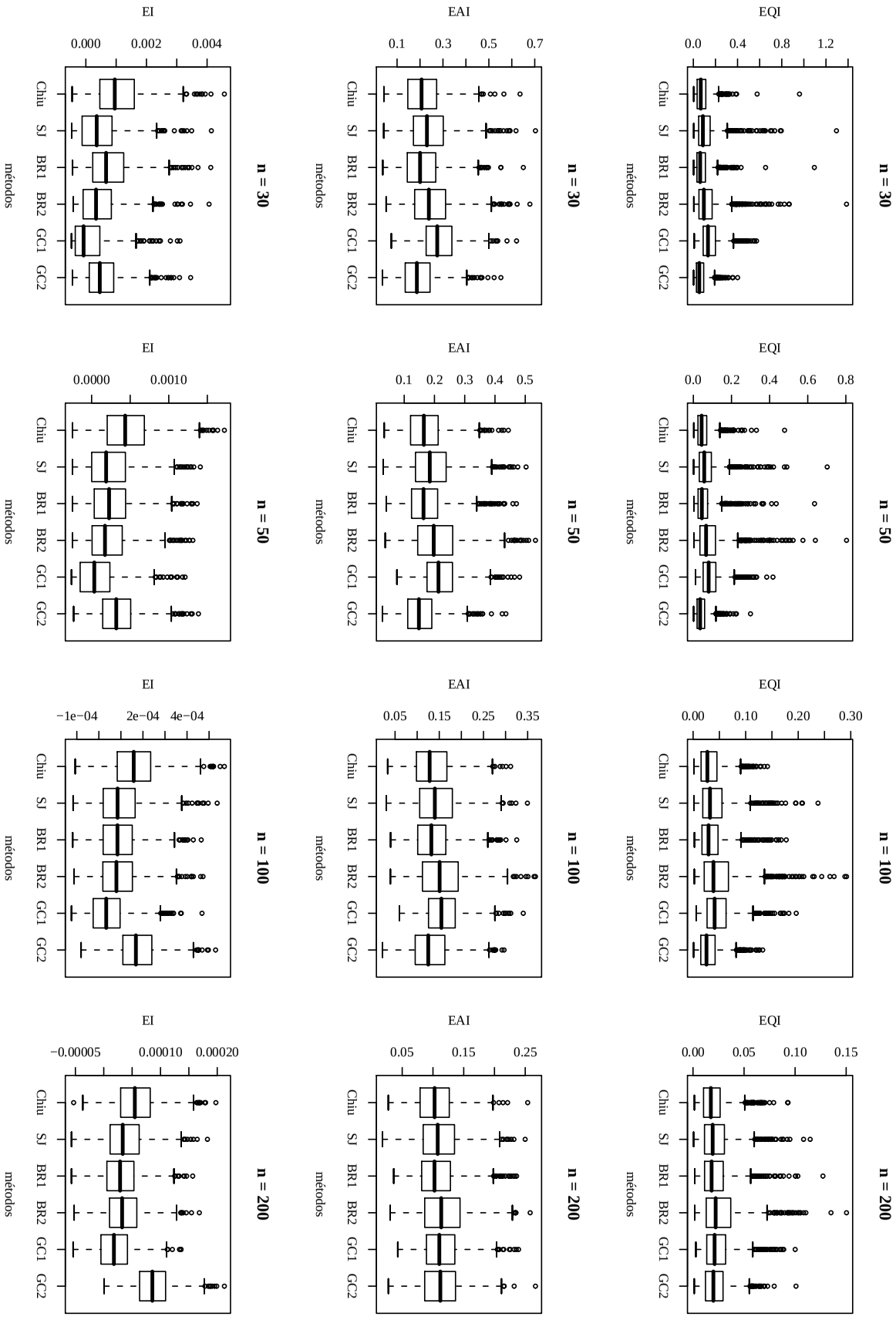

Figura 6.6: Boxplot do Erros Integrados ao estimar a distribuição Weibull(1,6)

6.4 $X \sim$ Qui-Quadrado(7)

Novamente, neste caso espera-se que os métodos de janela variável apresentem melhor desempenho, por causa da assimetria da distribuição. Os resultados relativos à metodologia *plug-in* são apresentados a seguir.

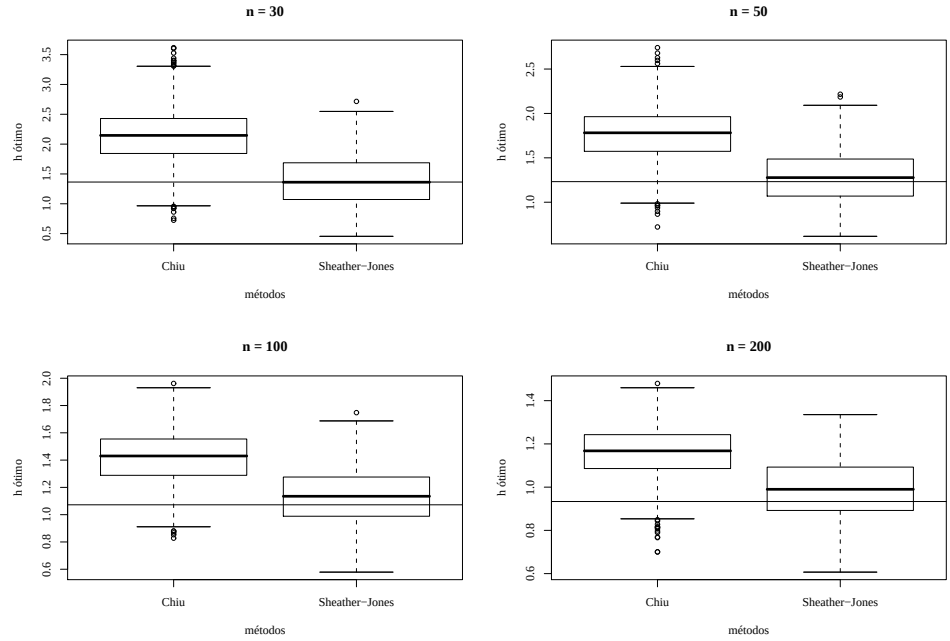

Figura 6.7: Boxplot das estimativas da janela ótima para a distribuição Qui-Quadrado(7). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

Tabela 6.7: Estatísticas Descritivas para as estimativas da janela ótima para a distribuição Qui-Quadrado(7)

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	1,3643	Chiu	2,1443	0,4458	0,7800
		SJ	1,3935	0,4078	0,0292
50	1,2318	Chiu	1,7721	0,3001	0,5403
		SJ	1,2846	0,2949	0,0528
100	1,0723	Chiu	1,4205	0,1940	0,3482
		SJ	1,1327	0,2082	0,0604
200	0,9335	Chiu	1,1597	0,1196	0,2262
		SJ	0,9865	0,1397	0,0530

A Figura 6.7 e a Tabela 6.7 mostram que, neste caso, ambos estimadores tendem, em

média, a superestimar a janela ótima teórica. Entretanto, a superestimação é maior no método de Chiu (1991).

No estimador de Sheather e Jones (1991) houve um pequeno incremento na superestimação com o aumento da amostra, isto é, o vício do estimador se tornou maior com aumento do amostra. Em relação à variabilidade, ambos os estimadores apresentaram desempenho similar para amostras de tamanho 50 e 100.

Uma comparação global entre as metodologias implementadas pode ser feita através da Tabela 6.8, que apresenta o EQMI, e da Figura 6.8, que apresenta os gráficos *boxplot* dos erros integrados.

Tabela 6.8: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0059	0,0038	0,0024	0,0014
Sheather-Jones	0,0073	0,0044	0,0026	0,0015
Brewer (Chiu) (BR1)	0,0059	0,0040	0,0025	0,0014
Brewer (Sheather-Jones) (BR2)	0,0079	0,0053	0,0031	0,0017
Gangopadhyay-Cheung (GI(99; 0, 02)) (GC1)	0,0108	0,0064	0,0033	0,0017
Gangopadhyay-Cheung (GI(99; 0, 006)) (GC2)	0,0055	0,0035	0,0021	0,0013

A Tabela 6.8 e a Figura 6.8 mostram que para amostras de tamanho 200, as metodologias tiveram um desempenho semelhante, em termos de EQMI. Para amostras de tamanho 30 e 50, o estimador de Chiu (1991) e o estimador de Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 006) tiveram os melhores desempenhos, enquanto que o estimador de Sheather e Jones (1991) e Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 02) apresentaram desempenho inferior.

Assim como no caso Gama e Weibull, a metodologia de Brewer (2000) teve o mesmo comportamento em relação à janela piloto.

De acordo com os resultados apresentados, verificou-se que os diversos estimadores são equivalentes para amostras de tamanho 200, em termos de erro. Além disso, constatou-se que, sob as condições simuladas, as metodologias de janela variável não apresentaram

melhora significativa em relação aos métodos de janela fixa. Esse resultado, possivelmente, pode ser explicada pela fraca assimetria da distribuição qui-quadrado, isto é, a quantidade de suavização a ser feita não varia muito para diferentes partes da função.

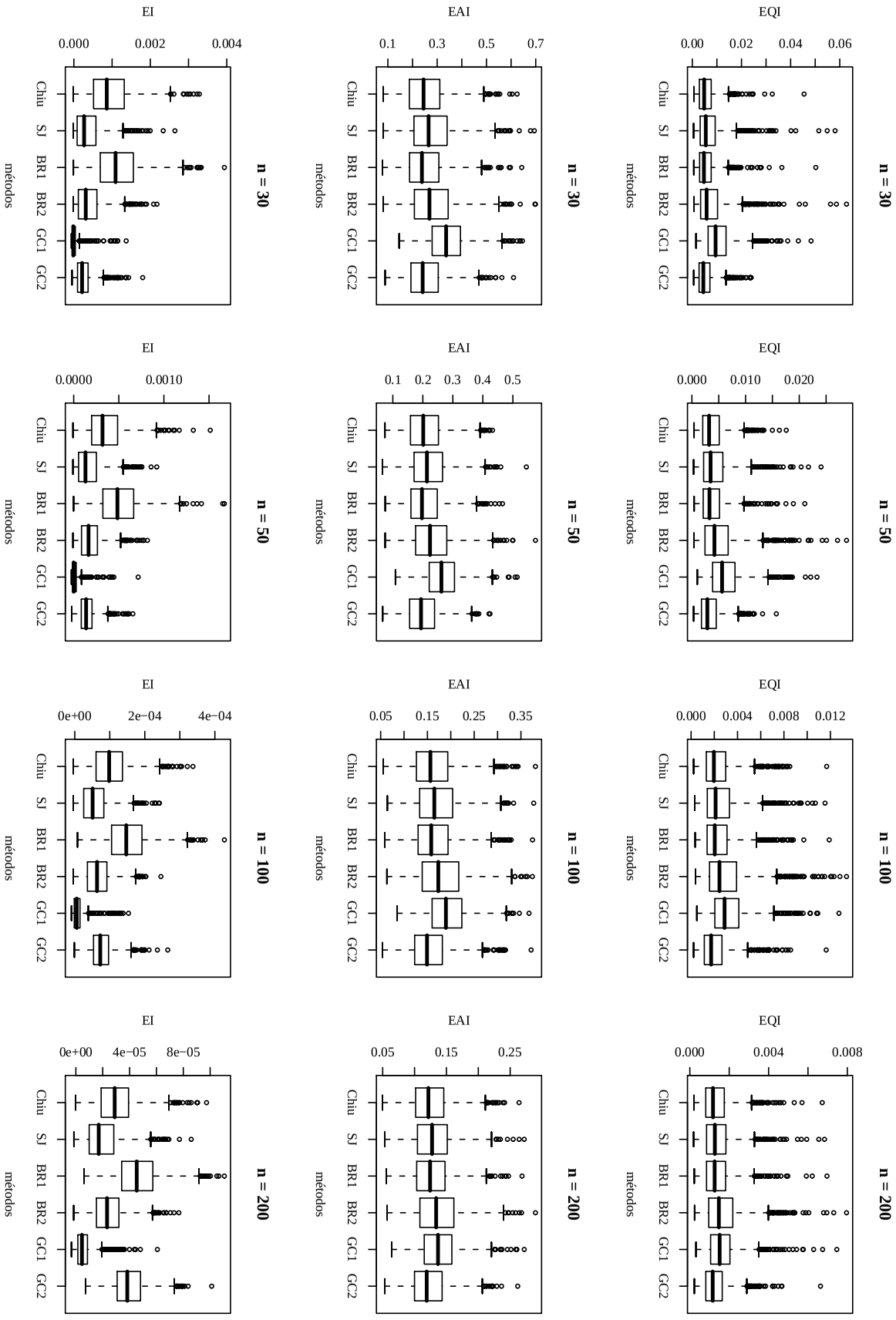

Figura 6.8: Boxplot do Erros Integrados ao estimar a distribuição Qui-Quadrado(7)

6.5 $X \sim \frac{1}{5}\text{Normal}(0,1)+\frac{1}{5}\text{Normal}\left(\frac{1}{4},\frac{4}{9}\right)+\frac{3}{5}\text{Normal}\left(\frac{13}{12},\frac{25}{81}\right)$

Assim como nos modelos assimétricos anteriores, neste caso espera-se que a metodologia de janela variável apresente melhores resultados do que os métodos de janela fixa. Os resultados relativos à metodologia *plug-in* são apresentados a seguir.

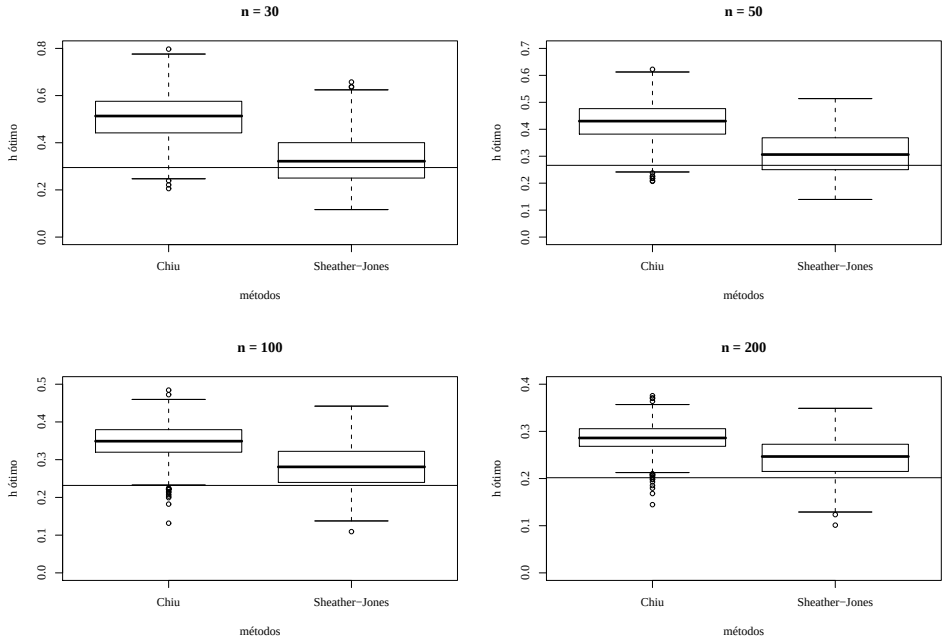

Figura 6.9: Boxplot das estimativas da janela ótima para a Mistura(1). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

Tabela 6.9: Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(1)

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	0,2948	Chiu	0,5103	0,1019	0,2155
		SJ	0,3288	0,1030	0,0340
50	0,2661	Chiu	0,4270	0,0698	0,1690
		SJ	0,3098	0,0771	0,0437
100	0,2317	Chiu	0,3465	0,0454	0,1148
		SJ	0,2805	0,0565	0,0488
200	0,2017	Chiu	0,2856	0,0292	0,0839
		SJ	0,2444	0,0405	0,0427

A Figura 6.9 e a Tabela 6.9 mostram que, neste caso, ambos os estimadores tendem,

em média, a superestimar a janela ótima teórica. Entretanto, a superestimação é maior no método de Chiu (1991).

Para amostras de tamanho 30, ambos os estimadores apresentaram variabilidade similar, enquanto que para amostras maiores, o método de Chiu (1991) apresentou menor variabilidade.

Uma comparação global entre os métodos implementados pode ser feita através da Tabela 6.10, que apresenta o EQMI, e da Figura 6.10, que apresenta os gráficos *boxplot* dos erros integrados.

Tabela 6.10: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0196	0,0128	0,0079	0,0045
Sheather-Jones	0,0245	0,0143	0,0085	0,0048
Brewer (Chiu) (BR1)	0,0185	0,0120	0,0077	0,0045
Brewer (Sheather-Jones) (BR2)	0,0268	0,0167	0,0098	0,0056
Gangopadhyay-Cheung (GI(99; 0, 2)) (GC1)	0,0246	0,0141	0,0077	0,0041
Gangopadhyay-Cheung (GI(99; 0, 06)) (GC2)	0,0146	0,0105	0,0082	0,0066

A Tabela 6.10 mostra que, para amostras de tamanho 100 e 200, as metodologias tiveram um desempenho semelhante, em termos de EQMI. Para amostras de tamanho 30 e 50, o estimador de Chiu (1991) e o estimador de Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 006) tiveram os melhores desempenhos, enquanto que o estimador de Sheather e Jones (1991) e Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 02) apresentaram desempenho inferior.

O desempenho da metodologia de Brewer (2000) apresentou o mesmo padrão de comportamento que teve nos modelos simulados anteriormente. As conclusões apresentadas podem ser reforçadas através da análise da Figura 6.10.

De acordo com os resultados apresentados, verificou-se que os diversos estimadores são equivalentes para amostras maiores que 100, em termos de erro. Como era de se esperar, constatou-se que, sob as condições simuladas, a metodologia de janela variável apresentou

uma melhora em relação aos métodos de janela fixa. Além disso, constatou-se novamente que o estimador de Chiu (1991) tende a apresentar maior vício e menor variabilidade que o estimador de Sheather e Jones (1991).

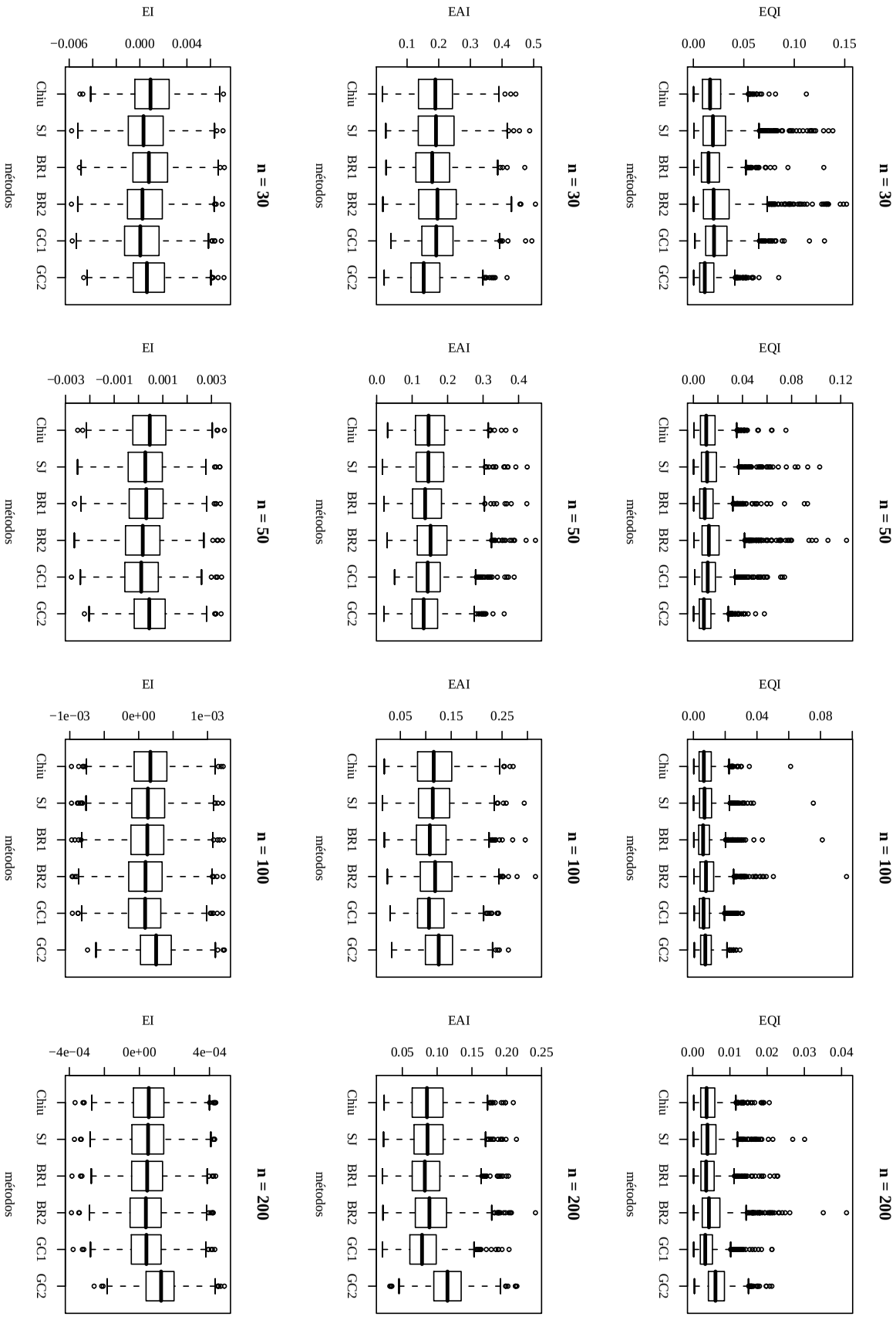

Figura 6.10: Boxplot do Erros Integrados ao estimar a Mistura(1)

6.6 $X \sim \frac{1}{2}\text{Normal}(2,1) + \frac{1}{2}\text{Normal}(5,1)$

Neste caso, espera-se que os métodos de janela variável apresentem melhor desempenho, devido a bi-modalidade da função densidade. Os resultados relativos à metodologia *plug-in* são apresentados a seguir.

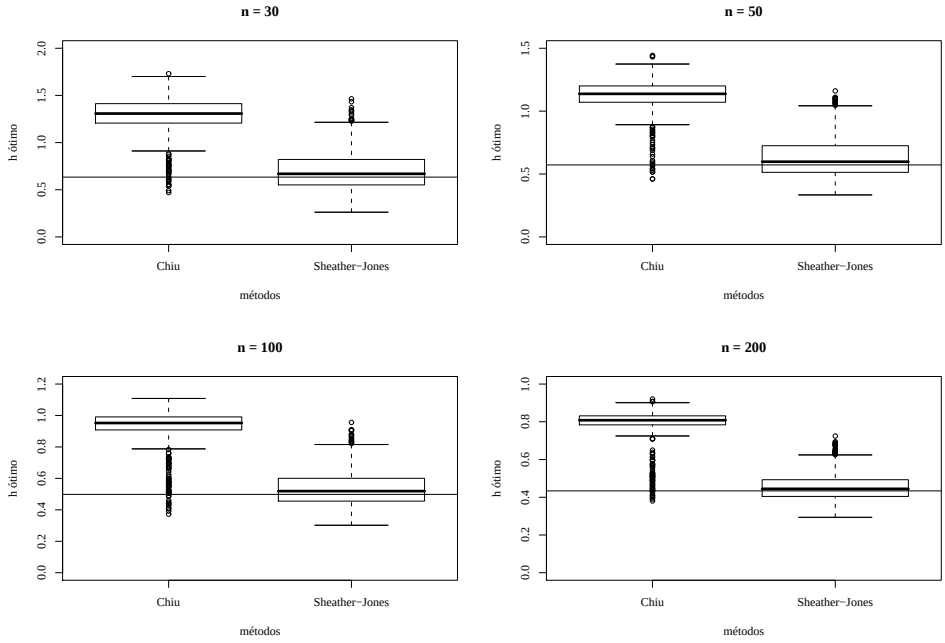

Figura 6.11: Boxplot das estimativas da janela ótima para a Mistura(2). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

Tabela 6.11: Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(2)

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	0,6339	Chiu	1,2935	0,1828	0,6596
		SJ	0,7034	0,2014	0,0695
50	0,5724	Chiu	1,1254	0,1247	0,5530
		SJ	0,6307	0,1571	0,0583
100	0,4983	Chiu	0,9300	0,1106	0,4317
		SJ	0,5379	0,1110	0,0396
200	0,4338	Chiu	0,7941	0,0791	0,3603
		SJ	0,4530	0,0710	0,0192

A Figura 6.11 e a Tabela 6.11 mostram que, neste caso, ambos os estimadores tendem,

em média, a superestimar a janela ótima teórica. Entretanto, a superestimação é maior no método de Chiu (1991), como já foi observado anteriormente, isto é, o estimador tende a apresentar maior vício.

O motivo de tal fato está vinculado ao limite de integração Λ, na expressão (4.4). Chiu (1991) propõe que este limite seja $\Lambda = \min\left\{\lambda : |\hat{\varphi}(\lambda)|^2 \leq \frac{c}{n}\right\}$, para alguma constante $c > 1$.

No entanto, ao analisarmos a Figura 6.12, que apresenta o comportamento de $|\hat{\varphi}(\lambda)|^2$ para diversos tamanhos amostrais, notamos que pode haver perda de informação ao tomarmos $\Lambda = \min\left\{\lambda : |\hat{\varphi}(\lambda)|^2 \leq \frac{c}{n}\right\}$.

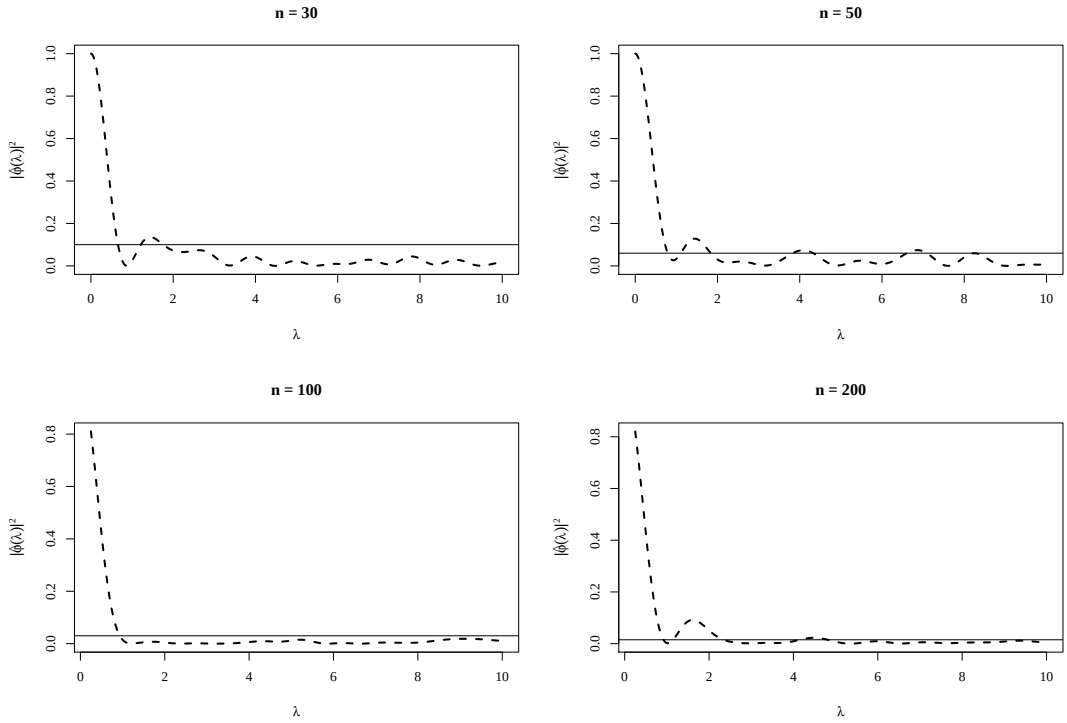

Figura 6.12: Função característica empírica pra diferentes tamanhos amostrais. A linha contínua representa o ponto de corte $\frac{c}{n}$, em que $c = 3$.

Portanto, neste caso o método tende a subestimar o valor de $R(f'')$, expressão (3.6). Conseqüentemente, o estimador propenderá a superestimar a janela ótima.

Em relação à variabilidade, ambos os estimadores apresentaram desempenho similar para amostras de tamanho 100, enquanto que para os casos 30 e 50, o método de Chiu (1991) obteve menor variabilidade. Todavia, este método apresentou um grande número de estimativas discrepantes, quando comparado com Sheather e Jones (1991). No caso de amostras de tamanho 200, o estimador Sheather e Jones (1991) apresentou menor dispersão.

Uma comparação global entre os métodos implementados pode ser feita através da Tabela 6.12, que apresenta o EQMI, e da Figura 6.13, que apresenta os gráficos *boxplot* dos erros integrados.

Tabela 6.12: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0132	0,0099	0,0071	0,0047
Sheather-Jones	0,0141	0,0090	0,0057	0,0033
Brewer (Chiu) (BR1)	0,0131	0,0089	0,0059	0,0035
Brewer (Sheather-Jones) (BR2)	0,0153	0,0104	0,0065	0,0037
Gangopadhyay-Cheung (GI(99; 0, 11)) (GC1)	0,0263	0,0151	0,0080	0,0040
Gangopadhyay-Cheung (GI(99; 0, 03)) (GC2)	0,0130	0,0079	0,0050	0,0031

A Tabela 6.12 mostra que para amostras de tamanho 30 e 50, o estimador de Chiu (1991) apresentou um bom desempenho, ou seja, não houve perda de informação ao determinar o limite de integração Λ. Contudo, para os tamanhos de amostra 100 e 200, essa determinação se mostrou deficitária, ocasionando um desempenho inferior às demais metodologias.

A Tabela 6.12 também mostra que o método de Brewer (2000), quando utilizado conjuntamente com Chiu (1991), apresentou bons resultados. Porém, quando utilizado com Sheather e Jones (1991), o estimador apresentou resultados inferiores, em termos de EQMI, para amostras de tamanho menores que 200.

A abordagem de Gangopadhyay e Cheung (2002) com *priori* GI$(99; 0, 03)$ apresentou os melhores resultados, em termos de erro, para amostras maiores que 30. Quando o método foi utilizado com uma distribuição *a priori* GI$(99; 0, 11)$, o estimador apresentou resultados inferiores.

As conclusões expostas são reforçadas através da análise da Figura 6.13.

De acordo com os resultados apresentados, verificou-se que o método de Chiu (1991) teve um comportamento inferior aos demais para amostras de tamanho 100 e 200. Novamente, o método de Gangopadhyay e Cheung (2002) apresentou bons resultados, quando

utilizado com IG$(99; 0, 03)$.

Neste caso, o método de Brewer (2000) utilizado em conjunto com Chiu (1991) apresentou melhora nos resultados quando comparado ao método Chiu (1991). Quando utilizado com Sheather e Jones (1991), o método teve resultados inferiores para amostras de tamanho 30 e 50.

Vale salientar que as metodologias de janela fixa tiveram um desempenho acima do esperado, uma vez que a suavização a ser feita varia em diferentes partes da função densidade. O estimador de Sheather e Jones (1991) apresentou resultados próximos à metodologia de janela variável para amostras maiores do que 30. Contrariamente, o estimador de Chiu (1991) apresentou um bom resultado para amostra de tamanho 30.

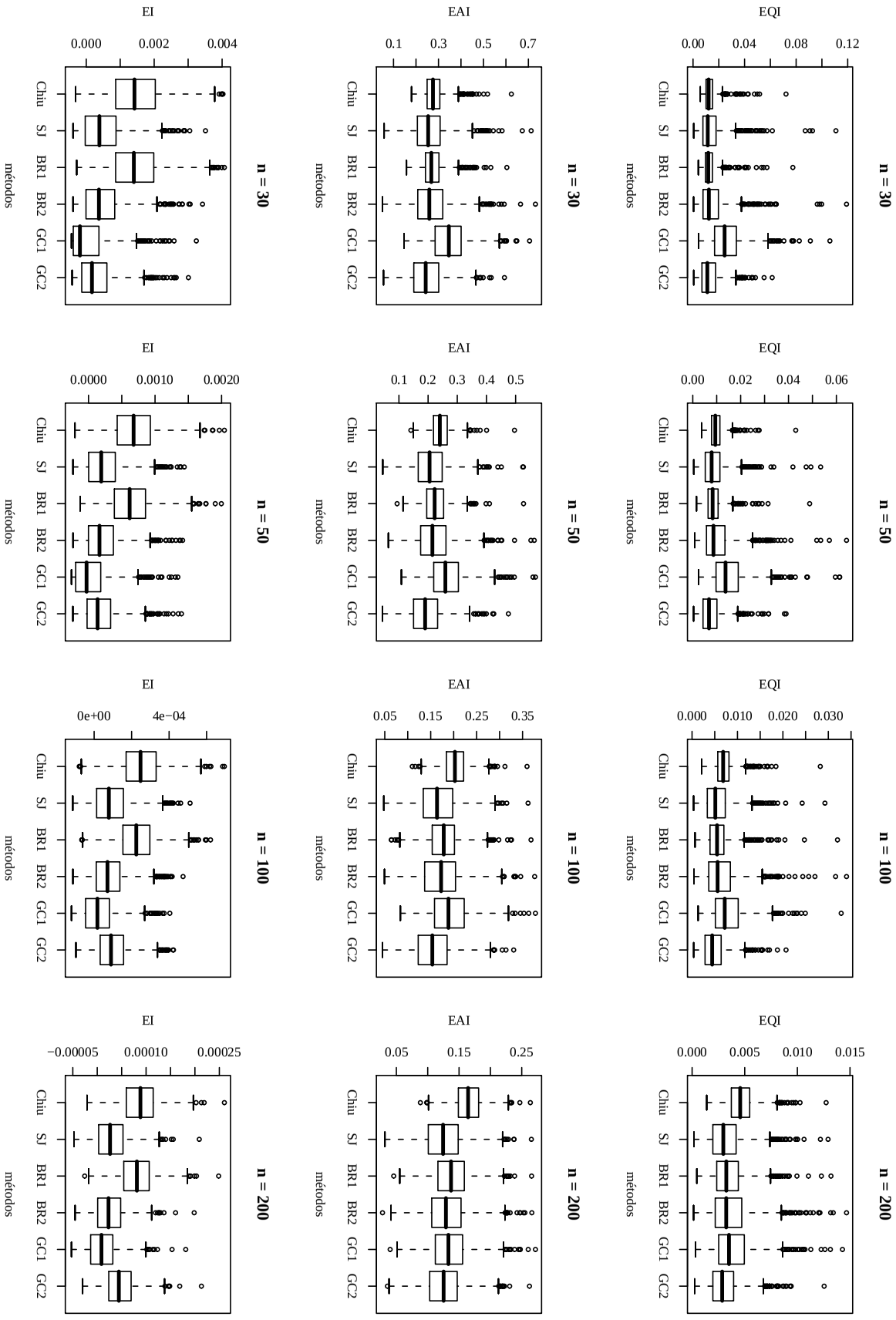

Figura 6.13: Boxplot do Erros Integrados ao estimar a Mistura(2)

6.7 $X \sim \frac{1}{2}\text{Normal}(2,1) + \frac{1}{2}\text{Normal}(6,1)$

Neste caso, também se espera que os métodos de janela variável apresentem melhor desempenho, devido à bi-modalidade da função densidade. Os resultados relativos à metodologia *plug-in* para os diferentes tamanhos amostrais que foram simulados são apresentados a seguir.

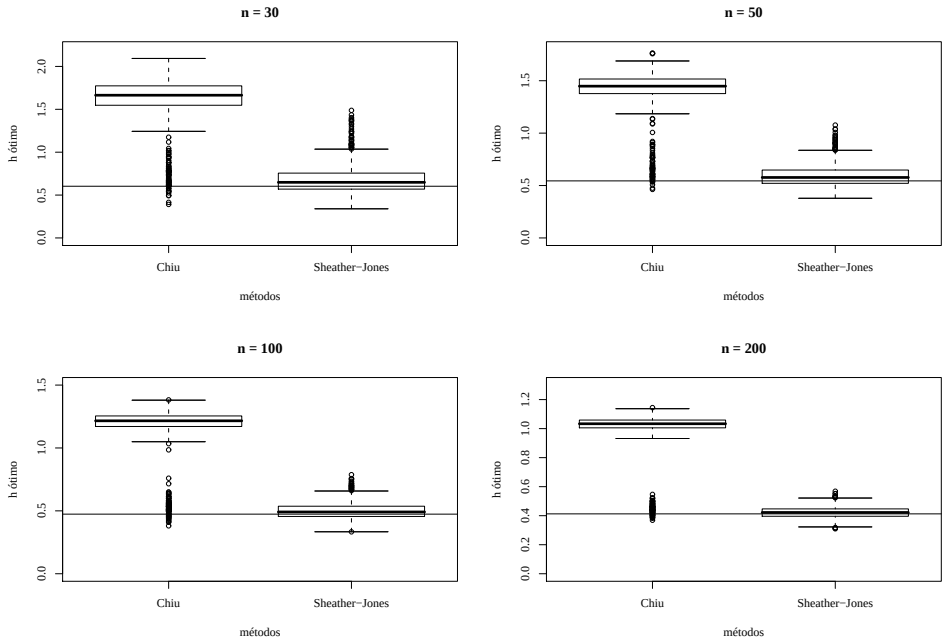

Figura 6.14: Boxplot das estimativas da janela ótima para a Mistura(3). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

Tabela 6.13: Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(3)

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	0,6029	Chiu	1,6052	0,2904	1,0023
		SJ	0,6805	0,1653	0,0776
50	0,5443	Chiu	1,4079	0,2114	0,8636
		SJ	0,5921	0,1033	0,0478
100	0,4739	Chiu	1,1594	0,2029	0,6855
		SJ	0,4988	0,0665	0,0249
200	0,4125	Chiu	0,9894	0,1651	0,5760
		SJ	0,4228	0,0398	0,0103

A Figura 6.14 e a Tabela 6.13 mostram que, neste caso, o método de Chiu (1991) tende a superestimar, excessivamente, a janela ótima teórica. O estimador também apresentou alta variabilidade e um excesso de estimativas discrepantes para a janela ótima. Neste caso, o método hipoteticamente ficou comprometido devido à forma de determinação do limite de integração Λ.

Neste caso, o estimador Sheather e Jones (1991) apresentou desempenho amplamente superior ao estimador Chiu (1991).

Uma comparação global entre as metodologias implementadas pode ser feita através da Tabela 6.14, que apresenta o EQMI, e da Figura 6.15, que apresenta os gráficos *boxplot* dos erros integrados.

Tabela 6.14: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0255	0,0209	0,0156	0,0111
Sheather-Jones	0,0155	0,0099	0,0063	0,0036
Brewer (Chiu) (BR1)	0,0243	0,0176	0,0117	0,0076
Brewer (Sheather-Jones) (BR2)	0,0162	0,0108	0,0069	0,0039
Gangopadhyay-Cheung (GI(99; 0, 12)) (GC1)	0,0281	0,0162	0,0085	0,0043
Gangopadhyay-Cheung (GI(99; 0, 03)) (GC2)	0,0143	0,0089	0,0058	0,0039

A Tabela 6.14 e a Figura 6.15 confirmam a hipótese de ineficiência do estimador de Chiu (1991) para estimar densidades bi-modais, devido à forma de especificação de Λ.

A Tabela 6.14 também mostra que, apesar do método de Chiu (1991) apresentar resultados ruins, o estimador de Brewer (2000) apresentou melhora quando utilizado conjuntamente com este. Entretanto, esta melhora é insuficiente para que o método se torne equivalente aos demais.

No caso de utilização conjunta com Sheather e Jones (1991), a abordagem não apresentou melhora em nenhum caso quando comparado ao método puro de Sheather e Jones (1991), porém apresentou melhores resultados que o estimador de Chiu (1991) e o estimador Brewer (2000) combinado com Chiu (1991).

Para amostras menores que 200, a abordagem de Gangopadhyay e Cheung (2002) com *priori* GI$(99; 0, 03)$ apresentou os melhores resultados, em termos de erro. Quando o método foi utilizado com uma distribuição *a priori* GI$(99; 0, 12)$, o estimador apresentou resultados inferiores para amostras de tamanho 30 e 50.

De acordo com os resultados apresentados, verificou-se a fragilidade (instabilidade) do método de Chiu (1991) para estimação de funções densidade que apresentam bimodalidade. O estimador Brewer (2000) apresentou uma melhora significativa quando utilizado com Chiu (1991). Entretanto, a melhora não foi suficiente para que o método conseguisse um desempenho equivalente aos métodos de Sheather e Jones (1991) e Gangopadhyay e Cheung (2002).

Vale salientar que o estimador de Sheather e Jones (1991) apresentou resultados acima dos esperados, próximos à metodologia de janela variável para os diferentes tamanhos amostrais.

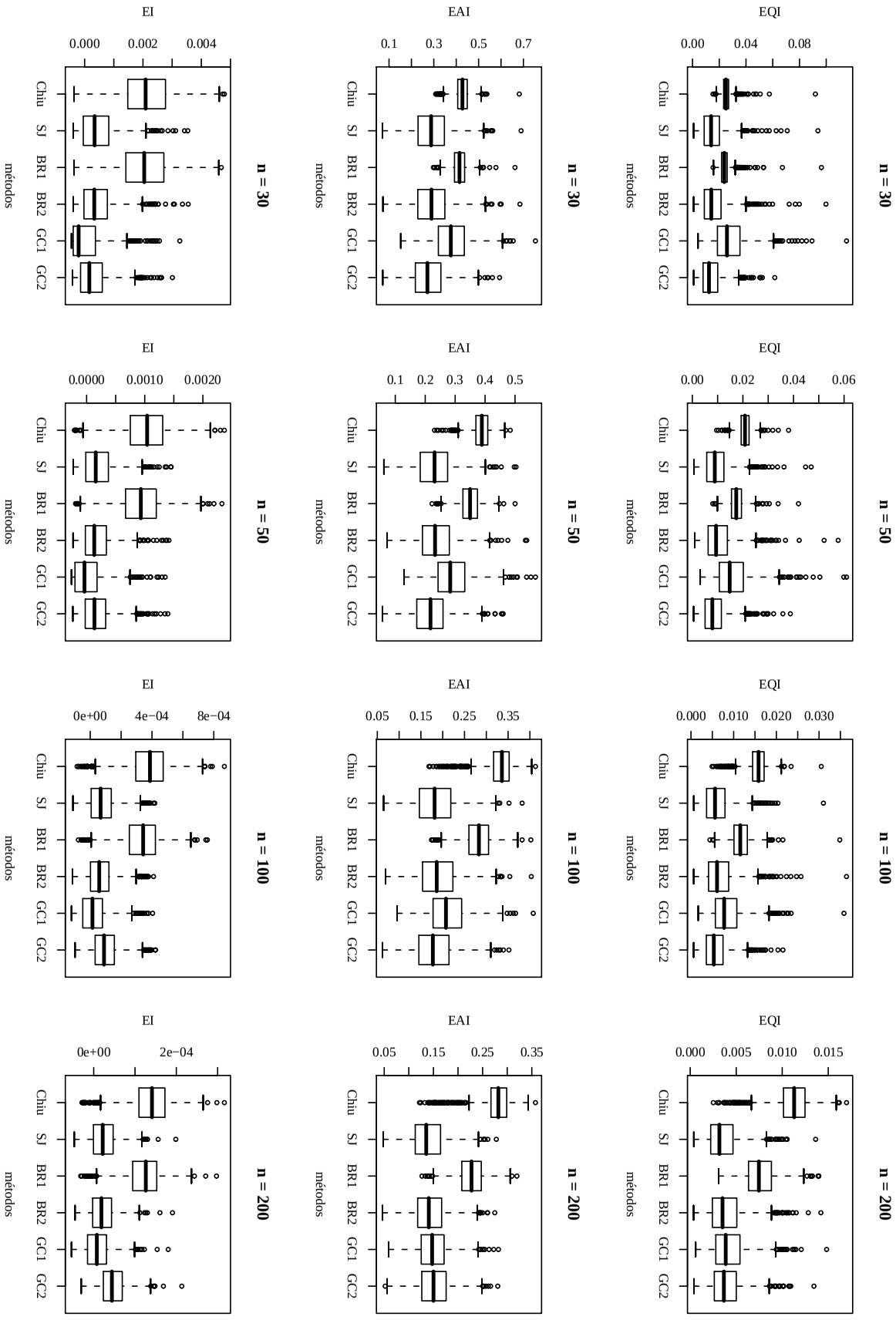

Figura 6.15: Boxplot do Erros Integrados ao estimar a Mistura(3)

6.8 $X \sim \frac{1}{2}\text{Normal}(2,1) + \frac{1}{2}\text{Normal}(8,1)$

Assim como no modelo anterior, neste caso espera-se que os métodos de janela variável apresentem melhor desempenho, devido à bi-modalidade da função densidade. Os resultados relativos à metodologia *plug-in* são apresentados a seguir.

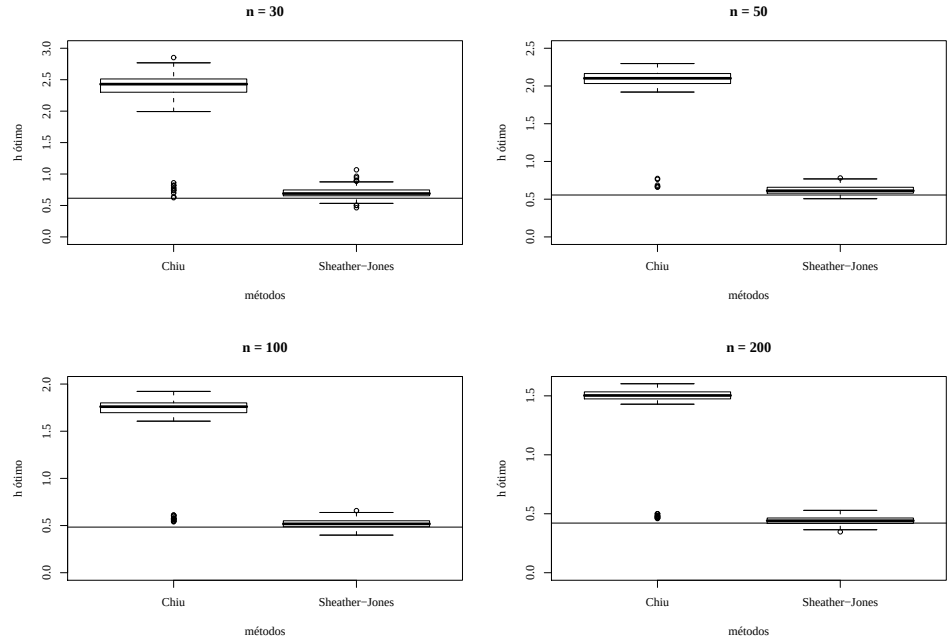

Figura 6.16: Boxplot das estimativas da janela ótima para a Mistura(4). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

Tabela 6.15: Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(4)

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	0,6152	Chiu	2,2707	0,5297	1,6555
		SJ	0,7050	0,1010	0,0898
50	0,5554	Chiu	2,0234	0,3452	1,4680
		SJ	0,6212	0,0580	0,0658
100	0,4835	Chiu	1,6018	0,4214	1,1183
		SJ	0,5200	0,0484	0,0365
200	0,4209	Chiu	1,3688	0,3626	0,9479
		SJ	0,4417	0,0341	0,0208

A Figura 6.16 e a Tabela 6.15 mostram que, assim como no modelo anterior, neste caso,

o método de Chiu (1991) tende a superestimar, excessivamente, a janela ótima teórica. Além disso, o estimador também apresentou alta variabilidade. Como era esperado, o método ficou comprometido pela forma de determinação do limite de integração Λ.

Novamente, o estimador Sheather e Jones (1991) apresentou desempenho superior ao estimador Chiu (1991).

Uma comparação global entre as metodologias implementadas pode ser feita através da Tabela 6.16, que apresenta o EQMI, e da Figura 6.17, que apresenta os gráficos *boxplot* dos erros integrados.

Tabela 6.16: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0480	0,0422	0,0326	0,0243
Sheather-Jones	0,0146	0,0095	0,0060	0,0034
Brewer (Chiu) (BR1)	0,0461	0,0365	0,0261	0,0185
Brewer (Sheather-Jones) (BR2)	0,0149	0,0101	0,0065	0,0037
Gangopadhyay-Cheung (GI(99; 0, 12)) (GC1)	0,0281	0,0163	0,0085	0,0043
Gangopadhyay-Cheung (GI(99; 0, 03)) (GC2)	0,0142	0,0089	0,0058	0,0038

A Tabela 6.16 confirma novamente a ineficiência do estimador de Chiu (1991) para densidades bi-modais. O método de Brewer (2000) apresentou uma melhora, quando utilizado conjuntamente com Chiu (1991). No caso de utilização conjunta com Sheather e Jones (1991), a abordagem não apresentou melhora em relação ao método "puro" de Sheather e Jones (1991). Contudo, o método apresentou melhores resultados que o estimador de Chiu (1991) e o estimador Brewer (2000) combinado com Chiu (1991).

A abordagem de Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 03) apresentou resultados equivalentes ao método de Sheather e Jones (1991). No entanto, quando o método foi utilizado com uma distribuição *a priori* GI(99; 0, 12), o estimador apresentou resultados inferiores para amostras de tamanho 30, 50 e 100.

As conclusões expostas são reforçadas através da análise da Figura 6.17.

De acordo com os resultados apresentados, verificou-se a fragilidade (instabilidade) do método de Chiu (1991) para a estimação de funções densidade bimodais. O estimador Brewer (2000) apresentou uma melhora significativa quando utilizada com Chiu (1991), entretanto insuficiente, pois o método não teve um desempenho equivalente aos métodos de Sheather e Jones (1991) e Gangopadhyay e Cheung (2002).

Vale salientar que o estimador de Sheather e Jones (1991) apresentou resultados acima dos esperados, devido à bimodalidade da função densidade. Além disso, o método obteve resultados equivalentes à metodologia de janela variável para os diferentes tamanhos amostrais.

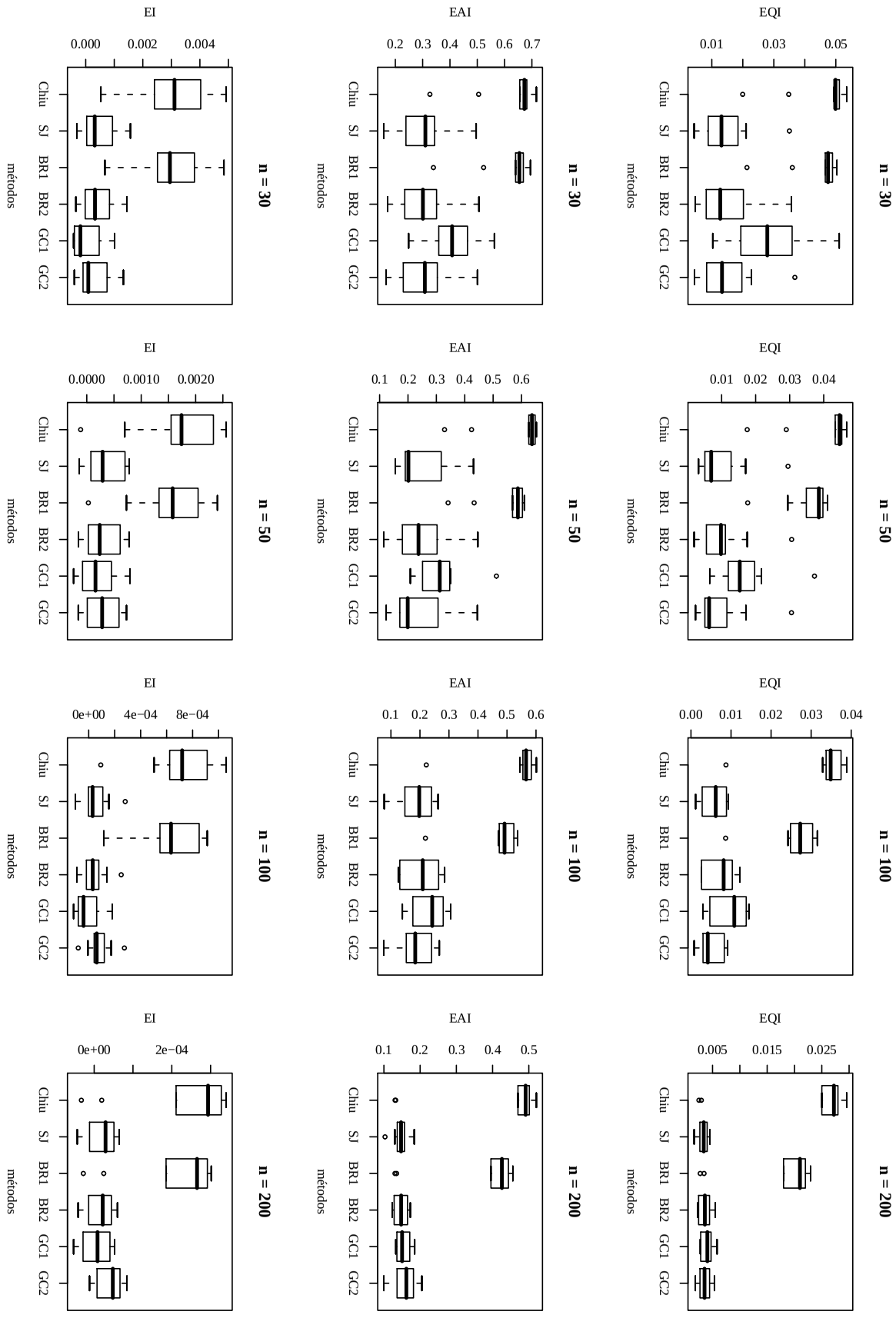

Figura 6.17: Boxplot do Erros Integrados ao estimar a Mistura(4). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

6.9 $X \sim \frac{1}{2}\text{Normal}(2,1) + \frac{1}{2}\text{Normal}(6,3)$

Novamente, espera-se que todos os métodos de janela variável apresentem melhor desempenho, devido à bi-modalidade da função densidade. Os resultados relativos à metodologia *plug-in* são apresentados a seguir.

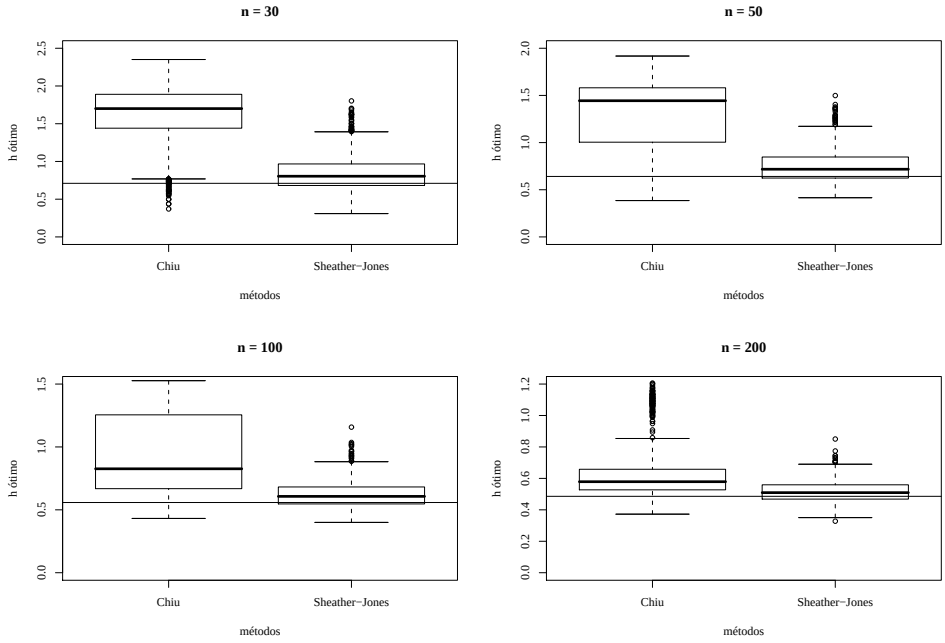

Figura 6.18: Boxplot das estimativas da janela ótima para a Mistura(5). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

Tabela 6.17: Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(5)

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	0,7101	Chiu	1,6030	0,4002	0,8902
		SJ	0,8484	0,2406	0,1383
50	0,6412	Chiu	1,3178	0,3475	0,6766
		SJ	0,7465	0,1727	0,1053
100	0,5582	Chiu	0,9291	0,3007	0,3743
		SJ	0,6215	0,1089	0,0633
200	0,4859	Chiu	0,6342	0,1785	0,1483
		SJ	0,5177	0,0698	0,0318

Como era de se esperar, a Figura 6.18 e a Tabela 6.17 mostram que o método de

Chiu (1991) tende a superestimar, excessivamente, a janela ótima e apresentar alta variabilidade. O estimador Sheather e Jones (1991) apresentou desempenho bem superior ao estimador Chiu (1991).

Uma comparação global entre as metodologias implementadas pode ser feita através da Tabela 6.18, que apresenta o EQMI, e da Figura 6.19, que apresenta os *boxplot* dos erros integrados.

Tabela 6.18: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0143	0,0105	0,0065	0,0033
Sheather-Jones	0,0124	0,0080	0,0050	0,0029
Brewer (Chiu) (BR1)	0,0139	0,0096	0,0058	0,0032
Brewer (Sheather-Jones) (BR2)	0,0131	0,0089	0,0056	0,0032
Gangopadhyay-Cheung (GI(99; 0, 09)) (GC1)	0,0242	0,0140	0,0073	0,0037
Gangopadhyay-Cheung (GI(99; 0, 02)) (GC2)	0,0112	0,0071	0,0046	0,0031

A Tabela 6.18 e a Figura 6.19 mostram que, para amostras de tamanho 100 e 200, o estimador de Chiu (1991) apresentou um bom desempenho, ou seja, não houve perda de informação ao determinar o limite de integração Λ. Todavia, este desempenho foi inferior ao desempenho dos métodos de Sheather e Jones (1991), Brewer (2000) combinado com Sheather e Jones (1991) e Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 02) para amostras de tamanho 30 e 50.

O método de Brewer (2000) apresentou bons resultados com ambos os estimadores *plug-in*. Para amostras de tamanho 200, todas as metodologias apresentaram desempenho semelhante, enquanto que para amostras de tamanho 30 e 50, o método de Chiu (1991) e Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 09) apresentaram resultados inferiores.

De acordo com os resultados apresentados, verificou-se que os diversos estimadores são equivalentes para amostras de tamanho 200, em termos de erro. Além disso, confirmou-se que sob as condições simuladas, as metodologias de janela variável apresentaram melhora

em relação aos métodos de janela fixa para amostras de tamanho 30, 50 e 100.

Vale ressaltar que o estimador de Sheather e Jones (1991) apresentou resultados acima dos esperados, devido à bimodalidade da função densidade.

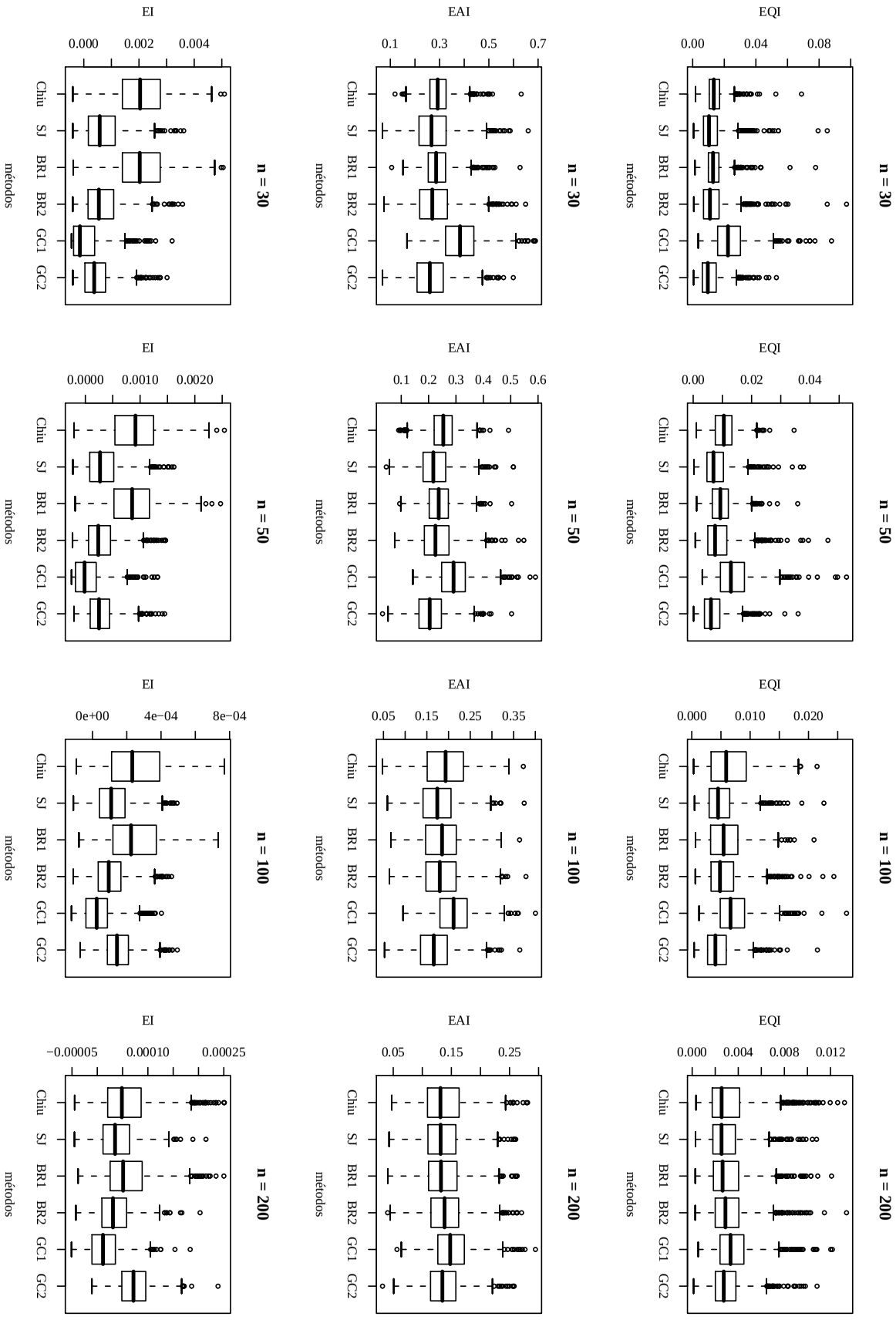

Figura 6.19: Boxplot do Erros Integrados ao estimar a Mistura(5)

6.10 $X \sim \frac{1}{2}\text{Normal}(2,1) + \frac{1}{2}\text{Normal}(8,3)$

Neste caso, esperasse que todos os métodos de janela variável apresentem melhor desempenho, devido à bi-modalidade da função densidade. Os resultados relativos à metodologia *plug-in* são apresentados a seguir.

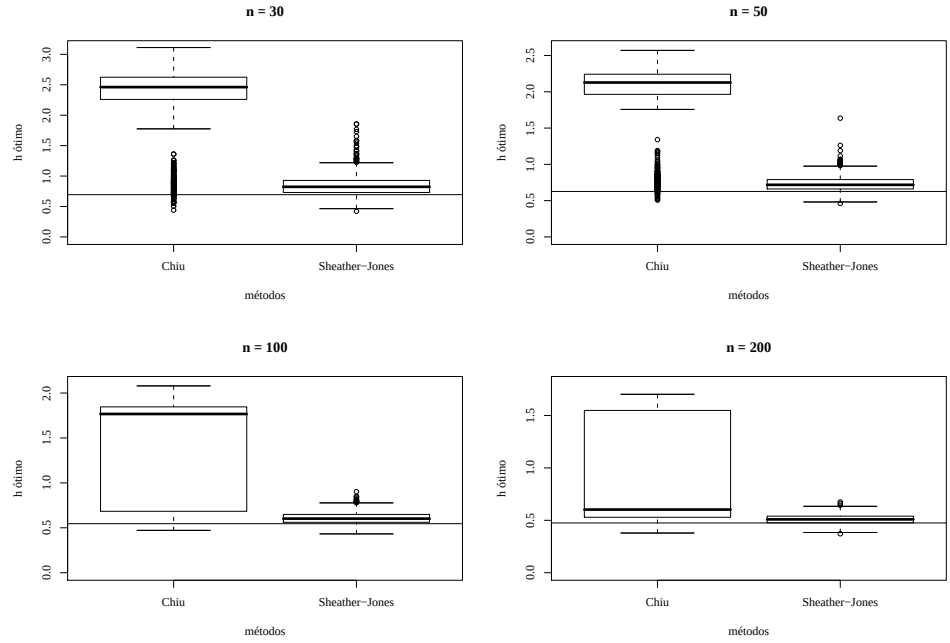

Figura 6.20: Boxplot das estimativas da janela ótima para a Mistura(6). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

Tabela 6.19: Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(6)

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	0,6941	Chiu	2,2663	0,6111	1,5722
		SJ	0,8467	0,1800	0,1526
50	0,6267	Chiu	1,9063	0,5600	1,2796
		SJ	0,7308	0,1103	0,1041
100	0,5456	Chiu	1,4283	0,5622	0,8827
		SJ	0,6070	0,0685	0,0614
200	0,4749	Chiu	1,0081	0,5111	0,5332
		SJ	0,5100	0,0463	0,0351

Como era de se esperar, a Figura 6.20 e a Tabela 6.19 mostram que o método de

Chiu (1991) apresenta resultados altamente insatisfatórios, com tendência a superestimar a janela ótima, além de possuir alta variabilidade. Novamente, o estimador Sheather e Jones (1991) se apresentou superior, em relação a vício e variabilidade, ao estimador Chiu (1991).

O estimador de Chiu (1991) também apresentou um excesso de valores discrepantes para amostras de tamanhos 30 e 50, ocasionado pela forma de determinação do limite de integração Λ.

Uma comparação global entre as metodologias implementadas pode ser feita através da Tabela 6.20, que apresenta o EQMI, e da Figura 6.21, que apresenta os gráficos *boxplot* dos erros integrados.

Tabela 6.20: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0273	0,0223	0,0154	0,0089
Sheather-Jones	0,0130	0,0084	0,0052	0,0031
Brewer (Chiu) (BR1)	0,0263	0,0196	0,0127	0,0073
Brewer (Sheather-Jones) (BR2)	0,0133	0,0089	0,0056	0,0033
Gangopadhyay-Cheung (GI(99; 0, 09)) (GC1)	0,0247	0,0143	0,0074	0,0038
Gangopadhyay-Cheung (GI(99; 0, 02)) (GC2)	0,0121	0,0078	0,0051	0,0036

A Tabela 6.20 confirma novamente a ineficência do estimador de Chiu (1991) para densidades bimodais, enquanto o método de Sheather e Jones (1991) apresentou resultados acima dos esperados para tais densidades.

O método de Brewer (2000) apresentou uma pequena melhora nos resultados, quando utilizado conjuntamente com Chiu (1991), em relação ao estimador de Chiu (1991) puro. Entretanto, esta melhora foi insuficiente quando comparado aos demais estimadores.

Para amostras de tamanho 100 e 200, o método de Sheather e Jones (1991) apresentou resultados equivalentes ao estimador de Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 02). No entanto, quando utilizado com uma distribuição *a priori* GI(99; 0, 19), o estimador apresentou resultados inferiores para amostras de tamanhos 30, 50 e 100.

Todas as conclusões expostas são reforçadas através da análise da Figura 6.21.

De acordo com os resultados apresentados, confirmou-se a fragilidade (instabilidade) do método de Chiu (1991) para funções densidade bimodais. O estimador Brewer (2000) apresentou bons resultados quando se utiliza uma boa estimava como janela piloto. Quando se utiliza uma estimativa ruim da janela piloto, o método tende a melhorar os resultados em comparação ao método utilizado na estimação da janela piloto.

Novamente, vale salientar que o estimador de Sheather e Jones (1991) apresentou resultados acima dos esperados, devido à diferença de suavização a ser feita na função densidade.

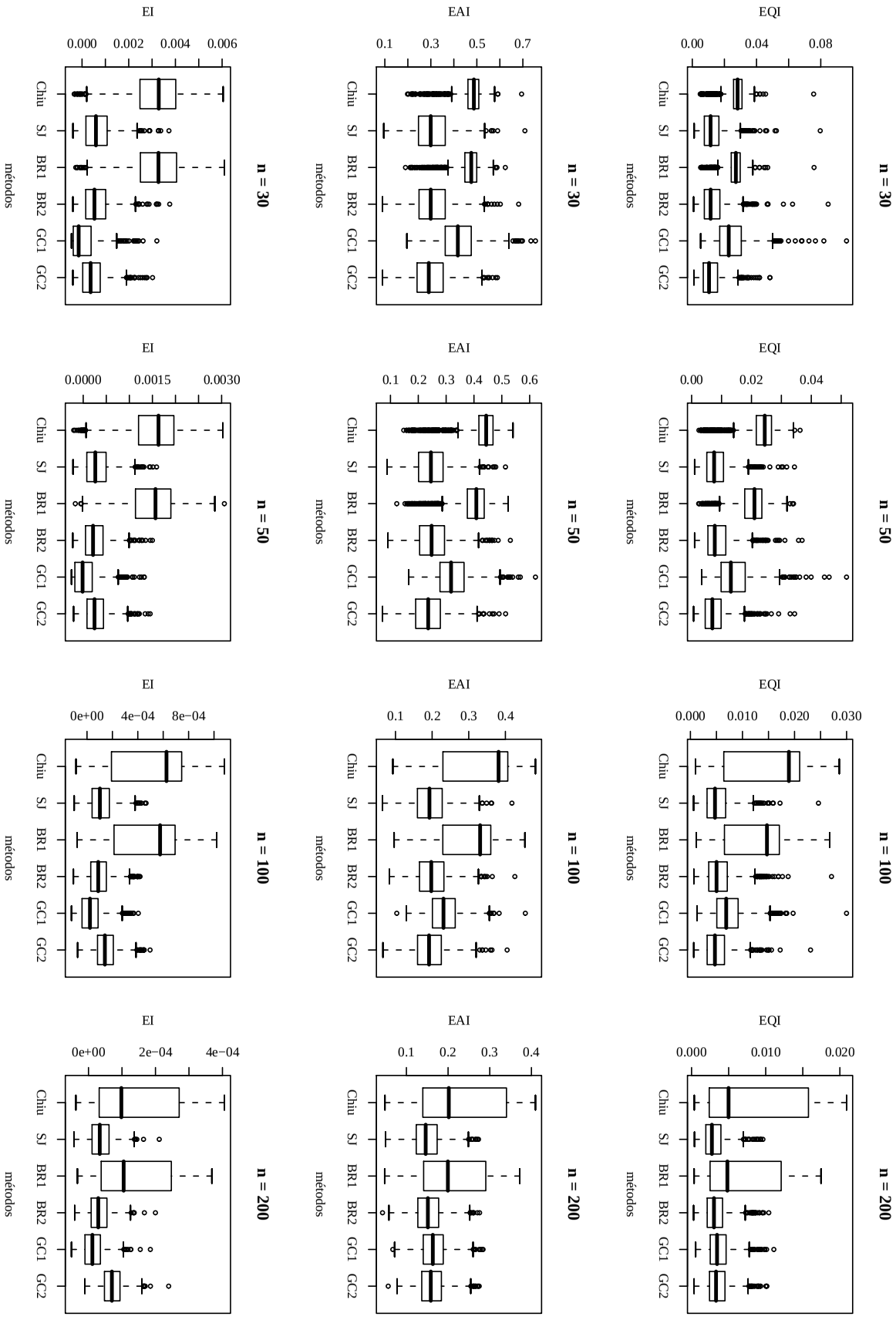

Figura 6.21: Boxplot do Erros Integrados ao estimar a Mistura(6)

6.11 $X \sim \frac{4}{5}\text{Normal}(0,1) + \frac{1}{5}\text{Normal}\left(2, \frac{1}{25}\right)$

Neste caso, espera-se que todos os métodos de janela variável apresentem melhores resultados, devido à bi-modalidade e ao comportamento de crescimento (decrescimento) das modas da função densidade. Os resultados relativos à metodologia *plug-in* são apresentados a seguir.

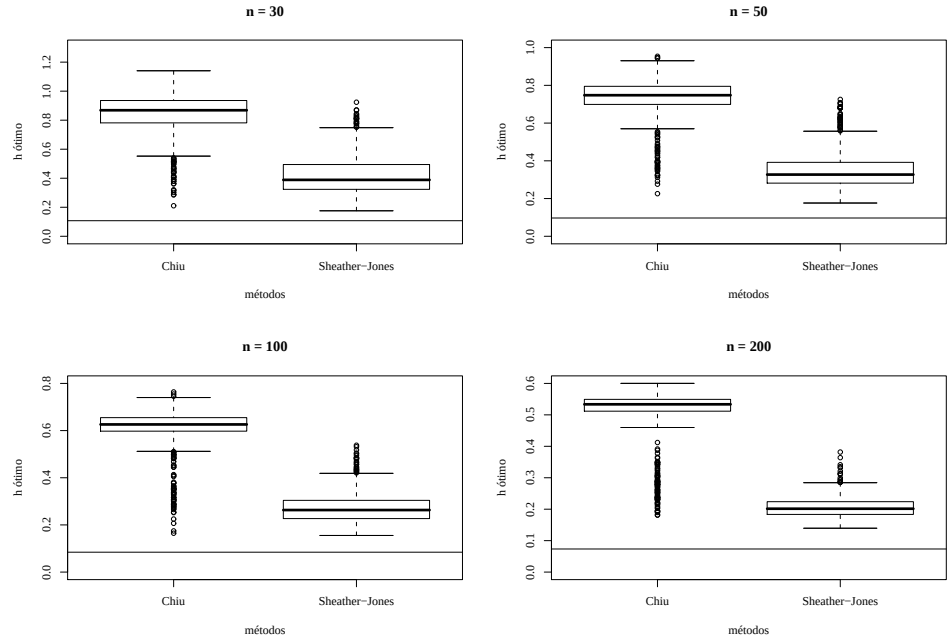

Figura 6.22: Boxplot das estimativas da janela ótima para a Mistura(7). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

A Tabela 6.21 e a Figura 6.22 mostram que, neste caso, ambos estimadores apresentaram tendência a superestimação. No entanto, o método de Chiu (1991) se mostrou mais viciado. Em relação à dispersão, para amostras de tamanhos 30 e 50, ambos os estimadores apresentaram variabilidade parecida, enquanto que para amostras maiores, o método de Chiu (1991) apresentou maior variabilidade. Sendo assim, o método de Sheather e Jones (1991) se mostrou superior, em termos de vício e variablidade, ao método de Chiu (1991).

Uma comparação global entre as metodologias implementadas pode ser feita através da Tabela 6.22, que apresenta o EQMI, e da Figura 6.23, que apresenta os gráficos *boxplot* dos erros integrados.

A Tabela 6.22 e a Figura 6.23 confirmam que o estimador Sheather e Jones (1991) apresenta melhores resultados que o estimador de Chiu (1991) para densidades bimo-

Tabela 6.21: Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(7)

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	0,1073	Chiu	0,8504	0,1321	0,7433
		SJ	0,4189	0,1339	0,3116
50	0,0968	Chiu	0,7368	0,0978	0,6400
		SJ	0,3478	0,0955	0,2571
100	0,0843	Chiu	0,6133	0,0780	0,5290
		SJ	0,2712	0,0600	0,1869
200	0,0734	Chiu	0,5099	0,0808	0,4365
		SJ	0,2056	0,0314	0,1322

Tabela 6.22: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0348	0,0312	0,0272	0,0226
Sheather-Jones	0,0278	0,0204	0,0140	0,0082
Brewer (Chiu) (BR1)	0,0346	0,0291	0,0235	0,0182
Brewer (Sheather-Jones) (BR2)	0,0278	0,0196	0,0126	0,0070
Gangopadhyay-Cheung (GI(99; 0, 32)) (GC1)	0,0301	0,0187	0,0114	0,0070
Gangopadhyay-Cheung (GI(99; 0, 09)) (GC2)	0,0252	0,0195	0,0164	0,0143

dais. Além disso, o método de Brewer (2000) combinado com Sheather e Jones (1991) apresentou melhor desempenho do que o combinado com Chiu (1991).

O estimador de Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 32) apresentou um baixo desempenho para amostras de tamanho 30. Todavia, para os demais tamanhos de amostra, o estimador obteve desempenho equivalente ou superior aos demais métodos. Utilizando uma *priori* GI(99; 0, 09), o estimador teve um bom desempenho para amostras de tamanhos 30 e 50 e um baixo desempenho para os demais tamanhos de amostra.

De acordo com os resultados apresentados, verificou-se a fragilidade (instabilidade) do método de Chiu (1991) para funções densidade bimodais. Neste caso, o estimador Brewer (2000) apresentou um desempenho melhor que os métodos utilizados na estimação da janela piloto. Novamente, o método de Gangopadhyay e Cheung (2002) apresentou um bom desempenho.

Como era esperado neste caso, as metodologias de janela variável apresentaram resultados equivalentes ou superiores aos métodos de janela fixa, em termos de erro. Isto porque a suavização a ser feita tem variações em diferentes partes da função densidade.

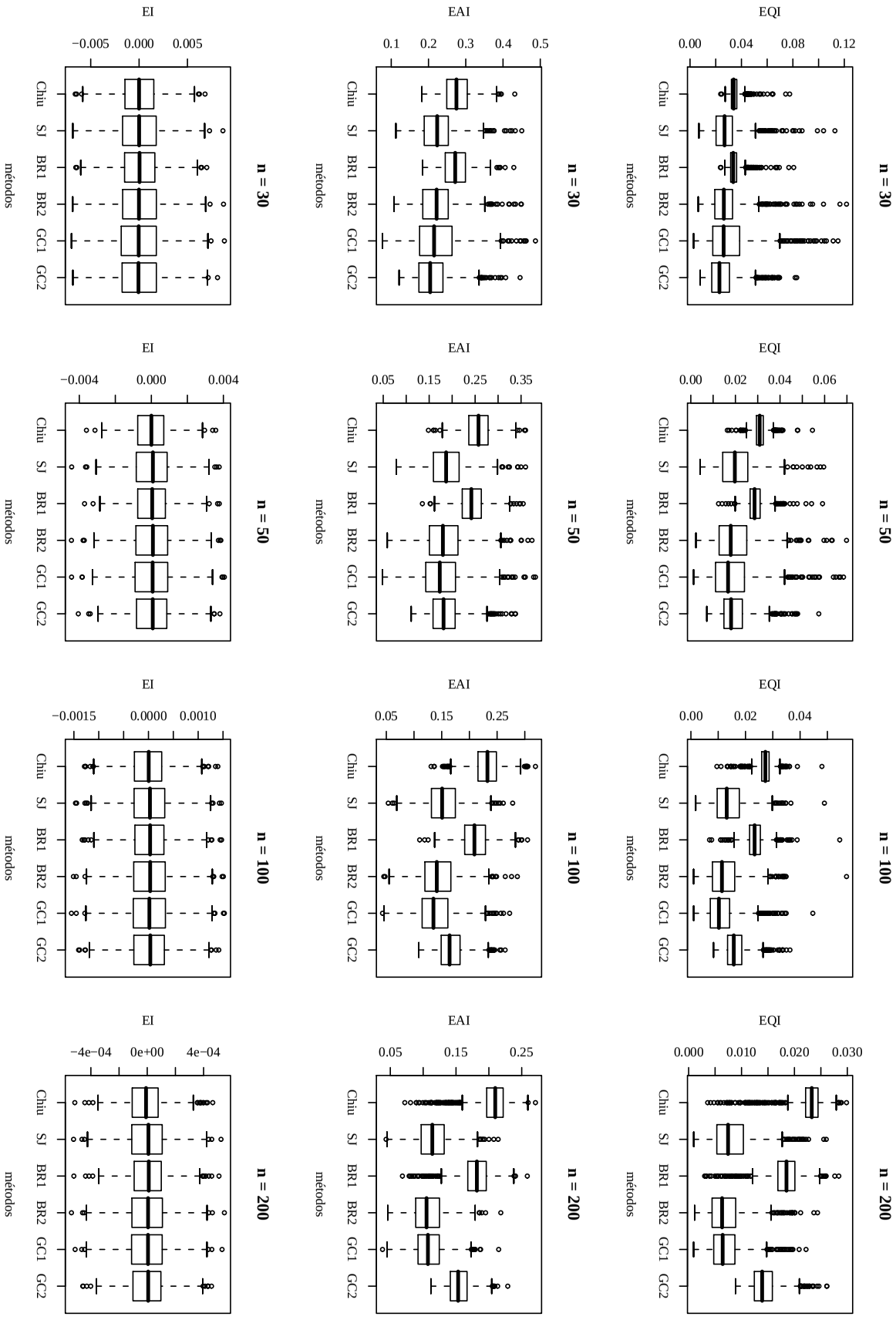

Figura 6.23: Boxplot do Erros Integrados ao estimar a Mistura(7)

6.12 $X \sim \frac{1}{4}\text{Normal}(3,1)+\frac{1}{2}\text{Normal}\left(6,\frac{1}{2}\right)+\frac{1}{4}\text{Normal}(9,1)$

Neste caso, espera-se que os métodos de janela variável apresentem melhor desempenho do que os métodos de janela fixa, devido à multi-modalidade da função densidade. Os resultados relativos à metodologia *plug-in* são apresentados a seguir.

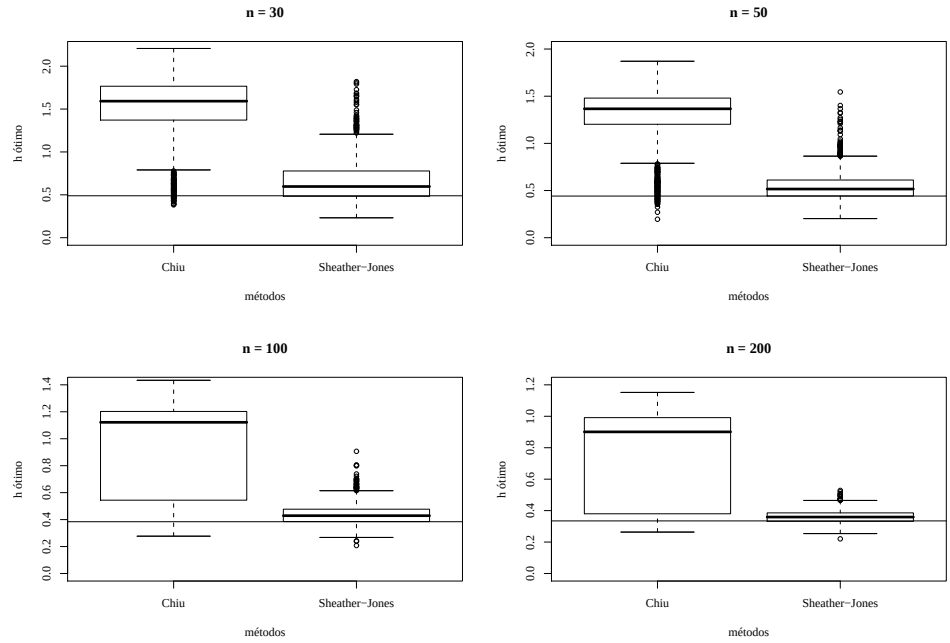

Figura 6.24: Boxplot das estimativas da janela ótima para a Mistura(8). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

Tabela 6.23: Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(8)

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	0,4889	Chiu	1,4987	0,4016	1,0098
		SJ	0,6646	0,2646	0,1757
50	0,4414	Chiu	1,2505	0,3528	0,8091
		SJ	0,5514	0,1644	0,1100
100	0,3843	Chiu	0,9588	0,3321	0,5745
		SJ	0,4401	0,0804	0,0558
200	0,3345	Chiu	0,6960	0,3086	0,3615
		SJ	0,3612	0,0430	0,0267

Como era de se esperar, a Figura 6.24 e a Tabela 6.23 mostram que o método de Chiu (1991) apresenta resultados insatisfatórios, com tendência a superestimação da janela ótima e alta dispersão. Novamente, o estimador de Sheather e Jones (1991) se apresentou superior, em relação a vício e variabilidade, ao estimador Chiu (1991).

Uma comparação global entre os métodos implementados pode ser feita através da Tabela 6.24, que apresenta o EQMI, e da Figura 6.25, que apresenta os gráficos *boxplot* dos erros integrados.

Tabela 6.24: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0265	0,0223	0,0166	0,0106
Sheather-Jones	0,0200	0,0131	0,0080	0,0046
Brewer (Chiu) (BR1)	0,0259	0,0203	0,0140	0,0085
Brewer (Sheather-Jones) (BR2)	0,0206	0,0138	0,0084	0,0049
Gangopadhyay-Cheung (GI(99; 0, 18)) (GC1)	0,0358	0,0209	0,0108	0,0054
Gangopadhyay-Cheung (GI(99; 0, 05)) (GC2)	0,0191	0,0118	0,0073	0,0046

A Tabela 6.24 mostra a superioridade do estimador de Sheather e Jones (1991) em relação ao estimador de Chiu (1991). Confirmando a ineficiência do estimador de Chiu (1991) para densidades multi-modais.

O método de Brewer (2000) apresentou melhora no desempenho, quando utilizado conjuntamente com Chiu (1991), em relação ao estimador de Chiu (1991) puro. Entretanto, esta melhora foi insuficiente quando comparado aos demais estimadores. No caso de utilização conjunta com Sheather e Jones (1991), o método apresentou o mesmo desempenho, em termos de erro, que o método de Sheather e Jones (1991) puro.

A abordagem de Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 05) apresentou resultados equivalentes ou superiores aos demais métodos. No entanto, quando o método foi utilizado com uma distribuição *a priori* GI(99; 0, 18), o estimador apresentou resultados inferiores para amostras de tamanhos 30, 50 e 100. Este comportamento pode ser atribuído à sensibilidade do método em relação à especificação dos hiperparâmetros da distribuição

a priori. Todas as conclusões expostas são reforçadas através da análise da Figura 6.25.

De acordo com os resultados apresentados, verificou-se a fragilidade (instabilidade) do método de Chiu (1991) para funções densidade multi-modais. O estimador Brewer (2000) apresenta bons resultados quando utilizado conjuntamente com Sheather e Jones (1991).

Como era esperado neste caso, as metodologias de janela variável apresentaram resultados equivalentes ou superiores aos métodos de janela fixa, em termos de erro. Isto porque a suavização a ser feita tem variações em diferentes partes da função densidade.

Vale salientar que o estimador de Sheather e Jones (1991) apresentou bons resultados para amostras menores que 200 e resultado equivalente às metodologias de janela variável para amostra de tamanho 200.

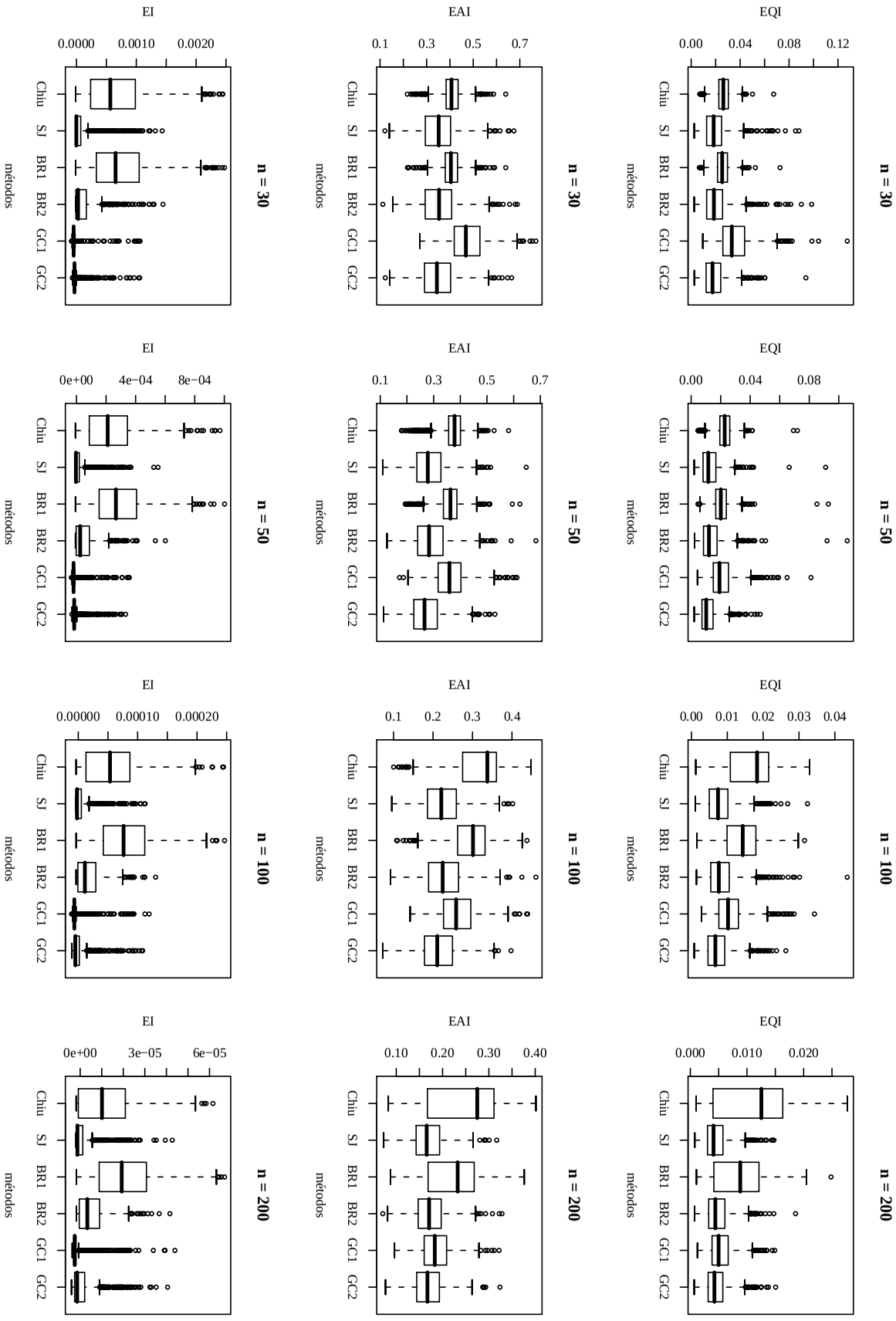

Figura 6.25: Boxplot do Erros Integrados ao estimar a Mistura(8)

6.13 $X \sim \frac{2}{3}\text{Normal}(0,1) + \frac{1}{3}\text{Normal}\left(0, \frac{1}{100}\right)$

Neste caso, espera-se que os métodos de janela variável apresentem melhor desempenho, devido ao crescimento e decrescimento acentuado da função densidade. Os resultados relativos à metodologia *plug-in* são apresentados a seguir.

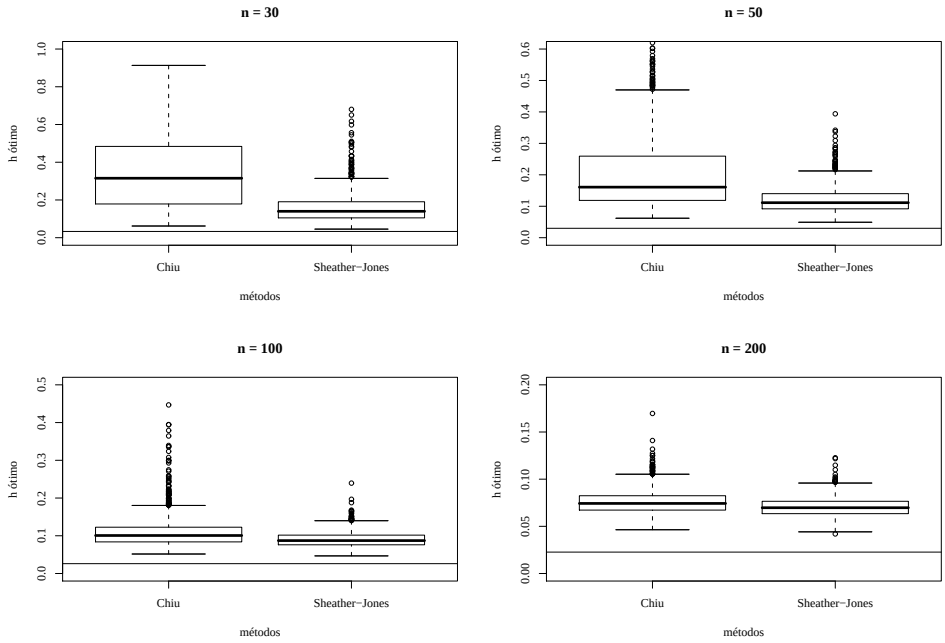

Figura 6.26: Boxplot das estimativas da janela ótima para a Mistura(9). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

Tabela 6.25: Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(9)

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	0,0331	Chiu	0,3390	0,1794	0,3059
		SJ	0,1596	0,0828	0,1265
50	0,0299	Chiu	0,2028	0,1147	0,1729
		SJ	0,1202	0,0424	0,0903
100	0,0260	Chiu	0,1112	0,0448	0,0852
		SJ	0,0904	0,0207	0,0644
200	0,0226	Chiu	0,0758	0,0130	0,0532
		SJ	0,0703	0,0103	0,0477

A Figura 6.26 e a Tabela 6.25 mostram que o método de Chiu (1991) apresentou tendência a superestimação da janela ótima e alta dispersão para amostras de tamanhos 50 e 100. Novamente, o estimador de Sheather e Jones (1991) se mostrou superior, em relação a vício e variabilidade, ao estimador Chiu (1991), sendo essa superioridade mais evidente nos casos de amostra de tamanhos 30 e 50.

Uma comparação global entre as metodologias implementadas pode ser feita através da Tabela 6.26, que apresenta o EQMI, e da Figura 6.27, que apresenta os gráficos *boxplot* dos erros integrados.

Tabela 6.26: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0813	0,0532	0,0275	0,0147
Sheather-Jones	0,0583	0,0387	0,0231	0,0132
Brewer (Chiu) (BR1)	0,0754	0,0447	0,0230	0,0125
Brewer (Sheather-Jones) (BR2)	0,0562	0,0366	0,0223	0,0130
Gangopadhyay-Cheung (GI(99; 3, 7)) (GC1)	0,0810	0,0479	0,0255	0,0131
Gangopadhyay-Cheung (GI(99; 1, 4)) (GC2)	0,0528	0,0335	0,0207	0,0136

A Tabela 6.26 mostra que, com exceção do método de Chiu (1991), para amostras de tamanho 200, as metodologias tiveram um desempenho semelhante, em termos de EQMI. Para amostras de tamanhos 30 e 50, o estimador de Sheather e Jones (1991), o estimador de Brewer (2000) combinado com Sheather e Jones (1991) e o estimador de Gangopadhyay e Cheung (2002) com *priori* GI(99; 1, 4) apresentaram bons resultados, enquanto que o estimador de Chiu (1991) e Gangopadhyay e Cheung (2002) com *priori* GI(99; 3, 7) apresentaram desempenho inferior.

O desempenho da metodologia de Brewer (2000) quando utilizada em conjunto com Chiu (1991) apresentou uma melhora em relação ao estimador de Chiu (1991) puro. Todas as conclusões apresentadas são reforçadas através da análise da Figura 6.27.

De acordo com os resultados apresentados, verificou-se que, sob as condições simuladas, as metodologias de janela variável apresentaram melhora em relação aos métodos de

janela fixa para amostras menores que 200. Para amostras de tamanhos 30 e 50, o estimador de Chiu (1991) demonstrou um desempenho inferior às demais metodologias. Este fato, novamente, pode ser atribuído a maneira de determinação do limite de integração Λ, que proporcionou alta variabilidade do estimador e tendência a superestimação.

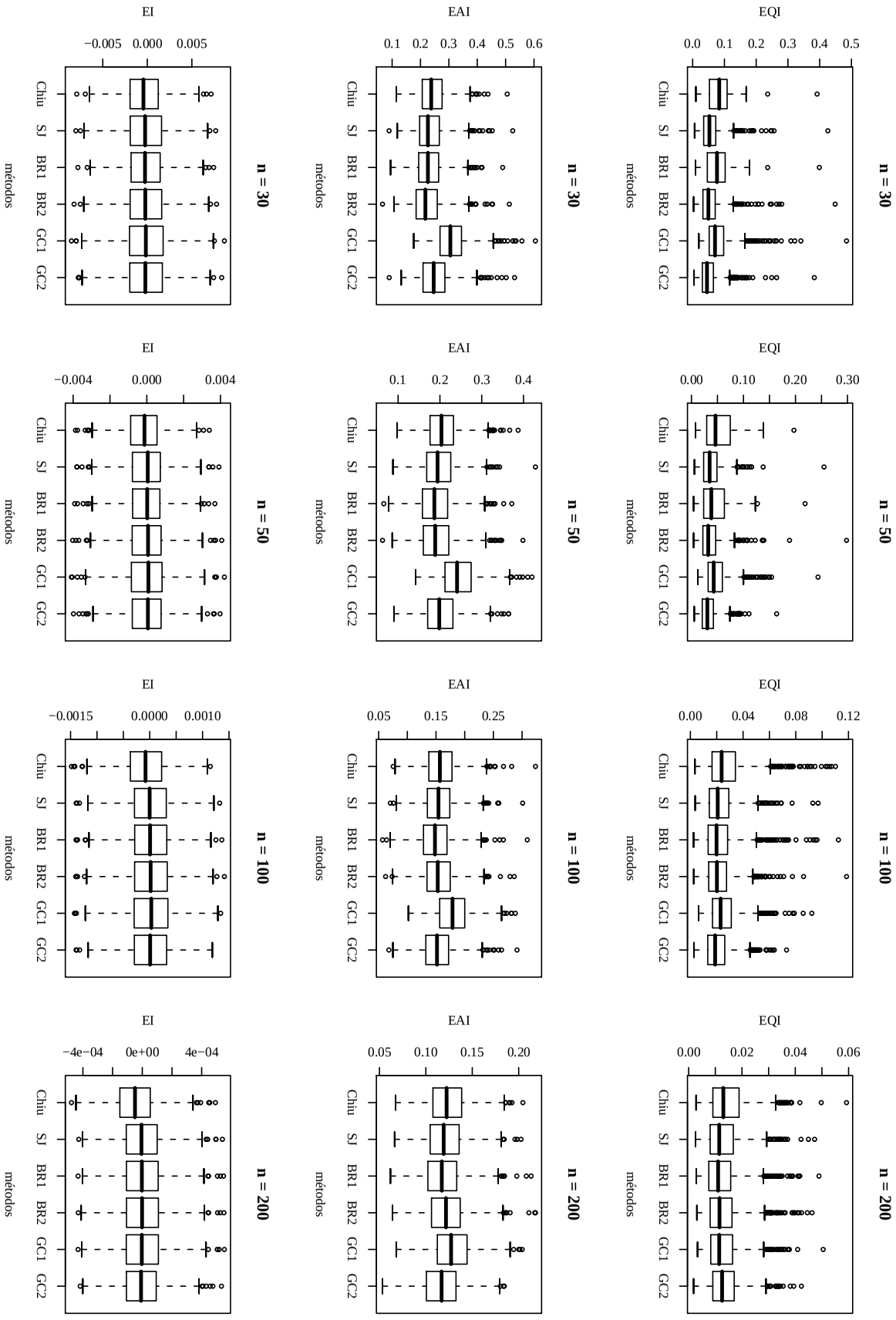

Figura 6.27: Boxplot do Erros Integrados ao estimar a Mistura(9)

6.14 $X \sim \frac{4}{5}\text{Weibull}\left(\frac{1}{100}, 6\right) + \frac{1}{5}\text{Weibull}(1, 6)$

Neste caso, espera-se que os métodos de janela variável apresentem melhor desempenho do que os métodos de janela fixa, devido à multi-modalidade e a assimetria da função densidade. Os resultados relativos à metodologia *plug-in* são apresentados a seguir.

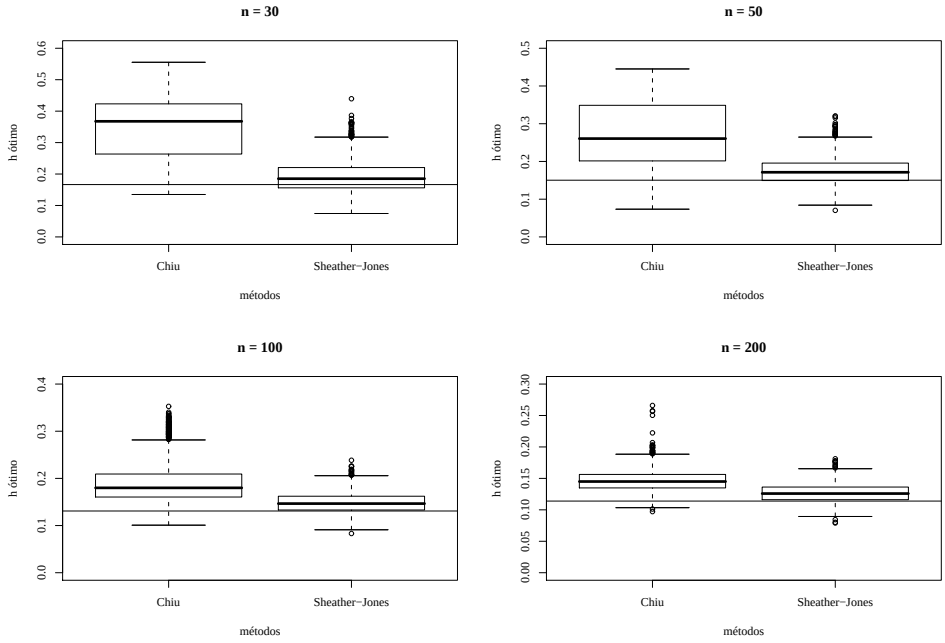

Figura 6.28: Boxplot das estimativas da janela ótima para a Mistura(10). A linha contínua representa a janela ótima teórica que minimiza o EQMIA.

Tabela 6.27: Estatísticas Descritivas para as estimativas da janela ótima para a Mistura(10)

n	h_{opt}	Método	Média ($\hat{h}_{\text{média}}$)	Desvio padrão	$\mid h_{opt} - \hat{h}_{\text{média}} \mid$
30	0,1664	Chiu	0,3471	0,0958	0,1807
		SJ	0,1937	0,0548	0,0273
50	0,1503	Chiu	0,2739	0,0805	0,1236
		SJ	0,1751	0,0382	0,0248
100	0,1308	Chiu	0,1943	0,0484	0,0635
		SJ	0,1478	0,0230	0,0170
200	0,1139	Chiu	0,1471	0,0183	0,0332
		SJ	0,1264	0,0151	0,0125

A Figura 6.28 e a Tabela 6.27 mostram que o método de Chiu (1991) apresentou ten-

dência a superestimação da janela ótima e alta variabilidade para amostras de tamanhos 50 e 100. Novamente, o estimador de Sheather e Jones (1991) se mostrou superior, em relação a vício e variabilidade, ao estimador Chiu (1991), sendo essa superioridade mais evidente nos casos de amostra de tamanhos 30 e 50.

Uma comparação global entre as metodologias implementadas pode ser feita através da Tabela 6.28, que apresenta o EQMI, e da Figura 6.29, que apresenta os gráficos *boxplot* dos erros integrados.

Tabela 6.28: Estimativas do Erro Quadrático Médio Integrado para diferentes tamanhos amostrais (30, 50, 100 e 200).

Método	EQMI			
	30	50	100	200
Chiu	0,0567	0,0532	0,0229	0,0124
Sheather-Jones	0,0526	0,0358	0,0211	0,0123
Brewer (Chiu) (BR1)	0,0547	0,0388	0,0219	0,0122
Brewer (Sheather-Jones) (BR2)	0,0563	0,0404	0,0237	0,0135
Gangopadhyay-Cheung (GI(99; 1, 55)) (GC1)	0,0987	0,0600	0,0305	0,0152
Gangopadhyay-Cheung (GI(99; 0, 39)) (GC2)	0,0471	0,0311	0,0188	0,0123

A Tabela 6.28 mostra que, com exceção do método de Chiu (1991), para amostras de tamanho 200, as metodologias tiveram um desempenho semelhante, em termos de EQMI. Para amostras de tamanhos 30 e 50, o estimador de Sheather e Jones (1991), o estimador de Brewer (2000) combinado com Chiu (1991) e o estimador de Gangopadhyay e Cheung (2002) com *priori* GI(99; 0, 39) apresentaram os melhores, enquanto que o estimador Gangopadhyay e Cheung (2002) com *priori* GI(99; 1, 55) apresentou um desempenho inferior aos demais.

O desempenho da metodologia de Brewer (2000) quando utilizada em conjunto com Chiu (1991) apresentou uma melhora em relação ao estimador de Chiu (1991) puro. Entretanto, quando combinado com Sheather e Jones (1991) o estimador de Brewer (2000) teve um desempenho inferior ao estimador puro para amostras de tamanho 30, 50 e 100. Todas as conclusões apresentadas são reforçadas através da análise da Figura 6.29.

De acordo com os resultados apresentados, verificou-se que, sob as condições simu-

ladas, as metodologias de janela variável apresentaram melhora em relação aos métodos de janela fixa para amostras menores que 200. Para amostras de tamanhos 30 e 50, o estimador de Chiu (1991) demonstrou um desempenho inferior às demais metodologias.

Como era esperado neste caso, as metodologias de janela variável apresentaram resultados equivalentes ou superiores aos métodos de janela fixa, em termos de erro. Isto porque a suavização a ser feita tem variações em diferentes partes da função densidade.

Vale salientar que o estimador de Sheather e Jones (1991) apresentou bons resultados para amostras menores que 200 e resultado equivalente às metodologias de janela variável para amostra de tamanho 200.

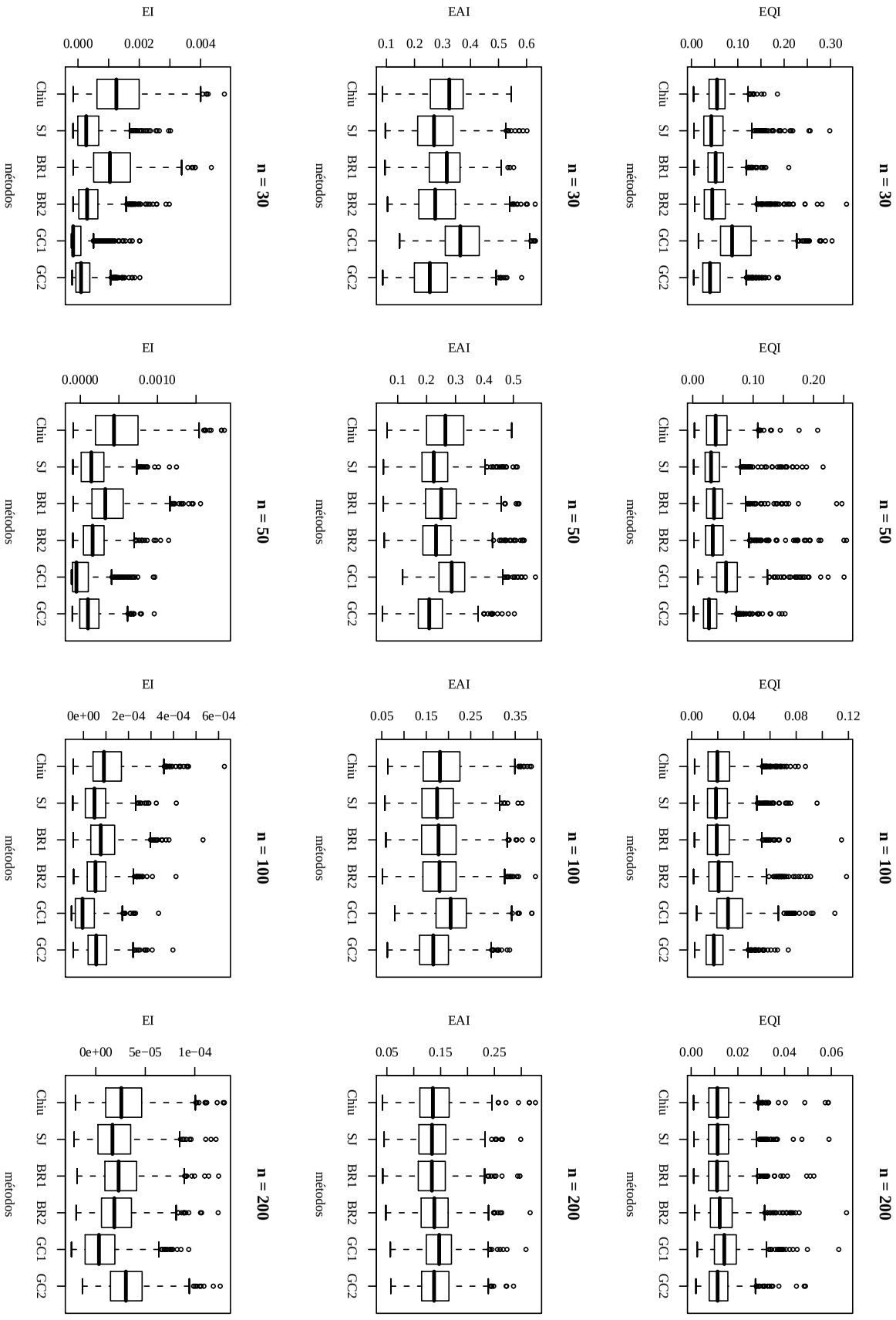

Figura 6.29: Boxplot do Erros Integrados ao estimar a Mistura(10)

6.15 Comentários

O Apêndice A apresenta, para uma amostra gerada aleatoriamente, as estimativas da função densidade para os diferentes modelos simulados e para os diferentes tamanhos de amostra estudados. Estas estimativas têm como objetivo apenas a ilustração das metodologias utilizadas em simulação, não sendo possível obter conclusões sobre o desempenho de cada método, pois trata-se de uma única amostra.

7 Aplicações

A estimação da função densidade de probabilidade pode ser empregada em diversos tipos de problemas, tais como classificação, regressão, confiabilidade, entre outros. No entanto, ela também é comumente utilizada em análises preliminares, com o intuito de explorar características relevantes de uma variável de interesse.

A fim de ilustrar esse tipo de análise exploratória, neste Capítulo serão apresentados exemplos de aplicações reais, com o intento de avaliar o comportamento do método do núcleo, assim como as metodologias de estimação da janela ótima estudadas nesta dissertação.

Como foi explicitado no Capítulo 5, o método de Sain e Scott (1996) também será utilizado nas aplicações, embora não tenha sido utilizado nas simulações devido ao alto tempo computacional que demanda para estimação das janelas ótimas. A implementação computacional deste método foi feita no *software* Matlab 7.0.

O Matlab é uma linguagem de alta performance utilizada para cálculos matemáticos e para a representação gráfica dos resultados. Este software integra computação numérica, visualização e programação em um ambiente de fácil interatividade. Além disso, permite a utilização de *toolboxes* que têm como objetivo disponibilizar soluções para problemas bem conhecidos. O nome do software é uma contração de *MATrix LABoratory*. O Matlab está disponível para Windows, Linux e Mac OSX. Para mais detalhes sobre o Matlab, acesse *http://www.mathworks.com/*.

Devido à necessidade de um método de otimização vetorial na implementação do algoritmo, um grande tempo foi dedicado ao desenvolvimento do programa e à sua otimização. No decorrer da implementação foram utilizados três algoritmos de otimização, dois estocásticos e um determinístico, sendo eles respectivamente:

- Algoritmo Genético (MAN *et al.*, 1996);
- Algoritmo C-PSO (BERGH; ENGELBRECHT, 2004);

- Algoritmo Elipsoidal (BLAND *et al.*, 1981).

Em geral, os resultados obtidos pelos três métodos foram equivalentes. No entanto o tempo computacional dos algoritmos estocásticos foi superior ao algoritmo elipsoidal. Por este motivo, nas aplicações foi utilizado o algoritmo elipsoidal.

A delimitação dos blocos foi feita de tal forma que o mínimo e o máximo da amostra sempre fossem o limite inferior do primeiro bloco e o limite superior do último bloco, respectivamente.

No primeiro exemplo, será analisado o comportamento do tempo (em minutos) entre erupções de um antigo gêiser. A segunda aplicação tem como intenção analisar o comportamento das despesas anuais com alimento de armazéns.

7.1 Aplicação 1

O objetivo desta aplicação é fazer uma análise exploratória do comportamento de tempo entre erupções sucessivas de um gêiser. Para tal avaliação, foram computados os tempos de 229 erupções sucessivas. Este banco de dados pode ser encontrado em Kitchens (2003).

Inicialmente, serão apresentadas as medidas descritivas através da Tabela 7.1 para sumariar o conjunto de dados. Além disso, será apresentado um histograma através da Figura 7.1 para avaliação do comportamento dos dados.

Tabela 7.1: Estatísticas Descritivas para o tempo (em minutos) entre erupções sucessivas.

Estatísticas	
Média	72,3143
Desvio padrão	13,8903
Coeficiente de Assimetria	-0,3375
Coeficiente de Curtose	-1,0217

A Tabela 7.1 mostra que o tempo médio entre erupções é de aproximadamente 72 minutos, com um desvio de 13,89 minutos. Pela Figura 7.1, podemos notar a bi-modalidade dos dados, sendo uma moda próxima de 50 minutos e outra em torno de 80 minutos.

Neste caso, espera-se que as metodologias de janela variável consigam captar melhor os detalhes da distribuição, pois se pressupõe que exista diferença na suavização a ser feita

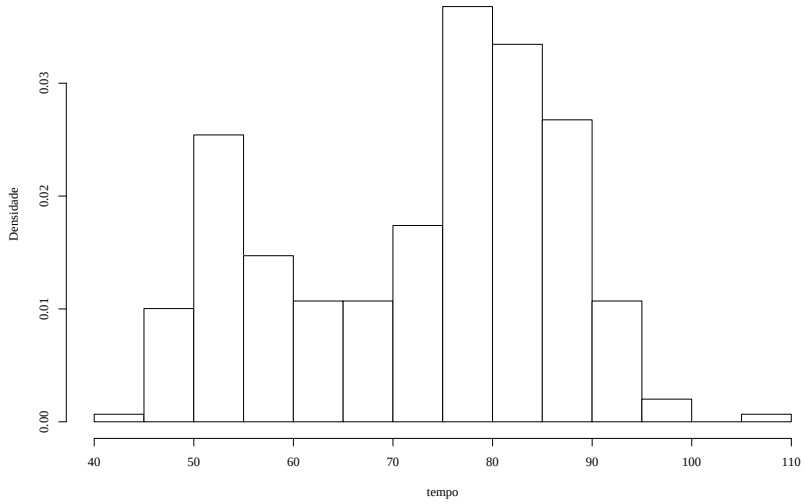

Figura 7.1: Histograma para o tempo (em minutos) sucessivo entre erupções.

perto das modas com a suavização na região de caudas. Dessa forma, aplicando o núcleo-estimador, empregando as diferentes metodologias implementadas nesta dissertação para estimar a densidade do tempo sucessivo entre erupções temos que:

- A estimativa da janela ótima pelo método de Chiu (1991) é $\hat{h}_{opt} = 2,7118$, enquanto que pelo método de Sheather e Jones (1991) é $\hat{h}_{opt} = 2,3893$.
- O método de Brewer (2000) foi utilizado conjuntamente com ambos os estimadores *plug-in*.
- No caso do estimador de Gangopadhyay e Cheung (2002) foram utilizadas duas distribuições *a priori*. Uma distribuição *a priori* não informativa GI$(0,01;100)$ e uma *priori* centrada próxima à janela ótima estimada pelo método de Sheather e Jones (1991) com um coeficiente de variação de aproximadamente 25%, que é dada por uma GI$(26;0,02)$.
- Para o método de Sain e Scott (1996), o número ótimo de blocos foi 12, mais informações sobre o comportamento do estimador são mostradas na Tabela 7.2. Para a otimização vetorial foi utilizado o algoritmo elipsoidal (BLAND *et al.*, 1981).

Tabela 7.2: Informações sobre as estimativas da janela ótima

Bloco	Limites dos blocos	Janela ótima
1	$[43,0000; 48,4167)$	10,4602
2	$[48,4167; 53,8333)$	2,6222
3	$[53,8333; 59,2500)$	4,8765
4	$[59,2500; 64,6667)$	3,7887
5	$[64,6667; 70,0833)$	13,8307
6	$[70,0833; 75,5000)$	2,6808
7	$[75,5000; 80,9167)$	2,2251
8	$[80,9167; 86,3333)$	3,3470
9	$[86,3333; 91,7500)$	3,6320
10	$[91,7500; 97,1667)$	12,7543
11	$[97,1667; 102,5833)$	10,7841
12	$[102,5833; 108,000)$	10,2382

As Figuras 7.3 e 7.2 mostram que todas as metodologias apresentaram estimativas muito próximas. A equivalência dos estimadores pode ser explicada, possivelmente, pelo tamanho da amostra de 229 observações. A utilização do método de Sheather e Jones (1991) como informação *a priori* para o estimador Gangopadhyay e Cheung (2002) apresentou resultados equivalentes aos demais, mostrando que a combinação pode proporcionar bons resultados.

Neste caso era esperado que estimador de Chiu (1991) apresenta-se um resultado insatisfatório, devido à bi-modalidade dos dados. Entretanto, o estimador obteve um resultado acima do esperado, apresentando um comportamento equivalente as demais metodologias. Em relação à abordagem de Brewer (2000), o estimador apresentou o mesmo comportamento dos estimadores *plug-in* utilizados na estimação da janela piloto.

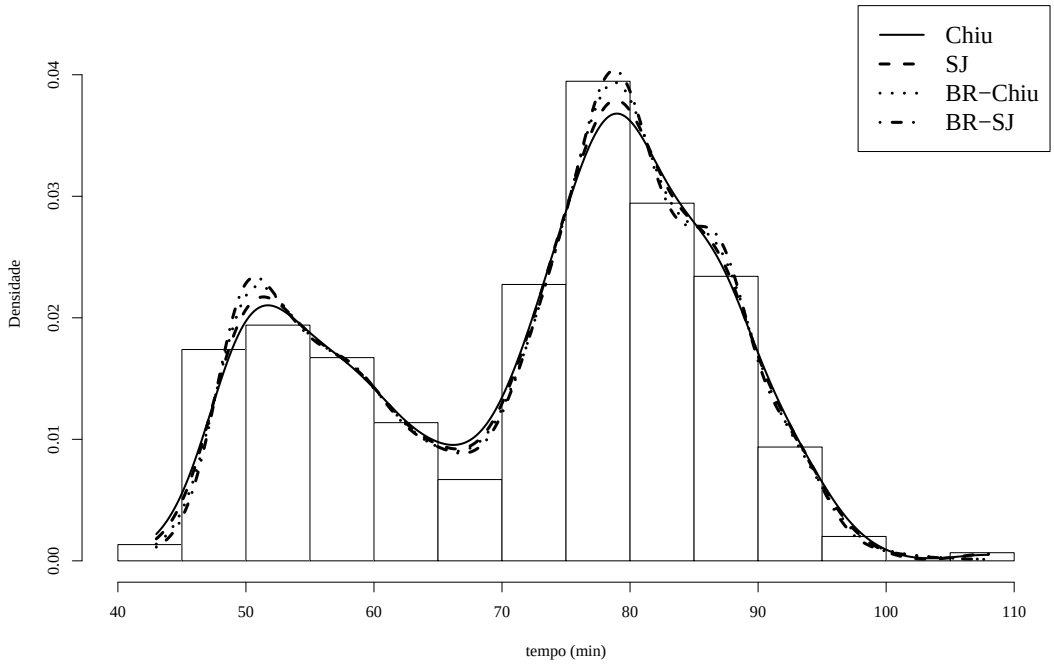

Figura 7.2: Núcleo-estimador para o tempo (em minutos) sucessivo entre erupções.

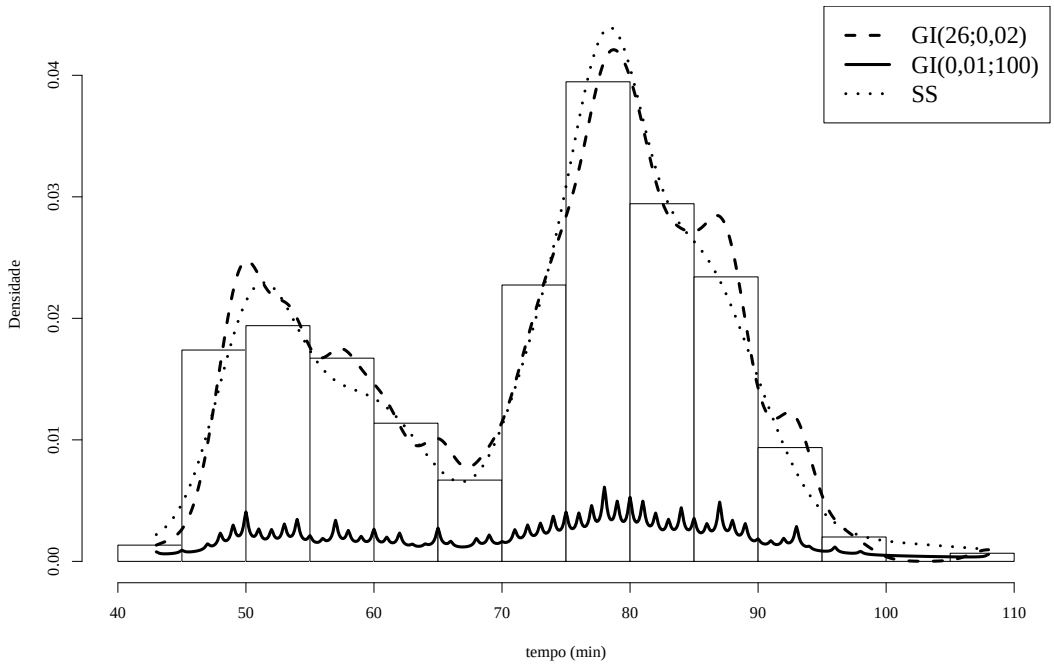

Figura 7.3: Núcleo-estimador para o tempo (em minutos) sucessivo entre erupções.

7.2 Aplicação 2

O objetivo desta aplicação é observar o comportamento das despesas anuais com alimento de 40 armazéns do estado de Ohio, EUA. Este banco de dados pode ser encontrado em Kitchens (2003).

Inicialmente, serão apresentadas as medidas descritivas através da Tabela 7.3 para sumariar o conjunto de dados. Além disso, será apresentado um histograma através da Figura 7.4 para avaliação do comportamento dos dados.

Tabela 7.3: Estatísticas Descritivas para as despesas anuais com alimento ($em1000$).

Estatísticas	
Média	3,6096
Desvio padrão	1,5093
Coeficiente de Assimetria	1,2998
Coeficiente de Curtose	1,3307

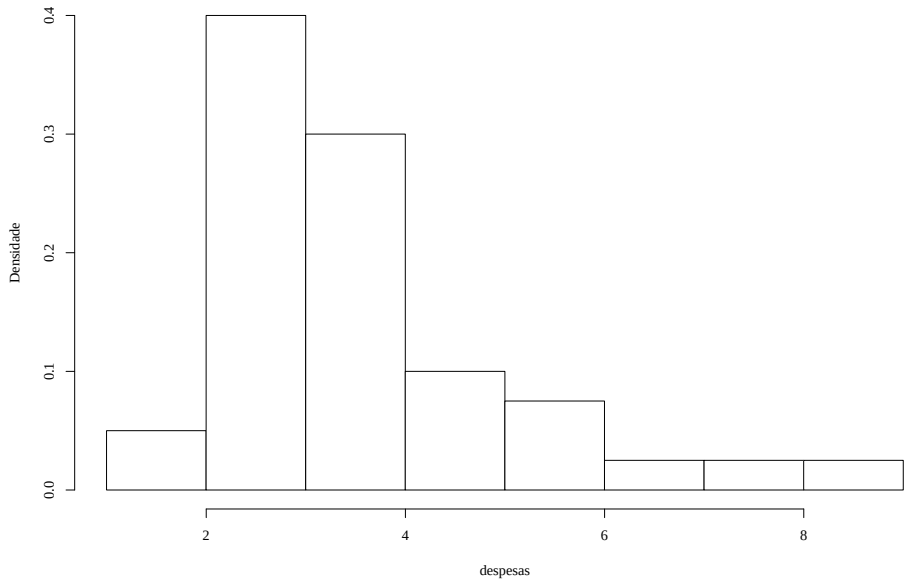

Figura 7.4: Histograma para as despesas.

A Figura 7.4 mostra a acentuada assimetria da distribuição dos dados. Neste caso, esperasse que as metodologias de janela variável consigam acompanhar melhor o crescimento abrupto da distribuição.

Dessa forma, aplicando o núcleo-estimador, empregando as diferentes metodologias

implementadas nesta dissertação, para estimar a densidade das despesas anuais de 40 armazéns temos que:

- A estimativa da janela ótima pelo método de Chiu (1991) é $\hat{h}_{opt} = 0,3556$, enquanto que pelo método de Sheather e Jones (1991) é $\hat{h}_{opt} = 0,3791$.
- O método de Brewer (2000) foi utilizado conjuntamente com ambos os estimadores *plug-in*.
- Assim como na aplicação anterior, para o estimador de Gangopadhyay e Cheung (2002) foram utilizadas duas distribuições *a priori*. Uma distribuição *a priori* não informativa GI$(0,01;100)$ e uma *priori* centrada próxima a janela ótima estimada pelo método de Sheather e Jones (1991) com um coeficiente de variação de aproximadamente 25%, ou seja, uma *priori* GI$(26;0,13)$.
- Para o método de Sain e Scott (1996), o número ótimo de blocos foi 7, mais informações sobre o comportamento do estimador são mostradas na Tabela 7.4. Para a otimização vetorial foi utilizado o algoritmo elipsoidal (BLAND *et al.*, 1981).

Tabela 7.4: Informações sobre as estimativas da janela ótima

Bloco	Limites dos blocos	Janela ótima
1	$[1,1800;2,1753)$	3,4544
2	$[2,1753;3,1706)$	0,3018
3	$[3,1706;4,1659)$	0,3812
4	$[4,1659;5,1611)$	1,7020
5	$[5,1611;6.1564)$	2,0426
6	$[6,1564;7,1517)$	2,2361
7	$[7,1517;8,1470)$	4,3381

A Figura 7.5 mostra que os estimadores de Chiu (1991) e Sheather e Jones (1991) apresentaram comportamento similar, com uma pequena diferença na região da moda. Novamente, o método de Brewer (2000) acompanhou o desempenho dos estimadores *plug-in* utilizados na estimação da janela piloto.

A Figura 7.6 mostra, também, que o estimador Gangopadhyay e Cheung (2002) apresentou um crescimento mais acentuado na região da moda, quando comparado com as demais metodologias, enquanto, que o estimador Sain e Scott (1996) apresentou um crescimento menos acentuado nesta região. Este fato pode ter sido ocasionado pelo tamanho

da amostra, já que Sain e Scott (1996) argumentam que o método é instável para pequenas amostras.

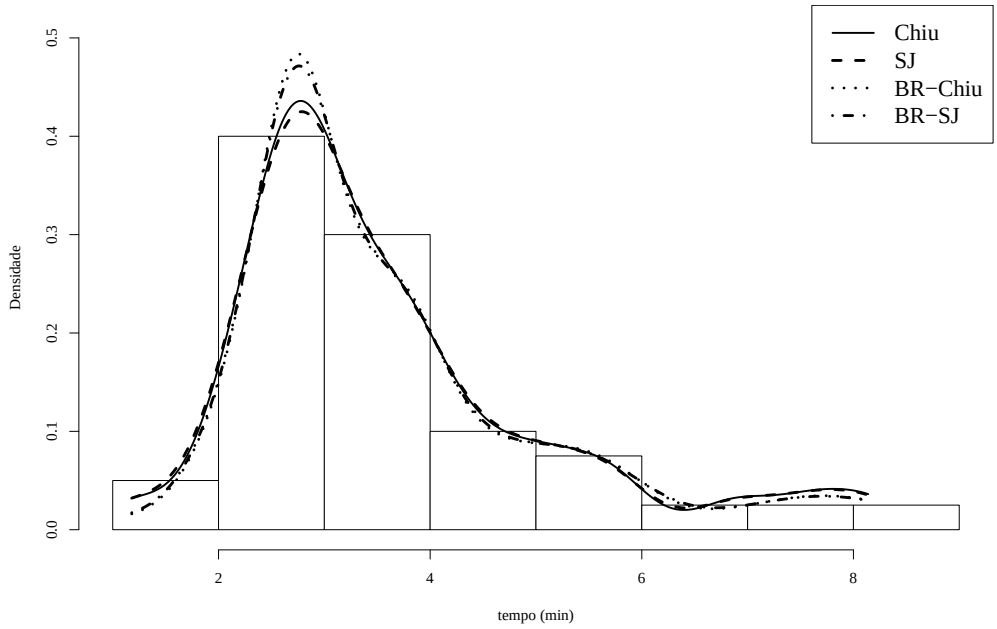

Figura 7.5: Núcleo-estimador para as despesas anuais de armazéns (em 1000 doláres).

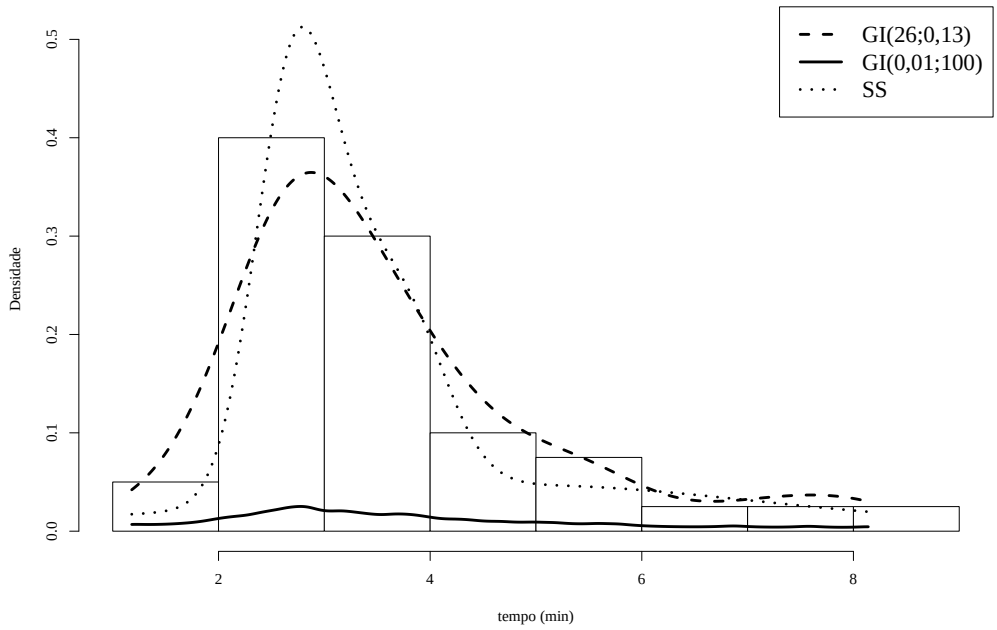

Figura 7.6: Núcleo-estimador para as despesas anuais de armazéns (em 1000 doláres).

8 Conclusões

O estudo apresentado nesta dissertação demonstra que não há uma metodologia proeminente na estimação de densidades, devido às particularidades de cada método. Em alguns casos simulados, houve equivalência de desempenho, em termos de erros, dos diferentes estimadores. Enquanto que em outros casos, algum estimador apresentou desempenho superior.

Com base nas simulações apresentadas, temos as seguintes conclusões:

O estimador de Chiu (1991), como era de se esperar, se mostrou deficitário para algumas densidades multi-modais, devido à forma de determinação do limite de integração Λ da expressão (4.4). Nos demais casos simulados, o método mostrou um bom desempenho.

O método de Sheather e Jones (1991) apresentou ineficiência no caso de densidades assimétricas que foram simuladas com tamanho de amostras 30 e 50. Nos demais casos simulados, o método demonstrou um bom desempenho, obtendo resultados acima do esperado em densidades multi-modais.

A metodologia de Sain e Scott (1996) não foi incluída nas simulações devido ao alto custo computacional, ou seja, a demanda de tempo para o seu processamento é muito alta. Este fato está associado à otimização vetorial necessária na estimação das janelas ótimas. Logo, poucas conclusões podem ser feitas sobre o método.

O estimador de Brewer (2000), em geral, teve desempenho similar aos estimadores *plug-in* utilizados na estimação da janela piloto. No caso de distribuições multi-modais, o método tende a apresentar resultados equivalentes ou superiores ao métodos de janela fixa. Assim como a metodologia de Sain e Scott (1996), a abordagem de Brewer (2000) com dependência entre vizinhos não foi incluída nas simulações devido ao gasto computacional e à necessidade de análises pontuais para cada amostra.

A metodologia de Gangopadhyay e Cheung (2002) demonstrou, em geral, um bom desempenho para *prioris* informativas. Contudo, se mostrou deficitária quando a dis-

tribuição *a priori* é não informativa. Esse problema pode ser contornado utilizando os estimadores *plug-in* para especificação da distribuição *a priori*, como mostrado nos exemplos de aplicação.

Tendo em vista as conclusões anteriores, um possível critério para escolha da metodologia de núcleo-estimador a ser utilizada é o histograma, ou seja, o usuário poderia obter informações prévias à respeito do comportamento da função densidade de probabilidade e assim optar pela escolha do método mais apropriada.

8.1 Propostas futuras

- Corrigir a forma de determinação do limite de integração Λ do estimador de Chiu (1991);
- Estudo sobre o número apropriado de estágios do estimador de Sheather e Jones (1991), aprofundando o estudo feito em Glória (2006);
- Estudo sobre sensibilidade do método de Brewer (2000) em relação à escolha dos hiperparâmetros da distribuição *a priori*;
- Estudo de simulação para o método de Brewer (2000) com dependência entre vizinhos, além de um estudo sobre o número apropriado de vizinhos na estimação das janelas ótimas;
- Estudo de simulação para a metodologia de Sain e Scott (1996);
- Adaptação da metodologia Sain e Scott (1996) para blocos de tamanhos diferentes;

APÊNDICE A – Resultados da Simulação

A.1 $X \sim \text{Normal}(0, 1)$

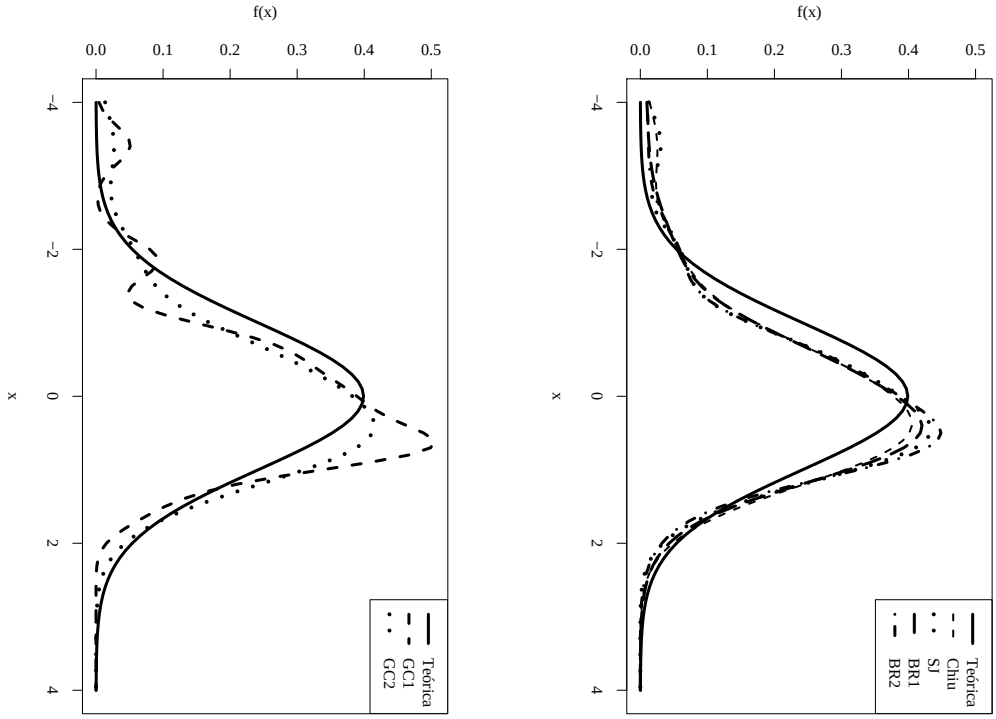

Figura A.1: Estimativa da distribuição Normal(0,1) com $n = 30$.

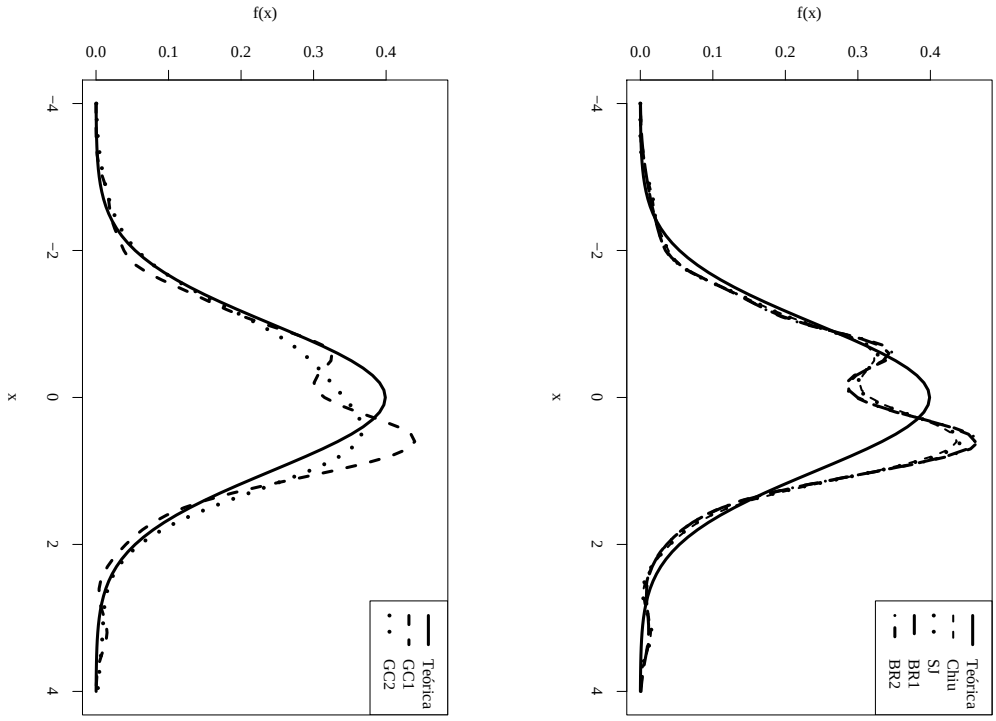

Figura A.2: Estimativa da distribuição Normal(0,1) com $n = 50$.

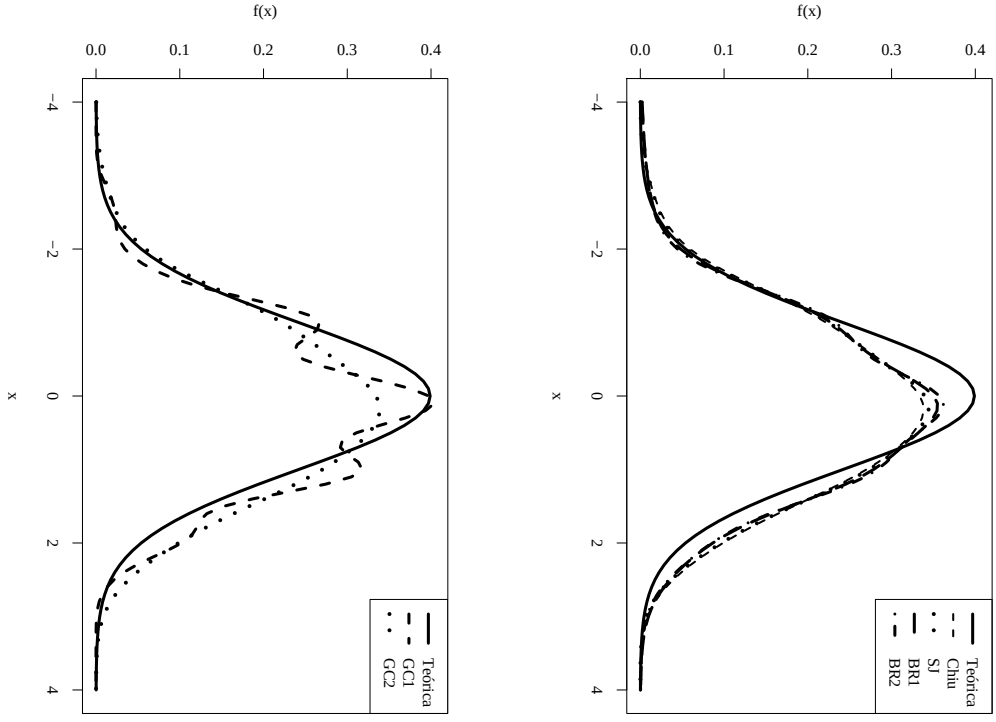

Figura A.3: Estimativa da distribuição Normal(0,1) com $n = 100$.

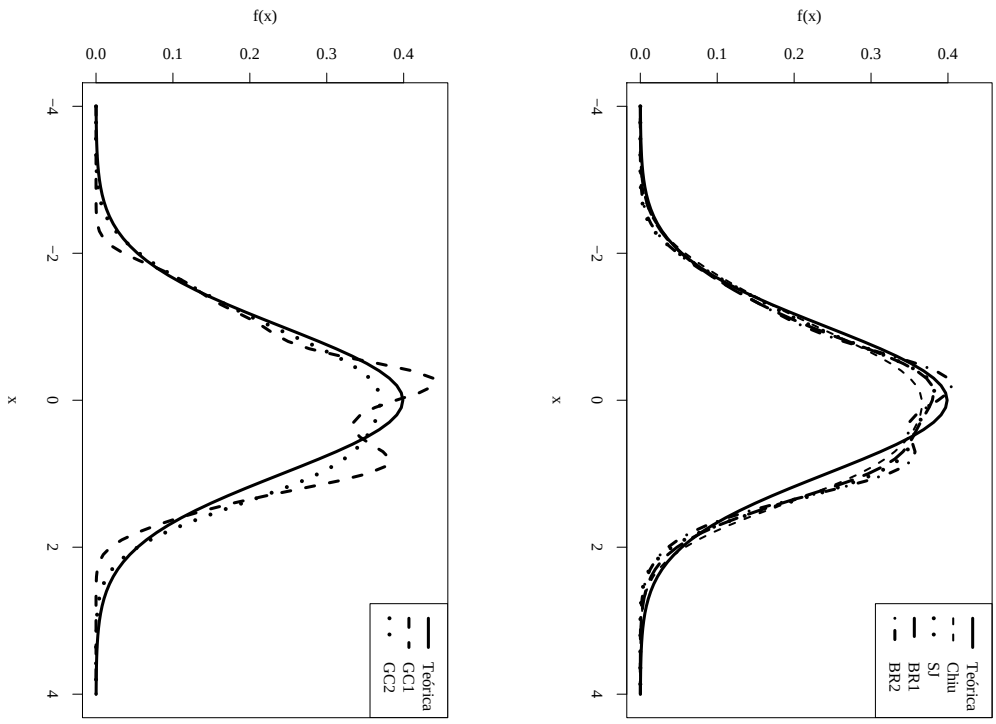

Figura A.4: Estimativa da distribuição Normal(0,1) com $n = 200$.

A.2 $X \sim$ Gama$(4, 2)$

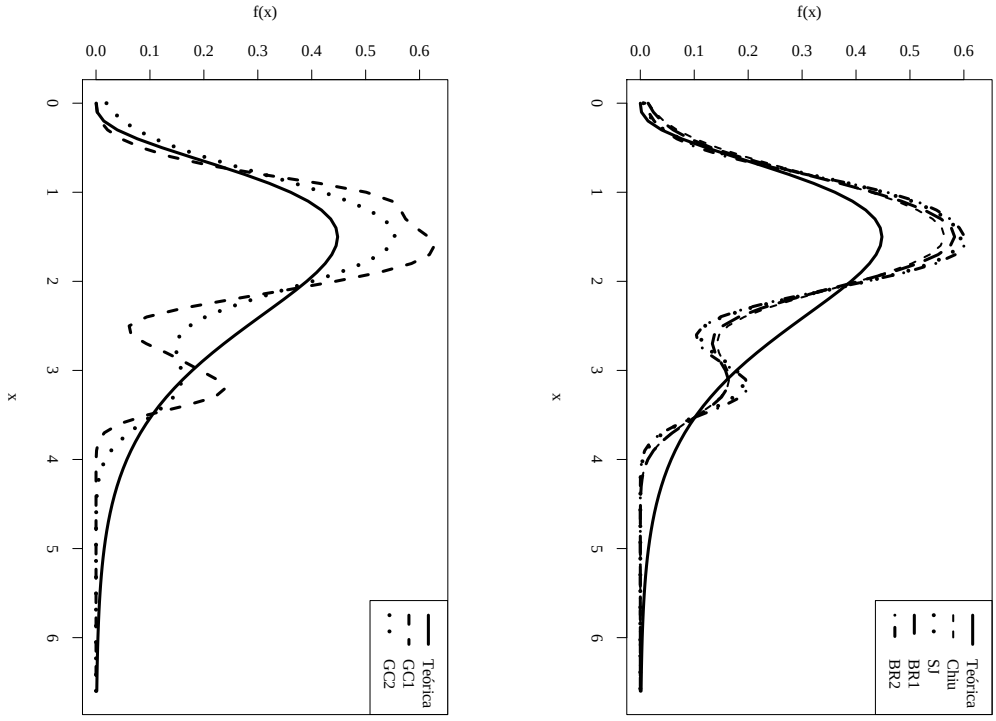

Figura A.5: Estimativa da distribuição Gama(4,2) com $n = 30$.

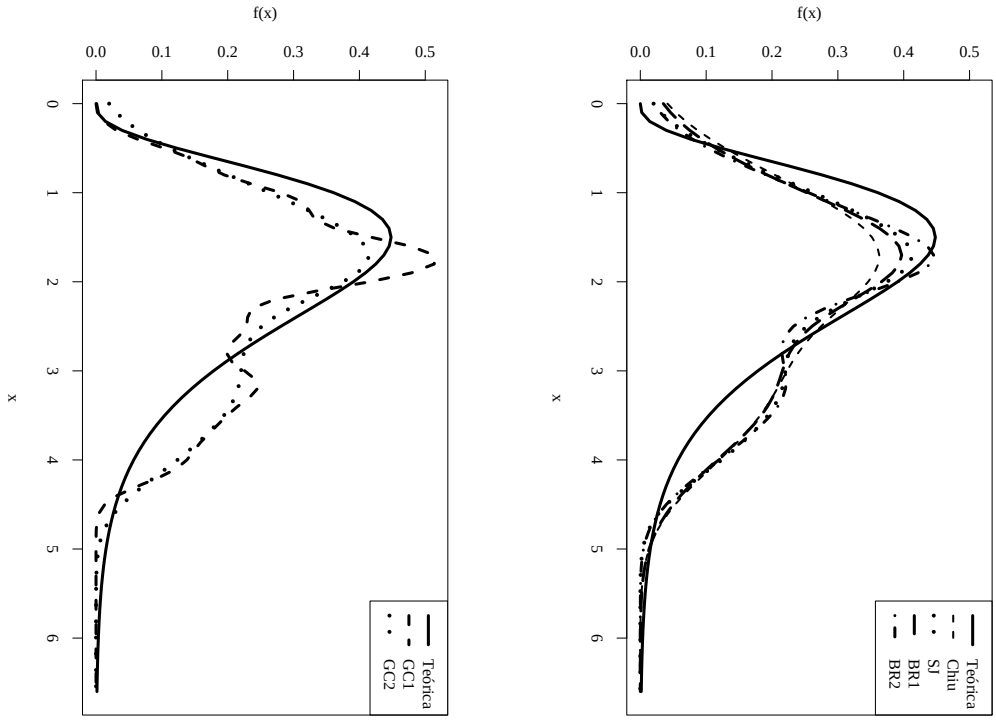

Figura A.6: Estimativa da distribuição Gama(4,2) com $n = 50$.

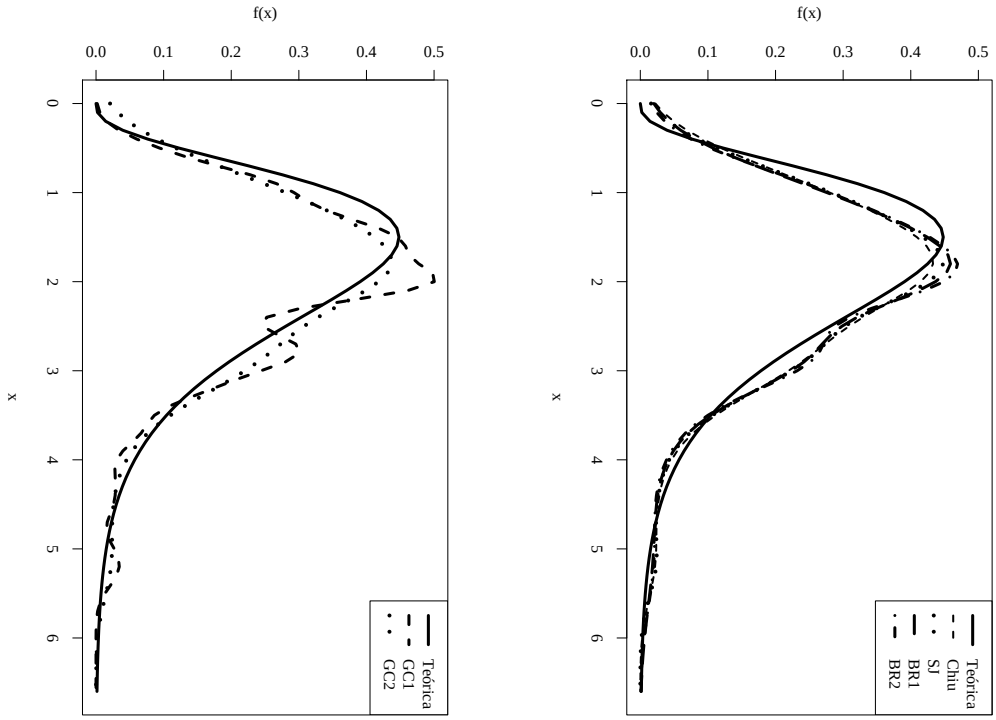

Figura A.7: Estimativa da distribuição Gama(4,2) com $n = 100$.

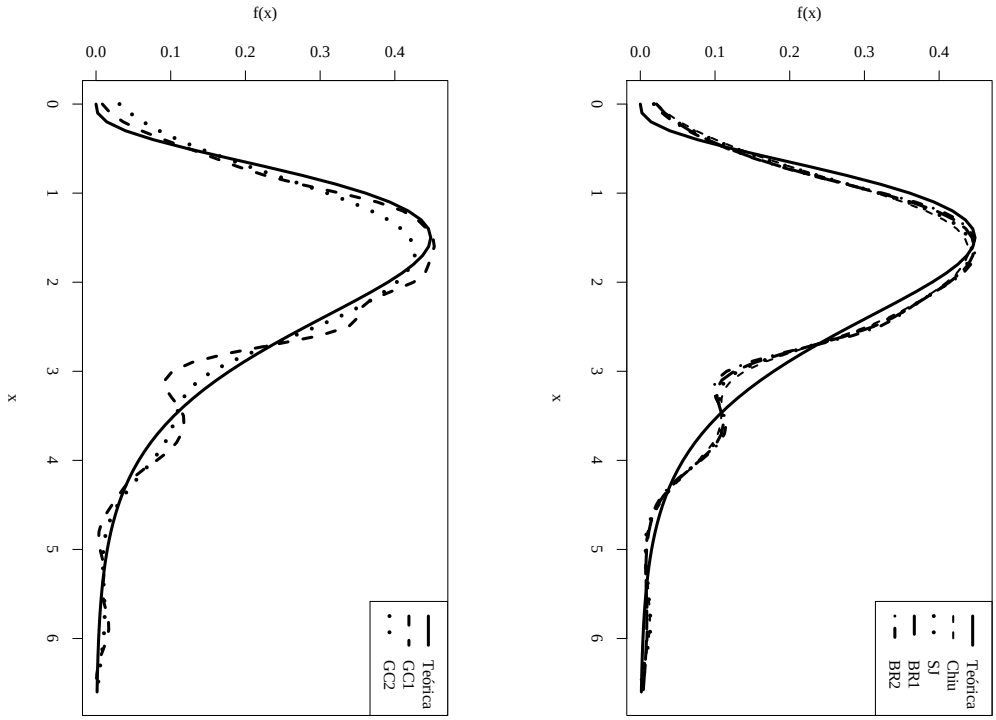

Figura A.8: Estimativa da distribuição Gama(4,2) com $n = 200$.

A.3 $X \sim$ Weibull$(1, 6)$

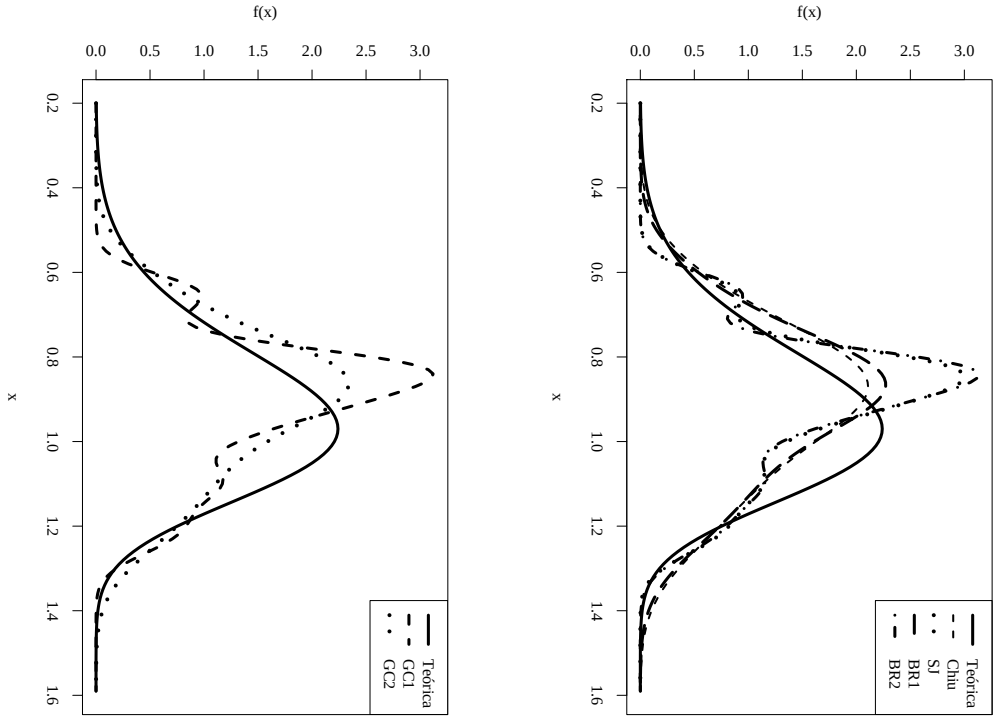

Figura A.9: Estimativa da distribuição Weibull(1,6) com $n = 30$.

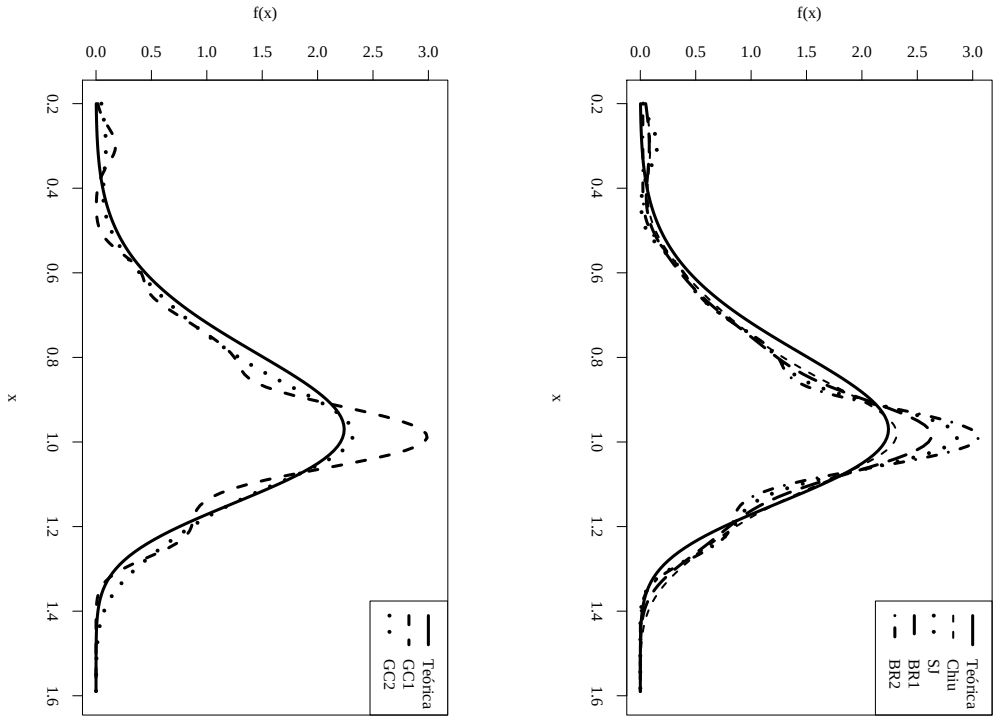

Figura A.10: Estimativa da distribuição Weibull(1,6) com $n = 50$.

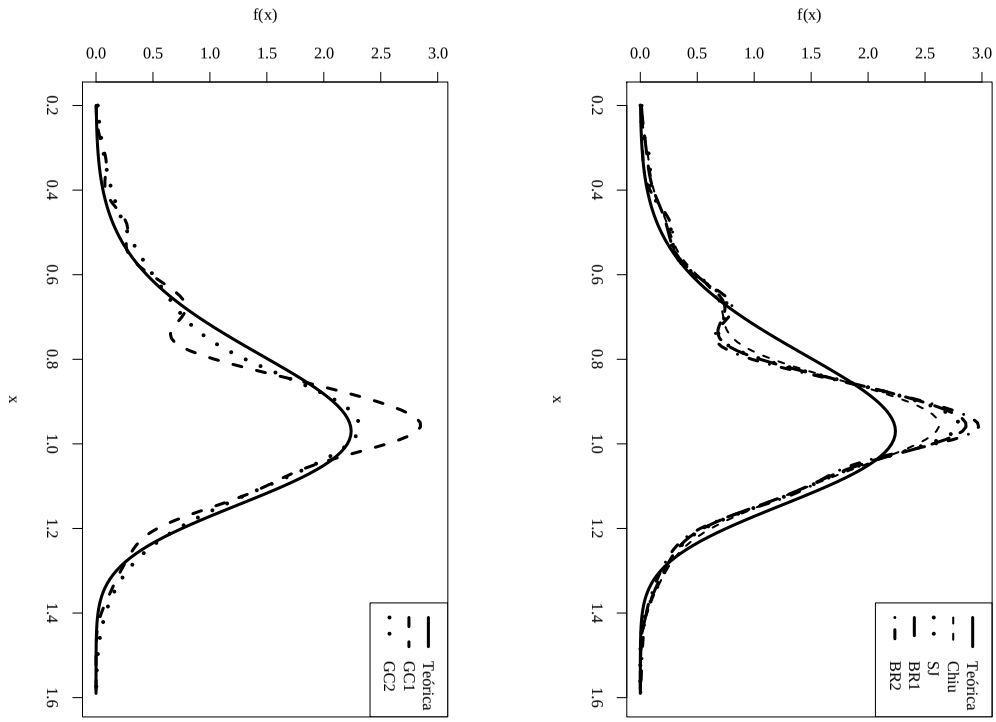

Figura A.11: Estimativa da distribuição Weibull(1,6) com $n = 100$.

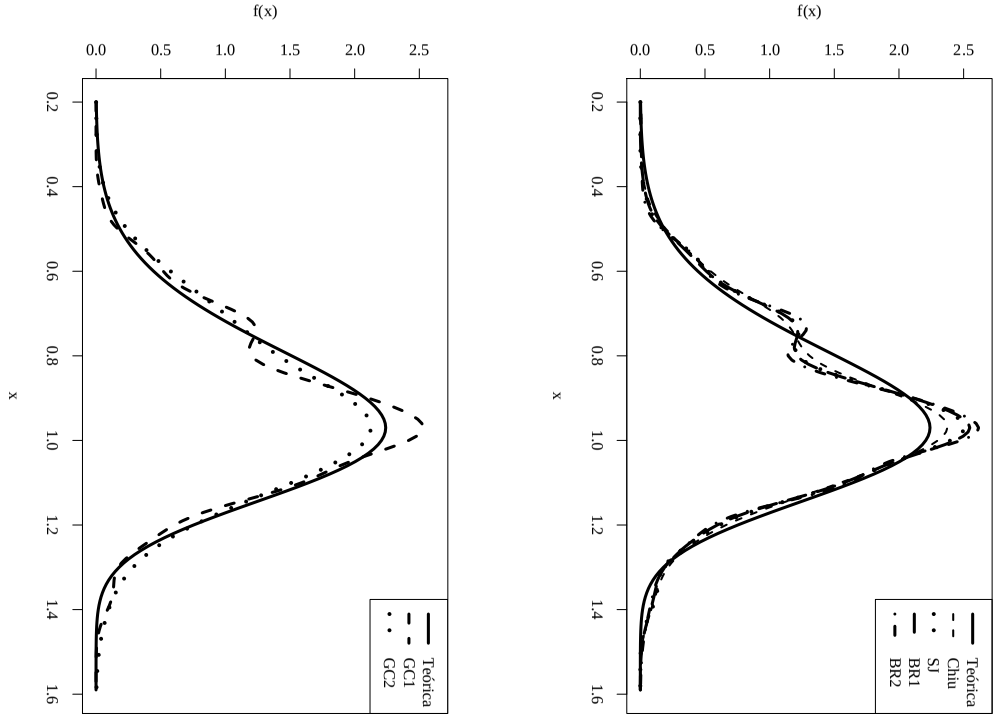

Figura A.12: Estimativa da distribuição Weibull(1,6) com $n = 200$.

A.4 $X \sim$ Qui-Quadrado(7)

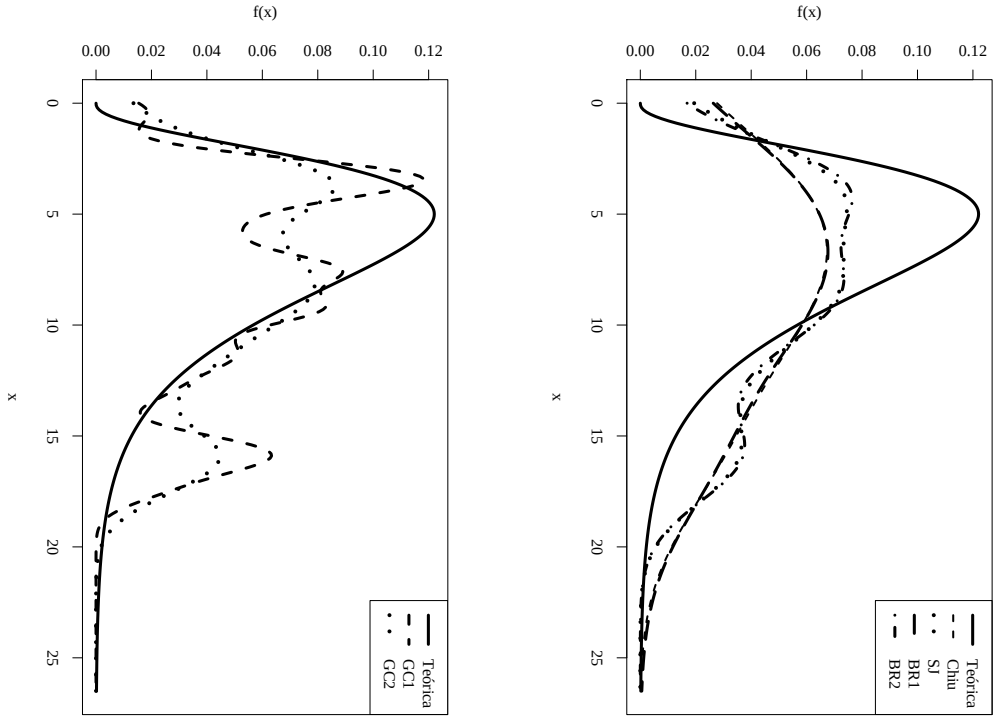

Figura A.13: Estimativa da distribuição Qui-Quadrado(7) com $n = 30$.

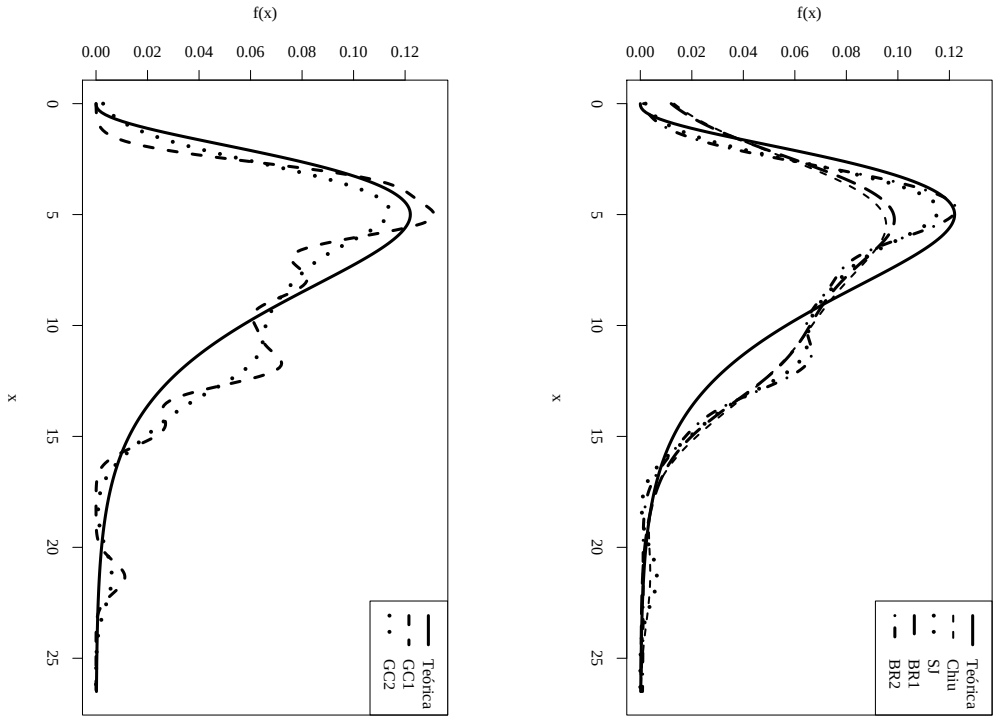

Figura A.14: Estimativa da distribuição Qui-Quadrado(7) com $n = 50$.

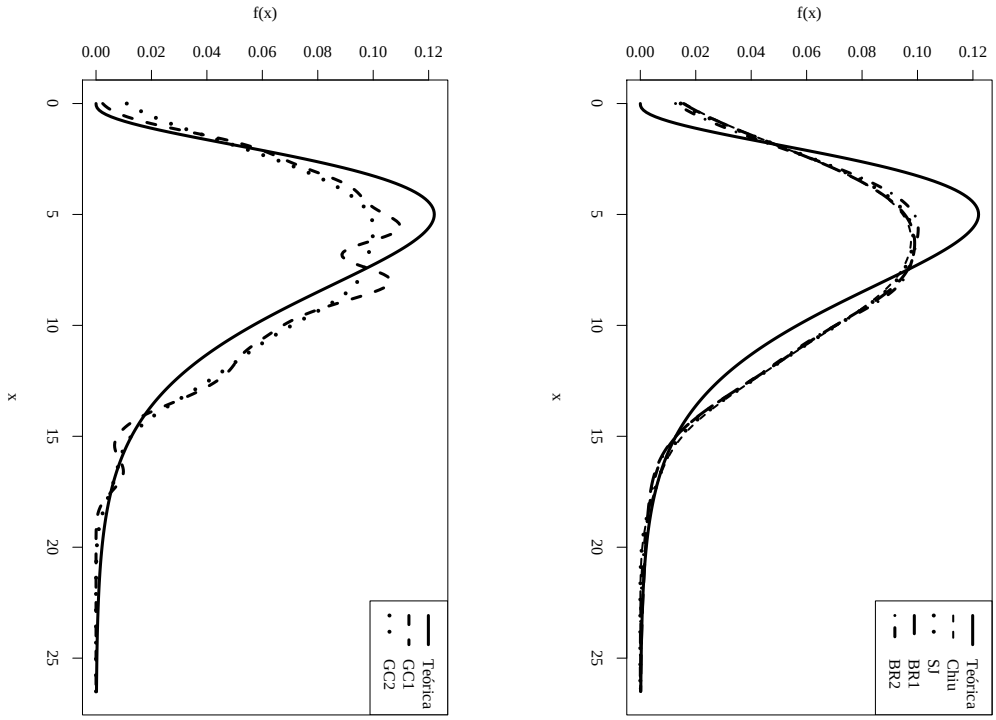

Figura A.15: Estimativa da distribuição Qui-Quadrado(7) com $n = 100$.

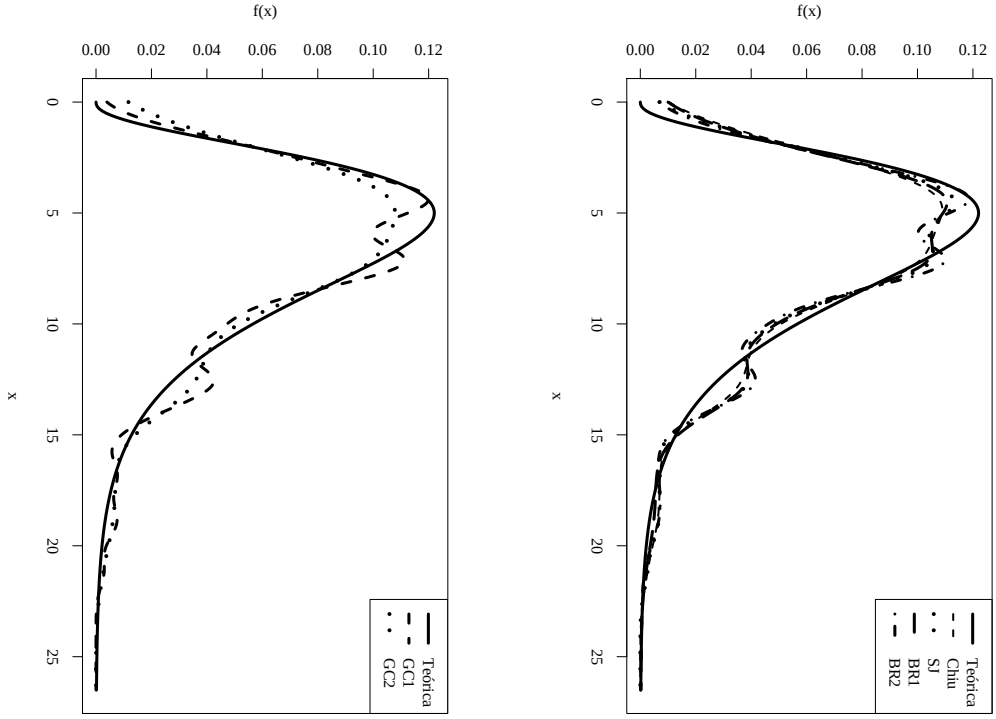

Figura A.16: Estimativa da distribuição Qui-Quadrado(7) com $n = 200$.

A.5 $X \sim Mistura(1)$

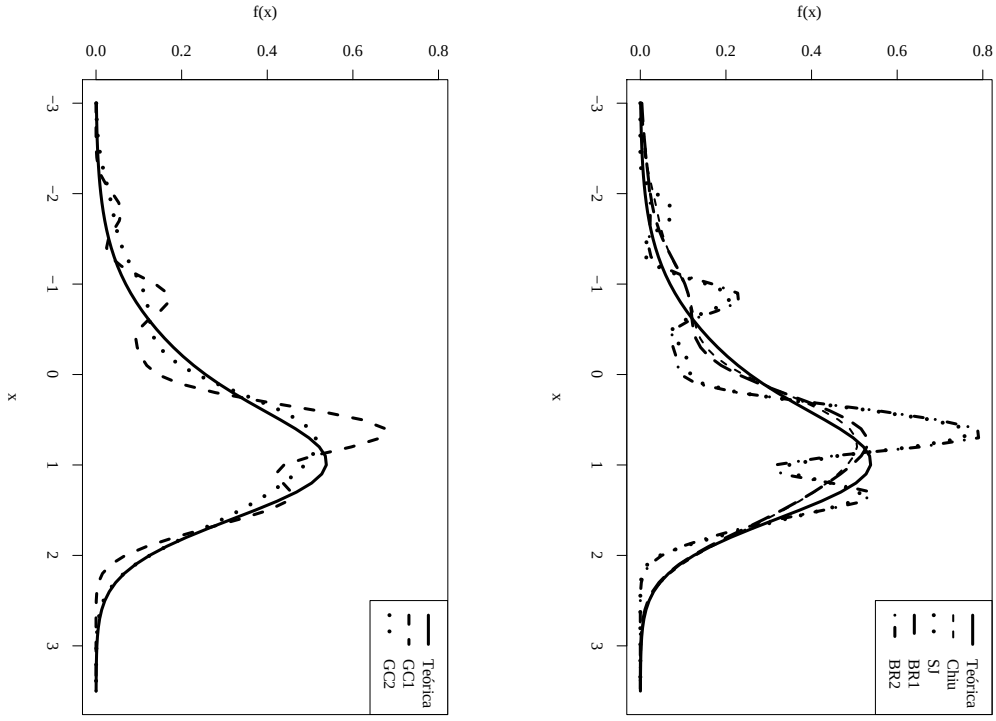

Figura A.17: Estimativa da distribuição Mistura(1) com $n = 30$.

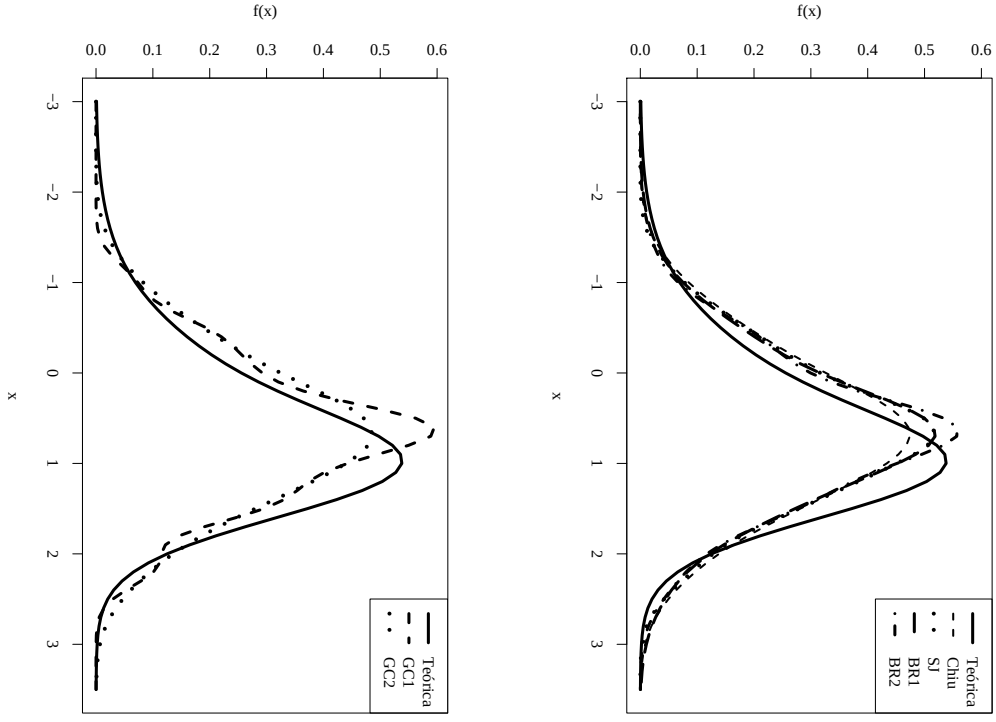

Figura A.18: Estimativa da distribuição Mistura(1) com $n = 50$.

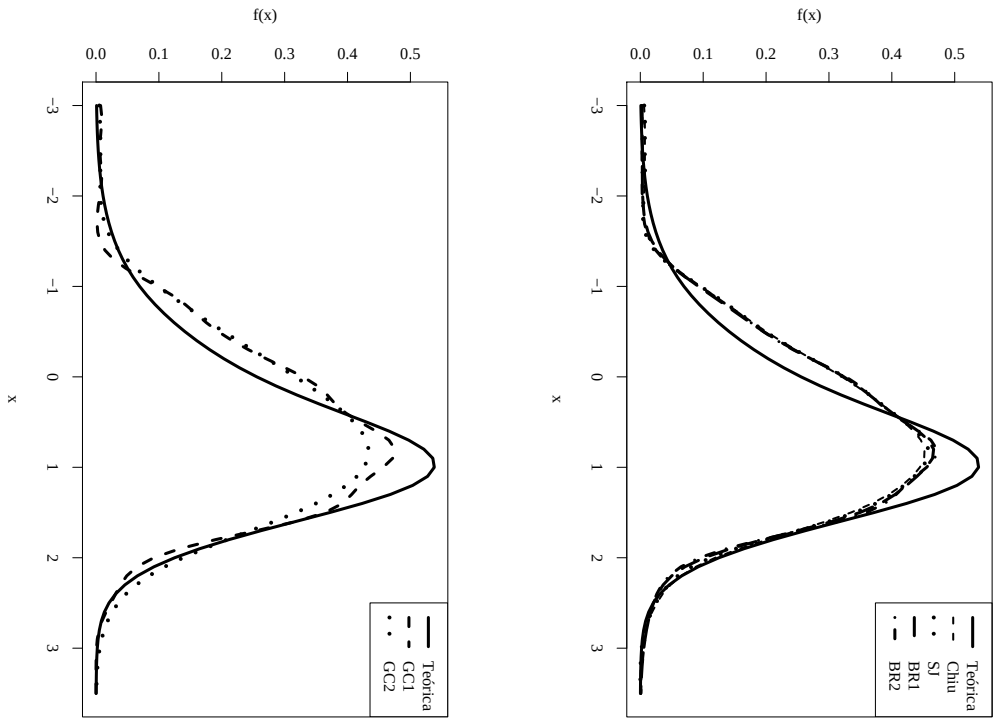

Figura A.19: Estimativa da distribuição Mistura(1) com $n = 100$.

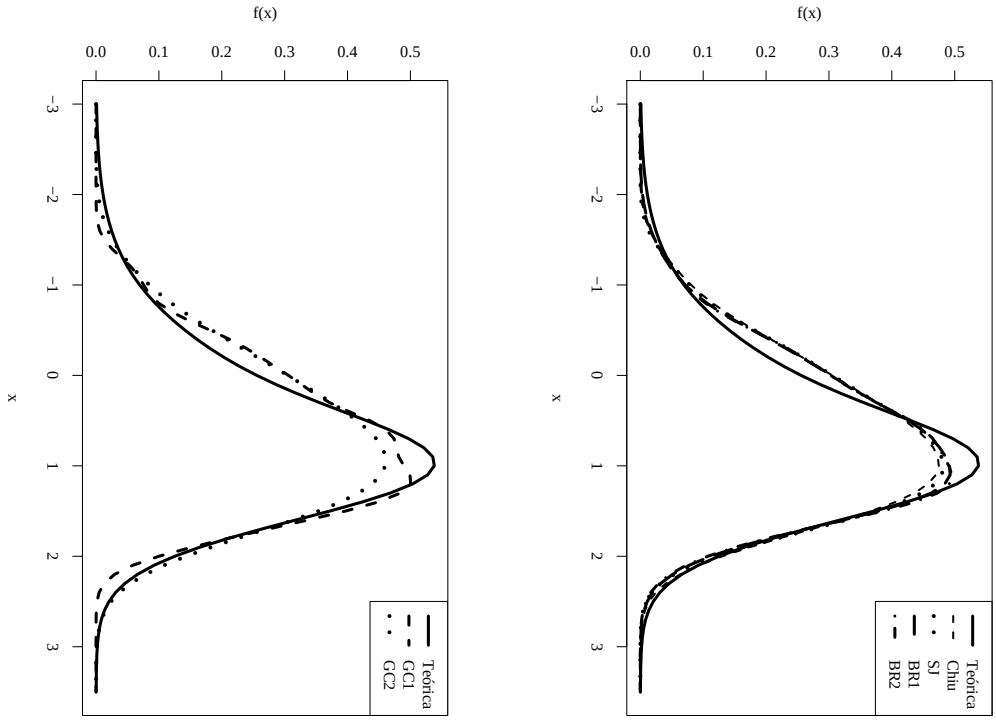

Figura A.20: Estimativa da distribuição Mistura(1) com $n = 200$.

A.6 $X \sim Mistura(2)$

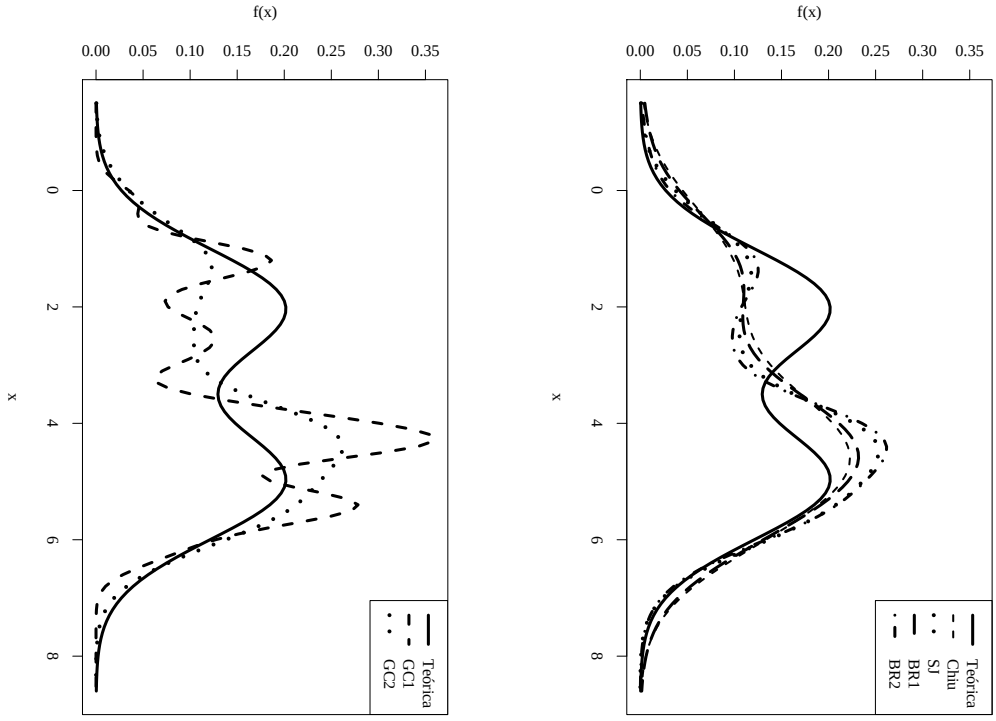

Figura A.21: Estimativa da distribuição Mistura(2) com $n = 30$.

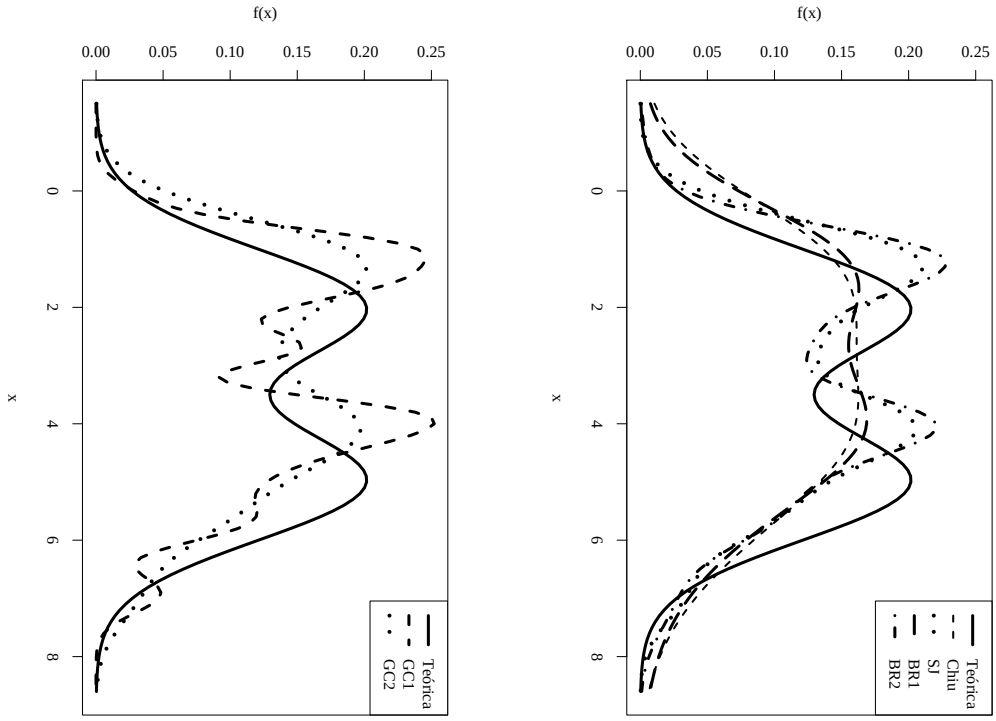

Figura A.22: Estimativa da distribuição Mistura(2) com $n = 50$.

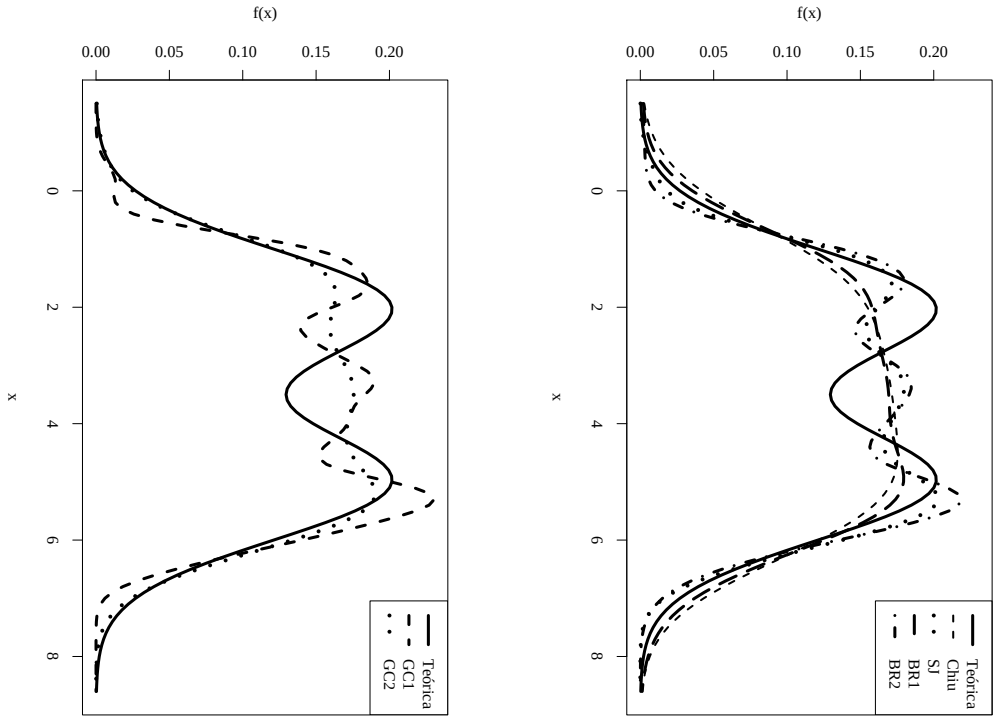

Figura A.23: Estimativa da distribuição Mistura(2) com $n = 100$.

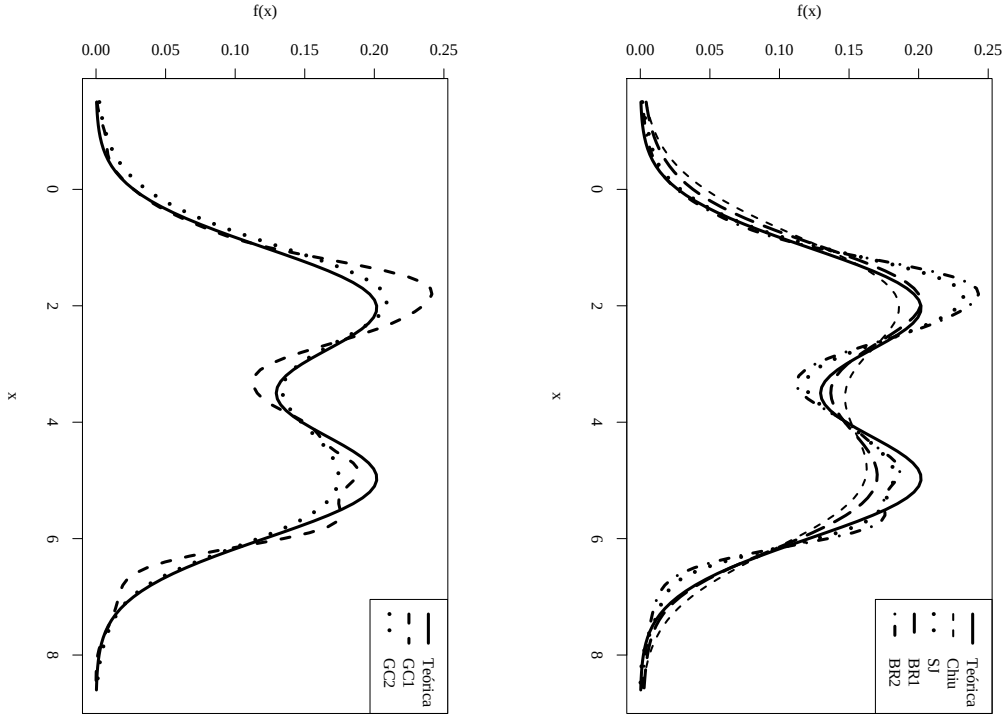

Figura A.24: Estimativa da distribuição Mistura(2) com $n = 200$.

A.7 $X \sim Mistura(3)$

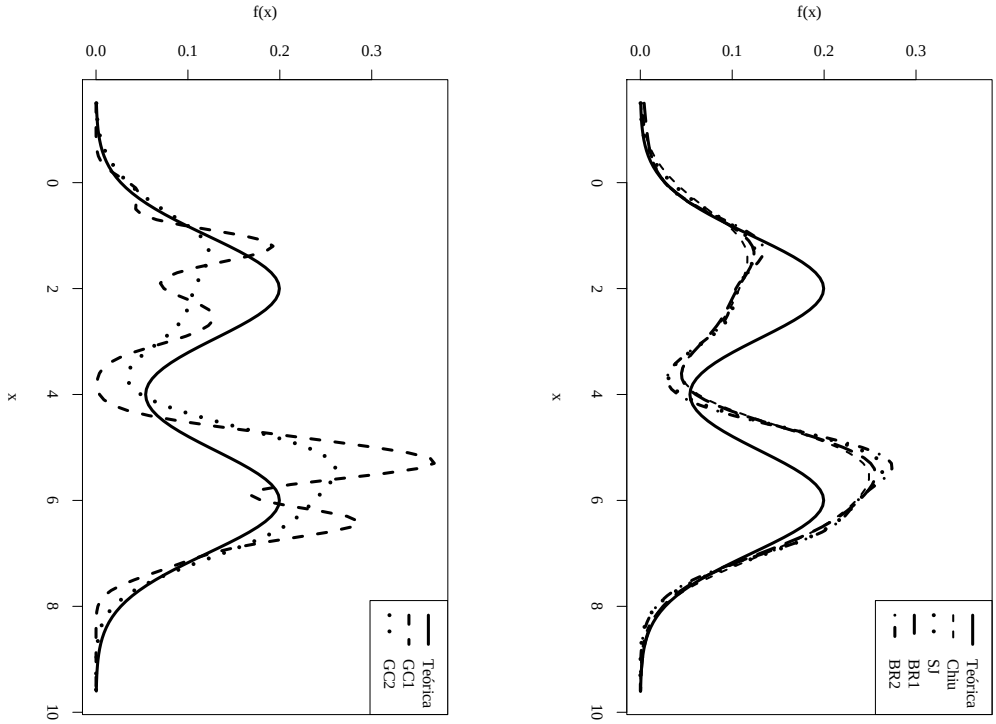

Figura A.25: Estimativa da distribuição Mistura(3) com $n = 30$.

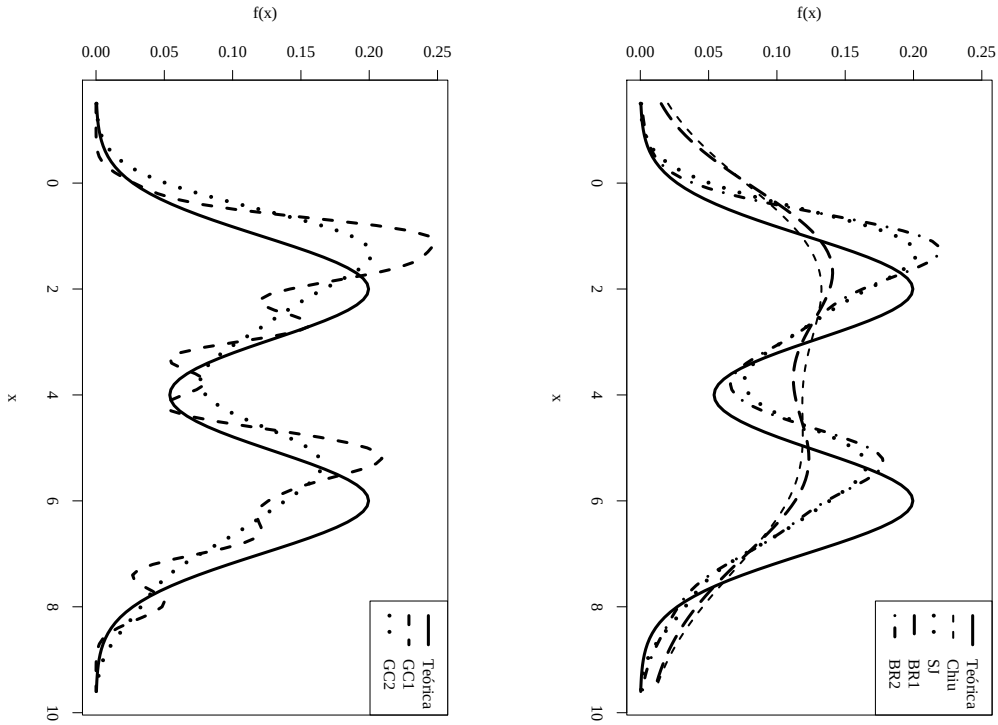

Figura A.26: Estimativa da distribuição Mistura(3) com $n = 50$.

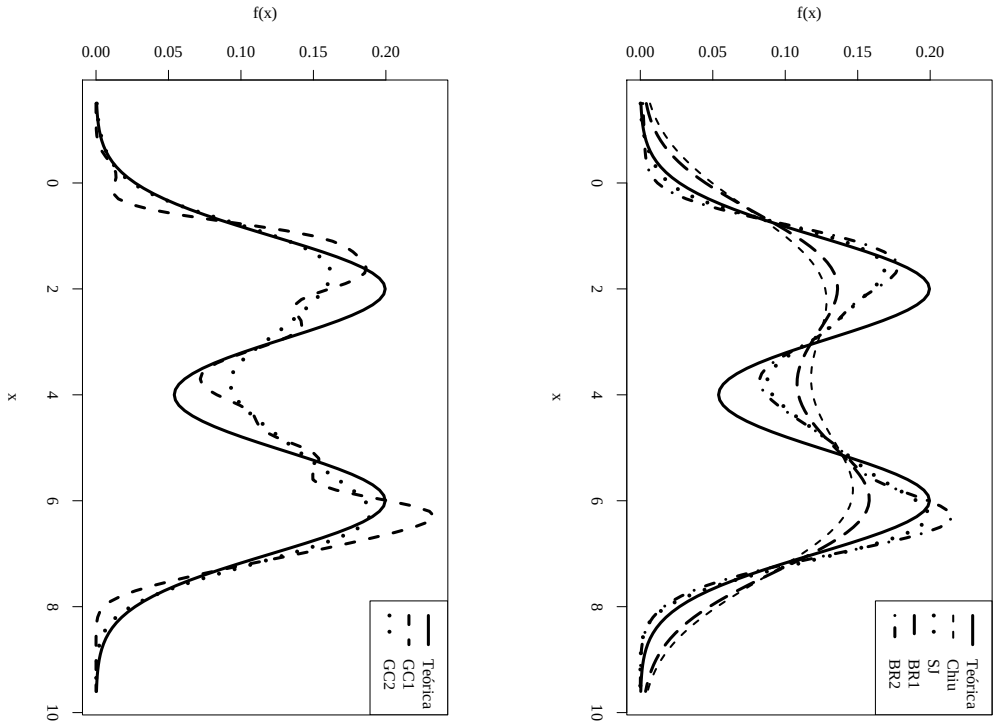

Figura A.27: Estimativa da distribuição Mistura(3) com $n = 100$.

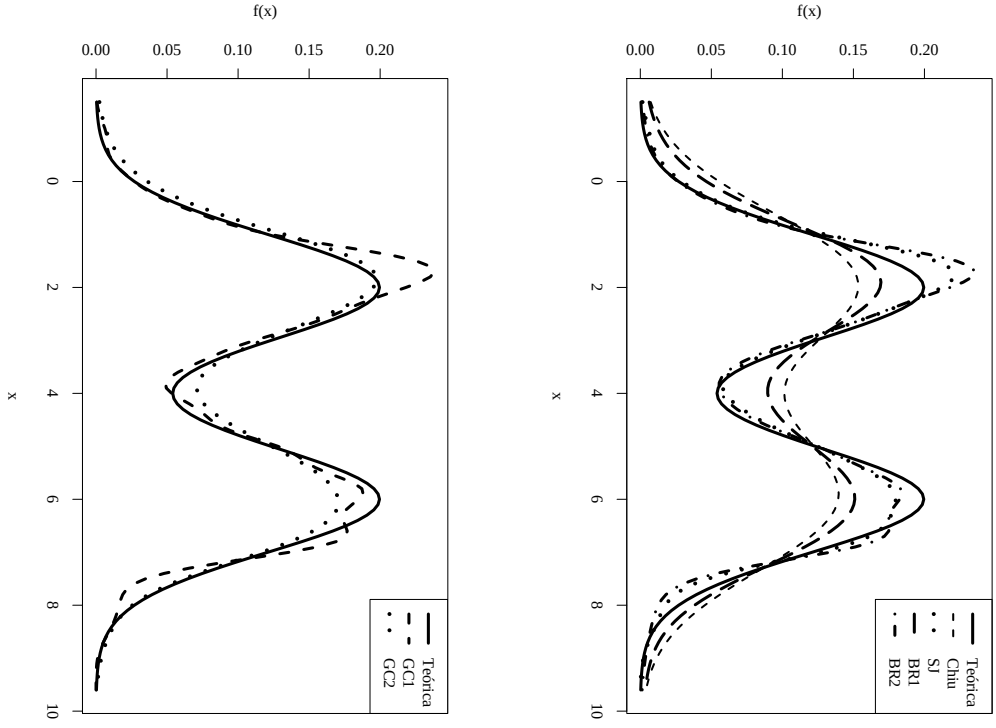

Figura A.28: Estimativa da distribuição Mistura(3) com $n = 200$.

A.8 $X \sim Mistura(4)$

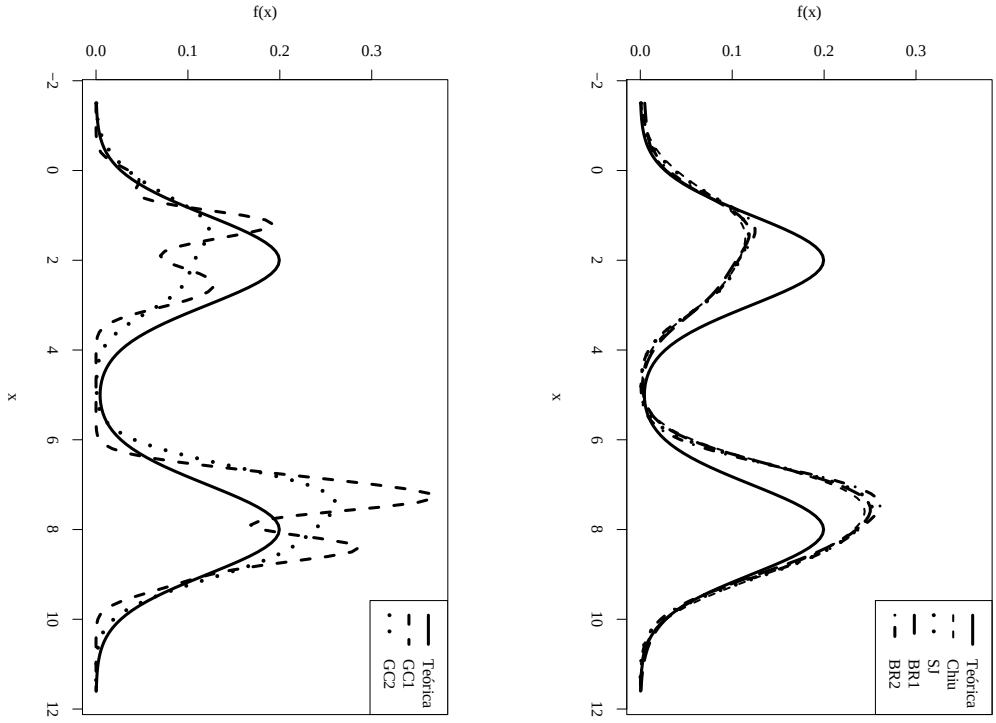

Figura A.29: Estimativa da distribuição Mistura(4) com $n = 30$.

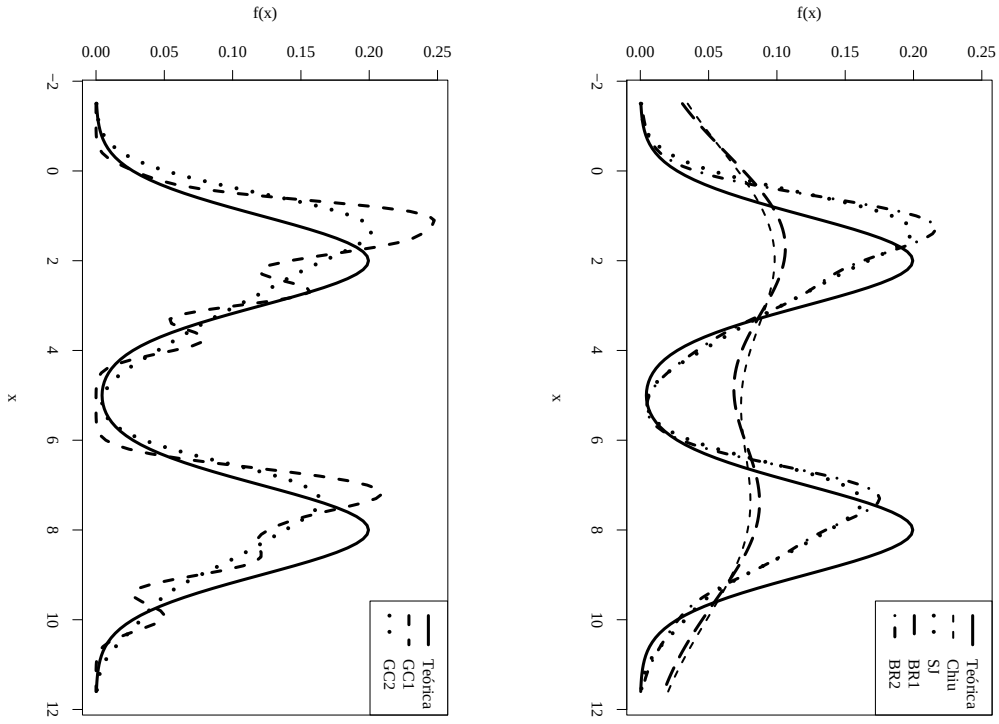

Figura A.30: Estimativa da distribuição Mistura(4) com $n = 50$.

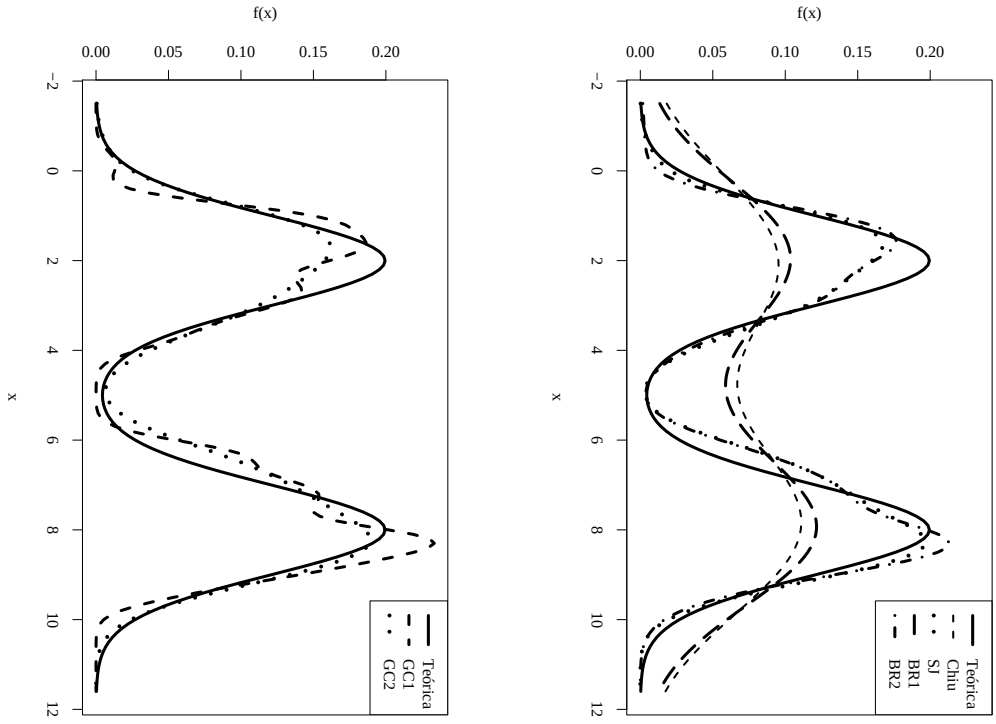

Figura A.31: Estimativa da distribuição Mistura(4) com $n = 100$.

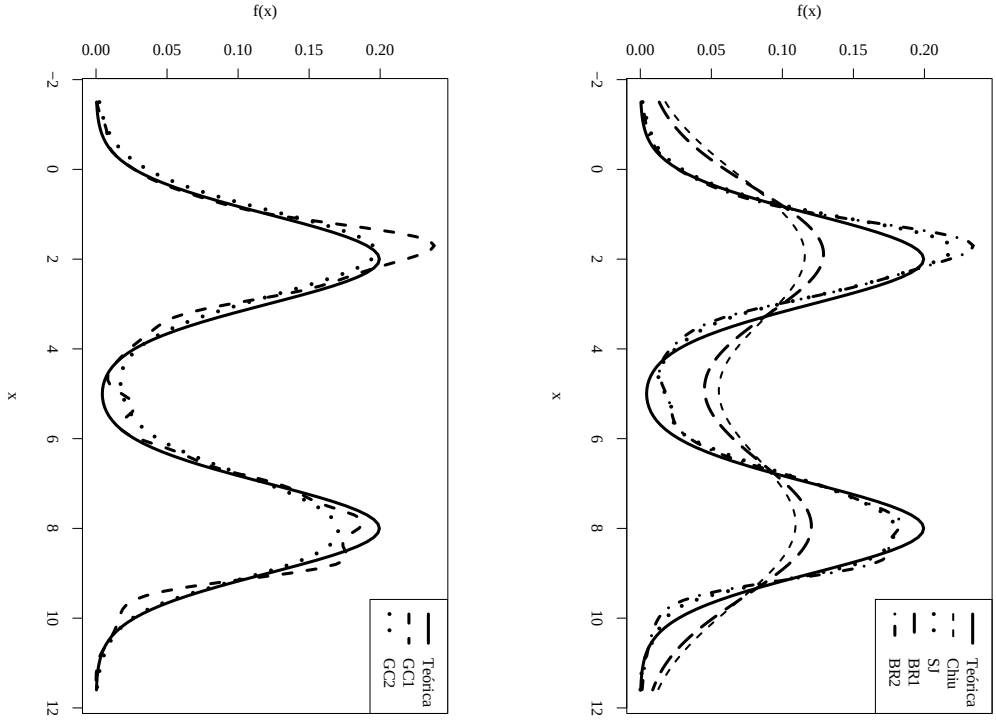

Figura A.32: Estimativa da distribuição Mistura(4) com $n = 200$.

A.9 $X \sim Mistura(5)$

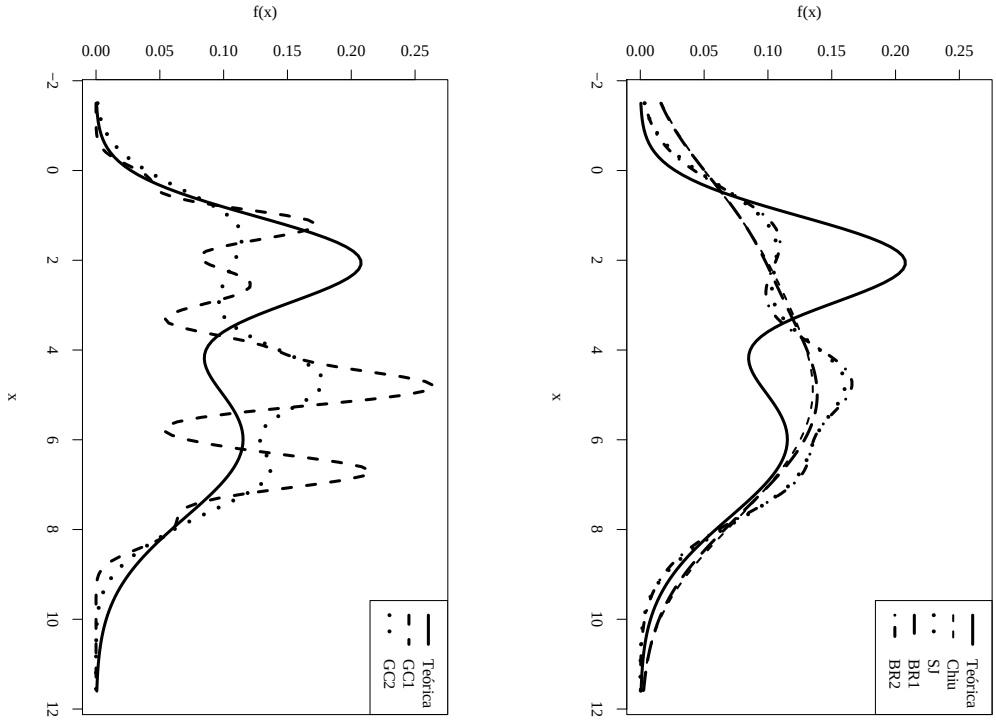

Figura A.33: Estimativa da distribuição Mistura(5) com $n = 30$.

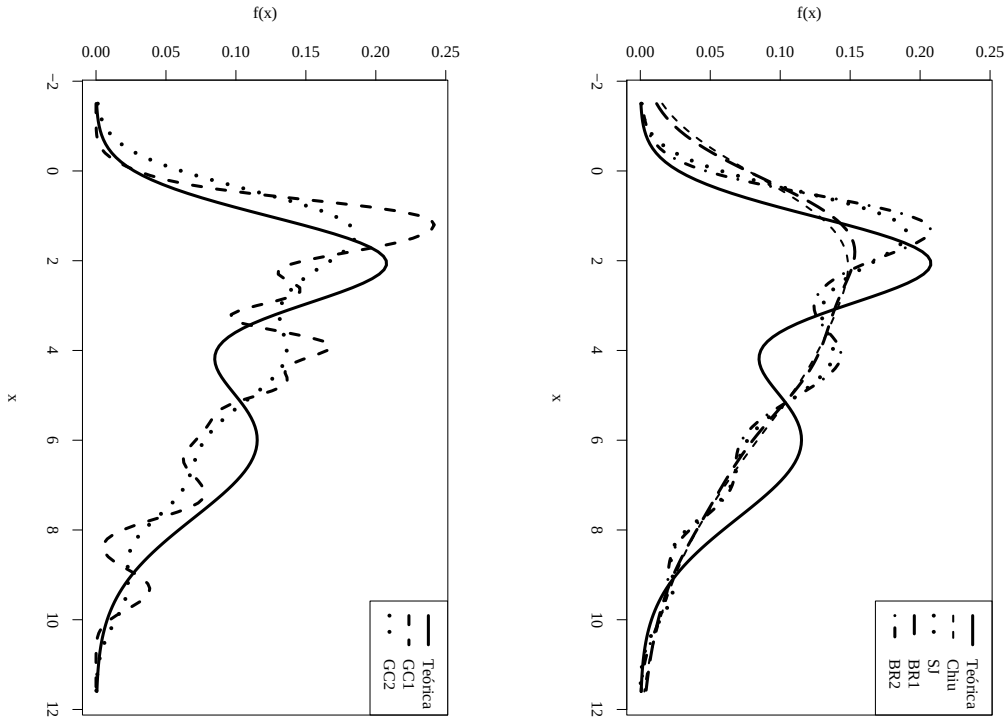

Figura A.34: Estimativa da distribuição Mistura(5) com $n = 50$.

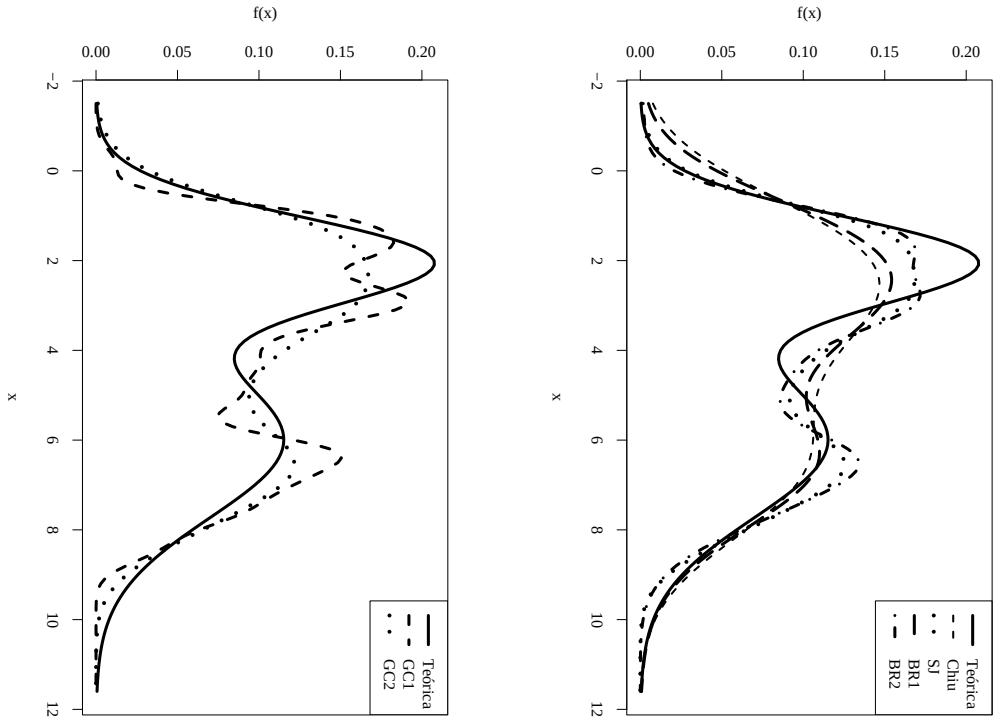

Figura A.35: Estimativa da distribuição Mistura(5) com $n = 100$.

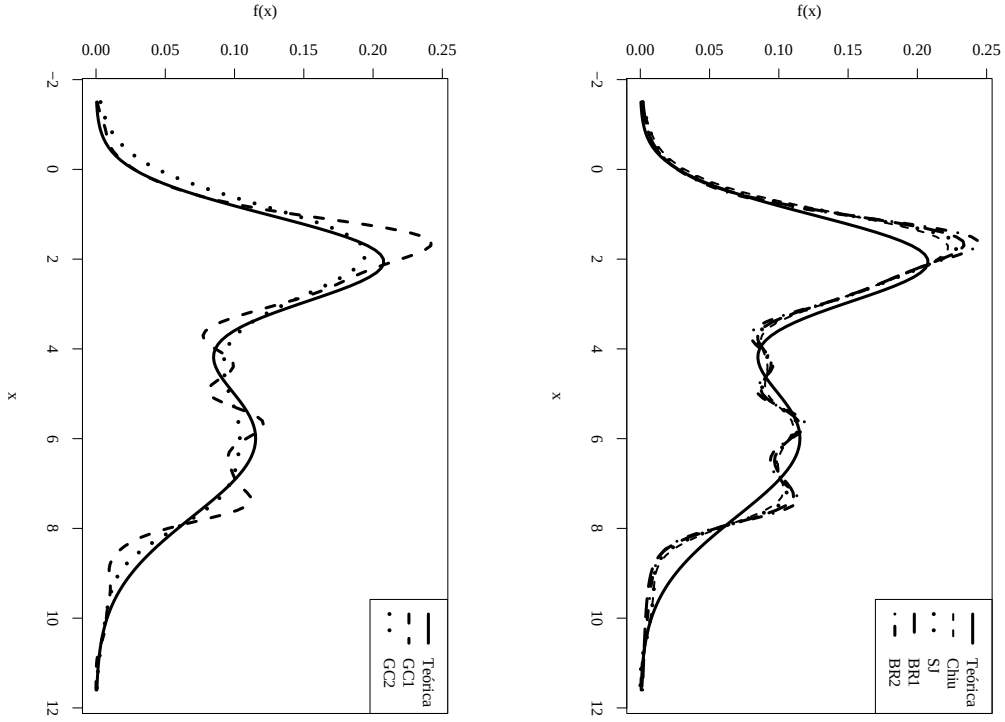

Figura A.36: Estimativa da distribuição Mistura(5) com $n = 200$.

A.10 $X \sim Mistura(6)$

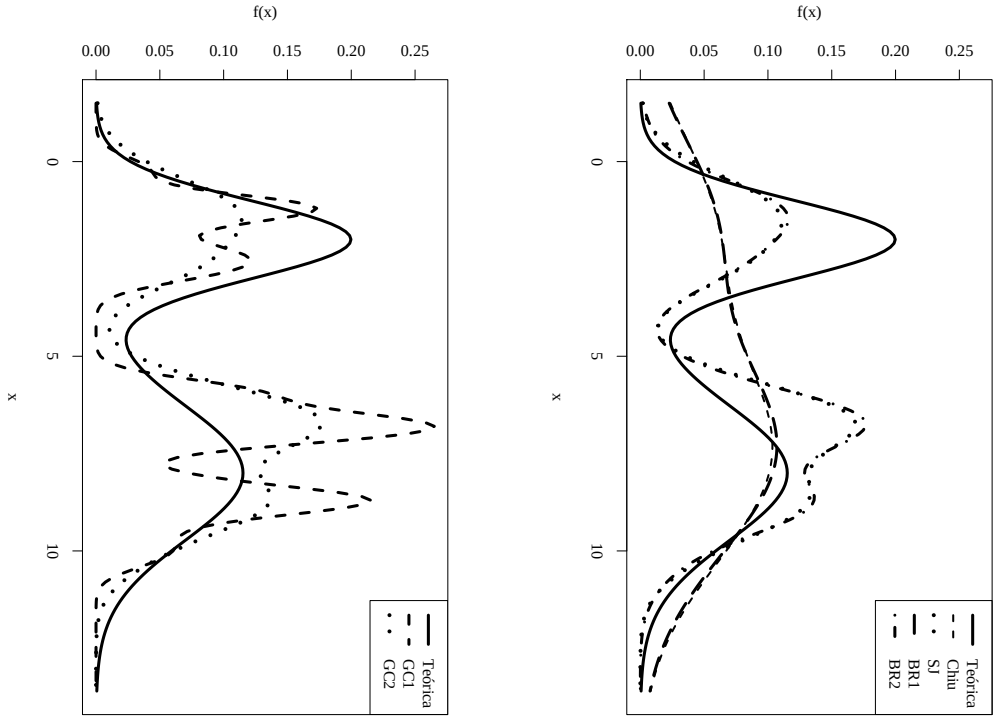

Figura A.37: Estimativa da distribuição Mistura(6) com $n = 30$.

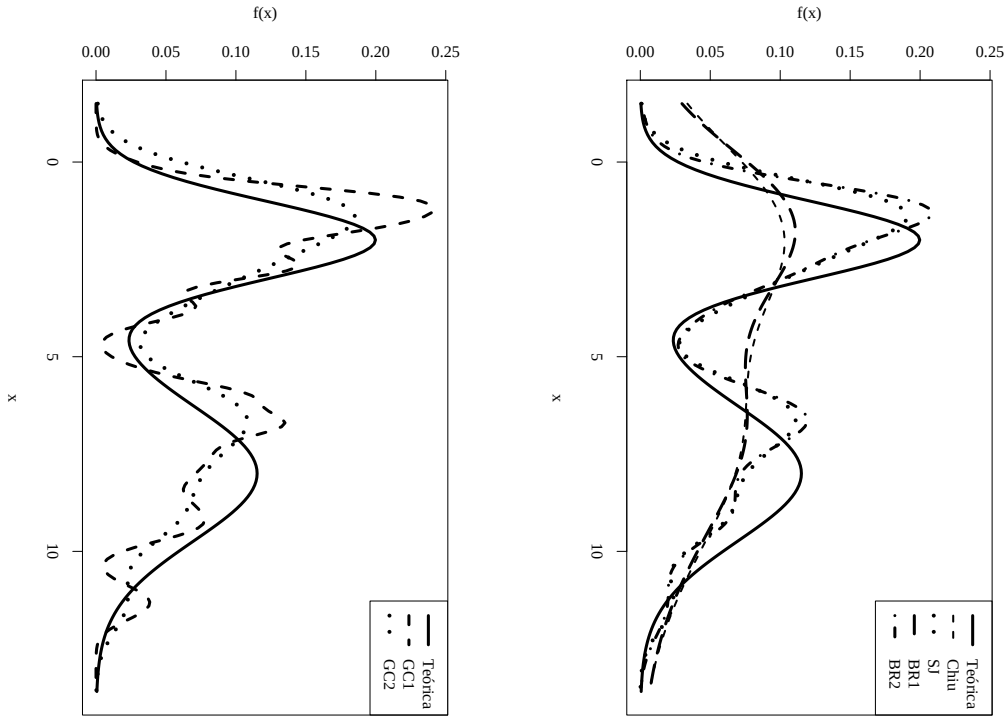

Figura A.38: Estimativa da distribuição Mistura(6) com $n = 50$.

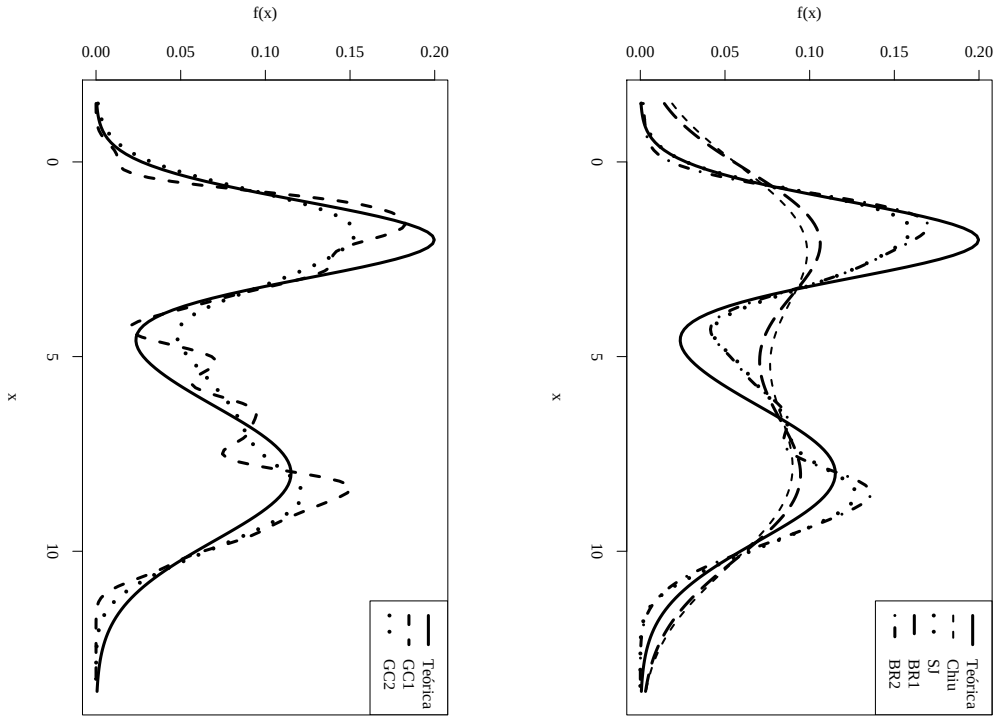

Figura A.39: Estimativa da distribuição Mistura(6) com $n = 100$.

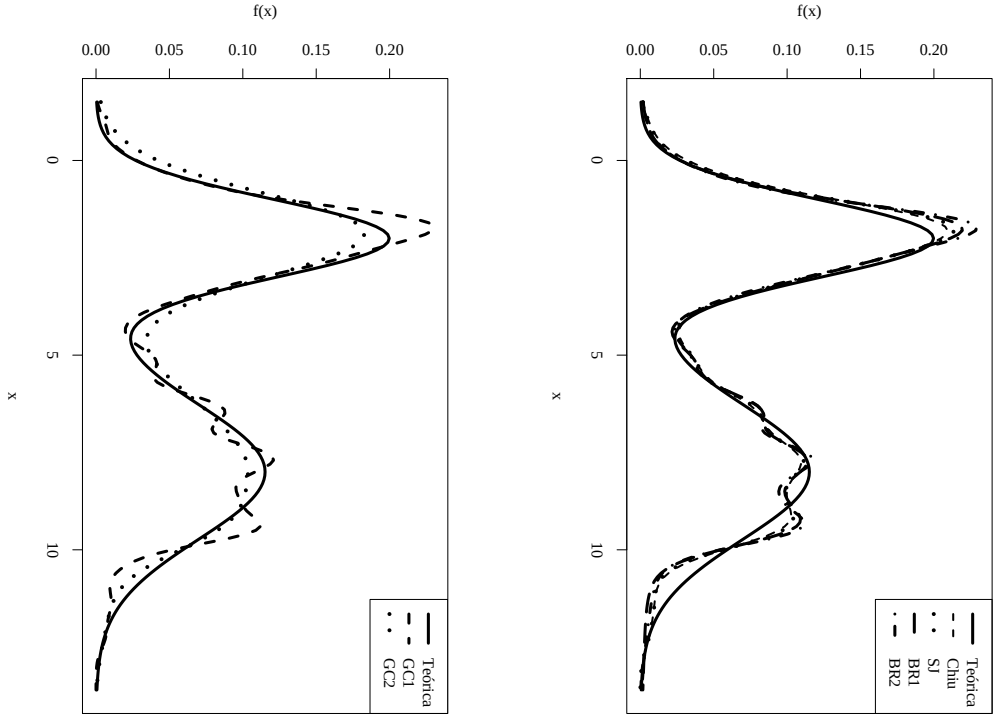

Figura A.40: Estimativa da distribuição Mistura(6) com $n = 200$.

A.11 $X \sim Mistura(7)$

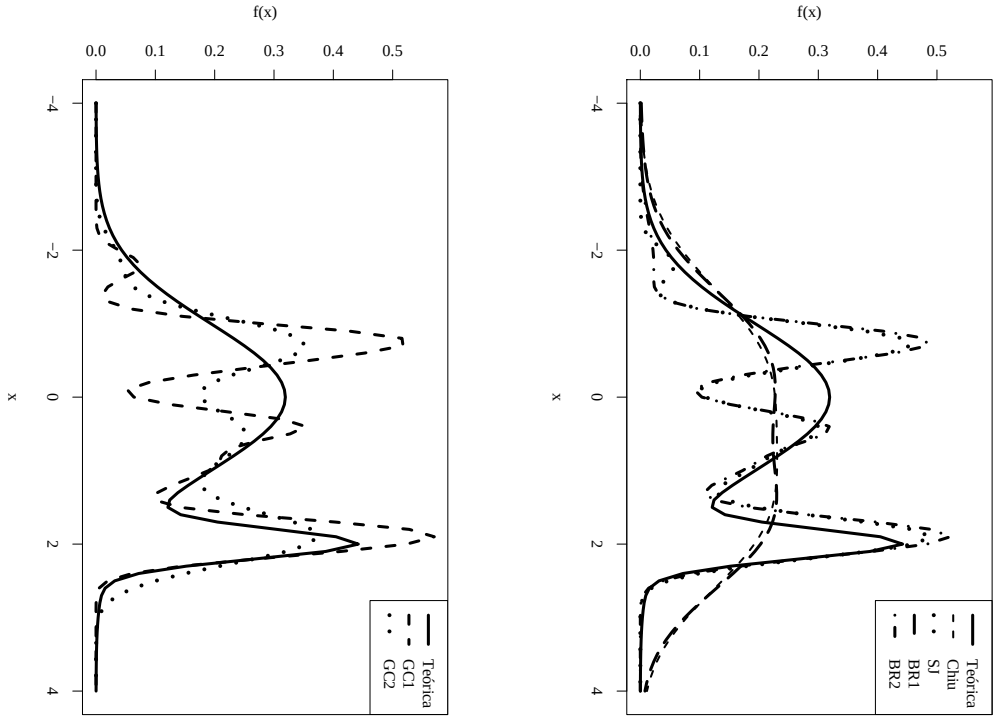

Figura A.41: Estimativa da distribuição Mistura(7) com $n = 30$.

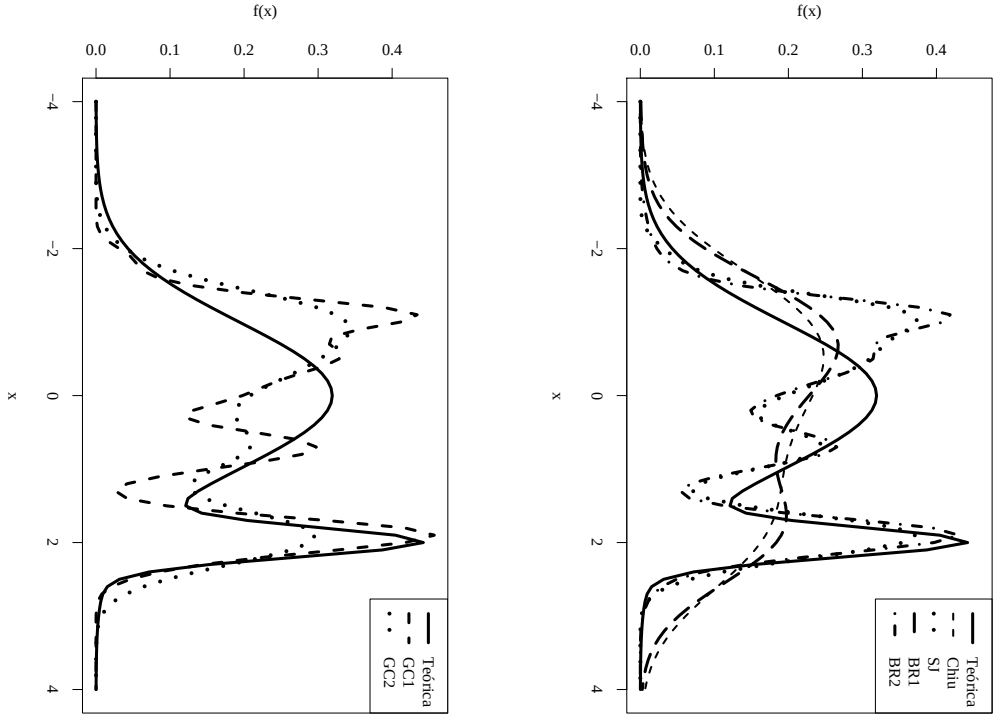

Figura A.42: Estimativa da distribuição Mistura(7) com $n = 50$.

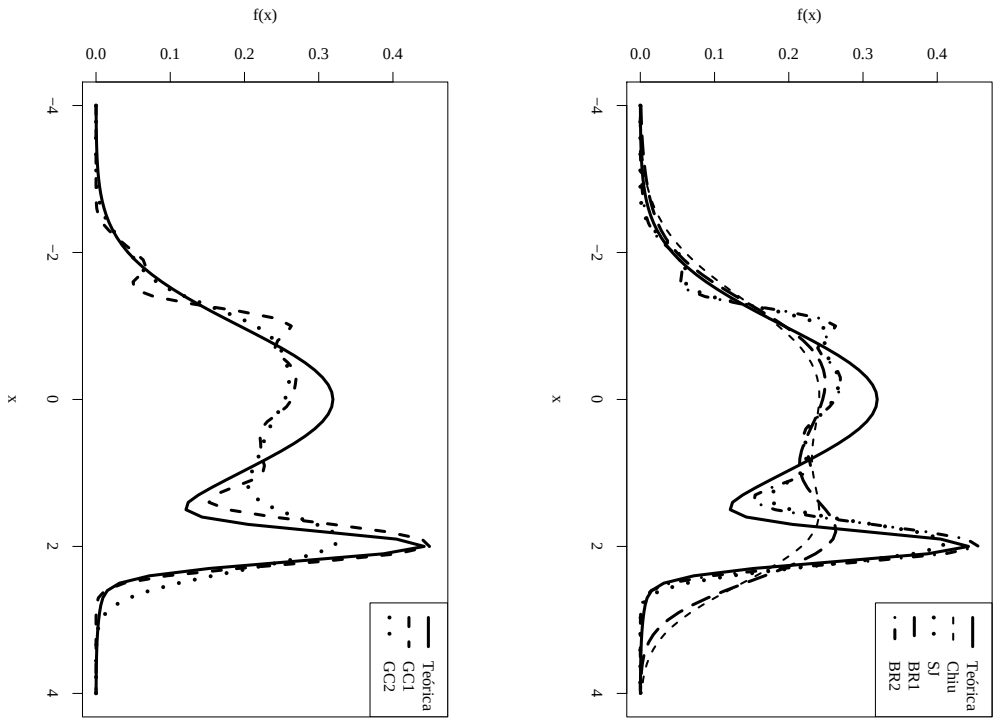

Figura A.43: Estimativa da distribuição Mistura(7) com $n = 100$.

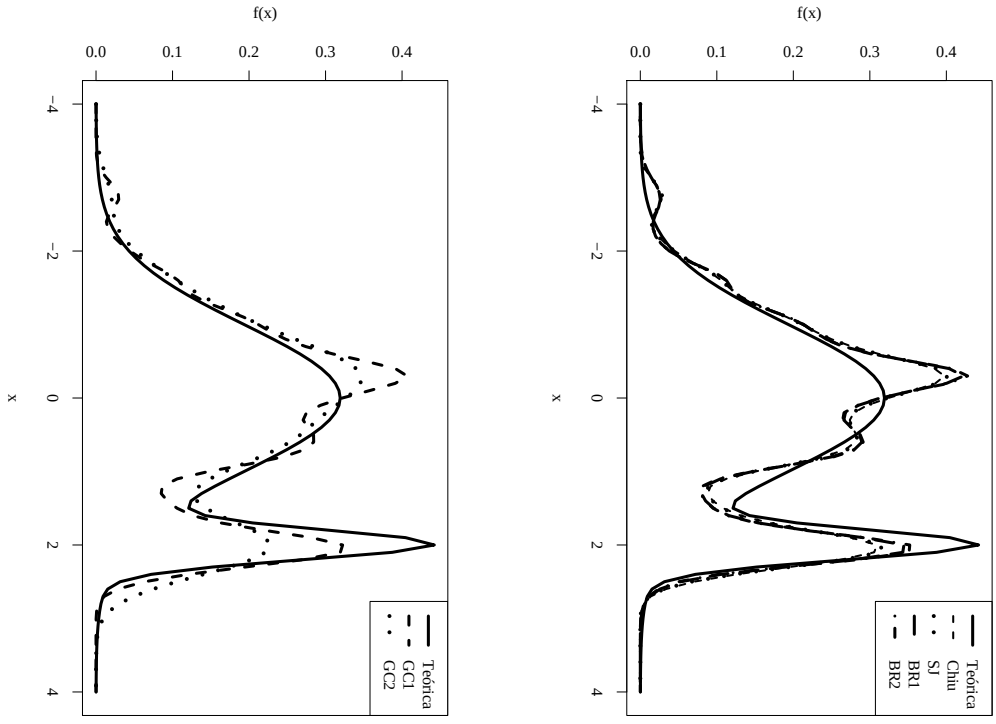

Figura A.44: Estimativa da distribuição Mistura(7) com $n = 200$.

A.12 $X \sim Mistura(8)$

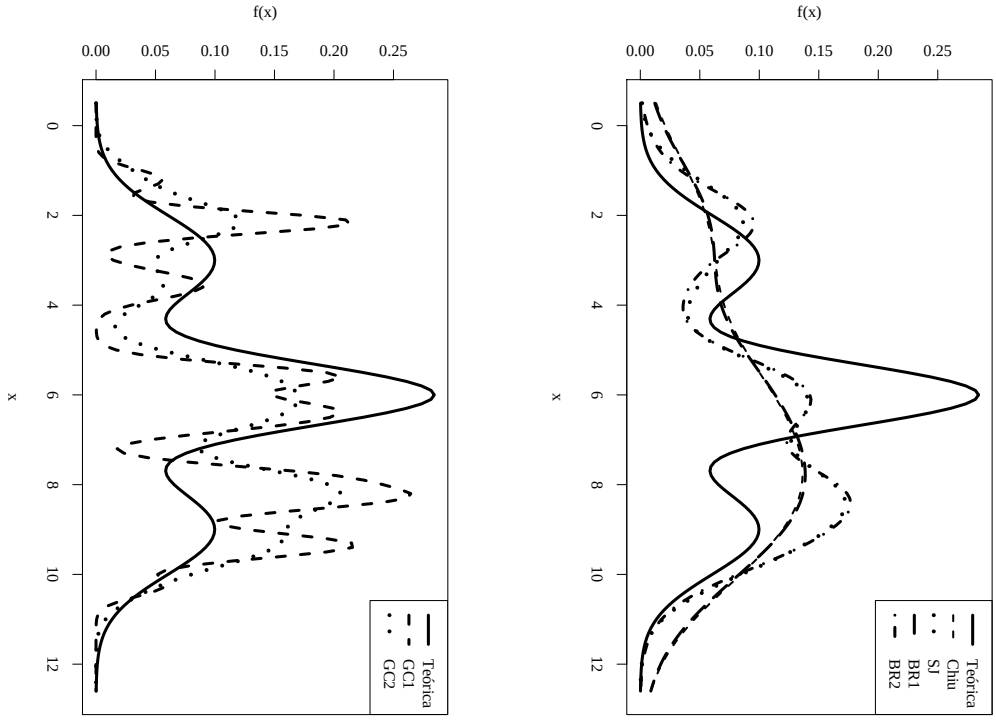

Figura A.45: Estimativa da distribuição Mistura(8) com $n = 30$.

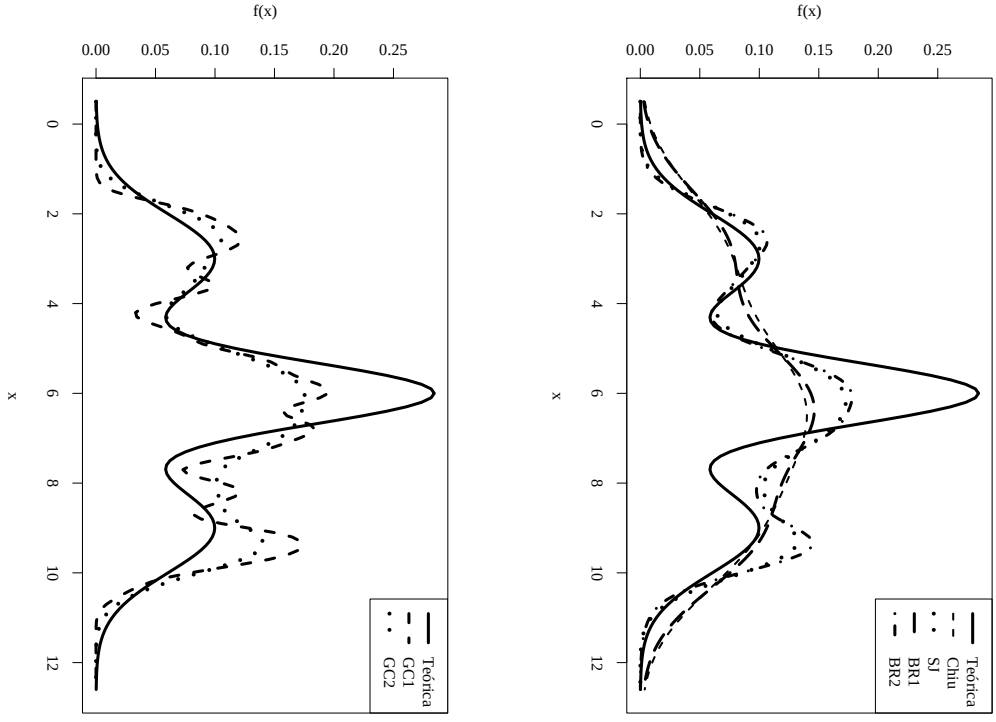

Figura A.46: Estimativa da distribuição Mistura(8) com $n = 50$.

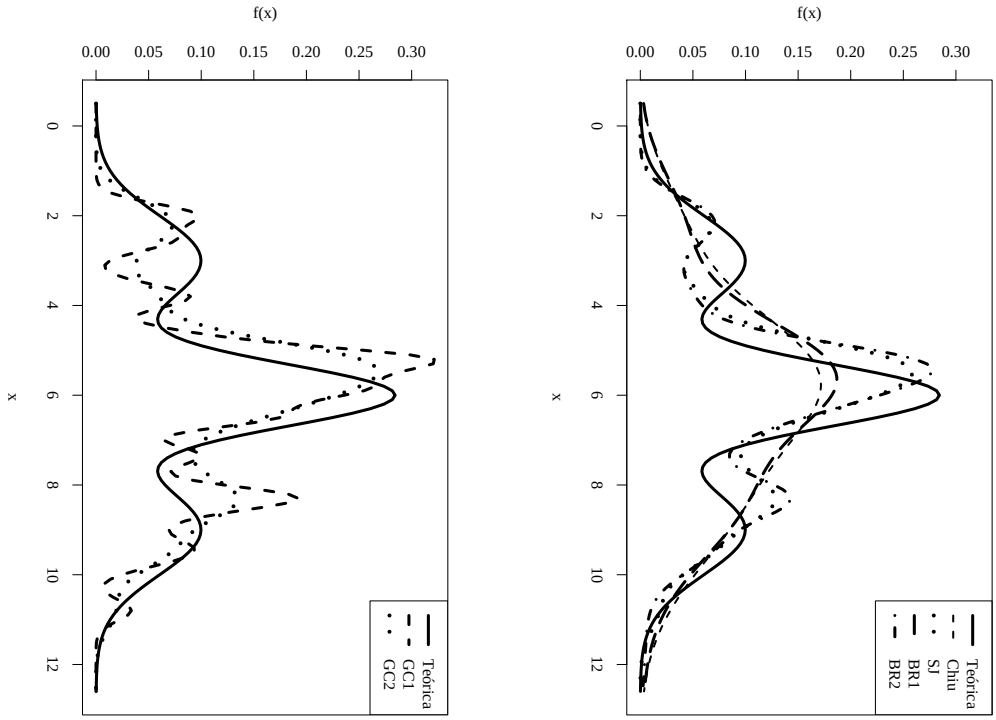

Figura A.47: Estimativa da distribuição Mistura(8) com $n = 100$.

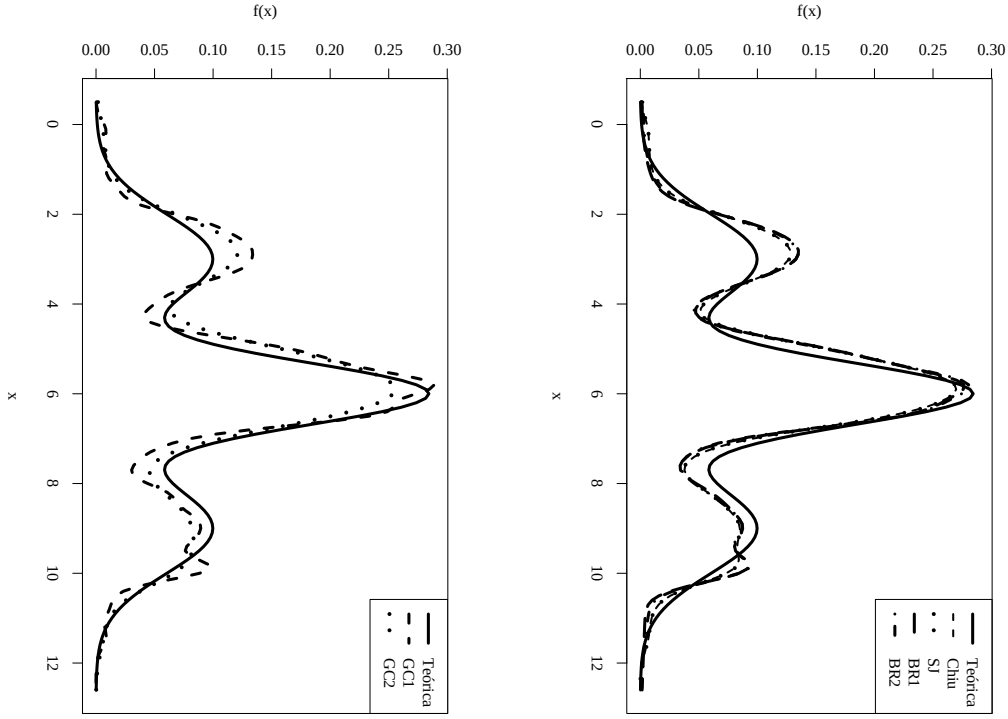

Figura A.48: Estimativa da distribuição Mistura(8) com $n = 200$.

A.13 $X \sim Mistura(9)$

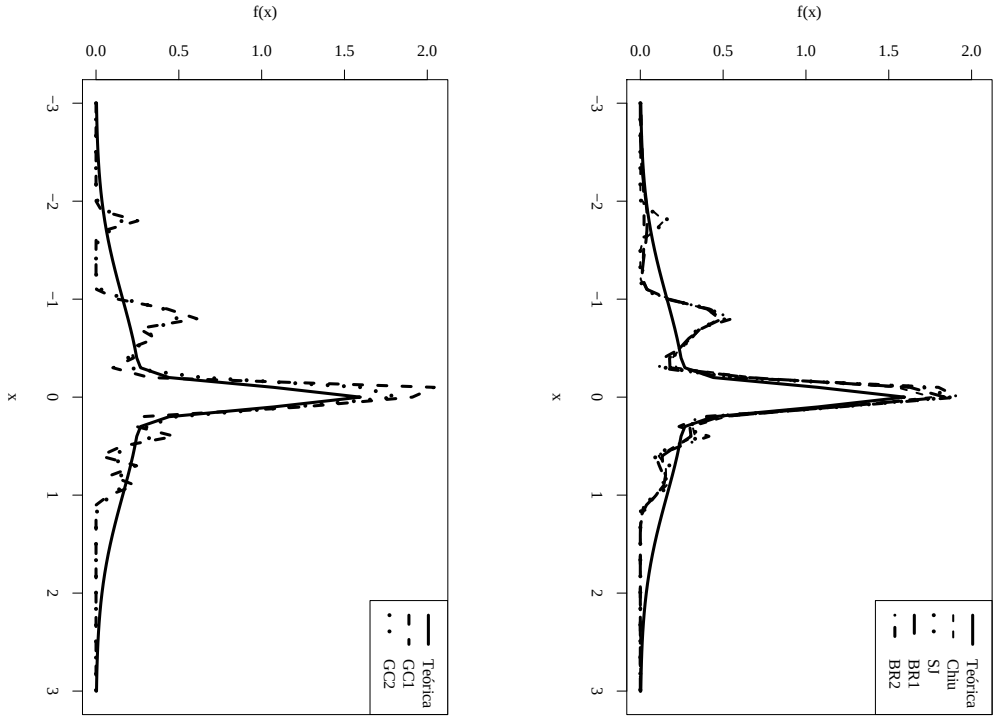

Figura A.49: Estimativa da distribuição Mistura(9) com $n = 30$.

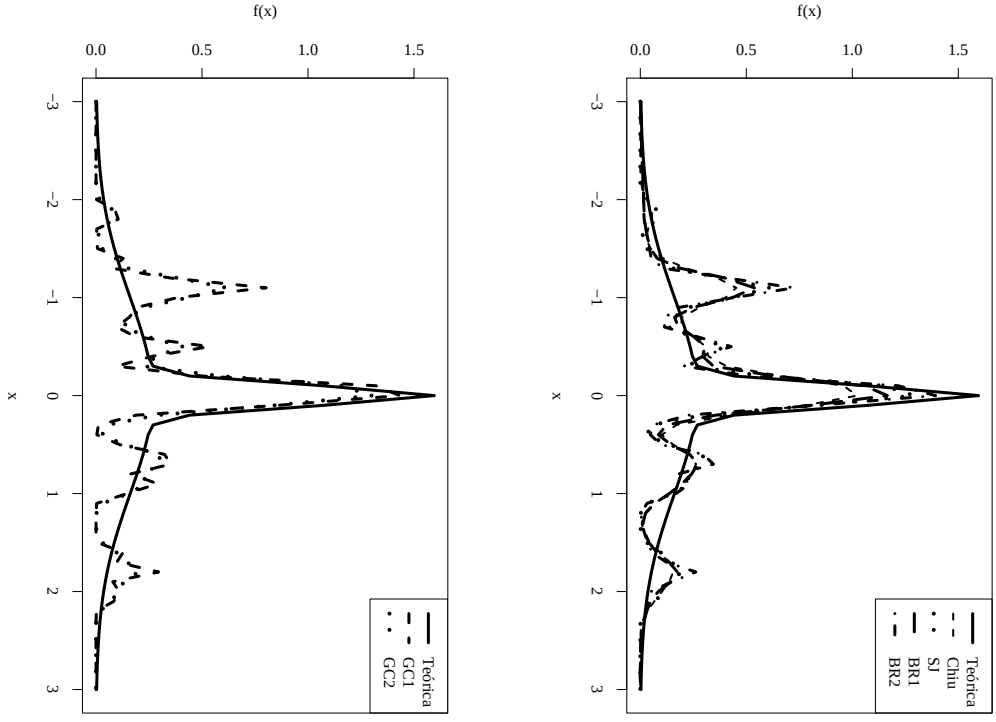

Figura A.50: Estimativa da distribuição Mistura(9) com $n = 50$.

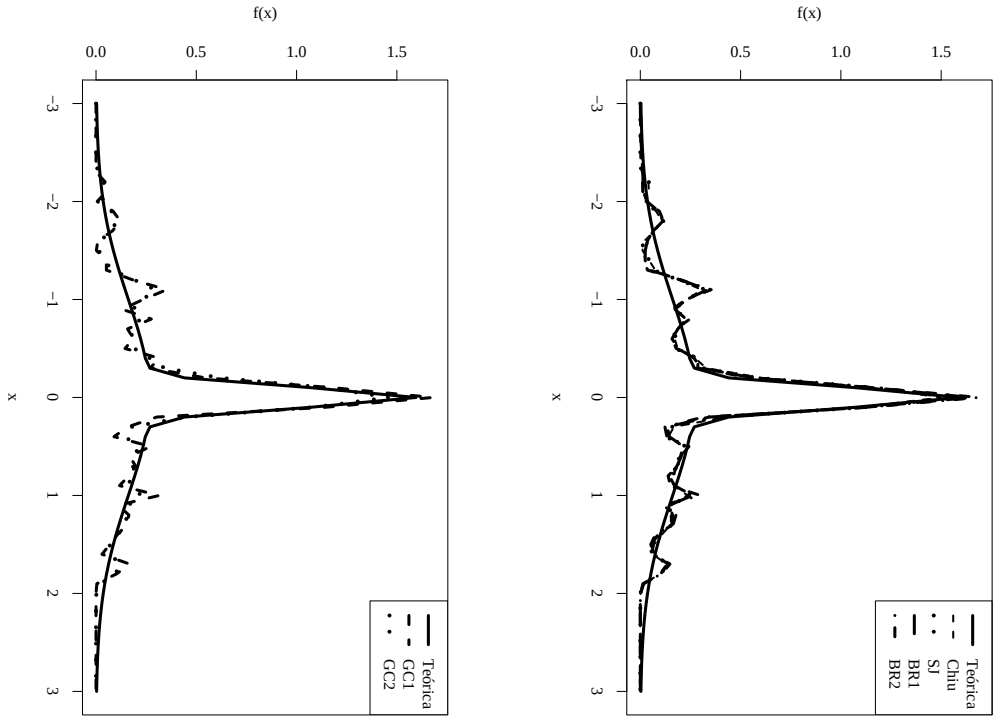

Figura A.51: Estimativa da distribuição Mistura(9) com $n = 100$.

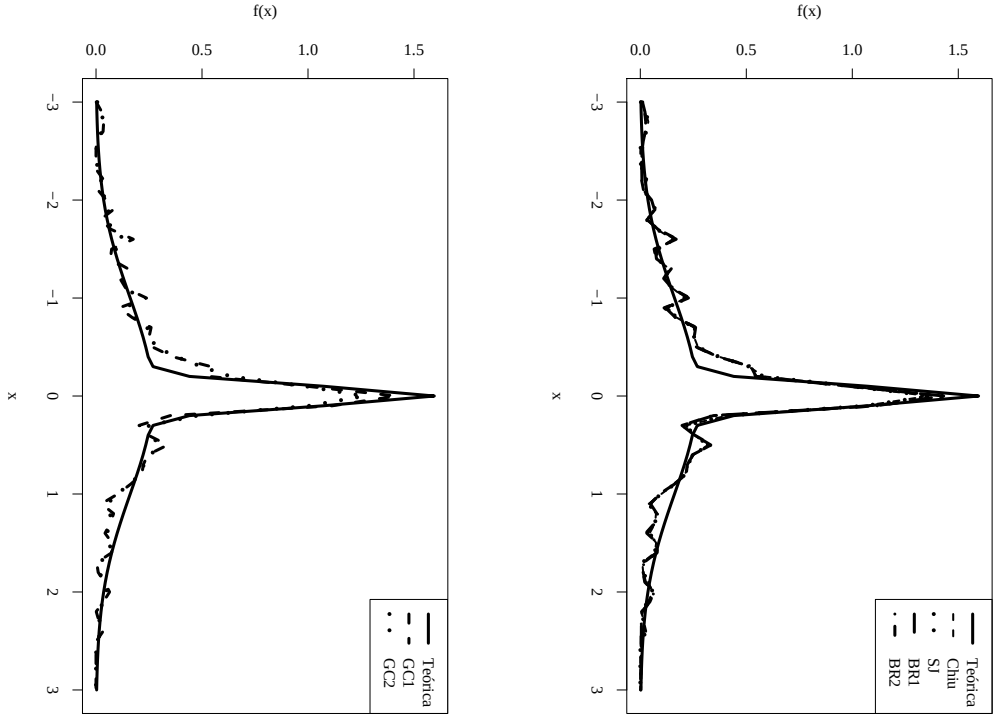

Figura A.52: Estimativa da distribuição Mistura(9) com $n = 200$.

A.14 $X \sim Mistura(10)$

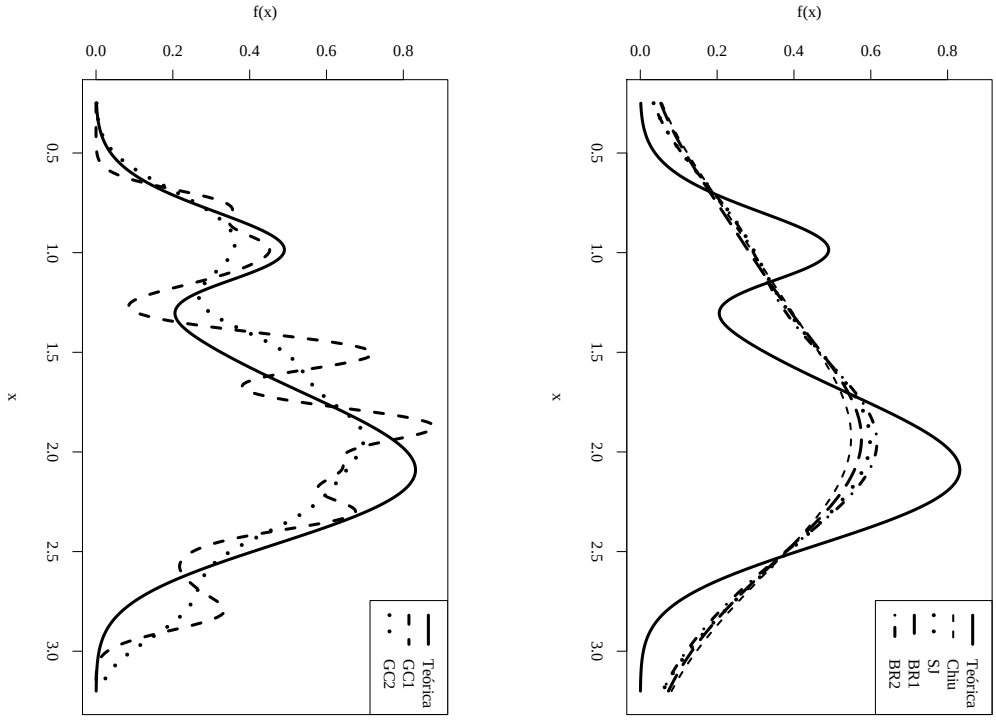

Figura A.53: Estimativa da distribuição Mistura(10) com $n = 30$.

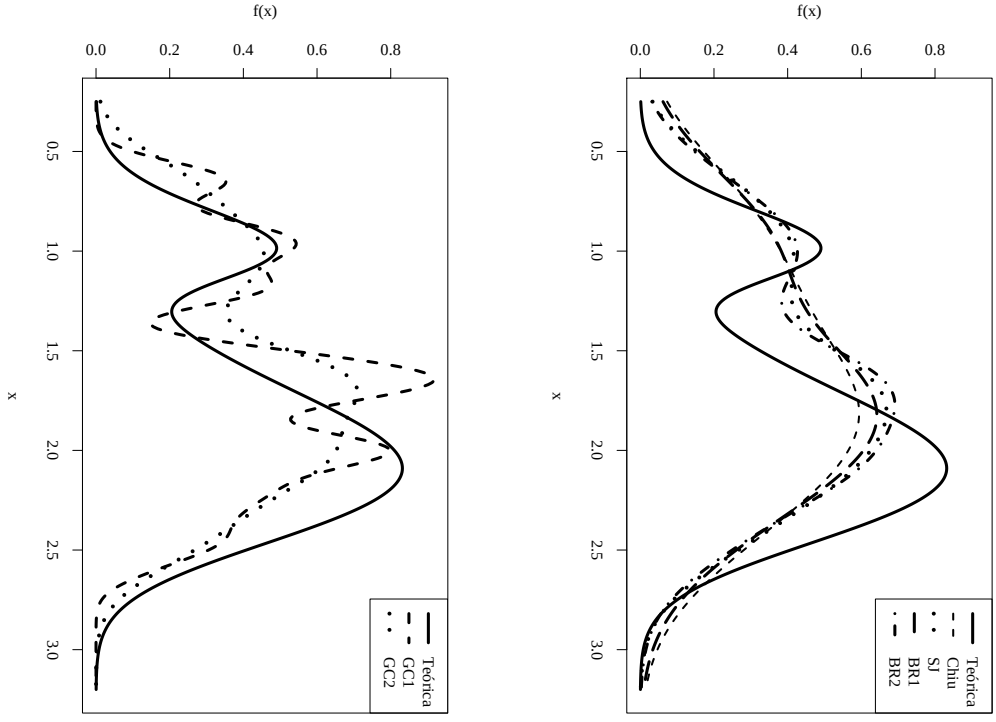

Figura A.54: Estimativa da distribuição Mistura(10) com $n = 50$.

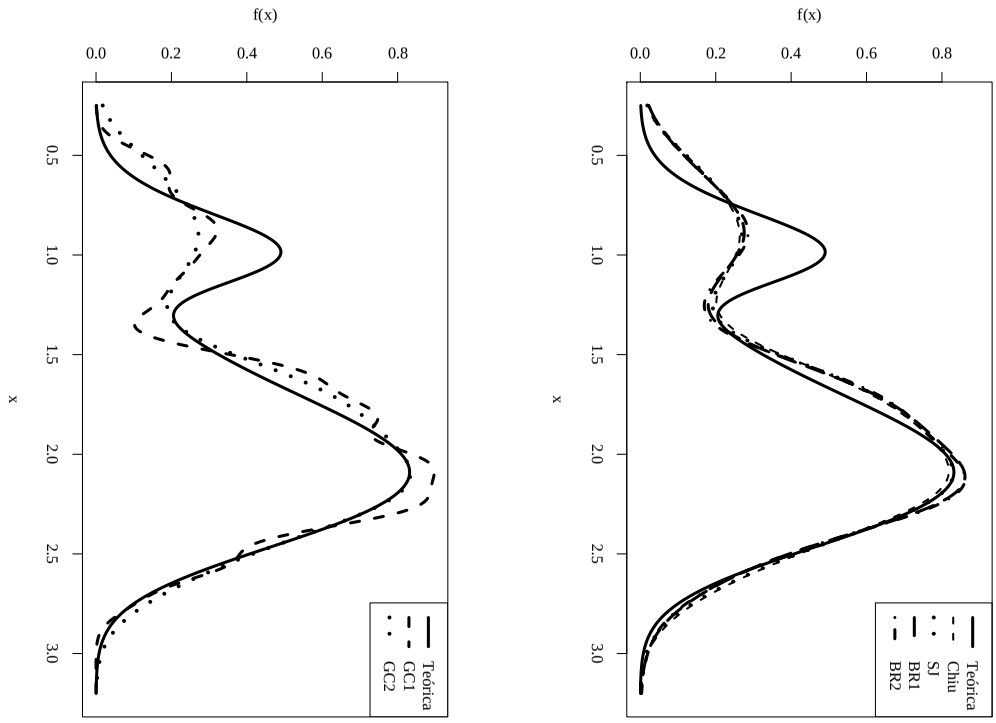

Figura A.55: Estimativa da distribuição Mistura(10) com $n = 100$.

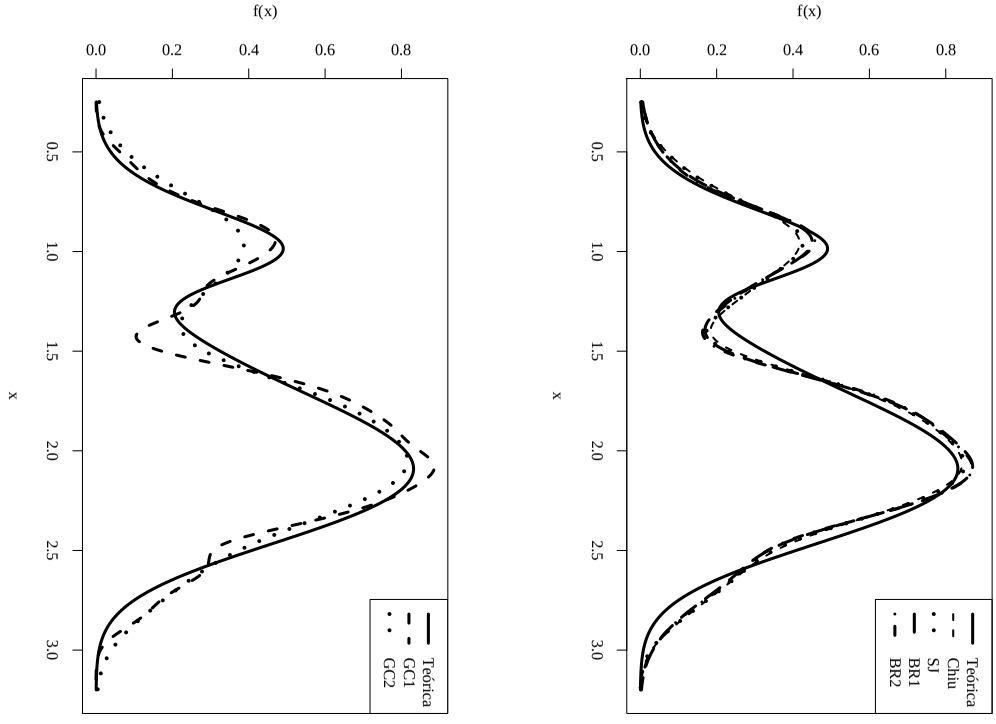

Figura A.56: Estimativa da distribuição Mistura(10) com $n = 200$.

Referências

ABRAMSON, I. On bandwidth variation in kernel estimates - a square root law. *The Annals of Statistics*, v. 10, p. 1217–1223, 1982.

BARTLETT, M. Statistical estimation of density functions. *Sankhyā Ser. A*, v. 25, p. 245–254, 1963.

BERGH, F. van den; ENGELBRECHT, A. A cooperative approach to particle swarm optimization. *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, v. 8(3), p. 225–239, 2004.

BERTRAND-RETALI, R. Convergence uniforme d'une estimateur de la densité par la method du noyau. *Revue Roumaine de Mathematiques Pures et Appliquées*, v. 23, p. 361–385, 1978.

BLAND, R. G.; GOLDFARB, D.; TODD, M. The ellipsoid method: a survey. *Operations Research*, v. 29(6), p. 1039–1091, 1981.

BOWMAN, A. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, v. 71, p. 353–360, 1984.

BOWMAN, A.; AZZALINI, A. *Applied Smoothing Techniques for Data Analysis*. [S.l.]: Oxford University Press, Oxford, 1997.

BREIMAN, L.; MEISEL, W.; PURCELL, E. Variable kernel estimates of multivariate densities. *Technometrics*, v. 19, p. 135–144, 1977.

BREWER, M. A bayesian model for local smoothing in kernel density estimation. *Computing and Statistics*, v. 10, p. 299–309, 2000.

CACOULLOS, T. Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, v. 18, p. 179–189, 1966.

CAVALCANTE, C. *Avaliação do Núcleo-Estimador na Estimação de Funções de Densidades Multivariadas*. 2004. Dissertação (Mestrado) — Universidade Federal de Minas Gerais.

CHIU, S. Bandwidth selection for kernel density estimation. *The Annals os Statistics*, v. 19, p. 1883–1905, 1991.

DAMASCENO, E. *Escolha do Parâmetro de Suavidade em Estimação Funcional*. 2000. Dissertação (Mestrado) — Universidade Federal de Minas Gerais.

DEHEUVELS, P. Estimation non paramétrique de la densité par histogrammes généralises ii. *Publications del l'Institute Statistique de l'Université Paris*, v. 22, p. 1–23, 1977.

DEVROYE, L.; GYöRFI, L. *Nonparametric Density Estimation: the L_1 View*. [S.l.]: New York: John Wiley, 1985.

DOORNIK, J. A. *Object-Oriented Matrix Programming using Ox 5th edition*. [S.l.]: London: Timberlake Consultants Press and Oxford: www.doornik.com, 2005.

DUIN, R. On the choice of smoothing parameter for parzen estimators of probability density functions. *IEEE Transactions on Computing*, C25, p. 1175–1179, 1976.

DUONG, T. *Bandwidth selectors for multivariate kernel density estimation*. 2004. Tese (Doutorado) — University of Western Australia.

DUONG, T.; HAZELTON, M. Plug-in bandwidth matrices for bivariate kernel density estimation. *Nonparametric Statistics*, v. 15(1), p. 17–30, 2003.

DUONG, T.; HAZELTON, M. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, v. 32, p. 485–506, 2005.

EPANECHNIKOV, V. Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, v. 14, p. 153–158, 1969.

FIX, E.; HODGES, J. Nonparametric discrimination: consistency properties. Report Number 4, USAF School of Aviation Medicine, Randolph Field, Texas. 1951.

FUKUNAGA, K. *Introduction to statistical pattern recognition*. [S.l.]: New York: Academic Press, 1972.

GANGOPADHYAY, A.; CHEUNG, K. Bayesian approach to the choice of smoothing parameter in kernel density estimation. *Nonparametric Statistics*, v. 14(6), p. 655–664, 2002.

GLóRIA, F. *Uma Avaliação do Desempenho de Núcleo-Estimadores no Controle de Processos Multivariados*. 2006. Dissertação (Mestrado) — Universidade Federal de Minas Gerais.

HABBEMA, J.; HERMANS, J.; BROEK, K. van den. A stepwise discriminant analysis program using density estimation. In COMPSTAT, ed. G. Bruckmann, Pyshica-Verlag, Viena, p. 101-110. 1974.

HALL, P. The influence of rounding errors on some nonparametric estimators of a density and its derivates. *Journal of Applied Mathematics*, v. 42, p. 390–399, 1982.

HALL, P.; MARRON, J. Estimation of integrated squared density derivates. *Statistics and Probability Letters*, v. 6, p. 109–115, 1987b.

HALL, P. *et al.* On optimal data-based bandwidth in kernel density estimation. *Biometrika*, v. 78, p. 263–269, 1991.

HAZELTON, M. An optimal local bandwidth selector for kernel density estimation. *Journal of Statistical Planning and Inference*, v. 77, p. 37–50, 1999.

JONES, M. Variable kernel density estimates. *Australian Journal of Statistics*, v. 32, p. 361–371, 1990.

JONES, M.; KAPPENMAN, R. On a class of kernel density estimate bandwidth selectors. *Scandinavian Journal of Statistics*, v. 19, p. 337–350, 1992.

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

JONES, M.; SHEATHER, S. Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivates. *Statistics and Probability Letters*, v. 11, p. 511–514, 1991.

KITCHENS, L. *Basic Statistics and Data Analysis*. [S.l.]: Duxbury, 2003.

LOFTSGAARDEN, D.; QUESENBERRY, C. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, v. 36, p. 1049–1051, 1965.

MAN, K.; TANG, K.; KWON, S. Genetic algorithms: Concepts and applications. *IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS*, v. 46(5), p. 519–534, 1996.

MINNOTTE, M. *A Test of Mode Existence with Applications to Multimodality*. 1992. Tese (Doutorado) — Rice University.

NADARAYA, E. On the integral mean square error of some nonparametric estimates for the density function. *Theory of Probability and Its Applications*, v. 19, p. 133–141, 1974.

PARZEN, E. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, v. 33, p. 1065–1076, 1962.

PAULINO, C.; TURKMAN, M.; MURTEIRA, B. *Estatística Bayesiana*. [S.l.]: Fundação Calouste Gulbenkian, 2003.

POLANSKY, A.; BAKER, E. Multistage plug-in bandwidth selection for kernel distribution function estimates. *Journal of the Statistical Computation and Simulation*, v. 65, p. 63–80, 2000.

ROSENBLATT, M. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, v. 27, p. 832–837, 1956.

RUDEMO, M. Empirical choice of histogram and kernel density estimators. *Scandinavian Journal of Statistics*, v. 9, p. 65–78, 1982.

SAIN, S. *Adaptive Kernel Density Estimation*. 1994. Tese (Doutorado) — Rice University.

SAIN, S.; SCOTT, D. On locally adaptive density estimation. *Journal of the American Statistical Association*, v. 91, p. 1525–1534, 1996.

SCHUCANY, W. Locally optimal window widths for kernel density estimation with large samples. *Statistics and Probability Letters*, v. 7, p. 401–405, 1989.

SCHUSTER, E.; GREGORY, G. On the nonconsistency of maximum likelihood density estimators. *Proceeddings of the Thirteenth Interface of Computer Science and Statistcs*, W. F. Eddy, p. 292–294, 1981. New York: Springer-Verlag.

SCOTT, D. *Multivariate Density Estimation: Theory, Practice and Visualization*. [S.l.]: New York: John Wiley, 1992.

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

SCOTT, D.; TERRELL, G. Biased and unbiased cross-validation in density estimation. *Jornal of the American Statistical Association*, v. 82, p. 1131–1146, 1987.

SHEATHER, S.; JONES, M. A reliable data-based bandwidth selection method for kernel density estimation. *Journal Royal Statistical Society B*, v. 53, p. 683–690, 1991.

SILVERMAN, B. Kernel density estimation using the fast fourier transform. *Applied Statistics*, v. 31, p. 93–97, 1982.

SILVERMAN, B. *Density estimation for statistics and data analysis*. [S.l.]: London: Chapman & Hall, 1986.

SIMONOFF, J. *Smoothing Methods in Statistics*. [S.l.]: New York: Springer-Verlag New York, Inc., 1996.

TERRELL, G.; SCOTT, D. Variable kernel density estimation. *The Annals of Statistics*, v. 20, p. 1236–1265, 1992.

WAND, M.; JONES, M. Multivariate plug-in bandwidth selection. *Computational Statistics*, v. 9, p. 97–116, 1994.

WAND, M.; JONES, M. *Kernel Smoothing*. [S.l.]: London: Chapman & Hall, 1995.

WAND, M.; MARRON, J.; RUPPERT, D. Transformations in density estimation. *Journal of the American Statistical Association*, v. 86, p. 343–361, 1991.

WOODROOFE, M. On choosing a delta sequence. *The Annals of Mathematical Statistics*, v. 41, p. 1665–1771, 1970.

ZHANG, X.; KING, M.; HYNDMAN, R. A bayesian approach to bandwidth selection for multivariate kernel density estimation. *Computational Statistics & Data Analysis*, v. 50, p. 3009–3031, 2006.