UNIVERSIDADE TÉCNICA DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

Bootstrap Robusto com base na Função de Influência

Maria da Conceição Esperança Amado
(Mestre)

Dissertação para Obtenção do Grau de *Doutor em Matemática*

Orientadora: Doutora Ana Maria Nobre Vilhena Nunes Pires de Melo Parente

Júri

Presidente: Reitor da Universidade Técnica de Lisboa

Vogais: Doutor Fernando Manuel Fialho Rosado
Doutora Maria Fernanda Neto Ramalhoto
Doutor João António Branco
Doutor António Pedro de Pinto de Brito Duarte Silva
Doutora Ana Maria Nobre Vilhena Nunes Pires de Melo Parente

Junho de 2003

Resumo

O objectivo deste trabalho é desenvolver procedimentos inferenciais robustos baseados em reamostragem quando o modelo gerador dos dados pertence a uma vizinhança de contaminação do modelo central Normal (univariado ou multivariado). Para isso propõe-se a utilização da função de influência (Hampel *et al.*, 1986) de um funcional de interesse na indução de um mecanismo de controlo no plano de reamostragem. Após uma breve introdução em que se procura estabelecer um enquadramento para o problema, descrevem-se dois dos conceitos mais relevantes de estatística robusta, a função de influência e o ponto de rotura, bem como alguns dos procedimentos inferenciais robustos para a estimação de parâmetros de localização e dispersão multivariados. Verifica-se que, quando o modelo gerador dos dados não é o modelo central, não é evidente o cálculo de medidas de precisão de estimadores robustos, assim como a determinação da sua distribuição, mesmo que assintótica. A utilização do método de reamostragem *bootstrap*, proposto por Efron (1979), é uma alternativa para a estimação dessas medidas. Este método apresenta, contudo, vários problemas quando aplicado a estimadores robustos, exigindo assim um procedimento que permita ultrapassar essas dificuldades. Foi com esse intuito que se desenvolveu um novo método de reamostragem que se denominou *bootstrap* robusto. Através de várias aplicações do novo método mostra-se que ele origina estimativas robustas estáveis e é computacionalmente menos dispendioso que o *bootstrap* usual de um estimador robusto. Este novo método permite efectuar inferências robustas nas mais variadas áreas da Estatística.

Palavras Chave: *Bootstrap*, Função de Influência, Estimação Robusta, Observações Atípicas, Métodos de Reamostragem, Análise Multivariada.

Abstract

The purpose of this work is to develop robust inferential procedures based on resampling when the data generating model belongs to a contamination neighborhood of the (univariate or multivariate) Normal central model. Having that in mind we considered the influence function (Hampel *et al.*, 1986) in the induction of a control mechanism in the resampling plan. After a brief introduction that provides a setting for the problem, we describe the two most important concepts of robust statistics, the influence function and the breakdown point, as well as some robust inferential procedures used in the estimation of multivariate location and scatter parameters. When the data generating model is not the central model, the precision measures of robust estimators, and their distribution, namely the asymptotic ones, are hard to obtain. The use of Efron's bootstrap (Efron, 1979) is an alternative for the estimation of these measures. However, a few problems arise when this method is applied to robust estimators, thus, a procedure to overcome those difficulties is needed. For that matter we developed a new resampling method, which was called robust *bootstrap*. Some applications of the new method show that it provides stable robust estimates and it is computationally less expensive in comparison to bootstrapping a robust estimator. This new method allows robust inferences in various statistical fields.

Key words: Bootstrap, Influence Function, Robust Estimation, Outliers, Resampling Methods, Multivariate Analysis.

Agradecimentos

Ao longo deste percurso, muitas pessoas estiveram ao meu lado, encorajando-me e incentivando-me para que a presente dissertação se tornasse uma realidade.

A minha primeira palavra de gratidão vai para a minha orientadora Professora Ana Maria Pires, que desde sempre acreditou em mim aceitando-me como aluna de Doutoramento. Foi o seu apoio, a preciosa ajuda e estímulo que constituíram um decisivo contributo para a minha formação e realização deste projecto. Agradeço-lhe, sobretudo, a sua dedicação e amizade.

Gostaria de ter uma palavra de agradecimento a todos os meus colegas da Secção de Estatística e Aplicações do Departamento de Matemática do IST, do qual tenho feito parte como Assistente, pois foram eles que asseguraram o trabalho lectivo da Secção enquanto estive de dispensa. Agradeço em especial, pela amizade e camaradagem, ao António Pacheco, Cláudia Nunes, Delfina Barbosa, Isabel Rodrigues, João Branco, Manuel Cabral Morais, Paulo Soares e Rosário Oliveira. Ao Paulo Soares não posso deixar de agradecer também as sugestões que me deu do ponto de vista informático.

Ao meu amigo de longa data, Francisco Saldanha da Gama, que sempre teve a palavra certa no momento certo, o meu muito obrigada.

Não posso também deixar de citar os meus pais e irmãos pelo seu carinho e apoio. Por fim, tenho de agradecer todo o carinho e compreensão que, ao longo desta caminhada, o meu marido Luís e o meu filho Alexandre me dedicaram.

Este trabalho foi elaborado com o apoio do Centro de Matemática e Aplicações e do projecto Praxis XXI PCEX/P/MAT/41/96.

Conceição Amado

ÍNDICE

Resumo **i**

Agradecimentos **v**

Lista de figuras **xi**

Lista de tabelas **xv**

Notação **xxv**

1 Introdução **1**

2 Estimação multivariada e inferência robusta **11**

2.1 Função de influência 13

2.2 Ponto de rotura 18

2.3 Estimação robusta de parâmetros de localização e dispersão multivariados 20

2.3.1 Estimação numa amostra multivariada 23

2.3.2 Estimação em várias amostras multivariadas 29

2.4 Inferência robusta não pontual 30

2.4.1 Testes de hipóteses robustos 34

2.4.2 Estimação regional robusta 40

2.4.3 Dificuldades em inferência robusta não pontual 42

3 Métodos de reamostragem **45**

3.1 Introdução 45

3.2 A abordagem usual 49

3.3 O *Jackknife* 51

3.3.1 A estimação *jackknife* do viés 51

3.3.2 A estimação *jackknife* da variância 53

3.3.3 *Deleted-d jackknife* 56

3.3.4 Observações finais sobre o *jackknife* 58

3.4 O *Bootstrap* 59

3.4.1 Generalidades e aspectos gerais 60

3.4.2 A estimação *bootstrap* da variância 62

3.4.3 A estimação *bootstrap* da distribuição amostral 64

3.4.4 Intervalos de confiança 66

3.4.5 Observações finais sobre o *bootstrap* 70

4 Bootstrap robusto **73**

4.1 O *bootstrap* e as observações discordantes 74

4.2 O método proposto: *bootstrap* robusto baseado na função de influência 79

4.3 Exemplos de aplicação e testes computacionais 85

4.3.1 Inferência para a localização univariada num modelo de localização e escala . 85

4.3.2 Intervalos de confiança robustos para o coeficiente de correlação linear 96

4.3.3 Estudo de eficiência de estimadores robustos multivariados em amostras finitas 112

4.3.4 Selecção robusta de variáveis em Análise Discriminante Linear 122

4.4 Generalização do *bootstrap* IF 146

4.4.1 Localização univariada 147

4.4.2 Coeficiente de correlação linear 170

5 Conclusões e trabalho futuro **181**

A Demonstrações **185**

A.1 Demonstração da proposição 2.1 185

A.2 Determinação do ponto $\mathbf{x}^o$ 187

B Resultados computacionais suplementares **189**

B.1 Resultados para a localização univariada 190

B.2 Resultados para o coeficiente de correlação linear 223

C Macros S **233**

C.1 *Bootstrap* IF para a localização univariada 233

C.2 *Bootstrap* IF para coeficiente de correlação linear 237

C.3 *Bootstrap* IF para selecção de variáveis em ADL 241

C.4 Cálculo do estimador S de localização e dispersão 245

C.5 *Bootstrap* IFG para localização univariada 251

Bibliografia **253**

LISTA DE FIGURAS

1.1 Desempenho do *bootstrap* na estimação do viés e da variância de $\bar{X}_n$. 4

1.2 Desempenho do *bootstrap* na estimação de quantis de ordem 5, 25, 50, 75 e 95 % da distribuição de $\bar{X}_n$. 5

1.3 Desempenho do *bootstrap* na estimação do viés e da variância de $\bar{X}_n$ quando existe uma observação atípica nos dados. 6

1.4 Desempenho do *bootstrap* na estimação do viés e da variância de $\bar{X}_n$ quando se aplica à amostra adulterada uma regra de rejeição robusta. 7

1.5 Desempenho do *bootstrap* na estimação do viés e da variância do estimador–M de localização, note-se que a linha horizontal representada no gráfico (b) não corresponde ao valor exacto da variância do estimador–M, continua a representar a variância de $\bar{X}_n$. 7

1.6 Diagramas de extremos-e-quartis para as estimativas *bootstrap* associadas às quatro situações simuladas. 8

1.7 Tempos de execução (em segundos) quando se utiliza o *bootstrap* com os três estimadores de localização: $\bar{X}_n$ (Md), estimador–M (M) e LTS. 9

2.1 Frequência relativa do acontecimento "*p-value*<0.05" em 1000 simulações para a hipótese $H_0 : \mu_1 = \mu_2$, variando o valor médio da segunda população ($\mu_2 = 0.0$, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5 e 4.0). 31

3.1 O método de reamostragem *bootstrap*. 61

4.1 Diagramas de extremos-e-quartis para comparação das 50000 estimativas *bootstrap* obtidas pelos vários métodos, o gráfico mais à direita representa em pormenor parte do gráfico esquerdo. 95

4.2 Exemplo de *Winsorization* de uma amostra bivariada com $n = 20$ e $\beta = 0.1$. As observações números 17 e 9 que têm associadas maior distância de Mahalanobis são deslocadas para a elipse que passa pela observação número 4, a qual corresponde a $d^2_{18:20}$, nessa deslocação a direcção original das observações é preservada. A amostra *Winsorized* é constituída então pelos pontos da figura à direita. 101

4.3 Comparação das funções peso associada à funções ψ para os três valores de c e de γ. 104

4.4 Características dos intervalos de confiança para ρ obtidos através dos métodos IFB e UB (não cobertura, média e desvio padrão) *versus* nível de contaminação. 108

4.5 Representação dos valores *jackknife-after-bootstrap* para os dados de Faculdades de Direito. 110

4.6 Representação das elipses de tolerância (a 97,5%) baseados nos estimadores RMCD para os dados de Faculdades de Direito. 110

4.7 Classificações correctas *versus* dimensionalidade (dados da qualidade do ar). 126

4.8 Frequência relativa do acontecimento "*p-value*<0.01" em 1000/100 simulações para a hipótese $H_0 : \alpha_2 = 0$, variando o vector médio da segunda população de acordo com $(2\cos\beta, 2\sin\beta)$ para $\beta = 0, 15, 30, \cdots, 90$ (em graus). A coluna da esquerda representa os valores obtidos para as distribuições NOR, NSC e NAC quando se utiliza o estimador não robusto de $\boldsymbol{\alpha}$. A coluna da direita representa os resultados para idênticas distribuições mas referentes ao estimador robusto. 133

4.9 Frequência relativa do acontecimento "*p-value*<0.05" em 1000/100 simulações para a hipótese $H_0 : \alpha_2 = 0$, variando o vector médio da segunda população de acordo com $(2\cos\beta, 2\sin\beta)$ para $\beta = 0, 15, 30, \cdots, 90$ (em graus). A coluna da esquerda representa os valores obtidos para as distribuições NOR, NSC e NAC quando se utiliza o estimador não robusto de $\boldsymbol{\alpha}$. A coluna da direita representa os resultados para idênticas distribuições mas referentes ao estimador robusto. 134

4.10 Frequência relativa do acontecimento "*p-value*<0.10" em 1000/100 simulações para a hipótese $H_0 : \alpha_2 = 0$, variando o vector médio da segunda população de acordo com $(2\cos\beta, 2\sin\beta)$ para $\beta = 0, 15, 30, \cdots, 90$ (em graus). A coluna da esquerda representa os valores obtidos para as distribuições NOR, NSC e NAC quando se utiliza o estimador não robusto de $\boldsymbol{\alpha}$. A coluna da direita representa os resultados para idênticas distribuições mas referentes ao estimador robusto. 135

4.11 % de inclusão de cada subconjunto nas 1000 simulações quando se considerou o vector valor médio da segunda população $\boldsymbol{\mu}_2 = (2, 0, 0)$ e a regra discriminante foi obtida por um estimador clássico. 140

4.12 % de inclusão de cada subconjunto nas 1000 simulações quando se considerou o vector valor médio da segunda população $\boldsymbol{\mu}_2 = (\sqrt{2}, \sqrt{2}, 0)$ e a regra discriminante foi obtida por um estimador clássico. 141

4.13 % de inclusão de cada subconjunto nas 1000 simulações quando se considerou o vector valor médio da segunda população $\boldsymbol{\mu}_2 = (2/\sqrt{3}, 2/\sqrt{3}, 2/\sqrt{3})$ e a regra discriminante foi obtida por um estimador clássico. 141

4.14 % de inclusão de cada subconjunto nas 100 simulações quando se considerou o vector valor médio da segunda população $\boldsymbol{\mu}_2 = (2, 0, 0)$ e a regra discriminante foi obtida por um estimador robusto. 142

4.15 % de inclusão de cada subconjunto nas 100 simulações quando se considerou o vector valor médio da segunda população $\boldsymbol{\mu}_2 = (\sqrt{2}, \sqrt{2}, 0)$ e a regra discriminante foi obtida por um estimador robusto. 142

4.16 % de inclusão de cada subconjunto nas 100 simulações quando se considerou o vector valor médio da segunda população $\boldsymbol{\mu}_2 = (2/\sqrt{3}, 2/\sqrt{3}, 2/\sqrt{3})$ e a regra discriminante foi obtida por um estimador robusto. 143

4.17 Comparação entre os vários métodos *bootstrap* na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$ para os dados da tensão sanguínea, excepto para R^U_{BOOT} onde a distribuição em causa é a de $\bar{X}_n - \mu$ (ou $\mu(F_n) - \mu(F)$). 167

4.18 Comparação entre os vários métodos *bootstrap* na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$ para os dados das Faculdades de Direito, excepto para R^U_{BOOT} onde a distribuição em causa é a de $R - \rho$ (ou $\rho(F_n) - \rho(F)$). 178

B.1 Desempenho dos vários estimadores R_{Boot} na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$ em termos dos indicadores globais, SQ, DA e KS, para as distribuições NOR, NSC–0.25 e NAC–0.1. 190

B.2 Desempenho dos vários estimadores R_{Boot} na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$ em termos dos indicadores globais, SQ, DA e KS, para as distribuições M4–0.1, M4–0.2 e T1. 191

B.3 Desempenho dos vários estimadores R_{Boot} na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$ em termos dos indicadores globais, SQ, DA e KS, para as distribuições LOGN1 e LOGN2. 192

B.4 Desempenho dos vários estimadores R_{Boot} na estimação da distribuição $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$ em termos dos indicadores globais, SQ, DA e KS, para as distribuições NOR, NSC–0.25 e NAC–0.1. 223

LISTA DE TABELAS

1.1 Dados gerados a partir de uma distribuição exponencial de valor esperado 10. 4

2.1 Estimativas relativas aos comprimentos e probabilidades de não cobertura de intervalos de confiança bilaterais para o valor médio de uma população F, com base em amostras de dimensão $n = 20$. 33

2.2 Estimativas relativas aos comprimentos e probabilidades de não cobertura de intervalos de confiança bilaterais para o valor médio de uma população F, com base em amostras de dimensão $n = 50$. 33

4.1 Probabilidades estimadas de não cobertura, assim como médias e desvios padrões dos comprimentos dos intervalos obtidos quando se aplica às reamostras *bootstrap* o estimador não robusto, $\bar{X}_n$. 92

4.2 Probabilidades estimadas de não cobertura, assim como médias e desvios padrões dos comprimentos dos intervalos obtidos quando se aplica às reamostras *bootstrap* o estimador robusto, LTS. 93

4.3 Intervalos de confiança a 95% e 99% de confiança para o valor médio, μ, da tensão sanguínea de um indivíduo. Os valores assinalados com * são os apresentados em Salibian-Barrera (2000, pág. 96). 94

4.4 Resultados da simulação para diferentes constantes de afinação para definição dos pesos do *bootstrap* IF quando se usa o estimador R. 103

4.5 Resultados do estudo de Monte Carlo para o coeficiente de correlação amostral de Pearson, R. 105

4.6 Resultados do estudo de Monte Carlo para os estimadores MCD. 105

4.7 Desempenho do método IFB para dados não contaminados e com crescente dimensão amostral. 107

4.8 Duas medidas, LSAT e GPA, do desempenho académico dos alunos universitários em 15 Faculdades de Direito Americanas, no ano de 1973. 110

4.9 Limites aproximados a 90% de confiança e estimativas pontuais para o coeficiente de correlação entre as variáveis LSTA e GPA, por diversos métodos. 111

4.10 Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis e $n_1 = n_2 = 50$, para a distribuição NOR. 117

4.11 Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis e $n_1 = n_2 = 50$, para a distribuição NSC. 117

4.12 Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis e $n_1 = n_2 = 50$, para a distribuição NAC. 118

4.13 Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis e $n_1 = n_2 = 50$, para a distribuição T3. 118

4.14 Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis, $n_1 = 100$ e $n_2 = 300$, para a distribuição NOR. 119

4.15 Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis, $n_1 = 100$ e $n_2 = 300$, para a distribuição NSC. 120

4.16 Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis, $n_1 = 100$ e $n_2 = 300$, para a distribuição NAC. 120

4.17 Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis, $n_1 = 100$ e $n_2 = 300$, para a distribuição T3. 121

4.18 Comparação das várias taxas de erro. 127

4.19 Associação dos *p-values* a cada uma das hipóteses consideradas. 137

4.20 Descrição do esquema de selecção de cada subconjunto para um dado valor de γ_*. . . 138

4.21 Frequência relativa do acontecimento "*p-value*$<\gamma_*$" em 1000 simulações para as hipóteses $H_0 : \alpha_2 = 0$, $H_0 : \alpha_3 = 0$, $H_0 : \alpha_2 = \alpha_3 = 0$ quando se utiliza o estimador clássico e para cada valor fixo de $\gamma_* = 0.1, 0.05, 0.01$. Caso a), $\boldsymbol{\mu} = (2, 0, 0)$. 144

4.22 Frequência relativa do acontecimento "*p-value*$<\gamma_*$" em 1000 simulações para as hipóteses $H_0 : \alpha_3 = 0$ quando se utiliza o estimador clássico e para cada valor fixo de $\gamma_* = 0.1, 0.05, 0.01$. Caso b), $\boldsymbol{\mu} = (\sqrt{2}, \sqrt{2}, 0)$. 144

4.23 Frequência relativa do acontecimento "*p-value*$<\gamma_*$" em 100 simulações para as hipóteses $H_0 : \alpha_2 = 0$, $H_0 : \alpha_3 = 0$, $H_0 : \alpha_{23} = 0$ quando se utiliza o estimador robusto e para cada valor fixo de $\gamma_* = 0.1, 0.05, 0.01$. Caso a), $\boldsymbol{\mu} = (2, 0, 0)$. 145

4.24 Frequência relativa do acontecimento "*p-value*$<\gamma_*$" em 100 simulações para as hipóteses $H_0 : \alpha_3 = 0$ quando se utiliza o estimador robusto e para cada valor fixo de $\gamma_* = 0.1, 0.05, 0.01$.Caso b), $\boldsymbol{\mu} = (\sqrt{2}, \sqrt{2}, 0)$. 145

4.25 Alguns dos intervenientes no modelo original e no modelo ponderado. 148

4.26 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal padrão e $c = 1.5$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC. 155

4.27 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal padrão e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC. 156

4.28 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal simetricamente contaminada e $c = 1.5$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC. 157

4.29 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal simetricamente contaminada e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC. 158

4.30 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4-0.1 e $c = 1.5$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC. 159

4.31 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.1 e $c = d = 2$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC. 160

4.32 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição T1 e $c = 1.5$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC. 161

4.33 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição T1 e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC. 162

4.34 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $c = 1.5$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC. 163

4.35 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC. 164

4.36 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN1 e $c = 1.5$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC. 165

4.37 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN1 e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC. 166

4.38 Percentis das várias distribuições *bootstrap* de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$ para os dados da tensão sanguínea, com $c = 1.5$. 167

4.39 Intervalos de confiança a 95% para o valor médio da tensão sanguínea. 168

4.40 Cobertura e (comprimento médio) de intervalos de confiança para a localização univariada a 95%. A negrito encontram-se os valores que apresentam diferenças estatisticamente significativas para 0.95. 169

4.41 Cobertura e (comprimento médio) de intervalos de confiança para a localização univariada a 99%. A negrito encontram-se os valores que apresentam diferenças estatisticamente significativas para 0.99. 169

4.42 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NOR e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 172

4.43 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NOR e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 173

4.44 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NSC–0.25 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 174

4.45 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NSC–0.25 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 175

4.46 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 176

4.47 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 177

4.48 Percentis das várias distribuições *bootstrap* $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$ para os dados das Faculdades de Direito, com $c = 5$. 178

4.49 Intervalos de confiança a 90% para o coeficiente de correlação entre as variáveis LSTA e GPA. 179

B.1 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal padrão e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 193

B.2 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal padrão e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 194

B.3 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal padrão e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 195

B.4 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal simetricamente contaminada e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. . . 196

B.5 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal simetricamente contaminada e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. . . 197

B.6 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal simetricamente contaminada e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. . . 198

B.7 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.1 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 199

B.8 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4-0.1 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 200

B.9 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4-0.1 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 201

B.10 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.2 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 202

B.11 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.2 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 203

B.12 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.2 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 204

B.13 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.2 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 205

B.14 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.2 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 206

B.15 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4-0.2 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 207

B.16 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição T1 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 208

B.17 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição T1 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 209

B.18 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição T1 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 210

B.19 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 211

B.20 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 212

B.21 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 213

B.22 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN1 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 214

B.23 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN1 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 215

B.24 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN1 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 216

B.25 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN2 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 217

B.26 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN2 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 218

B.27 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN2 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 219

B.28 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN2 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 220

B.29 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN2 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 221

B.30 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN2 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 222

B.31 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NOR e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 224

B.32 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NOR e $c = 5$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 225

B.33 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NOR e $c = 10$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 226

B.34 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NSC–0.25 e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 227

B.35 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NSC–0.25 e $c = 5$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 228

B.36 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NSC–0.25 e $c = 10$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 229

B.37 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 230

B.38 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $c = 5$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 231

B.39 Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $c = 10$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC. 232

Notação

$\mathbf{A}^T$, $\det(\mathbf{A})$, $\mathbf{A}^{-1}$:	a transposta, o determinante e a inversa de uma matriz $\mathbf{A}$.
$I\!R$ e $I\!R^m$:	o conjunto dos números reais e o espaço Euclidiano de dimensão m.
$I\!N$:	o conjunto dos números inteiros.
$\|\mathbf{x}\|$:	norma Euclidiana de um vector $\mathbf{x} \in I\!R^m$, $\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x}$.
v.a.:	variável aleatória.
i.i.d:	independentes e identicamente distribuídas.
$\Delta_{\mathbf{x}}$:	função distribuição de uma v.a. degenerada em $\mathbf{x} \in I\!R^m$.
H_0 e H_1:	hipótese nula e alternativa, respectivamente.
F_n:	função de distribuição empírica da amostra $\mathbf{x}_n$ que afecta massa n^{-1} a cada x_i.
$N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:	Distribuição Normal m-multivariada com vector médio $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$.
$N(\mu, \sigma^2)$:	Distribuição Normal univariada com valor médio μ e variância σ^2.
$\Phi(x)$:	função de distribuição da distribuição Normal padrão no ponto x.
$[i]$:	representa o maior inteiro menor ou igual a i.
$I_A(s)$:	função indicatriz relativa ao elemento s no conjunto A, isto é, $$I_A(s) = \begin{cases} 1, & se\ s \in A \\ 0, & se\ s \notin A \end{cases}.$$
$\rightarrow_{q.c.}$:	convergência quase certa.
$\rightarrow_p$:	convergência em probabilidade.
$\rightarrow_d$:	convergência em distribuição.
$a_n = O(b_n)$:	se $\|a_n\| \leq k\|b_n\|$ para todo o n e constante k (com $\{a_n\}$ e $\{b_n\}$ duas sequências de números).
$a_n = o(b_n)$:	se $a_n/b_n \rightarrow 0$ quando $n \rightarrow \infty$.

Para além desta notação observe-se que: as matrizes e vectores são representados a negrito, as primeiras por letras maiúsculas e os segundos por minúsculas. A convenção usual de utilizar letras maiúsculas para variáveis aleatórias e estimadores e minúsculas para concretizações e estimativas torna-se difícil de manter, especialmente em tópicos de Análise Multivariada. Espera-se que a distinção fique clara a partir do contexto.

O símbolo ■ representa o final de uma Demonstração, de um Exemplo, ou de uma Observação.

"Imagination is the beginning of creation. You imagine what you desire, you will what you imagine and at last you create what you will."

George Bernard Shaw (1856 - 1950)

CAPÍTULO 1

Introdução

A grande evolução de meios computacionais iniciada na década de oitenta e a que se continua ainda hoje a assistir, tornou possível quer o desenvolvimento dos chamados métodos computacionalmente intensivos, quer a análise de conjuntos de dados cada vez mais volumosos e complexos. Bancos, seguradoras e organismos oficiais constituem exemplos de entidades que processam um elevado número de variáveis e observações.

Simultaneamente surgem novas questões, nomeadamente as colocadas pela imunologia ou pelos testes sobre o desempenho de novas tecnologias, cujas respostas requerem a aplicação de métodos estatísticos mas cujo delineamento de amostragem possui custos muito elevados não permitindo recolher um grande número de observações. Nestes casos não é possível utilizar muitos dos procedimentos tradicionais baseados em teoria assintótica, o que tem conduzindo a um grande interesse pelos métodos de reamostragem.

Para além da possibilidade de análise de conjuntos de dados de pequena dimensão outra importante aplicação dos métodos de reamostragem encontra-se na resolução de problemas estatísticos onde não é possível efectuar suposições sobre o modelo gerador dos dados, inserindo-se assim na área da estatística não paramétrica ou semi-paramétrica. Note-se ainda a utilidade destes procedimentos em amostras de grandes dimensões: quando se obtém uma amostra de elevada dimensão ela pode ser dividida em subconjuntos e aplicando técnicas de reamostragem poder-se-á avaliar a estabilidade do procedimento de estimação.

Os métodos de reamostragem podem dividir-se em quatro tipos: permutação (Fisher, 1935), *jackknife* (Quenouille, 1949, 1956), validação cruzada (Stone, 1974) e *bootstrap* (Efron, 1979). Actualmente o *bootstrap* é, talvez, o método de reamostragem com mais sucesso na análise estatística, devido não só às suas inúmeras aplicações possíveis mas também às boas propriedades teóricas que

ao longo destas três décadas têm vindo a ser demonstradas. O *bootstrap* de Efron simplesmente replica um elevado número de vezes a amostra de trabalho criando uma "população" virtual. Em seguida são seleccionadas amostras dessa "população" construindo desse modo uma distribuição amostral empírica.

Encontram-se aplicações do *bootstrap* nas mais variadas áreas, entre elas podem-se citar:

- ambiente atmosférico (Hanna, 1989);
- engenharia biomédica (Banga e Ghorbel, 1993; Haynor e Woods, 1989; Maitra, 1997);
- finanças (Kuhle e Moorehead, 1990) e engenharia financeira, (Bhar e Chiarella, 1996; Ankembrand e Tomassini, 1996);
- geofísica (Fisher e Hall, 1989; Constable e Tauxe, 1990; Fisher e Hall, 1991; Tauxe *et al.*, 1991);
- metrologia (Ciarlini, 1997; Cox *et al.*, 1997);
- processamento de sinais (Krolik *et al.*, 1991; Maiwald e Böhme, 1994; Reid *et al.*, 1996);
- sistemas de potência (Herman e Heunis, 1999);
- visão computacional (Lange *et al.*, 1998; Katatani e Ohta, 1998).

Assiste-se, porém, a algum cepticismo relativamente à utilização destes métodos. A esse propósito Peterson (1991) comenta o seguinte:"*how to infer the "truth" from a sample of data that may be incomplete or drawn from an ill-defined population*". Por outro lado, num dos artigos iniciais, Diaconis e Efron (1983) argumentam que os métodos de reamostragem libertam os investigadores de duas limitações da estatística tradicional: "*the assumption that the data conform to a bell-shaped curve and the need to focus on statistical measures whose theoretical properties can be analyzed mathematically*".

No entanto a maior parte das razões de cepticismo evocadas por vários autores relativamente à utilização de métodos de reamostragem (entre eles Peterson, 1991 e Young, 1994) não são, de uma forma geral, características específicas destes métodos mas sim da maior parte dos procedimentos tradicionais em Estatística.

Existem, contudo, questões tais como a possível existência de observações atípicas na amostra de trabalho, que apesar de comuns à maior parte dos métodos estatísticos clássicos podem ter consequências especialmente gravosas quando se utilizam métodos de reamostragem, em particular o *bootstrap* de Efron. De facto, a referida replicação das amostras é realizada com reposição o que pode aumentar a percentagem de observações atípicas inicial. Uma alternativa para lidar com essa situação será a utilização de procedimentos robustos. Estes últimos permitem acomodar essas observações originando resultados perto do óptimo quando se verificam as suposições sobre o modelo central considerado mas não sofrem grande alteração quando se verificam alguns afastamentos a

esse modelo. Todavia, a utilização simultânea do *bootstrap* e de métodos robustos pode conduzir a outros problemas:

(i) instabilidade numérica, devido à diminuição do ponto de rotura do procedimento conjunto (*bootstrap* **mais** método robusto);

(ii) aumento do tempo de cálculo, devido à necessidade de resolução dum elevado número de problemas de optimização não convexa para obtenção das estimativas robustas.

É neste contexto que se insere este projecto cujo objectivo principal consiste em desenvolver procedimentos de reamostragem robustos que permitam ultrapassar estes dois problemas. Tendo em vista essa finalidade assim como a motivação dada pelos problemas do estudo realizado em Amado e Pires (1998), desenvolveu-se um método de reamostragem que se denominou *bootstrap* robusto. A ideia em que assenta este método é a reamostragem ponderada das observações que constituem a amostra de trabalho. A ponderação tem como propósito controlar observações que se afastam de uma determinada estrutura principal (modelo central) e é conseguida através da utilização da função de influência do funcional de interesse. A função de influência (Hampel, 1968) é uma importante ferramenta na estatística robusta e permite "medir" no funcional de interesse o "viés" originado por uma observação atípica.

Para esclarecer as questões abordadas, reamostragem *bootstrap*, robustez e tempo de cálculo, optou-se por apresentar um exemplo simples que pretende ilustrar essas questões mesmo antes de qualquer formalização do problema. De modo a permitir alguma comparação teórica optou-se por aplicar o método de reamostragem *bootstrap* a um problema de estimação da localização univariada com base num conjunto de dados artificiais gerados a partir de uma distribuição exponencial.

Exemplo 1.1 Dados exponenciais. Geraram-se $n = 20$ observações provenientes de um modelo com a seguinte função de densidade $f(x) = \lambda \exp(-\lambda x)$, $x \geq 0$, para $\lambda = 1/10$ (estas observações encontram-se na Tabela 1.1). Suponha-se que o interesse recai nas propriedades, viés e variância, do estimador média aritmética, $\bar{X}_n$, que tem como objectivo estimar o parâmetro de localização, $1/\lambda = 10$, que se vai assumir como sendo desconhecido. Para além disso, vai supor-se que não são possíveis cálculos teóricos nem aproximações. Nestas condições uma alternativa é utilizar o método *bootstrap*, o qual requer a geração com reposição de todas as amostras possíveis a partir da amostra original de dimensão $n = 20$. Todas essas amostras pertencem ao simplex n-dimensional $C_n = \{(i_1, i_2, \cdots, i_n), i_j \in I\!N, \sum i_j = n\}$, que no total corresponde a $\#C_n = \binom{2n-1}{n}$ amostras distintas. Utilizando a fórmula de Stirling obtém-se a aproximação $\#C_n \approx (n\pi)^{-1/2} 2^{2n-1}$ a qual resulta em $\#C_{20} \approx 69355338660$ para este exemplo. Sendo este número proibitivo de gerar, aplica-se uma técnica de simulação cujo nome se deve aos famosos casinos de Monte Carlo, o método de simulação Monte Carlo. Desta forma obtêm-se aproximações dos denominados estimadores *bootstrap*, estas aproximações são também, em geral, designadas por estimadores *bootstrap*. Para se verificar qual o comportamento dessas aproximações na estimação do viés e variância de $\bar{X}_n$, fazem-se quatro repetições do método de Monte Carlo para cada número fixo de réplicas,

Tabela 1.1: Dados gerados a partir de uma distribuição exponencial de valor esperado 10.

18.1518613	8.5621992	11.5023341	4.3625254	24.7947288	24.9311525
15.9424253	0.2227789	8.2393210	13.0200912	0.3219198	2.3834135
38.7998226	0.9650346	1.0048651	5.8146298	9.8587106	2.2537651
9.5367390	2.0056384				

$B = 10, 25, 50, 100, 250, 500, 600, 800, 1000$. Os gráficos da Figura 1.1 apresentam os valores empíricos do viés e da variância de $\bar{X}_n^*$ (análogo a $\bar{X}_n$ mas para a amostra *bootstrap* "*") obtidos após quatro repetições da simulação. Cada linha mostra como as estimativas do viés e da variância vão sendo alteradas quando o número de réplicas vai aumentando. As linhas horizontais representam os valores exactos.

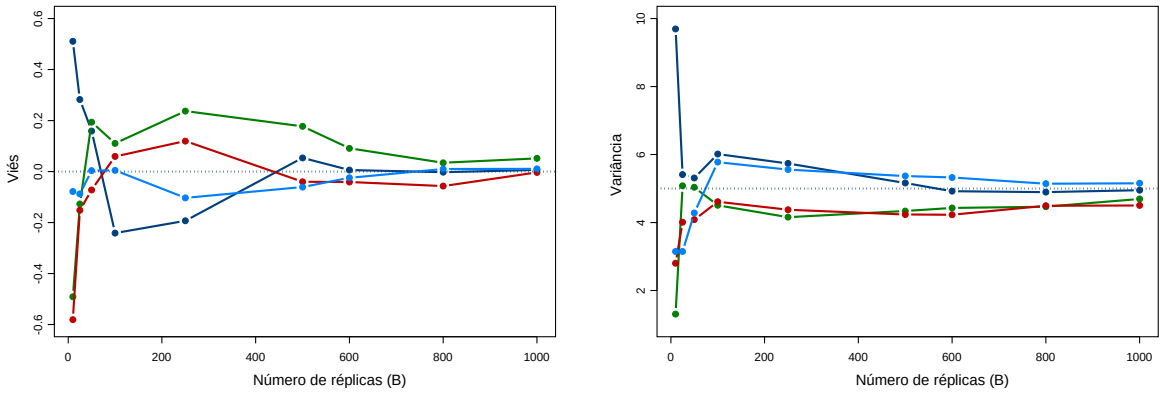


Figura 1.1: Desempenho do *bootstrap* na estimação do viés e da variância de $\bar{X}_n$.

Pela observação dos gráficos é patente a redução da variabilidade à medida que o número de réplicas *bootstrap*, B, aumenta assim como a "convergência" dos valores simulados para os valores exactos ($Viés\left(\bar{X}_n\right) = 0$ e $Var(\bar{X}_n) = 10^2/20$). Para além disso verifica-se que parece ser suficiente, para se obter uma boa aproximação, valores de B na ordem dos 500, 600.

Outra importante aplicação do método de reamostragem *bootstrap* reside na estimação da designada distribuição amostral do estimador de interesse. Através dela também se obtêm as medidas de centralidade e de precisão já referidas. A Figura 1.2 representa os quantis empíricos de $\bar{X}_n^*$ obtidos também após quatro simulações para cada valor de B. As linhas horizontais representam os valores dos quantis exactos (observe-se que nesta situação a distribuição de $\bar{X}_n$ possui uma forma conhecida).

Nesta situação também os quantis empíricos *bootstrap* "convergem" para os valores exactos à medida que o número de réplicas aumenta. Estes quantis poderiam ser utilizados para construir intervalos

de confiança ou testes de hipóteses para $1/\lambda$.

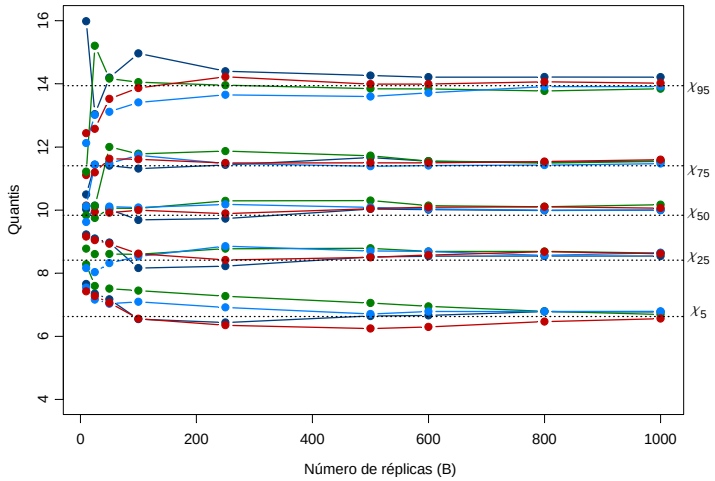


Figura 1.2: Desempenho do *bootstrap* na estimação de quantis de ordem 5, 25, 50, 75 e 95 % da distribuição de $\bar{X}_n$.

Considere-se agora a seguinte alteração aos dados, o primeiro valor passou de 18.1518613 para 181.518613, isto é, supõe-se a existência de uma observação atípica na amostra devido a erro grosseiro. Repetiram-se as simulações e obtiveram-se os resultados compilados na Figura 1.3.

O primeiro facto a salientar quando se observa a Figura 1.3 é relativo à estimação do viés, onde parece que a "convergência" se mantém, apesar de cada uma das linhas apresentar muito maior variabilidade que as análogas da Figura 1.1. Mas é nos valores empíricos da variância que se observa o efeito nefasto da presença dessa observação atípica, verificando-se um afastamento bastante acentuado relativamente ao "verdadeiro" valor.

Uma vez que este conjunto de dados apresenta uma observação atípica pode pensar-se em duas estratégias: (i) rejeição da observação com posterior aplicação do método clássico e (ii) utilização de um estimador robusto de localização. Relativamente à primeira estratégia aplica-se uma regra de detecção robusta (ver detalhes e discussão no Capítulo 4) a qual detecta a primeira observação como atípica. Quanto à segunda optou-se por utilizar um estimador–M de localização com a função biponderada de Tukey (ver Capítulo 2).

As Figuras 1.4 e 1.5 representam os resultados obtidos respectivamente por cada uma dessas estratégias. Em primeiro lugar deve observar-se que, apesar do objectivo comum ser a estimação do parâmetro de localização os gráficos das Figuras 1.4 e 1.5 não são comparáveis pelo facto dos segundos representarem a estimação do viés e da variância do estimador–M de localização, enquanto que os primeiros continuam a representar a estimação do viés e variância de $\bar{X}_n$. Pode, no entanto, verificar-se que o comportamento do procedimento em que é removida a observação atípica da

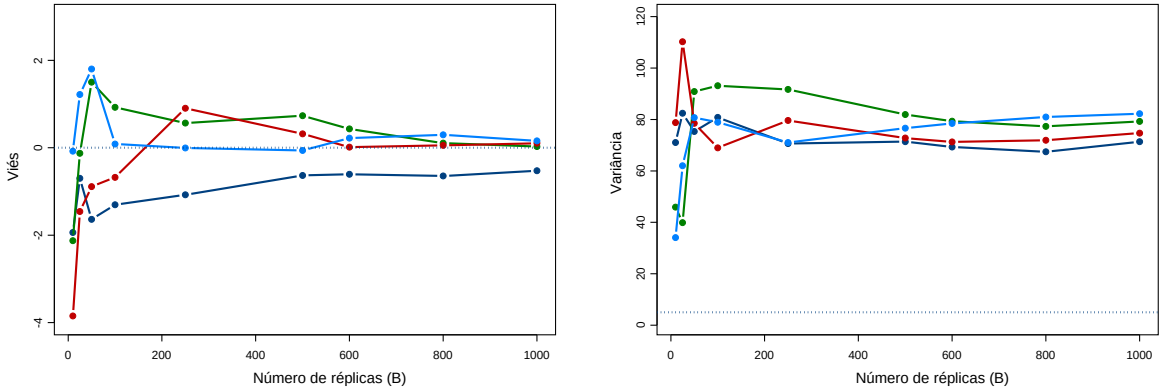


Figura 1.3: Desempenho do *bootstrap* na estimação do viés e da variância de $\bar{X}_n$ quando existe uma observação atípica nos dados.

amostra não é idêntico ao observado na Figura 1.1, por um lado devido à diminuição da dimensão da amostra de trabalho e por outro devido a um factor não controlável que é a variabilidade introduzida pelo passo de "limpeza". É também importante referir que, pelo facto da distribuição exacta de X não ser simétrica, o estimador–M não está a estimar o valor esperado $1/\lambda$ mas sim um outro parâmetro de localização (do tipo média populacional aparada).

Um defensor de regras de rejeição dirá que essa estratégia é a melhor, quando observamos os gráficos das Figuras 1.4 e 1.6 mas a questão essencial que se coloca à utilização desta estratégia é, por um lado, a não inclusão da variabilidade no procedimento de estimação devido a essa "limpeza" e por outro lado o efeito que determinadas observações atípicas têm de se "disfarçar" de "boas" observações quando de facto não o são, o que em geral não é detectado. Note-se ainda que a utilização de uma regra robusta de rejeição com posterior estimação clássica é ainda um procedimento robusto. Quando se utilizou simultaneamente o *bootstrap* e o estimador robusto não parece ter havido problemas de instabilidade numérica devido à presença da observação atípica mais que uma vez numa ou mais amostras *bootstrap*, este facto é justificado pelas boas propriedades de "resistência" que o estimador–M tem.

Suponha-se agora que o objectivo é a determinação de intervalos de confiança para o parâmetro de localização $1/\lambda$. Como se ignora *a priori* se observações atípicas estão presentes ou não na amostra de trabalho opta-se por utilizar para além do estimador clássico, $\bar{X}_n$, dois estimadores robustos de localização, o já referido estimador–M e o estimador denominado na literatura anglo-saxónica por *Least Trimmed Squares*, *LTS*, (Rousseeuw, 1984, 1985). A Figura 1.7 representa o tempo de execução mediano de quatro repetições do procedimento *bootstrap* para vários valores de B quando o objectivo é a determinação da distribuição amostral *bootstrap* de cada um dos três estimadores referidos. Esses cálculos foram realizados num *Pentium II* a 550 *MHz* com 128 *Mb* de memória *Ram*. As linhas horizontais representam os máximos dos tempos de execução medianos, ou seja o

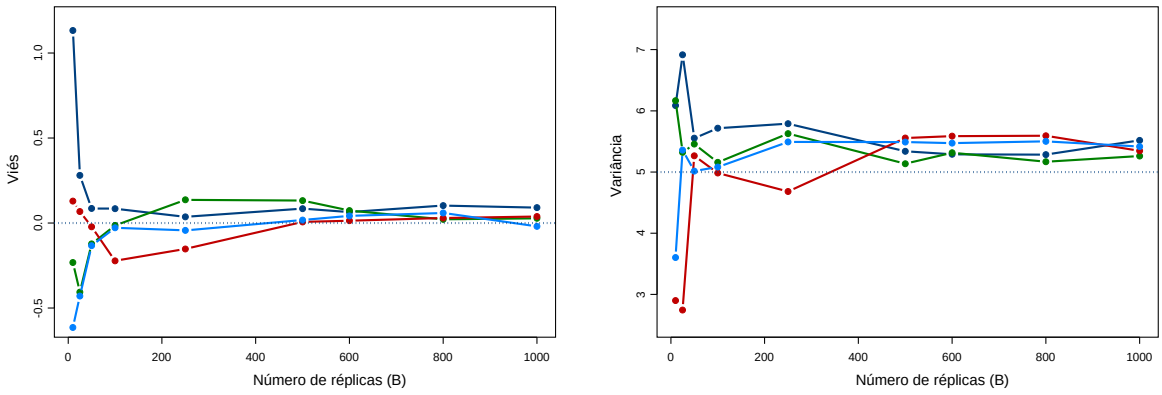
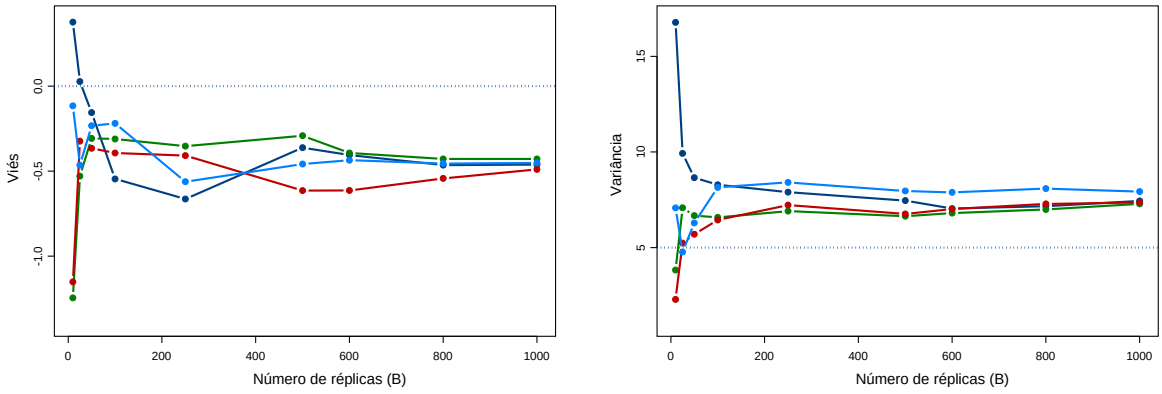


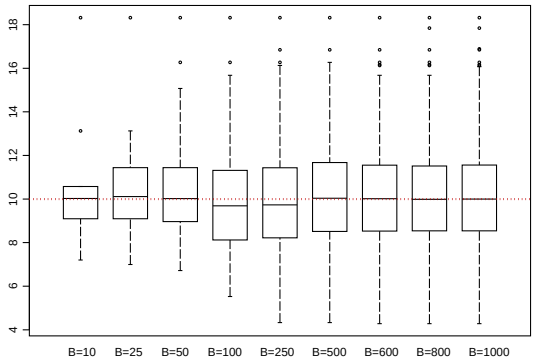
Figura 1.4: Desempenho do *bootstrap* na estimação do viés e da variância de $\bar{X}_n$ quando se aplica à amostra adulterada uma regra de rejeição robusta.



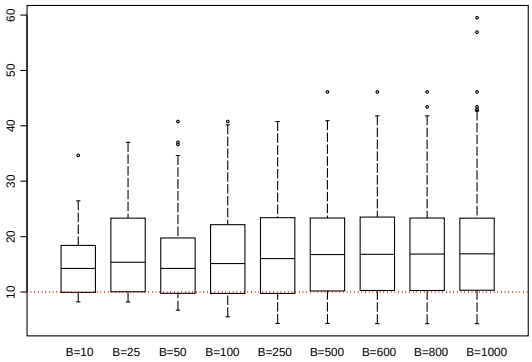
(a) Estimativas *bootstrap* do viés.

(b) Estimativas *bootstrap* da variância.

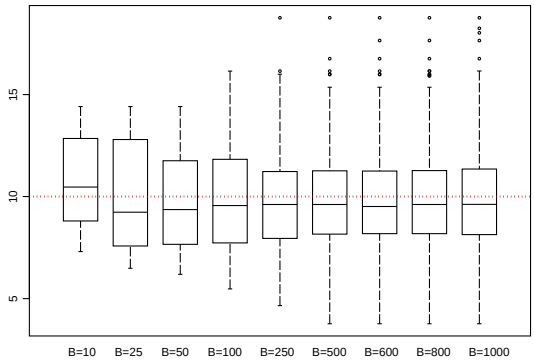
Figura 1.5: Desempenho do *bootstrap* na estimação do viés e da variância do estimador–M de localização, note-se que a linha horizontal representada no gráfico (b) não corresponde ao valor exacto da variância do estimador–M, continua a representar a variância de $\bar{X}_n$.



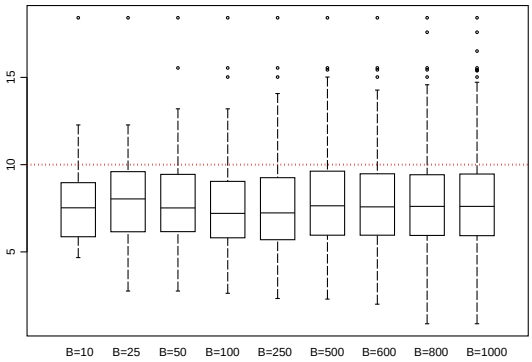
(a) Estimativas *bootstrap* de $\bar{X}_n$.



(b) Estimativas *bootstrap* de $\bar{X}_n$ quando a amostra possui uma observação atípica.



(c) Estimativas *bootstrap* de $\bar{X}_n$ quando se utilizou uma regra de rejeição robusta e se removeu essa observação.



(d) Estimativas *bootstrap* do estimador–M de localização.

Figura 1.6: Diagramas de extremos-e-quartis para as estimativas *bootstrap* associadas às quatro situações simuladas.

tempo de execução mediano associado à simulação com $B = 1000$ réplicas *bootstrap*. Se por um lado a utilização de estimadores robustos acautela a possibilidade de uma má estimação devido à possível presença de observações atípicas, por outro lado, e observando a Figura 1.7, verifica-se que essa precaução conduz a um elevado dispêndio de tempo de cálculo.

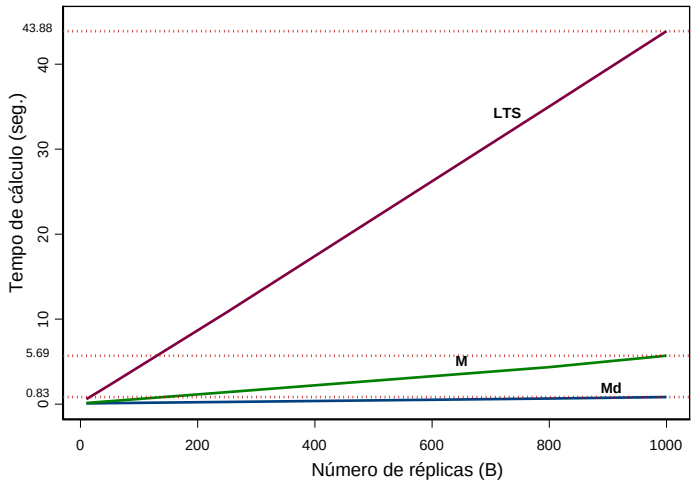


Figura 1.7: Tempos de execução (em segundos) quando se utiliza o *bootstrap* com os três estimadores de localização: $\bar{X}_n$ (Md), estimador–M (M) e LTS.

Para terminar refira-se que os métodos de estimação utilizados neste exemplo para um estimador genérico, T, podem também ser aplicados a qualquer transformação desse estimador. Por exemplo, poder-se-ía querer determinar características de centralidade ou precisão para $T = \ln \bar{X}_n$.

■

O exemplo anterior mostra que a existência de observações atípicas e o custo elevado de cálculo constituem dois aspectos que limitam a aplicação do *bootstrap* clássico de Efron. Neste trabalho propõe-se um procedimento que permite superar essas limitações. Esse novo procedimento de reamostragem permite efectuar inferências robustas quando o modelo gerador dos dados pertence a uma vizinhança de contaminação de um modelo central (univariado ou multivariado). Para isso recorre-se à utilização da função de influência (Hampel *et al.*, 1986) de um funcional de interesse para a indução de um mecanismo de controlo no plano de reamostragem, permitindo acomodar e controlar possíveis observações atípicas. Pode-se sintetizar os objectivos principais deste trabalho, nos seguintes pontos:

- construção de um procedimento de reamostragem robusto que não requeira excessivo tempo de cálculo;

- desenvolvimento de técnicas de reamostragem aplicáveis à Análise Multivariada. Neste sentido apresenta-se um critério robusto de selecção de variáveis em Análise Discriminante Linear. Este tema assume particular relevância por ser importante dispor de critérios objectivos de selecção. Embora para alguns métodos ditos clássicos de Análise Discriminante exista uma resposta (também clássica) para esta questão, ela não é necessariamente a melhor, nomeadamente em termos de robustez.

Tendo em conta os pontos expostos acima organizou-se este trabalho da seguinte forma – no Capítulo 2 são revistos e descritos os principais conceitos de inferência robusta numa perspectiva multivariada. Para além dos conceitos de função de influência e ponto de rotura de um funcional são ainda descritos vários procedimentos para a estimação robusta com alto ponto de rotura de parâmetros de localização e dispersão de populações multivariadas. Apresentam-se ainda os poucos procedimentos existentes relacionados com a inferência robusta não pontual assim como uma análise das dificuldades inerentes a este tópico.

No Capítulo 3 discute-se a motivação e importância dos métodos estatísticos de reamostragem, fazendo-se uma breve referência aos vários métodos existentes. Neste capítulo também se coligem as aplicações mais importantes relacionadas com dois deles: o *jackknife* e o *bootstrap*. Relativamente ao *jackknife* ainda se apresenta uma sua extensão, o denominado *deleted-d jackknife*. Quanto ao *bootstrap*, introduzem-se as ideias base assim como a sua aplicação a intervalos de confiança e de uma forma breve à Estatística Multivariada.

No Capítulo 4 descreve-se o novo método proposto, referido atrás como *bootstrap* robusto. Inicia-se esta parte da tese com uma discussão pormenorizada sobre a problemática das observações discordantes e do *bootstrap* clássico (Efron, 1979). A seguir são definidos alguns conceitos necessários para a construção desse novo procedimento, nomeadamente a função de influência empírica padronizada robusta. Após a descrição do *bootstrap* robusto apresenta-se a sua aplicação em três problemas distintos:

(i) estimação do parâmetro de localização num modelo de localização e escala univariado;

(ii) estimação do coeficiente de correlação numa população bivariada;

(iii) selecção de variáveis em Análise Discriminante Linear para dois grupos.

Neste capítulo é ainda desenvolvida uma generalização ao método proposto justificada pela relação que se mostra existir entre este método e os estimadores–W (Tukey, 1970). Aplica-se esta generalização aos problemas de estimação (i) e (ii).

Finalmente, no último capítulo, conclui-se esta dissertação efectuando-se uma síntese do trabalho apresentado e sugerindo-se também algumas direcções que poderão vir a ser desenvolvidas no futuro, na sequência do que aqui foi exposto e dos desafios encontrados ao longo deste trabalho.

"*Você acredita num Deus que joga aos dados, eu na lei e ordem completas.*"

Albert Einstein (carta a Max Born)

CAPÍTULO 2

Estimação multivariada e inferência robusta

A estimação, a predição e a tomada de decisões em situações de incerteza constituem os objectivos principais da estatística. Para cumprir esses objectivos, a abordagem usual utilizada pela estatística, quer esta seja frequencista quer seja Bayesiana, consiste normalmente em basear-se em modelos paramétricos. Essencialmente baseia-se na determinação e construção de procedimentos óptimos sob a hipótese de um modelo paramétrico exacto. Porém, se esse modelo paramétrico não for totalmente correcto as inferências resultantes podem ser completamente distorcidas ou até mesmo erradas. Ao utilizar-se um modelo paramétrico exacto está-se a assumir que a realidade se comporta idealmente, no entanto, os modelos reais apresentam geralmente alguns desvios em relação aos seus modelos idealizados. Uma abordagem estatística que permite lidar com esses afastamentos é fornecida pela estatística robusta. Esta parte da estatística tem como finalidade a construção de métodos inferenciais que permitam responder de forma adequada em relação a alguns afastamentos (nomeadamente de hipóteses mais comuns de normalidade, independência e linearidade) do modelo paramétrico exacto.

Para ultrapassar alguma da não adequabilidade dos modelos paramétricos existem outros métodos estatísticos, nomeadamente procedimentos adaptativos e procedimentos não paramétricos (distribuição livre). Porém, a resposta dada por cada um desses métodos é distinta da dada pela estatística robusta.

Os procedimentos adaptativos (Hogg, 1974; Takeuchi, 1971) são assintoticamente óptimos para uma classe alargada de distribuições. De um ponto de vista teórico, estes métodos não respondem à mesma questão que a estatística robusta, já que eles contêm a restrição da simetria das distribuições (ou simetria unimodal, no caso da estimação de parâmetros de localização). Para além dessa restrição, outras questões se colocam relativamente ao uso de procedimentos adaptativos, nomeadamente a sua baixa eficiência para dimensões pequenas ou moderadas e o facto de serem apenas baseados em resultados assintóticos.

Quanto aos procedimentos não paramétricos deve precisar-se que uma interpretação estrita da definição do termo "procedimentos estatísticos não paramétricos" não implica que a estimação ou inferência não estejam directamente ligadas a parâmetros. É claro que os procedimentos não paramétricos dizem respeito a "parâmetros", isto é, a características numéricas das distribuições (a mediana, por exemplo, é um importante "parâmetro" em estatística não paramétrica). O sentido de não paramétrico advém do facto dessa estimação ou inferência permitir ignorar determinadas considerações acerca da distribuição do modelo gerador dos dados (por exemplo, a normalidade). A distinção básica entre a estatística não paramétrica e a robusta é que esta última não permite relaxar completamente a distribuição do modelo gerador dos dados, dado que, a teoria da robustez assenta em vizinhanças de modelos paramétricos.

Para além disso, a maior parte dos métodos não paramétricos são apenas de distribuição livre com respeito a algumas das suas características. Por exemplo, a maior parte dos testes de hipóteses não paramétricos possui uma distribuição de probabilidades, sob a hipótese nula, que não depende do modelo gerador dos dados (distribuição livre) mas a sua potência é distribucionalmente dependente. Apesar da sua aplicabilidade estar sujeita à verificação de hipóteses muito gerais, a maior parte dos procedimentos não paramétricos possui certas restrições de simetria e de continuidade da distribuição do modelo subjacente aos dados. Se estas hipóteses não são verificadas, a eficiência do método diminui. Significa isto que, existe por vezes alguma falta de robustez nestes procedimentos. Uma combinação ideal pode ser obtida através da conjugação dos métodos não paramétricos e robustos mediante a utilização de técnicas de reamostragem a partir de estatísticas robustas. Esta ideia será desenvolvida mais adiante no Capítulo 4.

Verifica-se também que a teoria da rejeição de observações discordantes (na literatura anglo-saxónica designadas por *outliers*) está bastante ligada à teoria da robustez. Apesar destas duas teorias terem estado, durante bastante tempo, separadas assiste-se hoje em dia ao seu uso em forma de complementaridade. De facto, a identificação dessas observações discordantes torna-se mais eficiente quando se utilizam procedimentos robustos (ver por exemplo, Rousseeuw e Van Zomeren, 1990 e Becker e Gather, 2001).

Duas contribuições fundamentais para a teoria da robustez foram as dadas por Huber (1964) e Hampel (1968). Huber introduziu uma importante classe de estimadores, os designados estimadores–M. Outra importante noção que este autor introduziu e formalizou foi a de vizinhança de um modelo paramétrico através do "modelo para erros grosseiros" (na literatura anglo-saxónica designado por *gross error model*). Segundo a abordagem de Huber alcança-se robustez através da construção de um estimador minimax o qual minimiza um critério de perda (tal como o viés ou a variância) no pior caso de qualquer vizinhança do modelo paramétrico. Hampel, por seu lado, introduziu vários critérios que permitem qualificar um estimador como sendo robusto face a erros grosseiros. Esta abordagem também designada por abordagem infinitesimal tem como base três importantes conceitos, a robustez qualitativa, a função de influência e o ponto de rotura de um estimador. Nas secções seguintes estes conceitos irão ser abordados com mais detalhe.

Face ao exposto até aqui, considera-se assim que são necessários procedimentos robustos porque a

distribuição ideal do modelo probabilístico não consegue reflectir, por um lado, toda a realidade da Natureza, assim como toda a informação dada por um conjunto de pontos observados (a qual pode incluir arredondamentos, erros grosseiros nos dados, mudanças ligeiras nas distribuições marginais e nas estruturas de correlação dos dados, eliminação, adição, e identificação incorrecta de certos pontos). Estes aspectos tornam-se ainda mais relevantes em estatística multivariada devido à maior dificuldade em detectar qualquer um destes problemas quando a dimensão do espaço amostral aumenta.

Dada a importância para este trabalho do conceito de robustez providenciado pela função de influência inicia-se a próxima secção precisamente pela sua definição e descrição das suas propriedades. Apresenta-se ainda, o também importante, conceito de ponto de rotura. Em seguida faz-se uma breve descrição e discussão acerca de estimadores robustos de localização e dispersão multivariados de alto ponto de rotura e termina-se este capítulo com uma curta exposição sobre estimação robusta não pontual.

2.1 Função de influência

A função de influência foi introduzida por Hampel (1968, 1974), tendo sido amplamente desenvolvida posteriormente em Hampel *et al.* (1986). Esta função constitui uma das técnicas fundamentais usadas na teoria da robustez e na área de diagnóstico (neste âmbito ver os trabalhos de Cook e Weisberg, 1982; Critchley, 1985; Critchley e Vitiello, 1991; He e Simpson, 2000; Fung, 1996 e Lu *et al.*, 1997).

O objectivo fundamental desta função consiste em fornecer uma medida precisa da sensibilidade de um estimador a uma "pequena" quantidade de contaminação em cada ponto. O seu estudo permite também caracterizar várias propriedades de um estimador, nomeadamente de robustez. Para além disso, permite também a derivação de estimadores robustos com propriedades definidas à partida. Sob certas condições de regularidade, esta função também é utilizada na determinação da variância assintótica de um estimador.

A maior parte dos estimadores, T_n, dependem da amostra, $(x_1, x_2, \ldots, x_n)$, uni ou multivariada, apenas através da função de distribuição empírica,

$$F_n(x) = n^{-1} \sum_{i=1}^{n} \Delta_{x_i},$$

isto é, $T_n = T(F_n)$, onde Δ_u é a distribuição que coloca massa 1 no ponto u. Quando isto acontece diz-se que o estimador T_n é equivalente ao funcional $T(F)$. Se o funcional, $T(F)$, é diferenciável[1] em F então

$$T(G) = T(F) + dT_F(G - F) + R, \tag{2.1}$$

[1]Conceito a precisar mais adiante.

onde F e G são funções de distribuição, dT_F corresponde ao diferencial de T em F na direcção de $G - F$ e R corresponde ao termo geralmente designado por "resto". A expressão (2.1) é designada por expansão de Von Mises de primeira ordem do funcional T, podendo-se verificar a sua analogia com a expansão de Taylor de primeira ordem para funções usuais. Existem três noções de diferenciabilidade de funcionais: a compacta (ou de Hadamard), a de Fréchet e a de Gâteaux. Cada uma delas está relacionada com vários tipos de convergência de R à medida que G converge para F. Assim, a diferenciabilidade compacta requer que R convirja para zero uniformemente em conjuntos compactos. A diferenciabilidade à Fréchet requer que R convirja para zero uniformemente mas em conjuntos limitados. A diferenciabilidade à Gâteaux é um conceito mais "fraco" que os dois anteriores e está relacionado com a existência da derivada usual de uma função de valor real $T(F_\varepsilon)$ com respeito ao valor real ε.

Assim $T(F)$ é diferenciável à Gâteaux em F se existir uma função mensurável, ψ_F, tal que

$$\lim_{\varepsilon \to 0} \varepsilon^{-1} \left\{ T\left[F + \varepsilon\left(G - F\right)\right] - T(F) \right\} = \int \psi_F d(G - F). \tag{2.2}$$

Se existir a derivada compacta ou à Fréchet ela é igual à derivada à Gâteaux. Por esta razão, quando se pretende determinar derivadas de funcionais o meio de cálculo mais conveniente é o da derivada à Gâteaux. Sendo ψ_F convenientemente padronizada tal que $\int \psi_F dF = 0$ então pode escrever-se

$$dT_F(G - F) = \int \psi_F dG = 0.$$

Se $G = \Delta_x$, ou seja, se a distribuição é degenerada no ponto x, então de (2.2) tira-se que

$$\lim_{\varepsilon \to 0} \varepsilon^{-1} \left\{ T\left[(1-\varepsilon)F + \varepsilon\Delta_x\right] - T(F) \right\} = \psi_F(x). \tag{2.3}$$

Esta expressão não é mais do que a definição de função de influência, $IF(x; F, T)$, do funcional T em F definida por Hampel (Hampel *et al.*, 1986, pág. 84). A distribuição $(1-\varepsilon)F + \varepsilon\Delta_x$ para pequenos valores de ε é praticamente a função de distribuição F a menos do peso adicional no ponto x. Verifica-se portanto que esta função mede o efeito no funcional $T(F)$ de uma alteração infinitesimal no ponto x, ou seja, a influência em $T(F)$ devida ao ponto x.

Para mais detalhes sobre diferenciabilidade de funcionais estatísticos consultar, por exemplo, Fernholz (1983) e Shao e Tu (1995, Cap. 2).

As definições seguintes formalizam e generalizam a noção da função de influência de um funcional estatístico. Assuma-se que os estimadores considerados são equivalentes a funcionais, que m representa a dimensão do espaço amostral e que p denota a dimensão do espaço paramétrico.

Seja $\mathcal{X}$ um espaço amostral, $\mathcal{X} \subseteq I\!R^m$, $\mathcal{F}$ o espaço das funções de distribuição sobre $\mathcal{X}$ e $\mathbf{T}$ um estimador de um parâmetro $\boldsymbol{\theta}$, com $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq I\!R^p$. Seja $\mathcal{D}(\mathbf{T}) \subseteq \mathcal{F}$ o domínio do funcional $\mathbf{T}(F)$ e $\Delta_{\mathbf{x}}$ a função de distribuição degenerada no ponto $\mathbf{x}$.

Definição 2.1 – Função de influência: *A função de influência do funcional* $\mathbf{T}$ *em* F *é dada por*

$$IF(\mathbf{x}; \mathbf{T}, F) = \lim_{\varepsilon \to 0} \varepsilon^{-1} \left\{ \mathbf{T}[(1-\varepsilon)F + \varepsilon\Delta_{\mathbf{x}}] - \mathbf{T}(F) \right\}. \tag{2.4}$$

Esta definição aplica-se apenas a funcionais vectoriais porém ela é de fácil de generalização a funcionais definidos por matrizes. Para isso, bastará usar um operador que transforme a matriz num vector, como por exemplo, o operador *vecs* definido em Hampel *et al.* (1986, pág. 272), ou um outro que permita organizar os elementos distintos da matriz num vector.

A função de influência, $IF(\mathbf{x};\mathbf{T},F)$, de um funcional estatístico, $\mathbf{T}$, possui, sob condições de regularidade adequadas, as seguintes propriedades:

P.1 O seu valor esperado é nulo, isto é

$$E_F\left[IF\right] = \int IF(\mathbf{x};\mathbf{T},F)dF(\mathbf{x}) = \mathbf{0},$$

onde $\mathbf{0}$ corresponde ao vector nulo de dimensão p.

P.2 Para uma qualquer função de distribuição G próxima de F, a expansão de Von Mises de primeira ordem de $\mathbf{T}$ em F (avaliada em G) é dada por

$$\mathbf{T}(G) \simeq \mathbf{T}(F) + \int IF(\mathbf{x};\mathbf{T},F)dG(\mathbf{x}), \tag{2.5}$$

então para amostras de grandes dimensões, $(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n)$ pode tomar-se $G = F_n$ e a expressão (2.5) resulta no seguinte resultado importante:

$$\mathbf{T}(F_n) \simeq \mathbf{T}(F) + n^{-1}\sum_{j=1}^{n} IF(\mathbf{x}_i;\mathbf{T},F) \Leftrightarrow \tag{2.6}$$

$$\Leftrightarrow n^{1/2}\left(\mathbf{T}(F_n) - \mathbf{T}(F)\right) \simeq n^{-1/2}\sum_{j=1}^{n} IF(\mathbf{x}_i;\mathbf{T},F). \tag{2.7}$$

P.3 Pela propriedade anterior (2.7) e pelo Teorema do Limite Central, conclui-se que a variável aleatória $n^{1/2}\left(\mathbf{T}(F_n) - \mathbf{T}(F)\right)$ tem aproximadamente distribuição normal p-multivariada com valor esperado nulo e matriz de covariâncias assintótica ($VA(\mathbf{T},F) = \lim_{n\to\infty}(nVar(\mathbf{T}))$) dada por

$$VA(\mathbf{T},F) = \int IF(\mathbf{x};\mathbf{T},F)IF(\mathbf{x};\mathbf{T},F)^T dF(\mathbf{x}). \tag{2.8}$$

P.4 Sendo $\beta : \mathbf{\Theta} \subset I\!R^p \to \beta(\mathbf{\Theta}) \subset I\!R^k$ uma transformação regular do parâmetro $\boldsymbol{\theta}$, com matriz Jacobiana, $\mathbf{B}(\boldsymbol{\theta})_{(k\times p)} = \partial\beta(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, e $\beta(\mathbf{T})$ um estimador de $\beta(\boldsymbol{\theta})$ verificam-se as seguintes propriedades de equivariância,

$$IF(\mathbf{x};\beta(\mathbf{T}),F) = \mathbf{B}(\mathbf{T}(F))IF(\mathbf{x};\mathbf{T},F), \tag{2.9}$$

$$VA(\beta(\mathbf{T}),F) = \mathbf{B}(\mathbf{T}(F))VA(\mathbf{T},F)\mathbf{B}(\mathbf{T}(F))^T. \tag{2.10}$$

Para além destas propriedades, a função de influência permite através da sua forma obter informação acerca da robustez de $\mathbf{T}$. Assim, um funcional estatístico cuja função de influência seja limitada não é sensível a observações extremas. A definição 2.2 formaliza este conceito.

Definição 2.2 – Sensibilidade a grandes erros: *Dado um funcional* $\mathbf{T}$ *em* F *diz-se que a sua sensibilidade a grandes erros (não padronizada) é*

$$s_{np}(\boldsymbol{T}, F) = \sup_{\mathbf{x}} \|IF(\mathbf{x}; \boldsymbol{T}, F)\| \tag{2.11}$$

e $\boldsymbol{T}$ *diz-se* B*-robusto sse* $s_{np}(\boldsymbol{T}, F)$ *for finito (isto é, um estimador é* B*-robusto sse a sua função de influência for limitada).*

Quando se trabalha em problemas unipopulacionais, a função de influência não é mais do que o núcleo da primeira derivada do funcional que define o estimador. No caso multipopulacional, isto é, de estimadores de parâmetros que dependem de mais do que uma população, é necessário usar a noção de função de influência parcial (em analogia com a noção de derivada parcial). Este conceito foi introduzido por Pires (1995) e desenvolvido em Pires e Branco (2002) e generaliza a noção de função de influência ao caso multipopulacional. A ideia é avaliar a alteração em $\mathbf{T}$ devida à contaminação infinitesimal num determinado ponto para cada uma das populações. Considere-se agora o funcional $\mathbf{T}(\mathbf{F})$, com domínio $\mathcal{D}(\mathbf{T}) \subseteq \mathcal{F} \times \mathcal{F} \times \cdots \times \mathcal{F}$, onde $\mathbf{F} = F_1 \times F_2 \times \cdots \times F_g$ e g designa o número de populações.

Definição 2.3 – Funções de influência parciais: *Funções de influência parciais do funcional* $\mathbf{T}$ *são definidas por*

$$IF_i(\mathbf{x}, \mathbf{T}, \mathbf{F}) = \lim_{\varepsilon \to 0} \varepsilon^{-1} \left\{ \mathbf{T}[((1-\varepsilon)F_i + \varepsilon \Delta_{\mathbf{x}}) \times \mathbf{F}_{-i}] - \mathbf{T}(\mathbf{F}) \right\}, \tag{2.12}$$

para os valores de $\mathbf{x} \in \mathcal{X}$ *onde esse limite exista e* $i = 1, 2, \cdots, g$. $\mathbf{F}_{-i}$ *representa a seguinte medida produto* $\mathbf{F}_{-i} = F_1 \times F_2 \times \cdots \times F_{(i-1)} \times F_{(i+1)} \times \cdots \times F_g$.

Pires e Branco no seu artigo de 2002 derivaram as propriedades atrás referidas para o caso multipopulacional. Assim, sob determinadas condições de regularidade, a função de influência parcial, IF_i $(i = 1, 2, \cdots, g)$, possui as seguintes propriedades:

R.1 O seu valor esperado é nulo, isto é

$$E_{F_i}[IF_i] = \int IF_i(\mathbf{x}; \mathbf{T}, \mathbf{F}) dF_i(\mathbf{x}) = \mathbf{0},$$

onde $\mathbf{0}$ corresponde ao vector nulo de dimensão p.

R.2 Para uma qualquer função de distribuição multivariada, denotada por $\mathbf{G} = (G_1, G_2, \cdots, G_g)$, próxima de $\mathbf{F}$, a expansão de Von Mises de primeira ordem de $\mathbf{T}$ em $\mathbf{F}$ (avaliada em $\mathbf{G}$) é dada por

$$\mathbf{T}(\mathbf{G}) \simeq \mathbf{T}(\mathbf{F}) + \sum_{i=1}^{g} \int IF_i(\mathbf{x}; \mathbf{T}, \mathbf{F}) dG_i(\mathbf{x}), \tag{2.13}$$

tomando-se $G_i = F_{n_i}$ (n_i grande) e representando a função de distribuição empírica, de g amostras independentes, $\{\mathbf{x}_{ij}, i = 1, 2, \cdots, g; j = 1, 2, \cdots, n_i\}$, por $\mathbf{F}_{n_t} = (F_{n_1}, F_{n_2}, \cdots, F_{n_g})$ onde $n_t = \sum_{i=1}^{g} n_i$ a expressão (2.13) origina o seguinte resultado

$$\mathbf{T}(\mathbf{F}_{n_t}) \simeq \mathbf{T}(\mathbf{F}) + \sum_{i=1}^{g} \sum_{j=1}^{n_i} n_i^{-1} IF_i(\mathbf{x}_{ij}; \mathbf{T}, \mathbf{F}) \Leftrightarrow \tag{2.14}$$

$$\Leftrightarrow {n_t}^{1/2} \left(\mathbf{T}(\mathbf{F}_{n_t}) - \mathbf{T}(\mathbf{F})\right) \simeq \sum_{i=1}^{g} (n_i w_i)^{-1/2} \sum_{j=1}^{n_i} IF_i(\mathbf{x}_{ij}; \mathbf{T}, \mathbf{F}), \tag{2.15}$$

onde $w_i = n_i/n_t$.

R.3 Partindo da expressão (2.15) e respeitando as condições do Teorema do Limite Central pode concluir-se que a variável aleatória ${n_t}^{1/2} \left(\mathbf{T}(\mathbf{F}_{n_t}) - \mathbf{T}(\mathbf{F})\right)$ tem aproximadamente distribuição normal p-multivariada com valor esperado nulo e matriz de covariâncias assintótica dada por

$$VA(\mathbf{T}, \mathbf{F}) = \sum_{i=1}^{g} w_i^{-1} VA_i(\mathbf{T}, \mathbf{F}), \tag{2.16}$$

onde

$$VA_i(\mathbf{T}, \mathbf{F}) = \int IF_i(\mathbf{x}; \mathbf{T}, \mathbf{F}) IF_i(\mathbf{x}; \mathbf{T}, \mathbf{F})^T dF_i(\mathbf{x}). \tag{2.17}$$

R.4 As propriedades de equivariância apresentadas na propriedade **P.4** generalizam-se facilmente bastando, para tal, substituir nas expressões (2.9) e (2.10) IF por IF_i para $i = 1, 2, \cdots, g$.

Versões finitas da Função de Influência

A função de influência é um conceito assintótico, existem, porém, versões finitas deste conceito. Em geral a "conversão" da noção de função de influência para o caso de amostragem finita é realizada substituindo a função de distribuição F por F_n ou por F_{n-1}. Das várias versões existentes, as mais importantes são a curva de sensibilidade (Tukey, 1970), o método de reamostragem *jackknife* (Quenouille, 1949) e a denominada função de influência empírica (Hampel *et al.*, 1986, pág. 93).

A curva de sensibilidade de Tukey é uma versão univariada, e é obtida substituindo em (2.4) (neste caso considera-se $m = 1$ e $p = 1$) F por F_{n-1} e ε por n^{-1}. Omitindo-se o limite, essa substituição origina a seguinte expressão:

$$CS_{n-1}(x) = n \left\{T((n-1)/n \times F_{n-1} + n^{-1}\Delta_x) - T(F_{n-1})\right\}. \tag{2.18}$$

Pode também generalizar-se esta ideia para o caso multivariado, obtendo-se neste caso uma superfície de sensibilidade.

Obtém-se uma versão finita da função de influência por *jackknife*, para $j = 1, 2, \cdots, n$, calculando a seguinte diferença

$$\mathbf{T}_{-j} = n\mathbf{T}_n - (n-1)\mathbf{T}_{n-1}^{(j)}, \tag{2.19}$$

onde n representa a dimensão de uma amostra $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$, e $\mathbf{T}_{n-1}^{(j)} = \mathbf{T}(F_{n-1}^{(j)})$, onde $F_{n-1}^{(j)}$ denota a função de distribuição empírica omitindo a j-ésima observação. A expressão (2.19) pode também ser vista como uma aproximação de (2.4) onde se substitui F por F_{n-1} e ε por $-(n-1)^{-1}$. O método de reamostragem *jackknife* será descrito com mais detalhe no Capítulo 4.

A função de influência empírica apresentada em Hampel *et al.* (1986) não é mais do que o valor do funcional quando se adiciona uma observação, $\mathbf{x}$, isto é,

$$EIF(\mathbf{x}) = \mathbf{T}_n(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{n-1}, \mathbf{x}). \tag{2.20}$$

Quer esta última função, quer a "curva" de sensibilidade podem ser redefinidas tendo em conta não a adição de uma observação $\mathbf{x}$ mas sim substituindo uma observação existente por um valor arbitrário $\mathbf{x}$. Tem-se assim disponíveis mais duas versões finitas da função de influência.

Como foi exposto anteriormente, a função de influência mede o efeito de contaminações infinitesimais no funcional estatístico. De uma forma simples, estando numa situação univariada, pode dizer-se que a função de influência mede o comportamento do estimador (funcional) quando se faz percorrer a contaminação pela recta real. Mesmo quando o valor da função de influência não é um escalar pode definir-se uma distância baseada nela que possibilitará uma ordenação das observações segundo o valor da função. Logo, observações influentes (no sentido de Hampel) serão aquelas que produzem valores elevados da função de influência. Outra questão será definir o limiar valor da função de influência (ou da sua distância) a partir do qual se pode considerar que uma observação é influente. Retorna-se a este assunto no Capítulo 4.

2.2 Ponto de rotura

A função de influência é uma importante ferramenta na caracterização de estimadores robustos. Porém, este conceito é meramente local, isto é, a função de influência descreve o comportamento do estimador face a contaminações infinitesimais. É então natural que se pretenda dispor de um conceito global, ou seja, de uma medida que indique qual o afastamento máximo à distribuição do modelo que é permitido sem que o estimador se degrade. Essa medida global designa-se por ponto de rotura do estimador. Das várias definições formais existentes deste conceito pode dizer--se que basicamente variam consoante o tipo de distância que permite definir o afastamento entre distribuições de probabilidades consideradas, nomeadamente a distância de Prohorov, Levy ou a métrica de Lipschitz. Para mais detalhes sobre estes conceitos veja-se Hampel *et al.* (1986, Cap. 2).

O ponto de rotura acima referido é designado por ponto de rotura assintótico. No entanto, é mais simples e intuitivo calcular uma versão finita, isto é, que não envolva distâncias entre distribuições de probabilidade. Existem várias definições de ponto de rotura finito. As definições seguintes formalizam este conceito segundo duas abordagens que, embora praticamente equivalentes, são distintas.

Considere-se, $\mathbf{x}_n$ uma amostra de dimensão n, $\mathbf{T}_n(\mathbf{x}_n)$ um estimador avaliado em $\mathbf{x}_n$ e S_k o conjunto

de todas as amostras de dimensão n, $\mathbf{s}_{n_k}$, obtidas substituindo quaisquer k observações em $\mathbf{x}_n$ por valores arbitrários.

Para Hampel, o ponto de rotura finito de um estimador corresponde à maior proporção de observações que pode ser modificada sem que o estimador se degrade, veja-se a definição 2.4 para a formalização.

Definição 2.4 –**Ponto de rotura** (Hampel): *Ponto de rotura de amostra finita de um estimador* $\mathbf{T}_n$ *(segundo Hampel) é dado por*

$$\varepsilon_H^*(T_n(\mathbf{x})) = 1/n \max_{1\leq k\leq n} \left\{ \sup_{\mathbf{s}_{n_k}\in S_k} \|T_n(\mathbf{s}_{n_k}) - T_n(\mathbf{x})\| < \infty \right\}. \tag{2.21}$$

Uma outra definição ligeiramente diferente é a apresentada por Donoho e Huber (1983). Estes autores definiram ponto de rotura de um estimador, para amostras finitas, como sendo a mais pequena proporção de observações que têm que ser modificadas para obter uma perturbação infinita na estimativa.

Definição 2.5 – **Ponto de rotura** *(Donoho e Huber): Ponto de rotura de amostra finita de um estimador* $\mathbf{T}_n$ *(segundo Donoho e Huber) é dado por*

$$\varepsilon_{DHu}^*(T_n(\mathbf{x})) = \min_{1\leq k\leq n} \left\{ k/n; \sup_{\mathbf{s}_{n_k}\in S_k} \|T_n(\mathbf{s}_{n_k}) - T_n(\mathbf{x})\| = \infty \right\}. \tag{2.22}$$

Verifica-se então que $\varepsilon_{DHu}^* = \varepsilon_H^* + 1/n$. Por exemplo o ponto de rotura da mediana univariada, segundo Hampel, é dado por $\varepsilon_H^* = ([(n+1)/2] - 1)/n$ e segundo Donoho e Huber ele será dado por $\varepsilon_{DHu}^* = [(n+1)/2]/n$. Para grandes amostras, isto é, $n \to \infty$, o ponto de rotura da mediana é $1/2 = \lim_{n\to\infty} \varepsilon_{DHu}^* = \lim_{n\to\infty} \varepsilon_H^*$.

As definições 2.4 e 2.5 apenas contemplam o caso de "explosão" da estimativa, isto é, a quantidade mínima de contaminação que torna a estimativa ilimitada. No entanto outros problemas podem surgir, nomeadamente para estimadores de covariâncias, que podem deixar de ter sentido quando um elemento da sua diagonal for zero, o que torna a matriz singular. Este caso é denominado de "implosão". Assim sendo, é usual definir ponto de rotura finito para estimadores de escala como sendo a mínima quantidade de contaminação que torna o maior valor próprio da matriz de covariâncias infinito ou que torna o menor valor próprio zero. Nem sempre é possível determinar teoricamente o ponto de rotura de um estimador, nestes casos é comum ele poder ser estimado por meio de estudos de simulação devidamente planeados.

É desejável que um estimador robusto tenha um ponto de rotura positivo elevado. A quantificação significativa deste termo "elevado" é de 50%, pois para percentagens de contaminação superiores não se consegue distinguir na amostra qual é a parte contaminada da não contaminada.

As propriedades desejáveis para um estimador robusto são: um alto ponto de rotura, uma função de influência limitada e eficiência elevada sob o modelo paramétrico central considerado. No entanto,

é difícil encontrar estimadores que possuam estas três propriedades em simultâneo e, em geral, o que se consegue obter são estimadores B-robustos óptimos. Este conceito constitui o denominado princípio de optimalidade de Hampel (Hampel *et al.*, 1986, pág. 116). Recentemente o interesse foca-se em estimadores com alto ponto de rotura, não só pela menor sensibilidade que poderão ter a elevadas percentagens de contaminação como também pela constatação das boas propriedades de eficiência que as novas propostas têm demonstrado possuir. Para além destas razões pensa-se que também é crucial o desenvolvimento que se tem vindo a assistir ao nível da sua implementação computacional. Na secção seguinte descrevem-se alguns estimadores de localização e dispersão multivariados e abordam-se algumas questões relacionadas com o seu cálculo. Todavia, será dado mais relevo aos estimadores com alto ponto de rotura.

2.3 Estimação robusta de parâmetros de localização e dispersão multivariados

Muitos dos métodos de análise multivariada requerem a estimação de parâmetros de localização e de dispersão. Em analogia com o que acontece no caso univariado, os estimadores multivariados geralmente usados (vector média amostral e matriz de covariâncias amostral) são óptimos quando o modelo gerador dos dados é a normal multivariada, basta, no entanto, existir uma pequena percentagem de observações discordantes para que eles percam as suas boas propriedades. Ver, por exemplo, Rousseeuw e Leroy (1987).

Desde a década de 70 têm sido propostos vários métodos para calcular estimadores robustos multivariados. Contudo, os procedimentos robustos necessitam mais frequentemente de processos iterativos comparativamente aos métodos clássicos, o que aliado a problemas multidimensionais os fazem depender de recursos computacionais elevados. Este problema começou a ser resolvido com o rápido desenvolvimento dos meios computacionais a partir da década de 80, que permitiu impulsionar a aplicação e o desenvolvimento de novos procedimentos robustos em análise multivariada. Hoje em dia, vários *softwares* estatísticos incluem nas suas versões elementares alguns procedimentos robustos.

Alternativas robustas para estimação multivariada da localização e dispersão foram inicialmente propostas usando métodos de eliminação de observações discordantes, essa foi a abordagem de um dos primeiros trabalhos nesta área (Gnanadesikan e Kettenring, 1972). Seja $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_i, \cdots, \mathbf{x}_n)$ uma amostra onde cada $\mathbf{x}_i$ é um vector m-dimensional e considere-se ainda a seguinte distância amostral,

$$d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}),$$

onde $\bar{\mathbf{x}}$ e $\mathbf{S}$ correspondem respectivamente ao vector média e à matriz de covariâncias amostrais.

Gnanadesikan e Kettenring (1972) baseando-se na distância amostral, d_i^2, detectavam e eliminavam observações atípicas e recalculavam o vector amostral de localização e a matriz de covariâncias após essa eliminação. Repetiam o processo até obter estabilidade. Porém estes estimadores não possuem

boas propriedades tendo inclusive ponto de rotura igual a zero.

Os estimadores–M de localização e dispersão multivariados foram propostos por Maronna (1976) e constituem uma generalização das ideias de Huber (1964) ao caso multivariado. A abordagem seguida por Maronna foi a de considerar a distância d_i^2 e usá-la para "pesar" as observações. Estes estimadores assim obtidos são também uma generalização dos estimadores de máxima verosimilhança. Porém, esses estimadores possuem um ponto de rotura que diminui à medida que aumenta a dimensão do espaço o que constitui um grave inconveniente para dimensões elevadas.

Seja $SDP(m)$ a classe das matrizes Simétricas Definidas Positivas de dimensão m, $\Theta = I\!R^m \times SDP(m)$ e o conjunto dos pares $\boldsymbol{\theta} = (\mathbf{t}, \mathbf{C})$ (este último pode ser visto como um subconjunto aberto de $I\!R^{m+(1/2)m(m+1)}$).

Sejam ainda $X_1, X_2 \cdots, X_n$ variáveis aleatórias m-dimensionais independentes e identicamente distribuídas (i.i.d.). O objectivo é a estimação do parâmetro $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ o qual caracteriza uma distribuição elipticamente simétrica, $F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ cuja função de densidade é da forma

$$(\det \boldsymbol{\Sigma})^{-1/2} f\left[(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right], \tag{2.23}$$

onde $f : [0, \infty) \to [0, \infty)$ é uma função tal que $f(\|\mathbf{x}\|)$ é uma densidade em $I\!R^m$ e $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \Theta$.

Dado um qualquer vector $\mathbf{t}$ e uma matriz $\mathbf{C}$ pertencente a $SDP(m)$ define-se a distância quadrada de Mahalanobis entre $\mathbf{x}$ e $\mathbf{t}$ como sendo:

$$DM^2(\mathbf{x}; \mathbf{t}, \mathbf{C}) = (\mathbf{x}-\mathbf{t})^T \mathbf{C}^{-1} (\mathbf{x}-\mathbf{t}). \tag{2.24}$$

Definição 2.6 – Estimadores–M: *Dadas duas funções reais de variável real não negativas u_1, u_2 as estimativas–M de localização e dispersão multivariadas, $(\mathbf{t}, \mathbf{C})$, são definidas como solução de*

$$\mathbf{t} = \frac{\sum_i^n u_1(d_i^2)\mathbf{x}_i}{\sum_i^n u_1(d_i^2)},$$

$$\mathbf{C} = n^{-1} \sum_i^n u_2(d_i^2)(\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})^T, \tag{2.25}$$

onde $d_i^2 = DM^2(\mathbf{x}_i; \mathbf{t}, \mathbf{C})$ e $\mathbf{C} \in SDP(m)$.

Como referido, os estimadores–M podem ser vistos como uma generalização dos estimadores de máxima verosimilhança para distribuições elipticamente simétricas, por exemplo, para a normal multivariada os estimadores de máxima verosimilhança correspondem aos estimadores–M quando se consideram as funções $u_1 = u_2 \equiv 1$. Por outro lado, os estimadores–M não são mais do que versões ponderadas do vector média amostral, $\bar{\mathbf{X}}_n$, e da matriz de covariâncias amostral, $\mathbf{S}_n$, cujas ponderações são determinados iterativamente com base nos dados e usando uma distância adequada. Maronna (1976) mostrou que sob certas condições de regularidade estes estimadores são

consistentes e assintoticamente normais. A única desvantagem desta classe de estimadores reside no facto de possuir um ponto de rotura baixo para dimensões superiores a $m = 3$.

Um outro estimador baseado em versões ponderadas da média amostral e da matriz de covariâncias amostral é o denominado estimador de Stahel-Donoho. Neste caso as ponderações são obtidas utilizando uma medida de discordância associada a cada ponto amostral. Basicamente a ideia consiste em identificar um ponto amostral multivariado discordante como sendo um ponto univariado discordante nalguma projecção ortogonal.

Sejam μ e σ funcionais univariados de localização e escala respectivamente. Para um qualquer $\mathbf{y} \in I\!R^m$ defina-se a sua medida de discordância por

$$r(\mathbf{y}; \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n) = \sup_{a \in A_m} \frac{|\mathbf{a}^T\mathbf{y} - \mu(\mathbf{a}^T\mathbf{x}_1, \mathbf{a}^T\mathbf{x}_2, \cdots, \mathbf{a}^T\mathbf{x}_n)|}{\sigma(\mathbf{a}^T\mathbf{x}_1, \mathbf{a}^T\mathbf{x}_2, \cdots, \mathbf{a}^T\mathbf{x}_n)},$$

onde $A_m = \{\mathbf{a} \in I\!R^m : \|\mathbf{a}\| = 1\}$.

Definição 2.7 – Estimativas de Stahel-Donoho: *As estimativas de Stahel-Donoho de localização e dispersão multivariadas são o par* $(\mathbf{t}^*, \mathbf{C}^*)$ *tal que*

$$\begin{aligned} \mathbf{t}^* &= \frac{\sum_i^n w_i \mathbf{x}_i}{\sum_i^n w_i}, \\ \mathbf{C}^* &= \frac{\sum_i^n w_i (\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})^T}{\sum_i^n w_i}, \end{aligned} \tag{2.26}$$

onde $w_i = w(r(\mathbf{x}_i; \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n))$ *e* w *é uma função não negativa.*

Se a função $w(r)$ for contínua e limitada para $r \geq 0$ e a função $w(r)r^2$ também for limitada para valores de r não negativos e, para além disso, se os estimadores univariados tiverem ponto de rotura assintótico de 0.50, então o estimador de Stahel-Donoho tem um ponto de rotura assintótico de 0.50 para qualquer dimensão. A distribuição assintótica deste estimador permanece desconhecida, no entanto, existe a conjectura de que a sua taxa de convergência é de $n^{1/2}$, (Maronna e Yohai, 1995; Gervini, 2002; Zuo *et al.*, 2002).

A dificuldade de utilização desta classe de estimadores reside no seu cálculo numérico. Para dimensão $m = 2$ ainda se pode considerar uma grelha de pontos suficientemente fina em A_m, porém, para dimensões superiores o seu cálculo torna-se bastante complicado. Neste último caso podem usar-se algoritmos baseados em sub-amostragem o que torna o seu cálculo dispendioso.

Como inicialmente referido, o interesse recente foca-se na obtenção de estimadores com um alto ponto de rotura para amostras finitas. Estimadores com esta característica foram propostos por

Rousseeuw (1985), nomeadamente, os denotados respectivamente por MVE e MCD, sendo o primeiro o estimador do Elipsóide de Volume Mínimo e o segundo o estimador de matriz de Covariâncias de Determinante Mínimo. Uma outra classe de estimadores com esta característica é a classe dos estimadores S. Os estimadores S foram inicialmente propostos para a regressão por Rousseeuw e Yohai (1984) mas em 1987 Rousseeuw e Leroy introduziram os estimadores S para localização e dispersão. O estudo destes estimadores prosseguiu com Davies (1987), Lopuhaä (1989), He e Wang (1996) e He e Fung (2000). Como estimadores robustos multivariados com alto ponto de rotura são de difícil cálculo é usual obterem-se apenas soluções aproximadas baseadas em algoritmos de optimização combinatória.

Na subsecção seguinte serão apresentados os estimadores de localização e da matriz de dispersão multivariada numa e em mais do que uma população que possuem ponto de rotura elevado. Efectua-se ainda uma breve referência a algoritmos disponíveis para o seu cálculo.

2.3.1 Estimação numa amostra multivariada

Para além de um alto ponto de rotura outra propriedade com interesse é a equivariância para transformações afins. Nas duas definições seguintes considera-se que as observações $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ estão organizadas na matriz $\mathbf{X}$ de ordem $m \times n$.

Definição 2.8 – Equivariância afim (localização): *Um estimador de localização* $\mathbf{t} \in I\!R^m$ *diz-se equivariante para transformações afins sse para qualquer vector* $\mathbf{b} \in I\!R^m$ *e para qualquer matriz quadrada não singular de ordem* m, $\mathbf{A}$,

$$\mathbf{t}(\mathbf{AX} + \mathbf{b}) = \mathbf{At}(\mathbf{X}) + \mathbf{b}.$$

Definição 2.9 – Equivariância afim (dispersão): *Um estimador de dispersão* $\mathbf{C} \in SDP(m)$ *diz-se equivariante para transformações afins sse para qualquer vector* $\mathbf{b} \in I\!R^m$ *e para qualquer matriz quadrada não singular* $\mathbf{A}$ *de ordem* m,

$$\mathbf{C}(\mathbf{AX} + \mathbf{b}) = \mathbf{AC}(\mathbf{X})\mathbf{A}^T.$$

O conceito de equivariância afim para estimadores está para além do conceito de robustez. Convém notar que existem casos de estimadores que não são equivariantes para transformações afins mas têm bons desempenhos, como por exemplo a nova proposta de Maronna e Zamar (2002) que relaxa, em termos finitos, a equivariância afim da dispersão. O requisito de estimadores robustos que possuam esta propriedade deve-se ao facto dos parâmetros de localização e dispersão serem equivariantes para transformações afins assim como os correspondentes estimadores clássicos, $\bar{\mathbf{X}}_n$ e $\mathbf{S}_n$. Deste modo, é desejável que estimadores alternativos de $\boldsymbol{\theta}$ verifiquem esta propriedade.

Os estimadores que se apresentam seguidamente têm um alto ponto de rotura e para além disso são afinamente equivariantes. Denote-se um qualquer elipsóide por

$$E(\mathbf{t}, \mathbf{C}, c) = \left\{\mathbf{x} : DM^2\left(\mathbf{x}; \mathbf{t}, \mathbf{C}\right) \leq c^2\right\}.$$

O volume do elipsóide que contém h pontos, $E(\mathbf{t}, \mathbf{C}, c_0)$, é dado por

$$V = (\det(\mathbf{C})c_0^2)^{1/2},$$

onde $c_0^2 = DM_h^2$ corresponde à *h-ésima* distância quadrada de Mahalanobis ordenada.

Rousseeuw (1985) introduziu o estimador do elipsóide de volume mínimo, MVE, como uma alternativa robusta aos estimadores clássicos de $\boldsymbol{\theta}$. Encontrar estimativas MVE corresponde a encontrar o elipsóide que minimiza o volume V sujeito a que contenha um certo número de pontos. Em termos geométricos o estimador MVE de localização corresponde ao centro desse elipsóide e o estimador de dispersão não é mais do que a matriz definidora da forma do elipsóide.

Definição 2.10 – Estimativas de Elipsóide de Volume Mínimo (MVE): *Sendo* $\mathbf{t}^*$ *e* $\mathbf{C}^*$ *respectivamente o vector e matriz que determinam o elipsóide de menor volume que cobre exactamente* h *observações, para* $[(n+m)/2]+1 \leq h < n$*, então as estimativas MVE de* $\boldsymbol{\mu}$ *e* $\boldsymbol{\Sigma}$ *são definidas por*

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{MVE} &= \mathbf{t}^*, \\ \hat{\boldsymbol{\Sigma}}_{MVE} &= f_m \mathbf{C}^*, \end{aligned} \tag{2.27}$$

onde f_m *é um factor de consistência para o modelo normal.*

Em geral escolhem-se

$$h = [n/2] + 1,$$

e

$$f_m = \left\{ \frac{med\{DM^2(\mathbf{x}; \mathbf{t}^*, \mathbf{C}^*)\}}{F_{\chi_m^2}^{-1}(0.50)} \right\},$$

onde $F_{\chi_m^2}$ representa a função de distribuição de um qui-quadrado com m graus de liberdade.

Significa então que o cálculo exacto de uma estimativa MVE envolve a pesquisa de todos os elipsóides que contêm h pontos, porém essa pesquisa torna-se impraticável para dimensões médias ou grandes. Por esta razão Rousseeuw e Leroy (1987) propuseram um algoritmo de reamostragem para aproximar o MVE, que vai considerando uma série de subconjuntos de $m+1$ observações calculando a sua média e matriz de covariâncias. O volume do elipsóide equivalente é então aumentado ou diminuído até conter exactamente h observações. Este procedimento é repetido um grande número de vezes até se obter a solução que corresponde ao de menor volume que contém os h pontos. Para amostras de pequenas dimensões é possível gerar todos os subconjuntos de cardinal $m+1$ mas em médias ou grandes dimensões os subconjuntos são escolhidos aleatoriamente. Outros algoritmos foram propostos para o cálculo do MVE, nomeadamente um algoritmo genético (Burns, 1992) presente no software estatístico S-PLUS (S-PLUS, 1999). Um algoritmo exacto com boas propriedades de rapidez e de aplicação em dimensões médias/grandes ($n \leq 100$ e $m \leq 5$) baseado em pesquisa *branch-and-bound* foi proposto por Agulló (1996). Este procedimento selecciona o subconjunto óptimo sem necessidade da pesquisa de todos os subconjuntos de dimensão h.

O ponto de rotura do MVE é essencialmente $(n-h)/n$, e quando se fixa $h = [n/2]+1$ esse converge para 0.5 ($n \to \infty$). Significa então que as estimativas MVE só se poderão degradar se mais de metade das observações forem alteradas para posições arbitrárias.

As propriedades assintóticas dos estimadores com alto ponto de rotura são geralmente descritas em termos do seu comportamento quando aplicados a amostras aleatórias provenientes de uma população normal multivariada. Apesar do alto ponto de rotura o estimador MVE não possui "boas" propriedades assintóticas, possuindo uma taxa de convergência inferior à usual (Davies, 1992). Esta razão leva a que se dê preferência à utilização do MCD, já que Butler, Davies e Jhun (1993) mostraram que este estimador é assintoticamente Normal com taxa de convergência usual, tornando-se então mais eficiente que o MVE.

Encontrar estimativas MCD de localização e dispersão corresponde a procurar h pontos para os quais o determinante da matriz de covariâncias clássica é mínimo.

Definição 2.11 – Estimativas de Matriz de Covariâncias de Estimativas de Determinante Mínimo (MCD): *Seja um qualquer subconjunto de* $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ *de cardinal* h*, com* $[(n+m)/2]+1 \leq h < n$*. O estimador MCD obtém-se seleccionando o subconjunto* $B_h = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \cdots, \mathbf{x}_{i_h}\}$ *para o qual a matriz de covariâncias clássica tem o menor determinante. Assim as estimativas MCD são definidas por*

$$\hat{\boldsymbol{\mu}}_{MCD} = \frac{1}{h}\sum_{j=1}^{h} \mathbf{x}_{i_j}$$

$$\hat{\boldsymbol{\Sigma}}_{MCD} = f_m \frac{1}{h}\sum_{j=1}^{h} \left(\mathbf{x}_{i_j} - \hat{\boldsymbol{\mu}}_{MCD}\right)\left(\mathbf{x}_{i_j} - \hat{\boldsymbol{\mu}}_{MCD}\right)^T, \tag{2.28}$$

onde f_m *é um factor de consistência.*

O factor de consistência, f_m, pode ser escolhido de forma idêntica ao apresentado para o MVE e para h escolhe-se, em geral, o valor de $h = [(n+m)/2]+1 \approx 0.5n$. Esta escolha conduz ao mais alto ponto de rotura assintótico possível, 0.50.

A taxa de convergência do MVE é de $O(n^{-1/3})$ enquanto que o MCD possui uma taxa de convergência usual, isto é de $O(n^{-1/2})$. Significa então que o MVE tem uma eficiência assintótica de zero, ou seja, para amostras de grandes dimensões o seu desempenho é arbitrariamente inferior ao do MCD.

A função de influência do MCD assim como a variância assintótica dos elementos da matriz de dispersão foram obtidas por Croux e Haesbroeck (1999). Estes autores verificaram que a função de influência do MCD é limitada, característica requerida para um estimador B–robusto. Apesar das vantagens do MCD este estimador raramente era utilizado devido à sua dificuldade de cálculo. Porém, Rousseeuw e Van Driessen (1999) propuseram um algoritmo, FAST-MCD, que em geral é mais rápido que os algoritmos existentes anteriormente, inclusive em grandes dimensões. Este algoritmo encontra-se implementado no *software* estatístico S-PLUS.

Observe-se que, quer o MVE quer o MCD, possuem a propriedade de separação de um elipsóide. Ambos definem um elipsóide tal que os h pontos estão sobre ou dentro dele, enquanto que os restantes $n-h$ pontos estão sobre ou fora do elipsóide. Quaisquer observações discordantes deverão estar entre os pontos não cobertos e afastados do centro do elipsóide. Refira-se porém que a forma do elipsóide para o MVE poderá não ser a do elipsóide do MCD, mesmo tendo considerado o mesmo número de pontos. E que também não há razão para que os dois elipsóides cubram necessariamente o mesmo conjunto de pontos.

Por último, introduzir-se-á aqui os denominados estimadores S. Estes foram inicialmente propostos num contexto de regressão por Rousseeuw e Yohai (1984). Mais tarde Rousseeuw e Leroy (1987) propuseram os estimadores S de localização e escala multivariados. Antes de se definirem estes estimadores é necessário introduzir a seguinte função.

Considere-se a função $\rho : I\!R \to [0, \infty)$ que satisfaz as seguintes propriedades:

Q.1 ρ é simétrica, possui derivada contínua ψ e $\rho(0) = 0$.

Q.2 Existe uma constante positiva, $c_0 > 0$, tal que ρ é estritamente crescente no intervalo $[0, c_0]$ e constante em $[c_0, \infty)$.

Definição 2.12 – **Estimativas S**: *Considere-se uma função ρ satisfazendo as propriedades Q.1 e Q.2. As estimativas S de $\boldsymbol{\theta}$, $(\hat{\boldsymbol{\mu}}_S, \hat{\boldsymbol{\Sigma}}_S)$, são solução do seguinte problema de minimização*

$$\begin{aligned} & \textit{arg min} \det(\mathbf{C}) \\ \textit{sujeito a} \quad & \\ & n^{-1} \sum_{i=1}^{n} \rho(DM(\mathbf{x}_{i;}\mathbf{t}, \mathbf{C})) = b_0, \end{aligned}$$

para qualquer par $(\mathbf{t}, \mathbf{C}) \in \Theta$, onde $0 < b_0 < \sup(\rho)$ é uma constante previamente fixada.

Outra definição equivalente para o problema de minimização anterior é a descrita na definição seguinte.

Definição 2.13 – **Estimativas S:** *Dado o par $(\mathbf{t}, \mathbf{C})$, onde $\mathbf{C} \in SDP(m)$. Se $s \equiv s(\mathbf{t}, \mathbf{C})$ é solução de*

$$n^{-1} \sum_{i=1}^{n} \rho(DM(\mathbf{x}_{i;}\mathbf{t}, \mathbf{C})/s) = b_0$$

e $(\mathbf{t}^*, \mathbf{C}^*)$ minimiza $s(\mathbf{t}, \mathbf{C})$ sujeito ao $\det(\mathbf{C}) = 1$, então

$$\begin{aligned} \hat{\boldsymbol{\mu}}_S &= \mathbf{t}^*, \\ \hat{\boldsymbol{\Sigma}}_S &= s^2(\mathbf{t}^*, \mathbf{C}^*)\mathbf{C}^*. \end{aligned}$$

Esta última definição parece reflectir a existência de alguma relação entre os estimadores S e os estimadores–M. De facto, Lopuhaä (1989) mostrou a existência dessa relação no caso da estimação

multivariada do parâmetro $\boldsymbol{\theta}$, verificando que $(\hat{\boldsymbol{\mu}}_S, \hat{\boldsymbol{\Sigma}}_S)$ é solução das seguintes equações que definem um estimador–M,

$$\begin{aligned} \hat{\boldsymbol{\mu}}_S &= \frac{\sum_{i=1}^{n} h(DM(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_S, \hat{\boldsymbol{\Sigma}}_S))\mathbf{x}_i}{\sum_{i=1}^{n} h(DM(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_S, \hat{\boldsymbol{\Sigma}}_S))}, \\ \hat{\boldsymbol{\Sigma}}_S &= \delta \sum_{i=1}^{n} h(DM(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_S, \hat{\boldsymbol{\Sigma}}_S))(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_S)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_S)^T, \end{aligned} \tag{2.29}$$

onde $h(u) = \psi(u)/u$, $\psi = \rho'$ e δ é um escalar que depende do par $(\hat{\boldsymbol{\mu}}_S, \hat{\boldsymbol{\Sigma}}_S)$ e dos dados.

Um exemplo de uma função ρ que goza das propriedades Q.1 e Q.2 é a denominada função biponderada de Tukey dada por

$$\rho_B(u, c_0) = \begin{cases} \dfrac{u^2}{2} - \dfrac{u^4}{2c_0^2} + \dfrac{u^6}{6c_0^4} & se \quad |u| \leq c_0, \\ \\ \dfrac{c_0^2}{6} & se \quad |u| \geq c_0. \end{cases} \tag{2.30}$$

A sua derivada, ψ, é também conhecida como a função *psi* biponderada de Tukey e não é mais do que

$$\psi_B(u, c_0) = u(1 - (u/c_0)^2)^2 I_{[-c_0, c_0]}(u).$$

A constante b_0 deve ser escolhida de forma a obter estimadores consistentes à Fisher sob o modelo central. Se este pertencer à família das distribuições elipticamente simétricas (2.23) então $b_0 = E_{\mathbf{0},\mathbf{I}}(\rho(\|\mathbf{x}\|))$. Por outro lado, a constante c_0, associada à função ρ, pode ser escolhida de tal forma que $0 < b_0/\sup(\rho) = r \leq (n-m)/2n$, o qual conduz a um ponto de rotura (para amostras finitas) de $\epsilon^*{}_H = \lceil nr \rceil / n$ (Lopuhaä e Rousseeuw, 1991) e assintoticamente de $\min(b_0/\rho(c_0), 1 - b_0/\rho(c_0))$. Para $r = (n-m)/2n$ obtém-se o ponto de rotura máximo $[(n-m+1)/2]/n$, o que assintoticamente se traduz em 0.50. Assim, para se conseguir um ponto de rotura de υ bastará resolver a seguinte equação em ordem a c_0, $b_0 = E_{\mathbf{0},\mathbf{I}}(\rho(\|\mathbf{x}\|)) = \upsilon\rho(\infty)$. Os estimadores S para além de serem estimadores com um alto ponto de rotura são assintoticamente Normais e possuem a taxa de convergência usual, $O(n^{1/2})$.

Apesar das suas boas propriedades esta classe de estimadores é raramente usada já que, contrariamente aos estimadores MVE e MCD não existe implementação computacional no *software* estatístico de uso corrente, nomeadamente S-PLUS, SAS ou SPSS. Existe, no entanto, um algoritmo bastante eficaz para o seu cálculo, denominado SURREAL (Ruppert, 1992). Este algoritmo apesar de ser de sub-amostragem possui vários melhoramentos que permitem que o cálculo de estimativas S seja menos dispendioso. O método de Ruppert baseia-se em direcções de pesquisa aleatórias, que permitem acelerar a convergência para um mínimo global e também numa melhoria local que é fundamentada nas equações (2.29). Para o estudo de simulação que irá ser apresentado no Capítulo 4 foi implementada uma versão baseada neste algoritmo. O programa respectivo foi escrito em linguagem S (S-PLUS) e pode consultar-se no Apêndice C.

De um modo geral a determinação da constante $b_0 = E_{\mathbf{0},\mathbf{I}}(\rho(\|\mathbf{x}\|))$ baseia-se na distribuição normal multivariada. A proposição seguinte mostra como este cálculo poderá ser efectuado em qualquer *software* estatístico sem necessidade de cálculo integral explícito. A demonstração da proposição necessita do seguinte lema.

Lema 2.1 *(Lopuhaä, 1998): Seja $z : [0, \infty) \to I\!R$ e $\mathbf{x} = (x_1, x_2, \cdots, x_k)^T$. Então*

$$\int_{I\!R^k} z(\mathbf{x}^T\mathbf{x})d\mathbf{x} = \frac{2\pi^{k/2}}{\Gamma(k/2)} \int_0^\infty z(r^2)r^{k-1}dr.$$

Proposição 2.1 *Seja $\mathbf{x} = (x_1, x_2, \cdots, x_m)^T$ um vector aleatório m-dimensional com distribuição normal multivariada, $F_0 \equiv \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, de vector médio $\mathbf{0}$ e matriz de covariâncias $\mathbf{I}_m$. Considere-se ainda que ρ_B é a função ρ biponderada de Tukey definida em (2.30). Então*

$$\begin{aligned} b_0 = E_{F_0}[\rho_B(\|\mathbf{x}\|)] =& \frac{m}{2} F_{\chi^2_{(m+2)}}(c_0^2) - \frac{(m+2)m}{2c_0^2} \times F_{\chi^2_{(m+4)}}(c_0^2) \\ &+ \frac{(m+4)(m+2)m}{6c_0^4} \times F_{\chi^2_{(m+6)}}(c_0^2) + c_0^2/6 \times (1 - F_{\chi^2_{(m)}}(c_0^2)), \end{aligned} \tag{2.31}$$

onde $F_{\chi^2_l}(.)$ representa a função de distribuição de um qui-quadrado central com l graus de liberdade.

Demonstração: Ver apêndice A.

Dado que a eficiência associada a estimadores com alto ponto de rotura, sob o modelo central Normal, pode ser bastante baixa, é muitas vezes recomendado o uso de versões reponderadas desses estimadores. Essas versões mantêm o ponto de rotura do estimador inicial e normalmente apresentam maior eficiência. Lopuhaä (1999) deduziu as propriedades assintóticas dos estimadores reponderados de localização e escala multivariados.

Definição 2.14 – Estimativas reponderadas a um passo: *Dados $(\mathbf{t}^0, \mathbf{C}^0)$ estimadores iniciais do parâmetro $\boldsymbol{\theta}$ e $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ uma amostra de um vector aleatório m-dimensional, as estimativas reponderadas a um passo são dadas por*

$$\begin{aligned} \mathbf{t}^1 &= \frac{\sum_{i=1}^n w_i \mathbf{x}_i}{\sum_{i=1}^n w_i}, \\ \mathbf{C}^1 &= f_m \frac{\sum_{i=1}^n w_i (\mathbf{x}_i - \mathbf{t}^1)(\mathbf{x}_i - \mathbf{t}^1)^T}{\sum_{i=1}^n w_i}, \end{aligned} \tag{2.32}$$

onde $w_i = w\left\{(\mathbf{x}_i - \mathbf{t}^0)^T (\mathbf{C}^0)^{-1} (\mathbf{x}_i - \mathbf{t}^0)\right\}$, com w uma função de peso adequada, $w : [0, +\infty) \to I\!R$ e f_m um factor de consistência.

De um modo geral escolhe-se para w a função

$$w(u) = I_{[0,q_\alpha]}(u) \qquad \text{com} \qquad q_\alpha = G^{-1}(1-\alpha), \tag{2.33}$$

onde $G(y)$ corresponde à função de distribuição de $\|\mathbf{x}\|^2$ no ponto y. No caso particular de um modelo Normal m-multivariado ter-se-á $q_\alpha = F^{-1}_{\chi^2_m}(1-\alpha)$. Quanto ao factor de consistência ele é dado por

$$f_m = (1-\alpha)\left\{\frac{\pi^{m/2}}{\Gamma((m/2)+1)}\int_0^{\sqrt{q_\alpha}} r^{m+1}g(r^2)dr\right\}^{-1}, \quad \text{onde} \quad g = G'.$$

Como nota final refira-se ainda a proposta apresentada por Woodruff e Rocke (1994). Estes autores propõem a estimação robusta multivariada da localização e forma através de estimadores compostos. Segundo os autores, usando pontos iniciais, tal como o MVE e o MCD exactos, no cálculo de um estimador S obtém-se um estimador que possui as mesmas propriedades que o estimador S exacto. O algoritmo para o seu cálculo utiliza métodos de partição o que permite o cálculo em grandes dimensões. Posteriormente Rocke e Woodruff (1997) apresentaram a estratégia global, em forma de algoritmo, para a estimação desses estimadores compostos.

2.3.2 Estimação em várias amostras multivariadas

Vários tópicos de análise multivariada requerem a estimação de parâmetros de localização e dispersão em mais do que uma população com uma matriz de covariâncias comum. Um dos exemplos típicos é a análise discriminante. O uso de estimadores com um ponto de rotura elevado nesta situação não é frequente e basicamente o que se conhece até à data são os estudos de Todorov *et al.* (1990), Constable e Tauxe (1992) assim como o de He e Fung (2000). O método dos primeiros autores baseia-se na estimação por MVE e MCD dos parâmetros de cada população, sendo a matriz de covariâncias comum estimada através da combinação das várias estimativas individuais. A ideia de He e Fung é também a de combinação, porém a abordagem é combinar as observações de cada população numa só e apenas depois estimar a matriz de covariâncias comum. Estes autores utilizaram estimadores S quer na sua versão usual quer reponderada.

Suponha-se que temos g amostras aleatórias de dimensão n_i, com $i = 1, 2\cdots, g$, provenientes de g populações m-dimensionais de localizações distintas mas com matriz de covariâncias comum. O objectivo a atingir é a estimação dos parâmetros $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$.

He e Fung no seu artigo de 2000 consideraram duas versões de estimadores S. O primeiro método, designemo-lo por A, baseia-se simplesmente em calcular estimadores S de $\boldsymbol{\mu}_i$, $\hat{\boldsymbol{\mu}}_i^S$, para cada $i = 1, 2, \cdots, g$ e em seguida combinar numa só amostra as g amostras centradas nessas estimativas. Obtém-se então uma amostra cuja dimensão é agora de $\sum_{i=1}^{g} n_i$ e a qual servirá para estimar o par $(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\Sigma}}_S)$ usando um estimador S. Após este cálculo poderão ser efectuadas actualizações das estimativas de localização e novamente centrar e combinar as g amostras. Este procedimento pode ser repetido até obter convergência. Cada passo requer a estimação S, numa amostra, $g+1$ vezes.

A segunda versão para o cálculo de estimativas de $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ é uma generalização do problema de minimização definido em (2.12) e que se vai designar por método B. Considerando a amostra combinada (mas não centrada) o estimador S, $(\boldsymbol{\mu}_1^S, \boldsymbol{\mu}_2^S, \cdots, \boldsymbol{\mu}_g^S, \boldsymbol{\Sigma}^S)$, é obtido minimizando $s \equiv s(\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_g, \mathbf{C})$ para $\mathbf{t}_i \in I\!R^m$ e $\mathbf{C} \in SDP(m)$ com $\det(\mathbf{C}) = 1$ sujeito a

$$\left(\sum_{i=1}^{g} n_i\right)^{-1} \sum_{i=1}^{g} \sum_{j=1}^{n_i} \rho(DM(\mathbf{x}_{ij}; \mathbf{t}_i, \mathbf{C})/s) = b_0,$$

onde ρ é uma função que satisfaz as propriedades Q.1 e Q.2.

Ambas as versões, A e B, podem ser calculadas usando o algoritmo SURREAL de Ruppert. O ponto de rotura para amostras finitas de cada um dos estimadores obtidos por cada um destes dois métodos é igual a

$$\min\left\{\frac{[n_i/2] - m + 1}{n_i}, \quad i = 1, 2, \cdots, g\right\}.$$

Quando $n_i \to \infty, \ \forall i$, obtém-se o ponto de rotura assintótico de 0.50.

Dada a importância destes estimadores na análise multivariada e não existindo um estudo que envolve estes estimadores nos moldes que serão necessários para aplicações neste trabalho vai apresentar-se no Capítulo 4 um estudo de simulação para comparar o seu desempenho em amostras finitas.

2.4 Inferência robusta não pontual

Apesar da palavra "robustez" ter sido usada pela primeira vez em estatística em relação a testes de hipóteses, Box (1953), verifica-se que a investigação se tem concentrado principalmente na estimação pontual. Entende-se facilmente este facto, pois sendo a estimação pontual já tão complexa é natural que a complexidade aumente quando se pretende obter intervalos de confiança ou testes de hipóteses robustos.

A robustez para testes de hipóteses está relacionada com a estabilidade do nível de significância e da função potência. Para os intervalos de confiança esta noção traduzir-se-á na estabilidade da probabilidade de cobertura e do comprimento do intervalo. Os exemplos seguintes pretendem ilustrar a falta de robustez de alguns métodos geralmente utilizados na prática.

Exemplo 2.1 Testes t e Wilcoxon. Considerem-se duas amostras aleatórias independentes, $(X_{11}, X_{12}, \cdots, X_{1n_1})$ e $(X_{21}, X_{22}, \cdots, X_{2n_2})$, provenientes cada uma da população $F_i(\mu_i, \sigma_i)$, $i = 1, 2$, onde μ_i e σ_i, representam respectivamente os parâmetros de localização e escala da população i. Com base numa concretização de cada uma das amostras aleatórias pretende-se verificar se os parâmetros de localização de cada uma dessas populações podem ser considerados iguais. Para simplificar o problema assuma-se que os modelos geradores de cada uma das amostras, F_1 e F_2, têm idêntica escala ($\sigma_1 = \sigma_2$) podendo apenas diferir nas localizações. Nestas condições dois testes adequados para testar a hipótese nula, $H_0 : \mu_1 = \mu_2$, são os denominados teste-t e o teste de

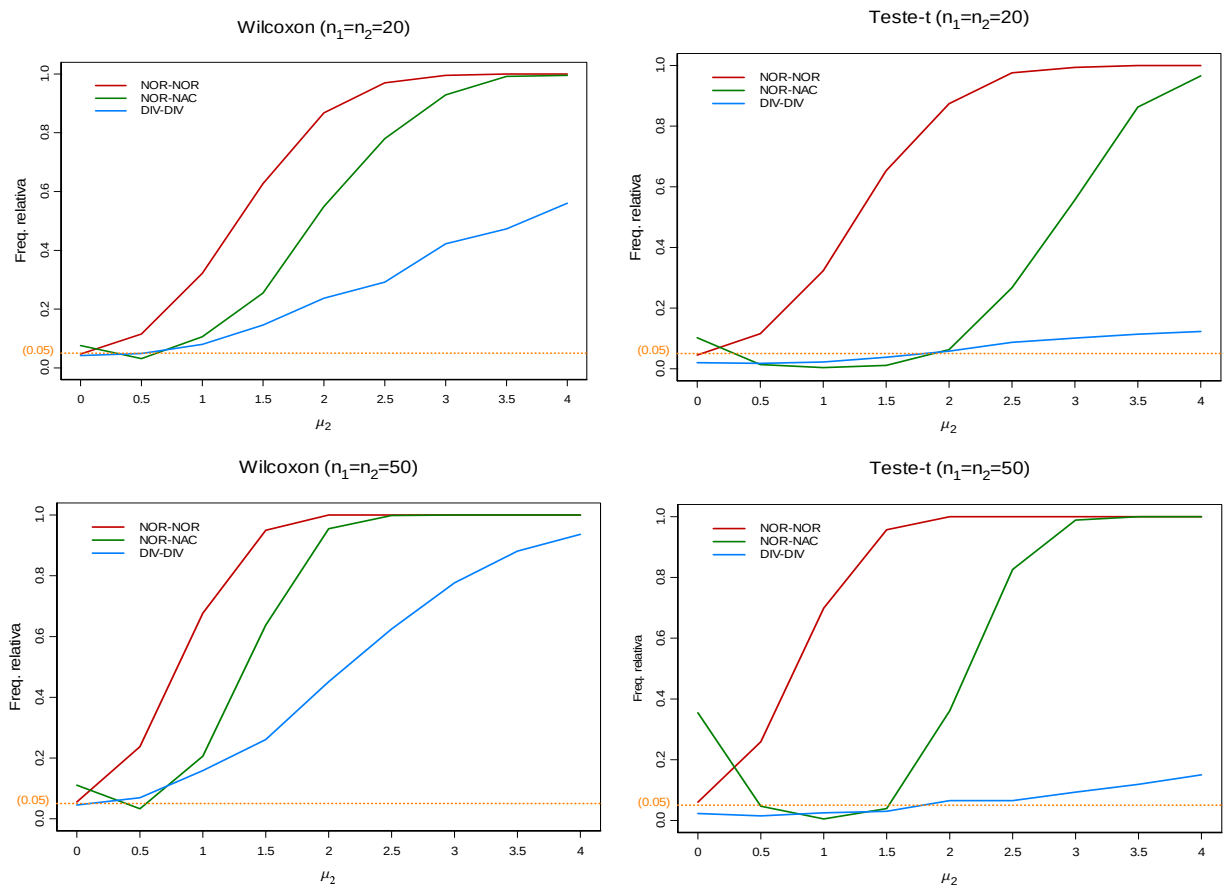


Figura 2.1: Frequência relativa do acontecimento "*p-value*<0.05" em 1000 simulações para a hipótese $H_0 : \mu_1 = \mu_2$, variando o valor médio da segunda população ($\mu_2 = 0.0$, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5 e 4.0).

Wilcoxon (Conover, 1980). Enquanto que o primeiro teste requer que as observações provenham de populações com distribuição Normal (e $\sigma_1 = \sigma_2$), o segundo teste não necessita dessa suposição (apenas de $F_1 = F_2$ e $\sigma_1 = \sigma_2$). Tendo como objectivo a avaliação do comportamento destes dois testes face a alguns desvios do modelo Normal procedeu-se a um pequeno estudo de simulação. Consideram-se então as seguintes distribuições:

- NOR-NOR: duas populações Normais com $F_1 \equiv N\left(0, \sigma^2\right)$ e $F_2 \equiv N\left(\mu_2, \sigma^2\right)$;
- NOR-NAC: a população um tem distribuição Normal, $F_1 \equiv N\left(0, \sigma^2\right)$ e a população dois tem distribuição Normal Assimetricamente ε-Contaminada com probabilidade $\varepsilon = 0.1$ à volta do ponto 10, isto é, $F_2 \equiv (1-\varepsilon)N\left(0, \sigma^2\right) + \varepsilon N\left(10, \sigma^2 = (0.01)^2\right)$;
- DIV-DIV: duas populações cujas distribuições são a Dividida (ou seja é a distribuição do quociente de uma Normal padrão por uma variável aleatória independente com distribuição uniforme no intervalo $[0, 1]$), às quais se aplica posteriormente uma transformação para que ambas tenham escala σ e parâmetros de localização $\mu_1 = 0$ para a primeira população e μ_2 para a segunda.

Fixou-se $\sigma = 2$, sem perda de generalidade, e variou-se o valor médio da segunda população $\mu_2 = 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5$ e 4.0. Para as dimensões n_1 e n_2 optou-se pelos seguintes valores: (i) $n_1 = n_2 = 20$ e (ii) $n_1 = n_2 = 50$. Geraram-se 1000 amostras de cada uma distribuições referidas para cada combinação (dimensão, parâmetro de localização μ) e avaliou-se a frequência relativa do acontecimento "*p-value*<0.05" nessas 1000 réplicas para a hipótese $H_0 : \mu_1 = \mu_2$.

Os resultados obtidos apresentam-se graficamente na Figura 2.1. Para a distribuição Normal o teste-t é óptimo e os resultados assim o reflectem. Por outro lado, para populações não Normais, é evidente a falta de robustez desse teste, quer relativa à estabilidade do nível de significância, quer da sua função potência. Para distribuições não Normais, e apesar dos resultados obtidos para o teste de Wilcoxon serem um pouco melhores que para o teste-t, verifica-se que este também não é uma boa alternativa pois a sua potência é bastante reduzida. ■

Exemplo 2.2 Intervalo de confiança para a localização univariada. Seja $(X_1, X_2, \cdots, X_n)$ uma amostra aleatória de dimensão n proveniente da população $F(\mu_i, \sigma_i)$, $i = 1, 2$, onde μ e σ são os parâmetros de localização e escala, respectivamente. Para comparar o desempenho, em termos da estabilidade e probabilidade de cobertura, da abordagem clássica e de uma possível abordagem robusta, realizou-se o estudo que se descreve em seguida. Geraram-se 1000 amostras de cada uma das seguintes distribuições:

- NOR: $N\left(0, \sigma^2 = 4\right)$;
- NAC: Normal Assimetricamente ε-Contaminada com probabilidade $\varepsilon = 0.1$ à volta do ponto 10, isto é, $(1-\varepsilon)N\left(0, \sigma^2 = 4\right) + \varepsilon N\left(10, \sigma^2 = (0.01)^2\right)$;
- DIV: distribuição dividida com $\mu = 0$ e $\sigma^2 = 4$.

Para cada uma dessas amostras, de dimensões $n = 20$ ou $n = 50$, construíram-se os intervalos de confiança bilaterais a $(1-\alpha) \times 100\% = 95\%$ para μ através de cada um dos seguintes métodos:

- Clássico: $\bar{X} \pm F^{-1}_{t(n-1)}\left(1 - \alpha/2\right) \frac{S}{\sqrt{n}}$,
- MedMad: $\hat{\mu} \pm F^{-1}_{t(n-1)}\left(1 - \alpha/2\right) \frac{\hat{\sigma}}{\sqrt{n}}$,

onde $F^{-1}_{t(n-1)}\left(1 - \alpha/2\right)$ corresponde ao quantil de probabilidade $(1 - \alpha/2)$ de uma distribuição $t - Student$ com $n - 1$ graus de liberdade, $\bar{X}_n$ e S_{n-1} correspondem, respectivamente, à média e desvio padrão da amostra aleatória, ou seja aos estimadores clássicos para μ e σ em populações Normais. O intervalo obtido pelo método MedMad não é mais que o intervalo clássico quando os respectivos estimadores clássicos são substituídos por dois estimadores robustos. Considerou-se para $\hat{\mu}$ a Mediana (Med) da amostra aleatória e para $\hat{\sigma}$ o estimador MAD, desvio mediano absoluto (Hoaglin *et al.*, 1992). Observe-se que o método MedMad é uma abordagem muito ingénua para a construção de intervalos de confiança robustos, já que a variável fulcral envolvida $(Med - \mu) / (MAD/\sqrt{n})$ não tem distribuição $t - Student$ com $n - 1$ graus de liberdade.

As Tabelas 2.1 e 2.2 apresentam, para cada combinação (distribuição, método, dimensão), as seguintes medidas:

- probabilidade estimada de não cobertura, isto é, contabiliza-se o número de vezes que $\mu = 0$ não pertence ao intervalo construído e divide-se este valor por 1000;
- média e o desvio padrão dos comprimentos dos 1000 intervalos construídos.

Tabela 2.1: Estimativas relativas aos comprimentos e probabilidades de não cobertura de intervalos de confiança bilaterais para o valor médio de uma população F, com base em amostras de dimensão $n = 20$.

Distribuição amostral	Método	(alvo: 0.050) Não Cobertura	Compr. do intervalo Média	Compr. do intervalo Desvio padrão
NOR	Clássico	0.054	1.8715	0.3069
	MedMad	0.136	1.8183	0.4843
NAC	Clássico	0.081	3.3745	0.1957
	MedMad	0.136	2.0427	0.5275
DIV	Clássico	0.027	56.6003	371.0930
	MedMad	0.079	4.2020	1.4121

Tabela 2.2: Estimativas relativas aos comprimentos e probabilidades de não cobertura de intervalos de confiança bilaterais para o valor médio de uma população F, com base em amostras de dimensão $n = 50$.

Distribuição amostral	Método	(alvo: 0.050) Não Cobertura	Compr. do intervalo Média	Compr. do intervalo Desvio padrão
NOR	Clássico	0.049	1.1316	0.1138
	MedMad	0.115	1.1141	0.1856
NAC	Clássico	0.485	2.0309	0.0719
	MedMad	0.162	1.2887	0.2080
DIV	Clássico	0.024	36.5366	128.2554
	MedMad	0.087	2.4844	0.5024

Em relação aos resultados das Tabelas 2.1 e 2.2 podem fazer-se as seguintes observações: o intervalo obtido pelo método clássico tem, como era de esperar, comportamento óptimo sob a distribuição Normal, no entanto, para as restantes distribuições, e apesar de se observar ainda alguma estabilidade no nível de confiança, é em relação ao comprimento do intervalo que se nota a sua falta de eficiência. Face ao exposto atrás acerca da utilização de quantis de uma t para a construção do intervalo MedMad, os desempenhos verificados relativamente à probabilidade estimada de cobertura

são esperados, verifica-se, no entanto, uma melhoria dos resultados nos comprimentos médios face ao intervalo clássico principalmente quando existe contaminação. ■

Com os Exemplos 2.1 e 2.2 fica patente a necessidade de procedimentos estatísticos que providenciem boas respostas quando há afastamentos dos modelos geradores dos dados à distribuição central (nos exemplos anteriores essa distribuição era a Normal). Na secção seguinte descrevem-se algumas das propostas robustas apresentadas.

2.4.1 Testes de hipóteses robustos

Vários autores têm proposto testes robustos ou pelo menos com várias propriedades desejáveis de robustez. De um modo geral existem dois tipos de abordagens, uma baseada em funções de influência (Hampel *et al.*, 1986) e a outra muito menos difundida assente em análises de vizinhanças de "encolhimento" (*shrinking neighborhoods*, Rieder (1978)). Relativamente à primeira abordagem referências gerais são dadas em Hampel *et al.* (1986) e Rousseeuw e Leroy (1987), para além dessas vejam-se também os trabalhos de Lambert (1982) e Ronchetti (1997).

Esta secção tem por objectivo dar uma noção sobre alguns testes robustos que têm sido propostos na literatura, assim como das dificuldades inerentes a este tópico sem, no entanto, entrar em detalhes e considerando apenas para o efeito a primeira abordagem referida acima.

Note-se ainda que os testes para os modelos lineares são os que têm tido maior divulgação na literatura. Veja-se, por exemplo, a classe de testes τ proposta por Ronchetti (1982), os testes propostos por Markatou e Hettmansperger (1990) e ainda os descritos por Markatou *et al.* (1991).

Testes de hipóteses em modelos paramétricos univariados

Um dos conceitos importantes presente na teoria dos testes de hipóteses é a denominada função potência assintótica. Esta função, para além de revelar o comportamento de uma estatística de teste para amostras de grandes dimensões, também é uma noção importante que permite comparar a eficiência relativa de dois testes. Para além disso, revela também os efeitos no nível e na potência do teste provocados por certos afastamentos da distribuição subjacente ao modelo gerador dos dados.

Assim, considere-se um modelo paramétrico, F_θ, onde $\theta \in \Theta \subseteq I\!R$, e uma amostra aleatória de dimensão n, $(X_1, X_2, \cdots, X_n)$, proveniente de F_θ. Suponha-se agora que se pretende testar a hipótese nula $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ aplicando a estatística de teste, T_n. Suponha-se ainda que a estatística T_n pode ser expressa por um funcional $T = T(F_n(x))$.

A hipótese H_0 é rejeitada a favor de H_1 quando a estatística de teste T_n exceder um ponto crítico:

$$T_n = T_n(X_1, X_2, \cdots, X_n) \geq k_n$$

O ponto crítico, k_n, é escolhido de tal modo que o tamanho do teste não ultrapasse um determinado nível, $\alpha : P_{\theta_0}(T_n \geq k_n) \leq \alpha$.

A função potência, dada por

$$\Pi_n(\theta) = P_\theta(T_n \geq k_n), \quad \theta > \theta_0,$$

pode ser alargada de modo a incluir $\theta = \theta_0$. Esta função, $\Pi_n(\theta)$, contém a informação necessária para averiguar o desempenho da sequência de estatísticas $\{T_n : n = 1, 2, \cdots\}$. $\Pi_n(\theta)$ depende da dimensão da amostra, no entanto, esta dependência pode ser removida considerando, por exemplo, alternativas da forma $\theta_n = \theta_0 + \delta n^{-1/2}$, onde $\delta > 0$. Define-se então função potência assintótica das estatísticas de teste $\{T_n\}$ para as alternativas $\{\theta_n\}$ como

$$\beta = \lim_{n\to\infty} (\Pi_n(\theta_n)).$$

Sob certas condições de regularidade, que incluem a Normalidade assintótica de T_n (quer sob H_0 quer sob H_1) o limite anterior pode ser calculado por

$$\beta = 1 - \Phi(\Phi^{-1}(1 - \alpha_0) - \delta\sqrt{E_T}),$$

onde

$$E_T = \lim_{n\to\infty} [\xi_n'(\theta_0)]^2 / \left(nVar_{F_{\theta_0}}(T_n)\right) = [\xi'(\theta_0)]^2 / VA(T, F_{\theta_0})$$

é a eficácia assintótica de Pitman do teste, $\xi_n(\theta) = E\left[T_n|\theta\right]$, $\xi(\theta) \equiv T(F_\theta)$,

$$VA(T, F_{\theta_0}) = \int IF\left(x; T, F_{\theta_0}\right)^2 dF_{\theta_0}(x)$$

designa a variância assintótica de T, α_0 é o nível assintótico nominal e $\Phi^{-1}\left(1 - \alpha_0\right)$ corresponde ao quantil de probabilidade $1 - \alpha_0$ de uma distribuição Normal padrão.

Mais detalhes e exemplos do cálculo da potência assintótica podem ser encontrados em Staudte e Sheather (1990, Cap. 5, 6) e Hampel *et al.* (1986, Cap. 3).

Rousseeuw e Ronchetti (1981) generalizaram o conceito de função de influência a testes de parâmetros unidimensionais com o propósito de investigar a robustez local de estatísticas de teste. Para além de outras propriedades, a função de influência para testes permite aproximar o nível assintótico e a potência assintótica sob distribuições contaminadas numa vizinhança das hipóteses.

Considere-se a seguinte vizinhança de contaminação-ε

$$\mathcal{V}_{\varepsilon,n}(F_{\theta_0}) = \{H : H = (1 - \varepsilon/\sqrt{n})F_{\theta_0} + (\varepsilon/\sqrt{n})H^*, \;\; H^* \text{ uma qualquer distribuição}\}.$$

Sob certas condições de regularidade, o nível assintótico, $NA(\varepsilon)$, e a potência assintótica, $PA(\varepsilon)$, sob contaminação, podem ser respectivamente expressos por,

$$NA(\varepsilon) = \alpha_0 + \varepsilon \int NIF\left(x; T, F_{\theta_0}\right) dH^*(x) + o(\varepsilon), \tag{2.34}$$

$$PA(\varepsilon) = \beta_0 + \varepsilon \int PIF\left(x; T, F_{\theta_0}\right) dH^*(x) + o(\varepsilon), \tag{2.35}$$

onde NIF e PIF correspondem respectivamente às funções de influência do nível e da potência definidas por

$$NIF(x;T,F_{\theta_0}) = \phi(\Phi^{-1}(1-\alpha_0))IF_{teste}(x;T,F_{\theta_0})/VA(T,F_{\theta_0})^{1/2}, \tag{2.36}$$

$$PIF(x;T,F_{\theta_0}) = \phi(\Phi^{-1}(1-\alpha_0) - \delta\sqrt{E_T})IF_{teste}(x;T,F_{\theta_0})/VA(T,F\theta_0)^{1/2}, \tag{2.37}$$

onde ϕ representa a função de densidade de probabilidade da Normal padrão e IF_{teste} denota a função de influência do teste, e não é mais do que a função de influência definida por (2.3) para um funcional $U(F) = \xi^{-1}(T(F))$ em que $\xi_n(\theta) \equiv E\left[T_n|\theta\right]$ (e $\xi(\theta) \equiv T(F_\theta)$). A função ξ_n definida de Θ para $I\!R$ verifica as seguintes condições: (i) $\xi_n(\theta)$ converge para $\xi(\theta)$, para todo o θ, e (ii) ξ possui primeira derivada e é estritamente monótona. Em Hampel *et al.* (1986, pág. 191) encontra-se a definição formal de função de influência de um teste. A função (2.36) pode interpretar-se como uma medida da influência que uma pequena contaminação no ponto x tem na probabilidade de rejeitar a hipótese nula.

As expressões (2.34) e (2.35) permitem determinar, ao longo da vizinhança, o nível assintótico máximo e a potência assintótica mínima do teste. A esse respeito ver por exemplo Hampel *et al.* (1986, pág. 199).

À semelhança da estimação pontual robusta, também aqui se pode tentar encontrar um compromisso entre robustez e eficiência através da pesquisa de um teste robusto óptimo. Este último deverá para tal maximizar a potência assintótica sujeito a que as funções de influência do nível e da potência sejam limitadas.

Para além das funções de influência outra análise robusta importante é a determinação do ponto de rotura de um teste. Esta medida descreve a fiabilidade global de um teste e indica a quantidade máxima de contaminação que pode ser tolerada pelo teste. He *et al.* (1990) introduziram o conceito de ponto de rotura para testes. Represente-se por Θ_0 e Θ_1 os conjuntos paramétricos associados respectivamente às hipóteses nula e alternativa e redefina-se a vizinhança de contaminação ε, $\mathcal{V}_\varepsilon(F_\theta)$, de um modelo central, F_θ, por

$$\mathcal{V}_\varepsilon(F_\theta) = \{H : H = (1-\varepsilon)F_\theta + \varepsilon H^*, \quad \text{com } H^* \text{ uma qualquer distribuição}\} \tag{2.38}$$

Então, define-se ponto de rotura da função potência de uma estatística de teste equivalente a um funcional, T, em $\theta \in \Theta_1$ por

$$\varepsilon_\theta^P(T) = \inf\{\varepsilon > 0 : T(F_{\theta_0}) \in T(\mathcal{V}_\varepsilon(F_\theta)) \quad \text{para algum } \theta_0 \in \Theta_0\}. \tag{2.39}$$

De forma análoga define-se ponto de rotura do nível de um teste por

$$\varepsilon_\theta^N(T) = \inf\{\varepsilon > 0 : T(F_\theta) \in T(\mathcal{V}_\varepsilon(F_{\theta_0})) \quad \text{para algum } \theta_0 \in \Theta_0\}. \tag{2.40}$$

A interpretação destas definições é a seguinte, para um dado θ, ε_θ^P corresponde à mínima proporção de observações discordantes que causará a confusão entre a hipótese nula com a alternativa θ. Quanto a ε_θ^N, pode dizer-se que é a mais pequena quantidade de contaminação que causará uma falha na rejeição de H_0 para o parâmetro θ.

Testes tipo M

Sejam $X_1, X_2, \cdots, X_n$, n variáveis aleatórias satisfazendo,

$$X_i = \mu + \sigma\epsilon_i, \qquad i = 1, 2, \cdots, n, \tag{2.41}$$

onde ϵ_i são variáveis aleatórias independentes, com distribuição comum F_ϵ e variância finita σ^2 (supostamente conhecida). A função de distribuição de X satisfaz $F_\mu(x) = F_\epsilon(x - \mu)$ onde μ é o parâmetro de localização. Pretende testar-se a hipótese nula $H_0 : \mu = \mu_0$ versus uma alternativa uni ou bilateral. Dada uma função ímpar genérica, ψ, considere-se um estimador definido pela equação

$$\sum_{i=1}^{n} \psi(x_i - T_n) = 0,$$

o qual não é mais do que um estimador–M univariado de localização. Então sob a hipótese nula a estatística

$$S_n = \sqrt{n}(T_n - \mu_0)$$

é assintoticamente normal.

A decisão a tomar pode então ser da forma: rejeita-se H_0 quando $S_n > k_n$ (ou quando $|S_n| > k_n$ se o teste for bilateral).

Para o mesmo modelo de localização com escala conhecida foram propostas outras estatísticas de teste (Sen, 1982), mas também testes do tipo M. Estas são da seguinte forma, sob H_0,

$$S_n = \sqrt{n}^{-1} \sum_{i=1}^{n} \psi(X_i - \mu_0),$$

onde ψ satisfaz $\int \psi(u - \mu_0) dF_{\mu_0} = 0$. Sob a hipótese nula,

$$S_n \to_d N\left(0, \int \psi^2(u - \mu_0) dF\mu_0\right).$$

O procedimento, no que concerne à tomada de decisão, é semelhante ao teste anterior.

O problema destas abordagens reside no desconhecimento da expressão da variância assintótica de S_n, sob a hipótese nula, quando a escala, σ, tem que ser estimada e F_{μ_0} não é simétrica.

Hampel *et al.* (1986) apresentam outras classes de testes, para além da classe M que foi apresentada acima, em problemas uniparamétricos em uma e duas populações (em geral são testes baseados em ordens).

Testes robustos em modelos paramétricos gerais

Ronchetti (1997) apresenta versões robustas do teste de Wald, do teste de *scores* e do teste de razão de verosimilhanças baseados em estimadores–M, isto é, em estimadores $\mathbf{T}_n$ de um parâmetro

$\boldsymbol{\theta}$ definidos por

$$\sum_{i=1}^{n} \psi(\mathbf{x}_i, \mathbf{T}_n) = \mathbf{0}. \tag{2.42}$$

Alguns dos testes robustos propostos na literatura para modelos lineares podem ser vistos como casos particulares daqueles propostos por Ronchetti (1997).

De modo a serem descritos os testes apresentados por esse autor é necessário introduzir alguns conceitos. Para tal admita-se um modelo paramétrico $F_{\boldsymbol{\theta}}$, com $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq I\!R^p$, e uma amostra i.i.d de vectores aleatórios $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ m-dimensionais. Seja $f_{\boldsymbol{\theta}}$ a função de densidade de $F_{\boldsymbol{\theta}}$ e denote-se a função *scores* por $s(\mathbf{x}, \boldsymbol{\theta}) = \partial \log f_{\boldsymbol{\theta}}(x)/\partial \boldsymbol{\theta}$. Seja $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$ uma partição de $\boldsymbol{\theta}$ com $\boldsymbol{\theta}^{(1)} \in I\!R^{p-q}$ e $\boldsymbol{\theta}^{(2)} \in I\!R^q$ $(q \leq p)$. Pretende-se, nestas condições, testar a hipótese nula $H_0 : \boldsymbol{\theta}^{(2)} = \mathbf{0}$, a qual pode ser vista como uma reformulação de uma hipótese linear geral.

Como referido atrás, os estimadores–M são consistentes e possuem, sob condições bastante gerais, uma distribuição assintótica Normal, (ver, por exemplo, Huber, 1981, Cap. 2). A função de influência do estimador definido por (2.42) é dada por,

$$IF(\mathbf{x}; \mathbf{T}, F_{\boldsymbol{\theta}}) = M(\psi, F_{\boldsymbol{\theta}})^{-1} \psi(\mathbf{x}, \boldsymbol{\theta}), \tag{2.43}$$

onde

$$M(\psi, F_{\boldsymbol{\theta}}) = -\int \frac{\partial \psi}{\partial \boldsymbol{\theta}} (\mathbf{x}, \boldsymbol{\theta})\, dF_{\boldsymbol{\theta}}(\mathbf{x}),$$

e a sua variância assintótica é expressa por

$$VA(\mathbf{T}, F_{\boldsymbol{\theta}}) = M(\psi, F_{\boldsymbol{\theta}})^{-1} Q(\psi, F_{\boldsymbol{\theta}}) (M(\psi, F_{\boldsymbol{\theta}})^{-1})^T, \tag{2.44}$$

com

$$Q(\psi, F_{\boldsymbol{\theta}}) = \int \psi(x, \boldsymbol{\theta}) \psi(x, \boldsymbol{\theta})^T dF_{\boldsymbol{\theta}}(\mathbf{x}).$$

Teste de tipo Wald

A estatística de teste para testar a hipótese H_0 acima expressa é dada pela seguinte forma quadrática

$$W_n^2 = (\mathbf{T}_n^{(2)})^T \left[VA(\mathbf{T}, F_{\boldsymbol{\theta}})^{(22)} \right]^{-1} \mathbf{T}_n^{(2)}, \tag{2.45}$$

onde $\mathbf{T}_n^{(2)}$ representa a componente do estimador $\mathbf{T}_n$ referente a $\boldsymbol{\theta}^{(2)}$ e $VA(\mathbf{T}, F_{\boldsymbol{\theta}})^{(22)}$ representa a partição da matriz $VA(\mathbf{T}, F_{\boldsymbol{\theta}})$ associada a $\boldsymbol{\theta}^{(2)}$.

Sob H_0, a distribuição assintótica de W_n^2 é um qui-quadrado central com q graus de liberdade.

No caso concreto do modelo de regressão linear este teste corresponde ao teste referido em Markatou *et al.* (1991).

Teste de tipo scores

Neste caso, a estatística de teste que permitirá testar H_0 é dada por,

$$S_n^2 = \mathbf{Z}_n^T \mathbf{C}^{-1} \mathbf{Z}_n, \tag{2.46}$$

sendo

$$\mathbf{Z}_n = n^{-1} \sum_{i=1}^{n} \psi(\mathbf{x}_i, \mathbf{T}_n^w)^{(2)},$$

$\mathbf{T}_n^w$ representa o estimador–M no modelo reduzido, ou seja, é a solução da equação abaixo

$$\sum_{i=1}^{n} \psi(\mathbf{x}_i, \mathbf{T}_n^w)^{(1)} = \mathbf{0} \quad com \quad \mathbf{T}_n^{w(2)} = \mathbf{0}.$$

A matriz quadrada, $\mathbf{C}$, de ordem q corresponde à matriz de covariâncias assintótica de $\mathbf{Z}_n$ e pode ser definida por $\mathbf{C} = \mathbf{M}^{(22.1)} \mathbf{V}^{(22)} (\mathbf{M}^{(22.1)})^T$ onde $\mathbf{M}^{(22.1)} = \mathbf{M}^{(22)} - \mathbf{M}^{(21)} (\mathbf{M}^{(11)})^{-1} \mathbf{M}^{(12)}$.

Sob H_0, a distribuição assintótica de S_n^2 é também um qui-quadrado central com q graus de liberdade.

Se a função ψ for a função *scores*, s, este teste acabado de descrever corresponde ao usual teste de *scores* (ou teste de Rao). Quando se considera o modelo de regressão linear, ter-se-á um teste do tipo *scores* conforme proposto por Markatou e Hettmansperger (1990).

Teste de tipo razão de verosimilhanças

Neste caso, a estatística de teste, que permitirá testar a hipótese em causa, terá a forma seguinte

$$R_n^2 = 2n^{-1} \sum_{i=1}^{n} [\xi(\mathbf{x}_i, \mathbf{T}_n) - \xi(\mathbf{x}_i, \mathbf{T}_n^w)], \tag{2.47}$$

onde $\xi(\mathbf{x}, \mathbf{0}) = \mathbf{0}$, $(\partial\xi/\partial\boldsymbol{\theta})(\mathbf{x}, \boldsymbol{\theta}) = \psi(\mathbf{x}, \boldsymbol{\theta})$ e T_n e T_n^w são os estimadores–M respectivamente no modelo completo e no modelo reduzido.

A distribuição de R_n^2 não é simples de determinar mesmo de uma forma assintótica. Ronchetti (1997) refere que sob uma condição adicional a distribuição assintótica desta estatística de teste, sob H_0, é uma combinação linear de qui-quadrados cada um com 1 grau de liberdade.

Quando ξ é o logaritmo da função de verosimilhança obtém-se o teste usual de razão de verosimilhanças. Este é também o teste τ proposto por Ronchetti (1982) para um modelo de regressão linear com uma escolha apropriada da função $\xi(\mathbf{x}, \boldsymbol{\theta})$.

Observações

Recentemente e com o propósito de testar o centro de uma população normal multivariada, Willems *et al.* (2002) propuseram a utilização de uma estatística de teste análoga à estatística clássica usada nestas situações: estatística T^2 de Hotelling. Esta estatística de teste é obtida substituindo os estimadores clássicos por estimadores MCD reponderados e foi denotada pelos autores por T_R^2. Sob

H_0, a distribuição assintótica de T_R^2 é aproximadamente proporcional a uma $F_{m,n'-m+1}$ onde m corresponde ao número de variáveis e n' é uma constante relacionada com a distribuição assintótica do estimador MCD.

Os testes apresentados nesta secção possuem boas propriedades locais de robustez (ver as referências citadas). He *et al.* (1990) estudaram também o comportamento global de alguns testes através do ponto de rotura definido atrás. Verificaram, por exemplo, que no caso de um modelo de localização univariado, o teste dos sinais é uniformemente mais robusto que o teste–M, e ainda, que para o teste da igualdade de várias médias (teste base da análise de variância a um factor), um teste de tipo *scores* tem melhor estabilidade global que um teste de tipo Wald.

2.4.2 Estimação regional robusta

Em vários problemas inferenciais o interesse reside na construção de regiões de confiança para um parâmetro de interesse, $\boldsymbol{\theta}$ (intervalos de confiança se $\boldsymbol{\theta} \in I\!R$), de um modelo sob consideração, $F_{\boldsymbol{\theta}}$.

Uma região de confiança com probabilidade de cobertura γ para um parâmetro $\boldsymbol{\theta}$ é um subconjunto de $\boldsymbol{\Theta}$, $RC_\gamma(\mathbf{x}_n)$ que depende apenas da amostra $\mathbf{x}_n$ e satisfaz

$$P(\boldsymbol{\theta} \in RC_\gamma(\mathbf{X}_n)) = \gamma.$$

Nesta definição está implícito que a probabilidade de cobertura não depende de quaisquer parâmetros perturbadores que possam existir no modelo. Numa interpretação frequencista o coeficiente de confiança corresponde ao valor esperado da frequência relativa da região que inclui, ou cobre, o verdadeiro valor do parâmetro $\boldsymbol{\theta}$, em repetições independentes do processo gerador dos dados $\mathbf{x}_n$.

Para cada amostra, $RC_\gamma(\mathbf{x}_n)$ constitui uma estimativa regional de $\boldsymbol{\theta}$. Como na estimação pontual e nos testes de hipóteses é importante verificar a robustez das regiões de confiança construídas. Isto envolve a avaliação do desempenho da região aleatória de confiança em termos de probabilidade de cobertura e do seu volume (comprimento ou amplitude, no caso de intervalos) face a algumas perturbações nos dados ou no modelo.

Gross (1976) efectuou um estudo de Monte Carlo bastante extensivo sobre intervalos de confiança. O seu objectivo consistiu em avaliar a robustez, quer em termos de validade (relativo à probabilidade de cobertura) quer em termos de eficiência (relativo ao comprimento do intervalo), dos intervalos construídos para um parâmetro de localização em populações univariadas simétricas com caudas pesadas. Para o efeito, utilizou vários estimadores robustos e concluiu que, para as situações estudadas, os melhores desempenhos foram obtidos para um estimador–M com função ψ tipo "onda" e biponderada de Tukey. Para além disso, também concluiu que os estimadores baseados em *jackknife* não tiveram um desempenho satisfatório (mesmo o relacionado com o estimador de "onda" atrás referido). Staudte e Sheather (1990, Cap. 5, 6) apresentam também vários intervalos de confiança robustos para o parâmetro de localização de uma população, assim como para a diferença desses parâmetros em duas populações.

De um modo geral, os procedimentos usuais de construção de regiões de confiança podem ser sintetizados nos seguintes:

1. Métodos de dualidade com testes de hipótese;
2. Métodos baseados em pivots;
3. Métodos baseados em ordenação estocástica (obtenção de funções de distribuição de estatísticas que são monótonas no parâmetro de interesse);
4. Métodos assintóticos (incluindo versões assintóticas dos pontos 1. e 2.).

Todavia, os três primeiros procedimentos apresentam dificuldades de aplicação. Por exemplo, no método pivotal a primeira dificuldade reside em encontrar a variável pivot e o segundo problema (válido para o primeiro e terceiro métodos), está relacionado com o desconhecimento da distribuição exacta da variável aleatória de interesse. Este último problema pode, em alguns casos, ser ultrapassado quando se trabalha em grandes amostras, dado que, sendo conhecida a distribuição assintótica desse pivot (ou variável de interesse), o problema fica solucionado.

Quando se trabalha com estimadores robustos o desconhecimento da distribuição exacta do estimador ou, de variáveis que permitam construir regiões de confiança, é uma dificuldade constante. Por esta razão, a construção de tais regiões é normalmente efectuada empregando procedimentos assintóticos. De seguida descrevem-se alguns métodos que têm sido propostos na literatura para a construção de regiões de confiança robustas. Sem perda de generalidade foque-se o interesse em parâmetros escalares, isto é, em $\boldsymbol{\theta} \in I\!R$. Seja T_n um estimador robusto de um parâmetro θ, assuma-se que a distribuição assintótica de T_n é Normal, tem-se então

$$\sqrt{n}(T_n - \theta) \underset{d}{\rightarrow} N(0, VA(T)). \tag{2.48}$$

Considere-se a construção de um intervalo bilateral para o parâmetro θ. Para uma probabilidade de cobertura, ou coeficiente de confiança γ, o objectivo é encontrar um percentil q_n tal que

$$\gamma = P(|T_n - \theta| < q_n). \tag{2.49}$$

Através de (2.48) e da resolução de (2.49) facilmente se obtém o pretendido. Logo, um intervalo de confiança robusto (aproximado) de nível $\gamma = 1 - \alpha$ para θ é

$$(t_n - \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{VA(T,F)}/n}, t_n + \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{VA(T,F)}/n} \tag{2.50}$$

onde t_n é o valor observado de T_n, $\sqrt{\widehat{VA(T,F)}/n}$ é a estimativa do seu erro padrão e $\Phi^{-1}(1-\alpha/2)$ é o quantil de ordem $1-\alpha/2$ de uma normal padrão.

A construção de intervalos robustos pode também ser efectuada usando o procedimento 1 (dualidade com testes de hipóteses), isto é, partindo das distribuições de estatísticas de teste, como, por exemplo, as referidas anteriormente, facilmente se podem obter regiões e intervalos de confiança

simultâneos para os parâmetros de interesse ou para combinações lineares deles. Neste sentido, Willems *et al.* (2002) apresentam intervalos de confiança robustos simultâneos para combinações lineares do centro de uma população normal. Esses intervalos assentam na aproximação da distribuição assintótica da estatística de teste apresentada atrás, T_R^2, denominada pelos autores por Hotelling robusta.

Conforme se pode constatar, a construção dos intervalos referidos, necessita do conhecimento da distribuição assintótica do estimador robusto, T_n, (ou pelo menos de uma aproximação dessa distribuição assintótica). Estes requisitos apresentam claras dificuldades dado que se torna, em muitos casos, difícil encontrar a distribuição (inclusive assintótica) de um estimador robusto.

Mesmo ultrapassando estas dificuldades e apesar dos intervalos (2.50) serem melhores que os seus análogos não robustos, eles ainda não são completamente satisfatórios, pelo facto, de não estar incorporado o viés devido à contaminação. Aliás, Fraiman *et al.* (2001) apresentam resultados numéricos onde a probabilidade de cobertura de intervalos de confiança robustos (usando o estimador–M de localização de Huber) se deteriora para amostras contaminadas de grandes dimensões, isto devido a essa não incorporação. Estes autores propõem que no cálculo de intervalos de confiança robustos deve ter incorporado o viés assintótico do estimador robusto considerado. O viés, ou desvio assintótico de um estimador robusto T_n é definido por $DA(T, F) = T_\infty - \theta$, onde T_∞ é o limite quase certo de T_n. Esse limite pode ser diferente de θ devido à tal possibilidade de existência de viés assintótico de T_n. Esta correcção introduz mais uma dificuldade na determinação dos intervalos. De facto, não existe nenhum método disponível para estimar o viés assintótico. Uma forma de lidar com este problema consiste em limitar $DA(T, F)$ utilizando o viés assintótico máximo na vizinhança de contaminação-ε considerada. Na literatura encontra-se disponível a expressão desse máximo para alguns estimadores robustos. Ver, por exemplo, Maronna *et al.* (1979), Martin *et al.* (1989), Croux *et al.* (1996) e Chen e Tyler (2002). Uma questão por resolver consiste em averiguar a possibilidade de estimar o viés referido através da aplicação de procedimentos de reamostragem.

Relativamente à estimação da variância, esta pode ser efectuada usando: i) aproximações empíricas da variância assintótica (determinando a expressão dessa variância, por exemplo, através da função de influência do estimador); ii) métodos de reamostragem, *jackknife* ou *bootstrap*. O método de reamostragem *bootstrap* pode também ser usado para estimar a distribuição do estimador robusto. No entanto o *bootstrap* usual quando aplicado a estimadores robustos pode conduzir à ocorrência de determinados problemas que serão discutidos mais detalhadamente no Capítulo 4.

2.4.3 Dificuldades em inferência robusta não pontual

Apesar das definições de funções de influência de um teste, do seu nível e da sua potência, de uma forma geral, pode afirmar-se que a construção de um teste robusto não é simples, e também que não é directo encontrar intervalos de confiança robustos. O problema reside frequentemente na determinação da distribuição associada ao estimador robusto utilizado. Na maior parte das vezes, é difícil obter as distribuições exactas de estimadores robustos e a distribuição assintótica de vários

estimadores robustos ainda não é conhecida. Mesmo quando a distribuição assintótica é conhecida outros problemas se colocam, nomeadamente, no que diz respeito à transposição desses resultados para o caso de amostras finitas. Uma abordagem para resolver esta última questão poderá consistir na utilização dos procedimentos robustos óptimos baseados na teoria assintótica e a partir deles, encontrar boas aproximações para amostras finitas. Exemplos de métodos que poderão permitir esta abordagem são: os métodos de aproximações empíricas de ponto de sela, métodos de expansão de "Edgeworth" e os métodos de reamostragem (principalmente o *bootstrap*). No Capítulo 4 será proposta uma alternativa para a realização de testes e intervalos de confiança robustos que mesmo sendo baseada em funções de influência conduz a um processo inferencial mais simples.

Esta tese tem precisamente por objectivo principal propor alternativas para a realização de testes robustos assim como para a construção de intervalos de confiança robustos. Os métodos alternativos que serão propostos para atingir tal objectivo conjugam as funções de influência com os métodos de reamostragem.

Antes de se poder prosseguir com a apresentação desses novos métodos considerou-se necessário efectuar uma introdução e expor algumas considerações sobre os métodos de reamostragem, originando por conseguinte o tema do próximo capítulo.

"My general feeling about bootstrapping is that I don't like it very much. It's easy for me to say that, because nowadays I dont't have to do practical problems for a living"

Henry Daniels, *Statistical Science*, Agosto 1993.

CAPÍTULO 3

Métodos de reamostragem

3.1 Introdução

As técnicas de reamostragem, cálculo repetido de estimativas da característica de interesse, um largo número de vezes, com repesagem ou não dos dados da amostra, são métodos, em geral não paramétricos, de estimação da distribuição amostral do estimador correspondente ou de algumas das suas características. A reamostragem é assim, simplesmente, um processo de estimação de probabilidades através de um vasto número de experiências numéricas. Esta ideia é bastante antiga, anterior mesmo à teoria das probabilidades. Porém o aparecimento desta criou condições para a libertação do trabalho monótono e exaustivo das experiências repetidas que juntamente com a estatística preparou o caminho para uma análise estatística elegante. Todavia, outras questões começaram a surgir, nomeadamente em problemas onde não é possível assumir determinados pressupostos. Por exemplo em certas situações não é razoável assumir um modelo numa família paramétrica, podendo apenas supor-se propriedades mais latas como a sua continuidade ou simetria. Os métodos de reamostragem surgiram, então, como uma ferramenta para superar tais limitações, estando adequados em problemas estatísticos envolvendo hipóteses distribucionais mínimas. O desenvolvimento tecnológico dos computadores nestas últimas décadas foi o grande impulsionador do corrente uso dos métodos de reamostragem em estatística. Por esta razão, estes métodos são muitas vezes designados na literatura por métodos computacionalmente intensivos. No entanto, o uso indiscriminado e intensivo de computadores não garante que a análise estatística seja realizada de forma correcta. É necessário que os problemas de inferência sejam bem equacionados. Se por um lado os métodos de reamostragem não necessitam de tantas considerações paramétricas como os restantes métodos estatísticos é aqui crucial que os dados sejam obtidos por um delineamento

correcto.

Em termos formais considere-se que se tem n observações, $\mathbf{x}_n$, pertencente a um espaço de resultados, $\mathcal{X}$, que é o gerador de todos os conjuntos de dados estatísticos que resultam da observação de uma variável aleatória X. Esse espaço de resultados ou amostral depende do delineamento da experiência em causa. Considere-se $f(x;\theta)$ a função de probabilidade ou densidade de uma variável aleatória X e denote-se por θ o parâmetro ou vector de parâmetros. Tendo disponível um elemento $\mathbf{x}_n \in \mathcal{X}$ coloca-se a seguinte questão:

Como extrair a informação contida em $\mathbf{x}_n$ de modo a deduzir as propriedades da população, $f(x;\theta)$, que o gerou?

A descrição do modelo probabilístico de $f(x;\theta)$ passa pela obtenção de determinadas características de $\mathbf{x}_n$ que ajudem nas inferências do parâmetro do modelo, θ. É neste ponto que a noção de estatísticas se torna necessária: funções, $T_n(\mathbf{X}_n)$, que não dependem de nenhum parâmetro desconhecido onde, $\mathbf{X}_n = (X_1, X_2, \cdots, X_n)$ representa uma amostra genérica aleatória de dimensão n. Estas funções são seleccionadas de acordo com algum princípio, por exemplo, máxima verosimilhança, suficiência, robustez. As estatísticas são então quantidades aleatórias que têm uma certa distribuição de probabilidade, a distribuição amostral da estatística. A maior parte das análises estatísticas requer algum conhecimento da distribuição amostral da estatística e esse depende do tipo de análise pretendida. Se o objectivo da inferência consistir na obtenção de intervalos ou regiões de confiança ou testes de hipóteses torna-se necessário conhecer a distribuição amostral da estatística ou os seus percentis. Por outro lado se o objectivo for a estimação pontual é desejável conhecer a precisão do estimador. Como a distribuição amostral de uma estatística assim como as suas características (por exemplo, precisão, centralidade) costumam depender da distribuição da população, $f(x;\theta)$, muitos procedimentos estatísticos são obtidos assumindo uma forma específica para f. Essa assunção, é em geral, feita quer por um conhecimento do fenómeno em causa quer por um acumulado de experiências passadas. Porém, existem muitas situações em que não é possível fazer quaisquer pressupostos sobre a sua forma. Nestes casos a distribuição amostral da estatística assim como as suas características irão ter de ser estimadas ou aproximadas. Convém, no entanto, referir que no caso da selecção de um estimador dentro de uma classe de estimadores, por vezes não é necessário estimar ou aproximar a sua medida de precisão pois, indiferentemente da população subjacente, pode existir um estimador, segundo algum critério de optimalidade, que é o melhor estimador a usar dentro da classe. É o caso de estimadores centrados de variância uniformemente mínima (UMVUE) ou dos estimadores BLUE[1] (*Best Linear Unbiased Estimator*) dentro da classe dos estimadores lineares. Quando não é possível recorrer a critérios de optimalidade, os dados estatísticos são os únicos a poder permitir estimar uma medida indicadora da sua precisão. É na parte de estimação da distribuição amostral da estatística ou das suas características que os métodos de reamostragem se enquadram.

[1] Note-se, contudo, que um estimador pode ser o melhor dentro desta classe mas é possível que todos os estimadores que pertencem à classe possuam más propriedades, é o caso, por exemplo, dos estimadores de mínimos quadrados para distribuições de caudas pesadas.

É no entanto importante salientar que se existir um elevado grau de certeza relativamente ao modelo gerador dos dados, restando apenas um problema de estimação, não há necessidade de optar por métodos de reamostragem.

O acesso generalizado a *software* estatístico de fácil uso, constitui, por um lado, um bom sinal do desenvolvimento da estatística mas, por outro lado, permite que a sua utilização indevida seja maior que nunca, em particular no que se refere aos métodos clássicos (os quais estão presentes nesse *software*, e são tipicamente baseados em hipóteses distribucionais específicas e em amostras "correctas"). Os procedimentos de reamostragem, apesar da sua aparente simplicidade também requerem uma correcta utilização, aliás como todos os métodos em estatística, caso contrário poder--se-ão obter resultados incorrectos e mesmo inadmissíveis. A ideia base destes procedimentos está na obtenção de conjuntos de dados estatísticos a partir de um conjunto de dados estatísticos original. É, contudo, o modo de obtenção desses conjuntos que vai caracterizar cada um deles. Os quatro procedimentos de reamostragem mais importantes podem resumir-se em:

1. *Jackknife* (proposto por Quenouille, 1949, 1956 e desenvolvido por Tukey, 1958);

   Este método foi introduzido como uma técnica de redução do viés de um estimador. A versão mais usada baseia-se no cálculo repetido do estimador em cada uma das possíveis subamostras obtidas através da eliminação de um valor da amostra original.

2. Permutação (Fisher, 1935 e Kempthorne, 1954);

   Os testes de permutação consistem em testar se um determinado termo pertence ao modelo estatístico. Com esse objectivo as observações são permutadas com respeito a esse termo e a estatística de teste é calculada para todas as possíveis configurações dos dados. Este processo origina uma distribuição que pode ser exacta ou empírica dos valores da estatística de teste. O valor actual é comparado com esses e a hipótese nula é rejeitada se o valor actual tem pouca probabilidade de ser excedido. Em situações de amostras de grandes dimensões são muitas vezes usadas considerações assintóticas.

3. Validação cruzada (Stone, 1974 e Geisser, 1975);

   Os dois principais objectivos deste método são a avaliação de um dado modelo estatístico e a selecção de um entre vários modelos competitivos. Numa situação ideal um modelo construído com base numa amostra seria validado aplicando-o a uma nova amostra proveniente da mesma população. Mas em muitas situações isto não é possível, sendo necessário usar a mesma amostra para construir e validar o modelo. A ideia da validação cruzada é a partição do conjunto de dados em amostra de treino e amostra de teste. A amostra de teste é usada para testar o modelo construído com base na amostra de treino. Este processo é repetido para um número (possivelmente grande) de divisões dos dados. Os resultados são cruzados para as várias subamostras permitindo avaliar ou seleccionar o modelo. Por exemplo, se o objectivo for o ajuste de um modelo de regressão a um conjunto de dados o procedimento de validação cruzada é o seguinte: o conjunto de dados é dividido aleatoriamente em duas partes. Uma

delas é usada para estimação. No exemplo referido a estimação consistirá no ajuste do modelo de regressão com todos os pormenores de análise: inclusão ou não de variáveis preditivas, transformações e mesmo técnicas de diagnóstico. A outra parte dos dados será usada para testar o modelo estimado, isto corresponderá no exemplo a prever os valores da segunda parte dos dados usando o modelo de regressão ajustado com base na primeira parte.

Outra forma de realizar a validação cruzada é deixar de fora um ponto amostral, ajustar o modelo aos restantes pontos e verificar como é que o modelo ajustado prevê o ponto excluído (este processo é denominado na literatura anglo-saxónica por *leave-one-out*). Repete-se este processo para cada um dos pontos da amostra, no final calcula-se a média dos erros de predição sendo esta medida designada por medida de validação cruzada do erro de predição. Este procedimento é muitas vezes confundido com o clássico *jackknife*, já que em ambos os casos é omitida uma observação. No entanto, a técnica de reamostragem do *jackknife* é bastante distinta da realizada pela validação cruzada. Para além disso este dois procedimentos têm objectivos distintos. Como atrás referido, com o *jackknife* pretende-se avaliar as propriedades de um estimador e com a validação cruzada avaliar a qualidade de um determinado modelo.[2]

4. *Bootstrap* (proposto por Efron, 1979);

   Enquanto que o *jackknife* e a validação cruzada reduzem a dimensão da amostra de trabalho, o *bootstrap* mantém a sua dimensão. Este método de reamostragem consiste na avaliação do estimador em todas as amostras possíveis, em geral, de dimensão igual à da amostra original, obtidas por selecção com reposição a partir da amostra disponível. Quando a amostra de trabalho tem grande dimensão é costume usar-se métodos de simulação para obter uma aproximação do estimador *bootstrap*.

Após o trabalho pioneiro de Quenouille, Tukey, Stone e Efron os métodos de reamostragem suscitaram um grande interesse na comunidade científica, mesmo para além dos estatísticos. Como consequência tem sido realizada muita investigação e publicada muita literatura. Hoje em dia, estas técnicas são bastante usadas e têm mostrado ser ferramentas essenciais na inferência estatística.

Quando se fala em métodos de reamostragem faz-se, muitas vezes, uma associação imediata com o método de simulação Monte Carlo. Apesar desta ligação não ser obrigatória, é uma das razões porque os métodos de reamostragem têm tanto sucesso, dado que muitas vezes não é possível calcular a estatística para todas as possíveis reamostras. Isso acontece, por exemplo, quando a dimensão da amostra é grande em reamostragens determinísticas mas também acontece para valores de n pequenos em planos de reamostragem estocásticos. É nesses casos que se recorre a simulações empíricas, denominadas Monte Carlo. O método de Monte Carlo escolhe aleatoriamente M reamostras de um elevado número de reamostras possíveis recalculando para cada uma delas o valor do estimador e usando a distribuição empírica dessas réplicas como uma aproximação da sua distribuição de amostragem. A validade deste procedimento pode ser justificada pela lei dos grandes números. A questão de quanto é que deverá ser o valor de M, apenas pode ser respondida

[2]Note-se que existe coincidência entre o *jackknife* e validação cruzada se o modelo for uniparamétrico.

tendo em conta o contexto do problema em questão. Tipicamente o valor de M varia de 50 a vários milhares. Uma boa abordagem sobre a aplicação destes métodos computacionalmente intensivos pode ser encontrada em Manly (1997).

Se o objectivo da análise estatística for a estimação ou aproximação de distribuições amostrais ou de algumas das suas características então os dois métodos de reamostragem mais adequados são o *jackknife* e o *bootstrap*. A parte principal deste capítulo focará estes dois métodos.

3.2 A abordagem usual

Um dos aspectos importantes quando se procede a uma análise estatística é a utilização das observações para aproximar ou estimar medidas de precisão de um dado estimador, tais como, a variância, o viés ou o erro quadrático médio. Antes de se introduzir os métodos de reamostragem *jackknife* e *bootstrap* descreve-se brevemente a abordagem tradicional utilizada na determinação de algumas dessas medidas.

A estimação usual, numa perspectiva frequencista é baseada na distribuição amostral do estimador. Um exemplo clássico é o dado pela estimação da variância de um estimador. Para isso considere-se a amostra aleatória $\mathbf{X}_n = (X_1, X_2, \cdots, X_n)$ proveniente de uma população desconhecida F e $T_n = T_n(\mathbf{X}_n)$ um estimador de um parâmetro $\theta \in \Theta$. Por definição, a variância de T_n é dada pela seguinte expressão

$$Var(T_n) = \int \left(T_n(\mathbf{x}_n) - \int T_n(\mathbf{y}_n) d \prod_{i=1}^{n} F(y_i) \right)^2 d \prod_{i=1}^{n} F(x_i), \tag{3.1}$$

onde $\mathbf{x}_n = (x_1, x_2, \cdots, x_n)$ e $\mathbf{y}_n = (y_1, y_2, \cdots, y_n)$.

Porém esta expressão não é de fácil cálculo para a generalidade dos estimadores, sendo muitas vezes impossível obter a partir dela uma fórmula mais explícita. Nos casos em que isso é possível, por exemplo para a média amostral, $T_n = \bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ obtém-se $Var(T_n) = n^{-1} Var(X_1)$, basta então estimar $Var(X_1)$. Esta estimação pode ser efectuada usando a variância amostral corrigida, $S_{n-1}^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$.

Não estando na presença de casos simples, a determinação de $Var(T_n)$ terá que ser efectuada usando aproximações. A abordagem tradicional reside na obtenção da variância assintótica do estimador, $VA(T_n, F)$. Como referido no capítulo anterior, sob certas condições de regularidade, $VA(T_n, F)$ pode ser obtida com base na função de influência do estimador. Outro procedimento usado e, de alguma forma relacionado com o anterior, é o denominado método delta. Na sua essência este método não é mais do que a expansão de uma função de uma variável aleatória em torno da sua média. Em geral, usa-se uma expansão de Taylor até à primeira ordem, calculando de seguida a variância. Por exemplo, se o objectivo for a aproximação da variância de $h(X)$, onde X é uma v.a.

com valor médio μ e h é uma função diferenciável, pode usar-se a seguinte aproximação

$$h(X) \approx h(\mu) + (X - \mu)\frac{\partial h}{\partial X}(\mu), \tag{3.2}$$

o que conduz a

$$Var(h(X)) \approx Var(X)(\frac{\partial h}{\partial X}(\mu))^2. \tag{3.3}$$

Contudo esta aproximação de $Var(h(X))$ só é boa se X tiver uma alta probabilidade de estar perto do seu valor médio μ. As expressões anteriores podem ser facilmente generalizadas para funções de vectores aleatórios, resultando

$$Var(h(\mathbf{X})) \approx \mathbf{H}'(\boldsymbol{\mu})Var(\mathbf{X})\mathbf{H}'(\boldsymbol{\mu})^T, \tag{3.4}$$

onde $\mathbf{X}$ é um vector aleatório de ordem m, $Var(\mathbf{X})$ é a sua matriz de covariâncias e $\mathbf{H}'$ corresponde à matriz de primeiras derivadas da função h.

O seguinte exemplo ilustra uma, entre muitas situações, onde não é possível obter uma expressão explícita a partir da equação (3.1).

Exemplo 3.1 Médias aparadas. Sejam $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ as n estatísticas ordinais de uma amostra aleatória de dimensão n na linha real ($\mathcal{X} = I\!R$), e β a percentagem que se "apara" de cada lado da amostra. A média β-aparada corresponde à média das restantes $(n-2\,[n\beta])$ estatísticas de ordem centrais, isto é

$$\bar{X}_n^\beta = \frac{1}{n - 2\,[n\beta]} \sum_{i=[n\beta]+1}^{n-[n\beta]} X_{i:n}.$$

Não existe uma forma explícita para a variância de $\bar{X}_n^\beta$ porém é possível mostrar que $VA(\bar{X}_n^\beta, F)$ é dada por

$$VA(\bar{X}_n^\beta, F) = 2(1-\beta)^{-2}\left[\int_0^{F^{-1}(1-\beta)} x^2 dF(x) + \beta F^{-1}(1-\beta)\right], \tag{3.5}$$

onde

$$F^{-1}(a) = \inf\{x : F(x) \geq a\}. \tag{3.6}$$

Sendo desconhecido o modelo gerador dos dados, uma estimativa de (3.5) pode ser obtida usando um estimador de F. ■

Pelo descrito pode constatar-se que a estimação de medidas de precisão de um estimador usando a abordagem tradicional contém algumas fraquezas:

1. a obtenção de estimativas precisas dessas medidas requer, em muitos casos, amostras de grandes dimensões;

2. para cada tipo de problema é, em geral, necessário obter uma expressão teoricamente;

3. em muitos problemas é difícil, e noutros mesmo impossível, deduzir essas fórmulas teoricamente (mesmo as aproximadas);

4. por vezes a expressão teórica é demasiado complicada para ser útil no cálculo da medida de precisão.

Nas secções seguintes irão ser abordadas duas metodologias estatísticas, o *jackknife* e o *bootstrap* que se têm revelado bastante eficazes em inferência estatística. Estes dois métodos de reamostragem são essencialmente de teor não paramétrico e têm como objectivo a estimação das características de precisão de estatísticas de interesse e também a estimação da distribuição de amostragem de uma variável aleatória, com base nos dados observados.

3.3 O *Jackknife*

Em 1949, Quenouille, inventou um método não paramétrico de estimação do viés de um estimador de correlação em séries temporais baseado na divisão da amostra em dois grupos. Em 1956 generalizou e explorou a aplicabilidade dessa ideia dividindo a amostra em várias subamostras. Este método, que seria mais tarde designado por *jackknife*, tem como versão de utilização mais frequente a divisão em n grupos de dimensão unitária, já que esta escolha elimina a arbitrariedade das subamostras.

3.3.1 A estimação *jackknife* do viés

Seja $\mathbf{X}_n = (X_1, X_2, \cdots, X_n)$ uma amostra de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) com distribuição F e $T_n = T_n(\mathbf{X}_n)$ um estimador de um parâmetro θ. O viés de T_n é definido por

$$Viés_F(T_n) = E_F(T_n) - \theta. \tag{3.7}$$

Denote-se por $T_{n-1,i} = T_{n-1}(X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$ o estimador com a mesma forma funcional de T_n mas baseado na subamostra que se obtém retirando da amostra original a i-ésima observação, $i = 1, 2, \cdots, n$. Seja $\bar{T}_n = n^{-1} \sum_{i=1}^{n} T_{n-1,i}$, então o estimador de jackknife do viés de T_n é dado por

$$[Viés(T_n)]_{Jack} = \widehat{Viés}(T_n) = (n-1)(\bar{T}_n - T_n). \tag{3.8}$$

De modo a avaliar a qualidade da estimação do viés de um estimador usando o *jackknife* é necessário saber quais as condições em que $[Viés(T_n)]_{Jack}$ é um estimador consistente.

Seja $T_n = h(\bar{\mathbf{X}}_n)$ uma estatística usada para estimar o parâmetro $\theta = h(\boldsymbol{\mu})$, onde $\boldsymbol{\mu} = E(\mathbf{X}_1)$ e h é uma função com domínio em $I\!R^m$ e conjunto de chegada $I\!R$. Assuma-se que h é diferenciável de segunda ordem e que as segundas derivadas parciais (que vão constituir a matriz Hessiana $\nabla^2 h(\mathbf{x}) = \partial^2 h(\mathbf{x})/\partial\mathbf{x}\partial\mathbf{x}^T$) são contínuas em $\boldsymbol{\mu}$. Nestas condições verifica-se a seguinte igualdade

$$\begin{aligned} T_n - \theta &= h(\bar{\mathbf{X}}_n) - h(\boldsymbol{\mu}) \\ &= [\nabla h(\boldsymbol{\mu})]^T(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) + 1/2(\bar{\mathbf{X}}_n - \boldsymbol{\mu})^T \nabla^2 h(\boldsymbol{\mu})(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) + V_n. \end{aligned} \tag{3.9}$$

O resto, V_n, é uma variável aleatória tal que $nV_n \to_p 0$, isto é, $V_n = o_p(n^{-1})$. Aplicando à equação anterior o operador valor médio obtém-se,

$$\begin{aligned} E\left(T_n - \theta\right) &= E\left(h(\bar{\mathbf{X}}_n) - h(\boldsymbol{\mu})\right) = \\ &= E\left([\nabla h(\boldsymbol{\mu})]^T(\bar{\mathbf{X}}_n) - \boldsymbol{\mu})\right) + \\ &\quad + E\left(1/2(\bar{\mathbf{X}}_n - \boldsymbol{\mu})^T \nabla^2 h(\boldsymbol{\mu})(\bar{\mathbf{X}}_n - \boldsymbol{\mu})\right) + \\ &\quad + E\left(V_n\right), \end{aligned} \tag{3.10}$$

o que é equivalente a

$$E\left(T_n - \theta\right) = E\left(h(\bar{\mathbf{X}}_n) - h(\boldsymbol{\mu})\right) = 0 + kn^{-1} + E\left(V_n\right), \tag{3.11}$$

onde $2k = tr(\nabla^2 h(\boldsymbol{\mu}) Var(\mathbf{X}_1))$. A expressão (3.11) define então o viés assintótico de T_n.

O estimador $[Viés(T_n)]_{Jack}$ pode ser usado para a construção de um novo estimador para θ, T_{Jack}. Este novo estimador é correntemente denominado na literatura por estimador de *jackknife* de viés corrigido e define-se por

$$T_{Jack} = T_n - [Viés(T_n)]_{Jack}. \tag{3.12}$$

Assim definido este estimador permite eliminar o termo de ordem $1/n$ de um viés que tenha a seguinte forma

$$E(T_n) = \theta + \frac{a_1}{n} + \frac{a_2}{n^2} + \ldots, \tag{3.13}$$

onde a_1, a_2 são desconhecidos mas não dependem de n. Esta aproximação é válida para muitos estimadores incluindo vários de máxima verosimilhança. Como $T_{n-1,i}$ são identicamente distribuídas, $i = 1, 2, \cdots, n$, então

$$E(T_{n-1,i}) = \theta + \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \ldots,$$

o que conduz a

$$\begin{aligned} E([Viés(T_n)]_{Jack}) &= (n-1)\left[E(\bar{T}_n) - E(T_n)\right] \\ &= (n-1)\left[a_1\left(\frac{1}{n-1} - \frac{1}{n}\right) + a_2\left(\frac{1}{(n-1)^2} - \frac{1}{n^2}\right) + \ldots\right] \\ &= a_1\frac{1}{n} + a_2\frac{2n-1}{(n-1)n^2} + O\left(\frac{1}{n^2}\right), \end{aligned}$$

indicando que o estimador $[Viés(T_n)]_{Jack}$, como estimador do viés do estimador T_n, estará correcto até à segunda ordem. Consequentemente,

$$Viés(T_{Jack}) = -a_2\frac{1}{(n-1)n} + O\left(\frac{1}{n^2}\right).$$

Esta última expressão confirma o que se referiu anteriormente acerca de T_{Jack}, isto é, trata-se de um estimador de T_n com redução do viés por eliminação do termo de ordem n^{-1}. Sob suposições mais fracas que (3.13) é possível mostrar que o estimador T_{Jack} possui a mesma distribuição assintótica que T_n tendo menor viés. Note-se que a redução do viés induzida pelo estimador T_{Jack} depende da suposição apresentada em (3.13) e também obviamente da boa estimação *jackknife* do viés, uma

vez que o termo subtraído ao estimador T_n, para obtenção do estimador T_{Jack}, não é mais do que uma estimativa do viés. O sucesso do estimador $[Viés(T_n)]_{Jack}$ depende da suavidade da estatística (estimador) T_n, a qual pode ser caracterizada pela diferenciabilidade da função que gera T_n. Para mais detalhes consultar os teoremas apresentados em Shao e Tu (1995, Cap. 2).

3.3.2 A estimação *jackknife* da variância

Iniciou-se a exposição do método *jackknife* pela estimação não paramétrica do viés de um estimador e consequentemente verificou-se a sua utilidade na redução do viés de estimadores. Apesar de este ser o objectivo inicial do método, o *jackknife* possui outra utilidade muito importante, a de permitir obter estimativas da variância de uma dada estatística. Esta proposta foi apresentada por Tukey (1958), tendo como ideia original a obtenção de uma variável pivot para estimação intervalar robusta. Este autor foi o responsável pela introdução dos conceitos de pseudo-valor e de estimador *jackknife*. A sua sugestão foi a de usar os valores recalculados, $T_{n-1,i}$, para definir n valores

$$\tilde{T}_{n,i} = nT_n - (n-1)T_{n-1,i} \qquad i = 1, 2, \cdots, n, \tag{3.14}$$

os quais designou por pseudo-valores. Tukey (1958) conjecturou que, num grande número de situações, estes n valores podiam ser tratados como aproximadamente independentes e identicamente distribuídos. Então a variável aleatória

$$\frac{\sqrt{n}\,(T_{Jack} - \theta)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\tilde{T}_{n,i} - T_{Jack}\right)^2}}, \tag{3.15}$$

teria aproximadamente uma distribuição *t-Student* com $n-1$ graus de liberdade e constituiria um pivot para a construção de um intervalo a $(1-\alpha)\times 100\%$ para θ. De um modo geral, esse intervalo de confiança só se tem revelado satisfatório para dimensões elevadas da amostra ($n \to \infty$).

O estimador T_{Jack} pode ser redefinido usando os pseudo-valores, $\tilde{T}_{n,i}$, isto é,

$$T_{Jack} = \frac{1}{n}\sum_{i=1}^{n}\tilde{T}_{n,i}\ .$$

Como se pode ver por (3.15), a fórmula de Tukey para estimar a variância de T_n é dada por

$$\begin{aligned}[Var(T_n)]_{Jack} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\tilde{T}_{n,i} - T_{Jack}\right)^2 = \\ &= \frac{n-1}{n}\sum_{i=1}^{n}\left(T_{n-1,i} - n^{-1}\sum_{j=1}^{n}T_{n-1,j}\right)^2 .\end{aligned} \tag{3.16}$$

Apresentam-se em seguida alguns exemplos de estimadores *jackknife*. No primeiro não há de facto qualquer necessidade de usar este método pois, como foi visto, esta situação é de fácil resolução na abordagem tradicional, no entanto constitui um exemplo clássico.

Exemplo 3.2 Estimador *jackknife* da variância da média aritmética. Dada uma amostra aleatória $\mathbf{X}_n = (X_1, X_2, \cdots, X_n)$ considere-se $T_n = \bar{X}_n$. Então

$$T_{n-1,i} = \frac{1}{n-1}\left(n\bar{X}_n - X_i\right), \qquad T_{Jack} = \bar{X}_n, \qquad n^{-1}\sum_{j=1}^{n} T_{n-1,i} = \bar{X}_n,$$

logo um estimador não paramétrico da variância de $\bar{X}_n$ é

$$\left[Var(\bar{X}_n)\right]_{Jack} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 = n^{-1}S_{n-1}^2,$$

expressão que coincide com a obtida pela abordagem clássica. Obviamente $\left[Viés(\bar{X}_n)\right]_{Jack} = 0$, quando $\bar{X}_n$ estima o valor médio de uma população num sentido não paramétrico.

■

Exemplo 3.3 Estimadores *jackknife* da variância e viés de médias aparadas. Considerem-se as condições do Exemplo 3.1. Se $(n-1)\beta = l$, for um inteiro, então

$$\left[Var(\bar{X}_n^{\beta})\right]_{Jack} = \frac{1}{n(n-1)(1-2\beta)^2}\sum_{i=1}^{n}\left(X_{w_i:n} - \bar{X}_n^{w}\right)^2.$$

O estimador *jackknife* do viés é dado por

$$\left[Viés(\bar{X}_n^{\beta})\right]_{Jack} = \frac{1}{1-2\beta}\left(\bar{X}_n^{\beta} - \bar{X}_n^{w}\right),$$

e finalmente

$$T_{Jack} = \frac{1}{1-2\beta}\left(\bar{X}_n^{w} - 2\beta\bar{X}_n^{\beta}\right),$$

onde

$$w_i = \begin{cases} l+1, & i \leq l+1 \\ i, & l+1 < i < n-l \\ n-l, & i \geq n-l \end{cases}.$$

e $\bar{X}_n^{w} = n^{-1}\sum_{i=1}^{n} X_{w_i:n}$ corresponde à média amostral *Winsorized*. Esta média é assim designada pois corresponde à média de uma amostra transformada, designada por amostra *Winsorized* (ver por exemplo, Hoaglin *et al.*, 1992). Quando $(n-1)\beta$ não é inteiro as expressões acima são ligeiramente mais complicadas.

■

Exemplo 3.4 Estimador *jackknife* da variância da variância amostral. Seja $\mathbf{X}_n = (X_1, X_2, \cdots, X_n)$ uma amostra aleatória de dimensão n e o estimador variância amostral

$$T_n = S_{n-1}^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Aplicando o *jackknife* tem-se

$$[Var(T_n)]_{Jack} = \frac{n^2}{(n-1)(n-2)^2}\left(\hat{\mu}_4 - \hat{\mu}_2^2\right),$$

onde $\hat{\mu}_k = n^{-1}\sum_{i=1}^{n}(x_i-\mu)^k$. Este constitui um bom resultado pois a variância de T_n é dada por

$$Var(T_n) = \frac{1}{n}\left(\mu_4 - \frac{(n-3)}{(n-1)}{\mu_2}^2\right),$$

em que $\mu_k = E(X_1-\mu)^k$ é o k-ésimo momento central de X_1.

Note-se que o estimador T_n assim definido não é um funcional, pois basta considerar a amostra de dimensão n, $\mathbf{x}_n = (x_1, x_2, \cdots, x_n)$, e a sua duplicada $\mathbf{x}_{2n} = (x_1, x_1, x_2, x_2, \cdots, x_n, x_n)$ para se verificar que ambas têm o mesmo valor de F_n mas diferentes valores de T_n. Só assintoticamente o estimador é equivalente a um funcional e para amostras finitas a diferença é apenas de ordem $1/n$. Observe-se ainda que a aplicação da expressão (3.16) não requer o conceito de funcional.

■

Para responder à questão da validade de $[Var(T_n)]_{Jack}$ como estimador de $Var(T_n)$, já que a conjectura de Tukey só é válida para determinadas estatísticas, consulte-se o Capítulo 2 de Shao e Tu (1995). O exemplo seguinte constitui a situação clássica do insucesso do *jackknife* como método para estimar a variância de uma estatística, neste caso, a mediana de uma amostra.

Exemplo 3.5 Estimadores *jackknife* da variância da mediana amostral. Considere-se a mediana amostral definida por

$$T_n = \begin{cases} X_{m:n}, & n = 2m-1 \\ \left(X_{m:n} + X_{m+1:n}\right)/2, & n = 2m \end{cases},$$

e suponha-se que se pretende estimar a mediana populacional $\theta = F^{-1}(0.5)$, com $F^{-1}(a)$ definida por (3.6).

Para amostras de dimensão par, $n = 2k$, com k inteiro, a expressão (3.16) conduz a

$$[Var\,(T_n)]_{Jack} = \frac{n-1}{4}\left(X_{k+1:n} - X_{k:n}\right)^2.$$

É possível mostrar, (Efron, 1982), que

$$n\,[Var(T_n)]_{Jack}\,/VA(T,F) \rightarrow_d Y,$$

onde $Y = \left(\chi_2^2/2\right)^2$ é uma variável aleatória de valor esperado 2 e variância 20, o que mostra que o estimador *jackknife* da variância da mediana amostral é inconsistente.

■

A justificação para o insucesso deste método no caso da mediana ou de outros quantis amostrais está relacionado com a falta de suavidade dos funcionais associados. Pelos teoremas apresentados em Shao e Tu (1995, Cap. 2) verifica-se que esse aspecto é crucial para obter boas propriedades dos estimadores *jackknife*. Assim sendo, em estudos teóricos acerca de $[Viés(T_n)]_{Jack}$ e de $[Var(T_n)]_{Jack}$ dever-se-á verificar se existe algum tipo de aproximação linear para a estatística T_n, usando alguma noção de diferenciabilidade do funcional equivalente a T_n.

3.3.3 *Deleted-d jackknife*

Para além do insucesso na estimação de variâncias de quantis amostrais, uma outra deficiência do *jackknife* é a de não poder originar directamente estimadores da distribuição amostral de uma estatística. Estas duas desvantagens podem ser evitadas quando o uso do *jackknife* não se baseia na eliminação de apenas uma observação mas sim de um grupo de observações. Esta técnica de reamostragem, denominada na literatura anglo-saxónica por *deleted-d jackknife*, foi proposta por Wu (1986) num contexto de regressão. Como anteriormente considere-se que se tem uma amostra aleatória $X_1, X_2, \cdots, X_n$, proveniente de uma população com função de distribuição F e um estimador T_n de um parâmetro θ. Denote-se por Cd um subconjunto qualquer de d elementos distintos, $1 \leq d \leq n$, provenientes do conjunto $\{1, 2, \cdots, n\}$ e represente-se por $T_{Cd} = T_r(X_i, i \in Cd^c)$ a estatística (que tem a mesma forma funcional que T_n) baseada nas restantes $r = n - d$ observações, onde Cd^c denota o complementar do conjunto Cd. No total existem $N = \binom{n}{d}$ diferentes subconjuntos Cd_j, com $j = 1, 2, \cdots, N$. Usando os vários valores eliminados, T_{Cd_j}, o estimador *deleted-d jackknife* da variância de T_n é dado por

$$[Var\,(T_n)]_{Jack-d} = \frac{n-d}{dN} \sum_{j=1}^{N} \left(T_{Cd_j} - N^{-1} \sum_{k=1}^{N} T_{Cd_k} \right)^2 . \tag{3.17}$$

Para que (3.17) seja consistente é necessário que o número de elementos d seja suficientemente grande. Esse valor suficientemente grande de d pode ser visto como uma correcção para a falta de suavidade do funcional que gera a estatística de interesse. Quando se escolhe $d = 1$ o estimador (3.17) é, como é natural, o estimador *jackknife* ordinário da variância definido por (3.16).

Outra importante vantagem deste método é a de permitir obter um estimador consistente da distribuição amostral de T_n. Como anteriormente considera-se $\theta(F)$ um funcional de interesse, por exemplo o valor médio ou o desvio padrão de F, e $T_n(\mathbf{X}_n)$ um estimador de $\theta(F)$, por exemplo a média ou o desvio padrão amostrais. O problema que agora se coloca é de estimar a distribuição amostral do estimador com base nos dados observados $(x_1, x_2, \cdots, x_n)$. A teoria tradicional do *jackknife* sugerida por Tukey (1958), consistia em tratar a variável aleatória

$$\frac{T_n\,(\mathbf{X}_n) - \widehat{Viés(T_n)} - \theta(F)}{\sqrt{\widehat{Var(T_n)}}},$$

como sendo uma *t-Student* com $n-1$ graus de liberdade. Porém, como já referido, isto só é válido em condições muito especiais (para mais detalhes sobre condições da validade da conjectura de Tukey

consultar, por exemplo, Thorburn, 1977 e Shi, 1984). O *deleted-d jackknife* permite colmatar esta falha do método original. Defina-se então

$$R_n(x) = R_{n,F}(x) = P\left(\sqrt{n}(T_n - \theta) \leq x\right),$$

esta função de distribuição pode ser estimada usando o denominado "histograma *jackknife* cumulativo",

$$\begin{aligned}[R_{n,F}(x)]_{Jack} &= P_*\left\{\sqrt{\frac{n(n-d)}{d}}\,(T_{Cd} - T_n) \leq x\right\} = \\ &= N^{-1}\sum_{j=1}^{N} I\left\{\sqrt{\frac{n(n-d)}{d}}\left(T_{Cd_j} - T_n\right) \leq x\right\}.\end{aligned} \tag{3.18}$$

Nesta expressão, $P_*(.) = P(.|\mathbf{X}_n)$ indica a probabilidade condicional de seleccionar a (re)amostra $\mathbf{X}_r$. Esta (re)amostra de dimensão $r = n - d$ é seleccionada segundo um esquema de amostragem aleatório simples sem reposição de $\mathbf{X}_n$. Pela teoria da amostragem finita significa então que $P_*\left(X_{Cd_j}\right) = N^{-1}$ e $f = r/n = (n-d)/n$ é o factor de correcção para populações finitas. Observe-se que $E(T_{Cd}|\mathbf{X}_n) = \bar{X}_n$ e $Var_*(T_{Cd}) = Var(T_{Cd}|\mathbf{X}_n) = \frac{1-f}{n-d}\hat{\sigma} = \frac{d}{n(n-d)}\hat{\sigma}$, onde $\hat{\sigma}$ é um estimador da raiz quadrada de $\sigma^2 = Var(X_1) < \infty$. Se estiver disponível um outro estimador da variância de T_n, $\widehat{Var_F(T_n)}$, também se pode definir um estimador *jackknife* da distribuição amostral da variável "studentized",

$$R'_n(x) = R'_{n,F}(x) = P\left\{\left(T_n - E_F(T_n)\right)/\sqrt{\widehat{Var_F(T_n)}} \leq x\right\},$$

sendo dado pela seguinte expressão,

$$\begin{aligned}R'_{nJack}(x) &= Pr_*\left\{\left(T_{Cd} - T_n\right)\Big/\sqrt{\widehat{Var_F\left(T_n\right)}} \leq x\right\} = \\ &= N^{-1}\sum_{j=1}^{N} I\left\{\left(T_{Cd_j} - T_n\right)\Big/\sqrt{\widehat{Var_F\left(T_n\right)}} \leq x\right\}.\end{aligned} \tag{3.19}$$

Observe-se que o estimador *deleted-d jackknife* da variância de T_n, (3.17), é exactamente a variância do respectivo histograma *jackknife*, porque

$$\begin{aligned}Var_*\left(\sqrt{\frac{n-d}{d}}\,(T_{Cd} - T_n)\right) &= \frac{n-d}{Nd}\sum_{j=1}^{N}\left(T_{Cd_j} - T_n - N^{-1}\sum_{k=1}^{N}(T_{Cd_k} - T_n)\right)^2 \\ &= [Var(T_n)]_{Jack-d},\end{aligned} \tag{3.20}$$

onde Var_* representa a variância calculada sob P_*.

Wu (1990) mostrou que o estimador *jackknife* de distribuição tem taxa de convergência de $O(n^{-1/2})$. Para funções suaves de médias Booth e Hall (1993) melhoraram esta taxa de convergência para $O(n^{-5/6})$ usando uma mistura do estimador *jackknife* com a distribuição normal (em limite). Outros resultados relativos ao histograma *jackknife* podem ser consultados em Shi (1991), Shao e Tu (1995) e Politis e Romano (1994).

3.3.4 Observações finais sobre o *jackknife*

O *jackknife* é um dos métodos de reamostragem que tem como objectivo principal a estimação do viés e da variância de um estimador. Como todos os métodos de reamostragem necessita de poucas hipóteses distribucionais. É um método não paramétrico, simples de aplicar em muitas e variadas situações. A extensão do exposto para o caso multivariado é bastante simples. Por exemplo, para estatísticas vectoriais, $\mathbf{T}_n$, bastará substituir nas expressões as quantidades univariadas pelas respectivas quantidades vectoriais, e o expoente dois pelo operador $\mathbf{x}\mathbf{x}^T$, onde $\mathbf{x}$ denota um vector genérico.

Os estimadores originais *jackknife* (*deleted-1 jackknife*) da variância e do viés de uma estatística T_n, $[Var(T_n)]_{Jack}$ e $[Viés(T_n)]_{Jack}$ respectivamente, devem o seu sucesso (ou melhor a sua consistência) à suavidade do funcional que a gera. O estimador $[Var(T_n)]_{Jack}$ é um estimador consistente para várias estatísticas, inclusive para estimadores–M e estatísticas de ordem lineares. Para mostrar a consistência do estimador $[Var(T_n)]_{Jack-d}$, com $d > 1$, (ver Shao e Tu (1995, Cap. 2)), são necessários menos requisitos de suavidade dos funcionais. Em contrapartida o estimador *jackknife* usual é de cálculo mais simples quando comparado com o *deleted-d jackknife*.

A utilização do *jackknife* em problemas mais complexos, onde existam observações não identicamente distribuídas (por exemplo, em modelos de regressão linear, lineares generalizados, não lineares ou de Cox), foi estudada por vários autores, tais como, Miller (1974), Hinkley (1977), Fox *et al.* (1980), Gong (1986) e Wu (1986). No contexto de modelos com estruturas de dados dependentes, como é o caso de séries temporais, existe também alguma literatura mas a investigação nesta área ainda não é muito avançada, ver por exemplo os trabalhos de Carlstein (1986) e Lele (1991).

Quando o estimador T_n é difícil de avaliar ou n é muito grande a aplicação do *jackknife* original pode constituir uma tarefa bastante pesada. Por esta razão têm sido propostas várias simplificações.

Uma extensão natural dos resultados (Quenouille, 1956) consiste em considerar que as n observações estão particionadas em aproximadamente g subconjuntos todos com a mesma cardinalidade, ou seja $n = gh$. Analogamente as estimativas eliminadas, $t_{n-h,i}$, são agora calculadas com base nas observações originais com o *i-ésimo* conjunto eliminado. O estimador *jackknife* da variância, assim obtido, é por vezes denominado por variância *jackknife* agrupada e será dado por

$$[Var(T_n)]_{gJack} = \frac{g-1}{g} \sum_{i=1}^{g} \left(T_{n-h,i} - g^{-1} \sum_{j=1}^{g} T_{n-h,j} \right)^2 .$$

A obtenção da estimativa $[Var(T_n)]_{gJack}$ requer apenas g avaliações da estatística. Se não existir uma forma simples de formar os g grupos estes podem ser escolhidos aleatoriamente.

Outra possibilidade é escolher aleatoriamente uma subamostra de dimensão h sem reposição e calcular h cópias do *jackknife* usual (*deleted-1 jackknife*) repetidamente deixando de fora em cada

uma delas um valor dessa subamostra. $[Var(T_n)]_{Jack}$ é então aproximada por

$$Var_{sJack} = \frac{n-1}{h} \sum_{t=1}^{g} h \left(T_{n-1,i_t} - h^{-1} \sum_{k=1}^{g} T_{n-1,i_k} \right)^2,$$

em que $\{i_1, i_2, \cdots, i_h\}$ indica os índices da subamostra seleccionada.

O cálculo de estimativas usando o *deleted-d jackknife* envolve ainda maior complexidade uma vez que são necessários $N = \binom{n}{d}$ avaliações da estatística em causa, o qual aumenta rapidamente com d. Para contornar esta dificuldade usam-se métodos de subamostragem, que conduzem a soluções aproximadas. Têm sido propostos métodos de subamostragem aleatória e de subamostragem equilibrada. A ideia é seleccionar de forma aleatória (ou equilibrada) uma amostra de subconjuntos, cada um deles de dimensão $n - d$, e avaliar a estatística nesses subconjuntos em vez de se proceder à avaliação da estatística em todos os subconjuntos de dimensão $n - d$. Para mais detalhes sobre estes dois procedimentos consultar Shao e Tu (1995, pág. 97). Se o cálculo de T_n requerer um grande número de iterações, por exemplo, no caso de estimadores robustos, então podem ser usadas aproximações dos seus pseudo-valores. Um exemplo sobre este tipo de método pode ser encontrado em Jorgensen (1993).

O S-PLUS e o SAS são exemplos de programas estatísticos que contêm funções para aplicação do método de reamostragem *jackknife* mas apenas na versão tradicional (*deleted-1 jackknife*).

Para finalizar esta secção, refira-se que outra aplicação também relevante do *jackknife* é em estatística robusta. Sendo importante tanto na possibilidade de poder ajudar na estimação de variâncias (viés) de estimadores robustos, como também de os pseudo-valores, $\tilde{T}_{n,i}$, serem indicativos da contribuição das observações no valor da estimativa, constituindo uma versão discreta da curva de influência de Hampel (como já mencionado no Capítulo 2).

3.4 O *Bootstrap*

Desde o trabalho de Efron (1979) que, o *bootstrap* tem sido o método de reamostragem que maior e mais rápido desenvolvimento teve em estatística. O resultado desse desenvolvimento traduz-se na extensa literatura a ele dedicada: várias centenas de artigos publicados (MathSciNet: <http://klymene.mpim-bonn.mpg.de/mathscinet>) assim como cerca de duas dezenas de livros. A um nível introdutório têm-se os artigos de Diaconis e Efron (1983) e Efron e Gong (1983). A relação do *bootstrap* com outros métodos de reamostragem é explorada em Efron (1982) e Hjorth (1994). Para além dos livros teóricos de Hall (1992a) e Shao e Tu (1995) têm sido também publicados outros livros focando aspectos mais práticos, entre eles salienta-se Efron e Tibshirani (1993) e Davison e Hinkley (1997).

O nome do procedimento de reamostragem *bootstrap* terá tido origem na expressão anglo-saxónica "*pull oneself up by ones bootstraps*", de tradução quase impossível mas que pretende transmitir a ideia de "colocar-se de pé pelos seus próprios meios". A analogia entre esta expressão e o método

de reamostragem em si terá a ver com a forma como este se realiza, já que o conjunto de dados de trabalho é usado para gerar novos conjuntos de dados. Na prática este método de reamostragem torna-se bastante útil na análise de conjuntos de observações cuja recolha seja muito dispendiosa e onde a informação *a priori* seja muito escassa, assim como em situações onde não estejam claras as hipóteses distribucionais e onde seja difícil obter observações posteriores.

O *bootstrap* é um método estatístico que se enquadra no "espírito da época", pois é computacionalmente intensivo e é delineado para obter a maior informação possível a partir de um conjunto de dados. Esta abordagem pode ser aplicada quer em problemas simples para a obtenção de medidas de incerteza e viés, quer em problemas mais complexos como regressão linear, análise discriminante, ou mesmo em processos estocásticos onde existe uma alta dependência dos dados. Uma vasta área de aplicação do *bootstrap* a situações reais foi já referida na introdução (ver Capítulo 1).

3.4.1 Generalidades e aspectos gerais

Os fundamentos base do *bootstrap* são conceptualmente simples. Seja F, em geral desconhecido, o modelo gerador da amostra aleatória $\mathbf{X}_n = (X_1, X_2, \cdots, X_n)$ isto é, X_i para $i = 1, 2, \cdots, n$ são variáveis aleatórias independentes e identicamente distribuídas. A partir de $(X_1, X_2, \cdots, X_n)$ calcula-se a estatística de interesse, $T_n(\mathbf{X}_n)$, que, sem perda de generalidade se considera como sendo uma função real de valores reais. Dada uma concretização de $(X_1, X_2, \cdots, X_n)$, $(x_1, x_2, \cdots, x_n)$, seja F_n a correspondente função de distribuição empírica. Chama-se amostra *bootstrap* $(X_1^*, X_2^*, \cdots, X_n^*) = \mathbf{X}_n^*$ à amostra de dimensão aleatória de dimensão n com distribuição comum F_n. Ou seja, cada variável aleatória X_j^* pode tomar o valor x_i com probabilidade $1/n$, $(i = 1, 2, \cdots, n)$. A estatística de interesse, T_n, avaliada na amostra *bootstrap* $\mathbf{X}_n^*$ é uma réplica *bootstrap* de T_n, e representa-se por $T_n^* = T_n(\mathbf{X}_n^*)$. Com base na amostra $(x_1, x_2, \cdots, x_n)$ o método de reamostragem *bootstrap* permite obter vários valores de T_n utilizando repetidamente esta amostra, criando, através de reamostragem com reposição, várias amostras de dimensão igual à original. A distribuição condicional da estatística aplicada à amostra *bootstrap* (dado $(x_1, x_2, \cdots, x_n)$) é denominada por distribuição *bootstrap*. Estas relações podem ser ilustradas pelo esquema seguinte:

$$\underset{\substack{\text{distribuição}\\ \text{desconhecida}}}{F} \overset{iid}{\rightarrow} \underset{\text{amostra aleatória}}{(X_1, X_2, \ldots, X_n)} \rightarrow \underset{\text{estatística de interesse}}{T_n(X_1, X_2, \ldots, X_n)},$$

perante o método de reamostragem *bootstrap* tem-se,

$$\underset{\substack{\text{distribuição}\\ \text{empírica}}}{F_n} \overset{iid}{\rightarrow} \underset{\substack{\text{amostra aleatória}\\ \textit{bootstrap}}}{(X_1^*, X_2^* \cdots, X_n^*)} \rightarrow \underset{\substack{\text{réplica } \textit{bootstrap}\\ \text{da estatística de interesse}}}{T_n\ (X_1^*, X_2^*, \ldots, X_n^*)} \quad .$$

Note-se que a utilização do sinal gráfico em forma de asterisco, "*", para representar amostras, estimadores ou outras quantidades relacionados com o *bootstrap* é a notação comum na literatura *bootstrap*. Neste trabalho essa notação será, sempre que possível, seguida.

Em termos mais formais suponha-se que temos uma amostra aleatória $\mathbf{X}_n$ (não necessariamente i.i.d) e $\mathcal{M}$ o modelo probabilístico que a originou. Em geral, $\mathcal{M}$ pode ser descrito pela distribuição conjunta de $\mathbf{X}_n$ ou por algumas quantidades que univocamente determinam a sua distribuição conjunta. Seja $V_n(\mathbf{X}_n, \mathcal{M})$ uma variável aleatória e suponha-se que se pretende estimar a sua distribuição através do método de reamostragem *bootstrap*. O primeiro passo deste procedimento consiste na estimação de $\mathcal{M}$ com base numa concretização de $\mathbf{X}_n$. Sejam $\mathbf{X}_n^*$ as variáveis geradas pelo modelo estimado $\hat{\mathcal{M}}$. A distribuição condicional de $V_n(\mathbf{X}_n^*, \hat{\mathcal{M}})$ dado $\mathbf{X}_n$ é o estimador *bootstrap* da distribuição de $V_n(\mathbf{X}_n, \mathcal{M})$. A ideia do *bootstrap* é "imitar" o terno $(\mathcal{M}, \boldsymbol{X}_n, V_n(\boldsymbol{X}, \mathcal{M}))$ através de $(\hat{\mathcal{M}}, \boldsymbol{X}_n^*, V_n(\boldsymbol{X}^*, \hat{\mathcal{M}}))$ onde as relações entre $\mathcal{M}$, $\boldsymbol{X}_n$ e $V_n(\boldsymbol{X}_n, \mathcal{M})$ são as mesmas que entre $\hat{\mathcal{M}}$, $\boldsymbol{X}_n^*$ e $V_n(\boldsymbol{X}_n^*, \hat{\mathcal{M}})$. Se $\mathcal{M} = \hat{\mathcal{M}}$ então a distribuição de $V_n(\boldsymbol{X}_n^*, \hat{\mathcal{M}})$ é idêntica à de $V_n(\boldsymbol{X}_n, \mathcal{M})$. Quando a distribuição condicionada de $V_n(\boldsymbol{X}, M)$ não é uma função explícita de $\boldsymbol{X}$ torna-se necessário utilizar um método de simulação (por exemplo, Monte Carlo) para calcular os estimativas *bootstrap*.

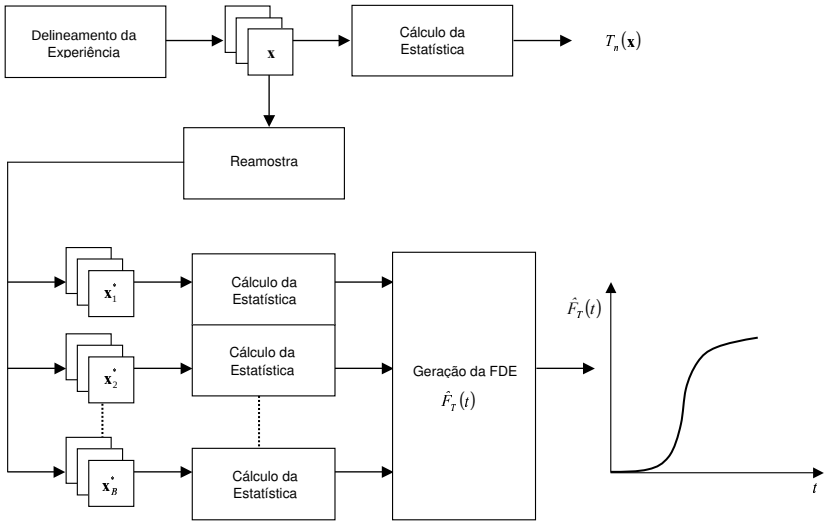


Figura 3.1: O método de reamostragem *bootstrap*.

Considere-se o caso simples de uma amostra $\boldsymbol{X}_n$ (i.i.d.) univariada proveniente de uma população F. A distribuição conjunta de $\boldsymbol{X}_n$ é determinada por F, neste caso $\mathcal{M} = F$. Não havendo nenhum conhecimento acerca de F, faz sentido fazer $\hat{\mathcal{M}} = F_n$ e obtém-se a chamada versão não paramétrica do *bootstrap*. Está-se no caso do *bootstrap* paramétrico quando F pertence a uma família paramétrica, neste caso $\mathcal{M} = F_\theta$, onde θ é parâmetro ou um vector de parâmetros e $\hat{\mathcal{M}} = F_{\hat{\theta}}$ onde $\hat{\theta}$ é um estimador de θ, por exemplo de máxima verosimilhança. Como é natural também se pode aplicar a versão não paramétrica a esta situação não tendo porém as mesmas propriedades de precisão que o paramétrico quando o modelo é correcto. A aplicação do método de Monte Carlo

no caso paramétrico é efectuada através da geração de amostras *bootstrap* $\boldsymbol{X}^* = (X_1^*, X_2^*, \ldots, X_n^*)$ a partir de $F_{\hat{\theta}}$ e no caso não paramétrico as amostras *bootstrap* são geradas a partir de F_n. Por vezes F_n é substituída por um estimador suave de F quando existe algum conhecimento acerca da *suavidade* de F. A Figura 3.1 ilustra genericamente o procedimento *bootstrap*.

A implementação do método *bootstrap* pode também ser vista de outra perspectiva, isto é, vendo--a como um simples método de substituição. Ou seja, se numa determinada expressão aparece a distribuição desconhecida F bastará colocar $\hat{F}$ no lugar de F para obter a aproximação *bootstrap* dessa expressão. Visto nesta perspectiva o princípio da substituição *bootstrap* não é mais do que o familiar procedimento de "colocar um estimador no lugar de algo desconhecido", correntemente denominado na literatura anglo-saxónica por *plug-in*.

3.4.2 A estimação *bootstrap* da variância

Seja $(X_1, X_2, \cdots, X_n)$ uma amostra aleatória de dimensão n proveniente da população F, com F desconhecido e suponha-se que se pretende determinar a variância do estimador T_n, $Var(T_n)$. Substituindo na expressão (3.1) F por um seu estimador $\hat{F}$, obtém-se o estimador *bootstrap* de $Var(T_n)$,

$$\begin{aligned}[Var(T_n)]_{Boot} &= \int \left(T_n(\mathbf{x}_n) - \int T_n(\mathbf{y}_n)\, d\prod_{i=1}^{n} \hat{F}(y_i) \right)^2 d\prod_{i=1}^{n} \hat{F}(x_i) \\ &= Var\left(T_n(X_1^*, X_2^*, \cdots, X_n^*) \,|\, (X_1, X_2, \cdots, X_n)\right),\end{aligned} \tag{3.21}$$

onde $(X_1^*, X_2^*, \cdots, X_n^*)$ é uma amostra aleatória *bootstrap* e $Var(.|X_1, X_2, \cdots, X_n)$ representa a variância condicionada, dado $(X_1, X_2, \cdots, X_n)$. Esta expressão corresponde à definição formal do estimador *bootstrap* da variância de T_n e na prática só é usada directamente quando se obtém uma função explícita de $(X_1, X_2, \cdots, X_n)$. O exemplo clássico é o da média de uma amostra, que é um dos poucos casos onde a determinação exacta dos momentos é possível, refira-se, mais uma vez, que nesta situação não há necessidade de utilização de métodos de reamostragem, apenas serve de ilustração.

Exemplo 3.6 Estimador *bootstrap* da variância da média aritmética. Considere-se a amostra aleatória $(X_1, X_2, \cdots, X_n)$ e o estimador $T_n = \bar{X}$. Como se sabe

$$Var\left(\bar{X}_n\right) = n^{-1} Var(X_1) = \int \left(x - \int y dF(y) \right)^2 dF(x),$$

se F for estimada pela função de distribuição empírica, isto é, $\hat{F} = F_n$, e se substituirmos F por F_n na expressão anterior obtém-se

$$\left[Var\left(\bar{X}_n\right)\right]_{Boot} = n^{-2} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2,$$

o qual coincide com a variância empírica de $\bar{X}_n$. Mas o estimador *bootstrap* da variância de $\bar{X}_n$ não é o estimador usual de $Var(\bar{X}_n)$ já que não é centrado. Recorde-se (Exemplo 3.2) que pelo contrário o estimador $\left[Var(\bar{X}_n)\right]_{Jack}$ é centrado.

■

Considere-se agora uma situação onde o método *jackknife* falha mas o *bootstrap* não, a estimação da variância da mediana amostral.

Exemplo 3.7 Estimador *bootstrap* da variância da mediana amostral. Seja $T_n = F_n^{-1}(0.5)$, com F_n^{-1} definida por (3.6) substituindo F por F_n. Para se determinar o estimador *bootstrap* da variância da mediana amostral vai construir-se em primeiro lugar a distribuição *bootstrap* da mediana. Seja $(X_1^*, X_2^*, \cdots, X_n^*)$ uma amostra *bootstrap* de dimensão n proveniente de F_n, isto é, $\hat{F} = F_n$ e assuma-se que $n = 2k - 1$. Denote-se por $(X_{1:n}^*, X_{2:n}^*, \cdots, X_{n:n}^*)$ a amostra *bootstrap* ordenada por ordem crescente. Como a dimensão da amostra é ímpar tem-se $T_n = F_n^{-1}(0.5) = X_{k:n}$. Dada a amostra observada $(x_1, x_2, \cdots, x_n)$ defina-se a variável aleatória N_i^* como sendo o número de vezes que x_i foi seleccionada na amostra *bootstrap* de dimensão n. Então o vector $(N_1^*, N_2^*, \cdots, N_n^*)$ tem uma distribuição $Multinomial(1/n, 1/n, \cdots, 1/n)$. Se considerarmos agora $N_{(i)}^*$ como sendo os valores de N_i^* induzidos pelas estatísticas ordinais $(x_{1:n}, x_{2:n}, \cdots, x_{n:n})$ então, para um inteiro j, $1 \leq j < n$, verifica-se

$$\begin{aligned} P\left(X_{k:n}^* > x_{j:n}\right) &= P\left(N_{(1)}^* + N_{(2)}^* + \cdots + N_{(j)}^* \leq k-1\right) \\ &= P\left(Bin(n, \frac{j}{n}) \leq k-1\right) \\ &= \sum_{i=0}^{k-1} \binom{n}{i} \left(\frac{j}{n}\right)^i \left(1 - \frac{j}{n}\right)^{n-i}, \end{aligned} \tag{3.22}$$

o que permite obter a distribuição *bootstrap* de $X_{k:n}$,

$$p_j = P\left(X_{k:n}^* = x_{j:n}\right) = \begin{cases} P\left(Bin\left(n, \frac{1}{n}\right) \geq k\right), & j = 1 \\ P\left(Bin\left(n, \frac{j}{n}\right) \geq k\right) - P\left(Bin\left(n, \frac{j-1}{n}\right) \geq k\right), & 1 < j < n \\ P\left(Bin\left(n, \frac{n-1}{n}\right) \geq k\right), & j = n \end{cases}. \tag{3.23}$$

Representando por $F_{X_{k:n}}$ a função de distribuição de $X_{k:n}$ tem-se, por definição,

$$Var\left(X_{k:n}\right) = \int \left[x - \int y dF_{X_{k:n}}(y)\right]^2 dF_{X_{k:n}}(x),$$

e substituindo na expressão anterior $F_{X_{k:n}}$ pela distribuição *bootstrap* de $X_{k:n}$, definida por (3.23), então

$$\left[Var\left(X_{k:n}\right)\right]_{Boot} = \sum_{j=1}^{n} p_j \left(X_{j:n} - \sum_{i=1}^{n} p_i X_{i:n}\right)^2.$$

■

Ao contrário do que acontece nos dois exemplos anteriores, muitas vezes a expressão (3.21) não é uma função explícita de $(X_1, X_2, \cdots, X_n)$ pelo que se torna necessário utilizar métodos de simulação de Monte Carlo. Como já discutido anteriormente a avaliação de $Var(T_n)$ definida em (3.1) é, salvo raras excepções, bastante difícil de efectuar directamente mesmo quando F é conhecida.

Nessa situação o método de Monte Carlo permite a obtenção de uma aproximação numérica para $Var(T_n)$. Esse algoritmo recria repetidamente novos conjuntos de dados a partir do modelo F, e calcula, para cada conjunto, o valor de T_n, no final determina o valor da variância da amostra constituída pelas várias avaliações de T_n. Esta ideia pode ser facilmente adaptada para o cálculo de uma aproximação de $[Var(T_n)]_{Boot}$ já que $\hat{F}$ é uma distribuição conhecida. Assim sendo, geram-se, independentemente, B amostras *bootstrap* a partir de $\hat{F}$, $\left\{X^*_{i,b}, i = 1, 2, \cdots, n; b = 1, 2, \cdots, B\right\}$, e para cada uma calcula-se $T^*_{n,b} = T_n(X^*_{1,b}, X^*_{2,b}, \cdots, X^*_{n,b})$ obtendo-se as B réplicas, $T^*_{n,1}, T^*_{n,2}, \cdots, T^*_{n,B}$, finalmente a aproximação Monte Carlo de $[Var(T_n)]_{Boot}$ é dada por

$$[Var(T_n)]^B_{Boot} = \frac{1}{B-1}\sum_{b=1}^{B}\left(T^*_{n,b} - B^{-1}\sum\nolimits_{j=1}^{B} T^*_{n,b}\right)^2. \tag{3.24}$$

Pela lei dos grandes números, $[Var(T_n)]_{Boot} = \lim\limits_{B\to\infty} [Var(T_n)]^B_{Boot}$.

A utilização de métodos de expansão em série de Taylor (método delta), assim como o *jackknife*, constituem também alternativas para a obtenção de aproximações de $[Var(T_n)]_{Boot}$. No entanto, tenha-se em atenção que, para estatísticas não regulares, tal como os quantis, pode não ser boa ideia a utilização do *jackknife*, pois como foi visto atrás o estimador *deleted-1 jackknife* da variância da mediana é inconsistente.

3.4.3 A estimação *bootstrap* da distribuição amostral

A grande utilidade do método *bootstrap* reside na estimação da distribuição de amostragem de uma variável aleatória de interesse, $T_n(X_1, X_2, \cdots, X_n, F)$, isto é, na estimação de

$$R_{n,F}(x) = P\left(T_n\left(X_1, X_2 \cdots, X_n, F\right) \leq x\right), \tag{3.25}$$

onde $(X_1, X_2, \cdots, X_n)$ é uma amostra aleatória de dimensão n proveniente de F. Para obter um estimador *bootstrap* de $R_{n,F}$ basta substituir em (3.25), F por um seu estimador $\hat{F}$, o que conduz a,

$$\begin{aligned} R_{n,Boot}(x) = R_{n,\hat{F}}(x) &= P\left(T_n\left(X^*_1, X^*_2, \cdots, X^*_n, \hat{F}\right) \leq x \middle| \left(X_1, X_2, \cdots, X_n\right)\right) = \\ &= P\left(T^*_n \leq x \middle| \left(\mathbf{X}_n\right)\right), \end{aligned} \tag{3.26}$$

onde $(X^*_1, X^*_2, \cdots, X^*_n)$ representa uma amostra *bootstrap* de dimensão n proveniente de $\hat{F}$.

De forma análoga ao que acontecia com a estimação *bootstrap* da variância de um estimador, é frequente que (3.26) não tenha uma expressão de fácil cálculo analítico. Mais uma vez a resolução pode passar pela aplicação de um procedimento de simulação, por exemplo o método de Monte Carlo. Aplicando-o obtém-se a seguinte aproximação Monte Carlo de $R_{n,Boot}(x)$,

$$R^B_{n,Boot}(x) = B^{-1}\sum_{b=1}^{B} I\left\{T^*_{n,b} \leq x\right\}, \tag{3.27}$$

em que $T^*_{n,b} = T_n(X^*_{1,b}, X^*_{2,b}, \cdots, X^*_{n,b}, \hat{F})$ é a b-ésima versão *bootstrap* da variável aleatória de interesse, $b = 1, 2, \cdots, B$.

Para além da aproximação Monte Carlo outros métodos são possíveis para determinar uma aproximação de $R_{n,Boot}$. As aproximações por expansão de Edgeworth e ponto de sela são dois deles. Para mais detalhes sobre este tipo de aproximações consulte-se Babu (1989), Hall (1992a), Davison e Hinkley (1988) e Davison e Hinkley (1997, Sec. 9.5).

Após a obtenção da distribuição *bootstrap* pode-se, consequentemente, calcular o valor médio e a variância *bootstrap* e usá-las como estimativas do valor médio e da variância de $T_n(X_1, X_2, \cdots, X_n, F)$. Naturalmente o estimador *bootstrap* da variância, $[Var(T_n)]_{Boot}$ definido por (3.21) (ou $[Var(T_n)]^B_{Boot}$, dado por (3.24)) é a variância da variável $T_n(X_1, X_2, \cdots, X_n, F)$ usando a distribuição $R_{n,Boot}$ (ou $R^B_{n,Boot}$).

Note-se ainda que é possível, para pequenos valores de n, obter sempre a distribuição *bootstrap* $R_{n,Boot}$, bastando para tal gerar todas as possíveis amostras com reposição a partir da amostra original.

A estimação de medidas de precisão de um estimador T_n pode também ser feita usando a distribuição *bootstrap*. Uma dessas medidas é o viés de um estimador. Para definirmos o estimador *bootstrap* desta quantidade considere-se que T_n é um estimador de um parâmetro desconhecido θ, e relembre-se a expressão do viés de T_n, definida em (3.7), reescrita agora da seguinte forma,

$$Viés\,(T_n) = \int x dR_{n,F}(x) - \theta,$$

onde $R_{n,F}(x)$ é dada por (3.25). Um estimador *bootstrap* do viés de T_n é então dado por

$$[Viés\,(T_n)]_{Boot} = \int x dR_{n,Boot}(x) - T_n. \tag{3.28}$$

Tal como nas situações anteriores, se o integral em (3.28) não for de fácil resolução pode usar-se a seguinte aproximação Monte Carlo,

$$\begin{aligned} [Viés\,(T_n)]_{Boot} &= \int x dR^B_{n,Boot}(x) - T_n = \\ &= B^{-1} \sum_{b=1}^{B} T^*_{n,b} - T_n, \end{aligned} \tag{3.29}$$

com $T^*_{n,b} = T_n(X^*_{1,b}, X^*_{2,b}, \cdots, X^*_{n,b}, \hat{F})$ constituindo a avaliação de T_n na b-ésima reamostra *bootstrap*, $b = 1, 2, \cdots, B$.

Para além da estimação do viés também se poderiam definir outras medidas de precisão. É, porém, noutras áreas que a aplicação do *bootstrap* se torna mais apelativa, nomeadamente intervalos de confiança, testes de hipóteses, modelos lineares assim como em problemas mais complexos como modelos multivariados, amostragem finita e séries temporais. Os já citados Efron e Tibshirani (1993), Davison e Hinkley (1997) e Hjorth (1994) constituiem boas referências da variedade de tópicos onde o *bootstrap* pode ser utilizado.

3.4.4 Intervalos de confiança

Existem vários métodos de construção de intervalos de confiança *bootstrap*. Entre os métodos mais utilizados encontram-se o método do percentil, o método do percentil-t, o método *bias-corrected* (Efron, 1982, Cap. 10) e o método *accelerated bias-corrected* (Efron, 1987). Comparações teóricas entre os vários tipo de intervalos *bootstrap* podem encontrar-se em Hall (1988) e Shao e Tu (1995, Cap. 4). Para além destes métodos existem outros mais sofisticados, nomeadamente o baseado no *bootstrap* iterativo e o *bootstrap* calibrado (veja-se Shao e Tu, 1995, Sec. 4.3). Como o objectivo é dar apenas uma ideia geral das técnicas mais utilizadas reduz-se aqui a apresentação aos quatro métodos inicialmente referidos. Porém antes de iniciar a sua descrição reveja-se o método tradicional baseado na normalidade assintótica.

Suponha-se que se tem $\mathbf{X}_n = (X_1, X_2, \cdots, X_n)$ uma amostra aleatória de dimensão n proveniente de uma população F. O objectivo é estimar um determinado parâmetro $\theta = \theta(F)$ associado com a distribuição F cuja forma é desconhecida. Considere-se ainda que $T_n = T(\mathbf{X}_n)$ é uma estatística adequada para estimar θ. De forma a simplificar a apresentação dos vários métodos considera-se apenas regiões de confiança em $I\!R$ (intervalos de confiança) bilaterais.

Se a distribuição assintótica de T_n for Normal e o viés de T_n for desprezável (comparado com a raiz quadrada de $Var_F(T_n)$) então um intervalo bilateral de confiança aproximado para θ, a $(1-\alpha) \times 100\%$, é dado por

$$\left(T_n - z\sqrt{Var_F\left(T_n\right)}, T_n + z\sqrt{Var_F\left(T_n\right)}\right)$$

onde $z = \Phi^{-1}(1-\alpha/2)$ é o percentil de probabilidade $1-\alpha/2$ de uma Normal padrão. Se o viés de T_n, $Viés_F\left(T_n\right)$ não for desprezável, então o intervalo de confiança deve ser ajustado de forma adequada. De uma forma geral um intervalo, a $(1-\alpha) \times 100\%$, é da forma

$$\left(T_n - Viés_F\left(T_n\right) - z\sqrt{Var_F\left(T_n\right)}, T_n - Viés_F\left(T_n\right) + z\sqrt{Var_F\left(T_n\right)}\right). \tag{3.30}$$

Os intervalos assim obtidos baseiam-se no facto de que se conhece a distribuição assintótica de T_n, todavia para formular este tipo de intervalos é necessário conhecer quer o $Viés_F\left(T_n\right)$ quer $Var_F\left(T_n\right)$. Naturalmente que estas duas últimas medidas podem ser estimadas porém só se pode utilizar este tipo de intervalos para grandes valores de n.

O método do percentil

As aproximações *bootstrap* de $Viés_F\left(T_n\right)$ e de $Var_F\left(T_n\right)$, dadas respectivamente por (3.28) e (3.21) (ou (3.29) e (3.24)) podem ser utilizadas para construir um intervalo de confiança para θ baseado na aproximação Normal, equação (3.30). Alternativamente pode pensar-se directamente na construção de intervalos de confiança baseados na distribuição exacta de T_n ou de $T_n - \theta$. Obviamente que essas distribuições não são conhecidas mas estão disponíveis as suas aproximações *bootstrap*, como se viu atrás. Sendo

$$R_{n,Boot} = P\left(T_n^* \leq x | \mathbf{X}_n\right)$$

a distribuição *bootstrap* de T_n, um intervalo *bootstrap* bilateral a $(1-\alpha)\times 100\%$ é da forma

$$\left(R_{n,Boot}^{-1}(\alpha/2), R_{n,Boot}^{-1}(1-\alpha/2)\right), \tag{3.31}$$

em que $R_{n,Boot}^{-1}(\alpha) = Q^*(\alpha)$ representa o percentil de probabilidade α da distribuição *bootstrap*. É por esta razão que este intervalo se diz construído pelo método (*bootstrap*) do percentil.

Exemplo 3.8 Intervalo de confiança *bootstrap* para a mediana. Pelo exemplo 3.7 e nas mesmas condições não é difícil obter um intervalo *bootstrap* a $(1-\alpha)\times 100\%$ de confiança para $T_n = F^{-1}(0.5) = X_{k:n}$, pois conhece-se a distribuição *bootstrap* de T_n, que como foi visto se baseia na distribuição Binomial. O limite inferior do intervalo *bootstrap* será $x_{j_1:n}$ com

$$j_1 = \inf\{j : P(X_{k:n}^* \geq j) > \alpha/2\}$$

e o limite superior do intervalo *bootstrap* será $x_{j_2:n}$ com

$$j_2 = \sup\{j : P(X_{k:n}^* \leq j) > \alpha/2\}.$$

Estas expressões podem ser facilmente implementadas num qualquer programa de estatística, permitindo determinar a sua solução para valores razoáveis de n.

■

Quando não é possível obter a distribuição *bootstrap* do estimador T_n recorre-se a técnicas de simulação de Monte Carlo e através de $R_{n,Boot}^B$ obtêm-se os percentis "empíricos" *bootstrap*. Para a determinação desses percentis procede-se da forma que em seguida se descreve. Em primeiro lugar ordenam-se por ordem crescente as B réplicas (valores de T_n para a reamostra *bootstrap*) *bootstrap*, obtendo-se $T_{1:B}^* \leq T_{2,B}^* \leq \cdots \leq T_{B:B}^*$. Se j for um inteiro positivo, então

$$R_{n,Boot}^B\left(T_{j:B}^*\right) = B^{-1}\sum_{b=1}^{B} I\left(T_{n,b}^* \leq T_{j:B}^*\right) = j/B.$$

Sendo assim, estimam-se os $\alpha/2$-ésimo e $1-\alpha/2$-ésimo quantis de T_n por

$$Q^*(\alpha/2) \cong T_{j_1:B}^*, \quad Q^*(1-\alpha/2) \cong T_{j_2:B}^*,$$

em que $j_1 = [B\alpha/2]+1$ e $j_2 = [B(1-\alpha/2)]+1$. Assim a versão Monte Carlo do intervalo percentil *bootstrap* definido em (3.31) é simplesmente

$$\left(T_{j_1:B}^*, T_{j_2:B}^*\right). \tag{3.32}$$

Em algumas situações utiliza-se a distribuição *bootstrap* de $T_n - \theta$,

$$R_{n,Boot}(x) = P\left(T\left(\mathbf{X}_n^*\right) - T_n \leq x\right),$$

ou a sua versão Monte Carlo, para construir o intervalo de confiança *bootstrap* pelo método do percentil (chamando-se a estes intervalos reflectidos).

O método *bias-corrected* percentil

Em situações onde o viés de T_n não é desprezável deve-se proceder à correcção do intervalo (3.31). Se se verificar que $P\left(T_n\left(\mathbf{X}_n^*\right) \leq T_n\left(\mathbf{X}_n\right)|\mathbf{X}_n\right) \neq 0.5$, então deve-se corrigir esse "viés" (Efron, 1981b). Seja $R_{n,Boot}$ a distribuição *bootstrap* de T_n dada uma amostra e $z_0 = \Phi^{-1}\left(R_{n,Boot}\left(T_n\right)\right)$. Então o intervalo *bootstrap* bilateral *bias-corrected* a $(1-2\alpha)\times 100\%$ de confiança é dado por

$$\left(R_{n,Boot}^{-1}\left(\Phi\left(2z_0 - z_\alpha\right)\right), R_{n,Boot}^{-1}\left(\Phi\left(2z_0 + z_\alpha\right)\right)\right) \tag{3.33}$$

em que z_α satisfaz $\Phi\left(-z_\alpha\right) = \alpha$. Os argumentos que justificam este intervalo são dados em Efron (1982, Cap. 10) ou em Shao e Tu (1995, Cap. 4). Note-se que, se $R_{n,Boot}(T_n) = 0.5$, então $z_0 = 0$ e o intervalo (3.33) não é mais do que o intervalo definido em (3.31).

O método *accelerated bias-corrected* percentil

Efron (1987) propôs, a partir do artigo de Schenker (1985), o método *accelerated bias-corrected* percentil, o qual não só tenta corrigir o viés mas também estabilizar a variância. O método assume a existência de uma transformação que vai conduzir à distribuição Normal padrão, não sendo, todavia, necessário conhecer essa transformação. O intervalo *bootstrap accelerated bias-corrected* percentil a $(1-2\alpha)\times 100\%$ de confiança é dado por

$$\left(R_{n,Boot}^{-1}\left(\Phi\left(z^\alpha\right)\right), R_{n,Boot}^{-1}\left(\Phi\left(z^{1-\alpha}\right)\right)\right),$$

com

$$z^\alpha = z_0 + \left(z_0 + z_\alpha\right)/\left(1 - a\left(z_0 + z_\alpha\right)\right),$$

representando a a constante de aceleração. Um problema na aplicação deste tipo de intervalos está na determinação dessa constante. Se $a = 0$ obtém-se o intervalo (3.33). Todavia a sua determinação ou estimação não é simples. Em Efron (1987) encontram-se algumas propostas para a sua aproximação. DiCiccio e Romano (1988) e Hall (1988) apresentam também procedimentos para a determinação do parâmetro a.

O método percentil-t

Para uma estatística geral pode considerar-se a distribuição amostral de uma variável "studentized", S_θ, dada por

$$R_n\left(x\right) = P\left(S_\theta \leq x\right),$$

com

$$S_\theta \equiv \frac{T_n - \theta\left(F\right)}{\sqrt{\widehat{Var_F\left(T_n\right)}}},$$

onde $\widehat{Var_F(T_n)}$ é um estimador consistente de $Var_F(T_n)$. Se a distribuição R_n for conhecida então um intervalo a $(1-\alpha)\times 100\%$ de confiança para $\theta = \theta(F)$ é

$$\left(T_n - q\left(1-\alpha/2\right)\sqrt{\widehat{Var_F\left(T_n\right)}}, T_n - q\left(\alpha/2\right)\sqrt{\widehat{Var_F\left(T_n\right)}}\right),$$

onde $q(\alpha)$ denota o quantil de probabilidade α da distribuição R_n. De um modo geral só são conhecidos estes quantis quando a amostra subjacente provém de uma população Normal. No entanto, a distribuição R_n e como consequência os seus quantis podem ser estimados usando o *bootstrap*. Em particular a distribuição *bootstrap* pode ser usada para aproximar R_n, sendo dada por

$$R_{n,Boot}(x) = P\left(T(\mathbf{X}_n^*) \leq x\sqrt{[\widehat{Var(T_n)}]_{Boot}} + T_n\right), \tag{3.34}$$

em que $[\widehat{Var(T_n)}]_{Boot}$ é uma estimativa de $Var(T_n)$ baseada na amostra *bootstrap*. Porém, se um estimador de $Var(T_n)$ não estiver disponível pode utilizar-se o estimador *bootstrap* como definido em (3.21) ou em (3.24). Em qualquer dos casos o intervalo bilateral percentil-t *bootstrap* a $(1-\alpha)\times 100\%$ de confiança é da forma

$$\left(T_n - q^*(1-\alpha/2)\sqrt{\widehat{Var_F(T_n)}}, T_n - q^*(\alpha/2)\sqrt{\widehat{Var_F(T_n)}}\right),$$

onde $q^*(\alpha)$ representa o quantil de probabilidade α da distribuição *bootstrap* $R_{n,Boot}$ definida por (3.34).

Este intervalo é mais preciso que o intervalo obtido pelo método do percentil (veja-se Hall, 1988 ou Shao e Tu, 1995, Cap. 4). Mas repare-se que este aumento de precisão é obtido à custa de um aumento de cálculo já que a construção do intervalo percentil-t requer o cálculo da $Var(T_n)$ para cada uma das reamostras *bootstrap*. Por esta razão este método diz-se que utiliza o procedimento *bootstrap* iterado (quando se estima a variância de T_n, por exemplo, através de *jackknife*) ou duplo (quando se estima a variância de T_n por *bootstrap*).

Dos vários métodos apresentados uns possuem melhores propriedades que outros assim como mais ou menos restrições ou requerimentos. Viu-se que o percentil-t necessita de um bom estimador da variância, que o método *accelerated bias-corrected* percentil precisa de um ajustamento analítico e o *bootstrap* iterado necessita de cálculos adicionais. Assim sendo, em muitas situações práticas continua a ser útil a utilização de intervalos *bootstrap* menos precisos mas mais simples de determinar.

Contudo, observe-se que, em várias situações paramétricas, o *bootstrap* origina estimativas regionais mais precisas do que os intervalos usuais baseados na normalidade assintótica dos estimadores de máxima verosimilhança, ver por exemplo Efron e Tibshirani (1993, Cap. 21) e Shao e Tu (1995, Cap. 3).

Testes de hipóteses

Como os intervalos de hipóteses e testes de hipóteses são problemas duais na teoria estatística, isto é, pode-se construir testes de hipóteses com base em intervalos de confiança e vice versa, não se irá desenvolver aqui este tópico. Não obstante, a existência dessa dualidade, é importante referir que existem situações onde é mais fácil construir um teste directamente do que através de um intervalo de confiança. Referências sobre testes de hipóteses *bootstrap* podem ser encontradas nos já citados Efron e Tibshirani (1993, Cap. 16), Shao e Tu (1995, Cap. 4) e Davison e Hinkley (1997, Cap. 4).

3.4.5 Observações finais sobre o *bootstrap*

O método *bootstrap* pode ser visto como sendo uma mistura de dois procedimentos: um é descrito pelo princípio da substituição, F substituído por $\hat{F}$, e, quando necessário, a seguir utiliza-se um algoritmo de aproximação de natureza numérica. Como em muitas situações a segunda fase é realizada usando métodos de Monte Carlo, o *bootstrap* é, por vezes, confundido com uma técnica de simulação.

No caso de independência e idêntica distribuição, Bickel e Freedman (1981) e Singh (1981) mostraram que para médias e quantis amostrais a distribuição *bootstrap* estima consistentemente a distribuição de amostragem. Ou seja, a diferença entre a distribuição amostral e a distribuição *bootstrap* converge para zero uniformemente, para todas as sequências amostrais, à medida que a dimensão da amostra aumenta. Babu e Singh (1983) estenderam estes resultados a uma classe alargada de estatísticas. O Capítulo 3 de Shao e Tu (1995), assim como as referências aí citadas, constitui uma boa fonte sobre o trabalho teórico que tem sido desenvolvido relativamente ao estudo da consistência e precisão dos estimadores *bootstrap*.

Quando se utiliza o método de simulação Monte Carlo para obter aproximações dos estimadores *bootstrap* uma importante questão é a escolha do valor de B. O valor de B deverá ser escolhido de forma a que o erro da simulação (da aproximação) seja desprezável quando comparado com o erro do estimador *bootstrap*. Efron (1987) considerou o coeficiente de variação (CV=quociente entre o desvio padrão e o valor médio) como uma medida da variabilidade da aproximação Monte Carlo para um dado valor de B. Assim para a estimação da variância o valor de B proposto por Efron é tal que, para um dado nível ε_0, se tem $CV\left(se^B_{Boot}\right) = \varepsilon_0$, onde $se^B_{Boot} = \sqrt{Var^B_{Boot}(T_n)}$. Em muitas situações essa relação pode reduzir-se a $B = 0.5\varepsilon_0^{-2}$. Por exemplo quando se fixa $\varepsilon_0 = 0.01$ então o valor de $B = 5000$, mas se $\varepsilon_0 = 0.1$ esse valor baixa para $B = 50$. De acordo com Efron e Tibshirani (1986) o valor mais comum para a estimação de momentos de um estimador é da ordem de 100 ou 200. Nesse mesmo artigo é apresentado outro método, também com base no coeficiente de variação, para a determinação do valor de B associado à construção de intervalos de confiança *bootstrap*. Efron e Tibshirani (1986) e Efron (1987) sugerem que o valor de B para a estimação da distribuição *bootstrap* ou de seus quantis deve ser da ordem de 1000 ou 2000. Babu e Singh (1983) mostraram que se $B \geq n \ln n$ continuava-se a manter a precisão de segunda ordem de $R^B_{n,Boot}$. Nesta linha, e estudando as taxas de convergência de $R^B_{n,Boot}$ e $\left(R^B_{n,Boot}\right)^{-1}$, Shi *et al.* (1990) propõem que para a estimação da distribuição *bootstrap* o valor de B seja aproximadamente $n^2 \ln \ln n$.

Quando se utiliza o método de Monte Carlo o processo de reamostragem torna-se computacionalmente dispendioso sendo, se possível, conveniente investir em métodos que reduzam esse custo. Existem técnicas padrão que melhoram a eficiência das aproximações Monte Carlo (Hammersley e Handscomb, 1964) nomeadamente a aproximação linear, a reamostragem equilibrada e a reamostragem por importância. Por exemplo a reamostragem por importância surgiu pela primeira vez no contexto da reamostragem *bootstrap* em Johns (1988) e Davison (1988). Veja-se Hall (1992b) e Shao e Tu (1995) para mais detalhes sobre a aplicação e discussão sobre a sua eficiência destas

técnicas de refinamento quando aplicadas ao *bootstrap*.

Uma das áreas da estatística onde o *bootstrap* tem tido bastantes aplicações é a regressão linear. Existem vários resultados na literatura relacionados com este modelo, vejam-se os trabalhos de Efron (1979), Freedman (1981), Freedman e Peters (1984), Wu (1986), Efron (1988) e de DeAngelis *et al.* (1993). Para além disso a estimação em regressão linear robusta através de *bootstrap* foi estudada por Shorack (1982) e mais recentemente por Salibian-Barrera e Zamar (2002a).

A necessidade da normalidade das populações subjacentes é uma constante na maior parte dos tópicos de análise multivariada. Como já discutido no capítulo anterior violações a esta suposição podem traduzir-se em maus desempenhos. Assim, em paralelo com a construção de procedimentos robustos, o *bootstrap* constitui uma alternativa não paramétrica para aproximar as distribuições das estatísticas em análise multivariada. Muitos dos resultados descritos nas secções anteriores podem ser alargados ao caso multivariado, mas existem tópicos únicos que têm que ser analisados separadamente, tais como a análise discriminante, modelos lineares multivariados e análise factorial, entre outros. Com base no estudo da distribuição *bootstrap* da variável aleatória $\sqrt{n}\,(\mathbf{S}_n - \mathbf{\Sigma})$, onde $\mathbf{\Sigma}$ denota a matriz de covariâncias populacional e $\mathbf{S}_n$ a sua usual versão amostral, realizado por Beran e Srivastava (1985), pode construir-se regiões de confiança e testes de hipóteses *bootstrap* para $\mathbf{\Sigma}$. Também Zhang e Boos (1992) propõem testes de hipóteses *bootstrap* para a matriz de covariâncias. Em modelos lineares multivariados pode encontrar-se o trabalho de Alemayeha (1987). Em análise factorial, de agrupamentos e componentes principais encontram-se os artigos de Chatterjee (1984), Peck *et al.* (1989) e Diaconis e Efron (1983), respectivamente. É, porventura, a análise discriminante o tópico de análise multivariada que mais aplicações teve do método *bootstrap*, essencialmente na estimação das taxas de erro de má classificação. Uma vez que este tema será retomado no capítulo seguinte serão aí apresentadas as referências da literatura *bootstrap* a ele dedicada.

Para finalizar refira-se que tem sido dedicado muito trabalho à justificação teórica da validade do *bootstrap*, mas como este procedimento é muito flexível, logo com um grande leque de aplicações, ainda resta muito a fazer. Em termos práticos, o *bootstrap* oferece o potencial e a flexibilidade de se obterem métodos inferenciais altamente precisos e pode mesmo eliminar a necessidade de assumir ou impor um modelo estatístico quando não há fundamentos para o fazer.

No capítulo seguinte descrevem-se algumas situações onde este método pode falhar e para uma delas apresenta-se uma proposta de resolução, assim como várias aplicações e análises de desempenho do método proposto.

"Quem sabia o que cada coisa significava realmente? Quão longe no escuro temos de cair, ou subir, antes de descobrirmos o significado e a verdade - mesmo assumindo que eles estão lá para os descobrirmos?"

Madeleine St John - "The essence of the thing"

CAPÍTULO 4

Bootstrap robusto

Os cálculos de reamostragem são baseados nos dados observados e, neste sentido, são condicionais a eles. Em particular, no caso não paramétrico não se usa nada mais que os próprios dados. Assim, problemas na amostra irão perturbar quer o funcionamento do *bootstrap*, quer os seus cálculos. Existem três situações amostrais evidentes que contribuem para essa perturbação: i) dados incompletos (provenientes de censuras ou perdas); ii) dados dependentes; e iii) dados discordantes. Estas situações podem, com maior ou menor dificuldade, ser contornadas. No caso de dados incompletos, o problema pode ser resolvido através da modificação dos estimadores de dados completos para dados incompletos, ou introduzindo mecanismos aleatórios de censura ou perda de dados (ver, por exemplo, Efron, 1981a; Hjort, 1992; Efron, 1994). Quando existe dependência nos dados, e no caso concreto do *bootstrap*, a situação pode ser resolvida efectuando reamostragem por blocos (ver, por exemplo, Hall, 1985; Carlstein, 1986; Liu e Singh, 1992; Hall *et al.*, 1995), ou então através da descrição da dependência através de um modelo estrutural envolvendo distúrbios independentes (ver, por exemplo, Freedman, 1984; Bose, 1988).

A última situação referida irá ser tratada com mais detalhe nas secções seguintes, uma vez que a possível existência de observações atípicas na amostra aliada ao método de reamostragem *bootstrap* constitui o interesse central deste capítulo assim como deste trabalho. Inicialmente introduzir-se-ão algumas questões referentes a observações discordantes e, em seguida apresentam-se os problemas que, em concreto, essas observações provocam quando se usam procedimentos estatísticos de reamostragem. Far-se-á ainda uma retrospectiva da pouca metodologia existente para a resolução desses problemas. De seguida apresenta-se um novo método o qual será denominado por *bootstrap* IF, para o distinguir do *bootstrap* clássico introduzido por Efron (1979). Este novo método de reamostragem, baseado no *bootstrap*, tendo como objectivo principal o de permitir lidar com a possível

existência de observações discordantes, resolve um problema de falta de robustez do método *bootstrap* clássico e permite efectuar inferências robustas não pontuais de forma mais simples que os métodos referidos no Capítulo 2.

Uma vez descritos os métodos existentes e proposto o novo, apresentar-se-ão vários estudos numéricos relativos a diversas situações de estimação com base em amostras com diferentes tipos de contaminação. Iniciar-se-á a exposição com um problema simples de estimação pontual e regional para a localização univariada. A seguir analisar-se-á o caso da estimação do coeficiente de correlação linear e, por último, abordar-se-á a selecção de variáveis em análise discriminante linear. Desta forma, pretende-se avaliar a qualidade e eficiência do novo método proposto e fazer algumas comparações com os métodos já existentes.

4.1 O *bootstrap* e as observações discordantes

As ideias e técnicas da chamada análise exploratória de dados têm recebido considerável atenção na literatura estatística (ver, por exemplo, Hoaglin, Mosteller e Tukey, 1983, 1985) e constituem uma parte fundamental na análise estatística aplicada. Para além de outros aspectos esta análise passa também pelo apuramento da possível existência de observações atípicas. Apesar de neste trabalho já se ter usado o termo observação atípica (ou discordante), até ao momento não foi discutido este conceito. Ao longo do tempo várias definições têm sido propostas relativamente ao conceito anglo-saxónico de *outlier* (ver, por exemplo, Beckman e Cook, 1983). Não obstante a dificuldade de definição e mesmo sendo uma noção vaga, vai considerar-se neste trabalho o conceito apresentado por Hampel *et al.* (1986, pág. 21) ou Hampel (2000, pág. 15), que é o seguinte: uma observação é atípica se não se ajustar ao padrão sugerido pela maioria das observações.

Em análise de dados, existem, fundamentalmente, duas estratégias para lidar com observações atípicas, a identificação e a acomodação. A primeira estratégia consiste simplesmente em tentar identificar todas as observações discordantes para estudo posterior. A identificação dessas observações pode conduzir à sua rejeição, à sua incorporação ou a um tratamento especial no caso de elas conterem alguma nova informação sobre o problema em estudo. A segunda estratégia constitui um dos objectivos da estatística robusta e consiste na possibilidade de, através de mudanças adequadas, quer no modelo e/ou no método de análise, permitir incorporar essas observações sem prejuízo para a análise estatística do padrão associado à maioria das observações. Apesar destas estratégias terem por base uma filosofia distinta assiste-se, hoje em dia, ao seu uso complementar, já que, e como foi referido no Capítulo 2, a detecção de observações discordantes torna-se mais eficiente e segura quando se usam métodos robustos. Isto acontece porque quando existem observações atípicas a análise estatística clássica é geralmente afectada por esses pontos e origina regras de decisão que podem induzir em erro.

Um método mais primitivo respeitante ao tratamento de observações atípicas, consiste no uso de "regras de rejeição de pontos discordantes". Para uma discussão e revisão geral destas regras veja-se

Barnett e Lewis (1994) e Davies e Gather (1993). Quando se considera o procedimento combinado de rejeição e posterior estimação usando as restantes observações pode e deve considerar-se esta classe de métodos como sendo robustos. Podem distinguir-se duas classes de regras de rejeição de observações atípicas, uma que se pode designar por objectiva e outra por subjectiva. Esta última classe é constituída por métodos de detecção/rejeição sem qualquer regra fixa ou expressão matemática envolvida na decisão (muitas vezes são métodos gráficos que servem para identificar as observações atípicas). À classe objectiva pertencem os procedimentos que são baseados em expressões matemáticas e regras bem definidas. Em qualquer das situações após a detecção (e não se tratando de erros possíveis de correcção) essas observações são removidas e são aplicados métodos estatísticos clássicos às observações restantes.

Apesar deste procedimento conduzir a estimativas robustas dever-se-á ter alguma precaução na análise dos resultados da sua aplicação, porque a rejeição e posterior estimação poderá conduzir a interpretações estatísticas erradas ou no mínimo questionáveis (Hampel, 2000, pág. 15). A grande questão que se coloca a este procedimento é que a análise estatística efectuada após a "limpeza" dos dados não tem em conta a variabilidade introduzida pelo passo de limpeza o que implica uma perda de eficiência do método usado. Sobre este assunto ver os trabalhos de Hampel (1985), Dupuis e Hamilton (2000) e Zuo *et al.* (2002). Existe ainda uma outra classe de regras de detecção/rejeição, constituída pelas denominadas regras Bayesianas de detecção de observações discordantes (ver nesta área os trabalhos de Chaloner e Brant, 1988 e de Weiss, 1996).

Como, infelizmente, não existem propostas na literatura para estimar de forma consistente a variabilidade das estimativas determinadas após a remoção de observações discordantes é necessário explorar outras alternativas. Assim, uma alternativa inferencial que permite lidar com observações discordantes e que constitui uma proposta melhor que o uso das regras robustas abruptas de detecção/rejeição (tradução de *hard rejection rule*) é o uso de uma classe de estimadores robustos com tratamento suavizado de possíveis pontos atípicos. Esses estimadores deverão rejeitar todos os pontos que são claramente atípicos e tratar de uma forma suave e contínua aqueles pontos que não são completamente bons nem completamente maus. Ou seja, estes estimadores rejeitam completamente observações distantes mas não de uma forma súbita, permitindo uma zona transitória de dúvida. Alguns estimadores–M, já mencionados, pertencem a esta classe. Até ao momento nada foi referido acerca dos métodos de reamostragem e a sua ligação com as observações discordantes. É esse o objectivo dos próximos parágrafos.

Como já referido nos Capítulos 1 e 3, em vários problemas estatísticos é impossível determinar soluções analíticas ou não é possível estabelecer as hipóteses que permitam fazer tal determinação. Nestas situações, os métodos de reamostragem têm um importante papel permitindo obter soluções aproximadas para esses problemas. O aumento do poder computacional tem sido o responsável pelo grande desenvolvimento desses métodos. Um desses procedimentos é o já mencionado *bootstrap*. Viu-se no capítulo anterior que existem duas versões do *bootstrap*, a não paramétrica (quando F é estimada por F_n) e a paramétrica (quando F_θ é estimada por $F_{\hat{\theta}}$). Em ambas a estatística de interesse T_n tem que ser calculada assumindo que os dados foram gerados por F_n ou $F_{\hat{\theta}}$. Todavia,

em muitas situações é necessário recorrer a aproximações Monte Carlo pois a estatística assim avaliada pode não ter expressão analítica explícita. Nesses casos, e como já explicado no capítulo anterior, gera-se um grande número de amostras provenientes de F_n ou $F_{\hat{\theta}}$ e para cada uma delas calcula-se o valor da estatística de interesse. Neste trabalho assumir-se-á que se está a trabalhar num contexto não paramétrico (sempre que não for esse o caso tal será referido).

Sendo a existência de observações discordantes numa amostra um problema na aplicação de qualquer procedimento estatístico, esse problema pode ter consequências mais graves quando se usam procedimentos de reamostragem *bootstrap*. A razão para esse agravamento tem a ver com a génese de amostragem do próprio método, a reposição. A presença de observações atípicas pode conduzir a distribuições amostrais estimadas com caudas mais pesadas, o que se deve à possibilidade de algumas reamostras *bootstrap* poderem ter um nível de contaminação superior ao da amostra original. Considere-se, como exemplo, a situação em que a amostra original de dimensão $n = 30$ possui uma observação discordante. Então espera-se que o número de reamostras *bootstrap*, de dimensão 30, que contêm essa observação discordante mais do que uma vez seja aproximadamente de 26%. Isto implica que cerca de 26% de valores estimados da estatística *bootstrap* podem ser bastante afectados por uma única observação estranha na amostra original, logo não se poderá confiar na distribuição amostral estimada, em especial nas caudas. Dado que a construção de intervalos de confiança ou testes de hipóteses requer quantis extremos este facto pode ser um problema quando se usa o *bootstrap* clássico. Athreya (1987) mostrou que o *bootstrap*, na sua versão original, pode falhar para distribuições de caudas longas. No seu artigo este autor considerou a estimação por *bootstrap* da média em distribuições sem variância. Estas distribuições originam amostras que possuem os mesmos problemas das amostras com observações discordantes.

Pelo exposto acima, fica claro que o uso de *bootstrap* com estimadores clássicos em distribuições de caudas pesadas ou com observações atípicas apresenta vários problemas, quer de natureza prática, quer teórica.

Uma importante questão a colocar quando se fazem cálculos por métodos de reamostragem é: O que seria diferente se uma ou mais observações fosse omitida da amostra original? Na presença de um modelo paramétrico existem algumas técnicas de diagnóstico para detectar e verificar qual a influência de uma ou mais observações (em geral de apenas uma) no cálculo estatístico. É exemplo disso a análise de resíduos em modelos de regressão. Mas quando se usa o *bootstrap* não paramétrico apenas se pode contar com os dados e como tal dever-se-á analisar cuidadosamente os resultados da simulação e verificar se ocorrem diferenças substanciais quando uma observação não foi considerada. Efron (1992) propôs uma técnica de diagnóstico de identificação de uma observação à qual denomino *jackknife-after-bootstrap*. Este processo, inicialmente sugerido para verificar a precisão das estimativas *bootstrap*, pode ser usado simultaneamente como técnica de diagnóstico para a detecção de observações discordantes. Este procedimento apesar de ser de reamostragem tem a vantagem do esforço de simulação adicional poder ser evitado (Efron, 1992, Davison e Hinkley, 1997, Cap. 3).

Na sequência da aplicação das técnicas de diagnóstico acima referidas dois possíveis resultados

podem ocorrer, um é a detecção (individual) de observações que parecem ser atípicas, outro é a não detecção de quaisquer observações com essa característica. Serão de seguida analisadas essas duas possibilidades.

Na primeira situação, uma questão que desde logo surge tem a ver com a remoção ou não das observações discordantes. Como já discutido atrás, a atitude que parece ser mais sensata é de nunca desprezar aquelas observações antes de se efectuar um exame cuidadoso para determinar a sua origem, já que nem sempre uma observação atípica é uma má observação e por outro lado a rejeição abrupta pode conduzir a perdas de eficiência.

Quanto à segunda situação, a não detecção, pode dever-se não só à inexistência de facto de observações atípicas mas sim a uma falha desses diagnósticos, quer por efeito de "disfarce" (tradução de *masking effect*) de algumas observações discordantes (isto é, observações que se "mascaram" de boas observações quando de facto não o são), quer ainda por dificuldades inerentes desses métodos, por exemplo, em casos multivariados.

Como foi já mencionado uma alternativa mais fiável do que o uso de técnicas de diagnóstico (identificação) é a utilização de métodos estatísticos robustos pois, e de uma forma geral, estes métodos são construídos de forma a descrever a estrutura que se ajusta à maioria das observações e a identificar os pontos que se desviam dessa estrutura principal. Todavia, no caso concreto do procedimento *bootstrap* esta alternativa tem também algumas desvantagens que se passam a descrever:

i) o ponto de rotura associado a todo o procedimento pode ser pequeno mesmo se baseado num estimador robusto com alto ponto de rotura. Este facto conduz a instabilidade numérica que se reflectirá numa estimação de pior qualidade, ver por exemplo Stromberg (1997), Amado e Pires (1998) e Singh (1998);

ii) aumento do custo de cálculo dado que, em geral, as estimativas robustas são solução de problemas de optimização não convexa sendo obtidas por procedimentos iterativos.

Antes de se passar à descrição do método que se irá propor com o objectivo de ultrapassar estas duas limitações é importante referir os trabalhos que têm sido publicados sobre este tema. Num contexto de estimação de variâncias de estimadores Ghosh *et al.* (1984), Shao (1990, 1992), assim como Stromberg (1997) notaram a falta de robustez do método *bootstrap* clássico. Ghosh *et al.* (1984) mostraram que é necessário impor uma condição nas caudas da distribuição da qual provêm os dados para que a estimativa *bootstrap* da variância da mediana convirja. Numa abordagem teórica a este problema Shao (1990) mostrou que truncando as caudas da distribuição *bootstrap* então a variância *bootstrap* converge para a variância assintótica da estimativa de interesse. No entanto não é evidente como se poderá implementar esse método num contexto finito. Stromberg (1997) propõe a estimação da variabilidade de estimadores robustos com base em procedimentos de reamostragem e recomenda o uso dos métodos que, no seu artigo, denominou por SBS e JK50. O primeiro consiste em efectuar o *bootstrap* como usualmente mas no último passo do procedimento em vez de se usar o estimador clássico da matriz de covariâncias usa-se um estimador S com ponto

de rotura de 0.5. O segundo método não é mais do que a estimação da matriz de covariâncias usando o *deleted-d jackknife*, como definido por Shao (1989), e tomando para d o valor correspondente a metade da dimensão da amostra de trabalho. Note-se porém que a ênfase do trabalho de Stromberg está no cálculo de estimativas da variabilidade de estimadores robustos e não em métodos gerais de robustificação do *bootstrap*.

Dada a importância que os quantis *bootstrap* têm no cálculo de intervalos de confiança e testes de hipóteses Singh (1998) avalia a robustez dessas estimativas. Apresenta uma expressão geral para o cálculo do ponto de rotura do quantil *bootstrap* de ordem p de uma estatística T_n e mostra que esse ponto de rotura pode ser bastante pequeno mesmo para estimativas robustas de localização com alto ponto de rotura. A proposta deste autor para a robustificação do *bootstrap* baseia-se na noção de *Winsorization*. A reamostragem *bootstrap* é então efectuada após a *Winsorization* da amostra usando estimadores robustos de localização e escala. Para certos estimadores L e M de localização univariados, Singh mostrou que os quantis *bootstrap* obtidos por esse método têm um maior ponto de rotura e convergem para os quantis da distribuição assintótica da estatística de interesse. Denomine--se este procedimento de robustificação do *bootstrap* por *bootstrap Winsorized*. Este método pode, então, constituir uma alternativa para ultrapassar a falta de robustez do *bootstrap* clássico. Apesar disso, esta proposta não resolve a questão do elevado tempo de cálculo assim como não é de fácil generalização para problemas multivariados, uma vez que requer a ordenação da amostra de trabalho.

Salibian-Barrera e Zamar (2002a) introduziram um novo método *bootstrap*, que denominaram *bootstrap* robusto, para estimar a distribuição e a variância de estimadores robustos MM de regressão (para definição destes estimadores consultar Yohai *et al.*, 1991). Neste trabalho aplicaram ainda o mesmo método para estimadores–M de localização quando calculados usando estimadores S de escala. Mais recentemente estes mesmos autores propuseram um procedimento que permite obter de forma rápida estimativas *bootstrap* da distribuição e de erros padrões de estatísticas definidas por equações de estimação, Salibian-Barrera e Zamar (2002b). Esta nova proposta baseia-se na aplicação do *bootstrap* em aproximações lineares das equações de estimação que definem as estatísticas. O já referido *bootstrap* robusto, Salibian-Barrera e Zamar (2002a), assim como o Newton-Raphson *bootstrap* de Salibian-Barrera (2002) são casos particulares deste método mais geral. O procedimento proposto por estes autores permite ultrapassar as duas limitações do *bootstrap* clássico referidas atrás. Uma outra alternativa é o método proposto na secção que se segue. Os últimos trabalhos referidos (Salibian-Barrera e Zamar, 2002a e Salibian-Barrera e Zamar, 2002b) foram desenvolvidos em simultâneo com o método que será em seguida descrito e apenas se tomou conhecimento desses trabalhos numa fase bem adiantada deste trabalho, como tal não poderiam deixar de ser referidos mas por questões óbvias de tempo não é possível compará-los com o método que aqui se propõe. Esta comparação constitui, desde já, um desafio para trabalho futuro.

4.2 O método proposto: *bootstrap* robusto baseado na função de influência

Nesta secção introduz-se um novo método inferencial de reamostragem *bootstrap* que permite resolver a falta de robustez do *bootstrap* clássico. A ideia principal é a de introdução de um mecanismo de controlo no plano de reamostragem. O objectivo é proteger o procedimento *bootstrap* contra um dado número de observações atípicas arbitrárias, tornando-o ao mesmo tempo computacionalmente menos dispendioso que a reamostragem de um estimador robusto.

Em geral, quando se usa o método de Monte Carlo para aproximar um estimador *bootstrap* o tipo de reamostragem usada é a uniforme; isto significa que a probabilidade de qualquer elemento da amostra ser seleccionado para uma reamostra é igual a n^{-1}, onde n é a dimensão da amostra de trabalho. Porém, outras técnicas de reamostragem podem ser utilizadas, como seja a reamostragem equilibrada ou por importância (para mais detalhes consultar Shao e Tu, 1995, pág. 206). Por exemplo, a reamostragem por importância é uma técnica padrão usada para melhorar a eficiência das aproximações Monte Carlo (Hammersley e Handscomb, 1964, pág. 57). De um modo geral, pode dizer-se que este método altera deliberadamente as probabilidades de reamostragem, atribuindo maior importância a determinadas observações amostrais que a outras. Essa alteração tem como finalidade aumentar a precisão do estimador, para um determinado número de simulações (ou equivalentemente, reduzir o número de simulações necessárias para obter uma dada precisão). O método que aqui se propõe consiste também numa alteração das probabilidades de reamostragem, mas com a finalidade de reamostrar com menor frequência observações mais influentes (no sentido da função de influência de Hampel, já definida no Capítulo 2) e, simultaneamente, reamostrar de igual modo as observações que pertencem à mesma estrutura principal. A ideia deste novo procedimento é bastante simples: em vez de o estimador *bootstrap* se basear na função de distribuição empírica F_n, baseia-se numa função cujas massas de probabilidade em cada ponto são calculadas com base numa determinada versão empírica da função de influência do funcional de interesse.

Para formalizar o método que será proposto, e doravante designado por *bootstrap* IF, é necessário introduzir algumas considerações e definições.

De forma a modelar a ocorrência de observações atípicas assumir-se-á que a actual distribuição dos dados (uni ou multivariados), F, pertence a uma vizinhança de contaminação-ε,

$$\mathcal{V}_\varepsilon(F_{\boldsymbol{\theta}}) = \{F \in \mathcal{F} : F(x) = (1-\varepsilon)F_{\boldsymbol{\theta}} + \varepsilon G\}, \tag{4.1}$$

onde $\mathcal{F}$ representa o conjunto de todas as funções de distribuição no espaço de interesse, $F_{\boldsymbol{\theta}}$ é um determinado modelo paramétrico "central", G é uma qualquer função de distribuição e $\varepsilon \in [0, 0.5[$ representa a proporção de contaminação.

Como foi já referido no Capítulo 2 existem várias versões finitas da função de influência de um funcional. Não obstante as diferenças entre elas, todas têm em comum o facto de tentarem, de algum modo, "medir" a influência que uma observação tem no valor da estimativa do funcional de interesse. No contexto deste trabalho torna-se necessário definir e usar uma versão empírica

robusta e padronizada da função de influência. Esta função empírica deverá possuir um alto ponto de rotura de forma a evitar efeitos de *masking* e por outro lado deverá ser padronizada para que não dependa da escala das observações. As definições seguintes, 4.1 e 4.2, formalizam estas ideias.

Seja $\mathcal{X}$ um espaço amostral, $\mathcal{X} \subseteq I\!R^m$ e $\mathcal{F}$ o espaço das funções de distribuição sobre $\mathcal{X}$. Considerem--se dois estimadores de um parâmetro $\boldsymbol{\theta}$, com $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq I\!R^p$, um robusto denotado por $\mathbf{T}_n^r$ e outro não robusto representado por $\mathbf{T}_n^{nr}$, os quais podem ser representados por funcionais Fisher consistentes, denotados, respectivamente, por $\mathbf{T}^r$ e $\mathbf{T}^{nr}$ (a definição de consistência à Fisher pode ser consultada em Hampel *et al.*, 1986, pág. 83).

Definição 4.1 – Função de Influência Padronizada (*SIF*): *A função de influência padronizada do funcional* $\mathbf{T}$ *(quer seja de* $\mathbf{T}^r$ *quer de* $\mathbf{T}^{nr}$*) em* F_θ *é dada por*

$$SIF(\mathbf{x};\mathbf{T},F_{\boldsymbol{\theta}}) = \left[IF\left(\mathbf{x};\mathbf{T},F_{\boldsymbol{\theta}}\right)^T V^{-1}(\mathbf{T},F_{\boldsymbol{\theta}}) IF\left(\mathbf{x};\mathbf{T},F_{\boldsymbol{\theta}}\right)\right]^{1/2}, \tag{4.2}$$

onde IF *representa a função de influência e*

$$VA\left(\mathbf{T},F_{\boldsymbol{\theta}}\right) = E_{F_{\boldsymbol{\theta}}}\left[IF\left(\mathbf{x};\mathbf{T},F_{\boldsymbol{\theta}}\right) IF\left(\mathbf{x};\mathbf{T},F_{\boldsymbol{\theta}}\right)^T\right] \tag{4.3}$$

é a variância assintótica de $\mathbf{T}_n$*, que se admite ser definida positiva.*

Analogamente ao que acontecia com a função de influência também é possível definir diversas versões finitas do conceito de SIF. Na Definição 4.2 apresenta-se uma delas. Antes porém assuma-se que, e como usualmente, $SIF(\mathbf{x};\mathbf{T},F_{\boldsymbol{\theta}})$ apenas depende de $F_{\boldsymbol{\theta}}$ através de um vector de parâmetros desconhecidos $\boldsymbol{\Omega}$, isto é, $SIF(\mathbf{x};\mathbf{T},F_{\boldsymbol{\theta}}) = SIF(\mathbf{x};\mathbf{T},\boldsymbol{\Omega})$.

Definição 4.2 – Função de Influência Empírica Padronizada e Robusta (*RESIF*): *A função de influência empírica padronizada e robusta é*

$$RESIF(\mathbf{x};\mathbf{T},\hat{\boldsymbol{\Omega}}^r) = SIF(\widehat{\mathbf{x};\mathbf{T}},F_{\boldsymbol{\theta}}) = SIF(\mathbf{x};\mathbf{T},\hat{\boldsymbol{\Omega}}^r), \tag{4.4}$$

onde $\hat{\boldsymbol{\Omega}}^r$ *é um estimador robusto do vector de parâmetros* $\boldsymbol{\Omega}$*.*

O que esta última definição significa é que se obtém uma versão finita da função de influência padronizada e robusta substituindo na expressão de SIF o vector de parâmetros desconhecidos por um estimador robusto desse vector.

Considere-se então que $SIF(\mathbf{x};\mathbf{T}^r,F_{\boldsymbol{\theta}})$ e $SIF(\mathbf{x};\mathbf{T}^{nr},F_{\boldsymbol{\theta}})$ dependem de $F_{\boldsymbol{\theta}}$ apenas através dos vectores de parâmetros desconhecidos $\boldsymbol{\Omega}_1$ e $\boldsymbol{\Omega}_2$, respectivamente. Face à Definição 4.2 podem ser consideradas duas funções de influência empíricas robustas e padronizadas, $RESIF(\mathbf{x};\mathbf{T}^r,\hat{\boldsymbol{\Omega}}_1^r)$ e $RESIF(\mathbf{x};\mathbf{T}^{nr},\hat{\boldsymbol{\Omega}}_2^r)$.

Observação 4.1 Outras duas funções de influência empíricas padronizadas podiam ser definidas, considerando na Definição 4.2 estimadores não robustos de $\boldsymbol{\Omega}_i$, $i = 1, 2$. Obviamente isso não faria qualquer sentido para a função de influência associada ao funcional $\mathbf{T}^r$. E para a SIF de $\mathbf{T}^{nr}$ não traria qualquer vantagem, em termos de robustez, já que se sabe que sofreria do efeito de *masking*.

■

Observação 4.2 Das duas funções, $RESIF(\mathbf{x}; \mathbf{T}^r, \hat{\mathbf{\Omega}}_1^r)$ e $RESIF(\mathbf{x}; \mathbf{T}^{nr}, \hat{\mathbf{\Omega}}_2^r)$, apenas a segunda é útil para o método que irá ser proposto, uma vez que de uma maneira geral, os estimadores robustos possuem funções de influência limitadas, o que conduz à impossibilidade de detecção de observações "perigosas" para o *bootstrap*. Observe-se que esses pontos podem não produzir efeitos nefastos num procedimento robusto normal mas, e como já referido, poderão causar o colapso do procedimento *bootstrap* ao aparecerem frequentemente. Uma observação atenta permitirá verificar que um alto valor da segunda função, $RESIF(\mathbf{x}; \hat{\mathbf{T}}^{nr}, \hat{\mathbf{\Omega}}_2^r)$, facilmente poderá reconhecer esse tipo de observações. Em 2000, Pison *et al.* usaram, ainda que num contexto diferente, uma função de influência empírica similar à que foi definida em 4.2 (embora não padronizada) e também recomendaram o uso da correspondente à $RESIF(\mathbf{x}; \mathbf{T}^{nr}, \hat{\mathbf{\Omega}}_2^r)$.

■

Observação 4.3 Atendendo ao objectivo final do uso da função $RESIF$ devem ser tomadas algumas precauções, que se passam a descrever:

1. O estimador robusto, $\hat{\mathbf{\Omega}}_2^r$, deve possuir um alto ponto de rotura, já que esta condição será essencial para o bom comportamento do procedimento global.
2. Naturalmente a SIF, equação (4.2), apenas está bem definida se (4.3) for não singular. No caso de isso acontecer, é possível dar uma definição alternativa utilizando a matriz inversa generalizada de Moore-Penrose.
3. É, ainda, importante para o bom funcionamento do método que as funções SIF e $RESIF$ sejam invariantes para certos grupos de transformações. Esses dependerão do problema e modelo em consideração.

■

Para clarificar o que foi referido nas observações considere-se o exemplo seguinte.

Exemplo 4.1 Localização multivariada. Denote-se por $\boldsymbol{\theta}$ o parâmetro de localização de uma distribuição m-multivariada. Considere-se que a distribuição dos dados, F, pertence à seguinte vizinhança de contaminação-ε

$$\mathcal{V}_\varepsilon(F_{\boldsymbol{\theta}}) = \{F \in \mathcal{F} : F(\mathbf{x}) = (1-\varepsilon)F_{\boldsymbol{\theta}} + \varepsilon G\},$$

onde $F_{\boldsymbol{\theta}}$ é a função de distribuição de uma $N_m(\boldsymbol{\theta}, \mathbf{\Sigma})$ e $\mathcal{F}$ é o espaço das funções de distribuição sobre $I\!R^m$ e $G \in \mathcal{F}$.

Seja $\mathbf{T}_n^{nr} = \bar{\mathbf{X}}_n$ o estimador óptimo sob o modelo central, então $\mathbf{\Omega}_2 = (\boldsymbol{\theta}, \mathbf{\Sigma})$ e considere-se que $\hat{\mathbf{\Omega}}_2^r = (\hat{\boldsymbol{\theta}}^r, \hat{\mathbf{\Sigma}}^r)$ representa um estimador robusto de $\mathbf{\Omega}_2$. Facilmente se verifica que

$$RESIF(\mathbf{x}; \mathbf{T}^{nr}, \hat{\mathbf{\Omega}}_2^r) = \left[(\mathbf{x} - \hat{\boldsymbol{\theta}}^r)^T \left(\hat{\mathbf{\Sigma}}^r\right)^{-1} (\mathbf{x} - \hat{\boldsymbol{\theta}}^r)\right]^{1/2}.$$

Mas esta expressão não é mais do que a denominada distância de Mahalanobis robusta, que é usada frequentemente na detecção de observações atípicas em conjuntos de dados multivariados (Rousseeuw e Leroy, 1987; Rousseeuw e Van Zomeren, 1990). Para além disso, esta distância é invariante no grupo de transformações afins. ■

Descreve-se em seguida o método proposto e aqui designado por *bootstrap* IF. Considere-se que $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ é uma amostra disponível gerada pelo modelo $\mathcal{X} \in I\!R^m$, que $\mathbf{T}_n$ é um estimador (robusto ou não) de uma característica populacional, $\boldsymbol{\theta} \in I\!R^p$, e $\mathbf{t}_n$ é o seu valor amostral. Assuma-se que a amostra dada é proveniente de uma distribuição desconhecida, F, que pertence à vizinhança de contaminação-ε, $\mathcal{V}_\varepsilon(F_{\boldsymbol{\theta}})$ definida em (4.1). O objectivo, em termos gerais, é efectuar inferências acerca do parâmetro $\boldsymbol{\theta}$. Denote-se por $\mathbf{T}^{nr}$ o funcional equivalente ao estimador não robusto $\mathbf{T}_n^{nr}$, este último estimador não é mais do que $\mathbf{T}_n$ se este for um estimador não robusto, caso $\mathbf{T}_n$ seja um estimador robusto então $\mathbf{T}_n^{nr}$ é um estimador não robusto do mesmo parâmetro de interesse, $\boldsymbol{\theta}$.

A primeira etapa é a dedução das funções SIF e $RESIF$. De seguida, e para cada ponto da amostra, $\mathbf{x}_i$, $i = 1, 2, \cdots, n$, determine-se o valor da função $RESIF_i = RESIF(\mathbf{x}_i; \mathbf{T}^{nr}, \hat{\boldsymbol{\Omega}}_2^r)$ e definam-se os pesos, w_i, como sendo

$$w_i = I_{[0,c]}\left(|RESIF_i|\right) + \psi\left(c, |RESIF_i|\right) \times I_{]c,+\infty]}\left(|RESIF_i|\right), \tag{4.5}$$

onde $c > 0$ é uma constante de afinação e ψ é uma função não negativa que deve verificar as seguintes condições:

$$\lim_{v \to +\infty} v^2 \psi(c, v) = 0, \tag{4.6}$$

para cada valor de c fixo, e

$$\frac{\partial}{\partial v} \psi\left(c, v\right)_{|v=c} = 0. \tag{4.7}$$

Por último obtenham-se as probabilidades de reamostragem $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ com $p_i = \frac{w_i}{\sum_{j=1}^n w_j}$, $i = 1, 2, \cdots, n$. Seja $G_{n\mathbf{p}}$ a função de distribuição que afecta as probabilidades $p_1, p_2, \ldots, p_n$ a $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Seja $(\mathbf{x}_1^*, \mathbf{x}_2^*, \cdots, \mathbf{x}_n^*)$ uma amostra *bootstrap* IF seleccionada, com reposição, a partir das observações $\{\mathbf{x}_i,\ i = 1, 2, \cdots, n,\}$ e de acordo com a função $G_{n\mathbf{p}}$. Determine-se e calcule-se a estimativa *bootstrap*, $\mathbf{t}_n^* = \mathbf{T}_n(\{\mathbf{x}_i^*,\ \ i = 1, 2, \cdots, n\})$. O estimador $\mathbf{T}_n$ é então avaliado para todas as reamostras possíveis geradas com reposição segundo o modelo $G_{n\mathbf{p}}$. Não sendo possível gerar todas as reamostras usar-se-á, e como usualmente, métodos de simulação Monte Carlo. Note-se ainda que se $\mathbf{p} = (n^{-1}, n^{-1}, \cdots, n^{-1})$ então $G_{n\mathbf{p}} = F_n$, pelo que neste caso o método *bootstrap* IF é equivalente ao *bootstrap* clássico.

Devido à introdução dos pesos (com a excepção para $p_i = 1/n$) este método não pode fornecer nem a distribuição do estimador robusto nem a do não robusto, fornece sim a distribuição de um novo estimador robusto definido implicitamente. A estimativa pontual correspondente na amostra original pode ser calculada como o parâmetro de localização da distribuição *bootstrap* (ver exemplos mais adiante).

A proposição 4.1 justifica a necessidade da condição (4.6) na definição dos pesos, w_i.

Proposição 4.1 *A função $\psi(c, v)$ tem que ser $o(v^{-2})$ quando $v \to +\infty$ de forma a proteger a distribuição bootstrap de efeitos nefastos produzidos pela possível existência de observações atípicas.*

Demonstração:

Seja $\{\mathbf{x}_i,\ i = 1, 2, \cdots, n\}$ uma amostra de dimensão n. Suponha-se que nessa amostra existem alguns pontos atípicos e represente-se por $A = \{a_j : a_j$ um ponto atípico$\}$ o conjunto constituído por esses pontos, com # A $= k < \frac{n}{2}$. Por simplicidade, mas sem perda de generalidade, assuma-se que todas as observações atípicas estão concentradas no mesmo ponto, ξ_{out}, e que $IF(\xi_{out}; T, F) \to +\infty$ quando $\xi_{out} \to +\infty$.

Considere-se que $P(\xi_{out}$ ser reamostrado$) = p$ com $p \leq \frac{1}{n} < \frac{1}{k}$ e represente-se por X a variável aleatória que representa o número de observações atípicas na reamostra de dimensão n. A v.a. X segue então uma distribuição $Bin(n, kp)$.

Escolha-se um percentil de ordem elevada desta distribuição, b_α, com $\alpha \approx 1$, (a ideia é que apenas poucas reamostras tenham mais que b_α pontos atípicos, de modo que apenas os percentis muito extremos da distribuição *bootstrap* sejam afectados). Nesta situação b_α será da ordem de $nkp + z_\alpha\sqrt{nkp}$ onde $z_\alpha = \Phi^{-1}(\alpha)$ (assumindo que kp é muito menor que um). Deste modo os k pontos atípicos irão afectar os percentis de ordem inferior a α da distribuição *bootstrap* no máximo por $b_\alpha v$, onde $v = IF(\xi_{out}; T, F)$.

Se p for da ordem de $|v|^{-1}$, então $b_\alpha v$ permanece ilimitada quando $\xi_{out} \to \infty$. Suponha-se agora que p é da ordem de $|v|^{-2}$, então $b_\alpha v$ é limitada mas $\lim\limits_{v \to +\infty} b_\alpha v \neq 0$, o que significa que os pontos atípicos poderão introduzir um viés elevado. Sendo assim, conclui-se que o valor de p tem que ter um decrescimento mais rápido que $|v|^{-2}$. ■

A segunda condição imposta à função ψ, (4.7), é apenas uma condição de suavidade de forma a preservar a eficiência do procedimento para conjuntos de dados "limpos" ou com um número moderado de pontos atípicos e pode também ser vista como uma rejeição suave de observações atípicas em oposição à rejeição abrupta já referida no início do capítulo.

Uma família flexível de funções, de onde se poderá escolher a função ψ , é o núcleo das funções de densidade de probabilidade da distribuição t–Student, e cuja forma limite corresponde à função de densidade Normal a menos duma constante. Isto é, funções da forma:

$$\Psi(x; l, d, \gamma) = \begin{cases} \left[1 + \frac{(x-l)^2}{\gamma d^2}\right]^{-\frac{\gamma+1}{2}} & 0 < \gamma < \infty \\ \exp[-\frac{(x-l)^2}{2d^2}] & \gamma = \infty, \end{cases} \tag{4.8}$$

onde l é o parâmetro de localização, d é o parâmetro de escala e γ corresponde ao parâmetro de forma.

A condição (4.6) é satisfeita se $\gamma > 1$ e a condição (4.7) se o parâmetro de localização corresponder à já referida constante de afinação, isto é se $l = c$. Usando esta família introduzem-se mais dois parâmetros que é necessário escolher para além da constante de afinação. Uma vez que o parâmetro

de escala não afecta a taxa de decrescimento é possível reduzir o número de parâmetros considerando $l = d = c$.

Como usualmente acontece na estimação robusta ao escolher a constante de afinação estabelece-se um compromisso entre eficiência e robustez. Valores elevados de c conduzem a procedimentos mais eficientes (em relação ao modelo central) mas são também menos robustos. Por outro lado, um baixo valor de c assegurará um procedimento mais robusto mas menos eficiente. Uma forma de resolver o problema da escolha desta constante consiste na determinação da probabilidade, sob o modelo central $F_{\boldsymbol{\theta}}$, de uma observação vir a ter peso inferior a um. Como é natural essa probabilidade deve ter um valor pequeno. No caso presente a fixação desta constante pode ser considerada como uma dificuldade do método proposto já que pode não ser conhecida a distribuição da SIF. Porém, esta dificuldade pode ser contornada através da realização de um estudo de simulação prévio e usando a ideia já referida. Ou seja, simula-se a distribuição da SIF sob o modelo paramétrico central, $F_{\boldsymbol{\theta}}$ e escolhe-se para valor de c um alto percentil dessa distribuição. Quando se está no modelo central, $F_{\boldsymbol{\theta}}$, a escolha de um percentil elevado da distribuição da SIF conduzirá quase sempre a pesos unitários para todas as observações o que torna o procedimento *bootstrap* IF praticamente equivalente ao *bootstrap* usual. As considerações seguintes poderão ser usadas para escolher o referido percentil.

Seja $p_c = P(SIF > c)$, então a probabilidade de ter que atribuir um peso mais baixo que o usual a pelo menos uma observação numa amostra de dimensão n é $p_{c,n} = 1 - (1 - p_c)^n$. Se Y representar o número de amostras *bootstrap* com pelo menos uma observação de baixo peso então $Y \sim Binomial(B, p_{c,n})$. Finalmente, fixe-se um risco, q, de uma proporção, a, de amostras *bootstrap* ser afectada, isto é, $P(Y \geq aB) = q$. Se, por exemplo, $a = 0.10$, $B = 1000$, $q = 0.05$ e $n = 100$ cálculos imediatos conduzem aos seguintes valores $p_{c,n} = 0.084785$ e $p_c = 0.0009$. Observe-se que estes cálculos não consideram a grandeza dos pesos e, da maneira como é proposto serem construídos (ver a expressão 4.5), o seu efeito, apesar de presente, pode ser bastante pequeno, já que os pesos são escolhidos de forma a decrescerem suavemente.

Para além da função de influência apresentada em (4.4) poderia ser usada outro tipo de funções de influência empíricas, como por exemplo as definidas no Capítulo 2. No entanto é a $RESIF$ aquela que apresenta melhores propriedades já que não depende da escala das observações e permite algum afastamento dos dados ao modelo central considerado. Por exemplo, a apresentada por Hampel *et al.* (1986, pág. 93) ou a curva de sensibilidade de Tukey dependem da escala das observações. No que diz respeito à obtida por *jackknife* depende também da escala das observações e além disso pode não reflectir de boa forma os casos atípicos, já que pode sofrer de efeito de *masking*. Além disso, em termos computacionais esta função requer reamostragem o que conduz a um aumento do tempo de cálculo do procedimento global *jackknife* mais *bootstrap*.

O método *bootstrap* IF constitui uma versão robusta do *bootstrap* clássico já que, de acordo com a expressão (4.5), os pontos discordantes terão pesos de reamostragem pequenos. Note-se ainda que pontos discordantes extremos receberão um peso praticamente nulo, não afectando os cálculos de reamostragem. Por outro lado, existe um menor dispêndio de tempo de cálculo comparado com

o uso do *bootstrap* clássico com um estimador robusto (dado que as estimativas robustas só são calculadas uma vez para a determinação da $RESIF$).

Com o plano de reamostragem proposto garante-se que uma grande quantidade de réplicas *bootstrap* ficam livres da influência nefasta de observações discordantes. Todavia, existe a possibilidade de que valores mais extremos possam ser afectados. Isto não é importante caso se pretenda apenas percentis da distribuição para intervalos de confiança, por exemplo, mas poderá afectar a média ou a variância *bootstrap*. Assim sendo, essas quantidades devem ser substituídas por medidas robustas, tais como a média ou a variância aparadas ou o intervalo interquartis (tal como foi também proposto por Stromberg, 1997, para a reamostragem de um estimador robusto).

O método *bootstrap* IF exposto é perfeitamente geral e aplicável a qualquer estimador. O facto de aliar algum conhecimento teórico à reamostragem pode constituir uma desvantagem, mas é precisamente esse conhecimento teórico que o torna mais coerente e possuindo as vantagens já referidas.

4.3 Exemplos de aplicação e testes computacionais

Nesta secção serão apresentadas três situações de estimação com o objectivo de avaliar a qualidade do procedimento proposto. O primeiro problema tem como objectivo fazer inferências sobre a localização univariada, o segundo problema estudado é bivariado e relativo a inferências sobre o coeficiente de correlação linear. O último problema analisado é a selecção de variáveis em análise discriminante linear. Para cada problema de estimação referido apresenta-se a função de influência do funcional em estudo, as distribuições usadas assim como a expressão de cálculo dos pesos de reamostragem para o *bootstrap* IF.

O comportamento do método *bootstrap* IF, quer para amostras infinitas, quer para amostras finitas, não é simples de estudar teoricamente. Nesta situação informação válida acerca do método pode ser obtida usando simulações devidamente delineadas. Assim, para cada exemplo de aplicação estudado serão apresentados estudos de simulação sendo a eficácia do método *bootstrap* IF e a também a sua importância quantificada por comparação com outros procedimentos. O método proposto e os outros procedimentos usados foram implementados em linguagem *S*. A maioria dos testes computacionais foram realizados num computador *Pentium II* a 550 *MHz* com 128 *Mb* de memória *Ram*. Esses programas encontram-se no Apêndice C, estando ainda disponíveis na página da Internet http://www.math.ist.utl.pt/∼camado/.

4.3.1 Inferência para a localização univariada num modelo de localização e escala

O problema aqui apresentado corresponde a uma situação de grande interesse prático e que à primeira vista parece ser simples: a determinação de intervalos de confiança para um parâmetro de

localização com escala desconhecida em distribuições univariadas. A questão principal consiste no desconhecimento acerca do modelo gerador dos dados. Não sabendo isso torna-se difícil efectuar quaisquer inferências sobre o parâmetro de interesse. Como também já mencionado, os métodos puramente de reamostragem contêm a fraqueza de apenas contarem com os dados e poderem conduzir a conclusões erradas se estes não forem de "qualidade". É por esta razão que considerações sobre robustez devem ser sempre feitas mesmo em problemas tão simples como este.

Sejam $X_1, X_2, \cdots, X_n$ n variáveis aleatórias reais satisfazendo

$$X_i = \mu + \sigma\epsilon_i, \qquad i = 1, 2, \cdots, n, \tag{4.9}$$

onde ϵ_i são variáveis aleatórias independentes e identicamente distribuídas não dependendo de nenhum parâmetro desconhecido. O interesse reside na estimação de μ sendo a escala, σ, um parâmetro perturbador. Para além disso vai supor-se que o modelo gerador dos dados, F, pertence à seguinte vizinhança de contaminação-ε,

$$\mathcal{V}_\varepsilon(F_{(\mu,\sigma)}) = \{F \in \mathcal{F} : F(x) = (1-\varepsilon)F_{(\mu,\sigma)} + \varepsilon G\}, \tag{4.10}$$

e $F_{(\mu,\sigma)}$ é o modelo paramétrico "central" que se vai assumir como sendo $N(\mu, \sigma^2)$, $\mathcal{F}$ é o espaço das funções de distribuição sobre $I\!R$, $G \in \mathcal{F}$ e $\varepsilon \in [0, 0.5[$ representa a percentagem de contaminação.

Considerem-se os seguintes estimadores para μ, $T_n^{nr} = \bar{X}_n$ e $T_n^r = LTS$, onde o primeiro é a usual média aritmética e o segundo é o já referido *Least Trimmed Squares*, estimador robusto de localização proposto por Rousseeuw (1984, 1985). Observe-se que este estimador é o equivalente univariado da componente de localização do estimador de matriz de covariâncias de determinante mínimo MCD.

Designe-se por T^{nr} o funcional equivalente a $\bar{X}_n$. Tem-se então que a função de influência de T^{nr} é dada por $IF(x; T^{nr}, F) = x - \mu$ e a variância assintótica é $VA(\bar{X}_n, F) = \sigma^2$. Logo

$$SIF(x; T^{nr}, F) = |x - \mu|/\sigma. \tag{4.11}$$

Esta função depende apenas do vector de parâmetros $\Omega = (\mu, \sigma)$. Seja

$$\hat{\mathbf{\Omega}}^r = (med(X), MAD), \tag{4.12}$$

um estimador robusto desse vector, onde $med(X)$ é o estimador mediana amostral e MAD é o, já referido, desvio mediano absoluto, isto é, $MAD = med\,|X - med(X)| \times k$, onde k é uma constante para obter consistência sob o modelo central.

Uma versão robusta e padronizada da função de influência de T^{nr} é então dada por

$$RESIF(x; T^{nr}, \hat{\mathbf{\Omega}}^r) = |x - med(X)|/MAD, \tag{4.13}$$

com $\hat{\mathbf{\Omega}}^r$ definido em (4.12).

Seja $(x_1, x_2, \cdots, x_n)$ uma amostra univariada proveniente do modelo (4.9) concretização de uma amostra aleatória de dimensão n, $(X_1, X_2, \cdots, X_n)$, que é proveniente de um população com distribuição $F \in \mathcal{V}_\varepsilon(F_{(\mu,\sigma)})$. Com base na amostra dada pretende-se determinar um intervalo com

$(1-\alpha)\times 100\%$ de confiança para μ. A obtenção deste intervalo será efectuada com base nos seguintes métodos:

- **Método clássico** (TI): intervalo t, isto é $\left(\bar{x}_n \pm F^{-1}_{t_{(n-1)}}\left(1-\frac{\alpha}{2}\right) s_{\bar{x}_n}\right)$, onde $\bar{x}_n$ é a média da amostra $(x_1, x_2, \cdots, x_n)$, $F^{-1}_{t_{(n-1)}}\left(1-\frac{\alpha}{2}\right)$ é o quantil de probabilidade $1-\frac{\alpha}{2}$ da distribuição t–Student com $n-1$ graus de liberdade e $s^2_{\bar{x}_n} = s^2_{n-1}/n$ é uma estimativa da variância de $\bar{X}_n$, sendo s^2_{n-1} a variância corrigida da amostra. Observe-se que quando se utiliza o estimador LTS este intervalo é análogo mas substituindo os estimadores clássicos de localização e escala, pelos estimadores robustos respectivos.

- ***Bootstrap* clássico ou uniforme** (UB): intervalo obtido pelo método do percentil usando o *bootstrap* clássico (veja-se expressão 3.31).

- ***Bootstrap* baseado na função de influência** (IFB): intervalo obtido também pelo método do percentil mas aplicando o *bootstrap* IF, com $RESIF$ definida pela expressão (4.13). As probabilidades de reamostragem são determinadas com base em (4.5) e escolhendo para valor da constante de afinação, c, um percentil elevado da distribuição da função SIF sob o modelo central $F_{(\mu,\sigma)} = N(\mu, \sigma^2)$. Observando-se que, neste caso, a distribuição de SIF^2 é um qui-quadrado central com um grau de liberdade, com SIF definida por (4.11), fixou--se $c = \sqrt{F^{-1}_{\chi^2_{(1)}}(0.99)}$. A função ψ foi escolhida na família das funções Ψ definida em (4.8), concretamente $\psi(x) = \exp\left[-\frac{1}{2}(x-c)^2 c^{-2}\right]$ (dado que o parâmetro de escala da família de funções, Ψ, definida em (4.8) não afecta a taxa de decrescimento dos pesos fixou-se $d = c$, reduzindo deste modo o número de parâmetros a ajustar).

  Observe-se ainda que, de acordo com a Observação 4.1, se utilizam as mesmas probabilidades de reamostragem independentemente do estimador (robusto ou não).

- ***Bootstrap Winsorized*** (WB): intervalo obtido segundo o método do percentil mas usando o método de reamostragem proposto por Singh (1998). Dada a amostra atrás referida considere-se que $(x_{1:n}, x_{2:n}, \cdots, x_{n:n})$ é a sua versão ordenada de forma crescente. A amostra *Winsorized*, $(x^w_1, x^w_2, \cdots, x^w_n)$, com proporção $\beta > 0$ de *Winsorization* é obtida a partir de

$$x^w_i = \begin{cases} x_{k+1:n}, & \text{se} \quad x_i \leq x_{k:n}, \\ x_{n-k:n}, & \text{se} \quad x_i \geq x_{n-k+1:n}, \\ x_i, & \quad\quad \text{caso contrário}, \end{cases}$$

  onde $k = [n\beta]$. Um estimador *bootstrap Winsorized* é então obtido usando a função de distribuição empírica da amostra $\{x^w_i\}_{i=1,2,\cdots,n}$ em vez da usual $F_n(x)$ como no *bootstrap* clássico. Singh propõe, no caso do estimador mediana, $\beta \leq \frac{1}{4}$ e no caso do estimador robusto de escala, intervalo interquartis, aconselha o uso de $\beta < \frac{1}{4}$. A ideia geral, quanto à escolha de um valor de β, que sai do trabalho de Singh é que isso depende de alguma forma do estimador em estudo, mas como é óbvio esse valor terá que verificar sempre $0 < \beta < \frac{1}{2}$. No estudo presente optou-se por fixar $\beta = 0.1$, já que este valor corresponde à máxima

contaminação determinística ensaiada. Este valor será o mesmo quer para o estimador $\bar{X}_n$ quer para o estimador LTS;

- ***Deleted-d jackknife*** (DJ): intervalo obtido usando os percentis da distribuição do estimador de interesse estimada através de uma adaptação do já referido histograma *jackknife* cumulativo (veja-se o Capítulo 3, expressão (3.18)). Esta adaptação é simples já que basta seguir um raciocínio semelhante ao método percentil do *bootstrap*. Recordando que o número de avaliações que devem ser feitas para determinar uma estimativa *deleted-d jackknife* é $N = \binom{n}{d}$ e que a complexidade computacional, para um dado n, aumenta rapidamente com d foi usado um método de subamostragem aleatório. Ou seja, escolhem-se aleatoriamente m subconjuntos de entre todos os subconjuntos de $\{1, 2, \cdots, n\}$ com cardinal d. Esta selecção aleatória pode ser realizada com ou sem reposição. Neste estudo optou-se por realizá-la com reposição já que o valor de m é, de uma forma geral, muito menor que N e nestes casos fazer amostragem com ou sem reposição origina praticamente os mesmos resultados. Quanto à escolha do valor para m teve-se em conta o exemplo em Shao e Tu (1995, pág. 200), assim estes autores aconselham o uso de $m = n^\delta$, $\delta > 1$, escolhendo para esse exemplo $\delta = 3$. É este valor que é usado também neste estudo, isto é, considerou-se $m = n^3$.

Em todos os métodos de reamostragem foi usada a versão não paramétrica apesar de se ter fixado a distribuição dos dados, uma vez que numa situação prática esta não é conhecida.

Como o objectivo do estudo é comparar os métodos acima em termos de robustez e eficiência, vão ser consideradas para $(X_1, X_2, \cdots, X_n)$ distribuições que possuem algumas observações atípicas com a excepção do primeiro caso o qual constituirá o denominado caso de controlo. Assim, procedeu-se à geração de observações pseudo-aleatórias de acordo com cada uma das quatro situações seguintes:

- Normal, NOR:
$$X_i \sim N(\mu = 0, \sigma^2 = 1), \quad i = 1, 2, \cdots, n;$$

- Normal Simetricamente ε-Contaminada com factor de escala f_e, NSC–ε:
$$X_i \sim (1-\varepsilon)N(0,1) + \varepsilon N(0, f_e), \quad i = 1, 2, \cdots, n,$$
com $f_e = 9$ e $\varepsilon = 0.25$.

- Normal Assimetricamente ε-Contaminada, NAC–ε:
$$X_i \sim N(0,1) \quad \text{para} \quad i = 1, 2, \cdots, n - [\varepsilon n] \quad \text{e} \quad X_i \sim \Delta_x \quad \text{para} \quad i = n - [\varepsilon n] + 1, \cdots, n,$$
com $\varepsilon = 0.1$ e $x = 10$;

- Dividida, DIV:
$$X_i \sim N(0,1)/Unif(0,1), \quad i = 1, 2, \cdots, n,$$
também denominada distribuição *Slash* de Tukey.

A distribuição NOR constitui o controlo, isto é, o parâmetro a estimar é o valor médio μ que foi fixado em zero. A segunda distribuição corresponde a gerar dados provenientes da vizinhança de contaminação-ε como definida em (4.1) com $F_{(\mu,\sigma)} = N(0,1)$, $G = N(0,9)$ e proporção de contaminação $\varepsilon = 0.25$. Observe-se ainda que o primeiro caso, NOR, corresponde ao caso particular de NSC-ε para $\varepsilon = 0$. Gerar dados a partir do modelo NSC é uma forma simples e razoável de modelar desvios simétricos ao modelo central. Note-se, no entanto, que os dados gerados poderão pertencer à parte central da amostra não sendo valores extremos e que, para além disso, o número de observações contaminantes em cada amostra é uma variável aleatória. Com a situação NAC pretende-se ilustrar uma contaminação assimétrica e determinística, onde $\varepsilon\times100\%$ das observações estão concentradas num mesmo ponto $x = 10$. No que diz respeito à distribuição DIV, pretende-se gerar amostras provenientes de uma distribuição de caudas muito pesadas, a qual constitui uma situação propícia ao "aparecimento" de observações discordantes. Em alternativa à distribuição Dividida poderia ter sido considerada a distribuição de Cauchy mas optou-se pela primeira por ser considerada mais desfavorável.

Quanto às dimensões das amostras a gerar optou-se pelos valores $n = 20$ e $n = 50$. Relativamente ao número de simulações, $nsimul$, foi necessário distinguir duas situações: quando se usa o estimador não robusto, $\bar{X}_n$, fixou-se $nsimul = 1000$, quando se usa o estimador robusto LTS reduziu-se esse número para metade, $nsimul = 500$, já que o tempo de cálculo nesta situação corresponde a cerca de quarenta vezes o da primeira (veja-se Exemplo 1.1).

A probabilidade de cobertura foi fixada em $1 - \alpha = 0.90$. Para avaliar a eficácia dos vários procedimentos (TI, UB, IFB, WB e DJ) calculam-se os comprimentos dos intervalos obtidos e a probabilidade estimada de não cobertura. Quando se considera o modelo Normal, NOR, sabe-se que o procedimento TI corresponde ao procedimento óptimo, isto é, o intervalo de confiança aleatório obtido por este método contém o valor esperado $\mu = 0$ com probabilidade $1 - \alpha$ e o valor esperado do seu comprimento é mínimo.

As Tabelas 4.1 e 4.2 apresentam as seguintes características sumárias, respectivamente para os estimadores $\bar{X}_n$ e LTS:

- a probabilidade estimada de não cobertura, isto é, o número de vezes que $\mu = 0$ não pertence ao intervalo construído dividido por $nsimul$;
- a média e o desvio padrão dos comprimentos dos $nsimul$ intervalos construídos.

A obtenção das características acima descritas pode ser realizada por dois processos:

1. para cada combinação, distribuição e dimensão, gera-se uma amostra obtendo-se para esta amostra cinco intervalos de confiança, um por cada método (TI, DJ, UB, IFB, WB). Deste modo os resultados obtidos são dependentes;
2. para cada combinação, distribuição e dimensão, geram-se cinco amostras, uma por cada método, obtendo-se os cinco intervalos de confiança. Esta situação conduz à independência

dos resultados;

O primeiro processo conduz a uma redução na variabilidade dos resultados, já que a amostra de trabalho é a mesma. Tem também a vantagem da geração de um menor número de amostras. A desvantagem associada a este critério diz respeito à maior complexidade na análise dos resultados, já que se espera que haja uma correlação positiva entre os resultados dos vários métodos. Este facto conduzirá à necessidade de determinar essas correlações e usar métodos adequados na comparação dos resultados.

O segundo processo tem a desvantagem clara da necessidade da geração de um maior número de amostras mas por outro lado a vantagem também evidente da redução do grau de complexidade da análise dos resultados, pois as comparações podem ser efectuadas directamente.

Neste estudo optou-se pelo segundo processo, uma vez que considerações relativas ao tempo computacional não são importantes na geração das amostras, isto é a geração de 1000 amostras ou de 5000 não influi grandemente no tempo de execução final. Como exemplo, tome-se a geração de uma amostra de dimensão $n = 50$ segundo a distribuição NSC–0.25 (a mais dispendiosa de cálculo de entre as três apresentadas): o tempo computacional para obter essa amostra é de apenas 0.01 segundos, o que é muito pequeno quando comparado com o tempo de execução dos métodos de reamostragem.

Quanto à análise dos resultados, a primeira comparação a fazer é relativa ao comportamento dos vários procedimentos quando se usa na fase de reamostragem o estimador clássico $\bar{X}_n$. Analisando a Tabela 4.1, um primeiro aspecto relevante tem a ver com o bom desempenho, para dados Normais, dos métodos TI, DJ, UB e IFB. As probabilidades estimadas de cobertura são semelhantes, assim como as médias e desvios padrões dos comprimentos dos intervalos nestes quatro procedimentos, facto este que se torna mais saliente para $n = 50$.

Quando os dados são não Normais é apenas o procedimento *bootstrap* IF, IFB, que tem um desempenho uniforme, em termos de probabilidade de cobertura, ao longo das várias distribuições consideradas (NSC-0.25, NAC-0.1 e DIV). Verifica-se ainda que esses valores melhoram à medida que n aumenta. Para o método IFB, e no que diz respeito ao comprimento médio do intervalo, verifica-se um aumento ao considerar-se a seguinte ordem das distribuições, NAC-0.1, NSC-0.25 e DIV, o que pode ser justificado pelo aumento da percentagem de observações atípicas de cada distribuição. Ou seja, em NSC-0.25 espera-se que sejam 25% as observações atípicas, cujo valor é superior às geradas por NAC-0.1 (que é de facto 10%) e que as observações geradas por DIV possuam um maior número de observações deste tipo relativamente a NSC e NAC. Independentemente desta ordem crescente dos comprimentos médios para as várias distribuições não Normais é para o método IFB que esses valores são menores quando comparados com os restantes métodos (mantendo simultaneamente a taxa de cobertura aceitável).

Outro facto a salientar é o fraco desempenho do método *bootstrap Winsorized*, WB. As probabilidades de não cobertura apresentam, independentemente da distribuição, valores superiores aos dos restantes, o que origina comprimentos médios inferiores. Relembre-se, porém, que este proce-

dimento foi proposto por Singh (1998) para estimadores robustos.

Quanto ao *deleted–$n/2$ jackknife*, DJ, verifica-se que o seu desempenho é relativamente semelhante ao do *bootstrap* clássico, UB, não havendo portanto vantagem na sua utilização em distribuições com observações atípicas.

Em segundo lugar analise-se o comportamento dos vários procedimentos relativos à estimação por LTS na fase de reamostragem. A Tabela 4.2 evidencia um pior desempenho dos vários métodos quando comparado com o desempenho para o estimador $\bar{X}_n$. Esta diferença é mais significativa na distribuição Normal, o que já era esperado. Usar o método TI para a construção de intervalos de confiança quando se usa o LTS não é razoável e os resultados assim o demonstram.

Quanto aos métodos de reamostragem, verifica-se que é para o *deleted–$n/2$ jackknife* que os resultados são piores. Em relação aos métodos *bootstrap* UB e IFB, os resultados são bastante semelhantes. Mais uma vez o BW não teve o desempenho esperado mesmo tratando-se de um estimador robusto, pois apesar da cobertura ser razoável observa-se um aumento dos comprimentos dos intervalos.

Como último comentário refira-se que as conclusões de Stromberg (1997) relativas ao *deleted-d jackknife* não se podem estender a esta situação. Aparentemente o *deleted-d jackknife* tem pior desempenho na estimação de quantis do que na estimação de medidas de precisão.

Como conclusão e resumo deste estudo, consideram-se simultaneamente as duas tabelas e apresenta-se, para cada distribuição, a ordenação por comprimento médio dos intervalos para os métodos com cobertura aceitável:

- Distribuição NOR:

$$TI(\bar{X}_n) \simeq DS(\bar{X}_n) \simeq UB(\bar{X}_n) \simeq IFB(\bar{X}_n) < UB(LTS) \simeq IFB(LTS) < WB(LTS);$$

- Distribuição NSC–0.25:

$$IFB(\bar{X}_n) < TI(\bar{X}_n) \simeq DJ(\bar{X}_n) < UB(\bar{X}_n) < UB(LTS) \simeq IFB(LTS) < WB(LTS);$$

- Distribuição NAC–0.1:

$$IFB(\bar{X}_n) < UB(LTS) \simeq IFB(LTS) < WB(LTS) < TI(LTS);$$

- Distribuição DIV:

$$IFB(\bar{X}_n) < UB(LTS) \simeq IFB(LTS) < WB(LTS) < DJ(\bar{X}_n) < TI(\bar{X}_n) \simeq TI(LTS).$$

Como ordenação comum às quatro distribuições consideradas tem-se:

$$IFB(\bar{X}_n) < UB(LTS) \simeq IFB(LTS) < WB(LTS),$$

concluindo-se que numa situação em que se desconhece a distribuição a escolha mais razoável recai sobre o método *bootstrap* IF, com a aplicação do estimador clássico, $\bar{X}_n$, nas reamostras.

Tabela 4.1: Probabilidades estimadas de não cobertura, assim como médias e desvios padrões dos comprimentos dos intervalos obtidos quando se aplica às reamostras *bootstrap* o estimador não robusto, $\bar{X}_n$.

				Comprimento do intervalo			
		(alvo: 0.100)		$n = 20$		$n = 50$	
Distribuição		Não cobertura			Desvio		Desvio
amostral	Método	$n = 20$	$n = 50$	Média	padrão	Média	padrão
NOR	TI	0.101	0.099	0.7646	0.1252	0.4727	0.0484
	DJ	0.133	0.107	0.7236	0.1179	0.4668	0.0477
	UB	0.126	0.105	0.7070	0.1165	0.4583	0.0475
	IFB	0.129	0.108	0.7061	0.1155	0.4584	0.0478
	WB	0.236	0.195	0.5570	0.1150	0.3764	0.0465
NSC–0.25	TI	0.083	0.097	1.2799	0.3621	0.8073	0.1322
	DJ	0.121	0.101	1.2471	0.3395	0.7948	0.1301
	UB	0.127	0.117	1.1811	0.3337	0.7813	0.1283
	IFB	0.112	0.100	1.0770	0.2644	0.7107	0.1020
	WB	0.234	0.235	0.7914	0.2203	0.7813	0.1283
NAC–0.1	TI	0.180	0.952	2.4879	0.0655	1.5049	0.0244
	DJ	0.226	0.979	2.3452	0.0765	1.4555	0.0311
	UB	0.402	0.985	2.2919	0.0767	1.4555	0.0375
	IFB	0.136	0.110	0.8299	0.2096	0.5410	0.0928
	WB	0.327	0.574	0.7158	0.1391	0.4937	0.0659
DIV	TI	0.049	0.061	22.7480	163.0601	26.4781	202.2434
	DJ	0.119	0.112	13.9845	68.4645	12.1911	35.4033
	UB	0.142	0.171	20.0932	141.5676	24.1843	181.1322
	IFB	0.120	0.117	1.8251	0.4686	1.1783	0.1883
	WB	0.302	0.322	1.8900	1.1634	1.1495	0.3450

Tabela 4.2: Probabilidades estimadas de não cobertura, assim como médias e desvios padrões dos comprimentos dos intervalos obtidos quando se aplica às reamostras *bootstrap* o estimador robusto, LTS.

				Comprimento do intervalo			
		(alvo: 0.100)		$n = 20$		$n = 50$	
Distribuição		Não cobertura			Desvio		Desvio
amostral	Método	$n = 20$	$n = 50$	Média	padrão	Média	padrão
	TI	0.464	0.506	0.7569	0.1273	0.4732	0.0480
	DJ	0.466	0.425	1.3387	0.4928	1.0803	0.3915
NOR	UB	0.072	0.074	1.3965	0.4955	1.1146	0.3887
	IFB	0.080	0.074	1.3944	0.4972	1.1135	0.3901
	WB	0.058	0.068	1.5558	0.4836	1.4252	0.4282
	TI	0.190	0.282	1.3381	0.3787	0.8094	0.1378
	DJ	0.422	0.376	1.4579	0.5349	1.1475	0.4183
NSC–0.25	UB	0.092	0.070	1.5165	0.5825	1.1341	0.4051
	IFB	0.100	0.078	1.5193	0.5945	1.1432	0.4098
	WB	0.056	0.066	1.8774	0.6479	1.5641	0.5518
	TI	0.000	0.026	2.4934	0.0669	1.5056	0.0248
	DJ	0.482	0.354	1.3295	0.4080	1.0300	0.3408
NAC–0.1	UB	0.086	0.070	1.4258	0.4969	1.0573	0.3673
	IFB	0.116	0.096	1.4426	0.5492	1.0839	0.4068
	WB	0.062	0.082	1.7006	0.5788	1.2954	0.4499
	TI	0.034	0.010	22.0112	102.0801	18.9579	87.4998
	DJ	0.316	0.282	2.3675	0.9287	1.6084	0.5565
DIV	UB	0.054	0.082	2.4999	0.9793	1.6231	0.5635
	IFB	0.068	0.084	2.4799	1.0614	1.6575	0.6237
	WB	0.048	0.080	3.2428	1.3103	2.1153	0.9197

Em seguida apresenta-se uma comparação entre os métodos anteriores e também o proposto por Salibian-Barrera (2000, 2002) num conjunto de dados apresentado e analisado em Salibian-Barrera (2000, pág. 95). Este método será denominado por *bootstrap* M, MB.

No estudo de simulação não foi considerado o método MB pois a simulação foi efectuada anteriormente ao seu conhecimento.

Exemplo 4.2 Dados de tensão sanguínea. Num programa de hipertensão observou-se durante dez meses a tensão sanguínea de vários pacientes (Rosner, 1977). Uma amostra desses valores para um determinado indivíduo foi a seguinte: $(40, 75, 80, 83, 86, 88, 90, 92, 93, 95)$. O objectivo é a estimação do valor médio, μ, da tensão sanguínea desse indivíduo. Com este objectivo construíram-se intervalos de confiança segundo os vários procedimentos *bootstrap* atrás descritos e usando o estimador $\bar{X}_n$ na fase de reamostragem. Na Tabela 4.3 apresentam-se os vários intervalos obtidos e ainda os valores calculados por Salibian-Barrera (2000, pág. 96) pelo método MB, assim como os valores obtidos para o intervalo clássico sem a observação aparentemente mais extrema (40), denotado por TI_{-40}. De forma a que os resultados fossem comparáveis aos deste autor efectuaram-se também 50000 réplicas *bootstrap*. Todos os restantes factores se mantiveram iguais aos do estudo de simulação apresentado atrás.

Tabela 4.3: Intervalos de confiança a 95% e 99% de confiança para o valor médio, μ, da tensão sanguínea de um indivíduo. Os valores assinalados com * são os apresentados em Salibian-Barrera (2000, pág. 96).

	95%		99%	
Método	Limite inferior	Limite superior	Limite inferior	Limite superior
TI	70.7	93.7	65.7	98.7
TI_{-40}	81.8	91.9	79.5	94.2
DJ	74.2	89.2	73.2	91.6
UB	71.5	89.7	67.4	90.9
IFB	83.3	90.7	82.1	91.6
WB	81.3	89.5	80.0	90.5
MB	79.7*	92.0*	78.5*	94.9*

Observando a Tabela 4.3 verifica-se que, para os níveis de confiança considerados, o intervalo com menor comprimento é o associado ao procedimento *bootstrap* IF. No caso de se assumir que os dados provêm de uma população com distribuição Normal com alguma percentagem de contaminação, e tendo em conta os resultados do estudo de simulação acima, pode dizer-se que os intervalos obtidos por IFB são os melhores.

Na Figura 4.1 apresentam-se representações gráficas das distribuições *bootstrap* estimadas pelos métodos UB, IFB, e WB. Verifica-se que a distribuição *bootstrap* estimada pelas duas versões robustas do *bootstrap* é bastante menos assimétrica que a correspondente ao *bootstrap* clássico. Para este método, UB, salienta-se o elevado peso da cauda esquerda da distribuição empírica

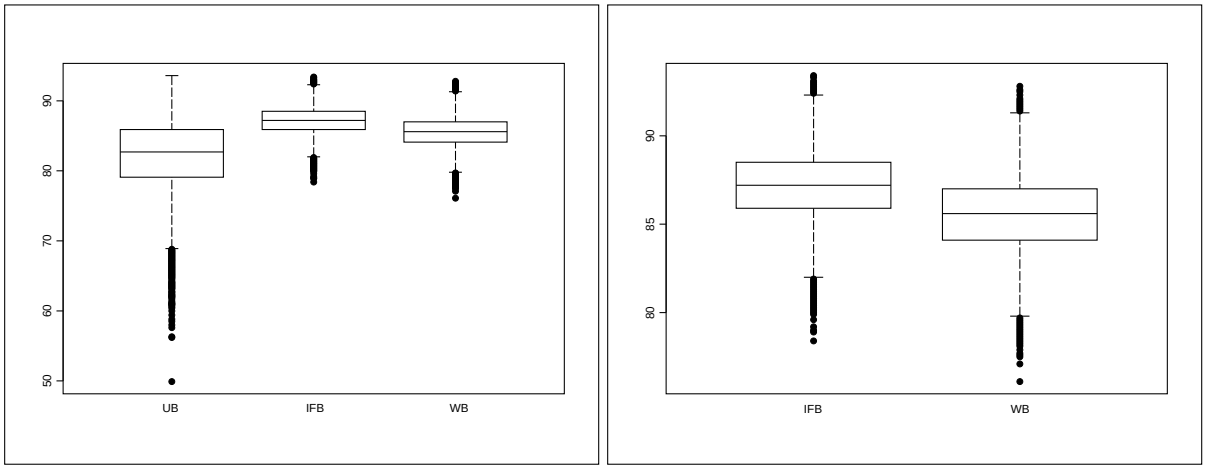


Figura 4.1: Diagramas de extremos-e-quartis para comparação das 50000 estimativas *bootstrap* obtidas pelos vários métodos, o gráfico mais à direita representa em pormenor parte do gráfico esquerdo.

respectiva, facto este que se deve à selecção da observação atípica um grande número de vezes. Pela observação do gráfico à direita, onde aparecem apenas os dois métodos robustos, verifica-se que as duas distribuições empíricas *bootstrap* têm um grau de assimetria semelhante, tendo apenas diferentes localizações. ■

Conclusões

Com este estudo analisou-se a estimação intervalar do valor médio de uma população F próxima do modelo central Normal, usando procedimentos de reamostragem e o clássico intervalo t. A consideração de várias distribuições para o modelo gerador dos dados permitiu avaliar quer a robustez dos vários procedimentos quer a sua eficiência. A principal conclusão retirada dos resultados obtidos é a seguinte:

- Tendo em conta a "qualidade" da probabilidade de cobertura e dos comprimentos dos intervalos, tudo indica que a melhor combinação que se pode fazer é a que corresponde ao método *bootstrap* IF com a utilização do estimador não robusto, $\bar{X}_n$, na fase de reamostragem.

Um estudo desta natureza é sempre necessariamente limitado. Embora não tivesse sido possível fazê-lo (por uma questão de tempo) considera-se que teria sido desejável:

- Aprofundar o estudo para outros valores de dimensão amostral, assim como variar a percentagem de contaminação das distribuições NSC–ε e NAC–ε;
- Estudar o comportamento do *bootstrap Winsorized* face a várias escolhas para o valor de β, dado que este método não teve o desempenho que se esperava quando se usou o estimador robusto.

4.3.2 Intervalos de confiança robustos para o coeficiente de correlação linear

Como segundo exemplo de aplicação do procedimento proposto, *bootstrap* IF, vai considerar-se a obtenção de intervalos de confiança para o coeficiente de correlação linear, ρ. A escolha deste exemplo deve-se por um lado ao interesse que este estimador tem suscitado na literatura relativa ao *bootstrap* (Efron, 1982; Dolker *et al.*, 1982; Efron, 1987; Silverman e Young, 1987; Young, 1988) e por outro lado à conhecida grande sensibilidade a observações atípicas do estimador usual, o coeficiente de correlação amostral, R.

Seja (X_1, X_2) um vector aleatório proveniente de uma população bidimensional F com valor médio $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$, onde $\sigma_{ii} = Var_F(X_i)$ e $\sigma_{ij} = Cov(X_i, X_j)$, para $i \neq j$ e $i, j = 1, 2$. O coeficiente de correlação linear entre as variáveis X_1 e X_2 é definido por

$$\rho = corr(X_1, X_2) = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}. \tag{4.14}$$

Este parâmetro é importante na caracterização da intensidade da ligação entre as duas variáveis e, ao contrário da covariância, não depende das unidades em que elas se exprimem. Considere-se como modelo central uma distribuição Normal bivariada de parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, isto é $F_{\boldsymbol{\theta}} \equiv N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, de acordo com a notação utilizada neste capítulo.

O estimador clássico de ρ é o denominado coeficiente de correlação linear amostral de Pearson, $T_n^{nr} = R$, definido por

$$R = \frac{\sum_{i=1}^{n} \left(X_{i1} - \bar{X}_1\right)\left(X_{i2} - \bar{X}_2\right)}{\sqrt{\sum_{i=1}^{n} \left(X_{i1} - \bar{X}_1\right)^2 \sum_{i=1}^{n} \left(X_{i2} - \bar{X}_2\right)^2}},$$

onde $(\bar{X}_1, \bar{X}_2)$ denota o vector média da amostra aleatória $((X_{11}, X_{12}), (X_{21}, X_{22}) \cdots, (X_{n1}, X_{n2}))$.

Considere-se também o estimador robusto de ρ, T_n^r =RMCD, que é obtido com base no estimador da matriz de Covariâncias de Determinante Mínimo na sua versão reponderada (veja-se o Capítulo 2), isto é, se $\widehat{\boldsymbol{\Sigma}}_{RMCD}$ denotar o estimador da matriz de covariâncias RMCD então um estimador robusto para ρ é obtido substituindo em (4.14) os valores de σ_{ij} pelos elementos correspondentes de $\widehat{\boldsymbol{\Sigma}}_{RMCD}$.

Denote-se por T^{nr} o funcional equivalente ao estimador R. A função de influência de T^{nr} no ponto (x_1, x_2) é dada para qualquer função de distribuição bidimensional, F, com momentos de segunda ordem finitos, por

$$IF(x_1, x_2; T^{nr}, F) = -0.5\rho({y_1}^2 + {y_2}^2) + y_1 y_2, \quad (|\rho| \neq 1)$$

onde $y_i = (x_i - \mu_i)/\sqrt{\sigma_{ii}}$, $i = 1, 2$, (Devlin *et al.*, 1975). A variância assintótica de R pode ser determinada por $Var\left[IF(x_1, x_2; T^{nr}, F)\right]$. Sob uma distribuição Normal bivariada e aplicando o Lema 3.6 apresentado em Pires (1990, pág. 85) obtém-se, após alguns cálculos, $VA(R, F) = (1-\rho^2)^2$.

A função de influência padronizada é então dada por

$$SIF(x_1, x_2; T^{nr}, F) = \frac{|IF\,(x_1, x_2; T^{nr}, F)\,|}{1-\rho^2}, \quad |\rho| \neq 1. \tag{4.15}$$

Convém notar que ao assumir a distribuição Normal para o cálculo da variância assintótica isso não significa que se está a assumir que os dados são normalmente distribuídos. Isso é apenas uma consequência da optimalidade do estimador R sob o modelo Normal, pelo que as observações influentes deverão ser referentes a esta distribuição. Como se pode observar a função (4.15) depende do seguinte vector de parâmetros, $\mathbf{\Omega} = (\boldsymbol{\mu}, \mathbf{\Sigma})$. Considere-se o seguinte estimador robusto de $\mathbf{\Omega}$,

$$\hat{\mathbf{\Omega}}^r = \left(\hat{\boldsymbol{\mu}}^r, \hat{\mathbf{\Sigma}}^r\right), \tag{4.16}$$

onde $\hat{\boldsymbol{\mu}}^r = (\hat{\mu_1}, \hat{\mu_2})$ e $\hat{\mathbf{\Sigma}}^r = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} \end{bmatrix}$ são os estimadores robustos RMCD (com ponto de rotura de 0.5) do centro e escala da distribuição bivariada F, e $\hat{\rho} = \hat{\sigma}_{12}/\sqrt{\hat{\sigma}_{11}\hat{\sigma}_{22}}$.

Então a função de influência empírica padronizada, $RESIF$, é obtida substituindo na expressão (4.15) o vector de parâmetros $\mathbf{\Omega}$ pelo definido em (4.16), isto é,

$$RESIF(x_1, x_2; T^{nr}, F) = \frac{1}{1-\hat{\rho}^2}\left|-0.5\hat{\rho}\left\{\frac{(x_1-\hat{\mu}_1)^2}{\hat{\sigma}_{11}} + \frac{(x_2-\hat{\mu}_2)^2}{\hat{\sigma}_{22}}\right\} + \frac{(x_1-\hat{\mu}_1)^2}{\hat{\sigma}_{11}} \times \frac{(x_2-\hat{\mu}_2)^2}{\hat{\sigma}_{22}}\right|. \tag{4.17}$$

A seguinte proposição apresenta duas propriedades importantes de SIF definida em (4.15).

Proposição 4.2 *Seja* $\mathbf{X} = (X_1, X_2)^T$ *um vector aleatório com função de distribuição* F*, com* $E(\mathbf{X}) = \boldsymbol{\mu}$*,* $Var(\mathbf{X}) = \mathbf{\Sigma}$ *e* $\rho = corr(X_1, X_2)$*, e seja* T^{nr} *o funcional equivalente a* R*. Então*

(a) A distribuição de $SIF(\mathbf{X}; T^{nr}, F)$ *não depende de* ρ *(desde que* $|\rho| \neq 1$*).*

(b) $SIF(\mathbf{x}; T^{nr}, F)$ *é invariante para transformações afins de* $\mathbf{X}$*.*

Demonstração:

Seja $\mathbf{Y} = (Y_1, Y_2)^T$ o vector correspondente à forma padronizada de $\mathbf{X}$. Então $E[\mathbf{Y}] = \mathbf{0}$ e $Var(\mathbf{Y}) = \mathbf{\Lambda}$, com

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Pode então escrever-se $\mathbf{Y} = \mathbf{\Lambda}^{1/2}\mathbf{Z}$, onde $\mathbf{Z} = (Z_1, Z_2)^T$ é tal que $E[\mathbf{Z}] = \mathbf{0}$ e $Var(\mathbf{Z}) = \mathbf{I}_2$, e

$$\mathbf{\Lambda}^{1/2} = \begin{bmatrix} \sqrt{1+\rho}+\sqrt{1-\rho} & \sqrt{1+\rho}-\sqrt{1-\rho} \\ \sqrt{1+\rho}-\sqrt{1-\rho} & \sqrt{1+\rho}+\sqrt{1-\rho} \end{bmatrix} \times \frac{1}{2}.$$

É fácil verificar que $SIF(\mathbf{X}; T^{nr}, F) = |\mathbf{Y}^T\mathbf{A}\mathbf{Y}|$, com

$$\mathbf{A} = \begin{bmatrix} -\rho & 1 \\ 1 & -\rho \end{bmatrix} \times \frac{0.5}{(1-\rho^2)}.$$

Assim $SIF(\mathbf{X}; T^{nr}, F) = |\mathbf{Z}^T \mathbf{\Lambda}^{1/2} \mathbf{A} \mathbf{\Lambda}^{1/2} \mathbf{Z}|$ mas

$$\mathbf{\Lambda}^{1/2} \mathbf{A} \mathbf{\Lambda}^{1/2} = \begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix}.$$

Portanto $SIF(\mathbf{X}; T^{nr}, F) = |Z_1 Z_2|$, o que completa a prova da primeira parte da proposição pois a distribuição de $\mathbf{Z}$ não depende de ρ.

Seja $F_{\mathbf{C},\mathbf{b}}$ a distribuição da variável aleatória $\mathbf{X}_{\mathbf{C},\mathbf{b}} = \mathbf{C}\mathbf{X} + \mathbf{b}$, onde $\mathbf{b} \in I\!R^2$ e $\mathbf{C}$ é uma matriz (2×2) não singular. Pretende-se mostrar que

$$SIF(\mathbf{x}; T^{nr}, F) = SIF(\mathbf{x}_{\mathbf{C},\mathbf{b}}; T^{nr}, F_{\mathbf{C},\mathbf{b}}),$$

para qualquer $\mathbf{x} \in I\!R^2$ e $\mathbf{x}_{\mathbf{C},\mathbf{b}} = \mathbf{C}\mathbf{x} + \mathbf{b}$.

Tendo em conta a prova de (a) é necessário apenas mostrar que $\mathbf{Z} = \mathbf{Z}_{\mathbf{C},\mathbf{b}}$ com $\mathbf{Z} = \mathbf{\Sigma}^{1/2}(\mathbf{x} - \mu)$ e $\mathbf{Z}_{\mathbf{C},\mathbf{b}} = \mathbf{\Sigma}^{1/2}_{\mathbf{C},\mathbf{b}}(\mathbf{x}_{\mathbf{C},\mathbf{b}} - \boldsymbol{\mu}_{\mathbf{C},\mathbf{b}})$ onde $\mathbf{\Sigma}_{\mathbf{C},\mathbf{b}} = \mathbf{C}\mathbf{\Sigma}\mathbf{C}^T$ e $\boldsymbol{\mu}_{\mathbf{C},\mathbf{b}} = \mathbf{C}\boldsymbol{\mu} + \mathbf{b}$. Escolhendo $\mathbf{\Sigma}^{1/2}_{\mathbf{C},\mathbf{b}}$ como $\mathbf{C}\mathbf{\Sigma}^{1/2}$ facilmente se obtém o pretendido. ■

Tal como na secção anterior também aqui se vai proceder a um estudo de simulação para avaliar o desempenho do método *bootstrap* IF. Com esse objectivo determinaram-se intervalos com $(1 - \alpha) \times 100\%$ de confiança para ρ baseados em amostras provenientes de populações bidimensionais. A obtenção desses intervalos será efectuada com base nos métodos que se passam a descrever:

- **Método clássico ($Fisher$)**: dado que a distribuição amostral exacta de R é bastante complicada, o procedimento usual para determinar intervalos de confiança para ρ baseia-se na distribuição da denominada transformação z de Fisher, isto é,

  $$Z = \frac{1}{2} \ln\left(\frac{1+R}{1-R}\right) \equiv \tanh^{-1}(R) \overset{a}{\sim} N\left(\tanh^{-1}(\rho), \frac{1}{n-3}\right).$$

  Então, os limites de confiança aproximados, (A, B), para $\tanh^{-1}(\rho)$ com $(1-\alpha) \times 100\%$ de confiança são, $A = Z - \Phi^{-1}(1-\alpha/2)/\sqrt{n-3}$ e $B = Z + \Phi^{-1}(1-\alpha/2)/\sqrt{n-3}$. Para ρ esses limites são dados por,

  $$\left(\frac{\exp(2A) - 1}{1 + \exp(2A)}, \frac{\exp(2B) - 1}{1 + \exp(2B)}\right);$$

- ***Bootstrap* clássico ou uniforme (UB)**: intervalo obtido pelo método do percentil usando o *bootstrap* clássico (expressão 3.31);

- ***Bootstrap* baseado na função de influência (IFB)**: intervalo obtido pelo método do percentil mas aplicando o *bootstrap* IF, com $RESIF$ definida pela expressão (4.17). As probabilidades de reamostragem são determinadas com base nos pesos

  $$w_i = I_{[0,c]}\left(|RESIF_i|\right) + \Psi\left(|RESIF_i|; c, c, \gamma\right) \times I_{]c,+\infty[}\left(|RESIF_i|\right), \tag{4.18}$$

  para $i = 1, 2, \ldots, n$, com Ψ definida em (4.8).

Numa primeira fase desta análise consideraram-se três possíveis valores para γ, ($\gamma = 2, 10, \infty$) e três valores possíveis para a constante de afinação, ($c = 5, 7, 10$). Verificou-se, através de integração numérica, que estes três valores para c correspondem respectivamente aos percentis de ordem 0.9977, 0.99977 e 0.99999 da distribuição de SIF sob o modelo Normal bivariado (o qual, de acordo com a Proposição 4.2, é simplesmente o valor absoluto do produto de duas variáveis aleatórias Normais padrão independentes). Os resultados desta análise preliminar encontram-se na Tabela 4.4 e serão discutidos mais adiante. Nesta aplicação também se fixou $d = c$ por razões análogas às descritas no estudo relativo à localização univariada.

Para uma mesma amostra, as probabilidades de reamostragem são iguais, independentemente do estimador que se pretende reamostrar, seja o estimador não robusto R, seja o estimador robusto RMCD. Para o estimador robusto RMCD essas probabilidades são determinadas usando $RESIF(\mathbf{x}; R, \hat{\mathbf{\Omega}}_2^r)$ e não $RESIF(\mathbf{x}; RMCD, \hat{\mathbf{\Omega}}_1^r)$, de acordo com o que foi referido na Observação 4.1;

Note-se ainda que pela Proposição 4.2 tanto a $RESIF$, como os pesos, w_i, são invariantes para transformações afins se o estimador robusto usado for equivariante para o mesmo tipo de transformações.

- ***Bootstrap Winsorized*** (WB): intervalo obtido usando o método do percentil mas adaptado ao método proposto por Singh (1998). Em primeiro lugar é preciso decidir como proceder à *Winsorization* de uma amostra multivariada, a começar pela sua ordenação, pois ao contrário do caso univariado, esta transformação não é única. Para a ordenação recorre-se ao conceito, designado na literatura anglo-saxónica, por *data depth* e que se pode traduzir por profundidade de um ponto. Este conceito fornece um modo de ordenação dos pontos (para além de uma nova definição de parâmetros de localização e escala da distribuição subjacente). Aqueles pontos que estão mais perto do "centro" da distribuição são designados por profundos, enquanto que os mais afastados são designados por superficiais e pertencem à cauda da distribuição. Liu e Singh (1993) apresentam quatro noções de profundidade, Mahalanobis, Tukey, *simplicial* e *majority*, descrevendo as suas versões populacionais e as respectivas formas amostrais.

  Nesta análise optou-se por usar a noção mais simples das quatro mencionadas mas numa versão robusta, a qual se designa por profundidade de Mahalanobis robusta. Seja $D(\mathbf{x}, \boldsymbol{\mu}_F)$ a distância de Mahalanobis entre $\mathbf{x}$ e $\boldsymbol{\mu}_F$ (o valor médio de F), isto é $D^2(\mathbf{x}, \boldsymbol{\mu}_F) = (\mathbf{x} - \boldsymbol{\mu}_F)^T \mathbf{\Sigma}_F^{-1} (\mathbf{x} - \boldsymbol{\mu}_F)$. Chama-se profundidade de Mahalanobis a $M_h D = \left[1 + D^2(\mathbf{x}, \boldsymbol{\mu}_F)\right]^{-1}$, sendo a sua versão amostral robusta dada por

$$m_h d = \left[1 + d^2(\mathbf{x}, \hat{\boldsymbol{\mu}})\right]^{-1}, \tag{4.19}$$

  onde $d^2(\mathbf{x}, \hat{\boldsymbol{\mu}}) = (\mathbf{x} - \hat{\boldsymbol{\mu}})^T \hat{\mathbf{\Sigma}}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}})$ e $\hat{\boldsymbol{\mu}}$ e $\hat{\mathbf{\Sigma}}$ são estimadores robustos, respectivamente, do vector valor médio, $\boldsymbol{\mu}_F$, e da matriz de covariâncias $\mathbf{\Sigma}_F$. Note-se que a ordenação com base nesta medida de profundidade é equivalente à ordenação com base na distância de Mahalanobis robusta.

O método que se propõe para a *Winsorization* da amostra $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$, com uma proporção β de *Winsorization*, é o seguinte:

(i) Determinam-se estimativas robustas do vector valor médio e da matriz de covariâncias, $\hat{\boldsymbol{\mu}}$ e $\hat{\boldsymbol{\Sigma}}$, respectivamente. Nas aplicações descritas neste capítulo optou-se pelos estimadores RMCD de ponto de rotura 0.50;

(ii) Calcula-se a distância de Mahalanobis robusta para cada observação, com base nas estimativas obtidas no passo (i), isto é,

$$d_i^2 = (\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}),$$

e ordenam-se estes por ordem crescente, $d_{1:n}^2, d_{2:n}^2, \cdots, d_{n:n}^2$;

(iii) Seja $k = [n\beta]$. Os k pontos correspondentes a $d_{n-k+1:n}^2, d_{n-k:n}^2, \cdots, d_{n:n}^2$ (ou, o que é equivalente, aos k menores valores de $m_h d$) devem ser deslocados até que a sua distância de Mahalanobis robusta seja igual a $d_{n-k:n}^2$, mas preservando a sua orientação. Para tal calcula-se

$$l_i = \sqrt{\frac{d_{n-k:n}^2}{d_i^2}}, \quad i = 1, 2, \cdots, n,$$

e finalmente obtém-se a amostra *Winsorized*, $(\mathbf{x}_1^W, \mathbf{x}_2^W, \cdots, \mathbf{x}_n^W)$:

$$\mathbf{x}_i^W = \begin{cases} \mathbf{x}_i, & \text{se} \quad l_i \geq 1 \\ \hat{\boldsymbol{\mu}} + l_i (\mathbf{x}_i - \hat{\boldsymbol{\mu}}), & \text{se} \quad l_i < 1 \end{cases}, \quad i = 1, 2, \cdots, n.$$

A Figura 4.2 ilustra a *Winsorization* de uma amostra bivariada.

Como na secção anterior, também se optou por fixar β em 0.1, já que este valor corresponde à máxima contaminação determinística que será testada. Este valor será o mesmo quer para o estimador R quer para o estimador RMCD;

- ***Bootstrap* de rejeição (RB)**: limites de confiança obtidos pelo método do percentil, mas usando um método de rejeição de observações discordantes. A regra utilizada é a seguinte:

  1. Determina-se a função $RESIF(\mathbf{x}; R, \hat{\boldsymbol{\Omega}}^r)$ como definida em (4.17);
  2. Fixa-se uma constante $c > 0$ e rejeita-se qualquer observação, $\mathbf{x}_j$, tal que

     $$RESIF(\mathbf{x}_j; R, \hat{\boldsymbol{\Omega}}^r) > c.$$

     O valor c deve ser escolhido tendo em consideração o já referido para o método *bootstrap* IF (ver Tabela 4.4, mais adiante, e respectiva discussão de resultados);
  3. Aplica-se o *bootstrap* usual às observações restantes.

O método assim construído pretende ilustrar um método de detecção e rejeição de observações atípicas. Note-se que nesta situação poder-se-ia ter optado por constituir reamostras *bootstrap* de dimensão igual à dimensão original, n. No entanto, optou-se por reamostrar n' ($n' \leq n$), observações, das n' que restaram após o processo de rejeição, por se entender que desta forma se considera a variabilidade introduzida pelo passo de rejeição.

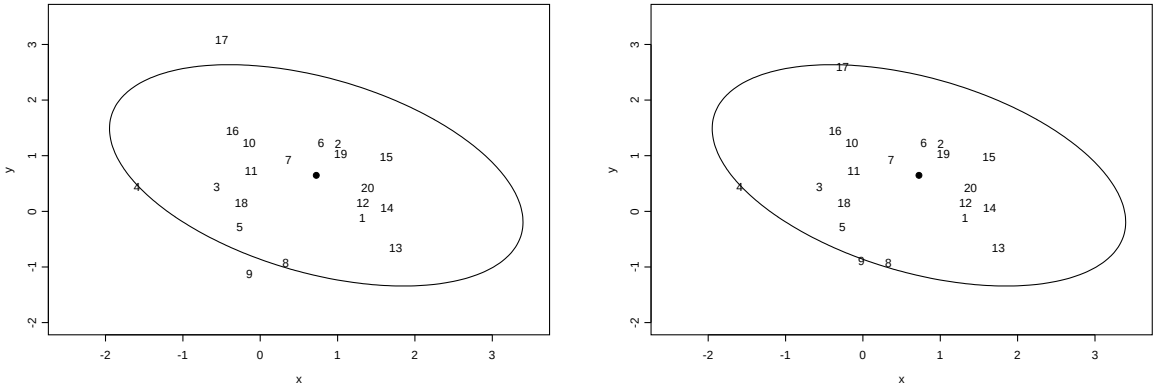


Figura 4.2: Exemplo de *Winsorization* de uma amostra bivariada com $n = 20$ e $\beta = 0.1$. As observações números 17 e 9 que têm associadas maior distância de Mahalanobis são deslocadas para a elipse que passa pela observação número 4, a qual corresponde a $d^2_{18:20}$, nessa deslocação a direcção original das observações é preservada. A amostra *Winsorized* é constituída então pelos pontos da figura à direita.

Tal como na secção anterior, o objectivo desta análise é comparar o desempenho dos vários métodos quer em termos de robustez, quer em termos de eficiência. Para isso realizou-se um estudo de simulação de Monte Carlo no qual se considerou para $(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n)$as seguintes distribuições:

- Normal bivariada, NOR: $\mathbf{X}_i \sim N_2(\mathbf{0}, \mathbf{\Lambda}), \quad i = 1, 2, \cdots, n;$
- Normal Simetricamente ε-Contaminada, NSC–ε:

$$\mathbf{X}_i \sim (1-\varepsilon) N_2(\mathbf{0}, \mathbf{\Lambda}) + \varepsilon N_2(\mathbf{0}, f_e \mathbf{\Lambda}), \quad i = 1, 2, \cdots, n,$$

  com $f_e = 9$ e $\varepsilon = 0.25$;
- Normal Assimetricamente ε-Contaminada, NAC–ε:

$$\mathbf{X}_i \sim N_2(\mathbf{0}, \mathbf{\Lambda}) \quad \text{para} \quad i = 1, 2, \cdots, n - [\varepsilon n] \quad \text{e} \quad \mathbf{X}_i \sim \Delta_{\mathbf{x}} \quad \text{para} \quad i = n - [\varepsilon n] + 1, \cdots, n,$$

  com $\varepsilon = 0.1$, $\mathbf{x} = (-5, 5)$ e n é a dimensão da amostra a gerar;

A distribuição de controlo é a distribuição Normal bivariada. Com as distribuições NSC e NAC ilustram-se duas situações de contaminação, a primeira de tipo simétrico e a segunda de tipo assimétrico. Em todas as distribuições a matriz de covariâncias $\mathbf{\Lambda}$ é dada por

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \text{com} \quad \rho = 0.5.$$

Usaram-se amostras de dimensão $n = 100$. Uma vez que as versões ponderadas do estimador MCD são mais eficientes e mantêm o elevado ponto de rotura foram usadas duas versões do estimador robusto RMCD, denotadas por RMCD50 e RMCD25, a primeira com ponto de rotura aproximado de

0.5 e a segundo com ponto de rotura aproximado de 0.25. Tendo em conta os tempos computacionais dos procedimentos de reamostragem quando se utilizam os estimadores robustos, foi necessário considerar duas situações para o número de simulações, $nsimul$. Assim, fixou-se $nsimul = 1000$ para o estudo com o estimador clássico R e $nsimul = 100$ quando se usam os estimadores robustos RMCD50 e RMCD25. Em ambos os casos (estimador robusto e não robusto), e para os métodos de reamostragem, foram usadas 1000 réplicas *bootstrap*.

Para cada distribuição (NOR, NSC–0.25, NAC–0.1) geraram-se as $nsimul$ amostras de dimensão $n = 100$ e construíram-se intervalos de confiança usando os cinco procedimentos descritos ($Fisher$, UB, IFB, WB, RB), fixando para nível de confiança $1-\alpha = 0.90$. Em todos os métodos *bootstrap* foi usada a versão não paramétrica apesar de se ter fixado a distribuição dos dados.

Os indicadores a observar são os seguintes:

- a probabilidade estimada de não cobertura, isto é, o número de vezes que $\rho = 0.5$ não pertence ao intervalo construído dividido por $nsimul$;
- a média e o desvio padrão dos comprimentos dos $nsimul$ intervalos construídos.

A obtenção dos indicadores acima descritos foi realizada segundo o processo de independência.

Estudo preliminar relativo às constantes no método IFB

Antes de se proceder à comparação dos vários métodos apresentados, e como mencionado atrás, estudou-se o comportamento do método *bootstrap* IF face a três valores para a constante de afinação, $c = 5, 7, 20$ e três valores de γ $(= 2, 10, \infty)$. A Figura 4.3 representa graficamente as várias funções ψ para as nove combinações possíveis das constantes.

Considerando as condições de delineamento anteriormente descritas, apresentam-se na Tabela 4.4 os indicadores acima referidos, para as nove combinações das três constantes de afinação e dos três valores de γ, obtidos para o método IFB quando se considera o estimador R. A partir destes resultados verifica-se que, como esperado, a escolha do valor de γ não é tão importante como a escolha da constante de afinação c. Todavia, devido ao comprimento médio dos intervalos construídos com $\gamma = 2$ (na distribuição NAC) e a uma ligeira melhoria na probabilidade estimada de cobertura (sob a distribuição NOR) recomenda-se o uso de $\gamma = 10$. A relação entre eficiência e robustez relacionada com a constante de afinação, discutida anteriormente, não é fácil de observar nestes resultados, pois é necessário comparar, ou o comprimento médio do intervalo para uma dada probabilidade de não cobertura, ou a probabilidade de não cobertura para um dado comprimento médio. No entanto, pode ser observado o seguinte: para a distribuição NOR existe uma tendência para atingir melhores probabilidades de cobertura para valores elevados de c (ignorando os comprimentos dos intervalos); e para a NAC pode ver-se que existe um aumento dos comprimentos médios dos intervalos à medida que c aumenta (ignorando as diferenças nas probabilidades de não cobertura). Contudo, verifica-se

Tabela 4.4: Resultados da simulação para diferentes constantes de afinação para definição dos pesos do *bootstrap* IF quando se usa o estimador R.

Distribuição amostral	Função de peso, ψ	(alvo: 0.100) não cobertura			Comprimento do intervalo Média			Desvio padrão		
	γ	c=5	c=7	c=10	c=5	c=7	c=10	c=5	c=7	c=10
NOR	∞	0.107	0.119	0.120	0.2433	0.2429	0.2440	0.033	0.033	0.034
	10	0.133	0.130	0.110	0.2427	0.2423	0.2430	0.034	0.034	0.033
	2	0.124	0.109	0.113	0.2434	0.2454	0.2426	0.032	0.034	0.034
NSC–0.25	∞	0.222	0.172	0.130	0.2675	0.2893	0.3147	0.049	0.053	0.058
	10	0.214	0.147	0.125	0.2700	0.2957	0.3174	0.048	0.053	0.058
	2	0.194	0.164	0.113	0.2761	0.2985	0.3187	0.047	0.054	0.060
NAC–0.1	∞	0.130	0.134	0.130	0.2454	0.2427	0.2501	0.035	0.035	0.047
	10	0.128	0.140	0.116	0.2425	0.2443	0.2889	0.034	0.039	0.084
	2	0.120	0.088	0.089	0.2747	0.3653	0.4925	0.053	0.079	0.087

que estas diferenças não são muito significativas, o que deve estar relacionado com o facto das três constantes de afinação corresponderem a percentis elevados.

Tendo em conta os aspectos discutidos, uma combinação razoável corresponde à dada por $c = 10$ e $\gamma = 10$. Este par de valores será usado no restante estudo de simulação onde se compara o desempenho do método *bootstrap* IF com os outros procedimentos ($Fisher$, UB, WB e RB). Observe-se que também se escolheu $c = 10$ para controlar a rejeição no método RB.

Principais resultados do estudo de Monte Carlo

Na Tabela 4.5 apresentam-se as estimativas das probabilidades de não cobertura e medidas descritivas do comprimentos dos intervalos (média e desvio padrão) para cada método e distribuição, quando se usa o estimador R.

Uma constatação que se pode fazer, quando se observam os resultados da Tabela 4.5, é a de que o método de Fisher, usado como controlo, é óptimo quando não existe contaminação mas degrada-se face a distribuições contaminadas. Verifica-se também que os métodos UB e WB (apesar deste último ter sido proposto para estimadores robustos) têm um desempenho razoável sob a distribuição de contaminação simétrica NSC. Este facto pode ser devido ao tipo de distribuição, pois a Normal simetricamente contaminada preserva o coeficiente de correlação e não dá tratamento preferencial a nenhuma direcção, no espaço das variáveis, para localizar as observações discordantes simuladas. Todavia, para a distribuição NAC ambos os métodos, UB e WB, falham. Quanto ao método RB verifica-se que piora para a distribuição NSC tendo, no entanto, um comportamento razoável para a distribuição NAC e NOR. Para o método IFB destaca-se a relativa estabilidade em termos de probabilidade de não cobertura para as três distribuições consideradas.

Na Tabela 4.6 apresentam-se resultados análogos para os estimadores MCD reponderados. Verifica-

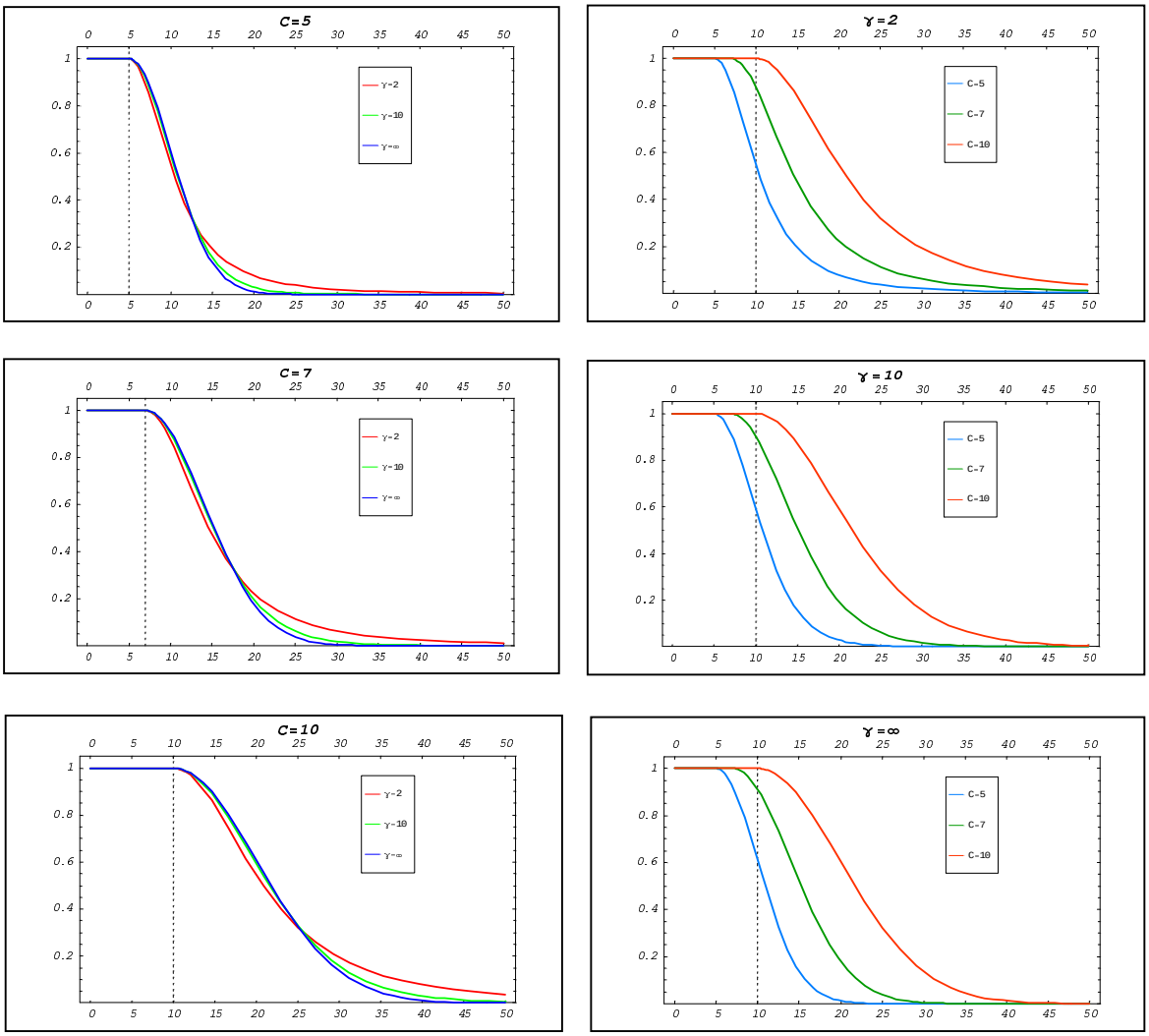


Figura 4.3: Comparação das funções peso associada à funções ψ para os três valores de c e de γ.

-se que para o estimador RMCD50 todos os métodos têm um desempenho similar: as probabilidades estimadas de não cobertura são sempre menores que a probabilidade fixada ($\alpha = 0.10$) e em todos os casos os comprimentos médios dos intervalos são muito maiores que os correspondentes na Tabela 4.5, mesmo para o UB. Este aspecto não é surpreendente e deve-se à pior eficiência do estimador RMCD50. Quando se observam as colunas relativas ao RMCD25 verifica-se um aumento de eficiência para todos os métodos (os comprimentos médios dos intervalos são menores). Por outro lado, a abordagem segundo o método *bootstrap* IF, IFB, apresenta um comportamento razoável ao longo de todas as distribuições, sendo melhor que o *Winsorized bootstrap*, WB, mas semelhante ao *bootstrap* usual, UB. Relembre-se que, devido ao número de simulações com que foram obtidos os resultados da Tabela 4.6, $nsimul = 100$, a incerteza relativa às estimativas das probabilidades de não cobertura é maior do que para as apresentadas na Tabela 4.5. Por exemplo, para o caso em que

Tabela 4.5: Resultados do estudo de Monte Carlo para o coeficiente de correlação amostral de Pearson, R.

Distribuição amostral	Método	(alvo: 0.100) Não Cobertura	Comprimento do intervalo Média	Comprimento do intervalo Desvio padrão
NOR	$Fisher$	0.104	0.2476	0.024
	UB	0.106	0.2433	0.035
	IFB(c=10,$\gamma = 10$)	0.110	0.2430	0.033
	WB	0.166	0.2245	0.028
	RB(c=10)	0.122	0.2422	0.033
NSC–0.25	Fisher	0.267	0.2456	0.036
	UB	0.124	0.3498	0.079
	IFB(c=10,$\gamma = 10$)	0.125	0.3174	0.058
	WB	0.165	0.2706	0.045
	RB(c=10)	0.256	0.2547	0.049
NAC–0.1	$Fisher$	1.000	0.2218	0.019
	UB	1.000	0.3669	0.034
	IFB(c=10,$\gamma = 10$)	0.116	0.2889	0.084
	WB	0.926	0.3592	0.041
	RB(c=10)	0.125	0.2563	0.037

Tabela 4.6: Resultados do estudo de Monte Carlo para os estimadores MCD.

Distribuição amostral	Método	(alvo: 0.10) Não cobertura RMCD50	(alvo: 0.10) Não cobertura RMCD25	Comprimento do intervalo RMCD50 Média	Comprimento do intervalo RMCD50 Desvio padrão	Comprimento do intervalo RMCD25 Média	Comprimento do intervalo RMCD25 Desvio padrão
NOR	UB	0.02	0.07	0.5669	0.127	0.3759	0.075
	IFB	0.03	0.07	0.5869	0.139	0.3886	0.071
	WB	0.01	0.05	0.5670	0.133	0.3783	0.084
NSC–0.25	UB	0.06	0.09	0.5304	0.124	0.3977	0.085
	IFB	0.05	0.10	0.5409	0.136	0.3906	0.080
	WB	0.02	0.03	0.5844	0.141	0.4314	0.096
NAC–0.1	UB	0.03	0.08	0.4739	0.096	0.3346	0.076
	IFB	0.03	0.15	0.6171	0.159	0.3891	0.093
	WB	0.02	0.07	0.9287	0.311	0.6524	0.223

a estimativa da probabilidade de não cobertura é de 0.15, situação que corresponde à combinação RMCD25, NAC e IFB, um intervalo aproximado a 95% de confiança para essa probabilidade é $(0.08; 0.22)$.

Como conclusão e resumo deste estudo, consideram-se simultaneamente as duas tabelas e apresenta-

-se, para cada distribuição, a ordenação por comprimento médio dos intervalos para os métodos com cobertura aceitável:

- Distribuição NOR:

$$Fisher(R) \simeq UB(R) \simeq IFB(R) < UB(RMCD25) \simeq IFB(RMCD25) < WB(RMCD25),$$

- Distribuição NSC–0.25:

$$IFB(R) < UB(R) < UB(RMCD25) < IFB(RMCD25),$$

- Distribuição NAC–0.1:

$$RB(R) < IFB(R) < UB(RMCD25) < IFB(RMCD25) < WB(RMCD25),$$

Como ordenação comum às três distribuições consideradas tem-se:

$$IFB(R) < UB(RMCD25) \simeq IFB(RMCD25) < WB(RMCD25),$$

concluindo-se que numa situação em que se desconhece a distribuição a escolha mais razoável recai sobre o método *bootstrap* IF, com a aplicação do estimador clássico, R, nas reamostras.

É importante referir que não ocorreram problemas numéricos quando se estimou a distribuição *bootstrap* dos estimadores RMCD. Este facto pode ser justificado pela dimensão das amostras de trabalho ($n = 100$), assim como pelo alto ponto de rotura (0.5 e 0.25) associado a estes estimadores. O único problema, já esperado, foi o excessivo tempo de cálculo requerido por este procedimento, o qual foi cerca de quarenta vezes superior ao do estimador não robusto.

Globalmente o método *bootstrap* IF associado ao estimador não robusto, R, surge como possuindo o melhor compromisso entre eficiência e robustez. Observe-se, porém, que caso se pretenda usar *bootstrap* com estes estimadores robustos e sob as distribuições usadas, dever-se-á optar pelo *bootstrap* clássico e o estimador RMCD25 já que a este método correspondem menores comprimentos médios dos intervalos. Todavia, na situação de UB com estimador robusto deve salvaguarde-se o facto de noutras condições esse procedimento conduzir aos já referidos problemas numéricos, uma vez que não há qualquer protecção contra o aumento da contaminação nas reamostras *bootstrap*.

Aumento da dimensão da amostra

Em seguida estuda-se o comportamento do método *bootstrap* IF face ao aumento da dimensão amostral. Com este estudo adicional pretende verificar-se se o procedimento IFB não conduz, sob o modelo central, a subcobertura quando aumenta a dimensão da amostra.

Para responder a esta questão delineou-se um estudo de Monte Carlo com as seguintes características: para o método *bootstrap* IF e usando o estimador clássico R geraram-se 1000 amostras de

Tabela 4.7: Desempenho do método IFB para dados não contaminados e com crescente dimensão amostral.

			Comprimento do intervalo			
	(alvo: 0.100)		IFB		*Fisher*	
Dimensão	Não Cobertura			Desvio		Desvio
da amostra	IFB	$Fisher$	Média	Padrão	Média	Padrão
$n = 100$	0.110	0.104	0.2430	0.0328	0.2476	0.0243
$n = 1000$	0.110	0.118	0.0777	0.0042	0.0781	0.0025
$n = 5000$	0.105	0.106	0.0347	0.0011	0.0348	0.0005
$n = 10000$	0.091	0.089	0.0246	0.0008	0.0247	0.0002
$n = 20000$	0.098	0.097	0.0174	0.0005	0.0174	0.0001

acordo com a distribuição NOR, considerada anteriormente. Para cada amostra determinou-se o intervalo a 90% de confiança para o coeficiente de correlação ρ, segundo o método do percentil baseado em 1000 réplicas *bootstrap*. Como controlo usou-se o método de Fisher.

Foi usado o IFB ($c = 10$,$\gamma = 10$) e dimensões amostrais de $n = 100, 1000, 5000, 10000$ e 20000. A Tabela 4.7 mostra os resultados das 1000 simulações para cada dimensão amostral. A sua análise permite verificar que não há qualquer problema de estimação quando se usa o método *bootstrap* IF em dados não contaminados à medida que n aumenta. Em termos de comprimento do intervalo verifica-se que o comprimento médio é semelhante para ambos os métodos (IFB e $Fisher$), sendo aproximadamente proporcional a $1/\sqrt{n}$. O desvio padrão dos comprimentos dos intervalos com IFB, parece revelar algum efeito de segunda ordem, pois é aproximadamente 1.5 a 5 vezes superior ao do método de Fisher, razão esta que aumenta com n.

Estudo empírico do ponto de rotura

No Capítulo 2 foi referido que uma propriedade importante dos estimadores robustos é o ponto de rotura não nulo. Recorde-se que numa amostra finita, o ponto de rotura de um estimador (ou de um procedimento relacionado, tal como um teste ou intervalo de confiança) é a mais pequena proporção de observações que têm que ser substituídas por valores arbitrários de modo a "forçar" o estimador a produzir valores arbitrariamente longe dos valores estimados para os dados originais (Donoho e Huber, 1983).

A determinação geral do ponto de rotura para o método *bootstrap* IF não é trivial, já que depende do estimador usado e também do objectivo da estimação (quantis, viés ou variância). No entanto, para se ter uma ideia do ponto de rotura do método IFB na estimação de intervalos de confiança para o coeficiente de correlação linear procedeu-se à sua avaliação por Monte Carlo. Com esse objectivo realizou-se o seguinte estudo:

- consideraram-se dois métodos *bootstrap*, o clássico, UB, e o *bootstrap* IF, IFB($c = 10$,$\gamma = 10$);

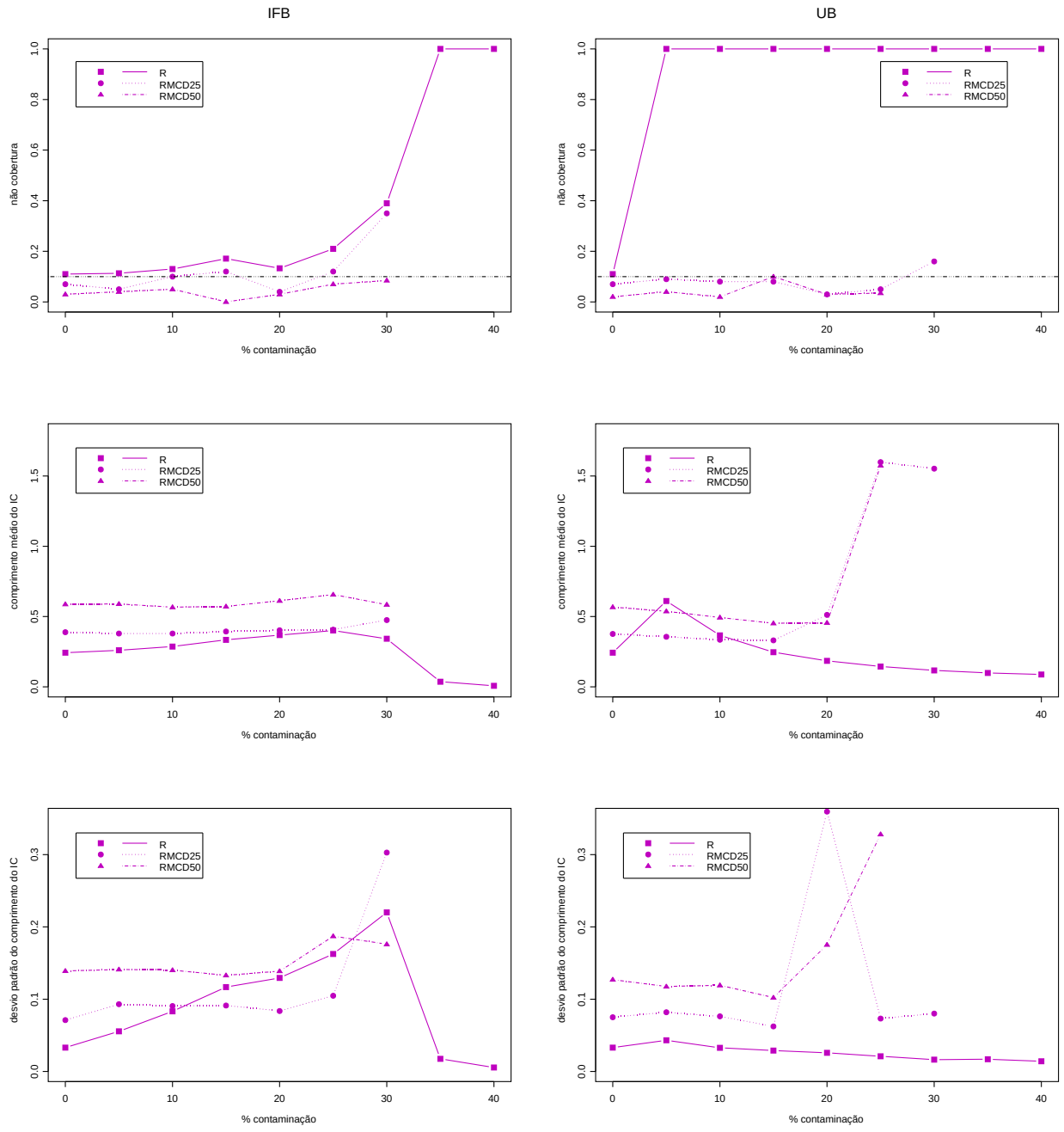


Figura 4.4: Características dos intervalos de confiança para ρ obtidos através dos métodos IFB e UB (não cobertura, média e desvio padrão) *versus* nível de contaminação.

- usaram-se os três estimadores previamente considerados (R, RMCD50, RMCD25);

- fez-se variar a proporção de contaminação, ε, de 0% a 40% com incrementos de 5% e para cada situação geraram-se 1000 amostras (para R) e 100 (para RMCD50 e RMCD25), cada uma com dimensão $n = 100$, a partir da distribuição NAC–ε conforme está definida atrás;

- para cada combinação (amostra gerada com ε fixo, estimador, método) determinaram-se os limites de confiança a 90% pelo método do percentil.

Os indicadores observados foram os descritos anteriormente: probabilidades estimadas de não cobertura, média e desvio padrão dos comprimentos do intervalos.

Os resultados obtidos apresentam-se na Figura 4.4. Note-se que não foi possível obter todos os resultados para os estimadores robustos, RMCD25, RMCD50, sob valores altos de contaminação devido a problemas de singularidade das matrizes de covariâncias.

Analisando os gráficos apresentados verifica-se que o ponto de rotura empírico para o método IFB e R se encontra entre 0.30 e 0.35, enquanto que para o método UB e RMCD50 (ou RMCD25) se encontra entre 0.20 e 0.25 (na coluna direita, onde se apresentam os gráficos relativos ao método UB, observa-se que o comprimento médio do intervalo para 25% de contaminação é praticamente 2, tornando o intervalo inútil). A partir destes resultados é mais uma vez evidente que o melhor método é o *bootstrap* IF e R (até 30% de contaminação).

Seguidamente apresenta-se uma análise de um conjunto de dados, bastante citado na literatura sobre *bootstrap* e a estimação do coeficiente de correlação e conhecidos como "dados de Faculdades de Direito" (ver Efron, 1982, pág. 10, Efron, 1992, Shao e Tu, 1995).

Exemplo 4.3 Dados de Faculdades de Direito. Considere-se a amostra apresentada na Tabela 4.8, relativa a 15 Faculdades de Direito Americanas e às duas variáveis seguintes: GPA (*Grade-Point Average*) e LSAT (*Law School Admission Test*).

Uma vez que é razoável supor a existência de uma relação aproximadamente proporcional entre estas duas variáveis, pretende-se saber qual é a intensidade dessa relação estimando o coeficiente de correlação ρ(LSAT, GPA).

Após alguns diagnósticos, Efron (1992) sugere que pelo menos um ponto é atípico. No gráfico da Figura 4.5, representam-se os resultados obtidos após a aplicação da técnica de diagnóstico *jackknife-after-bootstrap*, proposta por Efron (1992), para o *bootstrap* usual e o estimador R. Verifica-se de facto que o primeiro ponto da amostra tem um grande impacto na estimação de R (denote-se esse ponto por A). Na Figura 4.6 consideram-se duas técnicas de diagnóstico baseadas em regras robustas usando os estimadores RMCD50 e RMCD25. Mais uma vez o ponto A surge como discordante.

Aplicaram-se então os métodos IFB ($c = 10$,$\gamma = 10$) e UB com os três estimadores usados no estudo de simulação anterior, R, RMCD50 e RMCD25 para calcular o intervalo de confiança para o coeficiente de correlação entre as variáveis LSAT e GPA. Foram ainda calculados os intervalos baseados em: (i) Fisher's $tanh^{-1}(R)$ sem o ponto A e (ii) método UB e R sem o ponto A.

A Tabela 4.9 mostra nove intervalos a 90% de confiança para ρ(LSAT,GPA). Todos os limites *bootstrap* foram obtidos com base em $B = 1000$ réplicas *bootstrap* (não paramétrico e método do percentil). Não foi possível aplicar o método *bootstrap* com o estimador RMCD50, quer para UB ou IFB, porque para um grande número de reamostras a matriz de covariâncias estimada era

Tabela 4.8: Duas medidas, LSAT e GPA, do desempenho académico dos alunos universitários em 15 Faculdades de Direito Americanas, no ano de 1973.

i	(LSAT,GPA)	i	(LSAT,GPA)	i	(LSAT,GPA)
1	(576,3.39)	6	(580,3.07)	11	(653,3.12)
2	(635,3.30)	7	(555,3.00)	12	(575,2.74)
3	(558,2.81)	8	(661,3.43)	13	(545,2.76)
4	(578,3.03)	9	(651,3.36)	14	(572,2.88)
5	(666,3.44)	10	(605,3.13)	15	(594,2.96)

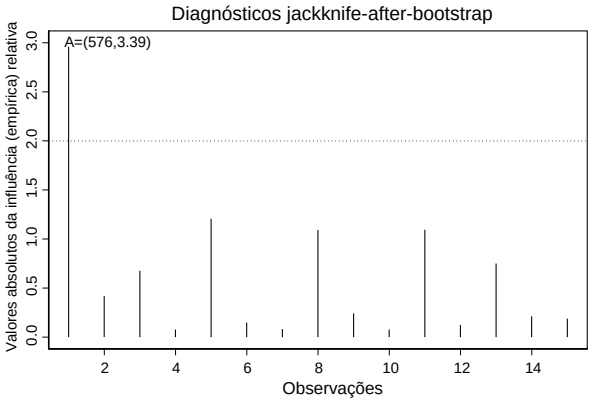


Figura 4.5: Representação dos valores *jackknife-after-bootstrap* para os dados de Faculdades de Direito.

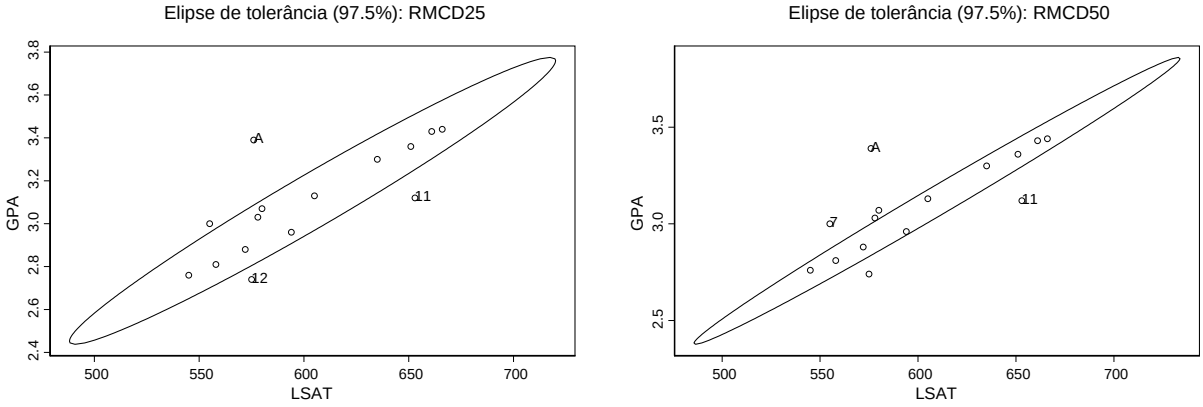


Figura 4.6: Representação das elipses de tolerância (a 97,5%) baseados nos estimadores RMCD para os dados de Faculdades de Direito.

Tabela 4.9: Limites aproximados a 90% de confiança e estimativas pontuais para o coeficiente de correlação entre as variáveis LSTA e GPA, por diversos métodos.

			Estimação pontual	
Método	Limite inferior	Limite superior	Média	Média aparada–5%
Fisher	0.509	0.907	$r = 0.776$	
Fisher sem o ponto A	0.735	0.959	$r_{-A} = 0.893$	
UB e R	0.522	0.947	0.770	0.778
UB e R sem o ponto A	0.779	0.968	0.886	0.890
UB e RMCD25	0.302	0.998	0.857	0.891
IFB e R	0.868	0.985	0.937	0.941
IFB e RMCD25	0.498	0.999	0.916	0.946
Bootstrap $BC_a(a = 0)$	0.488*	0.900*	—	—
Bootstrap-t	−0.060*	0.940*	—	—

* De Shao e Tu (1995, pág. 169)

singular). Da observação da tabela constata-se que o menor comprimento do intervalo é obtido para IFB e R. Este intervalo é bastante diferente do segundo e do quarto, o que mostra que deve haver outros pontos, para além de A, a afectar o valor de R. Assumindo que os dados provêm de uma distribuição Normal bivariada com alguma contaminação, e tendo em conta os resultados do estudo de Monte Carlo, é razoável concluir que esse intervalo é o melhor. A Tabela 4.9 também apresenta estimativas pontuais associadas a cada método de estimação. Para os métodos *bootstrap* indicam-se duas estimativas: a média aritmética e a média aparada a 5%, como sugerido no final da Secção 4.2. Interessa ainda observar que a aproximação dada pela transformação z de Fisher só é razoável para $n > 25$, sendo natural que possua um desempenho menos bom para valores de n inferiores.

■

Conclusões

Nesta secção avaliou-se a estimação intervalar do coeficiente de correlação linear usando procedimentos de reamostragem *bootstrap* e o método clássico de Fisher. A consideração de várias distribuições para o modelo gerador dos dados permitiu avaliar quer a robustez dos vários procedimentos quer a sua eficiência. As principais conclusões retiradas dos diversos estudos realizados foram as seguintes:

- O método *bootstrap* IF com o estimador não robusto R na fase de reamostragem apresenta melhores desempenhos globais, quer relativos às probabilidades de cobertura, quer relativos aos comprimentos dos intervalos. Esta conclusão é semelhante à principal conclusão da secção anterior.
- O método IFB não apresenta problemas de sub-cobertura em amostras de grandes dimensões

no modelo sem contaminação.

Um estudo desta natureza é sempre necessariamente limitado. Embora não tivesse sido possível fazê-lo (por uma questão de tempo) considera-se que teria sido desejável:

- Aprofundar o estudo para outros valores de dimensões amostrais, assim como variar a percentagem de contaminação da distribuição NSC–ε.
- Variar o valor de β para o *bootstrap Winsorized* assim como utilizar outras versões amostrais de profundidade para a ordenação da amostra, uma vez que, mais uma vez, este método não teve o desempenho esperado, mesmo para os estimadores robustos.
- Estudar o comportamento do coeficiente de correlação de Spearman quando se aplica o método de reamostragem *bootstrap* IF.

4.3.3 Estudo de eficiência de estimadores robustos multivariados em amostras finitas

Nesta secção apresenta-se um estudo de Monte Carlo para comparar o desempenho, em amostras finitas, dos vários estimadores de alto ponto rotura descritos no Capítulo 2. A inclusão desta análise precede o estudo da selecção de variáveis em Análise Discriminante (descrito na secção seguinte) dada a importância destes estimadores para toda a Análise Multivariada, em particular para a Análise Discriminante. Para além disso, e uma vez que o método bootstrap IF requer, no cálculo da função $RESIF$, estimadores de alto ponto de rotura, o estudo que aqui se realiza é fundamental para a aplicação deste método de reamostragem à selecção de variáveis em Análise Discriminante.

Segundo Hampel *et al.* (1986) os requisitos fundamentais que um bom estimador robusto deve possuir são: ter uma função de influência limitada, ter um ponto de rotura positivo e o mais elevado possível e finalmente possuir uma boa eficiência sob o modelo central considerado. Estimadores com estas três características em simultâneo são difíceis de obter. Por um lado a eficiência entra muitas vezes em conflito com os outros dois requisitos, isto é, a obtenção de uma boa eficiência sob o modelo central é em geral paga num decréscimo do ponto de rotura ou numa função de influência ilimitada. Por outro lado, a existência de um estimador com um ponto de rotura positivo não é garantia de "robustez" já que o seu viés pode ser bastante elevado, mesmo que seja limitado. É então importante avaliar o comportamento dos estimadores de alto ponto de rotura que foram descritos atrás.

Croux e Haesbroeck (1999) compararam a eficiência em amostras finitas dos estimadores MCD e S. O objectivo foi a estimação da matriz de covariâncias numa amostra. Do seu estudo de simulação estes autores concluíram que o estimador MCD com ponto de rotura de 0.25 é mais eficiente que o mesmo estimador com ponto de rotura de 0.50, sob o modelo normal. Aqueles autores concluíram ainda que as versões reponderadas, a um passo, possuem melhor desempenho que as

não reponderadas. Quanto à comparação do estimador S com o MCD, os resultados mostraram que o primeiro com um ponto de rotura de 0.25 possui uma eficiência finita superior ao segundo (mesmo considerando a sua versão reponderada).

Como foi referido no Capítulo 2, He e Fung (2000) usaram dois estimadores S para estimar a localização e dispersão em várias populações quando estas possuem matrizes de covariâncias comuns. Do estudo de simulação que realizaram concluíram que as versões reponderadas dos estimadores S (independentemente do método usado, A ou B) têm um melhor desempenho. E que o método A (a um passo) não é tão sensível às diferenças entre as dispersões nas populações.

O objectivo desta secção é fazer uma generalização do estudo realizado por He e Fung (2000). Assim para além dos estimadores S consideraram-se os estimadores MCD e MVE, comparando as suas eficiências finitas na estimação dos parâmetros de localização e dispersão em mais do que uma população com uma matriz de covariâncias comum.

Distribuições

Sem perda de generalidade considera-se o caso de duas populações, isto é $g = 2$. A hipótese básica neste estudo supõe que as distribuições de cada população possuem localizações distintas mas matrizes de covariâncias idênticas. Para além disso assume-se que o modelo central é a normal multivariada. O objectivo é verificar qual o comportamento (em termos de eficiência) dos estimadores atrás referidos quando se verificam afastamentos ao modelo central. Assim foram consideradas as seguintes distribuições para modelo gerador dos dados, $((\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_{n_1})\,|G_1)$ e $((\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_{n_2})\,|G_2)$:

- Normal, NOR:
$$\mathbf{X}_j|G_i \sim N_m(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad i = 1, 2, \quad j = 1, 2, \cdots, n_i;$$

- Normal Simetricamente ε-Contaminada com factor de escala f_e, NSC:
$$\mathbf{X}_j|G_i \sim (1-\varepsilon)N_m(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + \varepsilon N_m(\boldsymbol{\mu}_i, f_e\boldsymbol{\Sigma}_i), \quad i = 1, 2, \quad j = 1, 2, \cdots, n_i;$$

- Normal Assimetricamente ε-Contaminada, NAC:
$$\mathbf{X}_j|G_1 \sim N_m(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \text{ para } j = 1, 2, \cdots, n_1 - [\varepsilon n_1] \;\text{ e }\; \mathbf{X}_j|G_1 \sim \Delta_{\mathbf{x}^o} \text{ para } j = n_1 - [\varepsilon n_1] + 1, \cdots, n_1,$$
$$\mathbf{X}_j|G_2 \sim N_m(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \quad j = 1, 2, \cdots, n_2;$$

- t–Student com 3 graus de liberdade[1], T3:
$$\mathbf{X}_j|G_i \sim t_{m,3}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad i = 1, 2, \quad j = 1, 2, \cdots, n_i;$$

[1] $\mathbf{Y} \sim t_{m,\upsilon}$ se $\mathbf{Y} = V^{-1}\mathbf{X} + \boldsymbol{\mu}$ com $\mathbf{X} \sim N_m(\mathbf{0}, \boldsymbol{\Sigma})$ e $\upsilon V^2 \sim \chi^2_\upsilon$.

Em todos os casos os parâmetros escolhidos foram

$$\begin{array}{ll} \boldsymbol{\mu}_1 = \mathbf{0} & \boldsymbol{\Sigma}_1 = \mathbf{I}_m \\ \boldsymbol{\mu}_2 = (\mu, 0, \cdots, 0)^T & \boldsymbol{\Sigma}_2 = \mathbf{I}_m. \end{array}$$

De notar que a escolha da matriz identidade não implica perda de generalidade uma vez que qualquer outra situação ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 \neq \mathbf{I}_m$) pode ser reduzida a esta por aplicação de uma transformação linear conveniente, para a qual os estimadores usados são equivariantes. As localizações de cada grupo estão afastadas de μ, tendo-se optado por escolher $\mu = 2$.

O número de variáveis escolhido foi de $m = 2, 5$ e 10 para cada $n_1 = n_2 = 50$ e $n_1 = 100$ e $n_2 = 300$.

Para a distribuição NSC fixou-se $\varepsilon = 0.1$ e $f_e = 9$, para qualquer $i = 1, 2$. Esta distribuição é vulgarmente usada em estudos de simulação para avaliação de robustez pois reflecte uma situação de observações discordantes em termos de dispersão.

Representou-se por NAC a distribuição contaminada (apenas num dos grupos) com $n_1 \times \varepsilon$ pontos concentrados em $\mathbf{x}^o$, correspondendo esta situação à simulação de erros grosseiros. No estudo presente fixou-se $\varepsilon = 0.1$. O ponto $\mathbf{x}^o$ foi escolhido de forma a que quando se usa a distribuição NAC a direcção discriminante entre os dois grupos seja ortogonal à direcção discriminante sem contaminação. Escolheu-se então o ponto $\mathbf{x}^o = (22, 2)$, para $m = 2$ e $\mathbf{x}^o = (22, 2, 0, 0, \cdots, 0)$ para $m = 5, 10$ (os cálculos apresentam-se na Secção A.2 no Apêndice A).

Estimadores

- Estimadores S com ponto de rotura 0.25 e 0.50 denotados por S25 e S50 respectivamente;
- Estimadores S, MCD e MVE nas suas versões reponderadas (a um passo) que serão representados por, RS25, RS50, RMCD25, RMCD50, RMVE50;
- Estimadores anteriores usando o método A, que serão representados de igual forma acrescentando apenas a letra "A" no final, por exemplo S25A ou RS25A.

Quando se estima a matriz de covariâncias comum, $\boldsymbol{\Sigma}_C$, considera-se, como usualmente, a matriz combinada,

$$\hat{\boldsymbol{\Sigma}}_C = \frac{(n_1 - 1)\,\hat{\boldsymbol{\Sigma}}_1 + (n_2 - 1)\,\hat{\boldsymbol{\Sigma}}_2}{n_1 + n_2 - 2}.$$

Como controlo obtiveram-se as estimativas correspondentes aos estimadores óptimos (máxima verosimilhança) quando o modelo central é a Normal. Neste estudo esse estimador será referido como "Clássico".

Forma de obtenção de resultados

Para cada distribuição e fixos os vários intervenientes (dimensões e parâmetros dessa distribuição) gera-se uma amostra e calcula-se o valor de cada uma das estimativas com base nela. Como referido

anteriormente esta forma de simulação conduz a resultados dependentes mas reduz a variabilidade das diferenças dos métodos de estimação dois a dois, já que a amostra é a mesma. A opção por este processo de obtenção de resultados, que é diferente da que foi adoptada nos estudos de simulação anteriores, foi tomada por se pretender efectuar algumas comparações com os resultados de He e Fung (2000).

Critérios de avaliação

Na comparação de estimadores vários critérios podem ser adoptados. Neste estudo optou-se por usar a estimativa do erro quadrático médio (MSE). Denotando por $\mathbf{T}$ um qualquer estimador de um parâmetro $\boldsymbol{\theta}$ define-se $\widehat{MSE(\mathbf{T})}$ como

$$\widehat{MSE(\mathbf{T})} = (ns)^{-1} \sum_{k=1}^{ns} \|\mathbf{t}_k - \boldsymbol{\theta}\|^2,$$

onde $\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_{ns}$ são as ns estimativas obtidas pelo método de Monte Carlo. Para cada grupo, geraram-se 100 amostras (isto é, $ns = 100$).

Assim,

- para os vectores de localização a estimativa de MSE é dada por

$$\widehat{MSE(\hat{\boldsymbol{\mu}}_g)} = ns^{-1} \sum_{k=1}^{ns} \sum_{i=1}^{m} \left(\hat{\mu}_{g_i}^{(k)} - \mu_{g_i} \right)^2,$$

  onde $\hat{\mu}_{g_i}^{(k)}$ corresponde à i-ésima componente da estimativa de localização do grupo g obtida a partir da k-ésima amostra desse grupo e μ_{g_i} é a i-ésima componente do vector $\boldsymbol{\mu}_g$, $g = 1, 2$.

- para a matriz de covariâncias a estimativa de MSE é dada por

$$\widehat{MSE(\hat{\boldsymbol{\Sigma}}_C)} = ns^{-1} \sum_{k=1}^{ns} \sum_{i=1}^{m} \sum_{j=i}^{m} \left(\hat{\sigma}_{ij}^{(k)} - I_{ij} \right)^2,$$

  onde $\hat{\sigma}_{ij}^{(k)}$ corresponde ao elemento (i, j) da matriz de covariâncias estimada a partir da k-ésima amostra e I_{ij} é o elemento (i, j) da matriz identidade de ordem m.

Algoritmos

Para o cálculo do MCD usou-se o algoritmo FAST-MCD referido anteriormente e que está implementado no software estatístico S-PLUS. Para o cálculo do MVE usou-se o algoritmo genético que está implementado no mesmo software. Por defeito estes dois métodos estão implementados para a obtenção de estimativas associadas a estimadores de ponto de rotura de 50%. No caso do estimador MCD basta mudar, na macro de cálculo, um parâmetro de entrada para obter as estimativas associadas a estimadores com ponto de rotura de 25%. O procedimento implementado no SPLUS

para o MVE não permite essa mudança, o que reduziu a escolha ao MVE com ponto de rotura de aproximadamente 50%. As estimativas S obtiveram-se após a implementação em S-PLUS do algoritmo SURREAL. Neste caso, o resultado da Proposição 2.1 permite obter facilmente estimativas associadas aos estimadores com os dois pontos de rotura referidos.

Resultados

Os resultados da simulação estão resumidos nas Tabelas 4.10 a 4.17. Em cada tabela, e para cada parâmetro, realça-se a negrito o valor mínimo da estimativa de MSE. Uma vez que um dos objectivos deste estudo é comparar a precisão dos estimadores robustos foi salientado, a sublinhado, para cada parâmetro, o menor valor da estimativa de MSE dentro da classe dos estimadores robustos.

Antes de se proceder à análise dos resultados é necessário observar que os valores correspondentes às estimativas de localização aplicando o método A, descrito anteriormente, são sempre iguais às estimativas obtidas pelos estimadores "simples". Significa isto que, por exemplo, os valores de RMCD25 para $\hat{\boldsymbol{\mu}}_g$ são iguais às estimativas de RMCD25A para $\hat{\boldsymbol{\mu}}_g$, pois o procedimento relativo ao método A foi aplicado apenas a um passo, o que conduz à introdução somente de alterações na estimativa da matriz de covariâncias comum e não na estimativa de localização de cada grupo.

Na Tabela 4.10 apresentam-se os resultados para a distribuição Normal quando a relação entre as dimensões das amostras para cada população é de 1:1 ($n_1 = n_2 = 50$). Aí sobressai desde logo, a optimalidade dos estimadores clássicos, assim como a boa eficiência dos estimadores S de ponto de rotura 0.25. Outro facto saliente, mas esperado, é o crescimento de MSE com a dimensão do espaço. No entanto, e de um modo geral, realça-se a boa prestação de qualquer um dos estimadores usados.

Nas Tabelas 4.11, 4.12 e 4.13 encontram-se os $\widehat{MSE}$ para distribuições diferentes do modelo central. Para estas observa-se uma grande degradação dos estimadores clássicos, independentemente do número de variáveis considerado.

Tabela 4.10: Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis e $n_1 = n_2 = 50$, para a distribuição NOR.

Estimador	$m = 2$			$m = 5$			$m = 10$		
	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$
Clássico	**0.0408**	**0.0392**	**0.0511**	**0.0986**	**0.0983**	**0.1869**	**0.2028**	0.2034	0.6543
RMCD25	0.0552	0.0560	0.1426	0.1386	0.1372	0.3977	0.3238	0.2969	1.0989
RMCD25A	0.0552	0.0560	0.1201	0.1386	0.1372	0.3038	0.3238	0.2969	0.9537
S25	0.0429	0.0418	0.0582	0.1014	0.1019	0.1922	0.2058	**0.2028**	0.6388
RS25	0.0451	0.0430	0.0817	0.1057	0.1071	0.2135	0.2134	0.2104	0.6530
S25A	0.0429	0.0418	0.0585	0.1014	0.1019	0.1921	0.2058	**0.2028**	**0.6368**
RS25A	0.0445	0.0437	0.0838	0.1070	0.1072	0.2199	0.2133	0.2099	0.6649
RMCD50	0.0694	0.0794	0.2296	0.1930	0.2046	0.6136	0.3941	0.3835	1.4210
RMCD50A	0.0694	0.0794	0.1703	0.1930	0.2046	0.4453	0.3941	0.3835	1.2641
RMVE50	0.0482	0.0479	0.1023	0.1116	0.1082	0.2322	0.2314	0.2243	0.7233
RMVE50A	0.0482	0.0479	0.0948	0.1116	0.1082	0.2337	0.2314	0.2243	0.7134
S50	0.0702	0.0702	0.1442	0.1214	0.1236	0.2602	0.2263	0.2130	0.7328
RS50	0.0513	0.0501	0.1180	0.1128	0.1106	0.2461	0.2239	0.2180	0.6964
S50A	0.0702	0.0702	0.1378	0.1214	0.1236	0.2485	0.2263	0.2130	0.7025
RS50A	0.0482	0.0482	0.1095	0.1129	0.1113	0.2404	0.2216	0.2102	0.6993

Tabela 4.11: Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis e $n_1 = n_2 = 50$, para a distribuição NSC.

Estimador	$m = 2$			$m = 5$			$m = 10$		
	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$
Clássico	0.0712	0.0754	1.8142	0.1748	0.1884	5.2547	0.3878	0.3834	13.5314
RMCD25	0.0511	0.0563	0.0951	0.1594	0.1427	0.3431	0.3049	0.3394	1.0350
RMCD25A	0.0511	0.0563	0.0848	0.1594	0.1427	0.2980	0.3049	0.3394	0.9389
S25	**0.0402**	**0.0491**	0.4063	**0.1292**	0.1252	1.5615	0.2556	0.2705	5.0671
RS25	0.0427	0.0510	0.0738	0.1346	0.1212	0.3009	0.2505	0.2522	0.8960
S25A	**0.0402**	**0.0491**	0.2244	**0.1292**	0.1252	1.3445	0.2556	0.2705	4.2312
RS25A	0.0415	0.0514	**0.0663**	0.1336	**0.1194**	0.2774	0.2452	**0.2433**	0.7621
RMCD50	0.0539	0.0804	0.1446	0.2098	0.1824	0.5230	0.3838	0.4374	1.3320
RMCD50A	0.0539	0.0804	0.1223	0.2098	0.1824	0.3859	0.3838	0.4374	1.2442
RMVE50	0.0471	0.0543	0.1140	0.1374	0.1352	0.3468	0.2554	0.2667	0.8654
RMVE50A	0.0471	0.0543	0.0995	0.1413	0.1352	0.2800	0.2554	0.2667	0.7935
S50	0.0583	0.0726	0.2701	0.1486	0.1434	1.1417	**0.2410**	0.2627	2.4681
RS50	0.0439	0.0541	0.0858	0.1404	0.1276	0.2668	0.2455	0.2564	0.7878
S50A	0.0583	0.0726	0.2244	0.1404	0.1434	0.6160	**0.2410**	0.2627	1.8126
RS50A	0.0435	0.0546	0.0767	0.1353	0.1225	**0.2644**	0.2445	0.2495	**0.7362**

Tabela 4.12: Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis e $n_1 = n_2 = 50$, para a distribuição NAC.

Estimador	$m = 2$			$m = 5$			$m = 10$		
	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$
Clássico	6.4058	**0.0404**	828.5380	6.4168	**0.1047**	827.8508	6.6100	**0.1899**	825.0379
RMCD25	0.0466	0.0602	0.1083	0.1464	0.1416	0.3694	0.3150	0.2987	1.1065
RMCD25A	0.0466	0.0602	0.0948	0.1464	0.1416	0.3027	0.3150	0.2987	0.9942
S25	**0.0422**	0.0451	0.2982	0.6081	**0.1047**	78.8775	6.9903	0.1905	901.0940
RS25	0.0425	0.0450	0.0694	0.9776	0.1095	106.7202	6.8572	0.1932	816,4223
S25A	**0.0422**	0.0451	0.1931	0.6081	**0.1047**	50.3803	6.9903	0.1905	381.4608
RS25A	0.0431	0.0426	**0.0654**	**0.1155**	0.1056	**0.2282**	3.2631	0.1938	370.3501
RMCD50	0.0564	0.0818	0.1766	0.1806	0.1743	0.5635	0.4013	0.3897	1.4193
RMCD50A	0.0564	0.0818	0.1522	0.1806	0.1743	0.4406	0.4013	0.3897	1.3029
RMVE50	0.0450	0.0584	0.0790	0.1221	0.1121	0.2501	**0.2516**	0.2155	0.7946
RMVE50A	0.0450	0.0584	0.0928	0.1221	0.1121	0.2670	**0.2516**	0.2155	0.7689
S50	0.0542	0.0795	0.2092	0.1274	0.1184	0.3943	7.6079	0.2052	1152.9129
RS50	0.0457	0.0489	0.0966	0.1201	0.1108	0.2585	5.9690	0.2056	697.4287
S50A	0.0542	0.0795	0.1767	0.1274	0.1184	0.3495	7.6079	0.2052	26.9510
RS50A	0.0461	0.0445	0.0854	0.1213	0.1104	0.2484	0.2580	0.1957	**0.7464**

Tabela 4.13: Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis e $n_1 = n_2 = 50$, para a distribuição T3.

Estimador	$m = 2$			$m = 5$			$m = 10$		
	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$
Clássico	0.1035	0.7455	38.6207	0.3153	1.1469	1366.0941	0.6107	1.3421	175.8961
RMCD25	0.0587	0.0967	0.2014	0.1636	0.1850	0.6368	0.3280	0.3579	2.1189
RMCD25A	0.0587	0.0967	0.2207	0.1636	0.1850	0.6394	0.3280	0.3579	1.8149
S25	**0.0527**	0.1246	2.4665	0.1575	0.2960	7.9094	0.3303	0.5066	19.0375
RS25	0.0610	0.1478	0.6003	0.1694	0.2906	1.9445	0.3415	0.4604	4.9198
S25A	**0.0527**	0.1246	2.2338	0.1575	0.2960	17.8036	0.3303	0.5066	20.2716
RS25A	0.0635	0.1177	0.5268	0.1695	0.2471	1.8283	0.3488	0.4075	3.8672
RMCD50	0.0614	0.0981	**0.1956**	0.1703	0.2253	0.6559	0.3421	0.4007	1.7407
RMCD50A	0.0614	0.0981	0.2306	0.1703	0.2253	**0.5762**	0.3421	0.4007	**1.5638**
RMVE50	0.0589	0.1228	0.4019	0.1624	0.2514	1.3105	0.3360	0.3705	2.9100
RMVE50A	0.0589	0.1228	0.3283	0.1624	0.2514	0.9665	0.3360	0.3705	2.1053
S50	0.0582	**0.0808**	0.9713	**0.1402**	**0.1634**	4.8497	**0.2727**	**0.3102**	8.7574
RS50	0.0582	0.1047	0.2340	0.1603	0.2298	0.9824	0.3216	0.3721	2.5246
S50A	0.0582	**0.0808**	0.8382	**0.1402**	**0.1634**	6.0430	**0.2727**	**0.3102**	8.4051
RS50A	0.0590	0.0892	0.2266	0.1632	0.2053	0.9716	0.3237	0.3382	2.2087

Tabela 4.14: Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis, $n_1 = 100$ e $n_2 = 300$, para a distribuição NOR.

Estimador	$m = 2$			$m = 5$			$m = 10$		
	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$
Clássico	**0.0182**	**0.0070**	**0.0141**	**0.0589**	**0.0207**	**0.0536**	**0.0983**	**0.0352**	**0.1626**
RMCD25	0.0230	0.0083	0.0401	0.0703	0.0222	0.0908	0.1293	0.0386	0.2296
RMCD25A	0.0230	0.0083	0.0374	0.0703	0.0222	0.0823	0.1293	0.0386	0.2045
S25	0.0191	0.0077	0.0157	0.0607	0.0211	0.0545	0.0995	0.0357	0.1627
RS25	0.0191	0.0078	0.0305	0.0624	0.0223	0.0699	0.1027	0.0369	0.1809
S25A	0.0191	0.0077	0.0157	0.0607	0.0211	0.0545	0.0995	0.0357	**0.1626**
RS25A	0.0209	0.0079	0.0307	0.0633	0.0224	0.0699	0.1047	0.0369	0.1828
RMCD50	0.0275	0.0090	0.0623	0.0853	0.0239	0.1268	0.1709	0.0424	0.2945
RMCD50A	0.0275	0.0090	0.0520	0.0853	0.0239	0.1011	0.1709	0.0424	0.2296
RMVE50	0.0196	0.0092	0.0400	0.0645	0.0230	0.0797	0.1105	0.0382	0.1962
RMVE50A	0.0196	0.0092	0.0400	0.0645	0.0230	0.0798	0.1105	0.0382	0.1993
S50	0.0321	0.0117	0.0325	0.0707	0.0237	0.0668	0.1074	0.0382	0.1756
RS50	0.0203	0.0080	0.0358	0.0655	0.0223	0.0746	0.1063	0.0373	0.1872
S50A	0.0321	0.0117	0.0323	0.0707	0.0237	0.0663	0.1074	0.0382	0.1746
RS50A	0.0215	0.0080	0.0353	0.0648	0.0224	0.0733	0.1057	0.0373	0.1860

Quanto à distribuição NSC (com $n_1 = n_2 = 50$), Tabela 4.11, verifica-se que, e de uma forma geral, continuam a ser os estimadores S25 que conduzem a melhores resultados dentro dos estimadores robustos e não robustos usados. Mais em detalhe, pode verificar-se que, e apenas no que diz respeito aos estimadores S (já que são os únicos onde esta comparação faz sentido), a estimação da variância populacional melhora quando se utilizam estimadores robustos reponderados. Porém, o mesmo não ocorreu na estimação dos valores esperados, apesar da diferença ser ligeira.

A Tabela 4.12 evidencia desde logo um pior desempenho dos estimadores S não reponderados à medida que o número de variáveis aumenta (ou se reduz a relação $n_i : m$). Obtendo-se inclusive, para $m = 10$, um valor de $\widehat{MSE}$ de aproximadamente 1153, para a estimação de $\boldsymbol{\Sigma}_C$ quando se usa o estimador S50 (superior mesmo ao $\widehat{MSE}$ associado ao estimador clássico). Ainda para este número de variáveis observa-se que a estimação de $\boldsymbol{\Sigma}_C$ para o estimador RS25A não é muito boa (aliás para qualquer estimador S, mesmo com o método A, exceptuando o RS50A), estes resultados podem dever-se ao número de observações atípicas que é tolerado pelo estimador S usando o método A. Esta situação está de acordo com a observação de He e Fung (2000) relativa ao número de observações atípicas que é tolerado quando a relação entre as dimensões das amostras é próxima de 1 (neste caso tem-se mesmo 1). Comparando com os resultados que constam da Tabela 4.16 para o mesmo estimador, RS25A, observa-se que este facto não ocorreu pois aqui aquela relação é de 1:3.

Tabela 4.15: Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis, $n_1 = 100$ e $n_2 = 300$, para a distribuição NSC.

Estimador	$m = 2$			$m = 5$			$m = 10$		
	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$
Clássico	0.0369	0.0122	1.3809	0.0812	0.0305	3.6201	0.1800	0.0573	7.9396
RMCD25	0.0291	0.0093	0.0246	0.0666	0.0209	0.0779	0.1309	0.0399	0.2337
RMCD25A	0.0291	0.0093	0.0229	0.0666	0.0209	0.0751	0.1309	0.0399	0.2160
S25	0.0268	**0.0086**	0.2529	**0.0573**	0.0203	0.9188	0.1168	0.0382	2.3040
RS25	0.0272	**0.0086**	0.0219	0.0596	0.0203	0.0715	0.1103	**0.0376**	0.1926
S25A	0.0268	**0.0086**	0.2465	**0.0573**	0.0203	0.8940	0.1168	0.0382	2.2326
RS25A	0.0275	0.0089	0.0217	0.0585	0.0204	0.0704	**0.1096**	**0.0376**	**0.1912**
RMCD50	0.0337	0.0102	0.0379	0.0789	0.0223	0.1055	0.1688	0.0421	0.2853
RMCD50A	0.0337	0.0102	0.0330	0.0789	0.0223	0.0878	0.1688	0.0421	0.2329
RMVE50	0.0307	0.0100	0.0271	0.0616	0.0219	0.0758	0.1136	0.0398	0.2137
RMVE50A	0.0307	0.0100	0.0279	0.0616	0.0219	0.0766	0.1136	0.0398	0.2110
S50	0.0362	0.0123	0.1042	0.0675	0.0217	0.3104	0.1146	0.0388	0.8244
RS50	**0.0267**	0.0088	0.0215	0.0619	**0.0201**	0.0724	0.1111	0.0381	0.1969
S50A	0.0362	0.0123	0.0987	0.0675	0.0217	0.2398	0.1146	0.0388	0.9612
RS50A	0.0284	0.0089	**0.0212**	0.0610	0.0202	**0.0695**	0.1105	0.0382	0.1962

Tabela 4.16: Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis, $n_1 = 100$ e $n_2 = 300$, para a distribuição NAC.

Estimador	$m = 2$			$m = 5$			$m = 10$		
	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$
Clássico	6.3155	**0.0077**	200.8812	6.3906	**0.0157**	201.0456	6.3722	**0.0331**	200.9291
RMCD25	0.0259	0.0095	0.0387	0.0636	0.0175	0.0813	0.1311	0.0365	0.2397
RMCD25A	0.0259	0.0095	0.0365	0.0636	0.0175	0.0764	0.1311	0.0365	0.2208
S25	0.0241	0.0087	0.0631	1.0278	0.0159	37.3508	6.8325	0.0332	227.1024
RS25	**0.0234**	0.0089	0.0259	1.7165	0.0170	49.9537	6.7710	0.0345	206.4010
S25A	0.0241	0.0087	0.0358	1.0278	0.0159	21.6141	6.8325	0.0332	5.4592
RS25A	0.0238	0.0085	**0.0254**	0.0596	0.1610	**0.0565**	0.1180	0.0334	**0.1854**
RMCD50	0.0309	0.0108	0.0581	0.0744	0.0186	0.1152	0.1662	0.0384	0.3045
RMCD50A	0.0309	0.0108	0.0520	0.0744	0.0186	0.0932	0.1662	0.0384	0.2437
RMVE50	0.0269	0.0100	0.0396	0.0610	0.0168	0.0752	**0.1164**	0.0352	0.2076
RMVE50A	0.0269	0.0100	0.0397	0.0610	0.0168	0.0768	**0.1164**	0.0352	0.2108
S50	0.0355	0.0139	0.0342	0.0634	0.0185	0.0831	7.6694	0.0349	289.6951
RS50	0.0255	0.0094	0.0342	**0.0578**	0.0170	0.0695	5.9200	0.0347	178.8206
S50A	0.0355	0.0139	0.0311	0.0634	0.0185	0.0769	7.6694	0.0349	25.3594
RS50A	0.0255	0.0091	0.0330	0.0599	0.0168	0.0693	0.1194	0.0336	0.1899

Tabela 4.17: Desempenho, em termos de $\widehat{MSE}$, dos vários estimadores para $m = 2, 5$ e 10 variáveis, $n_1 = 100$ e $n_2 = 300$, para a distribuição T3.

Estimador	$m = 2$			$m = 5$			$m = 10$		
	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_C$
Clássico	0.0632	0.6009	54.4273	0.1432	0.6802	357.0222	0.2615	0.7047	158.3950
RMCD25	0.0370	0.0279	0.1805	0.0809	0.0387	0.2552	0.1552	0.0537	0.4567
RMCD25A	0.0370	0.0279	0.1903	0.0809	0.0387	0.2600	0.1552	0.0537	0.4291
S25	**0.0297**	0.0585	3.0073	0.0756	0.1182	18.2203	0.1418	0.1505	61.8907
RS25	0.0343	0.0787	0.6914	0.0859	0.1232	1.7141	0.1584	0.0985	2.0011
S25A	**0.0297**	0.0585	2.8727	0.0756	0.1182	21.0404	0.1418	0.1505	255.9835
RS25A	0.0338	0.0672	0.6356	0.0856	0.1140	1.7434	0.1562	0.0985	2.1758
RMCD50	0.0348	0.0227	**0.1446**	0.0801	0.0339	**0.1835**	0.1575	0.0572	0.4059
RMCD50A	0.0348	0.0227	0.1746	0.0801	0.0339	0.2024	0.1575	0.0572	**0.3746**
RMVE50	0.0386	0.0321	0.2509	0.0869	0.0449	0.3754	0.1608	0.0574	0.5692
RMVE50A	0.0386	0.0321	0.2222	0.0869	0.0449	0.3460	0.1608	0.0574	0.5155
S50	0.0359	**0.0201**	0.9713	**0.0686**	**0.0295**	28.7828	**0.1238**	**0.0434**	17.7821
RS50	0.0335	0.0338	0.2403	0.0796	0.0794	1.1188	0.1521	0.0650	0.9521
S50A	0.0359	**0.0201**	0.9153	**0.0686**	**0.0295**	1.9475	**0.1238**	**0.0434**	57.9740
RS50A	0.0341	0.0293	0.2401	0.0850	0.1005	1.9848	0.1541	0.0660	1.1750

Para a distribuição T3, Tabela 4.17, as estimativas de MSE relativas à estimação de $\boldsymbol{\Sigma}_C$ apresentam melhores resultados para os estimadores RMCD50, quer o reponderado simples quer o obtido pelo método A. Estes resultados são distintos dos obtidos para as restantes distribuições. Todavia, os estimadores RS25A e RS50A têm também bons desempenhos.

Os resultados obtidos quando a relação entre as dimensões das amostras para cada população é de 1:3 encontram-se nas Tabelas 4.14 a 4.17. A informação que consta na Tabela 4.14 é semelhante à dada pela Tabela 4.10, devendo-se apenas observar que alguns valores são idênticos apenas devido ao arredondamento efectuado.

Relativamente às restantes distribuições os comentários efectuados anteriormente continuam, de um modo geral, válidos. Exceptua-se o já referido valor de $\widehat{MSE}$ para $\boldsymbol{\Sigma}_C$ na Tabela 4.16 para $m = 10$ e RS25A.

Face aos resultados e comentários apresentados podem salientar-se os seguintes aspectos:

- Independentemente do número de variáveis e da relação entre as dimensões amostrais de cada grupo, é a utilização do método A de He e Fung (2000) que produz os melhores resultados para a estimação da matriz de covariâncias comum. Para os vectores médios não se pode tirar conclusão idêntica já que os resultados do método A são equivalentes à estimação "simples".

- Quando se tem amostras de dimensão $n_1 = n_2 = 50$ e $m = 2$ ou $m = 5$ e se pretende usar um estimador de ponto de rotura de 25% a melhor opção é o estimador RS25A; querendo usar um estimador com maior ponto de rotura deve optar-se por RS50A.

- Quando se tem amostras de dimensão $n_1 = n_2 = 50$ e $m = 10$ e se pretende utilizar um estimador com um ponto de rotura de 25% aconselha-se o estimador RMCD25A; se, no entanto, for requerido um maior ponto de rotura o conselho recai no estimador RS50A ou no estimador RMVE50A;
- Aconselha-se a utilização de idênticos estimadores quando a relação entre as dimensões das amostras seja de 1:3 (naturalmente para os mesmos valores de m).

Observação 4.4 Deve ser feita uma chamada de atenção relativa à necessidade da distinção entre algoritmo e método de estimação. Significa isto que se pode ter um estimador com boas propriedades mas ter um algoritmo de cálculo muito ineficiente e, por outro lado, ter-se um estimador com propriedades menos boas mas com um bom algoritmo de cálculo. O exemplo desta segunda situação encontra-se no estimador MVE, já que se sabe que em termos assintóticos possui menor eficiência que o MCD, quando o modelo central é a Normal multivariada, contudo verificou-se que, em praticamente todas as situações, os seus resultados foram melhores que os do MCD. Naturalmente que estes resultados podem não ser apenas devidos ao algoritmo de cálculo mas também a questões de eficiência finita. Trata-se de uma questão que tem interesse investigar mas que sai fora do âmbito desta tese.

■

Observação 4.5 Neste estudo é apenas o estimador S aquele que aparece na sua versão simples, isto é, sem qualquer reponderação, e quando se aplica o método A a centragem da amostra conjunta é realizada utilizando as estimativas S simples enquanto que nas restantes situações se utilizam os estimadores (quer MVE quer MCD) na sua versão reponderada, por isso conjectura-se que a aplicação do método A com a centragem nas estimativas S reponderadas produza ainda melhores resultados.

■

4.3.4 Selecção robusta de variáveis em Análise Discriminante Linear

Como último exemplo de aplicação do método de reamostragem proposto considera-se o problema da selecção robusta de variáveis em Análise Discriminante. Este tema assume particular relevância, uma vez que, com a crescente facilidade no acesso à informação, os utilizadores são muitas vezes confrontados com um excesso de variáveis, as quais perturbam o sistema de estimação ou predição, introduzindo-lhe ruído. Importa então dispor de critérios objectivos de selecção. Embora para alguns métodos de Análise Discriminante, ditos clássicos, exista uma resposta (também clássica) para esta questão, ela não é necessariamente a melhor, nomeadamente em termos de robustez. A aplicação do método *bootstrap* IF à selecção de variáveis adequadas em Análise Discriminante Linear permite introduzir um novo critério de selecção que tem a vantagem de ser robusto. Uma vez que a aplicação deste método de reamostragem requer alguns conceitos formais da Análise Discriminante

far-se-á nesta secção um curto resumo acerca deste tópico. Por último, apresentam-se dois estudos de Monte Carlo onde se ilustra a aplicação do método de reamostragem proposto.

Em Análise Multivariada as entidades são vectores representados por pontos num espaço multidimensional. O principal objectivo da Análise Discriminante (AD) é definir regiões nesse espaço de forma a que novos pontos (e consequentemente as entidades correspondentes) possam ser classificados conforme a região a que pertencem. De um modo geral, pode-se dizer que a Análise Discriminante permite construir uma regra estatística de decisão que conduz, baseando-se nos dados observados, à identificação da população (ou grupo) a que pertence uma entidade com determinadas características. Uma boa regra deverá conduzir a erros de afectação (classificação) de observações futuras o mais pequenos possíveis. O problema da Análise Discriminante coloca-se, como é óbvio, quando o número de populações é superior ou igual a dois, já que o caso igual a um pode ser encarado como uma análise de observações discordantes e não de classificação. Essa regra de decisão, que também é designada por regra de classificação, permite não só classificar mas também averiguar se existem diferenças entre as populações e caso existam em que direcções são mais importantes.

Formalmente tem-se, com respeito ao universo ou população de interesse, g grupos bem definidos, $G_1, G_2, \cdots, G_g$, $g \geq 2$. Estes grupos são conhecidos *a priori* e constituiem uma partição do universo. Cada entidade do universo é caracterizada por um ponto, $\mathbf{x} = (x_1, \ldots, x_m)$, que é uma concretização de um vector de m variáveis aleatórias $(X_1, \ldots, X_m)^T \in I\!R^m$ previamente definidas (sem perda de generalidade assume-se que as variáveis aleatórias assumem valores reais). Este vector aleatório m-dimensional tem, condicional ao grupo, $\mathbf{X}|G_i$, função de densidade de probabilidade $f_i(\mathbf{x})$ (f.d.p.), com $i = 1, 2, \cdots, g$. Para além disso supõe-se que a entidade tem uma probabilidade *a priori*, π_i de provir do grupo i, onde $\sum_{i=1}^{g} \pi_i = 1$ e $\pi_i \geq 0$, para $i, j = 1, 2, \cdots, g$. A f.d.p. de $\mathbf{X}$ na população é então dada pela mistura finita $f_{\mathbf{X}}(\mathbf{x}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{x})$. O objectivo é encontrar uma regra de decisão, δ, aplicação de $I\!R^m$ em $\{1, 2, \cdots, g\}$, que permita classificar as entidades e satisfaça determinado critério de optimalidade. Uma regra de decisão, δ, conduz a uma partição em k regiões $R_{\delta(1)}, R_{\delta(2)}, \cdots, R_{\delta(g)}$, onde $R_{\delta(i)} = \{\mathbf{x} \in I\!R^m : \delta(\mathbf{x}) = i\}$, onde $\delta(\mathbf{x}) = i$ significa que as entidades associadas ao vector $\mathbf{x}$ são afectas ao grupo G_i. Para medir a qualidade de uma regra de decisão define-se a matriz dos custos de má classificação, $C = [c(j|i)]$, onde $c(j|i)$ elemento genérico dessa matriz, representa o custo de classificar em G_j uma entidade do grupo G_i $(1 \leq j, i \leq g)$, naturalmente que se $i = j$ então $c(j|i) = 0$.

O risco de Bayes de uma regra de decisão δ representa o seu custo médio e é dado por

$$R(\delta) = \sum_{i=1}^{g} \pi_i \sum_{j=1}^{g} C(j|i) \int_{R_{\delta(j)}} f_i(\mathbf{x}) d\mathbf{x}.$$

A regra, δ^s, que minimiza $R(\delta)$ é denominada regra de decisão de Bayes. Esta regra generaliza vários critérios de decisão propostos na literatura, sendo a sua particularização efectuada através da relaxação de alguns dos intervenientes ou das suas relações. Por exemplo, se os custos de afectação forem iguais $c(j|i) = k$, para todo o $i \neq j$, a regra de decisão de Bayes consiste na afectação de

cada $\mathbf{x}$ ao grupo de probabilidade *a posteriori* máxima, ou seja,

$$\delta^s(\mathbf{x}) = \underset{i=1,2,\cdots,g}{\arg\max}\, \pi_i f_i(\mathbf{x}).$$

Quando se ignoram os custos e as probabilidades *a priori*, π_i, obtém-se a regra dada pelo critério da máxima verosimilhança. Na prática, as quantidades em jogo na construção da regra de decisão de Bayes são: os custos $c(j|i)$, as probabilidades *a priori* π_i e as densidades por grupo $f_i(\mathbf{x})$. Na maior parte dos casos, os custos são supostos iguais e todos iguais a um ($c(j|i) = 1$, $i \neq j$). A possibilidade de estimar as probabilidades *a priori* depende do esquema de amostragem adoptado na constituição da amostra de treino (misto ou separado). Por último, a estimação das densidades de cada grupo, $f_i(\mathbf{x})$, constitui o fundamento da estimação de uma regra estatística de decisão, distinguindo-se os vários métodos existentes pela forma de modelar e estimar essas densidades.

Existe, porém, uma regra discriminante que é baseada num critério distinto das referidas. Trata-se da regra discriminante de Fisher, Fisher (1936), cujo critério é a maximização da separação entre os grupos. Este critério, sendo historicamente o primeiro, continua a ser um dos mais aplicados, não só pela sua simplicidade (devido a assumir que as fronteiras das regiões são lineares), como também pelo facto de não assumir à partida nenhuma forma distribucional para as densidades de cada grupo. A função discriminante linear de Fisher, para dois grupos, é dada por

$$z(\mathbf{x}) = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{x}, \tag{4.20}$$

com $\boldsymbol{\alpha} \propto \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, onde $\boldsymbol{\mu}_i$ representa o valor médio do grupo i, $i = 1, 2$ e $\boldsymbol{\Sigma}^{-1}$ a matriz de covariâncias suposta comum aos dois grupos. A correspondente regra de decisão é dada por $\delta^F(\mathbf{x}) = 1$, se $z(\mathbf{x}) \geq 0$ e $\delta^F(\mathbf{x}) = 2$, se $z(\mathbf{x}) < 0$. Fisher escolheu como ponto de separação, $-\alpha_0$, o ponto médio entre as projecções das médias dos grupos, isto é, $-\alpha_0 = \frac{1}{2}\boldsymbol{\alpha}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. De um modo geral esta constante dependerá das probabilidades *a priori* de cada grupo e dos custos relativos dos dois tipos de má classificação. Apesar de nenhum pressuposto distribucional ter sido feito, as distribuições envolvidas estão sumariadas em termos dos seus primeiros e segundos momentos. Se as distribuições subjacentes puderem ser descritas completamente por esses momentos espera-se que a regra linear de Fisher seja a óptima (com a restrição adicional da matriz de covariâncias ser comum). Distribuições com esta propriedade são designadas por elípticas, sendo a distribuição Normal multivariada um caso particular.

Para mais detalhes sobre estimação de regras discriminantes consultar, por exemplo, Lachenbruch (1975), Hand (1981, 1997), McLachlan (1992), Huberty (1994), Pires (1995), ou ainda Anderson (1984, Cap. 6) ou Seber (1984, Cap. 6), entre outros livros de Análise Multivariada que contemplam este tópico.

Em relação a propostas robustas para a Análise Discriminante refiram-se os trabalhos de Ahmed e Lachenbruch (1977), Campbell (1978, 1982), Randles *et al.* (1978), Posse (1992), Todorov *et al.* (1994), Pires (1995) e mais recentemente os trabalhos de He e Fung (2000) e de Croux e Dehon (2001).

Uma vez que o objectivo de uma regra discriminante é a classificação correcta do maior número de futuras entidades é necessário averiguar a sua qualidade. A análise de desempenho de uma regra discriminante pode ser efectuada tendo em conta vários critérios. Um deles diz respeito à forma efectiva como uma entidade é afecta ao seu verdadeiro grupo. Como exemplos de medidas avaliadoras deste tipo tem-se o *Brier score* (Brier, 1950; Hand, 1997) e as taxas de erro de má classificação (óptima, actual e actual esperada). Para uma revisão geral acerca da definição e estimação destas taxas de erro, incluindo os métodos de validação cruzada e *bootstrap*, ver, por exemplo, Lachenbruch e Mickey (1968), Efron (1979, 1983), Chatterjee (1983), McLachlan (1980, 1987), McLachlan (1992, Cap. 10), Pires (1995, Cap. 3) e Hand (1997, Cap. 7).

Após uma breve formalização e referências gerais acerca de AD retorne-se ao principal objectivo desta secção, a selecção de variáveis. Como já referido, em muitos estudos práticos existe a tentação de incluir toda a informação disponível, não só pelo desconhecimento de quais as variáveis adequadas a incluir no modelo, como também por se pensar que um maior número de variáveis se traduz num melhor ajustamento. Porém esta opção conduz em muitas situações, não a um melhor modelo mas sim à introdução de confusão ou ruído. Além disso, pode também provocar o agravamento quer de problemas matemáticos (indeterminação, não convergência), quer de problemas estatísticos, nomeadamente sobrestimação. No caso particular da Análise Discriminante, e apesar de em termos teóricos (populações conhecidas) se mostrar que o número de variáveis redundantes não influencia as probabilidades de classificação, observa-se que em casos práticos (amostras finitas) à medida que aumenta o número de variáveis redundantes para a classificação aumentam também as probabilidades de classificação incorrecta. Esta situação é ilustrada nos Exemplos 4.4 e 4.5. Com o Exemplo 4.5 também se mostra o comportamento das taxas de erro face à existência de uma observação atípica nos dados de trabalho.

Exemplo 4.4 Qualidade do ar (Sequeira, 1997). Este problema diz respeito à discriminação entre ar de boa qualidade (grupo G_1) e ar de má qualidade (grupo G_2) baseada em seis variáveis obtidas por imagem de satélite (valores das bandas, $\{b_i, i = 1, 2, \cdots, 6\}$).

A amostra de trabalho consiste em 19 observações dessas variáveis para cada grupo. Após a estimação da função discriminante linear de Fisher com base em vários subconjuntos das variáveis, calcularam-se as taxas de erro aparentes e as taxas de validação cruzada.

O gráfico da Figura 4.7 mostra o desempenho da regra usada *versus* dimensionalidade, (em termos do número de observações correctamente classificadas). Pela sua observação verifica-se, por um lado, a característica optimista das taxas de erro aparentes e por outro o fenómeno, já referido, de que um aumento no número de variáveis não se traduz necessariamente num aumento do número de classificações correctas.

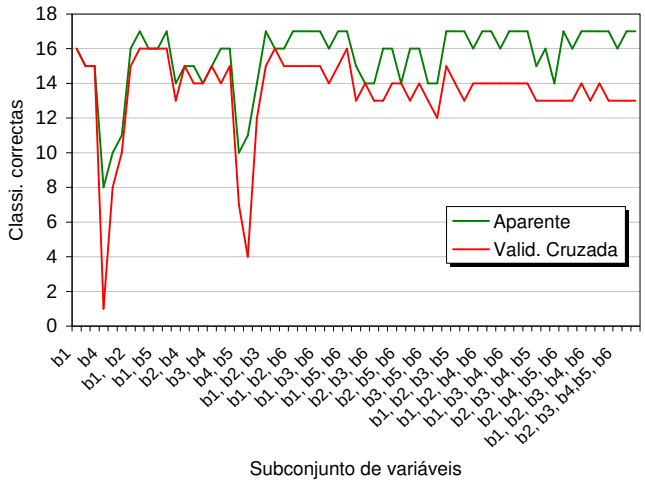


Figura 4.7: Classificações correctas *versus* dimensionalidade (dados da qualidade do ar).

■

Exemplo 4.5 Dimensionalidade, observações atípicas e classificação. Considere-se $\mathbf{X}|G_1 \sim N_{10}(\mathbf{0}, \mathbf{I})$ e $\mathbf{X}|G_2 \sim N_{10}(\boldsymbol{\mu}_2, \mathbf{I})$, com $\boldsymbol{\mu}_2 = (2, 0, \cdots, 0)^T$ e ainda $\pi_1 = \pi_2$ e $c(1|2) = c(2|1)$. Nestas condições a função discriminante linear de Fisher é óptima e as regiões de classificação são

$$R_{1opt} = \{\mathbf{x} : x_1 < 1\} \quad R_{2opt} = \{\mathbf{x} : x_1 > 1\},$$

obtendo-se as seguintes probabilidades de erro óptimas (ou taxas de erro óptimas), **independentemente do número de variáveis no modelo**,

$$e_{opt}(1) = \int_{R_{2opt}} f_1(\mathbf{x})d\mathbf{x} = e_{opt}(2) = \int_{R_{1opt}} f_2(\mathbf{x})d\mathbf{x} = 1 - \Phi(1) = 0.1587,$$

o que conduz à mesma taxa óptima total

$$e_{opt} = \pi_1 e_{opt}(1) + \pi_2 e_{opt}(2) = 0.1587.$$

Com o objectivo de avaliar a influência do número de variáveis na classificação geram-se aleatoriamente $n_1 = n_2 = 50$ observações de cada grupo e calcularam-se diversas taxas de erro associadas à função discriminante linear de Fisher estimada. Esta foi determinada substituindo em (4.20) os parâmetros $\boldsymbol{\mu}_i$ e $\boldsymbol{\Sigma}$ pelos estimadores clássicos, $\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{x}}_i$, média amostral do grupo i e $\hat{\boldsymbol{\Sigma}} = \mathbf{S}_C$ a matriz de covariâncias amostral combinada. Repetiu-se o mesmo exercício substituindo, na amostra de treino de G_1, uma observação pelo ponto atípico $\mathbf{x} = (10, 10)$. Os resultados, referentes às taxas de erro aparente, e_{ap}, validação cruzada, e_{vc}, e actual, e_{act}, para várias situações encontram-se na Tabela 4.18.

	Taxas de erro		
Situação	e_{ap}	e_{vc}	e_{act}
x_1	0.10	0.10	0.1599
x_1 e x_2	0.10	0.10	0.1602
$x_1, x_2, \cdots, x_{10}$	0.08	0.11	0.1842
x_1 e uma obs. atípica	0.11	0.12	0.1587
x_1, x_2 e uma obs. atípica	0.17	0.18	0.1897
$x_1, x_2, \cdots, x_{10}$ e uma obs. atípica	0.08	0.17	0.1947

Tabela 4.18: Comparação das várias taxas de erro.

Estes resultados mostram que o aumento do número de variáveis redundantes piora a classificação. Também se confirma que há um aumento das taxas de erro quando surge uma observação atípica. Observa-se mais uma vez a característica optimista da taxa aparente.

■

A selecção de variáveis é um aspecto importante não só pelo papel que tem no desempenho de uma regra discriminante, como já exemplificado, mas também porque o custo para recolher e processar um largo conjunto de variáveis pode ser elevado. Existem duas estratégicas básicas para proceder a essa selecção: (i) seleccionar um subconjunto a partir do conjunto de variáveis originais; e (ii) definir novas variáveis como (transformações) combinações das variáveis originais. A primeira tem a vantagem de, após a selecção, apenas necessitar das variáveis escolhidas para a classificação de futuras entidades. Quanto à segunda, para além de requerer todas as variáveis para futuras classificações tem ainda a desvantagem da interpretação. No entanto, a estratégia a optar deve depender do objectivo final pretendido. Quando se procede à selecção de um subconjunto de variáveis a partir do inicial é necessário definir um critério de selecção e a escolha ideal deveria ser feita avaliando esse critério para cada subconjunto possível. Porém, mesmo para valores moderados de m, a avaliação de todos os 2^m modelos possíveis torna-se proibitiva, tornando-se necessário usar outras estratégias de procura de um subconjunto adequado de variáveis. Neste contexto têm sido propostos vários algoritmos de selecção, sendo os mais mediáticos os denominados procedimentos por passos (*stepwise*, na literatura anglo-saxónica) nas suas versões de inclusão progressiva, eliminação regressiva ou ambas. Uma vez que este é um problema de optimização combinatória, vários métodos de investigação operacional têm também sido explorados, é o caso, por exemplo, dos algoritmos *branch-and-bound*, *simulated annealing* e genéticos. Os critérios de selecção usados nesses procedimentos são baseados ou na determinação de taxas de erro ou no cálculo de medidas de separabilidade.

Para referências gerais acerca da selecção de variáveis em AD ver, por exemplo, os livros de Hand (1981, Cap. 6), Hand (1997, Cap. 9), Huberty (1994, Cap. XVI), e McLachlan (1992, Cap. 12). Uma revisão global acerca deste tópico é feita em McKay e Campbell (1982a,b). Propostas mais recentes são apresentadas por Chatterjee e Samanta (1997), Duarte Silva (2001) e Cadima *et al.* (2002).

Testes para informação não adicional

A partir daqui apenas será discutida a situação onde uma entidade é classificada num de $g = 2$ grupos considerando a função discriminante linear de Fisher, expressa por $z(\mathbf{x}) = \boldsymbol{\alpha}^T\mathbf{x} + \alpha_0$, onde

$$\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} \tag{4.21}$$

e $\alpha_0 = -\frac{1}{2}\boldsymbol{\alpha}^T(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \ln(k)$, onde k é uma constante suposta conhecida e que depende das probabilidades *a priori* e dos custos.

Seja $\mathcal{S} = \{s : s$ é um subconjunto não vazio de $\{1, 2, \cdots, m\}\}$ e, para cada s em $\mathcal{S}$, seja m_s o número de elementos de s. Dado $s = \{k_1, k_2, \cdots, k_{m_s}\}$, seja $\mathbf{X}^{(s)} = \left(X_{k_1}, X_{k_2}, \cdots, X_{k_{m_s}}\right)^T$ o correspondente subvector do vector completo das m variáveis $\mathbf{X}$. Denotando por $\bar{s} = \{k_{m_s+1}, k_{m_s+2}, \cdots, k_m\}$ o complementar do conjunto s, $\mathbf{X}^{(\bar{s})}$ constitui o vector das $m - m_s$ variáveis que não estão em $\mathbf{X}^{(s)}$ e, organizando os indicies dos elementos em $\mathbf{X}$, tem-se $\mathbf{X} = \left(\mathbf{X}^{(s)T}\mathbf{X}^{(\bar{s})T}\right)^T$. De forma a simplificar a notação represente-se esta partição por

$$\mathbf{X} = \left(\mathbf{X}^{(1)T}\mathbf{X}^{(2)T}\right)^T, \tag{4.22}$$

onde as m_s variáveis de s surgem em primeiro lugar e as restantes $m - m_s$ variáveis (de $\bar{s}$) estão em seguida. Os vectores $\boldsymbol{\mu}_i$, $i = 1, 2$, e a matriz $\boldsymbol{\Sigma}$ são reorganizados da forma correspondente:

$$\boldsymbol{\mu}_i = \begin{pmatrix} \boldsymbol{\mu}_{i1}^T & \boldsymbol{\mu}_{i2}^T \end{pmatrix}^T, \quad i = 1, 2,$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Do mesmo modo faça-se $\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_1^T & \boldsymbol{\alpha}_2^T \end{pmatrix}^T$ e $\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\delta}_1^T & \boldsymbol{\delta}_2^T \end{pmatrix}^T$.

Teste F

Na selecção de variáveis a hipótese básica a testar é que $\mathbf{X}^{(2)}$ não é importante para a discriminação. Em AD linear isto equivale a testar $H_0 : \boldsymbol{\alpha}_2 = \mathbf{0}$. A Proposição 4.3, a seguir, mostra que testar esta hipótese é equivalente a testar a igualdade das distâncias de Mahalanobis entre os dois grupos usando apenas o subvector $\mathbf{X}^{(1)}$ ou todas as variáveis, $H_0 : \Delta_{m_s}^2 = \Delta_m^2$. Uma estatística adequada para testar esta hipótese nula é dada por

$$F = \frac{n - m - 1}{m - m_s}\left(\frac{D_m^2 - D_{m_s}^2}{q + D_{m_s}^2}\right), \tag{4.23}$$

onde $q = n(n-2)/n_1n_2$, D_m^2, $D_{m_s}^2$ são as correspondentes distâncias de Mahalanobis amostrais e $n = n_1 + n_2$. Quando a hipótese nula é verdadeira, F tem distribuição F de Snedecor com $m - m_s$ e $n - m - 1$ graus de liberdade.

Proposição 4.3 *Em AD linear testar* $H_0 : \boldsymbol{\alpha}_2 = \left(\alpha_{k_{m_s+1}}, \alpha_{k_{m_s+2}}, \cdots, \alpha_{k_m}\right) = \mathbf{0}$ *é equivalente a testar* $H_0 : \Delta_{m_s}^2 = \Delta_m^2$.

Demonstração: Pretende-se mostrar que:

(i) $\boldsymbol{\alpha}_2 = \mathbf{0} \quad \Rightarrow \Delta^2_{m_s} = \Delta^2_m$.

(ii) $\Delta^2_{m_s} = \Delta^2_m \quad \Rightarrow \boldsymbol{\alpha}_2 = \mathbf{0}$

Dada a função discriminante linear de Fisher, $z(\mathbf{x}) = \boldsymbol{\alpha}^T\mathbf{x} + \alpha_0$, com $\boldsymbol{\alpha}$ dado por (4.21) e considerando a partição (4.22) pode escrever-se

$$\boldsymbol{\alpha}^T\mathbf{x} = \boldsymbol{\alpha}_1^T\mathbf{x}^{(1)} + \boldsymbol{\alpha}_2^T\mathbf{x}^{(2)}.$$

Como $\Delta^2_m = \boldsymbol{\delta}^T\boldsymbol{\alpha} = \boldsymbol{\delta}_1^T\boldsymbol{\alpha}_1 + \boldsymbol{\delta}_2^T\boldsymbol{\alpha}_2 = \Delta^2_{m_s} + \boldsymbol{\delta}_2^T\boldsymbol{\alpha}_2$, conclui-se que $\Delta^2_m - \Delta^2_{m_s} = \boldsymbol{\delta}_2^T\boldsymbol{\alpha}_2$, o que mostra (i).

Por outro lado

$$\begin{aligned}
\Delta^2_m &= \begin{pmatrix} \boldsymbol{\delta}_1^T & \boldsymbol{\delta}_2^T \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix} = \\
&= \begin{pmatrix} \boldsymbol{\delta}_1^T & \boldsymbol{\delta}_2^T \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{2.1}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{2.1}^{-1} \\ -\boldsymbol{\Sigma}_{2.1}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \boldsymbol{\Sigma}_{2.1}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix} = \\
&= \boldsymbol{\delta}_1^T\Sigma_{11}^{-1}\boldsymbol{\delta}_1 + \delta_{2.1}^T\boldsymbol{\Sigma}_{2.1}^{-1}\boldsymbol{\delta}_{2.1} = \Delta^2_{m_s} + \boldsymbol{\delta}_{2.1}^T\boldsymbol{\Sigma}_{2.1}^{-1}\boldsymbol{\delta}_{2.1},
\end{aligned} \tag{4.24}$$

onde

$$\boldsymbol{\delta}_{2.1} = \boldsymbol{\delta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1$$

e

$$\boldsymbol{\Sigma}_{2.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$$

corresponde à matriz de covariâncias das variáveis $\mathbf{X}^{(2)}$ condicionadas a $\mathbf{X}^{(1)}$ (Seber, 1984, pág. 52). A expressão (4.24) pode ser escrita como

$$\Delta^2_m - \Delta^2_{m_s} = \boldsymbol{\delta}_{2.1}^T\boldsymbol{\Sigma}_{2.1}^{-1}\boldsymbol{\delta}_{2.1}. \tag{4.25}$$

Do mesmo modo, a expressão para o vector $\boldsymbol{\alpha}$, quando se considera a partição em causa, é dada por

$$\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix} = \boldsymbol{\Sigma}_{2.1}^{-1}\boldsymbol{\delta}_{2.1}. \tag{4.26}$$

Por (4.25) $\Delta_m - \Delta_{m_s} = 0$ implica que $\boldsymbol{\delta}_{2.1} = \mathbf{0}$, já que a matriz $\boldsymbol{\Sigma}_{2.1}^{-1}$ é definida positiva. Logo por (4.26), $\boldsymbol{\alpha}_2 = \mathbf{0}$, o que mostra (ii).

Teste *bootstrap*

Em Pires (1995, Cap. 6) é proposto um teste *bootstrap* para avaliar a importância de informação adicional em AD linear, com hipótese nula, $H_0 : \boldsymbol{\alpha}_2 = \mathbf{0}$. Para cada amostra *bootstrap* calcula-se

a função discriminante linear e registam-se as estimativas dos coeficientes, $\hat{\boldsymbol{\alpha}}$, obtendo-se a matriz de observações

$$\mathbf{B}^* = \begin{bmatrix} \hat{\alpha}^{(11)} & \cdots & \hat{\alpha}^{(1m)} \\ \hat{\alpha}^{(21)} & \cdots & \hat{\alpha}^{(2m)} \\ \vdots & & \vdots \\ \hat{\alpha}^{(B1)} & \cdots & \hat{\alpha}^{(Bm)} \end{bmatrix},$$

onde m corresponde ao número de variáveis originais e B ao número de réplicas *bootstrap*. Para testar a hipótese H_0 considera-se a submatriz $\mathbf{B}_2^*$ constituída pelos coeficientes associados às variáveis em causa. Em seguida calcula-se o vector médio, $\hat{\boldsymbol{\mu}}_{*2}$, e a matriz de covariâncias, $\hat{\boldsymbol{\Sigma}}_{*2}$, correspondentes. Para cada linha de $\mathbf{B}_2^*$, $\hat{\boldsymbol{\alpha}}_2^{(b)}$, $b = 1, 2, \cdots, B$, determina-se a distância ao quadrado, $d_{b*}^2 = \left(\hat{\boldsymbol{\alpha}}_2^{(b)} - \hat{\boldsymbol{\mu}}_{*2}\right)^T \hat{\boldsymbol{\Sigma}}_{*2}^{-1} \left(\hat{\boldsymbol{\alpha}}_2^{(b)} - \hat{\boldsymbol{\mu}}_{*2}\right)$ e para o ponto $\mathbf{0}$ tem-se $d_{0*}^2 = \boldsymbol{\mu}_{*2}{}^T \boldsymbol{\Sigma}_{*2}^{-1} \boldsymbol{\mu}_{*2}$.

O *p-value* do teste será então dado pelo quociente, B^*/B, onde $B^* = \#\left\{b : d_{b*}^2 > d_{0*}^2\right\}$. Observe-se ainda que, se $\hat{\boldsymbol{\alpha}}_2$ for constituído apenas por uma componente a variância não tem influência e $B^* = \#\left\{b : |\hat{\boldsymbol{\alpha}}_2^{(b)} - \hat{\boldsymbol{\mu}}_{*2}| = |\hat{\boldsymbol{\mu}}_{*2}|\right\}$.

O que se propõe nesta secção é uma extensão "robusta" do teste *bootstrap* acabado de descrever. A ideia é aplicar o método bootstrap robusto, IF, para criar a matriz $\mathbf{B}^*$ e a selecção de um determinado subconjunto dependerá do *p–value* do teste associado às variáveis presentes na hipótese nula. Descreve-se em seguida a referida proposta.

Seja $\{\mathbf{X}_{ij}, i = 1, 2; j = 1, 2, \cdots, n_i\}$ uma amostra aleatória m-dimensional proveniente das populações multivariadas F_i, $i = 1, g$ (com $g = 2$ grupos). Seja $\mathbf{T}_n^{nr}$ o estimador habitual da direcção $\boldsymbol{\alpha}$,

$$\mathbf{T}_n^{nr} : \hat{\boldsymbol{\alpha}} = \mathbf{S}_C^{-1} \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) \tag{4.27}$$

e $\mathbf{T}^{nr}$ o funcional equivalente.

As funções de influência parciais de $\mathbf{T}^{nr}$ no ponto $\mathbf{x}$ são dadas por

$$IF_i\left(\mathbf{x}; \mathbf{T}^{nr}, F_1, F_2\right) = \upsilon_i \boldsymbol{\alpha} + \left[(-1)^{i-1} - \upsilon_i \boldsymbol{\alpha}^T \mathbf{z}_i\right] \boldsymbol{\Sigma}^{-1} \mathbf{z}_i, \tag{4.28}$$

onde $\mathbf{z}_i = \mathbf{x} - \boldsymbol{\mu}_i$, $\upsilon_i = n_i/(n_1 + n_2)$, n_i é a dimensão da amostra do grupo i, e cada F_i tem associada o vector médio $\boldsymbol{\mu}_i$ e a matriz de covariâncias $\boldsymbol{\Sigma}$ (Pires, 1995, Cap. 3).

A função de influência padronizada é então dada por

$$SIF(\mathbf{x}; \mathbf{T}^{nr}, F_1, F_2) = \left[IF_i\left(\mathbf{x}; \mathbf{T}^{nr}, F_1, F_2\right)^T \mathbf{V}_i(\hat{\boldsymbol{\alpha}})^{-1} IF_i\left(\mathbf{x}; \mathbf{T}^{nr}, F_1, F_2\right)\right]^{1/2} \tag{4.29}$$

com

$$\mathbf{V}_i\left(\hat{\boldsymbol{\alpha}}\right) = \boldsymbol{\alpha}\boldsymbol{\alpha}^T + \left(\frac{1}{\upsilon_1 \upsilon_2} + \Delta_m^2\right) \boldsymbol{\Sigma}^{-1}.$$

Repare-se que, para cada i, a função (4.29) depende do vector de parâmetros, $\boldsymbol{\Omega} = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$. Considere-se um estimador robusto de $\boldsymbol{\Omega}$, $\hat{\boldsymbol{\Omega}}^r = \left(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}\right)$, $i = 1, 2$, onde $\hat{\boldsymbol{\mu}}_i$ e $\hat{\boldsymbol{\Sigma}}$ são estimadores robustos da localização e dispersão da distribuição F_i.

Então, para cada observação $\mathbf{x}_{ij}$, a função de influência empírica padronizada, $RESIF_{ij}$, é obtida através da expressão (4.29) mas substituindo o vector de parâmetros desconhecidos, $\mathbf{\Omega}$, pelas estimativas definidas por $\hat{\mathbf{\Omega}}^r$.

Estudo de Monte Carlo

Com o propósito de avaliar o critério de selecção proposto, procedeu-se a um estudo de simulação semelhante aos anteriores, contemplando vários métodos e também várias distribuições. O objectivo é verificar qual a sensibilidade destes métodos quando as distribuições das populações se afastam da normalidade. Não obstante a importância que tem a dimensão do espaço, neste estudo apenas foram consideradas duas situações, $m = 2$ e $m = 3$. A razão desta escolha deve-se ao objectivo principal deste estudo, pois o que está aqui em causa é apenas o critério de selecção de um determinado subconjunto e não o procedimento de selecção. Obviamente com este número de variáveis o método de selecção é baseado na avaliação do critério em todos os subconjuntos. A conjugação do critério proposto com um bom algoritmo de pesquisa é um estudo que fica desde já em aberto.

Como o delineamento deste estudo é em alguns aspectos condicionado pelo número de variáveis é necessário considerar duas situações distintas: I) $m = 2$; II) $m = 3$. As distribuições consideradas nos dois estudos são praticamente as mesmas do estudo de simulação referente à eficiência de estimadores robustos multivariados em amostras finitas, (Secção 4.3.3, pág. 113), pelo que não se acha necessário estar a descrevê-las novamente. Refira-se apenas que na situação II) foram utilizadas todas as distribuições aí referidas enquanto que na primeira situação, I), não se usou a distribuição T3.

Situação I) $m = 2$

Neste caso a experiência foi delineada de forma a que no caso de controlo apenas a variável X_1 do vector aleatório $\mathbf{X} = (X_1, X_2)$ seja importante para a discriminação. A hipótese nula a testar é $H_0 : \boldsymbol{\alpha}_2 = \mathbf{0}$, o que nesta situação concreta se reduz a $H_0 : \alpha_2 = 0$.

Os parâmetros das distribuições para cada grupo, $\boldsymbol{\mu}_i$, $i = 1, 2$ e $\mathbf{\Sigma}_i$, associados a este caso foram os seguintes:

Localização	Dispersão
$\boldsymbol{\mu}_1 = \mathbf{0}$	$\mathbf{\Sigma}_1 = \mathbf{I}_2$
$\boldsymbol{\mu}_2 = (2\cos\beta, 2\sin\beta)$	$\mathbf{\Sigma}_2 = \mathbf{I}_2$
onde $\beta = k\pi/12$, $k = 0, 1, \cdots, 6$	

A igualdade dos parâmetros de dispersão remete para a discriminante linear enquanto os parâmetros de localização foram escolhidos de tal modo que a distância entre os dois grupos é constante e igual a 2 mas a sua variação permite avaliar simultaneamente o nível de significância do teste ($k = 0$) e a sua função potência ($k \neq 0$).

Os estimadores de $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ utilizados foram o já referido $\mathbf{T}_n^{nr}$, expresso por (4.27), e um estimador robusto, $\mathbf{T}_n^{r}$, obtido por substituição de $\boldsymbol{\Sigma}$ e $\boldsymbol{\mu}_i$ pelos respectivos estimadores robustos RMCD50.

O teste da hipótese nula, $H_0 : \alpha_2 = 0$, foi realizado utilizando os seguintes métodos:

- **Teste F** (FT): método clássico utilizando a estatística dada por (4.23);
- ***Bootstrap* clássico ou uniforme** (UB): teste *bootstrap* descrito atrás;
- ***Bootstrap* baseado na função de influência** (IFB): teste semelhante ao imediatamente anterior mas aplicando o *bootstrap* IF, com $RESIF_i$, $i = 1, 2$, obtida através da substituição de estimadores robustos RMCD50 de localização e dispersão de cada grupo. As probabilidades de reamostragem de um ponto $\mathbf{x}_{ij}$, são determinadas com base nos pesos

$$w_{ij} = I_{[0,c]}\left(|RESIF_{ij}|\right) + \exp\left(\frac{(|RESIF_{ij}| - c)^2}{2c^2}\right) \times I_{]c,+\infty]}\left(|RESIF_{ij}|\right), \tag{4.30}$$

  para $j = 1, 2, \ldots, n_i$, $i = 1, 2$, (isto é, para parâmetros da função ψ escolheu-se $\gamma = \infty$ tomou-se $d = c$). O valor da constante de afinação, $c = 13.2$, foi escolhido após simulação da distribuição da SIF, sob uma distribuição Normal bivariada. As probabilidades de reamostragem são iguais, independentemente do estimador, clássico ou robusto, que se pretende reamostrar.
- ***Bootstrap Winsorized*** (WB): construção da matriz $\mathbf{B}^*$ obtida por reamostragem *bootstrap* das amostras *Winsorized* de cada grupo (obtidas através da aplicação de um procedimento análogo ao descrito no estudo sobre ρ, pág. 100). A proporção de *Winsorization* foi fixada em $\beta = 0.1$.

Quanto às dimensões das amostras foi considerada apenas a situação $n_1 = n_2 = 50$. As probabilidades *a priori*, π_i, $i = 1, 2$, são sempre iguais.

Para cada situação (num total de 96) foram geradas amostras independentes. De referir que, por simetria, os valores apresentados para $\beta = \pi/3, 5\pi/12, \pi/2$ foram obtidos ao testar a hipótese $H_0 : \alpha_1 = 0$ em $\beta = \pi/6, \pi/12, 0$. Do mesmo modo que no estudo de Monte Carlo relativo a ρ, fixou-se em 1000 o número de réplicas *bootstrap*, enquanto o número de simulações realizadas foi de 1000 ou 100 consoante o estimador utilizado na fase de reamostragem (não robusto e robusto, respectivamente).

Resultados

Como resultado síntese das 1000/100 simulações obteve-se a frequência relativa do acontecimento "*p-value*$< \gamma_*$", com $\gamma_* = 0.01, 0.05, 0.10$, o que corresponde à proporção de rejeições da hipótese nula, $H_0 : \alpha_2 = 0$, para um nível de significância de γ_*. Os resultados são apresentados nas Figuras 4.8, 4.9 e 4.10, segundo o método de estimação (clássico ou robusto).

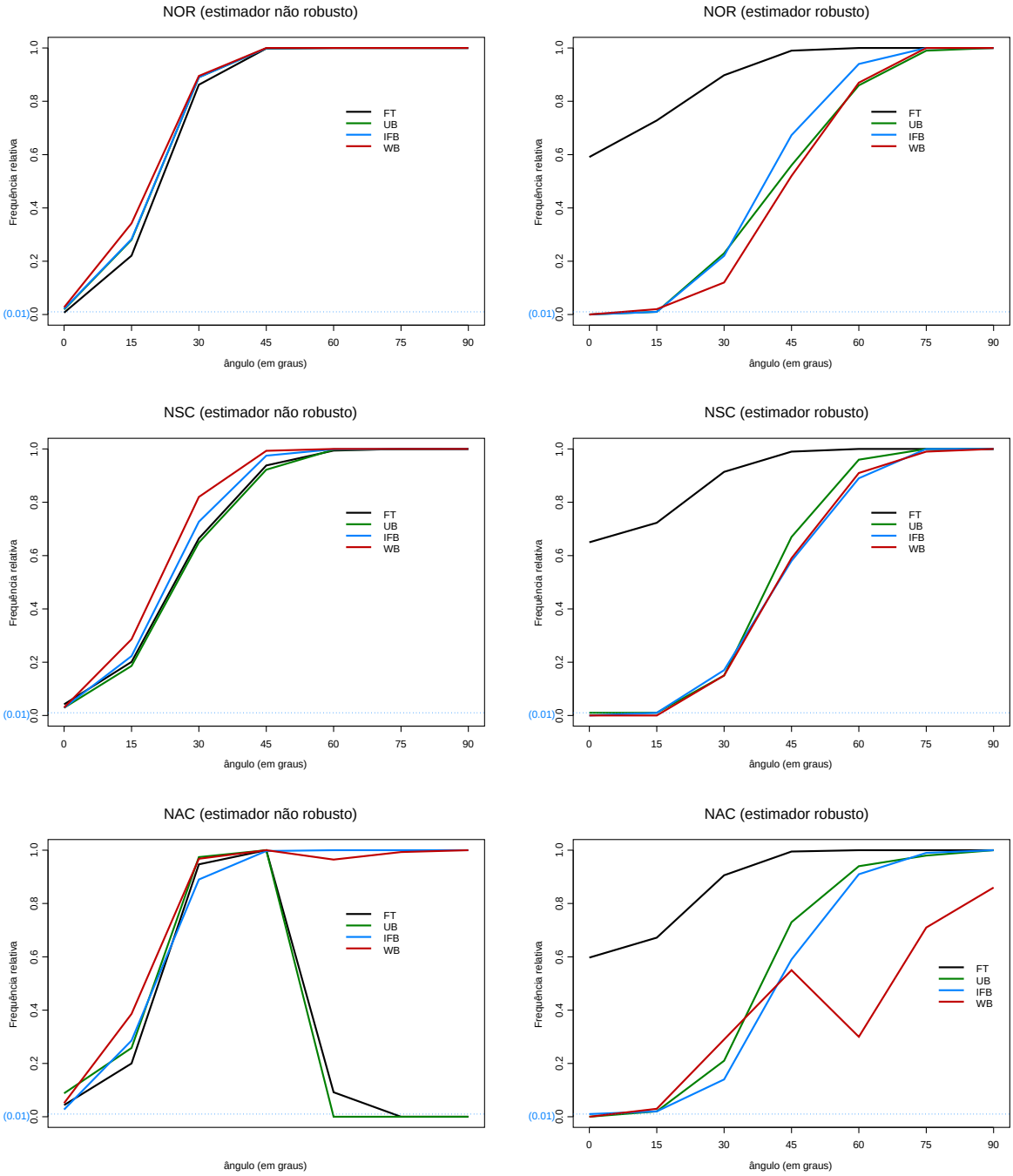


Figura 4.8: Frequência relativa do acontecimento "*p-value*<0.01" em 1000/100 simulações para a hipótese $H_0 : \alpha_2 = 0$, variando o vector médio da segunda população de acordo com $(2\cos\beta, 2\sin\beta)$ para $\beta = 0, 15, 30, \cdots, 90$ (em graus). A coluna da esquerda representa os valores obtidos para as distribuições NOR, NSC e NAC quando se utiliza o estimador não robusto de $\boldsymbol{\alpha}$. A coluna da direita representa os resultados para idênticas distribuições mas referentes ao estimador robusto.

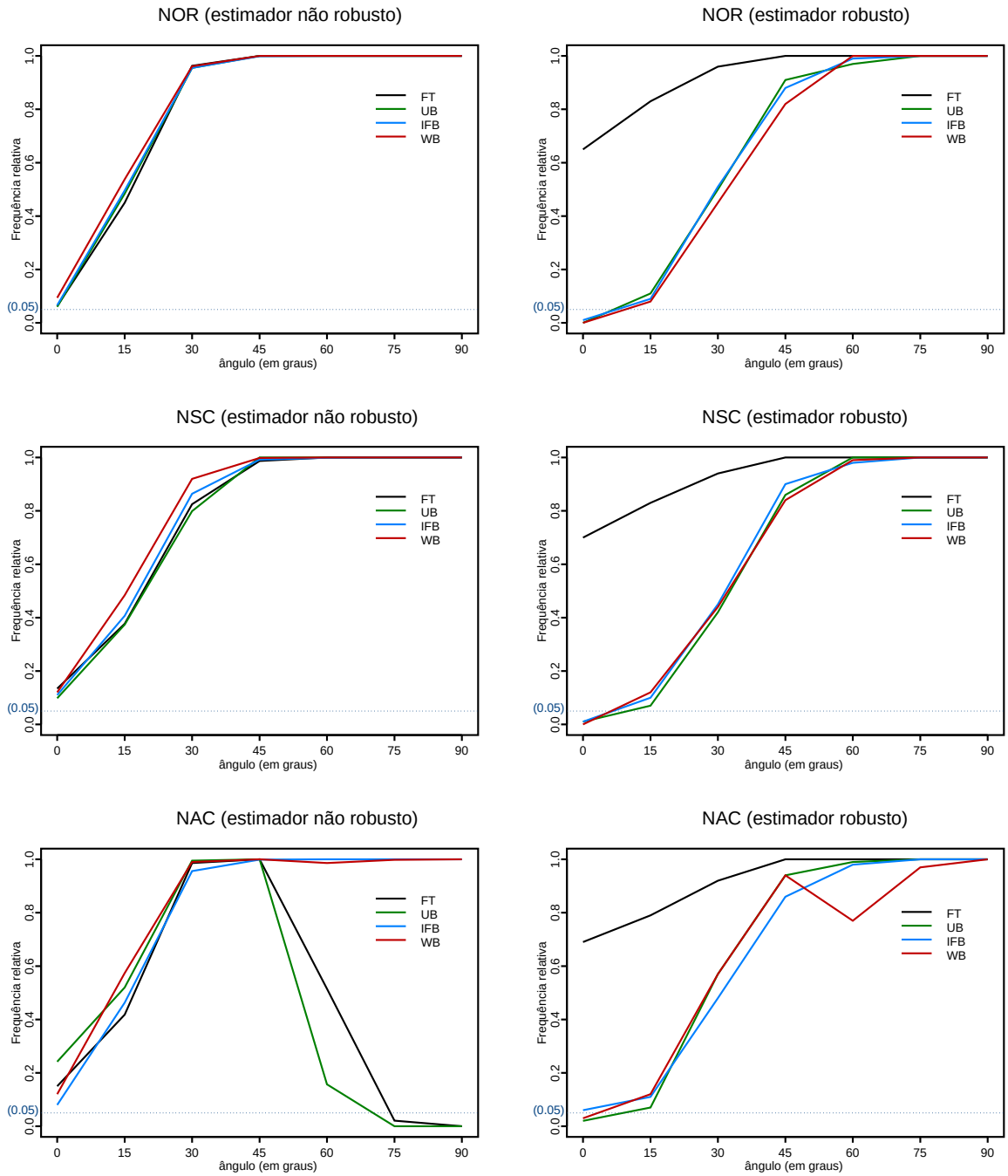


Figura 4.9: Frequência relativa do acontecimento "*p-value*<0.05" em 1000/100 simulações para a hipótese $H_0 : \alpha_2 = 0$, variando o vector médio da segunda população de acordo com $(2\cos\beta, 2\sin\beta)$ para $\beta = 0, 15, 30, \cdots, 90$ (em graus). A coluna da esquerda representa os valores obtidos para as distribuições NOR, NSC e NAC quando se utiliza o estimador não robusto de $\boldsymbol{\alpha}$. A coluna da direita representa os resultados para idênticas distribuições mas referentes ao estimador robusto.

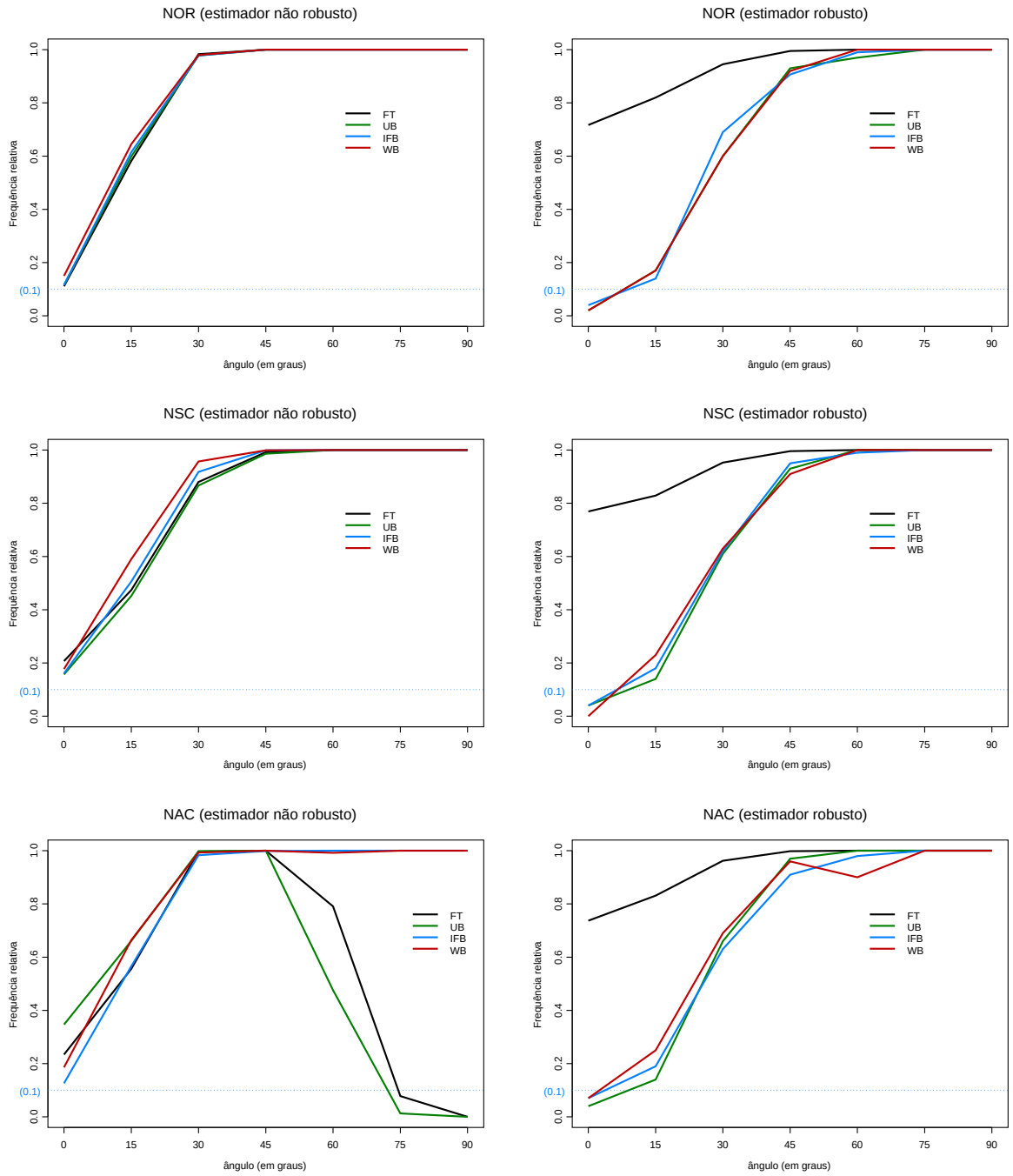


Figura 4.10: Frequência relativa do acontecimento "*p-value*<0.10" em 1000/100 simulações para a hipótese $H_0 : \alpha_2 = 0$, variando o vector médio da segunda população de acordo com $(2\cos\beta, 2\sin\beta)$ para $\beta = 0, 15, 30, \cdots, 90$ (em graus). A coluna da esquerda representa os valores obtidos para as distribuições NOR, NSC e NAC quando se utiliza o estimador não robusto de $\boldsymbol{\alpha}$. A coluna da direita representa os resultados para idênticas distribuições mas referentes ao estimador robusto.

Para dados normais e estimador clássico, o teste F teve o melhor desempenho como era de esperar, embora todos os outros tenham registado resultados muito próximos. Quando o estimador em consideração é robusto o teste F não se mostrou adequado, mesmo para dados normais, mas em contrapartida registaram-se bons desempenhos dos testes baseados em reamostragem.

No que diz respeito à distribuição NSC verifica-se que, para o estimador não robusto, todos os métodos apresentam bom comportamento, embora naturalmente inferior ao dos dados normais. Mesmo para o teste F isso acontece, porventura devido à pouca sensibilidade da distância de Mahalanobis a esta distribuição. Relativamente ao estimador robusto observam-se mais uma vez os resultados disparatados fornecidos pelo teste F e o comportamento razoável dos restantes métodos.

Os resultados obtidos, para a distribuição NAC com o estimador clássico, não constituíram surpresa já que a contaminação assimétrica é muito desfavorável, fazendo com que o desempenho dos métodos puramente clássicos, FT e UB sejam afectados de forma drástica. A reamostragem do estimador robusto conduziu a bons desempenhos, tendo sido o UB e o IFB os melhores de entre os três. Deve notar-se que na utilização do estimador robusto com o método UB não se observaram problemas relacionados com o ponto de rotura, provavelmente devido ao uso de estimadores de substituição com ponto de rotura de 0.5.

Quando se consideram todas as distribuições em simultâneo verifica-se que com o estimador não robusto os únicos métodos aceitáveis são o IFB e o WB, enquanto que com o estimador robusto são o IFB e o UB. Por outro lado, quando se comparam estes métodos entre si, verifica-se que os que usam na fase de reamostragem o estimador não robusto têm muito melhor potência do que os que usam o estimador robusto. Conclui-se assim que o mais razoável é a escolha entre IFB e WB com um estimador não robusto na fase de reamostragem, com uma ligeira vantagem do método IFB, por este apresentar níveis de significância estimados mais próximos dos pré-fixados.

Situação II) $m = 3$

O delineamento deste tipo de estudo de simulação torna-se mais complexo quando o número de variáveis é maior que dois devido ao elevado número de opções que se podem tomar para controlo das hipóteses de interesse. Assim sendo, optou-se por considerar, para parâmetros da distribuição (NOR, NSC, NAC ou T3) de cada grupo, quatro casos simples:

Localização	Dispersão
$\boldsymbol{\mu}_1 = \mathbf{0}$	$\boldsymbol{\Sigma}_1 = \mathbf{I}_2$
a) $\boldsymbol{\mu}_2 = (2, 0, 0)$	$\boldsymbol{\Sigma}_2 = \mathbf{I}_2$
b) $\boldsymbol{\mu}_2 = (\sqrt{2}, \sqrt{2}, 0)$	
c) $\boldsymbol{\mu}_2 = (2/\sqrt{3}, 2/\sqrt{3}, 2/\sqrt{3})$	

Tal como na situação I) optou-se pela igualdade dos parâmetros de dispersão e por manter a distância entre os dois grupos constante e igual a 2. Com o caso a) pretende-se ilustrar um modelo

onde a hipótese $H_0 : \alpha_2 = \alpha_3 = 0$ é verdadeira; no caso b) apenas a variável X_3 não tem interesse para a discriminação, ou seja, $H_0 : \alpha_3 = 0$; e, por último, na situação c) todas as variáveis são importantes para a discriminação.

Com o propósito de melhorar a eficiência da função que determina os pesos de reamostragem no *bootstrap* IF, utilizaram-se como estimadores robustos de $\mathbf{\Omega} = (\boldsymbol{\mu}_i, \mathbf{\Sigma})$ os estimadores de ponto de rotura perto de 0.5 que melhor desempenho tiveram no estudo da Secção 4.3.3, ou seja, estimadores reponderados e o método A de He e Fung (2000). Numa tentativa de melhorar a eficiência dos estimadores S com o método A procedeu-se à alteração referida na Observação 4.5. Tendo em conta a Observação 4.4 escolheu-se o estimador RMCD50A em vez do RMVE50A. Estas opções são denotadas por IFB-S e IFB-MCD, respectivamente.

A regra discriminante linear foi estimada utilizando os mesmos estimadores da situação I), bem como idênticas dimensões amostrais e probabilidades *a priori*. Relativamente aos testes de hipóteses foram considerados o teste F e os testes *bootstrap* baseados em UB, IFB-MCD e IFB-S. O método WB foi abandonado não só pela sua falta de competitividade com o método IFB mas também devido a limitações de tempo para as simulações.

Para cada combinação (distribuição, vector de localização, teste, estimador), num total de 72, foram geradas amostras independentes. Para o estimador clássico da regra discriminante fizeram-se $nsimul = 1000$ réplicas para cada uma dessas combinações e no caso dos testes *bootstrap*, para cada uma dessas réplicas fixou-se em 1000 o número de reamostras *bootstrap*. Quando se utiliza um estimador robusto foram apenas realizadas $nsimul = 100$ réplicas, mantendo no entanto 1000 reamostras *bootstrap*.

Uma vez que o número de variáveis é pequeno os testes foram aplicados a todas as hipóteses nulas possíveis. Na Tabela 4.19 apresentam-se os 6 subconjuntos do conjunto $\{1, 2, 3\}$ para os quais os testes fazem sentido, as hipóteses associadas e a notação utilizada para os *p-values*. Observe-se que o subconjunto com todas as variáveis não pode ser testado directamente.

Tabela 4.19: Associação dos *p-values* a cada uma das hipóteses consideradas.

Subconjunto	Teste *bootstrap*	Teste F	*p-value*
$\{1\}$	$H_0 : \alpha_2 = \alpha_3 = 0$	$H_0 : \Delta_{\{1\}} = \Delta_{\{123\}}$	p_1
$\{2\}$	$H_0 : \alpha_1 = \alpha_3 = 0$	$H_0 : \Delta_{\{2\}} = \Delta_{\{123\}}$	p_2
$\{3\}$	$H_0 : \alpha_1 = \alpha_2 = 0$	$H_0 : \Delta_{\{3\}} = \Delta_{\{123\}}$	p_3
$\{1, 2\}$	$H_0 : \alpha_3 = 0$	$H_0 : \Delta_{\{12\}} = \Delta_{\{123\}}$	p_4
$\{1, 3\}$	$H_0 : \alpha_2 = 0$	$H_0 : \Delta_{\{13\}} = \Delta_{\{123\}}$	p_5
$\{2, 3\}$	$H_0 : \alpha_1 = 0$	$H_0 : \Delta_{\{23\}} = \Delta_{\{123\}}$	p_6

Mais uma vez os testes descritos foram implementados em linguagem S e são aplicáveis a qualquer hipótese com qualquer número de variáveis, estando também as macros preparadas para gerar to-

dos os subconjuntos de um conjunto. A determinação de todos os subconjuntos de um conjunto é feita usando a conversão de um inteiro em binário e enumerando esses subconjuntos através da ordem lexicográfica crescente. Cada subconjunto corresponde a uma lista de zeros e uns que é posteriormente convertida para os inteiros apropriados. Para mais detalhes acerca da implementação computacional ver a descrição das macros no Apêndice C.

Tendo em conta as estimativas dos *p-values* (p_i) definidos na Tabela 4.19, considerou-se, para cada valor de $\gamma_* = 0.01, 0.05, 0.1$, o esquema descrito na Tabela 4.20 para a selecção de cada subconjunto[2] de variáveis.

Tabela 4.20: Descrição do esquema de selecção de cada subconjunto para um dado valor de γ_*.

Se:	Subconjunto seleccionado
$\hat{p}_1 > \gamma_*$ e $\hat{p}_2 < \gamma_*$ e $\hat{p}_3 < \gamma_*$ e $\hat{p}_6 < \gamma_*$	$\{1\}$
$\hat{p}_1 < \gamma_*$ e $\hat{p}_2 > \gamma_*$ e $\hat{p}_3 < \gamma_*$ e $\hat{p}_5 < \gamma_*$	$\{2\}$
$\hat{p}_1 < \gamma_*$ e $\hat{p}_2 < \gamma_*$ e $\hat{p}_3 > \gamma_*$ e $\hat{p}_4 < \gamma_*$	$\{3\}$
$\hat{p}_1 > \gamma_*$ e $\hat{p}_2 > \gamma_*$ e $\hat{p}_3 < \gamma_*$ e $\hat{p}_4 > \gamma_*$	$\{1, 2\}$
$\hat{p}_1 > \gamma_*$ e $\hat{p}_2 < \gamma_*$ e $\hat{p}_3 > \gamma_*$ e $\hat{p}_5 > \gamma_*$	$\{1, 3\}$
$\hat{p}_1 < \gamma_*$ e $\hat{p}_2 > \gamma_*$ e $\hat{p}_3 > \gamma_*$ e $\hat{p}_6 > \gamma_*$	$\{2, 3\}$
$\hat{p}_4 < \gamma_*$ e $\hat{p}_5 < \gamma_*$ e $\hat{p}_6 < \gamma_*$	$\{1, 2, 3\}$
qualquer outra situação	outro (de entre os referidos)

Resultados

Como resultado das 1000/100 simulações obtiveram-se, para cada distribuição, estimador (clássico/robusto), tipo de teste e valor de γ_*:

- A percentagem de inclusão de cada subconjunto, isto é, o número de vezes que cada conjunto é seleccionado a dividir pelo número de simulações;
- A frequência relativa do acontecimento "*p-value*$<\gamma_*$" nas *nsimul* simulações para cada hipótese H_0 verdadeira. Se um teste é adequado esta frequência relativa deve ser próxima do nível de significância nominal, γ_*.

As Figuras 4.11 a 4.16 representam os resultados relativos à percentagem de inclusão de cada subconjunto para cada situação simulada.

Pela observação da Figura 4.11, é fácil verificar que, como era esperado, a percentagem de inclusão da variável X_1 é a maior, e que o teste F tem o melhor desempenho quando se considera a distribuição Normal para modelo gerador dos dados. No entanto, para esta distribuição, também se

[2] Uma vez que as sete condições apresentadas não são exaustivas pode ocorrer outra situação para os valores de $\hat{p}_i$ e neste caso considera-se que será escolhido um subconjunto genérico designado por "outro".

observa um bom desempenho dos restantes métodos. Para as distribuições NSC e T3 o comportamento dos testes *bootstrap* é sensivelmente melhor que o do teste F. É, todavia, na distribuição NAC que se observa o comportamento óptimo do IFB em contraste com os resultados desastrosos quer do teste F (FT), quer do *bootstrap* usual (UB).

Relativamente à Figura 4.12, verifica-se que para a distribuição Normal todos os testes têm um desempenho adequado, seleccionando correctamente as variáveis X_1 e X_2, como sendo as mais importantes para a discriminação, um número elevado de vezes. É nas distribuições contaminadas que se observa o melhor desempenho do IFB face aos restantes métodos utilizados. Para a distribuição NSC e T3 observa-se para cada valor fixo γ_*, quer para o teste F quer para o teste UB, alguma confusão entre a selecção das variáveis X_1 e X_2 conjuntamente e separadamente. Para a distribuição NAC é visível a confusão destes testes entre a selecção correcta dessas duas variáveis e a selecção incorrecta de apenas X_2, selecção incorrecta esta que aumenta com o valor de γ_*. No que diz respeito ainda à distribuição NAC, também se verifica um comportamento distinto quando comparamos estes resultados com os apresentados na Figura 4.11. Este facto justifica-se pelo uso do ponto de contaminação $(22, 2, 0)$, o qual nesta situação não corresponde à troca dos eixos importantes para a discriminação. Para finalizar a análise refira-se que em nenhuma situação foi seleccionada apenas a variável X_3, o que está de acordo com o delineamento.

Os comentários efectuados para a Figura 4.12 aplicam-se também à Figura 4.13. Refira-se apenas que, neste caso e de acordo com o delineamento da simulação, são as três variáveis que são importantes para a discriminação.

Em termos gerais, e para o estimador clássico, verifica-se que diminuir o nível de significância individual implica uma tendência para incluir mais as variáveis correctas. Por exemplo, considere-se que se testa a hipótese $H_0 : \alpha_3 = 0$ e observe-se a Figura 4.12: neste caso H_0 é verdadeira e quando diminui o nível de significância, γ_*, diminui o número de rejeições e portanto as variáveis (X_1, X_2) são mais vezes seleccionadas.

As Figuras 4.14, 4.15 e 4.16 representam os resultados obtidos para o estimador robusto. Para a situação onde se considera $\boldsymbol{\mu}_2 = (2, 0, 0)$, Figura 4.14, verifica-se que são os testes *bootstrap* os que melhor desempenho tiveram, independentemente da distribuição utilizada. Quando se considera, $\boldsymbol{\mu}_2 = (\sqrt{2}, \sqrt{2}, 0)$, Figura 4.15, verifica-se que o teste F mantém o tipo de comportamento à medida que varia o γ_*, independentemente da distribuição considerada. Relativamente aos testes *bootstrap*, constata-se que à medida que diminui o valor de γ_* diminui também a percentagem de inclusão das variáveis correctas, existindo alguma confusão entre a selecção apenas da variável X_1 ou apenas da variável X_2. A Figura 4.16 revela um bom desempenho do teste F mas um pior desempenho dos testes *bootstrap*. Este estranho comportamento dos testes *bootstrap* perante o estimador robusto deve-se provavelmente a problemas de eficiência e deverá ser analisado mais em pormenor em trabalho futuro.

Nas Tabelas 4.21 a 4.24 encontram-se as frequências relativas dos acontecimentos "*p-value*$<\gamma_*$" nas *nsimul* simulações, para cada hipótese H_0. Uma vez que o objectivo é controlar o nível de

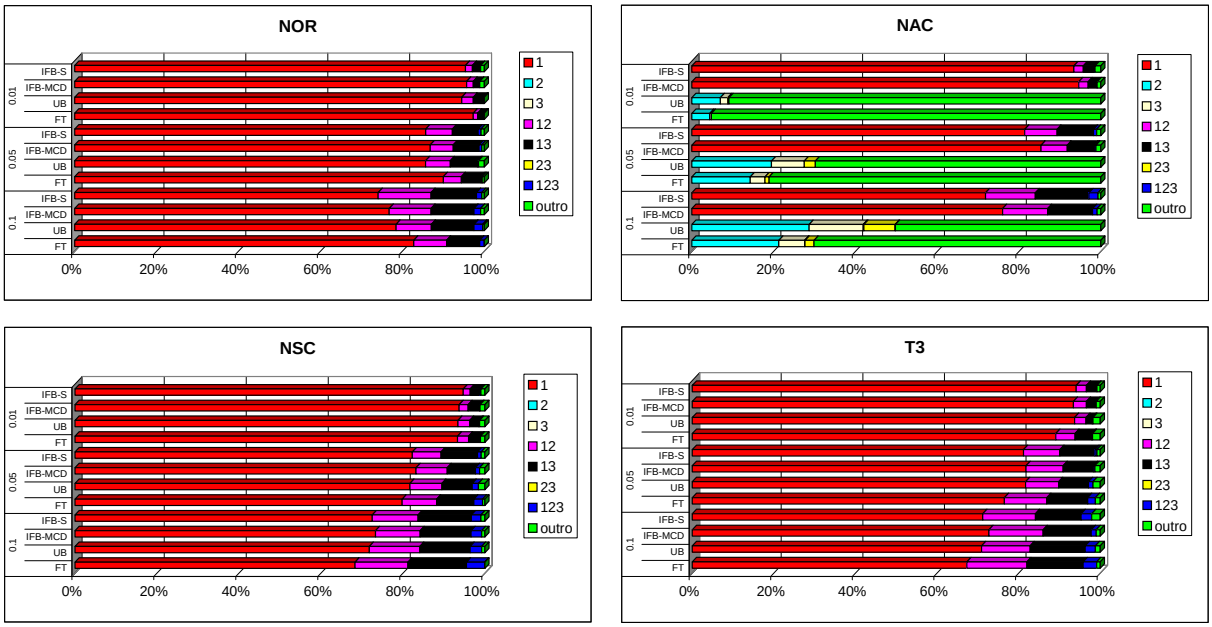


Figura 4.11: % de inclusão de cada subconjunto nas 1000 simulações quando se considerou o vector valor médio da segunda população $\boldsymbol{\mu}_2 = (2, 0, 0)$ e a regra discriminante foi obtida por um estimador clássico.

significância consideram-se apenas os casos em que H_0 é verdadeira. Isto é, para $\boldsymbol{\mu}_2 = (2, 0, 0)$ são verdadeiras, $H_0 : \alpha_2 = 0$, $H_0 : \alpha_3 = 0$ e $H_0 : \alpha_2 = \alpha_3 = 0$, enquanto que para $\boldsymbol{\mu}_2 = (\sqrt{2}, \sqrt{2}, 0)$ é verdadeira $H_0 : \alpha_3 = 0$. Quando se utiliza o estimador clássico, Tabelas 4.21 e 4.22, observa--se o comportamento esperado do teste F (FT) sob a distribuição Normal e a sua degradação face às distribuições contaminadas. Relativamente aos testes *bootstrap* observa-se a boa prestação do *bootstrap* usual (UB) quando o modelo gerador dos dados é a Normal (NOR) mas um pior desempenho quando o modelo é contaminado. É nesta situação que o *bootstrap* IF possui um melhor comportamento face aos restantes métodos, apesar de se verificar um ligeiro afastamento relativamente ao valor nominal γ_*.

Quando se utiliza um estimador robusto para determinar a regra discriminante, Tabelas 4.23 e 4.24, constata-se o mau desempenho do teste F, que em todas situações ultrapassou em muito o valor nominal. Este facto poderá ajudar a explicar a boa percentagem de inclusões observada para aquele nas Figuras 4.14, 4.15 e 4.16. Para os testes *bootstrap* verifica-se um comportamento semelhante, independentemente do tipo de reamostragem (usual ou IF) e de distribuição, tendo-se obtido valores estimados do nível de significância sempre inferiores aos nominais.

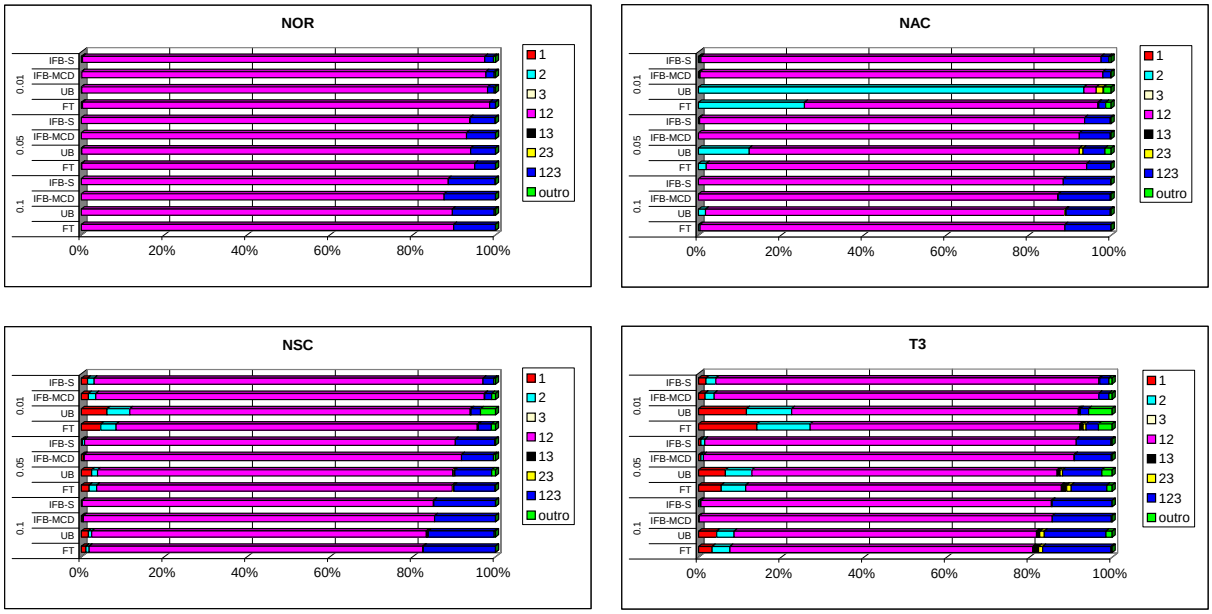


Figura 4.12: % de inclusão de cada subconjunto nas 1000 simulações quando se considerou o vector valor médio da segunda população $\boldsymbol{\mu}_2 = (\sqrt{2}, \sqrt{2}, 0)$ e a regra discriminante foi obtida por um estimador clássico.

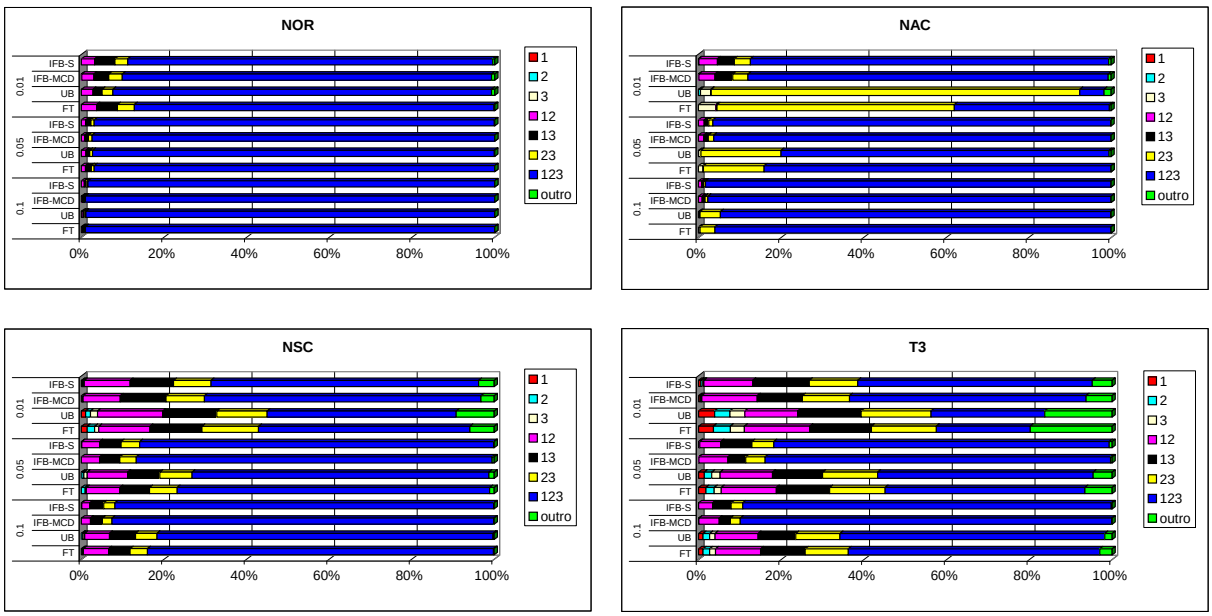


Figura 4.13: % de inclusão de cada subconjunto nas 1000 simulações quando se considerou o vector valor médio da segunda população $\boldsymbol{\mu}_2 = (2/\sqrt{3}, 2/\sqrt{3}, 2/\sqrt{3})$ e a regra discriminante foi obtida por um estimador clássico.

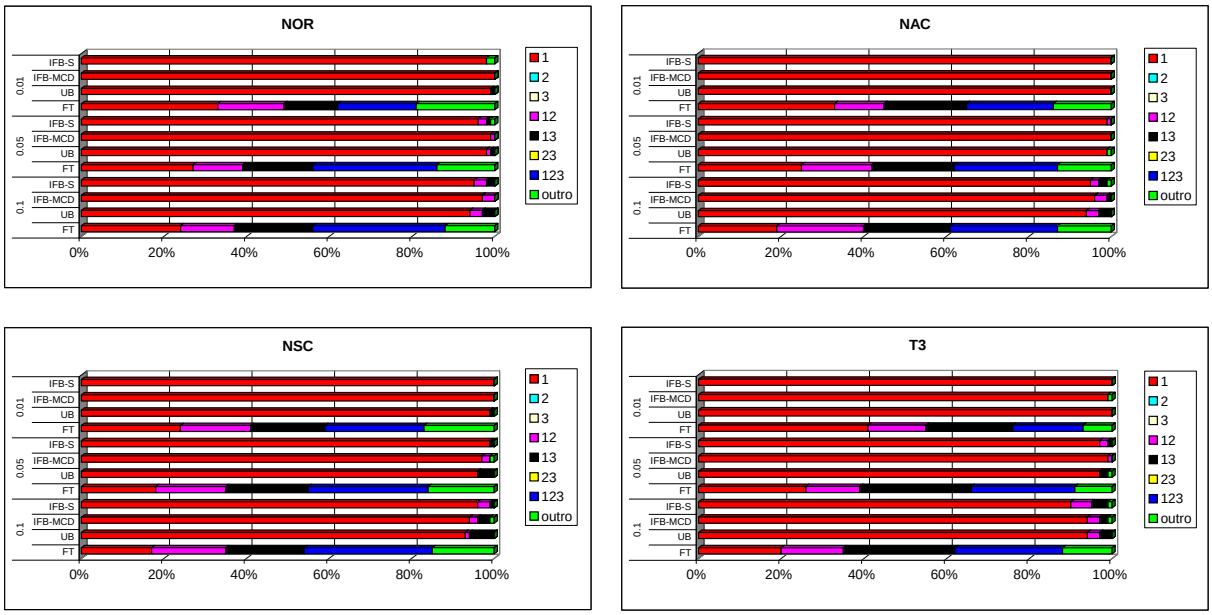


Figura 4.14: % de inclusão de cada subconjunto nas 100 simulações quando se considerou o vector valor médio da segunda população $\boldsymbol{\mu}_2 = (2, 0, 0)$ e a regra discriminante foi obtida por um estimador robusto.

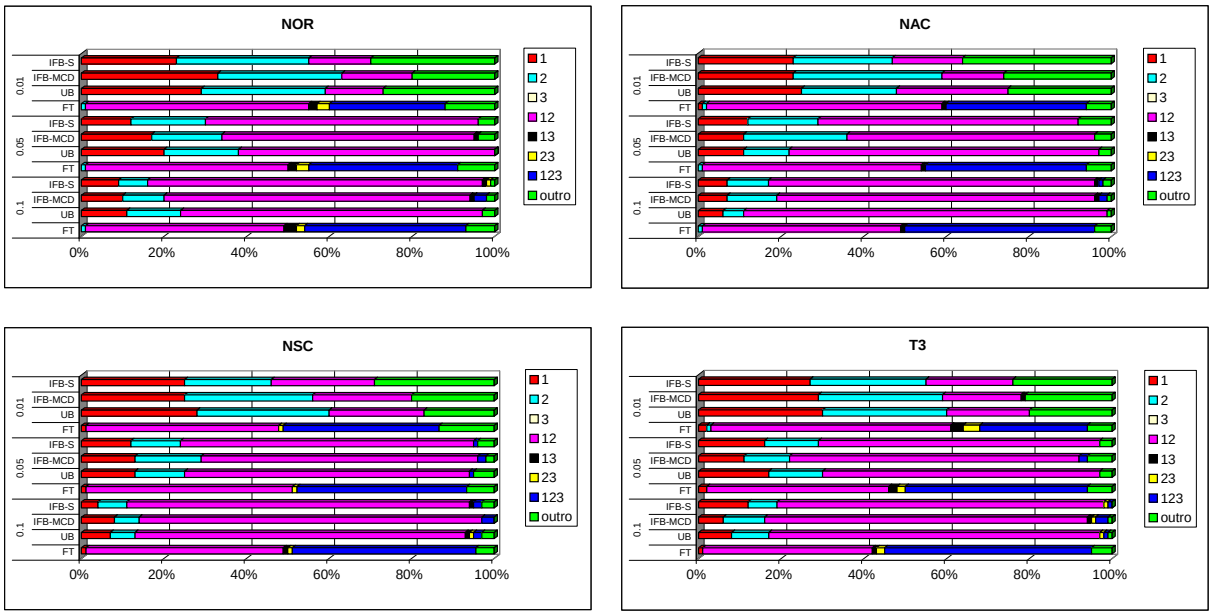


Figura 4.15: % de inclusão de cada subconjunto nas 100 simulações quando se considerou o vector valor médio da segunda população $\boldsymbol{\mu}_2 = (\sqrt{2}, \sqrt{2}, 0)$ e a regra discriminante foi obtida por um estimador robusto.

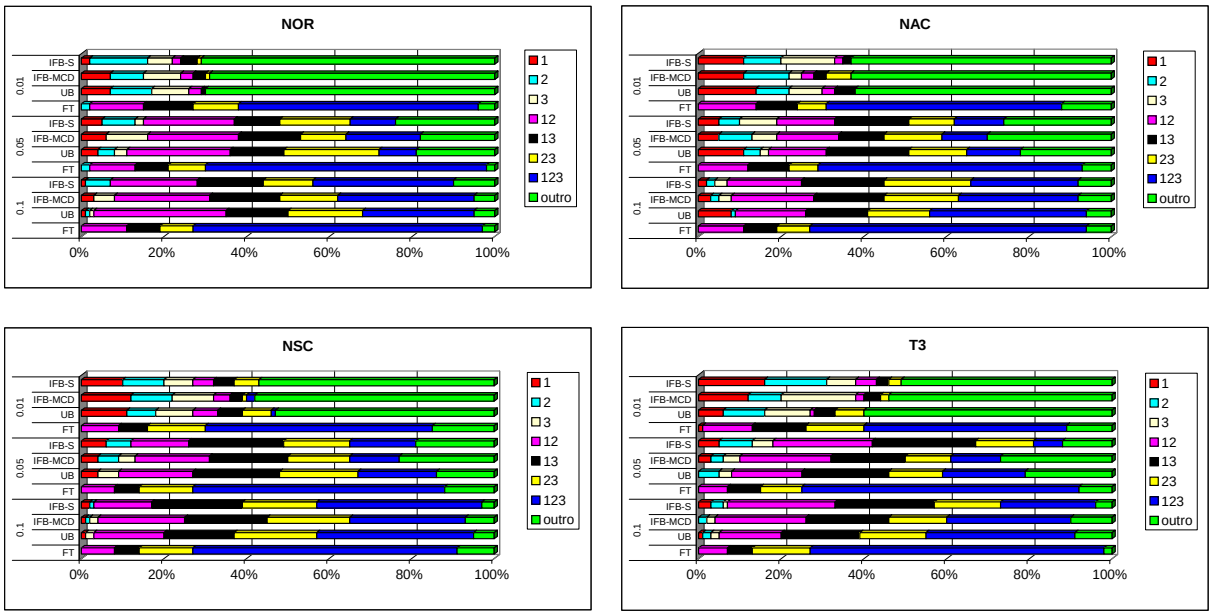


Figura 4.16: % de inclusão de cada subconjunto nas 100 simulações quando se considerou o vector valor médio da segunda população $\boldsymbol{\mu}_2 = (2/\sqrt{3}, 2/\sqrt{3}, 2/\sqrt{3})$ e a regra discriminante foi obtida por um estimador robusto.

Tabela 4.21: Frequência relativa do acontecimento "*p-value*$<\gamma_*$" em 1000 simulações para as hipóteses $H_0 : \alpha_2 = 0$, $H_0 : \alpha_3 = 0$, $H_0 : \alpha_2 = \alpha_3 = 0$ quando se utiliza o estimador clássico e para cada valor fixo de $\gamma_* = 0.1, 0.05, 0.01$. Caso a), $\boldsymbol{\mu} = (2, 0, 0)$.

		$\gamma = 0.1$				$\gamma = 0.05$				$\gamma = 0.01$			
Dist.	H_0	FT	UB	IFB-MCD	IFB-S	FT	UB	IFB-MCD	IFB-S	FT	UB	IFB-MCD	IFB-S
	$\alpha_2 = 0$	0.091	0.107	0.118	0.142	0.047	0.063	0.063	0.072	0.012	0.028	0.016	0.017
NOR	$\alpha_3 = 0$	0.092	0.125	0.121	0.124	0.051	0.069	0.070	0.070	0.015	0.025	0.016	0.020
	$\alpha_2 = \alpha_3 = 0$	0.096	0.142	0.137	0.144	0.052	0.092	0.079	0.092	0.010	0.033	0.026	0.035
	$\alpha_2 = 0$	0.173	0.151	0.135	0.135	0.105	0.093	0.086	0.079	0.028	0.032	0.023	0.020
NSC	$\alpha_3 = 0$	0.187	0.151	0.150	0.153	0.112	0.089	0.079	0.098	0.029	0.024	0.030	0.026
	$\alpha_2 = \alpha_3 = 0$	0.209	0.187	0.177	0.197	0.134	0.120	0.108	0.109	0.043	0.037	0.037	0.028
	$\alpha_2 = 0$	0.237	0.369	0.121	0.144	0.153	0.223	0.068	0.087	0.044	0.074	0.023	0.025
NAC	$\alpha_3 = 0$	0.085	0.210	0.121	0.153	0.046	0.107	0.070	0.097	0.004	0.021	0.022	0.029
	$\alpha_2 = \alpha_3 = 0$	0.203	0.410	0.140	0.187	0.104	0.299	0.091	0.115	0.028	0.126	0.027	0.034
	$\alpha_2 = 0$	0.180	0.144	0.145	0.155	0.123	0.092	0.095	0.094	0.049	0.031	0.032	0.025
T3	$\alpha_3 = 0$	0.171	0.160	0.130	0.138	0.119	0.084	0.078	0.091	0.043	0.017	0.024	0.026
	$\alpha_2 = \alpha_3 = 0$	0.235	0.187	0.196	0.199	0.161	0.121	0.130	0.118	0.072	0.028	0.036	0.041

Tabela 4.22: Frequência relativa do acontecimento "*p-value*$<\gamma_*$" em 1000 simulações para as hipóteses $H_0 : \alpha_3 = 0$ quando se utiliza o estimador clássico e para cada valor fixo de $\gamma_* = 0.1, 0.05, 0.01$. Caso b), $\boldsymbol{\mu} = (\sqrt{2}, \sqrt{2}, 0)$.

		$\gamma = 0.1$				$\gamma = 0.05$				$\gamma = 0.01$			
Dist.	H_0	FT	UB	IFB-MCD	IFB-S	FT	UB	IFB-MCD	IFB-S	FT	UB	IFB-MCD	IFB-S
NOR	$\alpha_3 = 0$	0.101	0.101	0.124	0.113	0.050	0.060	0.070	0.061	0.014	0.016	0.020	0.021
NSC	$\alpha_3 = 0$	0.176	0.164	0.147	0.150	0.103	0.094	0.078	0.095	0.034	0.025	0.018	0.026
NAC	$\alpha_3 = 0$	0.112	0.110	0.127	0.116	0.058	0.063	0.075	0.062	0.019	0.018	0.019	0.019
T3	$\alpha_3 = 0$	0.189	0.168	0.143	0.147	0.110	0.109	0.090	0.084	0.048	0.025	0.024	0.024

Tabela 4.23: Frequência relativa do acontecimento "*p-value*<γ_*" em 100 simulações para as hipóteses $H_0 : \alpha_2 = 0$, $H_0 : \alpha_3 = 0$, $H_0 : \alpha_{23} = 0$ quando se utiliza o estimador robusto e para cada valor fixo de $\gamma_* = 0.1, 0.05, 0.01$. Caso a), $\boldsymbol{\mu} = (2, 0, 0)$.

		$\gamma = 0.1$				$\gamma = 0.05$				$\gamma = 0.01$			
Dist.	H_0	FT	UB	IFB-MCD	IFB-S	FT	UB	IFB-MCD	IFB-S	FT	UB	IFB-MCD	IFB-S
NOR	$\alpha_2 = 0$	0.45	0.03	0.03	0.03	0.42	0.01	0.01	0.02	0.35	0.00	0.00	0.00
	$\alpha_3 = 0$	0.51	0.03	0.00	0.02	0.47	0.01	0.00	0.01	0.32	0.01	0.00	0.00
	$\alpha_2 = \alpha_3 = 0$	0.71	0.02	0.01	0.03	0.71	0.01	0.00	0.00	0.62	0.00	0.00	0.00
NSC	$\alpha_2 = 0$	0.49	0.01	0.02	0.03	0.46	0.00	0.02	0.00	0.41	0.00	0.00	0.00
	$\alpha_3 = 0$	0.50	0.06	0.03	0.01	0.49	0.04	0.00	0.01	0.42	0.01	0.00	0.00
	$\alpha_2 = \alpha_3 = 0$	0.79	0.04	0.03	0.01	0.77	0.01	0.03	0.00	0.70	0.01	0.00	0.00
NAC	$\alpha_2 = 0$	0.47	0.03	0.03	0.02	0.42	0.00	0.00	0.01	0.33	0.00	0.00	0.00
	$\alpha_3 = 0$	0.47	0.03	0.01	0.02	0.45	0.00	0.00	0.00	0.41	0.00	0.00	0.00
	$\alpha_2 = \alpha_3 = 0$	0.71	0.00	0.00	0.02	0.64	0.00	0.00	0.00	0.54	0.00	0.00	0.00
T3	$\alpha_2 = 0$	0.41	0.03	0.03	0.05	0.38	0.00	0.01	0.02	0.31	0.00	0.00	0.00
	$\alpha_3 = 0$	0.53	0.03	0.02	0.04	0.52	0.02	0.00	0.01	0.38	0.00	0.00	0.00
	$\alpha_2 = \alpha_3 = 0$	0.61	0.03	0.02	0.04	0.54	0.02	0.00	0.01	0.44	0.00	0.00	0.00

Tabela 4.24: Frequência relativa do acontecimento "*p-value*<γ_*" em 100 simulações para as hipóteses $H_0 : \alpha_3 = 0$ quando se utiliza o estimador robusto e para cada valor fixo de $\gamma_* = 0.1, 0.05, 0.01$.Caso b), $\boldsymbol{\mu} = (\sqrt{2}, \sqrt{2}, 0)$.

		$\gamma = 0.1$				$\gamma = 0.05$				$\gamma = 0.01$			
Dist.	H_0	FT	UB	IFB-MCD	IFB-S	FT	UB	IFB-MCD	IFB-S	FT	UB	IFB-MCD	IFB-S
NOR	$\alpha_3 = 0$	0.44	0.00	0.04	0.02	0.41	0.00	0.01	0.00	0.33	0.00	0.00	0.00
NSC	$\alpha_3 = 0$	0.47	0.04	0.03	0.03	0.42	0.01	0.02	0.01	0.39	0.00	0.00	0.00
NAC	$\alpha_3 = 0$	0.47	0.00	0.03	0.02	0.40	0.00	0.00	0.00	0.35	0.00	0.00	0.00
T3	$\alpha_3 = 0$	0.53	0.02	0.05	0.02	0.48	0.00	0.02	0.00	0.33	0.00	0.01	0.00

4.4 Generalização do *bootstrap* IF

O novo método proposto, *bootstrap* IF, não tem qualquer restrição no número de variáveis nem no número de parâmetros desde que esse número seja finito. Admite-se também que o procedimento possa ser aplicado quer a variáveis contínuas, discretas ou mistas, embora estes dois últimos casos constituam uma questão ainda em aberto, já que todas as experiências realizadas dizem respeito a variáveis contínuas. Pelos exemplos de aplicação e simulação apresentados nas secções anteriores pode verificar-se que o método *bootstrap* IF possuiu alguma sensibilidade relativamente à escolha da constante de afinação para determinação das probabilidades de reamostragem. De facto, para garantir que não existe sub-cobertura dos intervalos de confiança, sob o modelo central, necessita de constantes de afinação correspondentes a elevados percentis da distribuição da SIF. Seria preferível dispor de um método, que apesar de necessitar dessa constante de afinação seja mais flexível relativamente à sua escolha. É com esse objectivo que nesta secção se estuda e discute uma generalização para o método proposto. Este novo procedimento será designado por *bootstrap* baseado na função de influência generalizado e terá a seguinte notação, IFBG. Para além de um enquadramento teórico, apresentam-se resultados numéricos, para dados simulados e dados reais, quando o objectivo é a estimação de distribuições amostrais num modelo univariado de localização e escala e da correlação numa situação bivariada.

Na exposição seguinte, e para simplificar a apresentação, não se representam os vectores/matrizes a carregado, sempre que for necessário e que não se depreenda do texto, será indicado que se trata de um escalar, de um vector ou de uma matriz.

Recorde-se a definição de função de influência empírica padronizada, $RESIF$, assim como a definição dos pesos, w_i, $i = 1, 2, \cdots, n$, para uma dada amostra $(x_1, x_2, \cdots, x_n)$. Sejam $RESIF(x; F)$ e $w(x; F)$ os funcionais análogos, onde, por uma questão de simplicidade, se omitem quais os estimadores, robustos ou não, assim como o modelo em consideração. Então $RESIF_i = RESIF(x_i; F_n)$ e $w_i = w(x_i; F_n)$, para $i = 1, 2, \cdots, n$. Nas definições anteriores F representa o modelo gerador dos dados, o qual se assume que tem uma função de densidade f (com respeito a alguma medida, λ). Defina-se agora a função de distribuição teórica ponderada, F_w, como sendo a função de distribuição com densidade

$$f_w(x) = \frac{w(x; F) f(x)}{\int_{\mathcal{X}} w(x; F) f(x) d\lambda(x)}, \tag{4.31}$$

onde $\mathcal{X}$ representa o espaço amostral.

Recorda-se também que no procedimento proposto existem dois funcionais envolvidos, um denotado por T^{nr} equivalente a um estimador clássico e outro representado por T^r equivalente a um estimador robusto, ambos correspondendo a estimadores de um mesmo vector de parâmetros desconhecidos, θ, e no qual reside o interesse.

Seja $F_n = \frac{\sum_{i=1}^n \Delta_{x_i}}{n}$ a função de distribuição empírica, já referida acima, e

$$F_{w_n,n} = \frac{\sum_{i=1}^n w_i \Delta_{x_i}}{\sum_{i=1}^n w_i} = \sum_{i=1}^n p_i \Delta_{x_i} \tag{4.32}$$

a correspondente função de distribuição empírica ponderada.

Pelo teorema de Glivenko-Cantelli sabe-se que $F_n \rightarrow_{q.c.} F$, uniformemente em x. A relação entre F_n e F é semelhante à relação entre $F_{w_n,n}$ e F_w. A proposição 4.4 mostra que também $F_{w_n,n} \rightarrow_{q.c.} F_w$.

Proposição 4.4 *Para cada x, $F_{w_n,n}$ converge quase certamente para F_w se Ω_n^r convergir quase certamente para Ω e a função dos pesos for contínua nos parâmetros desconhecidos.*

Demonstração: Observe-se que

$$F_{w_n,n}(x) = \frac{\sum_{i=1}^n w(X_i;\Omega_n^r)\,\Delta_{X_i}(x)/n}{\sum_{i=1}^n w(X_i;\Omega_n^r)/n}, \tag{4.33}$$

onde $w(X_i;\Omega_n^r) \equiv w(X_i;F_n)$. O Lema A de Serfling (1980, pág. 253) e que não é mais do que uma extensão da Lei Forte dos Grandes Números, permite afirmar que o numerador de (4.33) converge quase certamente $(q.c.)$ para $E[w(X;\Omega)\,\Delta_X(x)]$, enquanto que o denominador converge $q.c.$ para $E[w(X;\Omega)]$, com $w(X;\Omega) \equiv w(X;F)$. Mas

$$E[w(X;\Omega)\,\Delta_X(x)] = \int_{-\infty}^{x} w(u;F)\,f(u)du \tag{4.34}$$

e

$$E[w(X;\Omega)] = \int_{\mathcal{X}} w(u;F)\,f(x)dx > 0,$$

o que demonstra a proposição.[3] ■

Esta proposição vem reforçar a importância da "suavidade" dos pesos w. De notar que se F_w for contínua então a convergência referida na proposição é uniforme em x.

Após as definições gerais das funções envolvidas no método *bootstrap* IF vai ilustrar-se as relações que existem entre elas e os estimadores que estão em causa. Para apresentar os conceitos fundamentais da generalização do método *bootstrap* IF tem-se em conta o modelo mais simples, o da localização univariada.

4.4.1 Localização univariada

Sejam $X_1, X_2, \cdots, X_n$, n variáveis aleatórias reais satisfazendo

$$X_i = \mu + \sigma\epsilon_i, \qquad i = 1, 2, \cdots, n, \tag{4.35}$$

onde ϵ_i são variáveis aleatórias independentes e identicamente distribuídas, com dispersão unitária e localização nula. O interesse reside na estimação de μ, sendo σ um parâmetro perturbador. Como habitualmente F representa a distribuição (comum) de X_i, $i = 1, 2, \cdots, n$, que se admite contínua com densidade f.

[3] Quando x é um vector os limites de integração em (4.34) devem ser entendidos como limites simples entre $-\infty$ e a componente respectiva de x.

Definam-se os pesos w_i $(i = 1, 2, \cdots, n)$ como registado na expressão (4.5) com $RESIF$ definida por (4.13). Seja $\mu(F) = \int xf(x)dx$ o funcional valor esperado e

$$\mu_w(F) = \int xw(x;F)\, f(x)dx \Big/ \int w(x;F)\, f(x)dx$$

o funcional valor esperado ponderado com função de pesos $w(x;F)$. Escolhendo uma função de pesos que decresça suficientemente rápido (pelo menos da ordem de $1/|x|$, quando $|x| \to \infty$) garante-se que $\mu_w(F)$ existe para qualquer distribuição F, mesmo quando $\mu(F)$ não existe. De notar que $\mu_w(F) \equiv \mu(F_w)$. Quando se consideram os estimadores verifica-se o mesmo tipo de coincidência, $\mu_w(F_n) \equiv \mu(F_{w_n,n})$. A Tabela 4.25 mostra as várias fases e analogias entre o modelo original e o modelo ponderado.

Por outro lado, se a função de pesos decrescer tão ou mais rapidamente que $1/x^2$, quando $|x| \to \infty$ garante-se que existe $Var(F_w)$, e que, pelo Teorema de Limite Central,

$$\sqrt{n}\left(\mu\left(F_{w_n,n}\right) - \mu\left(F_w\right)\right) \to_d N\left(0, Var\left(F_w\right)\right).$$

Tabela 4.25: Alguns dos intervenientes no modelo original e no modelo ponderado.

	Modelo original	Modelo ponderado
Função de densidade	f	f_w
Função de distribuição	F	F_w
Função de distribuição empírica	F_n	$F_{w_n,n}$
Funcionais	$\mu_w(F) \equiv \mu(F_w)$	
Estimadores	$\mu_w(F_n) \equiv \mu(F_{w_n,n})$	

O método proposto, *bootstrap* IF, simplesmente reamostra de forma usual mas a partir de $F_{w_n,n}$. Pelas propriedades do *bootstrap* usual sabe-se que a distribuição de $\mu(F_{w_n,n}) - \mu(F_w)$ pode ser estimada por $R_{n,boot}\left(\mu^*(F_{w_n,n}) - \mu(F_{w_n,n})\right)$, onde μ^* representa o estimador *bootstrap*. No entanto a distribuição que se quer estimar é a de $\mu_w(F_n) - \mu_w(F)$ que à partida não coincide com a de $\mu(F_{w_n,n}) - \mu(F_w)$, apesar da coincidência entre as estimativas. Interessa, então, determinar a relação entre estas duas distribuições ou ainda se eventualmente a distribuição de $\mu(F_{w_n,n}) - \mu(F_w)$ é mais facilmente relacionável com a distribuição de $\mu_{l(w)}(F_n) - \mu_{l(w)}(F)$, onde $l(w)$ é uma função dos pesos originais. Como se verá é este caso que se verifica, com $l(w) = \sqrt{w}$. Para chegar a essa conclusão faz-se uso das relações entre os estimadores–W de localização univariada (média ponderada das observações com função peso dependendo das observações) e os estimadores–M de localização univariada (Hampel *et al.*, 1986, Cap. 2).

Diz-se que μ_n^M é um estimador–M da localização univariada, μ, se for solução da equação

$$\sum_{i=1}^{n} \xi\left(X_i - \mu_n^M\right) = 0,$$

onde $\xi(u)$ é uma função ímpar cujas propriedades determinam o comportamento do estimador (ver Hampel *et al.*, 1986, Cap. 2). O funcional equivalente é definido como solução de

$$\int \xi\left(x-\mu^M(F)\right)dF = 0.$$

Tem-se obviamente $\mu_n^M = \mu^M(F_n)$. Se F for simétrica $\mu(F) \equiv \mu^M(F)$.

Outra forma de expressar μ_n^M é através de

$$\mu_n^M = \frac{\sum_{i=1}^n \frac{\xi\left(X_i-\mu_n^M\right)}{\left(X_i-\mu_n^M\right)} X_i}{\sum_{i=1}^n \frac{\xi(X_i-\mu_n^M)}{(X_i-\mu_n^M)}} = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}, \tag{4.36}$$

onde $w_i = \xi\left(X_i-\mu_n^M\right)/\left(X_i-\mu_n^M\right)$. Note-se que esta expressão não pode ser utilizada directamente para obter μ_n^M, já que os pesos dependem do estimador, mas é geralmente utilizada para obter uma aproximação do estimador–M a partir de uma solução inicial (estimador–W a um passo). A expressão (4.36) mostra que os estimadores–W completamente iterados são uma variante dos estimadores–M com $\xi(u) = uw(u)$. Logo possuem as mesmas propriedades que os estimadores–M, em particular têm a mesma função de influência e a mesma variância assintótica (Hampel *et al.*, 1986, pág. 116). Sob algumas condições adicionais esta característica estende-se também aos casos em que se realiza apenas uma iteração. A variância assintótica de um estimador–M de localização em F é dada por

$$VA\left(\mu^M, F\right) = \frac{\int \xi^2\left(x-\mu\right)dF}{\left(\int \xi'\left(x-\mu\right)dF\right)^2}, \tag{4.37}$$

se ξ for contínua, caso contrário surgem outras parcelas no denominador com efeito negativo (isto é, aumentam a variância). Este é mais um factor a reforçar a importância da suavidade dos pesos.

Substituindo $\xi(u) = uw(u)$ e $\xi'(u) = \frac{\partial \xi(u)}{\partial u} = w(u) + uw'(u)$, em (4.37), vem

$$VA(\mu_n^M, F) = \frac{\int u^2 w^2(u) dF}{\left(\int (w(u) + uw'(u))dF\right)^2}, \tag{4.38}$$

com $u = x - \mu$, pelo que

$$Var_F\left(\mu_n^M\right) \approx \frac{VA\left(\mu_n^M, F\right)}{n}. \tag{4.39}$$

Considerando agora o funcional μ em $F_{w_n,n}$, tem-se

$$E\left(\mu(F_{w_n,n})\right) = \int x f_w(x) dx$$

e

$$Var\left(\mu(F_{w_n,n})\right) = \frac{\int \left(x-\mu(F_w)\right)^2 f_w(x) dx}{n},$$

onde f_w é dada por (4.31). Estas quantidades podem escritas como

$$E\left(\mu(F_{w_n,n})\right) = \frac{\int x w(x) f(x) dx}{\int w(x) f(x) dx}$$

e

$$Var\left(\mu(F_{w_n,n})\right) = \frac{\int (x-\mu_w)^2\, w(x) f(x) dx}{n \int w(x) f(x) dx}. \tag{4.40}$$

Comparando as expressões (4.40) e (4.39) verifica-se que existem grandes semelhanças especialmente se no modelo não ponderado se tomar a raiz dos pesos. É esta a ideia motivadora do *bootstrap* IF generalizado, ou seja, reamostra-se o estimador usual com os pesos w e no final faz-se a devida correcção para que o processo reflicta o estimador com pesos $\sqrt{w}$, o qual se passa a representar por $\mu_{\sqrt{w}}(F_n)$, no modelo original. Desta forma a variância do estimador *bootstrap* assente nas reamostras ponderadas fica mais parecida com a variância do estimador $\mu_{\sqrt{w}}(F_n)$. A correcção a fazer é do tipo multiplicativo, admitindo que $\mu_{\sqrt{w}}(F) \approx \mu_w(F)$ (o que acontece se F for aproximadamente simétrica) e pode ser efectuada por duas vias:

I. alterando o número de elementos a reamostrar, isto é, em vez de se reamostrar n elementos de n reamostra-se n_{novo} elementos de n, com

$$n_{novo} = \frac{\left[\int \left(\sqrt{w(u)} + u\left(\sqrt{w(u)}\right)'\right) dF\right]^2}{\int w(u) dF} \times n$$

II. efectuando a reamostragem de n observações em n e no final corrigindo a variância através do factor multiplicativo n/n_{novo}.

Observe-se que embora n_{novo} dependa de quantidades desconhecidas é muito simples de estimar a partir de uma amostra concreta. De facto basta fazer

$$\hat{n}_{novo} = \frac{\left[\sum \sqrt{w(u_i)} + \sum u_i \left(\sqrt{w(u_i)}\right)'\right]^2}{\sum w(u_i)}, \tag{4.41}$$

com $u_i = (x_i - T_n)/S_n$, onde S_n é um estimador de escala robusto (sugere-se o MAD) e T_n é um estimador de localização robusto (sugere-se a mediana). Esta padronização é necessária para garantir a equivariância de todo o procedimento relativamente a mudanças de escala e é considerada em toda a literatura sobre métodos robustos (consulte-se, por exemplo, Hampel *et al.*, 1986, Cap. 2). De notar que se $w(u_i) \equiv 1$ todo o procedimento se reduz ao *bootstrap* usual e obtém-se naturalmente $\hat{n}_{novo} = n$.

Para avaliar o desempenho do método *bootstrap* IF com esta proposta de generalização fez-se um estudo de simulação, cujo delineamento e resultados são apresentados em seguida. Tanto o delineamento como o tipo de resultados foram inspirados no estudo de Srivastava e Chan (1989).

Estudo de Monte Carlo

Como já referido a distribuição de $T_n - \mu$, $R_n(x) = R_{n,F}(x) = P\{(T_n - \mu) \leq x\}$, onde T_n é um estimador genérico, pode ser estimada por *bootstrap* tendo em conta a seguinte aproximação de

Monte Carlo

$$R^B_{n,boot}(x) = B^{-1}\sum_{b=1}^{B} I\left(T^*_{n,b} - T_n \leq x\right),$$

onde $T^*_{n,b}$ representa a avaliação de T_n na b-ésima reamostra *bootstrap*. Analogamente ao que acontecia com o histograma *jackknife* cumulativo, obtido através do *deleted-d jackknife*, também o histograma *bootstrap* $\left\{T^*_{n,b} - T_n : b = 1, 2, \cdots, B\right\}$ pode ser utilizado para estimar a densidade de $T_n - \mu$.

O objectivo deste estudo de simulação é então estimar a distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, que se representa por $R_n(x)$, usando o método IFBG. Esta estimação é bastante importante pois a qualidade dos intervalos de confiança e testes de hipóteses *bootstrap* depende da qualidade da estimação *bootstrap* da distribuição.

Seja $(X_1, X_2, \cdots, X_n)$ uma amostra aleatória de dimensão n proveniente da população F, que é desconhecida. Considere-se o estimador $T_n = \bar{X}_n$ e o método *bootstrap* IF com a função $RESIF$ dada por (4.13). Seja $(x_1, x_2, \cdots, x_n)$ uma concretização da amostra aleatória referida e denote-se por $med = med(x_1, x_2, \cdots, x_n)$ a sua mediana e por MAD o seu desvio mediano absoluto. De acordo com o descrito atrás tenha-se em conta os pesos, w_i, para $i = 1, 2, \cdots, n$, definidos por

$$w_i = I_{[0,c]}\left(\left|\frac{x_i - med}{MAD}\right|\right) + \psi\left(c, \left|\frac{x_i - med}{MAD}\right|\right) \times I_{]c,+\infty]}\left(\left|\frac{x_i - med}{MAD}\right|\right), \tag{4.42}$$

onde $\psi(v) = \exp\left[-\frac{1}{2}(v-c)^2 c^{-2}\right]$ e c é a usual constante de afinação.

De modo a verificar o comportamento do método para constantes de afinação com valor inferior ao aplicado na simulação da Secção 4.3.1 escolheram-se para c os valores 1.5 e 2.0. Uma vez que o objectivo é também verificar qual o comportamento "robusto" do *bootstrap* IF generalizado, foram tidas em conta várias distribuições para o modelo gerador dos dados. A primeira distribuição constitui o caso de controlo, as quatro seguintes ilustram situações de contaminação simétrica e por fim as três últimas representam situações de contaminação assimétrica.

- Normal, NOR: $N(\mu = 0, \sigma^2 = 1)$;
- Normal Simetricamente ε-Contaminada com factor de escala f_e, NSC–ε:

  $$(1-\varepsilon)N(0,1) + \varepsilon N(0, f_e),$$

  com $\varepsilon = 0.25$ e $f_e = 9$;
- Normal simetricamente ε-contaminada com probabilidade $\varepsilon/2$ à volta do valor -4 e com probabilidade $\varepsilon/2$ à volta do valor 4. Esta distribuição é a empregue em Salibian-Barrera (2002), isto é,

  $$(1-\varepsilon)N(0,1) + \varepsilon/2N(-4, 0.1) + \varepsilon/2N(4, 0.1),$$

  - para $\varepsilon = 0.1$ denotar-se-á esta distribuição por M4–0.1;

  - para $\varepsilon = 0.2$ denotar-se-á esta distribuição por M4–0.2;

- t-Student com um grau de liberdade, T1: t-Student com um grau de liberdade, ou seja, a já mencionada distribuição de Cauchy;

- Normal Assimetricamente ε-Contaminada com probabilidade ε à volta do valor 10, NAC–ε:

$$(1-\varepsilon)N(0,1) + \varepsilon N(10, 0.1), \quad \text{com} \quad \varepsilon = 0.1;$$

- Lognormal padrão:

$$X \sim Lognormal(0, 1, \sigma) \quad \text{ou} \quad \ln(X) \sim N(0, \sigma)$$

  - para $\sigma = 1$ denotar-se-á esta distribuição por LOGN1;
  - para $\sigma = 2$ denotar-se-á esta distribuição por LOGN2.

Foram geradas amostras de dimensão $n = 20, 50$ e 100. Para cada combinação (dimensão, distribuição) foram geradas 500 amostras ($nsimul = 500$). Para cada amostra gerada usam-se 5000 réplicas *bootstrap* a partir do método *bootstrap* IF generalizado, IFBG, com probabilidades de reamostragem definidas pela normalização dos pesos w_i definidos por (4.42).

Como foi referido atrás o estimador *bootstrap* IF no modelo ponderado está relacionado com o estimador $\mu_{\sqrt{w}}(F_n) = \frac{\sum_{i=1}^{n}\sqrt{w_i}x_i}{\sum_{i=1}^{n}\sqrt{w_i}}$, com w_i $(i = 1, 2, \cdots, n)$ definidos por (4.42) no modelo original. Portanto é com a distribuição deste estimador $(R_n(x))$ que se deve fazer a comparação do desempenho do IFBG. Os autores Srivastava e Chan (1989) comparam as suas distribuições *bootstrap* com a distribuição exacta do estimador que estudaram, a variância amostral. Porém, no caso presente, a forma exacta do estimador de interesse não é simples de determinar para as distribuições mencionadas, pelo que se decidiu estimá-la através de um estudo de simulação independente. Para isso gerou-se um número elevado de amostras (5000) de cada uma das oito distribuições acima descritas com cada uma das dimensões consideradas ($n = 20, 50, 100$) e cada uma das constantes de afinação e calculou-se a estimativa de interesse (média ponderada) para cada uma. Esta amostra de 5000 observações foi depois centrada na sua média aritmética. Em seguida calcularam-se os percentis de ordem p : $2.5, 5, 10, 25, 50, 75, 90, 95$ e 97.5 % (valores escolhidos por Srivastava e Chan, 1989). Repetiu-se este processo 100 vezes e no fim determinou-se a mediana de cada um dos percentis.

Neste estudo foram tidas em conta várias estimativas da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$ com base no *bootstrap* IF generalizado:

1. $R^{(1)}_{BOOT}$:

$$R^{(1)}_{BOOT}(x) = \frac{1}{B}\sum_{b=1}^{B} I\left\{\left(\mu(F^*_{w_n,n})^b - \mu^*_n\right) \times \sqrt{n/\hat{n}_{novo}} + \mu^*_n - \mu_{\sqrt{w}}(F_n) \leq x\right\}, \tag{4.43}$$

   onde B é o número de réplicas *bootstrap* (neste caso $B = 5000$), $\mu(F^*_{w_n,n})^b$ é a média aritmética da b-ésima amostra *bootstrap* IFG, μ^*_n é a aproximação Monte Carlo do estimador *bootstrap*,

ou seja a média dos 5000 valores de $\mu(F^*_{w_n,n})$ e $\mu_{\sqrt{w}}(F_n)$ é a estimativa ponderada para a amostra original. Repare-se que $\mu^*_n - \mu_{\sqrt{w}}(F_n)$ não é mais do que uma estimativa *bootstrap* do viés do estimador $\mu_w(F_n)$ como estimador de $\mu_{\sqrt{w}}(F)$;

2. $R^{(2)}_{BOOT}$:

$$R^{(2)}_{BOOT}(x) = \frac{1}{B}\sum_{b=1}^{B} I\left\{\left(\mu(F^*_{w_n,n})^b - \mu^*_n\right) \times \sqrt{n/\hat{n}_{novo}} + \mu^*_n - \mu_w(F_n) \leq x\right\}, \tag{4.44}$$

esta distribuição empírica está centrada em $\mu_w(F_n)$ em vez de $\mu_{\sqrt{w}}(F_n)$;

3. $R^{(k)}_{BOOT}, \quad k = 3, 4, \cdots, 12$:

$$R^{(k)}_{BOOT}(x) = \frac{1}{B}\sum_{b=1}^{B} I\left\{\left(\mu(F^*_{w_n,n})^b - \mu^*_n\right) \times \sqrt{n/\hat{n}_{novo}} \times \sqrt{fac_k} + \mu^*_n - \mu_w(F_n) \leq x\right\}, \tag{4.45}$$

onde o factor de correcção fac_k é dado por

$$\left\{[Var(\mu(F_{w_n,n}))]^B_{Boot} + n^{-1}a_k d^2_{est}\right\} \Big/ [Var(\mu(F_{w_n,n}))]^B_{Boot},$$

$[Var(\mu(F_{w_n,n}))]^B_{Boot}$ é o estimador *bootstrap* da variância do estimador média usual na amostra ponderada, a_k são constantes que variam de 25 a 250 com incrementos de 25 e por último $d^2_{est} = \mu_{\sqrt{w}}(F_n) - \mu_w(F_n)$.

R_n foi também estimada utilizando o *bootstrap* IF, $R^{(IF)}_{BOOT}$. Todavia devido ao custo computacional optou-se pela realização desses cálculos apenas para as distribuições NOR, NSC e NAC, representando uma situação "normal", uma de contaminação simétrica e uma de contaminação assimétrica.

Resultados do estudo Monte Carlo

Definidas as formas das estimativas da distribuição *bootstrap*, o desempenho de cada uma delas (ou mais concretamente, das correcções efectuadas) vai ser avaliado através da proximidade entre os respectivos percentis amostrais e os estimados ($\widehat{F^{-1}_{\mu_{\sqrt{w}}}}(p)$, para um dado p) no estudo de simulação independente sem *bootstrap*. Para cada um das 500 simulações/amostras calcula-se $R^{(k)}_{BOOT}\left(\widehat{F^{-1}_{\mu_{\sqrt{w}}}}(p)\right)$, com $k = IF, 1, 2, \cdots, 12$. Os resultados finais são a média, ME_p, e o desvio padrão, SD_p desses 500 valores. É de notar que se a distribuição *bootstrap* estiver a ser estimada correctamente se deve ter $ME_p \approx p$, $\forall p$.

De forma a obter uma medida de desempenho global calcularam-se os seguintes indicadores:

- Relativamente a ME_p,

1. Máximo dos desvios absolutos (uma medida inspirada na estatística de Kolmogorov-Smirnov): $KS = \max_p |ME_p - p|$;
2. Soma dos desvios absolutos: $DA = \sum_p |ME_p - p|$;
3. Raiz quadrada da soma dos quadrados dos desvios: $SQ = \sqrt{\sum_p (ME_p - p)^2}$.

- Relativamente a SD_p, determinou-se o desvio padrão combinado: $SC = \sqrt{\sum_p SD_p^2 \big/ 9}$.

As Tabelas 4.26 a 4.37 apresentam os valores dos indicadores ME_p e SD_p para cada ordem p fornecidos por três das doze variantes estudadas do método *bootstrap* IF generalizado, $R_{BOOT}^{(k)}$, $k = IF, 1, 2, 3$, para as distribuições NOR, NSC, M4–01, T1, NAC e LOGN1. Os restantes resultados ($R_{BOOT}^{(k)}$, $k = 4, 5, \cdots, 12$, para aquelas distribuições e $R_{BOOT}^{(k)}$, $k = 1, 2, \cdots, 12$, para as distribuições M4–02 e LOGN2) apresentam-se no Apêndice B.

Pelos resultados obtidos verifica-se que as correcções propostas traduziram-se numa melhoria em relação ao desempenho do método *bootstrap* IF, em especial para o valor mais baixo da constante de afinação. De um modo geral é para a distribuição NSC que se nota uma maior diferença entre os resultados de IFB e IFBG.

Como era esperado, verifica-se para todas as distribuições uma melhoria de resultados à medida que aumenta a dimensão da amostra.

De uma forma global nota-se que as três primeiras distribuições R_{BOOT} têm bons desempenhos quando as amostras provêm de modelos simetricamente contaminados, havendo apenas um aumento na variabilidade, SC. Relativamente aos resultados em apêndice, Tabelas B.1 a B.30, verifica-se que quando os dados provêm da distribuição Normal não há grande vantagem em aumentar a constante a_k da expressão (4.45). Porém quando se têm dados com a contaminação simétrica NSC–0.1, os resultados melhoram (nomeadamente os valores de KS e SC) para as distribuições empíricas $R_{BOOT}^{(k)}$, $k = 5, 6, 7$, com excepção do caso $n = 20$ e $c = 1.5$.

Em relação aos valores obtidos para as distribuições assimétricas, Tabelas 4.34 a 4.37 salientam-se os valores superiores quer de KS quer de SC quando comparados com os das distribuições anteriores (sem contaminação e com contaminação simétrica). Para a distribuição NAC–0.1, Tabelas 4.34 e 4.35, verifica-se que para um mesmo valor n os resultados pioram bastante quando se passa de $c = 1.5$ para $c = 2$, mostrando o efeito da diminuição da "robustez" quando aumenta a constante de afinação. No geral, são as distribuições *bootstrap* $R_{BOOT}^{(2)}$ e $R_{BOOT}^{(3)}$ que têm melhor desempenho. No entanto, quando se comparam com as restantes distribuições *bootstrap* (ver Apêndice B, B.1–B.30) verifica-se que a distância KS diminui com o aumento de a_k mas os desvios SC aumentam. Um bom compromisso será optar por $R_{BOOT}^{(3)}$ com $c = 1.5$.

No que diz respeito à distribuição LOGN1, Tabelas 4.36 e 4.37 os resultados mostram que os percentis determinados por $R_{BOOT}^{(1)}$ pioram quando n aumenta. Este comportamento já não se verifica para as restantes distribuições empíricas $R_{BOOT}^{(k)}$, $k = 2, 3 \cdots, 12$ (para $k > 3$ ver resultados

que se encontram no Apêndice B). Para o mesmo valor de n os resultados melhoram ligeiramente com o aumento de c. A distribuição R_{BOOT} que melhor se comporta é a $R_{BOOT}^{(11)}$.

Não é de estranhar que para a distribuição LOGN2 (ver Apêndice B) os resultados piorem, pois esta distribuição possui a cauda mais pesada que a LOGN1. Nesta situação não parece ser muito importante a constante de afinação, já que os resultados são muito semelhantes para diferentes valores de c com n fixo. Aqui também se mantém a distribuição $R_{BOOT}^{(11)}$ como a que tem melhor desempenho.

Tabela 4.26: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal padrão e $c = 1.5$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC.

					NOR $n = 20$ $c = 1.5$						
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R_{BOOT}^{(IF)}$	ME	1.97	3.96	8.23	22.63	49.80	77.09	91.71	96.06	98.04	2.37
	SD	1.60	2.50	3.73	4.95	4.05	4.86	3.72	2.51	1.61	3.50
$R_{BOOT}^{(1)}$	ME	2.50	4.75	9.34	23.90	49.98	76.07	90.76	95.39	97.60	1.10
	SD	1.91	2.81	3.95	4.78	3.40	4.58	3.80	2.67	1.78	3.45
$R_{BOOT}^{(2)}$	ME	2.46	4.68	9.26	23.77	49.93	76.15	90.80	95.41	97.60	1.23
	SD	1.81	2.66	3.71	4.08	0.92	4.02	3.70	2.65	1.80	3.01
$R_{BOOT}^{(3)}$	ME	2.49	4.73	9.32	23.84	49.93	76.08	90.73	95.37	97.57	1.16
	SD	1.83	2.67	3.72	4.07	0.92	4.01	3.72	2.67	1.83	3.02
					NOR $n = 50$ $c = 1.5$						
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R_{BOOT}^{(IF)}$	ME	2.21	4.44	9.10	23.88	49.95	76.01	90.89	95.58	97.83	1.12
	SD	1.17	1.81	2.58	3.19	2.78	3.21	2.58	1.78	1.14	2.37
$R_{BOOT}^{(1)}$	ME	2.44	4.83	9.66	24.48	50.03	75.45	90.33	95.16	97.57	0.52
	SD	1.14	1.73	2.41	2.83	2.42	3.04	2.46	1.75	1.15	2.20
$R_{BOOT}^{(2)}$	ME	2.44	4.82	9.64	24.46	50.05	75.52	90.38	95.21	97.60	0.54
	SD	1.13	1.69	2.28	2.33	0.75	2.38	2.27	1.66	1.11	1.83
$R_{BOOT}^{(3)}$	ME	2.45	4.83	9.65	24.48	50.05	75.50	90.37	95.19	97.59	0.52
	SD	1.13	1.69	2.28	2.32	0.75	2.38	2.27	1.66	1.11	1.83
					NOR $n = 100$ $c = 1.5$						
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R_{BOOT}^{(IF)}$	ME	2.16	4.37	9.07	24.11	50.16	76.11	90.97	95.67	97.91	1.11
	SD	0.83	1.29	1.92	2.61	2.62	2.52	1.87	1.26	0.79	1.88
$R_{BOOT}^{(1)}$	ME	2.58	5.03	9.95	25.04	50.19	75.25	90.13	95.06	97.51	0.25
	SD	0.87	1.32	1.86	2.41	2.38	2.36	1.82	1.28	0.84	1.79
$R_{BOOT}^{(2)}$	ME	2.55	4.97	9.87	24.90	50.09	75.19	90.13	95.06	97.50	0.19
	SD	0.82	1.20	1.59	1.60	0.65	1.58	1.57	1.18	0.81	1.27
$R_{BOOT}^{(3)}$	ME	2.55	4.97	9.88	24.90	50.09	75.18	90.12	95.05	97.50	0.18
	SD	0.82	1.20	1.59	1.60	0.65	1.58	1.57	1.18	0.81	1.27

Tabela 4.27: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal padrão e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC.

NOR $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R_{BOOT}^{(IF)}$	ME	2.22	4.31	8.85	23.38	49.96	76.53	91.26	95.76	97.84	1.62
	SD	1.68	2.50	3.57	4.15	2.47	4.18	3.56	2.49	1.67	3.06
$R_{BOOT}^{(1)}$	ME	2.51	4.75	9.44	23.99	49.88	75.74	90.58	95.30	97.56	1.01
	SD	1.80	2.61	3.62	4.10	2.42	3.96	3.61	2.62	1.80	3.07
$R_{BOOT}^{(2)}$	ME	2.51	4.75	9.44	24.02	49.98	75.87	90.64	95.32	97.57	0.98
	SD	1.80	2.59	3.56	3.79	0.97	3.77	3.61	2.63	1.80	2.89
$R_{BOOT}^{(3)}$	ME	2.52	4.77	9.47	24.05	49.98	75.84	90.61	95.30	97.56	0.95
	SD	1.80	2.59	3.55	3.77	0.97	3.75	3.60	2.63	1.80	2.88
NOR $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R_{BOOT}^{(IF)}$	ME	2.40	4.77	9.57	24.33	49.97	75.56	90.44	95.26	97.63	0.67
	SD	1.18	1.75	2.37	2.57	1.62	2.54	2.32	1.70	1.13	1.98
$R_{BOOT}^{(1)}$	ME	2.45	4.83	9.68	24.52	50.09	75.49	90.38	95.22	97.60	0.49
	SD	1.15	1.68	2.29	2.45	1.52	2.44	2.24	1.66	1.12	1.91
$R_{BOOT}^{(2)}$	ME	2.45	4.83	9.67	24.50	50.10	75.50	90.39	95.23	97.61	0.50
	SD	1.14	1.67	2.25	2.28	0.79	2.24	2.21	1.65	1.12	1.79
$R_{BOOT}^{(3)}$	ME	2.45	4.83	9.67	24.51	50.10	75.50	90.39	95.23	97.61	0.50
	SD	1.14	1.67	2.25	2.28	0.79	2.23	2.21	1.65	1.12	1.79
NOR $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R_{BOOT}^{(IF)}$	ME	2.39	4.73	9.57	24.53	50.08	75.53	90.41	95.26	97.68	0.53
	SD	0.80	1.18	1.64	1.79	1.24	1.86	1.67	1.21	0.80	1.41
$R_{BOOT}^{(1)}$	ME	2.58	5.00	9.94	24.91	50.05	75.21	90.07	95.01	97.52	0.21
	SD	0.84	1.22	1.68	1.82	1.16	1.71	1.63	1.20	0.79	1.39
$R_{BOOT}^{(2)}$	ME	2.57	4.98	9.92	24.87	50.02	75.18	90.06	95.00	97.52	0.18
	SD	0.82	1.19	1.62	1.61	0.66	1.60	1.63	1.21	0.80	1.29
$R_{BOOT}^{(3)}$	ME	2.57	4.98	9.92	24.87	50.02	75.18	90.06	95.00	97.52	0.18
	SD	0.82	1.19	1.62	1.61	0.66	1.60	1.63	1.21	0.80	1.29

Tabela 4.28: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal simetricamente contaminada e $c = 1.5$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC.

NSC–0.25 $n = 20$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(IF)}_{BOOT}$	ME	1.54	3.38	7.33	21.67	51.24	80.44	93.72	97.20	98.78	5.44
	SD	3.72	5.66	8.29	11.76	10.79	10.29	6.88	4.35	2.59	7.80
$R^{(1)}_{BOOT}$	ME	2.91	5.41	9.98	23.95	49.66	75.53	89.69	94.43	97.03	1.05
	SD	5.15	7.18	9.40	11.09	9.18	11.46	9.70	7.38	5.31	8.70
$R^{(2)}_{BOOT}$	ME	2.74	5.12	9.54	23.48	49.86	76.42	90.43	94.92	97.33	1.52
	SD	4.44	6.22	8.12	8.47	1.58	8.52	8.08	6.18	4.41	6.62
$R^{(3)}_{BOOT}$	ME	3.01	5.49	10.03	23.96	49.87	75.94	89.94	94.54	97.06	1.04
	SD	4.74	6.52	8.38	8.56	1.57	8.63	8.33	6.47	4.70	6.82
NSC–0.25 $n = 50$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(IF)}_{BOOT}$	ME	1.09	2.58	6.20	20.30	49.27	78.91	93.49	97.29	98.87	4.70
	SD	1.90	3.44	5.76	9.74	10.90	9.25	5.62	3.41	1.94	6.62
$R^{(1)}_{BOOT}$	ME	2.86	5.42	10.41	25.37	50.35	75.12	89.67	94.59	97.15	0.42
	SD	3.52	5.10	6.99	9.13	9.61	9.41	7.21	5.33	3.82	7.04
$R^{(2)}_{BOOT}$	ME	2.58	4.95	9.68	24.43	50.04	75.61	90.23	95.00	97.42	0.61
	SD	2.85	4.03	5.29	5.22	0.90	5.23	5.33	4.08	2.89	4.23
$R^{(3)}_{BOOT}$	ME	2.72	5.16	9.95	24.71	50.04	75.34	89.95	94.80	97.28	0.34
	SD	2.97	4.15	5.38	5.25	0.90	5.26	5.41	4.19	3.00	4.30
NSC–0.25 $n = 100$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(IF)}_{BOOT}$	ME	1.29	2.99	7.00	21.63	50.07	78.24	92.98	97.02	98.71	3.37
	SD	1.81	3.16	5.34	9.36	11.54	9.37	5.37	3.23	1.93	6.61
$R^{(1)}_{BOOT}$	ME	2.67	5.16	10.10	24.79	49.63	74.44	89.52	94.63	97.22	0.56
	SD	2.66	3.99	5.80	8.42	9.52	8.41	5.78	4.02	2.69	6.20
$R^{(2)}_{BOOT}$	ME	2.43	4.81	9.64	24.48	50.07	75.44	90.33	95.17	97.57	0.52
	SD	1.93	2.79	3.74	3.72	0.67	3.69	3.73	2.79	1.91	2.96
$R^{(3)}_{BOOT}$	ME	2.50	4.92	9.79	24.62	50.07	75.29	90.17	95.06	97.49	0.38
	SD	1.97	2.83	3.77	3.72	0.66	3.69	3.76	2.84	1.95	2.98

Tabela 4.29: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal simetricamente contaminada e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC.

NSC–0.25 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(IF)}_{BOOT}$	ME	1.58	3.41	7.37	21.40	50.41	79.55	93.05	96.80	98.52	4.55
	SD	3.45	5.66	8.66	12.78	12.21	12.06	8.21	5.32	3.32	8.71
$R^{(1)}_{BOOT}$	ME	2.48	4.88	9.48	23.63	49.79	75.91	90.12	94.88	97.35	1.37
	SD	4.22	6.33	9.05	11.65	10.27	12.44	9.83	7.10	4.94	8.86
$R^{(2)}_{BOOT}$	ME	2.42	4.63	8.98	22.89	49.88	76.97	90.95	95.40	97.66	2.11
	SD	4.13	5.84	7.99	9.06	1.95	9.15	8.05	5.84	4.05	6.66
$R^{(3)}_{BOOT}$	ME	2.67	5.03	9.53	23.46	49.88	76.38	90.40	94.99	97.40	1.54
	SD	4.32	6.07	8.22	9.12	1.94	9.25	8.29	6.11	4.25	6.83
NSC–0.25 $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(IF)}_{BOOT}$	ME	1.59	3.47	7.69	22.58	50.75	78.53	92.91	96.95	98.69	3.53
	SD	2.88	4.50	6.80	10.61	11.52	9.82	6.10	3.78	2.28	7.25
$R^{(1)}_{BOOT}$	ME	2.62	5.07	9.89	24.78	50.33	75.49	90.17	95.02	97.45	0.49
	SD	3.54	5.18	7.22	9.53	9.80	9.61	7.16	5.10	3.48	7.15
$R^{(2)}_{BOOT}$	ME	2.38	4.64	9.19	23.83	50.09	76.10	90.72	95.38	97.65	1.17
	SD	2.91	4.21	5.66	5.87	0.96	5.96	5.69	4.21	2.89	4.57
$R^{(3)}_{BOOT}$	ME	2.50	4.83	9.46	24.11	50.09	75.82	90.46	95.19	97.53	0.89
	SD	3.00	4.29	5.73	5.87	0.96	5.96	5.75	4.30	2.97	4.61
NSC–0.25 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(IF)}_{BOOT}$	ME	1.32	3.07	7.26	22.24	50.42	78.40	92.97	96.99	98.68	3.40
	SD	1.73	3.03	5.23	9.45	11.68	9.52	5.49	3.35	2.04	6.68
$R^{(1)}_{BOOT}$	ME	2.68	5.17	10.17	25.12	49.96	74.65	89.45	94.53	97.14	0.55
	SD	2.99	4.35	6.09	8.51	9.76	9.20	6.57	4.59	3.07	6.59
$R^{(2)}_{BOOT}$	ME	2.50	4.86	9.68	24.54	50.06	75.49	90.29	95.13	97.52	0.49
	SD	2.27	3.31	4.42	4.39	0.77	4.38	4.38	3.26	2.24	3.49
$R^{(3)}_{BOOT}$	ME	2.57	4.97	9.83	24.69	50.06	75.33	90.13	95.02	97.45	0.33
	SD	2.32	3.35	4.45	4.38	0.76	4.38	4.41	3.31	2.28	3.51

Tabela 4.30: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4-0.1 e $c = 1.5$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC.

M4–0.1 $n = 20$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.67	4.96	9.53	23.75	49.79	75.92	90.45	95.04	97.39	1.25
	SD	2.94	4.41	6.41	8.47	7.39	8.12	6.12	4.25	2.78	6.01
$R^{(2)}_{BOOT}$	ME	2.57	4.77	9.22	23.43	50.02	76.55	90.80	95.20	97.44	1.57
	SD	2.72	4.04	5.72	6.52	1.20	6.46	5.74	4.12	2.75	4.71
$R^{(3)}_{BOOT}$	ME	2.70	4.98	9.52	23.78	50.01	76.20	90.49	94.99	97.31	1.22
	SD	2.76	4.07	5.73	6.47	1.20	6.41	5.75	4.15	2.79	4.71
M4–0.1 $n = 50$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.62	5.02	9.80	24.37	49.34	74.74	89.76	94.72	97.21	0.66
	SD	2.24	3.38	4.90	6.88	7.32	7.06	5.08	3.53	2.38	5.11
$R^{(2)}_{BOOT}$	ME	2.57	4.95	9.68	24.39	49.90	75.62	90.36	95.08	97.42	0.62
	SD	2.06	3.04	4.14	4.32	0.77	4.32	4.18	3.08	2.11	3.33
$R^{(3)}_{BOOT}$	ME	2.64	5.06	9.83	24.56	49.90	75.45	90.21	94.97	97.35	0.45
	SD	2.07	3.05	4.12	4.27	0.77	4.27	4.16	3.08	2.12	3.32
M4–0.1 $n = 100$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.74	5.32	10.36	25.25	50.00	74.91	89.81	94.80	97.28	0.36
	SD	1.78	2.76	4.07	6.08	6.93	6.12	4.10	2.77	1.81	4.44
$R^{(2)}_{BOOT}$	ME	2.59	5.06	9.95	24.75	49.83	75.12	90.06	94.98	97.40	0.25
	SD	1.53	2.24	3.00	3.00	0.70	3.00	3.02	2.23	1.52	2.39
$R^{(3)}_{BOOT}$	ME	2.63	5.12	10.02	24.83	49.83	75.04	89.98	94.93	97.37	0.17
	SD	1.53	2.24	2.99	2.97	0.70	2.97	3.00	2.23	1.52	2.37

Tabela 4.31: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.1 e $c = d = 2$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC.

M4–0.1 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.62	4.86	9.40	23.83	50.17	76.06	90.52	95.08	97.38	1.17
	SD	2.47	3.67	5.21	6.59	5.33	6.96	5.36	3.70	2.44	4.89
$R^{(2)}_{BOOT}$	ME	2.62	4.84	9.29	23.53	50.07	76.30	90.69	95.16	97.42	1.47
	SD	2.46	3.67	5.16	5.95	1.30	5.95	5.11	3.61	2.40	4.26
$R^{(3)}_{BOOT}$	ME	2.67	4.92	9.42	23.70	50.07	76.13	90.56	95.07	97.37	1.30
	SD	2.47	3.66	5.13	5.86	1.30	5.87	5.08	3.60	2.40	4.23
M4–0.1 $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.59	4.94	9.71	24.46	49.98	75.64	90.34	95.05	97.41	0.64
	SD	1.82	2.72	3.88	4.84	4.15	4.83	3.85	2.71	1.80	3.58
$R^{(2)}_{BOOT}$	ME	2.56	4.88	9.61	24.30	49.95	75.76	90.44	95.11	97.44	0.76
	SD	1.78	2.66	3.72	3.94	0.83	3.95	3.67	2.65	1.78	2.97
$R^{(3)}_{BOOT}$	ME	2.58	4.91	9.66	24.35	49.95	75.72	90.40	95.08	97.42	0.72
	SD	1.78	2.65	3.70	3.91	0.83	3.92	3.66	2.65	1.78	2.96
M4–0.1 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.61	5.09	10.00	24.92	50.01	75.26	90.08	95.00	97.42	0.26
	SD	1.36	2.04	2.83	3.66	3.67	3.84	2.94	2.06	1.35	2.80
$R^{(2)}_{BOOT}$	ME	2.58	5.04	9.89	24.77	49.93	75.32	90.14	95.04	97.45	0.32
	SD	1.35	1.99	2.66	2.73	0.72	2.75	2.63	1.94	1.30	2.13
$R^{(3)}_{BOOT}$	ME	2.58	5.05	9.91	24.79	49.93	75.30	90.12	95.03	97.44	0.30
	SD	1.35	1.98	2.65	2.71	0.72	2.73	2.62	1.93	1.29	2.12

Tabela 4.32: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição T1 e $c = 1.5$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC.

T1 $n = 20$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.64	5.00	9.61	24.21	50.49	76.73	91.09	95.53	97.71	1.73
	SD	4.49	6.23	8.27	10.10	8.27	8.92	6.83	4.82	3.30	7.14
$R^{(2)}_{BOOT}$	ME	2.30	4.49	8.88	23.23	50.06	76.92	91.16	95.51	97.64	1.92
	SD	3.45	4.93	6.67	7.46	1.38	7.32	6.54	4.87	3.50	5.48
$R^{(3)}_{BOOT}$	ME	2.47	4.75	9.25	23.65	50.06	76.50	90.80	95.25	97.47	1.50
	SD	3.63	5.12	6.81	7.43	1.38	7.28	6.66	5.03	3.66	5.57
T1 $n = 50$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.65	5.07	9.86	24.60	49.83	75.03	89.84	94.74	97.24	0.40
	SD	2.88	4.17	5.80	7.67	7.93	8.02	6.10	4.37	3.00	5.88
$R^{(2)}_{BOOT}$	ME	2.54	4.86	9.58	24.32	50.03	75.70	90.41	95.14	97.48	0.70
	SD	2.48	3.52	4.70	4.81	0.79	4.85	4.71	3.54	2.47	3.78
$R^{(3)}_{BOOT}$	ME	2.63	5.00	9.77	24.51	50.03	75.51	90.22	95.00	97.39	0.51
	SD	2.54	3.58	4.75	4.80	0.79	4.85	4.77	3.59	2.53	3.82
T1 $n = 100$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.78	5.26	10.23	24.99	49.69	74.48	89.44	94.46	97.05	0.56
	SD	2.37	3.44	4.85	6.61	7.12	6.87	5.23	3.83	2.69	5.07
$R^{(2)}_{BOOT}$	ME	2.70	5.13	10.04	24.89	50.04	75.12	89.97	94.84	97.31	0.20
	SD	2.02	2.86	3.75	3.67	0.66	3.66	3.76	2.88	2.03	2.99
$R^{(3)}_{BOOT}$	ME	2.75	5.19	10.12	24.97	50.04	75.05	89.88	94.78	97.27	0.25
	SD	2.05	2.89	3.78	3.68	0.66	3.66	3.78	2.90	2.05	3.00

Tabela 4.33: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição T1 e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC.

T1 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.90	5.25	9.76	23.86	49.86	75.77	90.08	94.61	97.00	1.14
	SD	4.59	6.31	8.36	10.12	8.12	10.35	8.71	6.74	5.12	7.84
$R^{(2)}_{BOOT}$	ME	2.80	5.05	9.44	23.42	50.00	76.43	90.59	94.94	97.20	1.58
	SD	4.38	5.84	7.54	8.17	1.63	8.21	7.57	5.89	4.42	6.31
$R^{(3)}_{BOOT}$	ME	2.98	5.31	9.78	23.79	50.00	76.07	90.25	94.68	97.02	1.21
	SD	4.53	6.01	7.69	8.19	1.63	8.23	7.72	6.07	4.58	6.42
T1 $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.87	5.41	10.31	25.00	49.62	74.53	89.37	94.45	97.06	0.63
	SD	3.16	4.50	6.15	8.06	8.18	8.14	6.23	4.43	3.00	6.09
$R^{(2)}_{BOOT}$	ME	2.72	5.19	9.99	24.70	49.94	75.32	89.97	94.85	97.29	0.32
	SD	2.60	3.70	4.88	5.00	0.92	5.03	4.92	3.68	2.60	3.94
$R^{(3)}_{BOOT}$	ME	2.82	5.33	10.18	24.89	49.94	75.12	89.77	94.70	97.20	0.33
	SD	2.67	3.76	4.93	5.00	0.92	5.02	4.98	3.74	2.65	3.98
T1 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.76	5.29	10.30	25.19	49.95	74.82	89.78	94.80	97.32	0.30
	SD	2.43	3.56	5.07	7.03	7.42	6.65	4.66	3.22	2.12	5.05
$R^{(2)}_{BOOT}$	ME	2.55	4.97	9.88	24.78	50.02	75.23	90.14	95.04	97.45	0.23
	SD	1.80	2.61	3.51	3.51	0.74	3.54	3.51	2.64	1.81	2.80
$R^{(3)}_{BOOT}$	ME	2.59	5.04	9.97	24.87	50.02	75.14	90.05	94.98	97.41	0.14
	SD	1.83	2.64	3.53	3.51	0.74	3.54	3.53	2.66	1.84	2.81

Tabela 4.34: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $c = 1.5$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC.

NAC–0.1 $n = 20$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(IF)}_{BOOT}$	ME	3.29	5.17	8.97	22.26	49.33	78.92	94.63	98.56	99.80	4.63
	SD	9.13	10.50	12.11	13.65	10.54	5.76	2.77	1.23	0.40	8.69
$R^{(1)}_{BOOT}$	ME	4.23	6.42	10.62	23.83	48.83	76.51	92.68	97.46	99.36	2.68
	SD	8.79	9.91	11.19	11.97	8.75	5.12	3.85	2.93	2.22	7.99
$R^{(2)}_{BOOT}$	ME	2.53	4.31	7.95	20.38	45.26	73.86	91.16	96.58	98.93	4.74
	SD	4.48	5.07	5.82	6.01	1.95	4.62	5.47	4.54	3.46	4.75
$R^{(3)}_{BOOT}$	ME	3.27	5.16	8.88	21.27	45.79	73.91	91.00	96.40	98.79	4.21
	SD	6.27	6.98	7.73	7.67	3.25	4.38	5.48	4.67	3.63	5.79
NAC–0.1 $n = 50$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(IF)}_{BOOT}$	ME	2.90	5.14	9.74	24.98	52.44	79.63	93.76	97.58	99.12	4.63
	SD	7.49	8.35	9.41	10.31	7.94	4.30	2.09	1.11	0.53	6.73
$R^{(1)}_{BOOT}$	ME	3.37	6.07	11.20	26.52	52.18	77.65	92.03	96.48	98.54	2.65
	SD	4.19	5.37	6.69	7.78	6.25	3.77	2.18	1.35	0.77	4.85
$R^{(2)}_{BOOT}$	ME	2.35	4.63	9.18	23.58	48.98	75.31	90.78	95.76	98.13	1.42
	SD	1.29	1.90	2.54	2.64	0.77	2.54	2.58	1.87	1.22	2.04
$R^{(3)}_{BOOT}$	ME	2.55	4.88	9.47	23.87	49.14	75.30	90.70	95.68	98.07	1.13
	SD	2.02	2.68	3.31	3.37	1.30	2.50	2.63	1.97	1.34	2.45
NAC–0.1 $n = 100$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(IF)}_{BOOT}$	ME	2.29	4.72	9.69	25.59	52.77	79.06	93.08	96.98	98.73	4.06
	SD	1.62	2.59	3.85	5.54	5.10	3.17	1.64	0.96	0.52	3.24
$R^{(1)}_{BOOT}$	ME	3.11	5.89	11.20	26.93	52.59	77.56	91.72	96.09	98.18	2.59
	SD	2.36	3.37	4.54	5.72	4.80	2.90	1.63	1.04	0.63	3.43
$R^{(2)}_{BOOT}$	ME	2.40	4.79	9.53	24.36	49.68	75.39	90.56	95.39	97.78	0.64
	SD	0.95	1.41	1.88	1.92	0.67	1.90	1.88	1.38	0.90	1.51
$R^{(3)}_{BOOT}$	ME	2.45	4.85	9.62	24.44	49.72	75.37	90.52	95.36	97.76	0.56
	SD	1.04	1.53	2.02	2.05	0.71	1.91	1.92	1.43	0.94	1.58

Tabela 4.35: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC.

NAC–0.1 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R_{BOOT}^{(IF)}$	ME	6.89	8.93	12.35	22.62	46.61	80.71	97.94	99.55	99.87	7.94
	SD	14.03	16.09	18.47	20.99	17.23	6.39	3.13	1.56	0.61	13.28
$R_{BOOT}^{(1)}$	ME	7.19	9.30	12.91	23.49	46.47	78.67	97.04	99.23	99.72	7.04
	SD	13.61	15.22	17.10	18.99	15.34	5.71	4.09	2.40	1.19	12.30
$R_{BOOT}^{(2)}$	ME	2.90	4.09	6.43	14.89	36.90	71.84	94.62	98.35	99.40	13.10
	SD	6.49	7.41	8.48	9.48	5.84	5.17	7.28	4.18	2.01	6.62
$R_{BOOT}^{(3)}$	ME	5.08	6.57	9.23	17.93	39.13	72.30	94.22	97.97	99.18	10.87
	SD	9.66	10.76	12.00	13.00	8.62	4.67	7.41	4.58	2.39	8.82
NAC–0.1 $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R_{BOOT}^{(IF)}$	ME	7.99	11.37	17.01	32.63	60.18	87.75	98.74	99.87	99.98	12.75
	SD	14.95	17.35	20.03	22.43	17.33	6.06	0.88	0.34	0.18	14.01
$R_{BOOT}^{(1)}$	ME	8.86	12.09	17.49	32.31	58.19	85.11	97.82	99.63	99.90	10.11
	SD	15.25	16.97	18.76	19.87	14.93	5.31	1.26	0.80	0.47	13.00
$R_{BOOT}^{(2)}$	ME	2.22	3.70	6.79	17.76	42.71	74.94	94.24	98.34	99.40	7.29
	SD	4.17	4.94	5.81	6.31	2.66	4.26	5.26	3.52	2.10	4.54
$R_{BOOT}^{(3)}$	ME	4.44	6.34	9.85	21.13	45.22	75.64	94.02	98.10	99.24	4.78
	SD	8.18	9.22	10.29	10.66	6.15	3.18	4.91	3.55	2.28	7.16
NAC–0.1 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R_{BOOT}^{(IF)}$	ME	8.24	12.57	19.85	39.07	67.66	90.91	99.00	99.89	99.99	17.66
	SD	13.53	16.27	19.18	21.27	15.55	5.42	0.71	0.10	0.02	13.08
$R_{BOOT}^{(1)}$	ME	10.40	14.77	21.84	39.77	66.32	89.22	98.44	99.76	99.97	16.32
	SD	15.31	17.51	19.69	20.57	14.77	5.34	0.78	0.14	0.04	13.33
$R_{BOOT}^{(2)}$	ME	1.62	3.17	6.55	18.86	45.16	76.53	94.44	98.46	99.59	6.14
	SD	1.74	2.44	3.31	3.98	1.45	3.11	3.16	1.87	0.92	2.62
$R_{BOOT}^{(3)}$	ME	3.61	5.69	9.68	22.54	47.95	77.42	94.33	98.29	99.48	4.33
	SD	5.95	7.12	8.39	9.08	5.44	2.04	2.75	1.81	0.98	5.62

Tabela 4.36: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN1 e $c = 1.5$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC.

LOGN1 $n = 20$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.57	4.04	7.07	19.36	53.35	87.16	97.42	99.12	99.66	12.16
	SD	4.57	5.97	7.95	11.42	6.95	6.35	3.00	1.51	0.76	6.24
$R^{(2)}_{BOOT}$	ME	1.37	2.26	4.21	13.11	43.75	82.44	96.13	98.60	99.44	11.89
	SD	2.63	3.60	5.12	8.20	3.35	7.34	4.18	2.24	1.19	4.74
$R^{(3)}_{BOOT}$	ME	1.66	2.68	4.83	14.17	44.10	81.56	95.61	98.33	99.30	10.83
	SD	3.00	4.01	5.54	8.41	3.29	7.38	4.51	2.54	1.42	4.95
LOGN1 $n = 50$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.16	4.22	8.76	26.25	61.52	89.87	98.26	99.50	99.85	14.87
	SD	2.81	4.30	6.52	9.26	5.69	3.74	1.52	0.69	0.32	4.76
$R^{(2)}_{BOOT}$	ME	0.70	1.56	3.79	14.88	46.49	82.28	96.31	98.83	99.62	10.12
	SD	1.07	1.78	3.05	5.35	1.26	4.79	2.70	1.30	0.62	2.91
$R^{(3)}_{BOOT}$	ME	0.91	1.92	4.40	15.87	46.67	81.41	95.78	98.58	99.50	9.13
	SD	1.27	2.05	3.37	5.48	1.25	4.86	2.97	1.51	0.74	3.05
LOGN1 $n = 100$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.46	5.18	11.34	33.02	68.74	92.66	98.95	99.74	99.93	18.74
	SD	2.69	4.23	6.37	8.41	5.27	2.35	0.75	0.29	0.10	4.35
$R^{(2)}_{BOOT}$	ME	0.50	1.27	3.53	15.44	47.62	82.46	96.51	98.97	99.69	9.56
	SD	0.65	1.21	2.22	3.78	0.85	3.44	2.04	0.97	0.43	2.08
$R^{(3)}_{BOOT}$	ME	0.66	1.59	4.13	16.44	47.75	81.57	95.97	98.72	99.58	8.56
	SD	0.80	1.42	2.46	3.88	0.85	3.51	2.26	1.14	0.54	2.20

Tabela 4.37: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN1 e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem nas linhas ímpares a KS e nas linhas pares a SC.

LOGN1 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.84	4.51	7.82	20.36	52.34	85.05	96.65	98.81	99.55	10.05
	SD	4.66	6.02	8.01	11.20	6.95	6.11	3.17	1.66	0.86	6.21
$R^{(2)}_{BOOT}$	ME	1.59	2.68	5.01	14.83	44.73	80.93	95.34	98.24	99.30	10.17
	SD	2.74	3.72	5.24	8.14	2.96	6.94	4.26	2.39	1.32	4.68
$R^{(3)}_{BOOT}$	ME	1.87	3.07	5.56	15.63	44.96	80.30	94.90	97.99	99.16	9.37
	SD	3.07	4.10	5.63	8.33	2.97	6.98	4.55	2.66	1.54	4.89
LOGN1 $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.41	4.55	9.26	26.41	59.50	87.61	97.45	99.19	99.74	12.61
	SD	3.20	4.64	6.65	8.71	5.88	4.05	1.83	0.87	0.40	4.80
$R^{(2)}_{BOOT}$	ME	0.95	1.99	4.60	16.51	47.08	80.80	95.42	98.40	99.43	8.49
	SD	1.32	2.12	3.42	5.26	1.15	4.57	2.97	1.59	0.79	2.98
$R^{(3)}_{BOOT}$	ME	1.13	2.29	5.09	17.26	47.21	80.17	94.98	98.17	99.32	7.74
	SD	1.52	2.36	3.66	5.31	1.15	4.58	3.18	1.79	0.93	3.09
LOGN1 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.69	5.45	11.48	31.81	65.37	90.16	98.16	99.47	99.84	15.37
	SD	2.62	4.08	6.02	7.86	5.59	3.13	1.15	0.47	0.19	4.27
$R^{(2)}_{BOOT}$	ME	0.73	1.75	4.50	17.17	47.97	80.92	95.51	98.52	99.49	7.83
	SD	0.76	1.36	2.46	3.86	0.82	3.41	2.19	1.09	0.51	2.16
$R^{(3)}_{BOOT}$	ME	0.87	2.02	4.97	17.86	48.05	80.30	95.10	98.31	99.39	7.14
	SD	0.88	1.53	2.63	3.89	0.82	3.41	2.33	1.22	0.59	2.23

Exemplo 4.6 Aplicou-se o *bootstrap* IF com as várias correcções atrás descritas ao conjunto de dados do Exemplo 4.2 da Secção 4.3.1. A Tabela 4.38 mostra os resultados obtidos com $B = 50000$ reamostras *bootstrap*. Como termo de comparação ainda se apresentam os resultados do *bootstrap* usual, U e de IF também com $B = 50000$. A notação utilizada é a do estudo de simulação anterior.

Tabela 4.38: Percentis das várias distribuições *bootstrap* de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$ para os dados da tensão sanguínea, com $c = 1.5$.

p	R^U_{BOOT}	R^{IF}_{BOOT}	$R^{(1)}_{BOOT}$	$R^{(2)}_{BOOT}$	$R^{(3)}_{BOOT}$
2.5	-10.71	-6.14	-4.71	-5.35	-5.13
5	-8.91	-4.94	-3.57	-4.21	-3.90
10	-6.41	-3.54	-2.57	-3.21	-2.82
25	-3.01	-1.34	-0.99	-1.63	-1.12
50	0.48	0.36	0.72	0.08	0.73
75	3.68	1.86	2.29	1.65	2.42
90	5.78	3.06	3.72	3.08	3.96
95	6.78	3.76	4.43	3.79	4.73
97.5	7.48	4.36	5.00	4.37	5.35

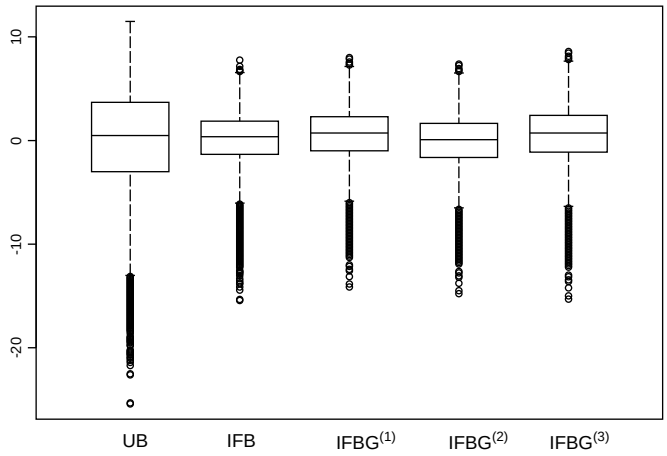


Figura 4.17: Comparação entre os vários métodos *bootstrap* na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$ para os dados da tensão sanguínea, excepto para R^U_{BOOT} onde a distribuição em causa é a de $\bar{X}_n - \mu$ (ou $\mu(F_n) - \mu(F)$).

Na Figura 4.17 apresentam-se os vários diagramas de extremos-e-quartis associados a cada um dos métodos *bootstrap* utilizados, UB (*bootstrap* usual), IFB (*bootstrap* IF) e $IFBG^{(k)}$, $k = 1, 2, 3$

Tabela 4.39: Intervalos de confiança a 95% para o valor médio da tensão sanguínea.

Método	Limite inferior	Limite superior
UB	71.5	89.7
IFB	80.1	90.6
$IFBG^{(1)}$	81.5	91.2
$IFBG^{(2)}$	81.8	90.6
$IFBG^{(3)}$	81.1	91.6
TI_{-40}	81.8	91.9
MB	79.7	92.0

(*bootstrap* IF generalizado). Nota-se que as distribuições associadas aos métodos *bootstrap* corrigidos possuem caudas bastantes inferiores às do método UB e ligeiramente inferiores às do método IFB.

As estimativas pontuais obtidas para este conjunto de dados são: $\bar{x} = 82.20$, sem a observação discordante (40) obtém-se $\bar{x}_{-40} = 86.89$ e $\mu_{\sqrt{w}}(F_n) = 86.21$ ($c = 1.5$).

Apresentam-se ainda, na Tabela 4.39, os intervalos de confiança, a 95%, para o valor médio da tensão sanguínea com base em cada um dos métodos referidos. De notar que o intervalo aqui apresentado para o método IFB é diferente do apresentado na Tabela 4.3 porque se baseia numa constante de afinação e numa estimativa pontual diferentes. Todos os métodos, à excepção do UB, fornecem intervalos bastante próximos do intervalo-t para a amostra sem observação atípica, mas os mais próximos são os obtidos pelos métodos $IFBG^{(1)}$ e $IFBG^{(3)}$.

■

Para finalizar esta secção apresenta-se um pequeno estudo de simulação realizado com o objectivo de comparar o *bootstrap* robusto proposto por Salibian-Barrera (2000) com o *bootstrap* generalizado, IFBG. Utilizou-se um delineamento semelhante ao desse autor (Salibian-Barrera, 2000, Sec. 3.6.2), ou seja, geraram-se amostras de dimensão $n = 20, 30$ e 50 provenientes da população $(1-\varepsilon)N(0,1)+\varepsilon N(-7, 0.1)$, com $\varepsilon = 0, 0.1, 0.2, 0.3$ e determinaram-se intervalos de confiança para a localização, μ, segundo os seguintes métodos:

- Método do percentil com $IFBG$, $c = 1.5; 2$ e $RESIF(x) = |x - \hat{\mu}_{LTS}|/\hat{\sigma}_{LTS}$;
- *Bootstrap* Robusto Studentized (SRB) (Resultados de Salibian-Barrera, 2000, pág. 129);
- *Bootstrap* Winsorized (WB) (Resultados de Salibian-Barrera, 2000, pág. 129).

Para cada combinação (n, ε) geraram-se 1000 amostras[4]. Cada intervalo *bootstrap* foi obtido com $B = 5000$ réplicas. As Tabelas 4.40 e 4.41 mostram as probabilidades de cobertura obtidas pelo

[4]Por limitações de tempo não se gerou um número de amostras igual ao do estudo de Salibian-Barrera (2000), que foi 3000.

método $IFBG$ para dois valores de c (1.5 e 2). Apresentam-se ainda os resultados obtidos quando se censuraram as reamostras *bootstrap* que possuíam mais de 50% de contaminação. Optou-se por efectuar esta censura já que nestas situações não existe nenhum método estatístico (inclusive robusto) que permita dar uma resposta correcta. Nestas tabelas são ainda reportados os resultados do estudo efectuado por Salibian-Barrera (2000) (métodos SRB e WB).

Tabela 4.40: Cobertura e (comprimento médio) de intervalos de confiança para a localização univariada a 95%. A negrito encontram-se os valores que apresentam diferenças estatisticamente significativas para 0.95.

		$IFBG$		Simulação censurada			
n	ε	$c = 2$	$c = 1.5$	$c = 2$	$c = 1.5$	SRB	WB
20	0.0	**0.922** (0.83)	**0.915** (0.85)	–	–	0.942 (1.01)	**0.938** (0.91)
	0.1	0.944 (1.14)	**0.923** (0.95)	–	–	0.950 (1.25)	**0.937** (1.11)
	0.2	0.955 (1.54)	**0.927** (1.13)	0.948 (1.58)	0.935 (1.12)	**0.962** (1.74)	**0.958** (1.49)
	0.3	**0.920** (2.08)	**0.890** (1.36)	0.954 (2.08)	0.938 (1.33)	**0.970** (2.56)	**0.972** (2.11)
30	0.0	0.939 (0.70)	0.930 (0.70)	–	–	0.949 (0.79)	0.948 (0.74)
	0.1	0.964 (0.93)	0.942 (0.79)	–	–	**0.958** (0.97)	0.950 (0.90)
	0.2	0.959 (1.29)	0.934 (0.90)	–	–	**0.964** (1.28)	**0.963** (1.18)
	0.3	0.961 (1.78)	0.933 (1.08)	0.975 (1.78)	0.951 (1.06)	**0.971** (2.01)	**0.971** (1.73)
50	0.0	0.941 (0.55)	0.943 (0.55)	–	–	0.946 (0.59)	0.942 (0.57)
	0.1	0.956 (0.70)	0.954 (0.60)	–	–	**0.958** (0.72)	**0.958** (0.70)
	0.2	**0.974** (0.98)	0.952 (0.71)	–	–	**0.967** (0.94)	**0.966** (0.91)
	0.3	**0.978** (1.41)	0.961 (0.83)	–	–	**0.980** (1.43)	**0.982** (1.32)

Tabela 4.41: Cobertura e (comprimento médio) de intervalos de confiança para a localização univariada a 99%. A negrito encontram-se os valores que apresentam diferenças estatisticamente significativas para 0.99.

		$IFBG$		Simulação censurada			
n	ε	$c = 2$	$c = 1.5$	$c = 2$	$c = 1.5$	SRB	WB
20	0.0	**0.973** (1.09)	**0.973** (1.11)	–	–	0.988 (1.43)	**0.968** (1.12)
	0.1	0.985 (1.52)	0.983 (1.25)	–	–	0.991 (1.81)	**0.983** (1.42)
	0.2	0.985 (2.06)	**0.973** (1.50)	0.978 (2.12)	0.981 (1.48)	**0.994** (2.52)	**0.985** (1.90)
	0.3	**0.948** (2.78)	**0.928** (1.82)	0.983 (2.78)	**0.978** (1.78)	**0.998** (3.80)	**0.995** (2.63)
30	0.0	0.982 (0.91)	0.986 (0.92)	–	–	0.988 (1.08)	**0.979** (0.93)
	0.1	0.989 (1.23)	0.982 (1.04)	–	–	0.993 (1.33)	**0.981** (1.14)
	0.2	0.989 (1.72)	**0.978** (1.19)	–	–	**0.997** (1.77)	0.988 (1.50)
	0.3	0.981 (2.38)	**0.966** (1.43)	0.995 (2.38)	0.984 (1.41)	**0.999** (2.68)	**0.995** (2.14)
50	0.0	0.986 (0.71)	0.990 (0.72)	–	–	0.989 (0.80)	**0.979** (0.73)
	0.1	0.992 (0.93)	0.991 (0.79)	–	–	0.992 (0.97)	0.991 (0.89)
	0.2	0.994 (1.30)	0.992 (0.93)	–	–	**0.999** (1.27)	0.993 (1.16)
	0.3	0.994 (1.88)	0.991 (1.10)	–	–	**0.999** (1.97)	**0.997** (1.70)

Quando se comparam os resultados, reportados nas Tabelas 4.40 e 4.41, verifica-se que a cobertura do método $IFBG$ é similar ao do WB sendo apenas pior que SRB para $n = 20$. Em termos de comprimento médio verifica-se que o método $IFBG$ com $c = 1.5$ conduziu a resultados melhores quer relativamente ao $IFBG$ com $c = 2$ quer aos métodos WB e SRB.

4.4.2 Coeficiente de correlação linear

Nesta secção repete-se o estudo da Secção 4.3.2 relativo ao comportamento do *bootstrap* IF na construção de intervalos de confiança para o coeficiente de correlação linear ρ mas aplicando as correcções anteriores para estimar a distribuição de $\rho_{\sqrt{w}}(F_n)$). Optou-se por, a título de experiência, usar as expressões obtidas para o caso da média univariada (embora sabendo que o mais correcto seria determinar expressões análogas para esta situação concreta). Os resultados obtidos, no estudo Monte Carlo descrito em seguida, foram encorajadores.

Delineamento do estudo de Monte Carlo

Foram consideradas novamente as distribuições NOR, NSC–0.25 e NAC–0.1 (descritas na Secção 4.3.2, pág. 101).

Utilizaram-se amostras de dimensão $n = 20, 100$ e fixou-se $\rho = 0.5$. Para constante de afinação escolheram-se três valores, $c = 2, 5$ e 10.

Para aplicar o método *bootstrap* IF generalizado seguem-se praticamente todos os passos já descritos para o *bootstrap* IF. Isto é, dada uma amostra bivariada de dimensão n, $((x_{11}, x_{12}), (x_{21}, x_{22}), \cdots, (x_{n1}, x_{n2}))$, proveniente de uma população bivariada desconhecida, F, e $\rho(F_n)$ o coeficiente de correlação amostral de Pearson, determina-se a versão empírica e padronizada da sua função de influência, $RESIF$ (expressão 4.17), e com ela definem-se as probabilidades de reamostragem partindo dos pesos w_i dados por (4.18). Neste caso escolheram-se para parâmetros da função ψ, $\gamma = \infty$, $d = c$ e os valores de c já referidos. Para efectuar a correcção do método IFBG reamostram--se, em cada réplica *bootstrap*, $\hat{n}_{novo}$ pontos de n, com $\hat{n}_{novo}$ dado por (4.41). Note-se que, também aqui, está em consideração a distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$ onde $\rho_{\sqrt{w}}(F_n)$ é o estimador que faz a ligação entre o modelo ponderado e o modelo original, e é dado por

$$\rho_{\sqrt{w}}(F_n) = \frac{\sum_{i=1}^{n} w_i (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n} w_i (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^{n} w_i (x_{i2} - \hat{\mu}_2)^2}},$$

com $\hat{\mu}_j = (\sum_{i=1}^{n} w_i x_{ij})/(\sum_{i=1}^{n} w_i)$, $j = 1, 2$.

Obtidas as reamostras *bootstrap* a partir da selecção com reposição de $\hat{n}_{novo}$ pontos amostrais, foram determinadas as seguintes aproximações Monte Carlo das distribuições *bootstrap* de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$:

1. $R^{(1)}_{BOOT}$:

$$R^{(1)}_{BOOT}(x) = \frac{1}{B}\sum_{b=1}^{B} I\left\{\left(\rho(F^*_{w_n,\hat{n}_{novo}})^b - \rho_{\sqrt{w}}(F_n)\right) \leq x\right\}, \tag{4.46}$$

onde B é o número de réplicas *bootstrap* (neste caso $B = 5000$), $\rho(F^*_{w_n,\hat{n}_{novo}})^b$ é a avaliação de R para a b-ésima amostra *bootstrap* IF, $\rho_{\sqrt{w}}(F_n)$ é a estimativa ponderada para a amostra original. Esta distribuição empírica está centrada em $\rho_{\sqrt{w}}(F_n)$;

2. $R^{(2)}_{BOOT}$:

$$R^{(2)}_{BOOT}(x) = \frac{1}{B}\sum_{b=1}^{B} I\left\{\left(\rho(F^*_{w_n,\hat{n}_{novo}})^b - \rho_{w}(F_n)\right) \leq x\right\}, \tag{4.47}$$

neste caso centrada em $\rho_w(F_n)$;

3. $R^{(k)}_{BOOT}, \quad k = 3, 4, \cdots, 12$:

$$R^{(k)}_{BOOT}(x) = \frac{1}{B}\sum_{b=1}^{B} I\left\{\left(\rho(F^*_{w_n,\hat{n}_{novo}})^b - \rho^*\right) \times \sqrt{n/\hat{n}_{novo}} \times \sqrt{fac_k} + \rho^* - \rho_w(F_n) \leq x\right\}, \tag{4.48}$$

onde ρ^* é a aproximação Monte Carlo do estimador bootstrap ou seja, a média dos B valores de $\rho(F^*_{w_n,\hat{n}_{novo}})$ e o factor de correcção fac_k é dado por

$$\left\{[Var(\rho(F_{w_n,n}))]^B_{Boot} + n^{-1}a_k d^2_{est}\right\} \Big/ [Var(\rho(F_{w_n,n}))]^B_{Boot},$$

$[Var(\rho(F_{w_n,n}))]^B_{Boot}$ é o estimador *bootstrap* da variância do estimador coeficiente de correlação usual na amostra ponderada, a_k são constantes que variam de 25 a 250 com incrementos de 25 e por último $d^2_{est} = \rho_{\sqrt{w}}(F_n) - \rho_w(F_n)$.

Para cada combinação, dimensão e distribuição, foram geradas 500 amostras e para cada uma delas aplicou-se o *bootstrap* IF generalizado com $B = 5000$ réplicas.

Tal como no estudo anterior também se estimou a distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$ através de um estudo de simulação independente com 5000 amostras de cada uma das três distribuições acima descritas com cada uma das dimensões consideradas ($n = 20, 100$) e calculou-se a estimativa de interesse, coeficiente de correlação ponderado, para cada uma. Esta amostra de 5000 observações foi depois centrada na sua média aritmética. Em seguida calcularam-se os percentis de ordem $p : 2.5, 5, 10, 25, 50, 75, 90, 95$ e 97.5 %. Este processo foi repetido 20 vezes e no fim determinou-se a mediana de cada um dos percentis.

Na fase do estudo relativa ao *bootstrap* obtiveram-se, para cada ordem p, os indicadores já descritos na pág. 153, ME_p e SD_p. Como medidas globais foram também calculados valores de KS, DA, SQ e SC. Os resultados obtidos encontram-se nas Tabelas 4.42 a 4.47 e nas Tabelas B.31 a B.39 do Apêndice B.

Verifica-se que para as distribuições, NOR e NAC–0.1 os resultados são melhores para a constante de afinação $c = 5$. Pelo contrário para a contaminação simétrica, NSC–0.25, é a constante $c = 10$

Tabela 4.42: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NOR e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

						NOR $n = 20$ $c = 2$					
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	1.02	2.78	7.34	24.50	58.81	85.17	94.09	96.38	97.55	10.17
	SD	2.54	4.42	7.64	11.94	9.38	9.31	5.99	4.46	3.48	7.22
$R^{(2)}_{BOOT}$	ME	0.76	2.25	6.43	23.81	59.52	85.75	94.05	96.15	97.21	10.75
	SD	1.64	2.93	5.23	8.20	4.80	8.83	6.41	5.19	4.37	5.71
$R^{(3)}_{BOOT}$	ME	0.90	2.48	6.82	24.13	59.24	85.27	93.68	95.88	97.00	10.27
	SD	1.93	3.31	5.71	8.48	4.95	9.01	6.63	5.43	4.62	5.96
						NOR $n = 20$ $c = 5$					
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.64	5.14	9.83	24.88	54.09	80.46	91.87	95.14	96.84	5.46
	SD	3.27	4.78	6.53	7.90	4.73	8.53	6.63	5.22	4.19	5.99
$R^{(2)}_{BOOT}$	ME	2.49	4.96	9.61	24.63	53.92	80.43	91.88	95.13	96.82	5.43
	SD	2.62	4.04	5.68	6.85	2.57	7.87	6.32	5.03	4.07	5.30
$R^{(3)}_{BOOT}$	ME	2.55	5.03	9.68	24.69	53.89	80.31	91.77	95.04	96.74	5.31
	SD	2.78	4.18	5.80	6.91	2.70	7.98	6.42	5.14	4.19	5.40
						NOR $n = 20$ $c = 10$					
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.79	5.02	9.50	24.10	53.45	80.97	92.47	95.57	97.11	5.97
	SD	3.06	4.36	6.01	7.29	2.51	7.68	6.21	4.83	3.80	5.37
$R^{(2)}_{BOOT}$	ME	2.72	4.92	9.38	23.93	53.25	80.79	92.34	95.48	97.03	5.79
	SD	2.78	4.02	5.65	6.88	1.90	7.73	6.31	4.92	3.90	5.23
$R^{(3)}_{BOOT}$	ME	2.74	4.94	9.39	23.94	53.25	80.78	92.33	95.46	97.02	5.78
	SD	2.82	4.06	5.69	6.90	1.91	7.74	6.33	4.95	3.94	5.25

que permite obter melhores resultados. Em qualquer um destes casos observa-se que os valores correspondentes aos percentis extremos estão mais próximos dos correctos do que os valores correspondentes aos percentis centrais. Este aspecto é importante, tanto para a construção de regiões de confiança como para a realização de testes de hipóteses.

Para $n = 20$ e escolhendo a constante de afinação $c = 5$ verifica-se que a melhor opção é usar a distribuição $R^{(2)}_{BOOT}$ independentemente do modelo gerador dos dados. De qualquer modo, não há grande diferença entre esta e as restantes $R^{(k)}_{BOOT}$, com $k > 1$. Também para $n = 100$ (e fixando o mesmo valor de c) se verifica que é a distribuição empírica $R^{(2)}_{BOOT}$ a que tem melhor comportamento.

Tabela 4.43: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NOR e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

						NOR $n = 100$ $c = 2$					
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	1.74	3.83	8.46	23.99	51.77	78.20	92.26	96.29	98.16	3.20
	SD	1.37	2.33	3.72	5.53	5.53	4.75	3.27	2.20	1.43	3.69
$R^{(2)}_{BOOT}$	ME	1.63	3.66	8.25	23.88	52.03	78.55	92.45	96.39	98.20	3.55
	SD	0.98	1.64	2.51	2.87	0.78	3.11	2.85	2.04	1.37	2.18
$R^{(3)}_{BOOT}$	ME	1.65	3.69	8.30	23.92	52.03	78.49	92.40	96.35	98.17	3.49
	SD	1.00	1.66	2.53	2.89	0.79	3.13	2.87	2.06	1.38	2.19
						NOR $n = 100$ $c = 5$					
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.64	5.07	9.92	25.28	51.12	76.56	90.83	95.47	97.64	1.56
	SD	1.48	2.17	2.96	3.23	1.35	3.32	3.24	2.40	1.67	2.54
$R^{(2)}_{BOOT}$	ME	2.62	5.05	9.88	25.22	51.06	76.51	90.80	95.45	97.63	1.51
	SD	1.43	2.09	2.85	3.05	0.88	3.27	3.23	2.41	1.67	2.46
$R^{(3)}_{BOOT}$	ME	2.62	5.05	9.88	25.22	51.06	76.51	90.80	95.45	97.63	1.51
	SD	1.43	2.10	2.85	3.05	0.88	3.27	3.24	2.41	1.67	2.46
						NOR $n = 100$ $c = 10$					
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.53	4.89	9.80	24.95	51.14	76.72	90.97	95.48	97.65	1.72
	SD	1.42	2.10	2.89	3.05	0.88	3.13	3.10	2.32	1.60	2.41
$R^{(2)}_{BOOT}$	ME	2.53	4.89	9.80	24.95	51.14	76.72	90.97	95.48	97.65	1.72
	SD	1.42	2.09	2.88	3.04	0.87	3.13	3.10	2.32	1.60	2.41
$R^{(3)}_{BOOT}$	ME	2.53	4.89	9.80	24.95	51.14	76.72	90.97	95.48	97.65	1.72
	SD	1.42	2.09	2.88	3.04	0.87	3.13	3.10	2.32	1.60	2.41

Tabela 4.44: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NSC–0.25 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NSC–0.25 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	0.64	1.61	4.62	19.98	61.19	88.86	95.53	97.15	97.90	13.86
	SD	1.31	2.46	4.87	10.20	10.76	9.76	6.07	4.64	3.87	6.83
$R^{(2)}_{BOOT}$	ME	0.56	1.45	4.32	19.62	62.12	89.43	95.73	97.22	97.94	14.43
	SD	1.05	1.95	3.92	8.07	7.29	8.80	5.69	4.48	3.77	5.60
$R^{(3)}_{BOOT}$	ME	0.62	1.57	4.54	19.90	61.77	89.00	95.47	97.03	97.78	14.00
	SD	1.16	2.11	4.12	8.17	7.33	8.98	5.86	4.65	3.92	5.73
NSC–0.25 $n = 20$ $c = 5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.15	4.38	8.78	23.52	55.73	83.36	92.68	95.24	96.56	8.36
	SD	3.65	5.49	8.20	12.19	11.55	13.00	8.66	6.64	5.41	8.88
$R^{(2)}_{BOOT}$	ME	1.87	3.92	8.12	22.88	56.32	84.50	93.45	95.71	96.81	9.50
	SD	2.84	4.32	6.36	9.17	5.84	10.46	7.36	5.90	5.02	6.74
$R^{(3)}_{BOOT}$	ME	2.09	4.23	8.53	23.24	56.03	83.81	92.94	95.34	96.52	8.81
	SD	3.08	4.59	6.71	9.42	6.25	10.81	7.64	6.16	5.26	7.02
NSC–0.25 $n = 20$ $c = 10$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.59	4.76	8.99	22.90	55.68	83.17	91.96	94.39	95.68	8.17
	SD	3.69	5.54	8.07	10.93	9.04	12.97	10.30	8.76	7.66	8.95
$R^{(2)}_{BOOT}$	ME	2.47	4.56	8.71	22.66	55.66	83.54	92.43	94.89	96.19	8.54
	SD	3.40	4.92	7.05	9.42	5.97	11.39	8.74	7.20	6.19	7.50
$R^{(3)}_{BOOT}$	ME	2.60	4.74	8.93	22.82	55.55	83.26	92.16	94.65	95.96	8.26
	SD	3.55	5.12	7.29	9.56	6.08	11.55	9.05	7.52	6.51	7.71

Tabela 4.45: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NSC–0.25 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NSC–0.25 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	0.46	1.33	4.03	17.85	52.01	84.68	96.27	98.57	99.40	9.68
	SD	0.61	1.35	2.88	6.70	8.23	6.67	3.28	1.81	1.02	4.51
$R^{(2)}_{BOOT}$	ME	0.40	1.21	3.77	17.34	52.02	85.22	96.51	98.70	99.45	10.22
	SD	0.45	1.01	2.16	4.29	1.33	4.55	2.65	1.47	0.82	2.51
$R^{(3)}_{BOOT}$	ME	0.42	1.24	3.83	17.44	51.99	85.08	96.44	98.66	99.43	10.08
	SD	0.46	1.03	2.18	4.30	1.35	4.56	2.68	1.50	0.84	2.52
NSC–0.25 $n = 100$ $c = 5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	1.74	3.75	8.21	23.27	51.49	78.84	92.59	96.46	98.15	3.84
	SD	1.52	2.67	4.40	7.25	8.40	7.78	5.13	3.34	2.18	5.32
$R^{(2)}_{BOOT}$	ME	1.60	3.51	7.82	22.73	51.39	79.35	93.07	96.77	98.34	4.35
	SD	1.15	2.00	3.19	4.07	1.17	4.51	3.79	2.56	1.70	2.94
$R^{(3)}_{BOOT}$	ME	1.64	3.57	7.92	22.83	51.37	79.22	92.96	96.69	98.29	4.22
	SD	1.17	2.02	3.21	4.08	1.21	4.52	3.82	2.59	1.73	2.96
NSC–0.25 $n = 100$ $c = 10$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.71	5.23	10.16	25.42	51.48	77.10	90.77	95.07	97.22	2.10
	SD	2.16	3.32	4.82	6.51	6.67	7.39	5.62	4.08	2.90	5.13
$R^{(2)}_{BOOT}$	ME	2.60	5.05	9.88	25.05	51.28	77.30	91.12	95.35	97.45	2.30
	SD	1.83	2.79	3.92	4.35	1.49	4.76	4.34	3.18	2.25	3.40
$R^{(3)}_{BOOT}$	ME	2.63	5.10	9.93	25.10	51.28	77.23	91.05	95.30	97.41	2.23
	SD	1.85	2.81	3.93	4.36	1.51	4.77	4.37	3.22	2.27	3.42

Tabela 4.46: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NAC–0.1 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	1.20	2.95	7.69	24.69	58.11	85.04	94.04	96.42	97.61	10.04
	SD	2.12	3.85	6.80	10.43	7.51	8.48	5.71	4.32	3.39	6.37
$R^{(2)}_{BOOT}$	ME	0.97	2.54	7.00	23.90	58.10	85.03	93.79	96.13	97.31	10.03
	SD	1.41	2.71	5.05	7.62	3.86	8.41	6.15	4.84	3.98	5.33
$R^{(3)}_{BOOT}$	ME	1.07	2.70	7.24	24.13	57.98	84.73	93.57	95.95	97.16	9.73
	SD	1.62	2.98	5.36	7.80	3.96	8.50	6.33	5.04	4.18	5.50
NAC–0.1 $n = 20$ $c = 5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.63	5.17	9.88	25.10	54.25	80.47	91.74	94.81	96.41	5.47
	SD	3.59	5.04	6.77	8.41	5.56	9.73	7.61	6.12	5.10	6.68
$R^{(2)}_{BOOT}$	ME	2.53	5.08	9.82	25.15	54.50	80.85	92.06	95.03	96.55	5.85
	SD	2.97	4.34	6.00	7.39	2.92	8.50	6.74	5.43	4.48	5.72
$R^{(3)}_{BOOT}$	ME	2.60	5.17	9.91	25.19	54.42	80.66	91.86	94.87	96.42	5.66
	SD	3.10	4.49	6.12	7.45	3.13	8.74	6.99	5.69	4.76	5.89
NAC–0.1 $n = 20$ $c = 10$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	0.17	0.59	2.25	15.88	52.63	77.51	86.35	89.24	91.04	9.12
	SD	0.36	0.91	2.35	7.77	19.91	24.11	20.93	18.78	17.08	15.37
$R^{(2)}_{BOOT}$	ME	0.49	1.17	3.37	20.20	62.45	88.02	94.83	96.51	97.37	13.02
	SD	0.86	1.73	3.38	6.77	9.18	11.60	8.96	7.71	6.86	7.20
$R^{(3)}_{BOOT}$	ME	0.85	1.55	3.83	19.80	58.27	81.88	89.13	91.34	92.67	8.27
	SD	1.67	2.40	3.80	6.88	14.33	18.92	16.55	14.93	13.61	12.09

Tabela 4.47: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NAC–0.1 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	1.98	4.03	8.77	24.27	51.60	78.19	92.11	96.21	98.12	3.19
	SD	1.49	2.35	3.63	5.14	4.68	4.34	3.25	2.27	1.51	3.43
$R^{(2)}_{BOOT}$	ME	1.88	3.87	8.55	24.06	51.59	78.29	92.15	96.21	98.11	3.29
	SD	1.14	1.80	2.71	3.07	0.84	3.31	3.06	2.22	1.51	2.35
$R^{(3)}_{BOOT}$	ME	1.89	3.90	8.58	24.09	51.58	78.26	92.12	96.19	98.09	3.26
	SD	1.16	1.82	2.73	3.09	0.84	3.31	3.07	2.23	1.52	2.36
NAC–0.1 $n = 100$ $c = 5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.54	5.08	9.92	25.14	51.21	76.85	91.05	95.52	97.59	1.85
	SD	1.42	2.15	2.97	3.25	1.29	3.42	3.23	2.40	1.72	2.55
$R^{(2)}_{BOOT}$	ME	2.52	5.06	9.89	25.09	51.16	76.81	91.02	95.50	97.58	1.81
	SD	1.38	2.09	2.88	3.10	0.92	3.39	3.24	2.42	1.72	2.49
$R^{(3)}_{BOOT}$	ME	2.52	5.06	9.89	25.09	51.16	76.81	91.02	95.50	97.58	1.81
	SD	1.39	2.09	2.89	3.10	0.92	3.39	3.24	2.42	1.73	2.49
NAC–0.1 $n = 100$ $c = 10$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	0.03	0.11	0.45	4.82	28.82	58.47	74.66	80.82	84.77	21.18
	SD	0.03	0.08	0.29	3.77	20.37	30.81	29.32	26.57	23.98	19.78
$R^{(2)}_{BOOT}$	ME	0.52	1.15	2.40	13.69	59.85	91.56	98.24	99.25	99.60	16.56
	SD	0.82	1.69	2.94	3.86	5.67	6.65	3.83	2.71	2.05	3.80
$R^{(3)}_{BOOT}$	ME	0.80	1.33	2.46	12.46	49.97	79.64	90.09	93.23	95.04	12.54
	SD	1.12	1.81	2.65	3.47	14.78	19.02	14.75	11.94	9.84	10.86

Exemplo 4.7 Considerando novamente os dados do Exemplo 4.3 (Secção 4.3.2) aplicou-se quer o *bootstrap* IF usual quer o IF generalizado, para estimação do coeficiente de correlação, $\rho_{\sqrt{w}}$, entre as variáveis LSAT e GPA.

Tabela 4.48: Percentis das várias distribuições *bootstrap* $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$ para os dados das Faculdades de Direito, com $c = 5$.

p	R^U_{BOOT}	R^{IF}_{BOOT}	$R^{(1)}_{BOOT}$	$R^{(2)}_{BOOT}$	$R^{(3)}_{BOOT}$
2.5	-0.30	-0.21	-0.25	-0.28	-0.27
5	-0.24	-0.15	-0.18	-0.21	-0.20
10	-0.17	-0.10	-0.12	-0.15	-0.13
25	-0.08	-0.02	-0.02	-0.05	-0.03
50	0.02	0.04	0.05	0.02	0.05
75	0.10	0.09	0.10	0.07	0.10
90	0.15	0.12	0.13	0.10	0.13
95	0.17	0.13	0.14	0.11	0.15
97.5	0.18	0.15	0.15	0.12	0.16

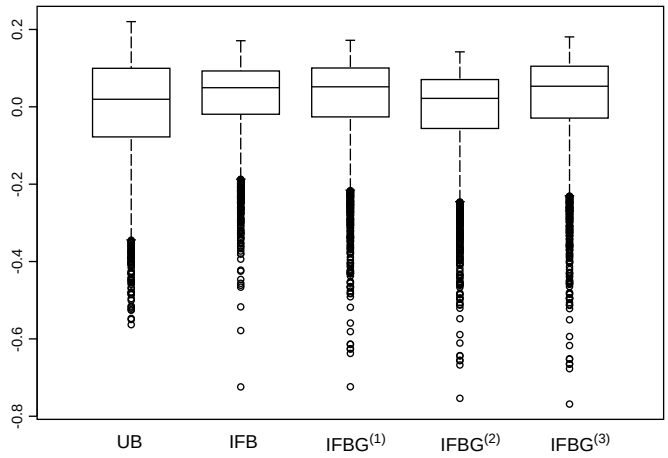


Figura 4.18: Comparação entre os vários métodos *bootstrap* na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$ para os dados das Faculdades de Direito, excepto para R^U_{BOOT} onde a distribuição em causa é a de $R - \rho$ (ou $\rho(F_n) - \rho(F)$).

Na Tabela 4.48 encontram-se os valores de diversos percentis das distribuições *bootstrap* com $B = 5000$ reamostras. Como já esperado os percentis UB são os de maior variabilidade. Entre os métodos *bootstrap* IF corrigidos não se observam grandes diferenças.

Tabela 4.49: Intervalos de confiança a 90% para o coeficiente de correlação entre as variáveis LSTA e GPA.

Método	Limite inferior	Limite superior
UB	0.53	0.94
IFB	0.68	0.96
$IFBG^{(1)}$	0.65	0.97
$IFBG^{(2)}$	0.62	0.94
$IFBG^{(3)}$	0.63	0.98
Fisher sem o ponto A	0.73	0.96

Na Figura 4.18 apresentam-se os vários diagramas de extremos-e-quartis associados a cada um dos métodos *bootstrap*.

As estimativas pontuais obtidas para este conjunto de dados são: $r = 0.78$, r sem observação discordante, $r_{-A} = 0.89$, e $\rho_{\sqrt{w}}(F_n) = 0.83$.

Para finalizar apresentam-se na Tabela 4.49 diversos intervalos de confiança a 90% para a correlação entre as variáveis LSAT e GPA. Note-se que o intervalo aqui apresentado para o método IFB é diferente do apresentado na Tabela 4.3 porque se baseia numa constante de afinação e estimativa pontual diferentes. Os métodos que fornecem intervalos mais próximos do intervalo Fisher para a amostra sem observação atípica são os métodos IFB e $IFBG^{(1)}$.

■

Conclusões

Nesta secção propôs-se um método para aperfeiçoar as estimativas *bootstrap* IF. Os resultados dos dois estudos de Monte Carlo apresentados permitem tirar as seguintes conclusões:

- Quando se estima a localização univariada as correcções propostas possuem bom desempenho e permitem considerar na função peso valores da constante de afinação inferiores ao utilizado ($c = \sqrt{F^{-1}_{\chi^2_{(1)}}(0.99)} \approx 2.58$) na Secção 4.3.1. O comportamento dos IFBG é bastante bom em distribuições Normais e simetricamente contaminadas, mesmo para dimensões amostrais baixas ($n = 20$). No caso de contaminação assimétrica (distribuições NAC–0.1, LOGN1 e LOGN2) os resultados são piores, sendo necessário fazer correcções mais drásticas, como por exemplo a empregue em $R^{(11)}_{BOOT}$. Naturalmente, os resultados melhoram com o aumento da dimensão amostral.

- As correcções propostas também produziram bons resultados para o coeficiente de correlação, não obstante o uso das expressões obtidas para a média (fica em aberto o estudo com as expressões de correcção correspondentes).

"*Experience is not what happens to you. It is what you do with what happens to you.*"

Aldous Huxley

CAPÍTULO 5

Conclusões e trabalho futuro

Ao longo deste trabalho abordou-se o problema da robustez do método de reamostragem *bootstrap*. Este capítulo contém um breve resumo do trabalho apresentado, salientando-se os principais resultados obtidos. Seguidamente, e porque um trabalho deste género resulta, inevitavelmente, na percepção da existência de outros problemas interessantes, apresentam-se algumas direcções e desafios para trabalho futuro.

A simplicidade e versatilidade do método de reamostragem *bootstrap* constitui um dos grandes atractivos para a sua aplicação, principalmente em problemas complexos de estimação. Tal como a maior parte das melhores ideias em matemática, o princípio que rege o *bootstrap*, a reamostragem a partir da função de distribuição empírica, é simples, elegante e para além disso, poderoso. No entanto, subjacente a esse princípio existe uma complexidade teórica suficiente para suscitar o interesse de muitos estatísticos, dos mais aplicados aos mais teóricos.

Quando se procede a uma análise estatística, surge muitas vezes a questão: o estimador de interesse tem distribuição Normal? Se não tiver as teorias baseadas na normalidade podem ser piores que as baseadas em *bootstrap*. Outra questão que se levanta é: uma vez que os cálculos de reamostragem são condicionais à amostra os seus resultados poderão depender bastante de algumas observações discordantes, o que fazer então? A tentativa de resposta a estas questões constituiu a motivação deste trabalho. E se à partida estas pareciam ter uma resposta simples, que passaria pelo utilização conjunta do *bootstrap* e procedimentos robustos, revelou-se um problema de investigação interessante e que culminou (ou se iniciou!) na proposta de um novo método, o denominado *bootstrap* robusto com base na função de influência, IFB.

A ideia deste novo método de reamostragem, o *bootstrap* robusto, é conceptualmente oposta à abordagem tradicional com posterior diagnóstico, já que simplesmente procede à reamostragem mas salvaguardando-a à partida da selecção de observações atípicas arbitrárias. Por vezes, esta

protecção traduz-se numa perda de eficiência quando não existem de facto observações discordantes. No entanto, como foi verificado pelos vários estudos de simulação realizados, essa perda pode ser pequena. Os métodos sugeridos, IFB ou IFBG, não são "automáticos", dado que requerem o cálculo da função de influência do funcional de interesse e da escolha da constante de afinação (o método IFBG apesar de ser menos sensível a esta escolha, também a requer). Porém, para a maioria dos funcionais equivalentes a estimadores ditos clássicos, as funções de influência estão publicadas. Por exemplo, na área de análise multivariada podem referir-se os trabalhos de Radhakrishnan e Kshirsagar (1981), Cook e Weisberg (1982), Critchley (1985), Pires e Branco (2002) e Boente *et al.* (2002). Por sua vez, a escolha de uma constante de afinação apropriada pode basear-se numa simulação prévia ou em cálculos numéricos.

O aspecto mais interessante do plano de reamostragem sugerido é a habilidade de determinar estimativas robustas, pontuais ou regionais, utilizando um estimador clássico, auxiliado por um estimador robusto mas que necessita de ser calculado apenas num passo. Este método evita assim as dificuldades inerentes a reamostrar um estimador robusto e responde ao principal objectivo enunciado no início deste trabalho: a construção de inferências robustas através de um método de reamostragem computacionalmente pouco dispendioso.

A validação do método proposto foi realizada empiricamente, com base em estudos de simulação planeados com bastante cuidado. Esta opção deveu-se à complexidade teórica subjacente a este novo método. Note-se que até ao momento só estão estudadas propriedades teóricas para vectores de componentes de probabilidades de reamostragem diferentes de n^{-1}, quando essas são permutáveis, o qual não é a situação no *bootstrap* IF. Apesar de se terem feito algumas tentativas neste sentido, não se conseguiu progredir nos estudos teóricos. Este é um desafio que fica desde já em perspectiva.

Assumindo que o modelo gerador dos dados pertence a uma vizinhança de contaminação de um modelo central Normal construíram-se inferências robustas pontuais e regionais para a localização univariada e para o coeficiente de correlação. Propôs-se também um critério robusto para a selecção de variáveis. Em cada um desses estudos compararam-se duas estratégias:

(i) utilização do *bootstrap* IF com um estimador clássico;

(ii) utilização do *bootstrap* clássico com um estimador robusto.

Em todas as situações estudadas, foi a primeira estratégia a que conduziu a melhores resultados, sendo simultaneamente aquela que tem associada um menor dispêndio de tempo de cálculo. Note-se, contudo, que se o objectivo consistir no estudo da distribuição de um determinado estimador robusto, o método IFB parece não ser tão eficiente quanto o *bootstrap* clássico desde que não se verifiquem problemas de instabilidade numérica. Desejando trabalhar com estimadores robustos dever-se-á consultar as recentes propostas de Salibian-Barrera e Zamar (2002b).

No âmbito da aplicação do *bootstrap* IF a problemas multivariados, procedeu-se a um estudo que generaliza as ideias apresentadas em He e Fung (2000) a outros estimadores de alto ponto de rotura, para além do estimador S. As versões ponderadas dos vários estimadores estudados apresentaram

melhor eficiência finita que as versões não ponderadas, o que vem ao encontro de alegações que têm sido apresentadas na literatura. Na determinação dos estimadores ponderados foi utilizada uma função peso do tipo: $w(x) = 1$ se $0 < x < c$ e $w(x) = 0$, se $x \geq c$. Os resultados obtidos para as probabilidades de reamostragem no *bootstrap* IF, nomeadamente o bom desempenho no controlo de observações discordantes, levam a pensar que, também neste caso, se conseguirá aumentar a eficiência se a função de peso, w, decrescer suavemente (por exemplo, escolhendo funções pertencentes à família Ψ, ver pág. 83). Assim, seria desejável estudar, quer do ponto de vista teórico quer do ponto de vista aplicado, a eficiência de uma classe mais abrangente de estimadores ponderados.

A flexibilização da escolha da constante de afinação no *bootstrap* IF conduziu ao desenvolvimento de uma generalização do método, que se designou por IFBG. Com esta generalização obtiveram-se resultados encorajadores e que constituíram melhorias em relação ao IFB. Esta generalização encontra-se, todavia, numa fase embrionária, pois apenas se estudou a situação da localização univariada. É desejável, se possível, realizar estudos equivalentes para outras aplicações, em particular para o coeficiente de correlação e para o critério de selecção de variáveis.

No contexto do desenvolvimento de um critério de selecção de variáveis em Análise Discriminante Linear, surgiram naturalmente vários pontos que não foram aqui abordados mas que poderão constituir focos de interesse futuro, nomeadamente:

1. O estudo do ponto de rotura associado a este critério robusto de selecção;

2. O estudo do comportamento do critério de selecção para métodos de Análise Discriminante não lineares, assim como a generalização para mais do que duas populações (quer no caso linear, quer no não linear);

3. A utilização do *bootstrap* IF na estimação de uma regra discriminante robusta e a sua comparação com outras propostas robustas presentes na literatura (ver, por exemplo, Pires, 1995);

4. A conjunção de vários algoritmos de pesquisa onde o critério de selecção é baseado no *bootstrap* IF. Neste sentido poder-se-á estudar a possibilidade de adaptação deste critério em procedimentos de pesquisa que se têm revelado promissores neste campo, nomeadamente, os baseados nos algoritmos de *leaps and bounds* (Duarte Silva, 1998) ou em algoritmos genéticos e de *simulated annealing* (Cadima *et al.*, 2002).

Uma área estatística particularmente interessante e com grande aplicabilidade é constituída pelos modelos de regressão. Os pontos atrás apresentados poderiam ser reescritos substituindo análise discriminante por regressão, constituindo assim mais novas possibilidades de investigação futura.

Apesar de à partida se saber que, na maior parte das situações, o estimador *bootstrap* da distribuição é mais preciso que o histograma cumulativo *jackknife*, os resultados obtidos na estimação de quantis utilizando o *deleted-d jackknife* ficaram aquém do esperado. Esta constatação leva a pensar numa adaptação do método IF também para este método de reamostragem. Conjectura-se que esse novo método, o qual se poderia denominar *deleted-d jackknife* IF possa ser mais eficiente que o *bootstrap*

IF quando o objectivo consistir na obtenção da variância de um estimador, como acontece já no caso clássico.

Os procedimentos robustos, como foi dito no Capítulo 2, não pertencem à classe dos métodos não paramétricos pois assume-se que se está a trabalhar numa determinada vizinhança de um modelo paramétrico central. Inevitavelmente, a maior parte dos métodos robustos supõem que esse modelo central é a Normal (univariada ou multivariada). Este trabalho não constituiu também excepção a essa suposição. Avaliar o comportamento do IFB ou do IFBG para outros modelos centrais contínuos, assim como considerar também modelos discretos e mistos, sempre que essa aplicação tenha sentido, pode ser um campo promissor para investigação futura.

O método *bootstrap* IFB (ou IFBG) é dependente da aplicação o que torna bastante complexo o estudo das suas propriedades teóricas gerais. A justificação da validade teórica deste método poderá ser iniciada pela demonstração da convergência da distribuição *bootstrap* IFBG para funções regulares da média amostral. Estes estudos teóricos constituiem uma grande área inexplorada.

APÊNDICE A

Demonstrações

Neste apêndice apresentam-se a demonstração da Proposição 2.1 e a determinação do ponto $\mathbf{x}^o$ associado à distribuição NAC, usada no estudo de simulação do Capítulo 4.

A.1 Demonstração da proposição 2.1

$$
\begin{aligned}
b_0 &= E_{F_0}[\rho_B(\|\mathbf{x}\|)] \\
&= \int\limits_{\|\mathbf{x}\|<c_0} \left\{ \frac{\|\mathbf{x}\|^2}{2} - \frac{\|\mathbf{x}\|^4}{2c_0^2} + \frac{\|\mathbf{x}\|^6}{6c_0^4} \right\} f_0(\mathbf{x}) d\mathbf{x} \\
&+ \int\limits_{\|\mathbf{x}\|>c_0} \frac{c_0^2}{6} f_0(\mathbf{x}) d\mathbf{x}
\end{aligned}
\tag{A.1}
$$

com $f_0 = (2\pi)^{-m/2} \times \exp\left\{-(1/2)\|\mathbf{x}\|^2\right\}$.

É imediato que o segundo integral origina (recorde-se que $c_0 > 0$)

$$c_0^2/6 \times (1 - F_{\chi^2_{(m)}}(c_0^2))$$

Separe-se e denote-se respectivamente por A, B e C cada uma das três expressões constituintes do

primeiro integral

$$A = \int\limits_{\|\mathbf{x}\|<c_0} \frac{\|\mathbf{x}\|^2}{2} f_0(\mathbf{x}) d\mathbf{x},$$

$$B = \int\limits_{\|\mathbf{x}\|<c_0} \frac{\|\mathbf{x}\|^4}{2c_0^2} f_0(\mathbf{x}) d\mathbf{x}, \tag{A.2}$$

$$C = \int\limits_{\|\mathbf{x}\|<c_0} \frac{\|\mathbf{x}\|^6}{6c_0^4} f_0(\mathbf{x}) d\mathbf{x}.$$

Então

$$A = \frac{(2\pi)^{-m/2}}{2} \int\limits_{\|\mathbf{x}\|<c_0} \|\mathbf{x}\|^2 \exp\left\{-(1/2)\|\mathbf{x}\|^2\right\} d\mathbf{x}.$$

Considerando $z(\|\mathbf{x}\|^2) = \|\mathbf{x}\|^2 \exp\left\{-(1/2)\|\mathbf{x}\|^2\right\}$ e aplicando o Lema 2.1 (baseado na mudança de coordenadas cartesianas para hiperesféricas) vem

$$A = \frac{2^{-m/2}}{\Gamma(m/2)} \int\limits_0^{c_0} r^{m+1} \exp\left\{-(1/2)r^2\right\} dr.$$

Fazendo uma mudança de variável ($y = r^2$) obtém-se

$$A = \frac{2^{-m/2}}{\Gamma(m/2)} \int\limits_0^{c_0^2} y^{\frac{m+1}{2}} \exp\left\{-(1/2)y\right\} (4y)^{(-1/2)} dy,$$

e finalmente

$$\begin{aligned} A =& \frac{m/2}{\Gamma(\frac{m+2}{2}) 2^{\frac{m+2}{2}}} \int\limits_0^{c_0^2} y^{\frac{m+2}{2}-1} \exp\left\{-(1/2)y\right\} dy \\ =& \frac{m}{2} F_{\chi^2_{(m+2)}}(c_0^2), \end{aligned}$$

como se queria mostrar.

A determinação de B e C é efectuada de forma semelhante tendo apenas em consideração que

$$z(\|\mathbf{x}\|^2) = \|\mathbf{x}\|^4 \exp\left\{-(1/2)\|\mathbf{x}\|^2\right\},$$

para o cálculo de B e

$$z(\|\mathbf{x}\|^2) = \|\mathbf{x}\|^6 \exp\left\{-(1/2)\|\mathbf{x}\|^2\right\},$$

para o cálculo de C. ■

A.2 Determinação do ponto $\mathbf{x}^o$

Considerem-se as distribuições NOR e NAC da Secção 4.3.3:

- Normal, NOR:

$$\mathbf{X}|G_i : N_m(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad i = 1, 2;$$

- Normal Assimetricamente ε-Contaminada, NAC:

$$\mathbf{X}|G_1 : N_m(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \quad \text{para} \quad j = 1, 2, \cdots, n_1 - [\varepsilon n_1] \quad \text{e} \quad \Delta_{\mathbf{x}^o} \quad \text{para} \quad j = n_1 - [\varepsilon n_1] + 1, \cdots, n_1,$$

$$\mathbf{X}|G_2 : N_m(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2);$$

Pretende-se determinar um ponto $\mathbf{x}^o$ tal que a direcção discriminante entre os dois grupos na situação contaminada seja ortogonal à direcção discriminante na situação não contaminada.

$\boldsymbol{\alpha} = (-\mu, 0, \cdots, 0)$ é o vector discriminante na situação não contaminada

e

$\boldsymbol{\alpha}^{Cont} = \left(\boldsymbol{\Sigma}_C^{Cont}\right)^{-1} \left(\boldsymbol{\mu}_1^{Cont} - \boldsymbol{\mu}_2\right)$ é o vector discriminante na situação contaminada.

Para que estes sejam ortogonais basta que a primeira componente de $\boldsymbol{\alpha}^{Cont}$ seja nula, isto é, $\alpha_1^{Cont} = 0$. Sem perda de generalidade pode fixar-se $\mathbf{x}^o = (a, b, 0, \cdots, 0)$. Tem-se então (admitindo que εn_1 é inteiro)

$$\boldsymbol{\mu}_1^{Cont} = \varepsilon \mathbf{x}^o,$$

$$\boldsymbol{\Sigma}_1^{Cont} = (1 - \varepsilon) \left[\mathbf{I}_m + \varepsilon \mathbf{x}^o (\mathbf{x}^o)^T, \right]$$

$$\boldsymbol{\Sigma}_C^{Cont} = (1 - \varepsilon) w_1 \left[\mathbf{I}_m + \varepsilon \mathbf{x}^o (\mathbf{x}^o)^T\right] + w_2 \mathbf{I}_m,$$

com $w_i = (n_i - 1)/(n_1 + n_2 - 2)$, $i = 1, 2$. A primeira linha de $\left(\boldsymbol{\Sigma}_C^{Cont}\right)^{-1}$ é, a menos de uma constante positiva, dada por

$$\left(1 - \varepsilon w_1 + \varepsilon(1 - \varepsilon) b^2 w_1, -\varepsilon(1 - \varepsilon) a b w_1, 0, \cdots, 0\right),$$

pelo que $\alpha_1^{Cont} = 0$ é equivalente a

$$\left(1 - \varepsilon w_1 + \varepsilon(1 - \varepsilon) b^2 w_1\right)(\varepsilon a - \mu) - \varepsilon^2 (1 - \varepsilon) a b^2 w_1 = 0 \Leftrightarrow$$

$$\Leftrightarrow b = \pm \sqrt{\frac{(1 - \varepsilon w_1)(\varepsilon a - \mu)}{\varepsilon(1 - \varepsilon)\mu w_1}}.$$

Para a existência de um ponto $\mathbf{x}^o$ nas condições requeridas é necessário que $a > \mu/\varepsilon$.

APÊNDICE B

Resultados computacionais suplementares

Neste apêndice apresentam-se alguns resultados suplementares, relativos à estimação da distribuição *bootstrap* pelo método IF generalizado, obtidos nos estudos de simulação descritos na Secção 4.4.

Relativamente à estimação do valor esperado univariado (Subsecção 4.4.1):

- nas Figuras B.1 a B.3 representam-se graficamente os indicadores, SQ, DA e KS, para as diferentes dimensões e distribuições;
- nas Tabelas B.1 a B.30 apresentam-se os resultados referentes a $R_{BOOT}^{(k)}$, $k = 4, 5, \cdots, 12$, para as distribuições NOR, NSC, M4–0.1, T1, NAC e LOGN1, assim como a $R_{BOOT}^{(k)}$, $k = 1, 2, \cdots, 12$, para as distribuições M4–0.2 e LOGN2.

Quanto à estimação do coeficiente de correlação (Subsecção 4.4.2):

- na Figura B.4 representam-se graficamente os indicadores, SQ, DA e KS, para as diferentes dimensões e distribuições;
- nas Tabelas B.31 a B.39 apresentam-se os resultados referentes a $R_{BOOT}^{(k)}$, $k = 4, 5, \cdots, 12$, para as distribuições NOR, NSC e NAC.

B.1 Resultados para a localização univariada

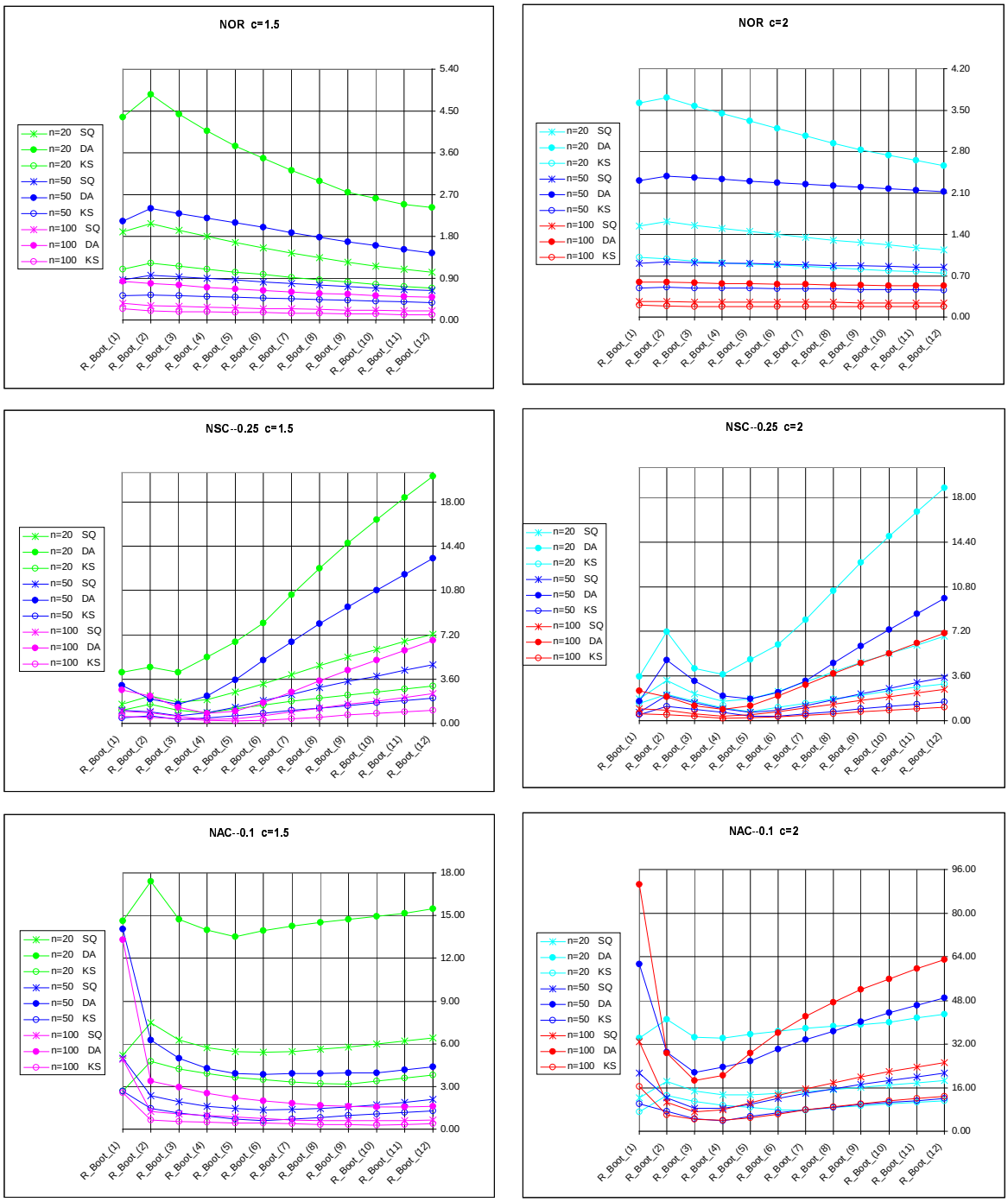


Figura B.1: Desempenho dos vários estimadores R_{Boot} na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$ em termos dos indicadores globais, SQ, DA e KS, para as distribuições NOR, NSC–0.25 e NAC–0.1.

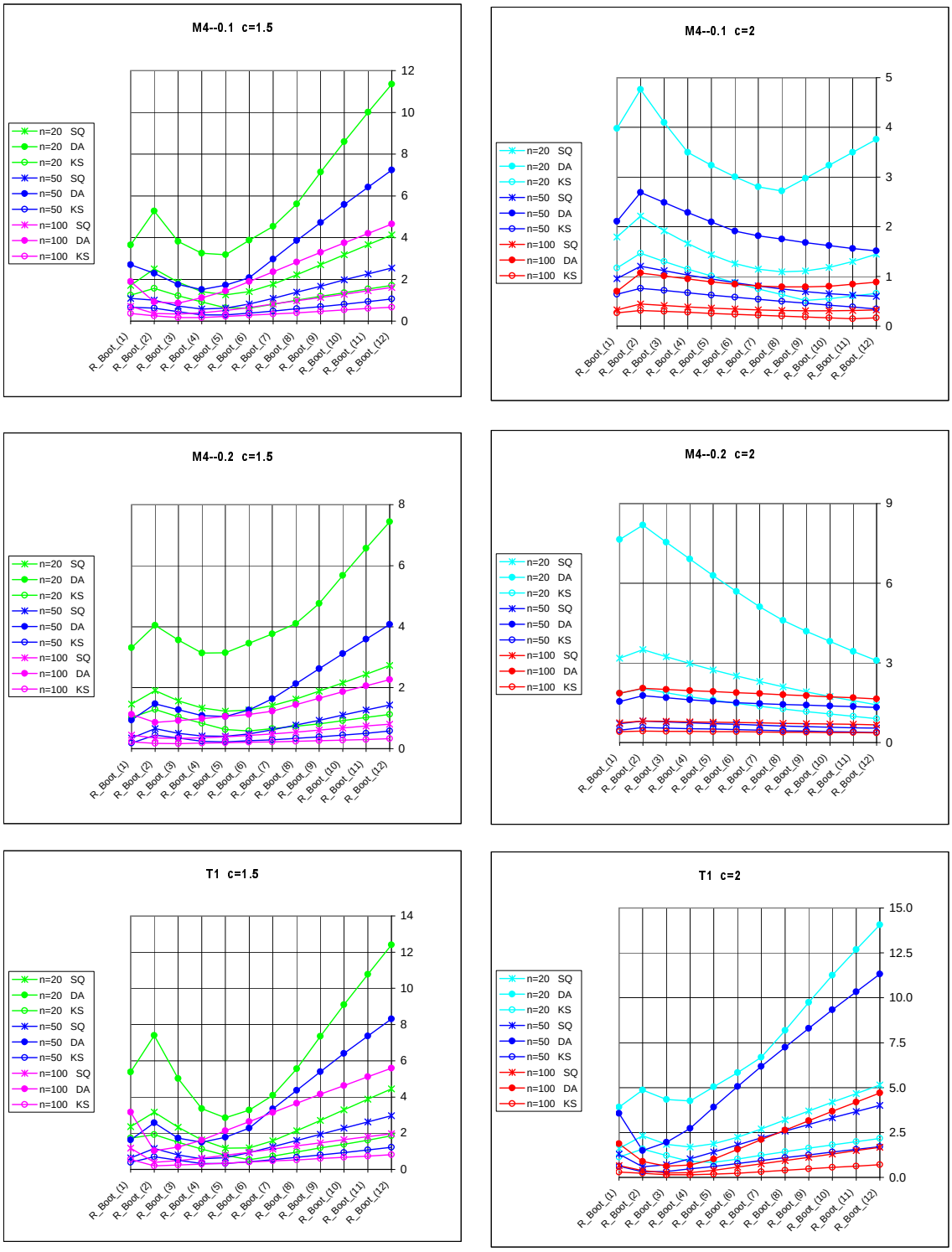


Figura B.2: Desempenho dos vários estimadores R_{Boot} na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$ em termos dos indicadores globais, SQ, DA e KS, para as distribuições M4–0.1, M4–0.2 e T1.

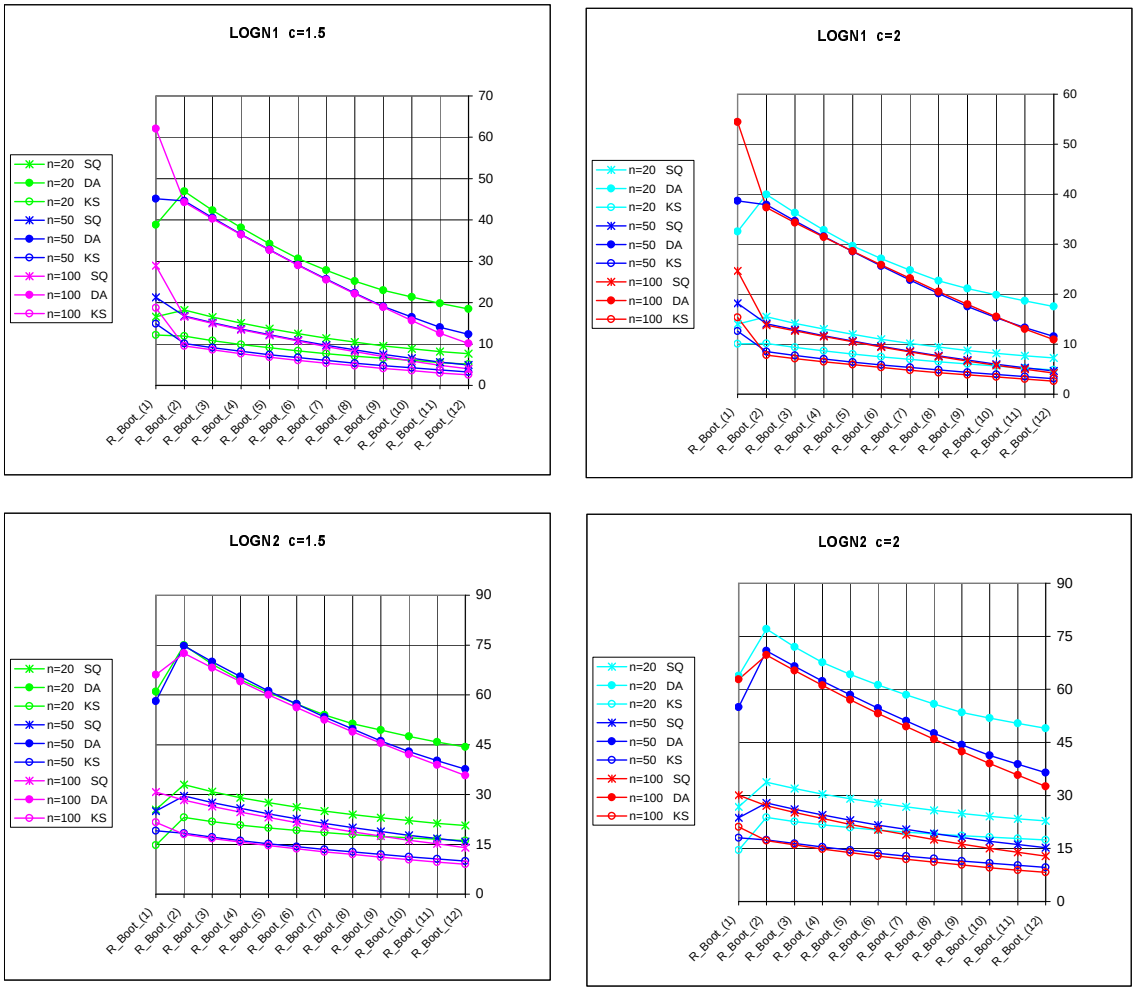


Figura B.3: Desempenho dos vários estimadores R_{Boot} na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$ em termos dos indicadores globais, SQ, DA e KS, para as distribuições LOGN1 e LOGN2.

Tabela B.1: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal padrão e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NOR $n = 20$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.53	4.77	9.38	23.90	49.94	76.02	90.67	95.32	97.54	1.10
	SD	1.85	2.70	3.74	4.07	0.92	4.01	3.73	2.69	1.85	3.03
$R^{(5)}_{BOOT}$	ME	2.56	4.82	9.44	23.96	49.94	75.96	90.61	95.28	97.51	1.04
	SD	1.88	2.72	3.76	4.08	0.92	4.02	3.76	2.72	1.88	3.05
$R^{(6)}_{BOOT}$	ME	2.59	4.86	9.50	24.02	49.94	75.90	90.55	95.23	97.48	0.98
	SD	1.92	2.75	3.78	4.09	0.92	4.04	3.79	2.75	1.91	3.07
$R^{(7)}_{BOOT}$	ME	2.62	4.91	9.55	24.08	49.94	75.84	90.49	95.18	97.44	0.92
	SD	1.95	2.79	3.81	4.10	0.92	4.05	3.82	2.79	1.95	3.10
$R^{(8)}_{BOOT}$	ME	2.66	4.96	9.61	24.13	49.94	75.79	90.44	95.14	97.41	0.87
	SD	1.98	2.83	3.84	4.12	0.92	4.07	3.86	2.83	1.99	3.13
$R^{(9)}_{BOOT}$	ME	2.69	5.00	9.66	24.18	49.94	75.74	90.38	95.09	97.38	0.82
	SD	2.02	2.87	3.88	4.13	0.92	4.09	3.90	2.87	2.03	3.15
$R^{(10)}_{BOOT}$	ME	2.72	5.04	9.71	24.23	49.94	75.69	90.33	95.05	97.34	0.77
	SD	2.06	2.91	3.91	4.15	0.92	4.11	3.94	2.92	2.07	3.18
$R^{(11)}_{BOOT}$	ME	2.75	5.08	9.77	24.27	49.94	75.64	90.28	95.01	97.31	0.73
	SD	2.11	2.96	3.95	4.17	0.92	4.13	3.98	2.96	2.11	3.22
$R^{(12)}_{BOOT}$	ME	2.79	5.12	9.82	24.32	49.94	75.60	90.23	94.97	97.28	0.68
	SD	2.15	3.00	3.99	4.19	0.92	4.15	4.02	3.01	2.15	3.25
NOR $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.53	4.78	9.49	24.07	49.98	75.81	90.59	95.29	97.55	0.93
	SD	1.79	2.59	3.54	3.76	0.97	3.74	3.59	2.63	1.80	2.88
$R^{(5)}_{BOOT}$	ME	2.54	4.80	9.52	24.10	49.98	75.79	90.57	95.27	97.54	0.90
	SD	1.80	2.59	3.54	3.75	0.97	3.73	3.58	2.63	1.81	2.87
$R^{(6)}_{BOOT}$	ME	2.56	4.82	9.54	24.12	49.98	75.76	90.54	95.26	97.53	0.88
	SD	1.80	2.59	3.54	3.75	0.97	3.72	3.58	2.63	1.81	2.87
$R^{(7)}_{BOOT}$	ME	2.57	4.83	9.56	24.15	49.98	75.74	90.52	95.24	97.52	0.85
	SD	1.80	2.59	3.54	3.74	0.97	3.72	3.58	2.63	1.81	2.87
$R^{(8)}_{BOOT}$	ME	2.58	4.85	9.58	24.17	49.98	75.72	90.50	95.22	97.51	0.83
	SD	1.80	2.60	3.54	3.74	0.97	3.72	3.58	2.63	1.81	2.87
$R^{(9)}_{BOOT}$	ME	2.59	4.86	9.60	24.19	49.98	75.70	90.47	95.21	97.50	0.81
	SD	1.81	2.60	3.55	3.74	0.97	3.72	3.58	2.63	1.82	2.88
$R^{(10)}_{BOOT}$	ME	2.60	4.88	9.62	24.21	49.98	75.67	90.45	95.19	97.49	0.79
	SD	1.81	2.61	3.55	3.74	0.97	3.72	3.58	2.64	1.82	2.88
$R^{(11)}_{BOOT}$	ME	2.61	4.90	9.64	24.23	49.98	75.65	90.43	95.18	97.48	0.77
	SD	1.82	2.62	3.56	3.75	0.97	3.72	3.59	2.64	1.83	2.88
$R^{(12)}_{BOOT}$	ME	2.63	4.91	9.66	24.25	49.98	75.64	90.41	95.16	97.47	0.75
	SD	1.83	2.63	3.57	3.75	0.97	3.72	3.60	2.65	1.84	2.89

Tabela B.2: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal padrão e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NOR $n = 50$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R_{BOOT}^{(4)}$	ME	2.46	4.84	9.67	24.50	50.05	75.49	90.35	95.18	97.58	0.50
	SD	1.13	1.69	2.28	2.32	0.75	2.38	2.27	1.66	1.11	1.83
$R_{BOOT}^{(5)}$	ME	2.46	4.85	9.68	24.51	50.05	75.47	90.34	95.17	97.58	0.49
	SD	1.13	1.69	2.28	2.32	0.75	2.37	2.27	1.67	1.11	1.83
$R_{BOOT}^{(6)}$	ME	2.47	4.86	9.70	24.53	50.05	75.46	90.32	95.16	97.57	0.47
	SD	1.14	1.69	2.28	2.31	0.75	2.37	2.27	1.67	1.11	1.83
$R_{BOOT}^{(7)}$	ME	2.48	4.87	9.72	24.54	50.05	75.44	90.31	95.15	97.56	0.46
	SD	1.14	1.69	2.28	2.31	0.75	2.37	2.27	1.67	1.12	1.83
$R_{BOOT}^{(8)}$	ME	2.49	4.89	9.73	24.56	50.05	75.43	90.29	95.14	97.55	0.44
	SD	1.14	1.69	2.28	2.31	0.75	2.37	2.27	1.67	1.12	1.83
$R_{BOOT}^{(9)}$	ME	2.50	4.90	9.75	24.57	50.05	75.41	90.28	95.12	97.54	0.43
	SD	1.15	1.70	2.28	2.31	0.75	2.37	2.27	1.68	1.12	1.83
$R_{BOOT}^{(10)}$	ME	2.51	4.91	9.76	24.59	50.05	75.40	90.26	95.11	97.54	0.41
	SD	1.15	1.70	2.29	2.31	0.75	2.37	2.28	1.68	1.12	1.83
$R_{BOOT}^{(11)}$	ME	2.51	4.93	9.78	24.60	50.05	75.38	90.25	95.10	97.53	0.40
	SD	1.15	1.70	2.29	2.32	0.75	2.38	2.28	1.69	1.13	1.84
$R_{BOOT}^{(12)}$	ME	2.52	4.94	9.79	24.62	50.05	75.37	90.23	95.09	97.52	0.38
	SD	1.15	1.71	2.29	2.32	0.75	2.38	2.28	1.69	1.13	1.84
NOR $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R_{BOOT}^{(4)}$	ME	2.45	4.83	9.67	24.51	50.10	75.49	90.38	95.23	97.60	0.49
	SD	1.14	1.67	2.25	2.28	0.79	2.23	2.20	1.65	1.12	1.79
$R_{BOOT}^{(5)}$	ME	2.46	4.84	9.68	24.52	50.10	75.49	90.38	95.22	97.60	0.49
	SD	1.14	1.67	2.25	2.28	0.79	2.23	2.20	1.65	1.12	1.79
$R_{BOOT}^{(6)}$	ME	2.46	4.84	9.68	24.52	50.10	75.48	90.37	95.22	97.60	0.48
	SD	1.14	1.67	2.25	2.28	0.79	2.23	2.20	1.65	1.12	1.79
$R_{BOOT}^{(7)}$	ME	2.46	4.84	9.69	24.52	50.10	75.48	90.37	95.22	97.60	0.48
	SD	1.14	1.66	2.25	2.28	0.79	2.23	2.20	1.65	1.12	1.79
$R_{BOOT}^{(8)}$	ME	2.46	4.85	9.69	24.53	50.10	75.47	90.37	95.21	97.60	0.47
	SD	1.14	1.66	2.25	2.27	0.79	2.22	2.20	1.65	1.12	1.78
$R_{BOOT}^{(9)}$	ME	2.46	4.85	9.69	24.53	50.10	75.47	90.36	95.21	97.59	0.47
	SD	1.14	1.66	2.25	2.27	0.79	2.22	2.20	1.65	1.12	1.78
$R_{BOOT}^{(10)}$	ME	2.47	4.85	9.70	24.54	50.10	75.47	90.36	95.21	97.59	0.47
	SD	1.14	1.66	2.25	2.27	0.79	2.22	2.20	1.65	1.12	1.78
$R_{BOOT}^{(11)}$	ME	2.47	4.85	9.70	24.54	50.10	75.46	90.35	95.20	97.59	0.46
	SD	1.14	1.66	2.25	2.27	0.79	2.22	2.20	1.65	1.12	1.78
$R_{BOOT}^{(12)}$	ME	2.47	4.86	9.71	24.54	50.10	75.46	90.35	95.20	97.59	0.46
	SD	1.14	1.66	2.24	2.27	0.79	2.22	2.20	1.65	1.12	1.78

Tabela B.3: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal padrão e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NOR $n = 100$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.56	4.98	9.88	24.91	50.09	75.18	90.11	95.05	97.50	0.18
	SD	0.82	1.20	1.59	1.60	0.65	1.58	1.57	1.18	0.81	1.27
$R^{(5)}_{BOOT}$	ME	2.56	4.99	9.89	24.92	50.09	75.17	90.11	95.04	97.49	0.17
	SD	0.82	1.20	1.59	1.59	0.65	1.58	1.58	1.18	0.81	1.27
$R^{(6)}_{BOOT}$	ME	2.56	5.00	9.90	24.93	50.09	75.16	90.10	95.03	97.49	0.16
	SD	0.82	1.20	1.59	1.59	0.65	1.58	1.58	1.18	0.81	1.27
$R^{(7)}_{BOOT}$	ME	2.57	5.00	9.91	24.94	50.09	75.15	90.09	95.03	97.48	0.15
	SD	0.82	1.20	1.59	1.59	0.65	1.58	1.58	1.18	0.81	1.27
$R^{(8)}_{BOOT}$	ME	2.57	5.01	9.92	24.94	50.09	75.14	90.08	95.02	97.48	0.14
	SD	0.82	1.20	1.59	1.59	0.65	1.58	1.58	1.18	0.81	1.28
$R^{(9)}_{BOOT}$	ME	2.58	5.02	9.93	24.95	50.09	75.14	90.07	95.01	97.48	0.14
	SD	0.82	1.20	1.59	1.59	0.65	1.58	1.58	1.18	0.81	1.28
$R^{(10)}_{BOOT}$	ME	2.58	5.02	9.93	24.96	50.09	75.13	90.06	95.01	97.47	0.13
	SD	0.82	1.20	1.59	1.59	0.65	1.58	1.58	1.18	0.81	1.28
$R^{(11)}_{BOOT}$	ME	2.58	5.03	9.94	24.97	50.09	75.12	90.06	95.00	97.47	0.12
	SD	0.82	1.20	1.59	1.59	0.65	1.58	1.58	1.18	0.81	1.28
$R^{(12)}_{BOOT}$	ME	2.59	5.03	9.95	24.97	50.09	75.11	90.05	95.00	97.46	0.11
	SD	0.82	1.21	1.59	1.59	0.65	1.58	1.58	1.19	0.81	1.28
NOR $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.57	4.99	9.92	24.87	50.02	75.18	90.06	95.00	97.52	0.18
	SD	0.82	1.19	1.62	1.61	0.66	1.60	1.63	1.21	0.80	1.29
$R^{(5)}_{BOOT}$	ME	2.57	4.99	9.92	24.87	50.02	75.18	90.06	95.00	97.52	0.18
	SD	0.82	1.19	1.62	1.61	0.66	1.60	1.62	1.21	0.80	1.29
$R^{(6)}_{BOOT}$	ME	2.57	4.99	9.92	24.87	50.02	75.18	90.06	95.00	97.52	0.18
	SD	0.82	1.19	1.62	1.61	0.66	1.60	1.62	1.21	0.80	1.29
$R^{(7)}_{BOOT}$	ME	2.57	4.99	9.92	24.87	50.02	75.18	90.05	95.00	97.51	0.18
	SD	0.82	1.19	1.62	1.61	0.66	1.60	1.62	1.21	0.80	1.29
$R^{(8)}_{BOOT}$	ME	2.57	4.99	9.93	24.87	50.02	75.18	90.05	95.00	97.51	0.18
	SD	0.82	1.19	1.62	1.61	0.66	1.60	1.62	1.21	0.80	1.29
$R^{(9)}_{BOOT}$	ME	2.57	4.99	9.93	24.88	50.02	75.18	90.05	95.00	97.51	0.18
	SD	0.82	1.19	1.61	1.61	0.66	1.60	1.62	1.21	0.80	1.29
$R^{(10)}_{BOOT}$	ME	2.57	4.99	9.93	24.88	50.02	75.17	90.05	95.00	97.51	0.17
	SD	0.82	1.19	1.61	1.61	0.66	1.60	1.62	1.21	0.80	1.29
$R^{(11)}_{BOOT}$	ME	2.57	4.99	9.93	24.88	50.02	75.17	90.05	95.00	97.51	0.17
	SD	0.82	1.19	1.61	1.60	0.66	1.60	1.62	1.21	0.80	1.29
$R^{(12)}_{BOOT}$	ME	2.57	4.99	9.93	24.88	50.02	75.17	90.05	94.99	97.51	0.17
	SD	0.82	1.19	1.61	1.60	0.66	1.60	1.62	1.21	0.80	1.29

Tabela B.4: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal simetricamente contaminada e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NSC–0.25 $n = 20$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	3.27	5.84	10.46	24.36	49.87	75.53	89.50	94.17	96.80	0.84
	SD	5.03	6.80	8.61	8.66	1.57	8.73	8.58	6.75	4.98	7.02
$R^{(5)}_{BOOT}$	ME	3.51	6.17	10.85	24.71	49.87	75.17	89.10	93.85	96.55	1.17
	SD	5.31	7.06	8.82	8.75	1.57	8.82	8.81	7.02	5.25	7.20
$R^{(6)}_{BOOT}$	ME	3.74	6.48	11.22	25.04	49.87	74.85	88.73	93.53	96.31	1.48
	SD	5.57	7.31	9.02	8.85	1.56	8.90	9.02	7.27	5.51	7.38
$R^{(7)}_{BOOT}$	ME	3.96	6.77	11.56	25.33	49.87	74.56	88.39	93.24	96.08	1.77
	SD	5.80	7.53	9.21	8.92	1.56	8.98	9.22	7.52	5.75	7.55
$R^{(8)}_{BOOT}$	ME	4.17	7.04	11.87	25.60	49.87	74.30	88.07	92.97	95.87	2.04
	SD	6.02	7.75	9.37	9.00	1.56	9.05	9.41	7.74	5.99	7.70
$R^{(9)}_{BOOT}$	ME	4.38	7.30	12.17	25.85	49.88	74.04	87.78	92.71	95.66	2.30
	SD	6.24	7.96	9.54	9.06	1.56	9.11	9.57	7.95	6.20	7.85
$R^{(10)}_{BOOT}$	ME	4.58	7.54	12.44	26.09	49.88	73.80	87.50	92.46	95.46	2.54
	SD	6.44	8.14	9.69	9.12	1.55	9.17	9.73	8.15	6.41	7.98
$R^{(11)}_{BOOT}$	ME	4.77	7.77	12.71	26.30	49.88	73.58	87.24	92.22	95.27	2.78
	SD	6.63	8.31	9.83	9.18	1.55	9.23	9.88	8.33	6.60	8.11
$R^{(12)}_{BOOT}$	ME	4.95	8.00	12.96	26.51	49.88	73.38	86.99	92.00	95.08	3.01
	SD	6.80	8.50	9.97	9.23	1.55	9.28	10.00	8.50	6.79	8.23
NSC–0.25 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.92	5.40	10.02	23.96	49.88	75.89	89.90	94.62	97.15	1.04
	SD	4.53	6.32	8.46	9.20	1.94	9.33	8.53	6.39	4.48	7.00
$R^{(5)}_{BOOT}$	ME	3.17	5.75	10.45	24.38	49.88	75.45	89.44	94.26	96.89	0.75
	SD	4.75	6.57	8.69	9.27	1.94	9.42	8.77	6.66	4.72	7.17
$R^{(6)}_{BOOT}$	ME	3.40	6.08	10.85	24.77	49.89	75.08	89.02	93.93	96.65	1.08
	SD	4.97	6.83	8.90	9.33	1.94	9.49	8.99	6.92	4.96	7.34
$R^{(7)}_{BOOT}$	ME	3.64	6.39	11.23	25.12	49.89	74.73	88.63	93.61	96.41	1.39
	SD	5.20	7.06	9.10	9.39	1.94	9.56	9.21	7.16	5.20	7.50
$R^{(8)}_{BOOT}$	ME	3.86	6.68	11.58	25.43	49.89	74.41	88.27	93.31	96.18	1.73
	SD	5.42	7.29	9.28	9.45	1.93	9.63	9.41	7.40	5.44	7.65
$R^{(9)}_{BOOT}$	ME	4.08	6.96	11.92	25.72	49.89	74.13	87.94	93.03	95.95	2.06
	SD	5.63	7.51	9.46	9.50	1.93	9.69	9.59	7.62	5.67	7.80
$R^{(10)}_{BOOT}$	ME	4.29	7.22	12.23	25.98	49.89	73.86	87.62	92.75	95.74	2.38
	SD	5.84	7.72	9.62	9.56	1.93	9.74	9.77	7.82	5.89	7.94
$R^{(11)}_{BOOT}$	ME	4.49	7.47	12.52	26.23	49.89	73.60	87.31	92.49	95.53	2.69
	SD	6.04	7.91	9.77	9.61	1.93	9.80	9.94	8.03	6.10	8.08
$R^{(12)}_{BOOT}$	ME	4.69	7.71	12.79	26.47	49.89	73.36	87.03	92.25	95.33	2.97
	SD	6.23	8.09	9.90	9.65	1.93	9.85	10.08	8.22	6.31	8.20

Tabela B.5: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal simetricamente contaminada e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NSC–0.25 $n = 50$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.87	5.36	10.22	24.95	50.04	75.09	89.69	94.60	97.14	0.40
	SD	3.09	4.27	5.48	5.28	0.90	5.29	5.51	4.31	3.13	4.39
$R^{(5)}_{BOOT}$	ME	3.01	5.56	10.47	25.19	50.04	74.86	89.44	94.40	97.00	0.60
	SD	3.23	4.39	5.58	5.33	0.89	5.33	5.61	4.43	3.25	4.47
$R^{(6)}_{BOOT}$	ME	3.15	5.75	10.70	25.40	50.04	74.65	89.20	94.21	96.86	0.80
	SD	3.36	4.52	5.69	5.37	0.89	5.38	5.71	4.56	3.37	4.56
$R^{(7)}_{BOOT}$	ME	3.28	5.93	10.93	25.60	50.04	74.45	88.97	94.03	96.72	1.03
	SD	3.49	4.65	5.80	5.42	0.89	5.42	5.82	4.69	3.50	4.66
$R^{(8)}_{BOOT}$	ME	3.42	6.11	11.14	25.80	50.04	74.25	88.76	93.85	96.58	1.24
	SD	3.63	4.78	5.91	5.46	0.89	5.47	5.92	4.82	3.64	4.75
$R^{(9)}_{BOOT}$	ME	3.56	6.28	11.35	25.98	50.04	74.07	88.55	93.67	96.45	1.45
	SD	3.76	4.91	6.02	5.51	0.88	5.52	6.04	4.94	3.76	4.84
$R^{(10)}_{BOOT}$	ME	3.69	6.45	11.55	26.16	50.04	73.90	88.35	93.50	96.31	1.65
	SD	3.89	5.04	6.12	5.56	0.88	5.58	6.14	5.07	3.89	4.94
$R^{(11)}_{BOOT}$	ME	3.81	6.62	11.75	26.32	50.04	73.74	88.15	93.33	96.19	1.85
	SD	4.01	5.16	6.22	5.61	0.88	5.62	6.24	5.19	4.01	5.02
$R^{(12)}_{BOOT}$	ME	3.94	6.78	11.94	26.48	50.04	73.58	87.96	93.17	96.06	2.04
	SD	4.14	5.28	6.33	5.65	0.88	5.66	6.34	5.30	4.13	5.11

NSC–0.25 $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.63	5.01	9.72	24.37	50.08	75.56	90.20	95.00	97.41	0.63
	SD	3.09	4.39	5.80	5.88	0.96	5.96	5.82	4.39	3.08	4.67
$R^{(5)}_{BOOT}$	ME	2.76	5.20	9.96	24.61	50.08	75.33	89.95	94.82	97.28	0.39
	SD	3.20	4.49	5.88	5.90	0.95	5.98	5.91	4.50	3.18	4.73
$R^{(6)}_{BOOT}$	ME	2.88	5.38	10.20	24.83	50.08	75.11	89.72	94.64	97.16	0.38
	SD	3.30	4.60	5.97	5.93	0.95	6.01	6.00	4.61	3.29	4.81
$R^{(7)}_{BOOT}$	ME	3.01	5.55	10.43	25.05	50.08	74.90	89.50	94.47	97.04	0.55
	SD	3.42	4.71	6.06	5.96	0.95	6.04	6.09	4.72	3.39	4.88
$R^{(8)}_{BOOT}$	ME	3.13	5.72	10.64	25.25	50.08	74.70	89.28	94.31	96.91	0.72
	SD	3.53	4.82	6.16	6.00	0.95	6.07	6.18	4.82	3.50	4.96
$R^{(9)}_{BOOT}$	ME	3.25	5.89	10.85	25.44	50.08	74.50	89.07	94.15	96.80	0.93
	SD	3.64	4.94	6.25	6.04	0.95	6.10	6.28	4.93	3.61	5.03
$R^{(10)}_{BOOT}$	ME	3.37	6.05	11.05	25.63	50.08	74.32	88.87	93.98	96.68	1.13
	SD	3.76	5.06	6.35	6.08	0.95	6.13	6.37	5.05	3.72	5.11
$R^{(11)}_{BOOT}$	ME	3.49	6.21	11.24	25.80	50.08	74.14	88.68	93.82	96.55	1.32
	SD	3.87	5.17	6.44	6.12	0.95	6.17	6.46	5.16	3.84	5.19
$R^{(12)}_{BOOT}$	ME	3.61	6.37	11.43	25.96	50.08	73.98	88.49	93.66	96.44	1.51
	SD	3.98	5.28	6.53	6.16	0.95	6.20	6.55	5.27	3.95	5.27

Tabela B.6: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a Normal simetricamente contaminada e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NSC–0.25 $n = 100$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.57	5.03	9.93	24.76	50.07	75.15	90.03	94.95	97.42	0.24
	SD	2.01	2.88	3.81	3.73	0.66	3.70	3.80	2.89	2.00	3.01
$R^{(5)}_{BOOT}$	ME	2.65	5.14	10.08	24.90	50.07	75.02	89.88	94.84	97.34	0.16
	SD	2.06	2.94	3.85	3.74	0.66	3.71	3.85	2.94	2.05	3.05
$R^{(6)}_{BOOT}$	ME	2.73	5.25	10.22	25.03	50.07	74.89	89.74	94.74	97.26	0.26
	SD	2.11	3.00	3.90	3.75	0.66	3.73	3.90	3.00	2.11	3.09
$R^{(7)}_{BOOT}$	ME	2.80	5.36	10.36	25.16	50.07	74.76	89.61	94.63	97.19	0.39
	SD	2.17	3.06	3.95	3.78	0.66	3.76	3.95	3.06	2.17	3.13
$R^{(8)}_{BOOT}$	ME	2.88	5.46	10.49	25.28	50.07	74.64	89.48	94.52	97.11	0.53
	SD	2.22	3.12	4.01	3.80	0.66	3.78	4.00	3.12	2.23	3.17
$R^{(9)}_{BOOT}$	ME	2.95	5.57	10.62	25.40	50.07	74.52	89.34	94.42	97.03	0.66
	SD	2.28	3.19	4.06	3.82	0.65	3.81	4.06	3.19	2.29	3.22
$R^{(10)}_{BOOT}$	ME	3.03	5.67	10.75	25.51	50.07	74.40	89.22	94.31	96.96	0.78
	SD	2.35	3.26	4.11	3.85	0.65	3.84	4.12	3.25	2.36	3.27
$R^{(11)}_{BOOT}$	ME	3.10	5.77	10.87	25.62	50.07	74.29	89.09	94.22	96.88	0.91
	SD	2.41	3.32	4.18	3.88	0.65	3.87	4.18	3.31	2.42	3.32
$R^{(12)}_{BOOT}$	ME	3.18	5.87	11.00	25.73	50.07	74.18	88.97	94.12	96.80	1.03
	SD	2.48	3.39	4.24	3.92	0.65	3.90	4.24	3.38	2.49	3.37

NSC–0.25 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.65	5.08	9.98	24.83	50.06	75.18	89.99	94.90	97.37	0.18
	SD	2.37	3.40	4.49	4.38	0.76	4.38	4.44	3.36	2.33	3.54
$R^{(5)}_{BOOT}$	ME	2.72	5.18	10.12	24.97	50.06	75.04	89.84	94.79	97.30	0.22
	SD	2.43	3.46	4.53	4.39	0.76	4.39	4.48	3.42	2.38	3.57
$R^{(6)}_{BOOT}$	ME	2.80	5.28	10.26	25.11	50.06	74.91	89.70	94.69	97.22	0.31
	SD	2.49	3.51	4.58	4.40	0.76	4.40	4.53	3.48	2.44	3.61
$R^{(7)}_{BOOT}$	ME	2.87	5.39	10.39	25.24	50.06	74.78	89.56	94.58	97.15	0.44
	SD	2.55	3.58	4.62	4.41	0.76	4.42	4.58	3.54	2.49	3.65
$R^{(8)}_{BOOT}$	ME	2.95	5.50	10.53	25.36	50.06	74.66	89.43	94.47	97.07	0.57
	SD	2.62	3.65	4.68	4.43	0.76	4.44	4.63	3.60	2.56	3.70
$R^{(9)}_{BOOT}$	ME	3.02	5.60	10.66	25.48	50.06	74.54	89.30	94.37	97.00	0.70
	SD	2.69	3.71	4.73	4.44	0.76	4.46	4.69	3.67	2.63	3.74
$R^{(10)}_{BOOT}$	ME	3.10	5.70	10.78	25.59	50.06	74.43	89.17	94.27	96.92	0.83
	SD	2.76	3.78	4.78	4.46	0.75	4.48	4.74	3.73	2.69	3.79
$R^{(11)}_{BOOT}$	ME	3.17	5.79	10.91	25.70	50.06	74.32	89.05	94.17	96.85	0.95
	SD	2.82	3.85	4.84	4.49	0.75	4.50	4.79	3.80	2.77	3.83
$R^{(12)}_{BOOT}$	ME	3.24	5.90	11.03	25.81	50.06	74.22	88.93	94.07	96.77	1.07
	SD	2.89	3.91	4.90	4.51	0.75	4.53	4.85	3.87	2.84	3.88

Tabela B.7: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.1 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

M4–0.1 $n = 20$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.84	5.19	9.82	24.08	50.01	75.89	90.20	94.79	97.17	0.92
	SD	2.82	4.14	5.78	6.45	1.20	6.39	5.79	4.20	2.85	4.73
$R^{(5)}_{BOOT}$	ME	2.98	5.40	10.09	24.36	50.01	75.60	89.92	94.58	97.03	0.64
	SD	2.90	4.21	5.83	6.44	1.20	6.39	5.85	4.28	2.92	4.77
$R^{(6)}_{BOOT}$	ME	3.12	5.60	10.34	24.61	50.01	75.35	89.67	94.38	96.89	0.62
	SD	2.98	4.30	5.91	6.45	1.20	6.40	5.92	4.36	3.01	4.82
$R^{(7)}_{BOOT}$	ME	3.26	5.79	10.59	24.85	50.01	75.10	89.42	94.19	96.76	0.81
	SD	3.08	4.41	5.99	6.47	1.19	6.42	6.00	4.46	3.10	4.89
$R^{(8)}_{BOOT}$	ME	3.40	5.98	10.83	25.07	50.01	74.88	89.19	93.99	96.62	1.01
	SD	3.19	4.52	6.07	6.48	1.19	6.45	6.08	4.57	3.20	4.95
$R^{(9)}_{BOOT}$	ME	3.54	6.17	11.05	25.29	50.01	74.67	88.96	93.81	96.48	1.19
	SD	3.31	4.63	6.16	6.51	1.19	6.47	6.17	4.68	3.31	5.02
$R^{(10)}_{BOOT}$	ME	3.67	6.35	11.27	25.48	50.01	74.47	88.74	93.64	96.35	1.36
	SD	3.43	4.74	6.25	6.54	1.19	6.50	6.26	4.79	3.43	5.09
$R^{(11)}_{BOOT}$	ME	3.81	6.52	11.48	25.67	50.01	74.29	88.53	93.46	96.21	1.54
	SD	3.55	4.85	6.34	6.57	1.19	6.53	6.35	4.91	3.55	5.17
$R^{(12)}_{BOOT}$	ME	3.94	6.68	11.69	25.84	50.01	74.11	88.33	93.29	96.08	1.71
	SD	3.67	4.96	6.42	6.60	1.19	6.56	6.43	5.02	3.67	5.24
M4–0.1 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.72	5.00	9.54	23.85	50.07	75.98	90.43	94.99	97.31	1.15
	SD	2.47	3.66	5.11	5.80	1.30	5.81	5.06	3.60	2.40	4.20
$R^{(5)}_{BOOT}$	ME	2.77	5.08	9.67	23.99	50.07	75.83	90.30	94.91	97.26	1.01
	SD	2.49	3.67	5.10	5.75	1.30	5.76	5.06	3.61	2.41	4.19
$R^{(6)}_{BOOT}$	ME	2.83	5.17	9.79	24.13	50.07	75.70	90.19	94.82	97.21	0.87
	SD	2.51	3.69	5.11	5.71	1.30	5.73	5.06	3.63	2.43	4.19
$R^{(7)}_{BOOT}$	ME	2.88	5.25	9.90	24.25	50.07	75.57	90.07	94.74	97.15	0.75
	SD	2.54	3.71	5.12	5.68	1.30	5.71	5.08	3.65	2.46	4.19
$R^{(8)}_{BOOT}$	ME	2.94	5.33	10.01	24.37	50.07	75.45	89.96	94.66	97.09	0.63
	SD	2.57	3.74	5.13	5.66	1.30	5.69	5.09	3.68	2.50	4.20
$R^{(9)}_{BOOT}$	ME	3.00	5.41	10.12	24.48	50.07	75.34	89.85	94.58	97.04	0.52
	SD	2.61	3.78	5.16	5.64	1.30	5.67	5.12	3.72	2.53	4.21
$R^{(10)}_{BOOT}$	ME	3.05	5.49	10.22	24.59	50.07	75.23	89.75	94.50	96.98	0.55
	SD	2.66	3.82	5.18	5.63	1.30	5.66	5.15	3.76	2.58	4.23
$R^{(11)}_{BOOT}$	ME	3.10	5.56	10.32	24.69	50.07	75.13	89.64	94.42	96.93	0.60
	SD	2.70	3.86	5.21	5.62	1.30	5.66	5.17	3.81	2.63	4.25
$R^{(12)}_{BOOT}$	ME	3.16	5.63	10.42	24.79	50.07	75.04	89.55	94.34	96.87	0.66
	SD	2.75	3.90	5.23	5.61	1.30	5.66	5.21	3.86	2.68	4.28

Tabela B.8: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4-0.1 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

M4–0.1 $n = 50$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.72	5.17	9.98	24.71	49.90	75.30	90.06	94.85	97.28	0.30
	SD	2.09	3.06	4.11	4.24	0.77	4.25	4.15	3.09	2.14	3.31
$R^{(5)}_{BOOT}$	ME	2.79	5.27	10.13	24.86	49.90	75.15	89.91	94.75	97.21	0.29
	SD	2.11	3.08	4.12	4.23	0.76	4.24	4.15	3.11	2.16	3.31
$R^{(6)}_{BOOT}$	ME	2.87	5.38	10.27	25.00	49.90	75.01	89.76	94.64	97.13	0.38
	SD	2.14	3.10	4.13	4.22	0.76	4.24	4.17	3.13	2.18	3.33
$R^{(7)}_{BOOT}$	ME	2.94	5.48	10.41	25.14	49.90	74.88	89.63	94.54	97.06	0.48
	SD	2.17	3.13	4.15	4.22	0.76	4.24	4.19	3.16	2.22	3.34
$R^{(8)}_{BOOT}$	ME	3.01	5.59	10.55	25.26	49.90	74.75	89.49	94.43	96.98	0.59
	SD	2.21	3.17	4.18	4.23	0.76	4.26	4.22	3.20	2.25	3.37
$R^{(9)}_{BOOT}$	ME	3.09	5.70	10.68	25.38	49.90	74.63	89.35	94.33	96.91	0.70
	SD	2.25	3.21	4.21	4.24	0.76	4.27	4.25	3.24	2.29	3.40
$R^{(10)}_{BOOT}$	ME	3.16	5.80	10.81	25.50	49.90	74.51	89.22	94.22	96.83	0.81
	SD	2.30	3.26	4.25	4.25	0.75	4.29	4.29	3.29	2.34	3.43
$R^{(11)}_{BOOT}$	ME	3.24	5.90	10.94	25.62	49.90	74.39	89.09	94.12	96.75	0.94
	SD	2.35	3.31	4.29	4.27	0.75	4.31	4.34	3.33	2.39	3.46
$R^{(12)}_{BOOT}$	ME	3.31	6.00	11.06	25.73	49.90	74.29	88.97	94.02	96.68	1.06
	SD	2.40	3.36	4.34	4.29	0.75	4.33	4.38	3.38	2.43	3.50
M4–0.1 $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.59	4.94	9.70	24.39	49.95	75.67	90.36	95.06	97.41	0.67
	SD	1.78	2.64	3.68	3.89	0.83	3.89	3.64	2.64	1.78	2.94
$R^{(5)}_{BOOT}$	ME	2.61	4.96	9.74	24.44	49.95	75.63	90.32	95.03	97.39	0.63
	SD	1.77	2.63	3.67	3.87	0.83	3.87	3.63	2.63	1.77	2.93
$R^{(6)}_{BOOT}$	ME	2.63	4.99	9.78	24.48	49.95	75.58	90.28	95.00	97.37	0.58
	SD	1.77	2.63	3.65	3.85	0.83	3.86	3.62	2.63	1.77	2.92
$R^{(7)}_{BOOT}$	ME	2.64	5.02	9.82	24.53	49.95	75.54	90.24	94.97	97.35	0.54
	SD	1.77	2.62	3.64	3.84	0.83	3.84	3.61	2.63	1.77	2.92
$R^{(8)}_{BOOT}$	ME	2.66	5.05	9.86	24.57	49.95	75.50	90.20	94.94	97.34	0.50
	SD	1.77	2.62	3.64	3.82	0.83	3.83	3.60	2.63	1.77	2.91
$R^{(9)}_{BOOT}$	ME	2.68	5.07	9.90	24.61	49.95	75.46	90.16	94.91	97.32	0.46
	SD	1.77	2.62	3.63	3.81	0.83	3.82	3.60	2.63	1.77	2.91
$R^{(10)}_{BOOT}$	ME	2.70	5.10	9.94	24.65	49.95	75.42	90.12	94.88	97.30	0.42
	SD	1.77	2.62	3.63	3.81	0.83	3.81	3.60	2.63	1.78	2.90
$R^{(11)}_{BOOT}$	ME	2.72	5.13	9.97	24.69	49.95	75.38	90.08	94.86	97.28	0.38
	SD	1.78	2.62	3.63	3.80	0.83	3.81	3.60	2.63	1.78	2.90
$R^{(12)}_{BOOT}$	ME	2.74	5.16	10.01	24.73	49.95	75.34	90.04	94.83	97.26	0.34
	SD	1.78	2.62	3.63	3.79	0.82	3.80	3.60	2.63	1.79	2.90

Tabela B.9: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4-0.1 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

M4–0.1 $n = 100$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.66	5.17	10.09	24.91	49.83	74.96	89.91	94.88	97.33	0.17
	SD	1.53	2.24	2.98	2.95	0.69	2.95	2.99	2.23	1.53	2.36
$R^{(5)}_{BOOT}$	ME	2.70	5.23	10.16	24.98	49.83	74.89	89.83	94.82	97.29	0.23
	SD	1.53	2.24	2.97	2.93	0.69	2.93	2.98	2.23	1.53	2.36
$R^{(6)}_{BOOT}$	ME	2.74	5.28	10.24	25.06	49.83	74.82	89.76	94.77	97.26	0.28
	SD	1.54	2.24	2.97	2.91	0.69	2.91	2.98	2.24	1.54	2.36
$R^{(7)}_{BOOT}$	ME	2.77	5.34	10.31	25.13	49.83	74.75	89.68	94.71	97.22	0.34
	SD	1.55	2.25	2.97	2.90	0.69	2.90	2.98	2.24	1.55	2.35
$R^{(8)}_{BOOT}$	ME	2.81	5.39	10.39	25.20	49.83	74.68	89.61	94.66	97.18	0.39
	SD	1.56	2.25	2.97	2.89	0.69	2.89	2.98	2.25	1.56	2.35
$R^{(9)}_{BOOT}$	ME	2.85	5.45	10.46	25.27	49.83	74.61	89.54	94.61	97.15	0.46
	SD	1.57	2.26	2.98	2.88	0.69	2.88	2.98	2.25	1.57	2.36
$R^{(10)}_{BOOT}$	ME	2.88	5.50	10.53	25.34	49.83	74.54	89.47	94.55	97.11	0.53
	SD	1.58	2.28	2.98	2.88	0.69	2.88	2.99	2.27	1.58	2.36
$R^{(11)}_{BOOT}$	ME	2.92	5.56	10.60	25.41	49.83	74.48	89.40	94.50	97.07	0.60
	SD	1.60	2.29	2.99	2.88	0.69	2.88	3.00	2.28	1.59	2.37
$R^{(12)}_{BOOT}$	ME	2.96	5.61	10.67	25.47	49.83	74.41	89.33	94.45	97.03	0.67
	SD	1.61	2.31	3.00	2.88	0.69	2.87	3.01	2.30	1.61	2.38
M4–0.1 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.59	5.07	9.93	24.81	49.93	75.28	90.10	95.02	97.44	0.28
	SD	1.34	1.98	2.64	2.70	0.72	2.72	2.61	1.92	1.29	2.11
$R^{(5)}_{BOOT}$	ME	2.60	5.08	9.95	24.83	49.93	75.26	90.09	95.01	97.43	0.26
	SD	1.34	1.97	2.63	2.69	0.72	2.70	2.61	1.92	1.29	2.10
$R^{(6)}_{BOOT}$	ME	2.61	5.09	9.97	24.85	49.93	75.24	90.07	94.99	97.42	0.24
	SD	1.34	1.97	2.62	2.67	0.72	2.69	2.60	1.92	1.29	2.10
$R^{(7)}_{BOOT}$	ME	2.62	5.10	9.99	24.87	49.93	75.22	90.05	94.98	97.41	0.22
	SD	1.34	1.96	2.61	2.66	0.72	2.68	2.59	1.91	1.29	2.09
$R^{(8)}_{BOOT}$	ME	2.63	5.11	10.01	24.89	49.93	75.20	90.03	94.97	97.40	0.20
	SD	1.34	1.96	2.60	2.65	0.72	2.67	2.58	1.91	1.28	2.08
$R^{(9)}_{BOOT}$	ME	2.63	5.13	10.03	24.91	49.93	75.18	90.01	94.96	97.40	0.18
	SD	1.34	1.95	2.59	2.64	0.72	2.66	2.58	1.91	1.28	2.08
$R^{(10)}_{BOOT}$	ME	2.64	5.14	10.04	24.92	49.93	75.16	90.00	94.94	97.39	0.16
	SD	1.33	1.95	2.59	2.63	0.72	2.65	2.57	1.91	1.28	2.07
$R^{(11)}_{BOOT}$	ME	2.65	5.15	10.06	24.94	49.93	75.15	89.98	94.93	97.38	0.15
	SD	1.33	1.95	2.58	2.62	0.72	2.64	2.56	1.90	1.28	2.07
$R^{(12)}_{BOOT}$	ME	2.66	5.16	10.08	24.96	49.93	75.13	89.96	94.92	97.37	0.16
	SD	1.33	1.95	2.58	2.62	0.72	2.64	2.56	1.90	1.28	2.06

Tabela B.10: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.2 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

M4–0.2 $n = 20$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.79	5.07	9.68	23.98	49.98	75.83	90.17	94.80	97.12	1.02
	SD	2.54	3.82	5.59	7.58	6.14	8.11	5.93	4.03	2.71	5.50
$R^{(2)}_{BOOT}$	ME	2.81	5.07	9.61	23.71	49.99	76.20	90.41	94.94	97.20	1.29
	SD	2.56	3.84	5.56	7.02	1.10	7.09	5.55	3.80	2.55	4.77
$R^{(3)}_{BOOT}$	ME	2.87	5.17	9.77	23.95	49.99	75.96	90.24	94.83	97.13	1.05
	SD	2.58	3.84	5.51	6.84	1.09	6.92	5.51	3.80	2.56	4.70
$R^{(4)}_{BOOT}$	ME	2.94	5.28	9.94	24.17	49.99	75.75	90.07	94.73	97.06	0.83
	SD	2.61	3.84	5.47	6.72	1.09	6.80	5.48	3.80	2.58	4.66
$R^{(5)}_{BOOT}$	ME	3.00	5.39	10.10	24.37	49.99	75.55	89.91	94.63	97.00	0.63
	SD	2.63	3.85	5.44	6.63	1.09	6.71	5.46	3.81	2.60	4.63
$R^{(6)}_{BOOT}$	ME	3.07	5.49	10.26	24.56	49.99	75.37	89.75	94.51	96.92	0.58
	SD	2.67	3.87	5.43	6.55	1.09	6.64	5.45	3.83	2.63	4.61
$R^{(7)}_{BOOT}$	ME	3.14	5.60	10.40	24.73	49.99	75.20	89.61	94.41	96.85	0.65
	SD	2.70	3.89	5.43	6.49	1.09	6.59	5.45	3.85	2.66	4.60
$R^{(8)}_{BOOT}$	ME	3.22	5.70	10.55	24.89	49.99	75.04	89.46	94.29	96.77	0.73
	SD	2.74	3.93	5.43	6.44	1.09	6.54	5.46	3.89	2.71	4.60
$R^{(9)}_{BOOT}$	ME	3.29	5.81	10.70	25.05	49.99	74.89	89.32	94.19	96.69	0.81
	SD	2.79	3.96	5.44	6.40	1.09	6.50	5.48	3.93	2.75	4.60
$R^{(10)}_{BOOT}$	ME	3.36	5.91	10.83	25.19	49.99	74.74	89.19	94.08	96.62	0.92
	SD	2.83	4.00	5.46	6.37	1.09	6.48	5.50	3.98	2.80	4.61
$R^{(11)}_{BOOT}$	ME	3.43	6.02	10.96	25.32	49.99	74.61	89.05	93.98	96.54	1.02
	SD	2.89	4.04	5.48	6.35	1.09	6.46	5.53	4.03	2.86	4.63
$R^{(12)}_{BOOT}$	ME	3.51	6.12	11.09	25.45	49.99	74.48	88.93	93.88	96.47	1.12
	SD	2.94	4.08	5.52	6.34	1.09	6.44	5.55	4.08	2.91	4.65

Tabela B.11: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.2 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

M4–0.2 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.32	4.44	8.79	23.25	50.17	76.85	91.21	95.55	97.68	1.85
	SD	1.90	2.94	4.41	6.00	4.99	6.23	4.46	2.94	1.90	4.26
$R^{(2)}_{BOOT}$	ME	2.32	4.42	8.72	22.96	50.04	77.01	91.28	95.59	97.69	2.04
	SD	1.89	2.92	4.40	5.68	1.16	5.64	4.35	2.91	1.88	3.77
$R^{(3)}_{BOOT}$	ME	2.35	4.46	8.80	23.12	50.04	76.86	91.19	95.54	97.66	1.88
	SD	1.88	2.91	4.34	5.51	1.16	5.47	4.29	2.88	1.87	3.69
$R^{(4)}_{BOOT}$	ME	2.38	4.52	8.89	23.27	50.05	76.72	91.10	95.48	97.63	1.73
	SD	1.89	2.90	4.29	5.37	1.16	5.35	4.25	2.86	1.86	3.64
$R^{(5)}_{BOOT}$	ME	2.41	4.57	8.98	23.41	50.05	76.59	91.00	95.42	97.59	1.59
	SD	1.89	2.89	4.25	5.27	1.16	5.27	4.21	2.85	1.86	3.59
$R^{(6)}_{BOOT}$	ME	2.44	4.62	9.07	23.53	50.05	76.47	90.91	95.37	97.56	1.47
	SD	1.90	2.89	4.22	5.19	1.16	5.20	4.18	2.85	1.86	3.56
$R^{(7)}_{BOOT}$	ME	2.48	4.68	9.16	23.64	50.05	76.36	90.82	95.31	97.52	1.36
	SD	1.91	2.89	4.20	5.13	1.16	5.14	4.17	2.85	1.87	3.54
$R^{(8)}_{BOOT}$	ME	2.51	4.73	9.25	23.74	50.05	76.26	90.74	95.25	97.49	1.26
	SD	1.92	2.89	4.19	5.09	1.16	5.09	4.17	2.86	1.88	3.52
$R^{(9)}_{BOOT}$	ME	2.55	4.79	9.33	23.84	50.05	76.17	90.65	95.19	97.45	1.17
	SD	1.95	2.90	4.19	5.05	1.16	5.05	4.17	2.88	1.90	3.52
$R^{(10)}_{BOOT}$	ME	2.59	4.84	9.42	23.93	50.05	76.07	90.58	95.13	97.41	1.07
	SD	1.96	2.91	4.19	5.02	1.16	5.02	4.18	2.90	1.92	3.52
$R^{(11)}_{BOOT}$	ME	2.62	4.90	9.50	24.03	50.05	75.99	90.50	95.08	97.37	0.99
	SD	1.99	2.94	4.20	4.99	1.16	5.00	4.19	2.93	1.95	3.52
$R^{(12)}_{BOOT}$	ME	2.66	4.96	9.57	24.11	50.05	75.90	90.42	95.02	97.33	0.90
	SD	2.02	2.96	4.22	4.98	1.16	4.98	4.21	2.96	1.99	3.53

Tabela B.12: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.2 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

M4–0.2 $n = 50$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.62	5.08	9.97	24.91	50.12	75.18	90.02	94.89	97.31	0.19
	SD	1.79	2.71	3.86	5.21	5.61	5.65	4.05	2.79	1.85	3.99
$R^{(2)}_{BOOT}$	ME	2.57	5.00	9.78	24.56	49.95	75.35	90.21	94.99	97.38	0.44
	SD	1.72	2.59	3.59	3.87	0.68	3.87	3.61	2.61	1.75	2.90
$R^{(3)}_{BOOT}$	ME	2.60	5.05	9.87	24.66	49.95	75.25	90.13	94.94	97.34	0.34
	SD	1.71	2.57	3.54	3.78	0.68	3.79	3.56	2.59	1.74	2.86
$R^{(4)}_{BOOT}$	ME	2.64	5.10	9.95	24.75	49.95	75.16	90.04	94.89	97.31	0.25
	SD	1.71	2.56	3.50	3.71	0.68	3.72	3.53	2.57	1.74	2.83
$R^{(5)}_{BOOT}$	ME	2.67	5.16	10.03	24.84	49.95	75.06	89.96	94.83	97.28	0.22
	SD	1.71	2.54	3.46	3.65	0.68	3.66	3.50	2.56	1.74	2.80
$R^{(6)}_{BOOT}$	ME	2.71	5.21	10.11	24.93	49.95	74.97	89.88	94.78	97.24	0.26
	SD	1.71	2.53	3.43	3.60	0.68	3.61	3.47	2.55	1.74	2.77
$R^{(7)}_{BOOT}$	ME	2.74	5.27	10.19	25.02	49.96	74.89	89.80	94.72	97.21	0.29
	SD	1.71	2.52	3.41	3.55	0.68	3.57	3.45	2.55	1.74	2.75
$R^{(8)}_{BOOT}$	ME	2.78	5.32	10.27	25.10	49.95	74.81	89.73	94.66	97.17	0.34
	SD	1.71	2.52	3.40	3.52	0.68	3.54	3.44	2.54	1.74	2.74
$R^{(9)}_{BOOT}$	ME	2.81	5.37	10.34	25.18	49.95	74.73	89.65	94.61	97.13	0.39
	SD	1.72	2.52	3.38	3.49	0.68	3.51	3.42	2.54	1.75	2.73
$R^{(10)}_{BOOT}$	ME	2.85	5.43	10.42	25.25	49.95	74.66	89.57	94.55	97.10	0.45
	SD	1.73	2.52	3.38	3.47	0.68	3.49	3.42	2.54	1.75	2.72
$R^{(11)}_{BOOT}$	ME	2.89	5.49	10.49	25.32	49.95	74.59	89.50	94.50	97.06	0.50
	SD	1.74	2.53	3.37	3.45	0.68	3.47	3.42	2.55	1.76	2.72
$R^{(12)}_{BOOT}$	ME	2.93	5.54	10.57	25.40	49.95	74.52	89.42	94.45	97.02	0.58
	SD	1.75	2.53	3.37	3.43	0.68	3.46	3.42	2.56	1.78	2.72

Tabela B.13: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.2 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

M4–0.2 $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.58	4.99	9.82	24.53	49.99	75.42	90.24	95.02	97.39	0.47
	SD	1.34	2.02	2.87	3.48	2.69	3.29	2.80	2.02	1.33	2.54
$R^{(2)}_{BOOT}$	ME	2.56	4.96	9.74	24.44	49.98	75.45	90.25	95.02	97.39	0.56
	SD	1.31	1.96	2.73	2.91	0.72	2.94	2.74	1.99	1.32	2.21
$R^{(3)}_{BOOT}$	ME	2.57	4.97	9.76	24.46	49.98	75.43	90.23	95.01	97.39	0.54
	SD	1.30	1.95	2.72	2.89	0.72	2.92	2.74	1.98	1.32	2.20
$R^{(4)}_{BOOT}$	ME	2.57	4.98	9.78	24.47	49.98	75.41	90.22	94.99	97.38	0.53
	SD	1.30	1.95	2.72	2.87	0.72	2.90	2.73	1.98	1.32	2.19
$R^{(5)}_{BOOT}$	ME	2.58	4.99	9.79	24.49	49.98	75.39	90.20	94.98	97.37	0.51
	SD	1.30	1.95	2.71	2.85	0.72	2.88	2.72	1.98	1.32	2.18
$R^{(6)}_{BOOT}$	ME	2.59	5.00	9.81	24.51	49.98	75.37	90.18	94.97	97.36	0.49
	SD	1.30	1.94	2.70	2.84	0.72	2.86	2.71	1.98	1.32	2.17
$R^{(7)}_{BOOT}$	ME	2.60	5.01	9.83	24.53	49.98	75.35	90.16	94.96	97.36	0.47
	SD	1.30	1.94	2.69	2.83	0.72	2.85	2.70	1.97	1.32	2.17
$R^{(8)}_{BOOT}$	ME	2.60	5.02	9.84	24.55	49.98	75.33	90.15	94.95	97.35	0.45
	SD	1.30	1.94	2.68	2.82	0.71	2.83	2.69	1.97	1.32	2.16
$R^{(9)}_{BOOT}$	ME	2.61	5.04	9.86	24.57	49.98	75.31	90.13	94.93	97.34	0.43
	SD	1.30	1.94	2.68	2.81	0.71	2.82	2.69	1.97	1.32	2.16
$R^{(10)}_{BOOT}$	ME	2.62	5.05	9.88	24.59	49.98	75.29	90.12	94.92	97.33	0.41
	SD	1.30	1.94	2.67	2.80	0.71	2.81	2.68	1.97	1.32	2.15
$R^{(11)}_{BOOT}$	ME	2.63	5.06	9.89	24.60	49.98	75.28	90.10	94.91	97.33	0.40
	SD	1.30	1.93	2.67	2.79	0.71	2.81	2.68	1.96	1.32	2.15
$R^{(12)}_{BOOT}$	ME	2.63	5.07	9.91	24.62	49.98	75.26	90.08	94.90	97.32	0.38
	SD	1.30	1.93	2.66	2.78	0.71	2.80	2.68	1.96	1.32	2.14

Tabela B.14: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4–0.2 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

M4–0.2 $n = 100$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.72	5.19	10.16	25.01	49.96	74.99	89.89	94.82	97.31	0.22
	SD	1.53	2.32	3.37	4.64	4.80	4.51	3.30	2.29	1.50	3.37
$R^{(2)}_{BOOT}$	ME	2.65	5.09	10.01	24.82	49.97	75.17	90.01	94.91	97.36	0.18
	SD	1.42	2.11	2.92	3.04	0.66	3.04	2.91	2.12	1.40	2.33
$R^{(3)}_{BOOT}$	ME	2.67	5.11	10.04	24.86	49.97	75.13	89.97	94.89	97.35	0.17
	SD	1.42	2.11	2.90	3.00	0.66	3.01	2.90	2.11	1.40	2.31
$R^{(4)}_{BOOT}$	ME	2.68	5.13	10.07	24.90	49.97	75.09	89.94	94.86	97.33	0.18
	SD	1.42	2.10	2.88	2.97	0.66	2.98	2.88	2.10	1.40	2.30
$R^{(5)}_{BOOT}$	ME	2.70	5.16	10.11	24.93	49.97	75.06	89.90	94.84	97.32	0.20
	SD	1.41	2.09	2.87	2.95	0.66	2.95	2.86	2.10	1.40	2.29
$R^{(6)}_{BOOT}$	ME	2.71	5.18	10.14	24.97	49.97	75.02	89.87	94.82	97.30	0.21
	SD	1.41	2.08	2.85	2.92	0.66	2.92	2.85	2.09	1.40	2.27
$R^{(7)}_{BOOT}$	ME	2.73	5.20	10.17	25.01	49.97	74.99	89.84	94.79	97.29	0.23
	SD	1.41	2.08	2.84	2.90	0.66	2.90	2.83	2.08	1.40	2.26
$R^{(8)}_{BOOT}$	ME	2.74	5.22	10.20	25.04	49.97	74.95	89.81	94.77	97.27	0.24
	SD	1.41	2.07	2.82	2.88	0.66	2.88	2.82	2.08	1.40	2.25
$R^{(9)}_{BOOT}$	ME	2.76	5.25	10.24	25.08	49.97	74.92	89.77	94.75	97.26	0.26
	SD	1.41	2.07	2.81	2.86	0.66	2.86	2.81	2.08	1.40	2.24
$R^{(10)}_{BOOT}$	ME	2.77	5.27	10.27	25.11	49.97	74.88	89.74	94.72	97.24	0.28
	SD	1.41	2.06	2.80	2.84	0.66	2.84	2.80	2.07	1.40	2.23
$R^{(11)}_{BOOT}$	ME	2.79	5.29	10.30	25.14	49.97	74.85	89.71	94.70	97.23	0.30
	SD	1.41	2.06	2.79	2.83	0.66	2.83	2.79	2.07	1.40	2.22
$R^{(12)}_{BOOT}$	ME	2.80	5.32	10.33	25.17	49.97	74.82	89.68	94.68	97.21	0.33
	SD	1.41	2.06	2.78	2.81	0.66	2.81	2.78	2.07	1.40	2.21

Tabela B.15: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição M4-0.2 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

M4–0.2 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.49	4.86	9.72	24.62	49.90	75.41	90.34	95.17	97.54	0.41
	SD	1.04	1.55	2.21	2.63	2.10	2.62	2.22	1.57	1.04	1.97
$R^{(2)}_{BOOT}$	ME	2.47	4.84	9.69	24.58	49.89	75.43	90.35	95.18	97.55	0.43
	SD	1.02	1.51	2.10	2.20	0.67	2.16	2.10	1.53	1.01	1.68
$R^{(3)}_{BOOT}$	ME	2.47	4.85	9.70	24.59	49.89	75.43	90.35	95.18	97.55	0.43
	SD	1.02	1.51	2.10	2.20	0.67	2.15	2.09	1.52	1.01	1.68
$R^{(4)}_{BOOT}$	ME	2.48	4.85	9.70	24.60	49.89	75.42	90.34	95.17	97.54	0.42
	SD	1.02	1.51	2.10	2.19	0.67	2.15	2.09	1.52	1.01	1.68
$R^{(5)}_{BOOT}$	ME	2.48	4.86	9.71	24.60	49.89	75.41	90.33	95.17	97.54	0.41
	SD	1.02	1.51	2.09	2.19	0.67	2.14	2.09	1.52	1.01	1.67
$R^{(6)}_{BOOT}$	ME	2.48	4.86	9.72	24.61	49.89	75.41	90.33	95.17	97.54	0.41
	SD	1.02	1.51	2.09	2.19	0.67	2.14	2.09	1.52	1.01	1.67
$R^{(7)}_{BOOT}$	ME	2.49	4.86	9.72	24.61	49.89	75.40	90.32	95.16	97.54	0.40
	SD	1.02	1.50	2.09	2.18	0.67	2.14	2.08	1.52	1.01	1.67
$R^{(8)}_{BOOT}$	ME	2.49	4.87	9.73	24.62	49.89	75.39	90.31	95.16	97.53	0.39
	SD	1.02	1.50	2.08	2.18	0.67	2.13	2.08	1.52	1.01	1.67
$R^{(9)}_{BOOT}$	ME	2.49	4.87	9.74	24.63	49.89	75.39	90.31	95.15	97.53	0.39
	SD	1.02	1.50	2.08	2.17	0.67	2.13	2.08	1.52	1.01	1.67
$R^{(10)}_{BOOT}$	ME	2.49	4.88	9.74	24.63	49.89	75.38	90.30	95.15	97.53	0.38
	SD	1.02	1.50	2.08	2.17	0.67	2.13	2.08	1.52	1.01	1.66
$R^{(11)}_{BOOT}$	ME	2.50	4.88	9.75	24.64	49.89	75.38	90.30	95.14	97.52	0.38
	SD	1.02	1.50	2.08	2.17	0.67	2.12	2.07	1.52	1.01	1.66
$R^{(12)}_{BOOT}$	ME	2.50	4.89	9.75	24.65	49.89	75.37	90.29	95.14	97.52	0.37
	SD	1.02	1.50	2.07	2.16	0.67	2.12	2.07	1.52	1.01	1.66

Tabela B.16: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição T1 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

T1 $n = 20$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.65	5.00	9.60	24.03	50.06	76.12	90.45	95.00	97.30	1.12
	SD	3.82	5.31	6.96	7.42	1.37	7.27	6.79	5.19	3.83	5.67
$R^{(5)}_{BOOT}$	ME	2.83	5.25	9.93	24.37	50.06	75.78	90.13	94.77	97.13	0.78
	SD	4.02	5.50	7.10	7.43	1.37	7.28	6.92	5.35	4.00	5.77
$R^{(6)}_{BOOT}$	ME	3.01	5.49	10.24	24.67	50.06	75.46	89.83	94.53	96.96	0.54
	SD	4.22	5.68	7.23	7.45	1.37	7.29	7.05	5.51	4.17	5.87
$R^{(7)}_{BOOT}$	ME	3.18	5.73	10.54	24.96	50.06	75.18	89.53	94.31	96.80	0.73
	SD	4.41	5.86	7.37	7.48	1.37	7.32	7.18	5.66	4.32	5.98
$R^{(8)}_{BOOT}$	ME	3.34	5.96	10.83	25.22	50.06	74.91	89.24	94.08	96.64	0.96
	SD	4.59	6.03	7.50	7.51	1.37	7.35	7.30	5.83	4.48	6.08
$R^{(9)}_{BOOT}$	ME	3.51	6.18	11.10	25.48	50.06	74.65	88.98	93.86	96.48	1.18
	SD	4.76	6.19	7.63	7.54	1.37	7.38	7.42	5.98	4.64	6.19
$R^{(10)}_{BOOT}$	ME	3.66	6.39	11.37	25.71	50.06	74.42	88.71	93.65	96.32	1.39
	SD	4.92	6.34	7.75	7.58	1.36	7.42	7.54	6.13	4.79	6.29
$R^{(11)}_{BOOT}$	ME	3.83	6.60	11.62	25.93	50.06	74.19	88.46	93.44	96.16	1.62
	SD	5.09	6.50	7.87	7.60	1.36	7.45	7.66	6.27	4.93	6.39
$R^{(12)}_{BOOT}$	ME	3.98	6.81	11.86	26.14	50.06	73.98	88.22	93.25	96.00	1.86
	SD	5.24	6.64	7.98	7.64	1.36	7.49	7.77	6.41	5.08	6.49

T1 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	3.16	5.56	10.09	24.12	50.00	75.75	89.93	94.43	96.84	0.88
	SD	4.69	6.19	7.85	8.21	1.63	8.26	7.87	6.25	4.76	6.53
$R^{(5)}_{BOOT}$	ME	3.33	5.79	10.38	24.41	50.00	75.46	89.63	94.21	96.67	0.83
	SD	4.86	6.36	8.00	8.24	1.62	8.29	8.02	6.42	4.93	6.65
$R^{(6)}_{BOOT}$	ME	3.50	6.01	10.65	24.68	50.00	75.20	89.36	93.99	96.50	1.01
	SD	5.03	6.53	8.14	8.28	1.62	8.33	8.16	6.60	5.11	6.76
$R^{(7)}_{BOOT}$	ME	3.66	6.22	10.90	24.92	50.00	74.95	89.11	93.78	96.34	1.22
	SD	5.20	6.70	8.28	8.32	1.62	8.37	8.30	6.76	5.28	6.87
$R^{(8)}_{BOOT}$	ME	3.82	6.42	11.15	25.14	50.00	74.73	88.87	93.58	96.18	1.42
	SD	5.37	6.87	8.41	8.36	1.62	8.41	8.42	6.92	5.45	6.98
$R^{(9)}_{BOOT}$	ME	3.97	6.62	11.37	25.36	50.00	74.53	88.64	93.38	96.03	1.62
	SD	5.53	7.03	8.53	8.39	1.62	8.45	8.54	7.07	5.60	7.08
$R^{(10)}_{BOOT}$	ME	4.12	6.81	11.60	25.55	50.00	74.33	88.42	93.20	95.88	1.81
	SD	5.69	7.18	8.65	8.43	1.62	8.48	8.66	7.21	5.75	7.18
$R^{(11)}_{BOOT}$	ME	4.26	6.99	11.80	25.74	50.00	74.15	88.21	93.02	95.74	1.99
	SD	5.83	7.32	8.76	8.47	1.61	8.52	8.77	7.35	5.91	7.28
$R^{(12)}_{BOOT}$	ME	4.40	7.16	12.00	25.91	50.00	73.97	88.01	92.84	95.59	2.16
	SD	5.98	7.46	8.86	8.51	1.61	8.55	8.87	7.49	6.05	7.38

Tabela B.17: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição T1 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

T1 $n = 50$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.73	5.14	9.95	24.68	50.03	75.33	90.04	94.85	97.29	0.33
	SD	2.61	3.65	4.80	4.82	0.78	4.86	4.82	3.67	2.59	3.86
$R^{(5)}_{BOOT}$	ME	2.82	5.28	10.12	24.85	50.03	75.17	89.86	94.72	97.20	0.32
	SD	2.67	3.73	4.86	4.84	0.78	4.88	4.89	3.75	2.67	3.91
$R^{(6)}_{BOOT}$	ME	2.92	5.41	10.29	25.01	50.03	75.01	89.69	94.59	97.10	0.42
	SD	2.75	3.81	4.93	4.87	0.78	4.90	4.96	3.83	2.75	3.96
$R^{(7)}_{BOOT}$	ME	3.01	5.53	10.45	25.16	50.03	74.86	89.53	94.45	97.01	0.55
	SD	2.83	3.89	5.00	4.89	0.78	4.93	5.03	3.91	2.82	4.02
$R^{(8)}_{BOOT}$	ME	3.10	5.66	10.61	25.30	50.03	74.72	89.37	94.33	96.91	0.67
	SD	2.91	3.98	5.08	4.92	0.78	4.96	5.11	4.00	2.91	4.08
$R^{(9)}_{BOOT}$	ME	3.20	5.78	10.76	25.44	50.03	74.58	89.22	94.20	96.82	0.80
	SD	3.00	4.07	5.15	4.95	0.78	4.99	5.19	4.09	3.00	4.15
$R^{(10)}_{BOOT}$	ME	3.29	5.91	10.91	25.57	50.03	74.44	89.07	94.08	96.73	0.93
	SD	3.09	4.16	5.24	4.97	0.78	5.02	5.26	4.18	3.08	4.21
$R^{(11)}_{BOOT}$	ME	3.38	6.03	11.05	25.70	50.03	74.32	88.92	93.96	96.63	1.08
	SD	3.18	4.25	5.31	5.00	0.77	5.05	5.34	4.27	3.17	4.27
$R^{(12)}_{BOOT}$	ME	3.47	6.15	11.20	25.82	50.03	74.19	88.78	93.84	96.54	1.22
	SD	3.27	4.34	5.39	5.03	0.77	5.08	5.41	4.36	3.27	4.34
T1 $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.92	5.47	10.37	25.08	49.94	74.94	89.58	94.55	97.10	0.47
	SD	2.74	3.84	4.98	4.99	0.92	5.03	5.03	3.82	2.72	4.02
$R^{(5)}_{BOOT}$	ME	3.02	5.61	10.54	25.25	49.94	74.76	89.41	94.41	97.00	0.61
	SD	2.82	3.92	5.04	5.00	0.92	5.04	5.09	3.89	2.79	4.06
$R^{(6)}_{BOOT}$	ME	3.12	5.75	10.72	25.42	49.94	74.60	89.24	94.27	96.90	0.76
	SD	2.90	4.00	5.10	5.01	0.92	5.05	5.15	3.98	2.88	4.12
$R^{(7)}_{BOOT}$	ME	3.22	5.88	10.89	25.57	49.94	74.45	89.07	94.14	96.79	0.93
	SD	2.99	4.08	5.16	5.02	0.91	5.06	5.21	4.07	2.96	4.17
$R^{(8)}_{BOOT}$	ME	3.32	6.01	11.05	25.72	49.94	74.30	88.91	94.01	96.70	1.09
	SD	3.07	4.16	5.23	5.04	0.91	5.08	5.28	4.15	3.05	4.22
$R^{(9)}_{BOOT}$	ME	3.42	6.14	11.20	25.86	49.94	74.16	88.74	93.88	96.60	1.26
	SD	3.16	4.24	5.29	5.06	0.91	5.09	5.35	4.24	3.14	4.28
$R^{(10)}_{BOOT}$	ME	3.51	6.27	11.36	26.00	49.94	74.02	88.59	93.76	96.50	1.41
	SD	3.25	4.33	5.36	5.08	0.91	5.11	5.42	4.33	3.23	4.34
$R^{(11)}_{BOOT}$	ME	3.61	6.39	11.51	26.13	49.94	73.89	88.45	93.63	96.40	1.55
	SD	3.34	4.42	5.43	5.10	0.91	5.14	5.48	4.42	3.32	4.40
$R^{(12)}_{BOOT}$	ME	3.70	6.51	11.66	26.25	49.94	73.76	88.30	93.51	96.31	1.70
	SD	3.43	4.50	5.50	5.13	0.91	5.16	5.55	4.50	3.42	4.46

Tabela B.18: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição T1 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

T1 $n = 100$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.79	5.25	10.20	25.04	50.04	74.97	89.80	94.72	97.23	0.29
	SD	2.08	2.91	3.80	3.68	0.66	3.66	3.80	2.93	2.09	3.02
$R^{(5)}_{BOOT}$	ME	2.83	5.31	10.29	25.11	50.04	74.89	89.72	94.66	97.18	0.34
	SD	2.11	2.94	3.82	3.69	0.66	3.67	3.83	2.97	2.12	3.04
$R^{(6)}_{BOOT}$	ME	2.88	5.37	10.37	25.19	50.04	74.82	89.64	94.60	97.14	0.40
	SD	2.15	2.98	3.85	3.70	0.66	3.68	3.85	3.01	2.15	3.06
$R^{(7)}_{BOOT}$	ME	2.92	5.44	10.45	25.26	50.04	74.75	89.57	94.53	97.09	0.47
	SD	2.18	3.01	3.88	3.71	0.66	3.69	3.88	3.04	2.19	3.09
$R^{(8)}_{BOOT}$	ME	2.96	5.49	10.52	25.33	50.04	74.68	89.49	94.47	97.05	0.53
	SD	2.22	3.05	3.91	3.72	0.66	3.70	3.91	3.08	2.22	3.11
$R^{(9)}_{BOOT}$	ME	3.01	5.55	10.60	25.40	50.04	74.61	89.41	94.41	97.00	0.60
	SD	2.26	3.08	3.94	3.73	0.66	3.71	3.94	3.11	2.26	3.14
$R^{(10)}_{BOOT}$	ME	3.05	5.61	10.67	25.46	50.04	74.54	89.34	94.35	96.96	0.67
	SD	2.30	3.11	3.97	3.75	0.66	3.73	3.97	3.15	2.30	3.16
$R^{(11)}_{BOOT}$	ME	3.10	5.67	10.74	25.53	50.04	74.47	89.26	94.30	96.92	0.74
	SD	2.34	3.15	4.01	3.77	0.65	3.74	4.01	3.19	2.33	3.19
$R^{(12)}_{BOOT}$	ME	3.14	5.73	10.81	25.59	50.04	74.41	89.19	94.24	96.88	0.81
	SD	2.38	3.20	4.04	3.78	0.65	3.76	4.04	3.22	2.37	3.22

T1 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.64	5.10	10.06	24.95	50.02	75.06	89.96	94.91	97.36	0.14
	SD	1.86	2.66	3.55	3.51	0.74	3.54	3.54	2.69	1.87	2.82
$R^{(5)}_{BOOT}$	ME	2.69	5.17	10.15	25.03	50.01	74.98	89.88	94.85	97.32	0.19
	SD	1.89	2.70	3.57	3.51	0.74	3.54	3.57	2.71	1.90	2.84
$R^{(6)}_{BOOT}$	ME	2.74	5.23	10.23	25.11	50.01	74.90	89.80	94.79	97.27	0.24
	SD	1.92	2.73	3.59	3.52	0.73	3.55	3.60	2.74	1.93	2.86
$R^{(7)}_{BOOT}$	ME	2.78	5.29	10.31	25.19	50.01	74.82	89.72	94.72	97.23	0.31
	SD	1.95	2.76	3.62	3.53	0.73	3.56	3.62	2.78	1.96	2.88
$R^{(8)}_{BOOT}$	ME	2.83	5.36	10.39	25.26	50.01	74.75	89.64	94.65	97.19	0.39
	SD	1.98	2.79	3.65	3.54	0.73	3.57	3.65	2.81	2.00	2.91
$R^{(9)}_{BOOT}$	ME	2.87	5.42	10.47	25.34	50.01	74.67	89.56	94.59	97.14	0.47
	SD	2.01	2.83	3.68	3.56	0.73	3.58	3.68	2.85	2.03	2.93
$R^{(10)}_{BOOT}$	ME	2.92	5.49	10.55	25.41	50.01	74.60	89.48	94.53	97.10	0.55
	SD	2.05	2.87	3.71	3.57	0.73	3.60	3.71	2.88	2.06	2.96
$R^{(11)}_{BOOT}$	ME	2.96	5.55	10.63	25.49	50.01	74.53	89.40	94.47	97.06	0.63
	SD	2.08	2.91	3.74	3.58	0.73	3.61	3.75	2.92	2.10	2.98
$R^{(12)}_{BOOT}$	ME	3.01	5.61	10.71	25.56	50.01	74.46	89.32	94.40	97.01	0.71
	SD	2.12	2.95	3.77	3.60	0.73	3.63	3.78	2.96	2.14	3.01

Tabela B.19: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NAC–0.1 $n = 20$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	3.85	5.78	9.54	21.87	46.12	73.89	90.84	96.24	98.68	3.88
	SD	7.68	8.40	9.10	8.77	4.03	4.30	5.54	4.79	3.76	6.60
$R^{(5)}_{BOOT}$	ME	4.33	6.28	10.05	22.31	46.35	73.84	90.69	96.10	98.58	3.65
	SD	8.76	9.48	10.12	9.53	4.55	4.26	5.61	4.91	3.88	7.22
$R^{(6)}_{BOOT}$	ME	4.72	6.70	10.46	22.67	46.52	73.79	90.56	95.98	98.49	3.48
	SD	9.64	10.33	10.89	10.11	4.93	4.24	5.67	5.02	3.99	7.72
$R^{(7)}_{BOOT}$	ME	5.06	7.05	10.82	22.98	46.67	73.74	90.44	95.86	98.41	3.33
	SD	10.35	11.01	11.50	10.57	5.23	4.23	5.74	5.12	4.09	8.12
$R^{(8)}_{BOOT}$	ME	5.36	7.36	11.13	23.26	46.79	73.69	90.32	95.76	98.33	3.21
	SD	10.95	11.58	12.01	10.96	5.46	4.23	5.80	5.21	4.18	8.46
$R^{(9)}_{BOOT}$	ME	5.63	7.62	11.40	23.50	46.90	73.63	90.21	95.66	98.26	3.13
	SD	11.46	12.04	12.44	11.29	5.65	4.24	5.86	5.30	4.26	8.75
$R^{(10)}_{BOOT}$	ME	5.86	7.88	11.65	23.71	46.99	73.58	90.10	95.56	98.20	3.36
	SD	11.89	12.47	12.81	11.56	5.82	4.26	5.92	5.38	4.34	9.00
$R^{(11)}_{BOOT}$	ME	6.08	8.10	11.88	23.91	47.08	73.53	90.00	95.47	98.13	3.58
	SD	12.28	12.82	13.14	11.81	5.96	4.28	5.97	5.45	4.41	9.23
$R^{(12)}_{BOOT}$	ME	6.28	8.31	12.10	24.10	47.15	73.47	89.90	95.38	98.07	3.78
	SD	12.62	13.15	13.43	12.02	6.09	4.30	6.03	5.53	4.48	9.43

NAC–0.1 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	6.80	8.43	11.20	19.84	40.41	72.49	93.91	97.68	98.99	9.59
	SD	12.16	13.29	14.48	15.18	10.13	4.50	7.53	4.88	2.71	10.41
$R^{(5)}_{BOOT}$	ME	8.18	9.88	12.68	21.21	41.31	72.59	93.64	97.45	98.82	8.69
	SD	14.03	15.11	16.23	16.64	11.12	4.43	7.63	5.13	2.97	11.56
$R^{(6)}_{BOOT}$	ME	9.33	11.05	13.86	22.28	41.98	72.65	93.41	97.26	98.67	8.02
	SD	15.48	16.50	17.53	17.68	11.80	4.40	7.72	5.33	3.21	12.42
$R^{(7)}_{BOOT}$	ME	10.29	12.02	14.83	23.15	42.53	72.67	93.21	97.09	98.54	7.79
	SD	16.62	17.60	18.54	18.50	12.33	4.38	7.80	5.50	3.40	13.11
$R^{(8)}_{BOOT}$	ME	11.12	12.85	15.65	23.89	42.98	72.69	93.02	96.94	98.41	8.62
	SD	17.56	18.49	19.37	19.16	12.74	4.38	7.87	5.65	3.58	13.67
$R^{(9)}_{BOOT}$	ME	11.84	13.59	16.35	24.52	43.36	72.69	92.85	96.80	98.29	9.34
	SD	18.36	19.25	20.03	19.70	13.07	4.39	7.94	5.78	3.73	14.13
$R^{(10)}_{BOOT}$	ME	12.49	14.22	16.97	25.08	43.69	72.68	92.69	96.67	98.18	9.99
	SD	19.03	19.89	20.61	20.15	13.35	4.40	8.00	5.90	3.88	14.53
$R^{(11)}_{BOOT}$	ME	13.07	14.78	17.54	25.57	43.99	72.67	92.54	96.55	98.09	10.57
	SD	19.63	20.43	21.11	20.54	13.59	4.41	8.05	6.01	4.02	14.88
$R^{(12)}_{BOOT}$	ME	13.58	15.29	18.04	26.01	44.25	72.65	92.40	96.44	97.99	11.08
	SD	20.14	20.91	21.54	20.87	13.80	4.43	8.11	6.10	4.14	15.18

Tabela B.20: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NAC–0.1 $n = 50$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.73	5.09	9.71	24.10	49.26	75.28	90.63	95.60	98.01	0.90
	SD	2.80	3.45	4.07	4.00	1.73	2.51	2.70	2.07	1.45	2.89
$R^{(5)}_{BOOT}$	ME	2.90	5.28	9.93	24.29	49.35	75.25	90.56	95.54	97.96	0.71
	SD	3.46	4.11	4.70	4.51	2.05	2.52	2.77	2.17	1.54	3.28
$R^{(6)}_{BOOT}$	ME	3.05	5.45	10.11	24.46	49.42	75.22	90.49	95.47	97.90	0.58
	SD	4.04	4.66	5.21	4.91	2.28	2.54	2.83	2.26	1.64	3.60
$R^{(7)}_{BOOT}$	ME	3.19	5.61	10.28	24.61	49.49	75.20	90.43	95.41	97.85	0.69
	SD	4.52	5.13	5.65	5.23	2.48	2.56	2.89	2.34	1.73	3.88
$R^{(8)}_{BOOT}$	ME	3.32	5.75	10.43	24.74	49.54	75.17	90.37	95.36	97.80	0.82
	SD	4.95	5.53	6.02	5.53	2.64	2.58	2.95	2.42	1.82	4.13
$R^{(9)}_{BOOT}$	ME	3.45	5.89	10.58	24.87	49.60	75.13	90.31	95.30	97.76	0.95
	SD	5.31	5.90	6.35	5.78	2.78	2.61	3.01	2.50	1.90	4.35
$R^{(10)}_{BOOT}$	ME	3.56	6.02	10.71	24.98	49.64	75.10	90.25	95.24	97.72	1.06
	SD	5.64	6.21	6.64	6.00	2.91	2.63	3.07	2.57	1.97	4.54
$R^{(11)}_{BOOT}$	ME	3.66	6.14	10.83	25.09	49.69	75.08	90.20	95.19	97.67	1.16
	SD	5.93	6.50	6.89	6.20	3.01	2.65	3.13	2.65	2.04	4.72
$R^{(12)}_{BOOT}$	ME	3.77	6.25	10.95	25.20	49.72	75.05	90.14	95.14	97.63	1.27
	SD	6.21	6.76	7.14	6.38	3.10	2.67	3.18	2.71	2.11	4.88
NAC–0.1 $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	6.34	8.44	12.12	23.42	46.75	75.96	93.79	97.89	99.10	3.84
	SD	11.17	12.27	13.29	13.34	8.04	2.91	4.78	3.60	2.42	9.10
$R^{(5)}_{BOOT}$	ME	7.91	10.13	13.89	25.11	47.83	76.15	93.58	97.70	98.98	5.41
	SD	13.41	14.45	15.36	15.09	9.22	2.88	4.73	3.67	2.53	10.48
$R^{(6)}_{BOOT}$	ME	9.24	11.53	15.35	26.46	48.65	76.26	93.38	97.53	98.88	6.74
	SD	15.10	16.08	16.92	16.35	10.03	2.91	4.72	3.73	2.61	11.51
$R^{(7)}_{BOOT}$	ME	10.39	12.72	16.58	27.59	49.30	76.33	93.19	97.38	98.78	7.89
	SD	16.45	17.38	18.11	17.33	10.64	2.99	4.73	3.79	2.69	12.33
$R^{(8)}_{BOOT}$	ME	11.38	13.77	17.63	28.52	49.86	76.37	93.01	97.23	98.68	8.88
	SD	17.55	18.43	19.07	18.09	11.13	3.07	4.76	3.86	2.77	12.98
$R^{(9)}_{BOOT}$	ME	12.28	14.69	18.56	29.34	50.34	76.40	92.85	97.10	98.59	9.78
	SD	18.48	19.30	19.87	18.71	11.52	3.15	4.78	3.92	2.84	13.53
$R^{(10)}_{BOOT}$	ME	13.09	15.51	19.38	30.05	50.76	76.42	92.69	96.96	98.51	10.59
	SD	19.27	20.03	20.54	19.22	11.85	3.23	4.82	3.97	2.90	13.99
$R^{(11)}_{BOOT}$	ME	13.82	16.25	20.13	30.70	51.12	76.43	92.55	96.84	98.43	11.32
	SD	19.95	20.66	21.11	19.66	12.11	3.31	4.85	4.02	2.95	14.39
$R^{(12)}_{BOOT}$	ME	14.49	16.94	20.80	31.28	51.45	76.43	92.41	96.73	98.35	11.99
	SD	20.53	21.22	21.61	20.04	12.35	3.39	4.89	4.07	3.01	14.73

Tabela B.21: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NAC–0.1 $n = 100$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.50	4.91	9.69	24.52	49.76	75.35	90.48	95.33	97.73	0.48
	SD	1.16	1.66	2.17	2.20	0.78	1.92	1.96	1.48	0.99	1.66
$R^{(5)}_{BOOT}$	ME	2.55	4.98	9.77	24.60	49.79	75.34	90.45	95.29	97.70	0.45
	SD	1.29	1.81	2.34	2.36	0.86	1.93	2.00	1.53	1.03	1.76
$R^{(6)}_{BOOT}$	ME	2.59	5.03	9.84	24.67	49.82	75.32	90.41	95.26	97.68	0.41
	SD	1.44	1.97	2.50	2.52	0.94	1.94	2.04	1.58	1.08	1.85
$R^{(7)}_{BOOT}$	ME	2.64	5.09	9.91	24.73	49.85	75.30	90.38	95.23	97.66	0.38
	SD	1.58	2.13	2.67	2.66	1.02	1.95	2.08	1.63	1.13	1.95
$R^{(8)}_{BOOT}$	ME	2.69	5.15	9.97	24.80	49.88	75.28	90.35	95.20	97.63	0.35
	SD	1.73	2.29	2.83	2.80	1.10	1.96	2.13	1.68	1.18	2.05
$R^{(9)}_{BOOT}$	ME	2.73	5.21	10.04	24.86	49.90	75.27	90.32	95.17	97.61	0.32
	SD	1.87	2.45	2.98	2.94	1.18	1.97	2.16	1.73	1.22	2.15
$R^{(10)}_{BOOT}$	ME	2.78	5.26	10.10	24.92	49.92	75.25	90.29	95.14	97.59	0.29
	SD	2.01	2.60	3.14	3.07	1.24	1.98	2.20	1.77	1.27	2.24
$R^{(11)}_{BOOT}$	ME	2.82	5.32	10.17	24.98	49.94	75.23	90.26	95.11	97.56	0.32
	SD	2.15	2.75	3.29	3.19	1.30	1.99	2.24	1.82	1.31	2.33
$R^{(12)}_{BOOT}$	ME	2.87	5.37	10.22	25.03	49.97	75.21	90.23	95.09	97.54	0.37
	SD	2.30	2.89	3.43	3.31	1.36	2.01	2.28	1.86	1.36	2.42

NAC–0.1 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	5.55	7.93	12.20	25.13	49.72	77.89	94.18	98.12	99.37	4.18
	SD	9.37	10.62	11.84	12.11	7.57	2.14	2.64	1.84	1.06	7.89
$R^{(5)}_{BOOT}$	ME	7.25	9.81	14.23	27.11	51.00	78.16	94.01	97.95	99.27	4.81
	SD	11.90	13.12	14.24	14.13	8.92	2.42	2.62	1.90	1.14	9.51
$R^{(6)}_{BOOT}$	ME	8.75	11.42	15.92	28.72	52.00	78.34	93.85	97.81	99.17	6.42
	SD	13.87	15.04	16.03	15.60	9.86	2.69	2.63	1.96	1.21	10.73
$R^{(7)}_{BOOT}$	ME	10.05	12.83	17.37	30.06	52.80	78.47	93.71	97.67	99.07	7.83
	SD	15.43	16.53	17.43	16.72	10.57	2.91	2.66	2.03	1.28	11.68
$R^{(8)}_{BOOT}$	ME	11.23	14.06	18.64	31.20	53.49	78.56	93.56	97.53	98.99	9.06
	SD	16.72	17.76	18.57	17.59	11.12	3.12	2.69	2.10	1.34	12.45
$R^{(9)}_{BOOT}$	ME	12.29	15.16	19.76	32.20	54.07	78.63	93.42	97.41	98.90	10.16
	SD	17.81	18.79	19.51	18.31	11.56	3.29	2.73	2.16	1.41	13.10
$R^{(10)}_{BOOT}$	ME	13.25	16.16	20.76	33.08	54.57	78.67	93.29	97.29	98.82	11.16
	SD	18.75	19.68	20.30	18.91	11.92	3.44	2.77	2.22	1.47	13.65
$R^{(11)}_{BOOT}$	ME	14.12	17.07	21.67	33.87	55.02	78.71	93.16	97.17	98.74	12.07
	SD	19.56	20.45	20.98	19.40	12.23	3.57	2.81	2.28	1.52	14.12
$R^{(12)}_{BOOT}$	ME	14.94	17.91	22.50	34.59	55.43	78.73	93.04	97.06	98.67	12.91
	SD	20.27	21.11	21.57	19.82	12.49	3.69	2.85	2.34	1.58	14.53

Tabela B.22: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN1 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

LOGN1 $n = 20$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	1.97	3.10	5.44	15.09	44.38	80.80	95.11	98.06	99.15	9.91
	SD	3.40	4.42	5.95	8.61	3.26	7.43	4.83	2.84	1.66	5.17
$R^{(5)}_{BOOT}$	ME	2.29	3.52	6.02	15.90	44.62	80.15	94.64	97.78	98.99	9.10
	SD	3.77	4.82	6.33	8.82	3.22	7.51	5.14	3.14	1.90	5.40
$R^{(6)}_{BOOT}$	ME	2.61	3.95	6.58	16.64	44.83	79.56	94.20	97.50	98.83	8.36
	SD	4.13	5.21	6.69	9.01	3.21	7.59	5.45	3.43	2.14	5.62
$R^{(7)}_{BOOT}$	ME	2.93	4.36	7.11	17.31	45.01	79.04	93.77	97.23	98.66	7.69
	SD	4.49	5.58	7.02	9.19	3.20	7.66	5.72	3.71	2.37	5.83
$R^{(8)}_{BOOT}$	ME	3.26	4.77	7.62	17.90	45.17	78.56	93.37	96.96	98.48	7.10
	SD	4.83	5.91	7.34	9.35	3.19	7.72	5.98	3.97	2.61	6.03
$R^{(9)}_{BOOT}$	ME	3.58	5.16	8.10	18.46	45.32	78.12	92.99	96.70	98.31	6.54
	SD	5.16	6.24	7.63	9.49	3.18	7.78	6.22	4.22	2.83	6.23
$R^{(10)}_{BOOT}$	ME	3.90	5.54	8.56	18.98	45.44	77.71	92.62	96.45	98.14	6.02
	SD	5.46	6.53	7.90	9.62	3.18	7.84	6.45	4.45	3.05	6.41
$R^{(11)}_{BOOT}$	ME	4.20	5.91	9.00	19.46	45.57	77.33	92.27	96.19	97.97	5.54
	SD	5.74	6.81	8.15	9.74	3.17	7.89	6.66	4.68	3.26	6.58
$R^{(12)}_{BOOT}$	ME	4.51	6.26	9.41	19.91	45.68	76.97	91.94	95.95	97.81	5.09
	SD	6.02	7.08	8.38	9.85	3.17	7.93	6.86	4.89	3.45	6.74

LOGN1 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.16	3.46	6.07	16.33	45.15	79.76	94.49	97.74	99.01	8.67
	SD	3.44	4.49	6.01	8.51	2.99	7.06	4.84	2.94	1.76	5.10
$R^{(5)}_{BOOT}$	ME	2.45	3.83	6.56	16.96	45.31	79.28	94.11	97.49	98.85	8.04
	SD	3.81	4.88	6.37	8.69	3.01	7.14	5.11	3.21	1.99	5.32
$R^{(6)}_{BOOT}$	ME	2.74	4.20	7.02	17.53	45.46	78.86	93.74	97.24	98.70	7.47
	SD	4.18	5.24	6.71	8.86	3.03	7.22	5.37	3.49	2.22	5.53
$R^{(7)}_{BOOT}$	ME	3.03	4.55	7.47	18.06	45.58	78.46	93.40	97.01	98.55	6.94
	SD	4.51	5.57	7.03	9.03	3.05	7.31	5.61	3.75	2.44	5.74
$R^{(8)}_{BOOT}$	ME	3.31	4.90	7.88	18.54	45.69	78.11	93.07	96.78	98.40	6.47
	SD	4.83	5.89	7.31	9.19	3.07	7.38	5.85	3.99	2.65	5.93
$R^{(9)}_{BOOT}$	ME	3.59	5.22	8.28	18.97	45.80	77.78	92.76	96.55	98.24	6.03
	SD	5.15	6.18	7.58	9.33	3.09	7.46	6.07	4.21	2.86	6.12
$R^{(10)}_{BOOT}$	ME	3.86	5.55	8.65	19.39	45.89	77.47	92.46	96.33	98.09	5.61
	SD	5.42	6.46	7.83	9.47	3.11	7.54	6.28	4.42	3.07	6.30
$R^{(11)}_{BOOT}$	ME	4.13	5.86	9.01	19.78	45.97	77.18	92.18	96.12	97.94	5.22
	SD	5.69	6.73	8.07	9.60	3.13	7.61	6.47	4.62	3.25	6.47
$R^{(12)}_{BOOT}$	ME	4.38	6.17	9.37	20.14	46.05	76.91	91.90	95.92	97.80	4.86
	SD	5.93	6.98	8.31	9.72	3.15	7.68	6.66	4.81	3.43	6.63

Tabela B.23: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN1 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

LOGN1 $n = 50$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	1.14	2.30	5.00	16.76	46.83	80.63	95.27	98.30	99.37	8.24
	SD	1.50	2.36	3.70	5.62	1.25	4.94	3.23	1.74	0.89	3.22
$R^{(5)}_{BOOT}$	ME	1.38	2.68	5.57	17.58	46.97	79.93	94.78	98.02	99.23	7.42
	SD	1.75	2.66	4.01	5.77	1.25	5.03	3.49	1.97	1.05	3.39
$R^{(6)}_{BOOT}$	ME	1.65	3.06	6.13	18.31	47.10	79.29	94.29	97.74	99.08	6.69
	SD	2.03	2.97	4.32	5.91	1.25	5.11	3.75	2.22	1.23	3.57
$R^{(7)}_{BOOT}$	ME	1.91	3.45	6.68	18.99	47.21	78.71	93.83	97.46	98.92	6.01
	SD	2.29	3.27	4.61	6.03	1.26	5.19	4.00	2.45	1.41	3.74
$R^{(8)}_{BOOT}$	ME	2.18	3.83	7.20	19.61	47.31	78.17	93.38	97.18	98.75	5.39
	SD	2.55	3.54	4.88	6.15	1.26	5.28	4.23	2.67	1.60	3.91
$R^{(9)}_{BOOT}$	ME	2.46	4.21	7.70	20.20	47.40	77.67	92.94	96.89	98.58	4.80
	SD	2.80	3.81	5.14	6.25	1.26	5.34	4.44	2.89	1.78	4.07
$R^{(10)}_{BOOT}$	ME	2.74	4.59	8.18	20.74	47.48	77.21	92.52	96.62	98.40	4.26
	SD	3.05	4.07	5.37	6.34	1.26	5.40	4.65	3.10	1.96	4.23
$R^{(11)}_{BOOT}$	ME	3.02	4.96	8.65	21.26	47.56	76.78	92.12	96.34	98.22	3.74
	SD	3.28	4.32	5.58	6.42	1.26	5.47	4.84	3.31	2.14	4.37
$R^{(12)}_{BOOT}$	ME	3.30	5.31	9.11	21.73	47.63	76.37	91.73	96.06	98.05	3.27
	SD	3.51	4.55	5.78	6.50	1.27	5.52	5.01	3.50	2.30	4.51
LOGN1 $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	1.33	2.60	5.57	17.94	47.33	79.60	94.57	97.93	99.19	7.06
	SD	1.75	2.61	3.90	5.40	1.16	4.63	3.39	1.99	1.09	3.23
$R^{(5)}_{BOOT}$	ME	1.53	2.91	6.05	18.56	47.43	79.09	94.16	97.69	99.06	6.44
	SD	1.97	2.86	4.15	5.50	1.17	4.68	3.60	2.20	1.26	3.37
$R^{(6)}_{BOOT}$	ME	1.75	3.23	6.51	19.14	47.52	78.61	93.77	97.46	98.93	5.86
	SD	2.20	3.12	4.39	5.62	1.18	4.75	3.81	2.40	1.42	3.52
$R^{(7)}_{BOOT}$	ME	1.96	3.55	6.96	19.68	47.60	78.17	93.39	97.22	98.79	5.32
	SD	2.42	3.37	4.62	5.73	1.19	4.82	4.00	2.60	1.58	3.67
$R^{(8)}_{BOOT}$	ME	2.19	3.86	7.39	20.17	47.67	77.76	93.03	96.98	98.65	4.83
	SD	2.66	3.61	4.85	5.83	1.19	4.89	4.19	2.79	1.74	3.82
$R^{(9)}_{BOOT}$	ME	2.42	4.17	7.81	20.64	47.75	77.37	92.68	96.74	98.51	4.36
	SD	2.87	3.84	5.06	5.93	1.20	4.95	4.38	2.98	1.91	3.96
$R^{(10)}_{BOOT}$	ME	2.64	4.48	8.22	21.09	47.81	77.01	92.34	96.51	98.36	3.91
	SD	3.08	4.07	5.27	6.03	1.20	5.02	4.54	3.16	2.06	4.10
$R^{(11)}_{BOOT}$	ME	2.88	4.79	8.61	21.51	47.87	76.66	92.01	96.28	98.21	3.49
	SD	3.29	4.28	5.46	6.12	1.21	5.08	4.70	3.33	2.22	4.24
$R^{(12)}_{BOOT}$	ME	3.10	5.09	8.98	21.90	47.92	76.34	91.69	96.05	98.06	3.10
	SD	3.49	4.48	5.65	6.20	1.22	5.15	4.87	3.50	2.38	4.37

Tabela B.24: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN1 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

LOGN1 $n = 100$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	0.84	1.93	4.71	17.35	47.86	80.76	95.44	98.45	99.46	7.65
	SD	0.98	1.65	2.71	4.00	0.85	3.59	2.47	1.33	0.65	2.33
$R^{(5)}_{BOOT}$	ME	1.05	2.28	5.29	18.17	47.96	80.03	94.91	98.17	99.33	6.83
	SD	1.16	1.88	2.96	4.11	0.85	3.68	2.70	1.51	0.79	2.47
$R^{(6)}_{BOOT}$	ME	1.27	2.65	5.87	18.93	48.05	79.35	94.40	97.89	99.18	6.07
	SD	1.35	2.11	3.19	4.23	0.85	3.77	2.91	1.70	0.93	2.61
$R^{(7)}_{BOOT}$	ME	1.51	3.02	6.43	19.63	48.12	78.73	93.90	97.60	99.02	5.37
	SD	1.54	2.34	3.42	4.34	0.85	3.86	3.11	1.89	1.07	2.76
$R^{(8)}_{BOOT}$	ME	1.75	3.39	6.97	20.27	48.19	78.15	93.42	97.30	98.86	4.73
	SD	1.74	2.56	3.64	4.44	0.85	3.94	3.31	2.08	1.22	2.90
$R^{(9)}_{BOOT}$	ME	2.00	3.76	7.48	20.87	48.26	77.61	92.96	97.01	98.69	4.13
	SD	1.93	2.77	3.83	4.53	0.84	4.00	3.51	2.27	1.37	3.03
$R^{(10)}_{BOOT}$	ME	2.26	4.14	7.99	21.45	48.32	77.11	92.50	96.71	98.50	3.55
	SD	2.12	2.96	4.03	4.61	0.84	4.06	3.68	2.44	1.52	3.16
$R^{(11)}_{BOOT}$	ME	2.52	4.50	8.48	21.98	48.38	76.64	92.06	96.41	98.32	3.02
	SD	2.30	3.15	4.21	4.68	0.84	4.12	3.84	2.61	1.66	3.28
$R^{(12)}_{BOOT}$	ME	2.79	4.87	8.96	22.48	48.43	76.20	91.63	96.12	98.13	2.52
	SD	2.48	3.34	4.37	4.75	0.83	4.17	3.99	2.78	1.80	3.40
LOGN1 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	1.04	2.30	5.43	18.51	48.13	79.74	94.70	98.09	99.29	6.49
	SD	1.01	1.70	2.81	3.95	0.82	3.45	2.48	1.37	0.69	2.32
$R^{(5)}_{BOOT}$	ME	1.21	2.59	5.88	19.10	48.20	79.22	94.29	97.87	99.17	5.90
	SD	1.15	1.89	2.98	4.02	0.82	3.50	2.65	1.52	0.80	2.42
$R^{(6)}_{BOOT}$	ME	1.39	2.88	6.32	19.66	48.26	78.73	93.91	97.64	99.04	5.34
	SD	1.30	2.08	3.17	4.10	0.82	3.57	2.80	1.67	0.92	2.53
$R^{(7)}_{BOOT}$	ME	1.58	3.17	6.76	20.18	48.32	78.28	93.55	97.40	98.91	4.82
	SD	1.47	2.27	3.35	4.18	0.82	3.63	2.96	1.83	1.03	2.65
$R^{(8)}_{BOOT}$	ME	1.77	3.46	7.18	20.67	48.37	77.85	93.18	97.17	98.78	4.33
	SD	1.63	2.45	3.53	4.26	0.82	3.69	3.11	1.98	1.16	2.76
$R^{(9)}_{BOOT}$	ME	1.97	3.76	7.59	21.13	48.42	77.45	92.82	96.94	98.63	3.87
	SD	1.80	2.64	3.70	4.34	0.82	3.75	3.26	2.13	1.29	2.87
$R^{(10)}_{BOOT}$	ME	2.18	4.05	7.98	21.57	48.47	77.06	92.47	96.71	98.49	3.43
	SD	1.96	2.81	3.86	4.41	0.82	3.81	3.41	2.28	1.42	2.99
$R^{(11)}_{BOOT}$	ME	2.39	4.34	8.38	21.98	48.52	76.70	92.13	96.48	98.35	3.02
	SD	2.13	2.98	4.02	4.49	0.82	3.87	3.54	2.43	1.54	3.09
$R^{(12)}_{BOOT}$	ME	2.60	4.62	8.75	22.38	48.56	76.34	91.79	96.24	98.21	2.62
	SD	2.29	3.14	4.17	4.55	0.82	3.92	3.67	2.57	1.66	3.20

Tabela B.25: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN2 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

LOGN2 $n = 20$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	3.69	4.52	6.05	12.00	38.06	89.71	98.65	99.62	99.88	14.71
	SD	7.79	8.68	10.05	13.60	15.46	7.44	2.69	1.21	0.56	8.98
$R^{(2)}_{BOOT}$	ME	1.90	2.37	3.26	6.97	26.90	84.55	97.72	99.31	99.77	23.10
	SD	4.54	5.16	6.16	8.98	11.80	10.03	4.23	1.98	0.94	6.90
$R^{(3)}_{BOOT}$	ME	2.31	2.85	3.88	7.99	28.21	83.41	97.28	99.13	99.69	21.79
	SD	5.12	5.76	6.79	9.52	11.69	10.18	4.69	2.30	1.14	7.19
$R^{(4)}_{BOOT}$	ME	2.73	3.35	4.49	8.91	29.24	82.46	96.85	98.94	99.61	20.76
	SD	5.65	6.32	7.34	10.02	11.62	10.31	5.12	2.63	1.35	7.47
$R^{(5)}_{BOOT}$	ME	3.15	3.83	5.07	9.76	30.10	81.64	96.44	98.75	99.51	19.90
	SD	6.14	6.82	7.85	10.45	11.56	10.42	5.50	2.95	1.56	7.74
$R^{(6)}_{BOOT}$	ME	3.57	4.30	5.63	10.54	30.84	80.93	96.05	98.55	99.42	19.16
	SD	6.59	7.27	8.30	10.84	11.49	10.51	5.85	3.25	1.77	7.98
$R^{(7)}_{BOOT}$	ME	3.97	4.76	6.16	11.24	31.49	80.29	95.67	98.36	99.32	18.51
	SD	7.01	7.69	8.70	11.16	11.43	10.59	6.18	3.53	1.98	8.20
$R^{(8)}_{BOOT}$	ME	4.37	5.20	6.67	11.90	32.06	79.73	95.32	98.16	99.21	17.94
	SD	7.39	8.06	9.06	11.44	11.38	10.65	6.48	3.80	2.19	8.41
$R^{(9)}_{BOOT}$	ME	4.75	5.61	7.17	12.52	32.57	79.21	94.98	97.98	99.11	17.43
	SD	7.74	8.40	9.40	11.70	11.32	10.70	6.75	4.04	2.39	8.60
$R^{(10)}_{BOOT}$	ME	5.11	6.02	7.64	13.09	33.03	78.73	94.64	97.79	99.01	16.97
	SD	8.06	8.72	9.71	11.94	11.27	10.75	7.00	4.28	2.58	8.78
$R^{(11)}_{BOOT}$	ME	5.47	6.42	8.08	13.62	33.46	78.28	94.32	97.60	98.90	16.54
	SD	8.35	9.02	9.99	12.14	11.21	10.79	7.24	4.51	2.75	8.95
$R^{(12)}_{BOOT}$	ME	5.82	6.81	8.51	14.12	33.84	77.87	94.01	97.42	98.80	16.16
	SD	8.64	9.30	10.25	12.33	11.16	10.82	7.45	4.71	2.93	9.11

Tabela B.26: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN2 e $n = 20$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

LOGN2 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	3.36	4.18	5.63	11.22	36.10	89.57	98.57	99.61	99.88	14.57
	SD	7.13	8.11	9.62	13.55	16.45	7.93	2.84	1.24	0.57	9.05
$R^{(2)}_{BOOT}$	ME	1.72	2.19	3.09	6.77	26.23	85.19	97.75	99.35	99.80	23.77
	SD	4.18	4.82	5.87	8.88	12.80	10.17	4.05	1.78	0.79	7.00
$R^{(3)}_{BOOT}$	ME	2.11	2.66	3.66	7.64	27.39	84.23	97.35	99.18	99.73	22.61
	SD	4.71	5.41	6.49	9.50	12.70	10.33	4.50	2.07	0.94	7.28
$R^{(4)}_{BOOT}$	ME	2.51	3.10	4.18	8.41	28.32	83.44	96.98	99.00	99.64	21.68
	SD	5.26	5.98	7.07	10.03	12.61	10.47	4.92	2.38	1.14	7.56
$R^{(5)}_{BOOT}$	ME	2.88	3.52	4.68	9.11	29.11	82.75	96.63	98.83	99.56	20.89
	SD	5.78	6.50	7.60	10.48	12.54	10.59	5.30	2.69	1.35	7.82
$R^{(6)}_{BOOT}$	ME	3.22	3.91	5.14	9.75	29.79	82.16	96.31	98.66	99.47	20.21
	SD	6.24	6.96	8.07	10.87	12.47	10.69	5.64	2.99	1.56	8.05
$R^{(7)}_{BOOT}$	ME	3.56	4.29	5.58	10.34	30.38	81.62	95.99	98.49	99.38	19.62
	SD	6.66	7.38	8.48	11.21	12.42	10.77	5.95	3.26	1.77	8.27
$R^{(8)}_{BOOT}$	ME	3.88	4.65	6.00	10.89	30.91	81.14	95.70	98.32	99.29	19.09
	SD	7.05	7.77	8.85	11.51	12.37	10.85	6.23	3.52	1.98	8.48
$R^{(9)}_{BOOT}$	ME	4.20	5.00	6.40	11.40	31.39	80.68	95.41	98.16	99.20	18.61
	SD	7.41	8.12	9.19	11.78	12.33	10.92	6.49	3.77	2.17	8.67
$R^{(10)}_{BOOT}$	ME	4.50	5.33	6.77	11.88	31.82	80.26	95.14	98.01	99.11	18.18
	SD	7.73	8.44	9.49	12.02	12.28	10.98	6.73	4.00	2.36	8.84
$R^{(11)}_{BOOT}$	ME	4.80	5.66	7.14	12.34	32.22	79.87	94.88	97.85	99.02	17.78
	SD	8.04	8.73	9.78	12.24	12.25	11.03	6.94	4.22	2.54	9.01
$R^{(12)}_{BOOT}$	ME	5.08	5.97	7.50	12.76	32.58	79.51	94.63	97.70	98.93	17.42
	SD	8.31	9.01	10.05	12.43	12.20	11.07	7.15	4.42	2.70	9.16

Tabela B.27: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN2 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

LOGN2 $n = 50$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	1.89	2.96	5.27	15.97	55.65	94.10	99.52	99.91	99.98	19.10
	SD	4.23	5.47	7.53	12.56	8.54	4.00	1.04	0.38	0.16	6.26
$R^{(2)}_{BOOT}$	ME	0.52	0.87	1.72	6.70	35.99	87.20	98.66	99.71	99.93	18.30
	SD	1.57	2.10	3.10	6.37	5.54	7.09	2.32	0.94	0.40	4.00
$R^{(3)}_{BOOT}$	ME	0.72	1.17	2.22	7.82	36.87	85.98	98.29	99.60	99.89	17.18
	SD	1.88	2.50	3.60	6.88	5.35	7.31	2.69	1.14	0.52	4.23
$R^{(4)}_{BOOT}$	ME	0.95	1.50	2.73	8.85	37.61	84.89	97.90	99.46	99.84	16.15
	SD	2.23	2.91	4.09	7.31	5.21	7.50	3.06	1.36	0.64	4.46
$R^{(5)}_{BOOT}$	ME	1.20	1.84	3.24	9.81	38.23	83.93	97.51	99.31	99.79	15.19
	SD	2.58	3.30	4.55	7.69	5.08	7.65	3.41	1.59	0.77	4.67
$R^{(6)}_{BOOT}$	ME	1.46	2.20	3.75	10.70	38.76	83.08	97.12	99.15	99.72	14.30
	SD	2.92	3.69	4.98	8.00	4.98	7.77	3.75	1.83	0.91	4.88
$R^{(7)}_{BOOT}$	ME	1.74	2.56	4.26	11.53	39.23	82.29	96.74	98.99	99.65	13.47
	SD	3.25	4.07	5.36	8.28	4.90	7.87	4.05	2.05	1.06	5.08
$R^{(8)}_{BOOT}$	ME	2.02	2.93	4.76	12.31	39.63	81.57	96.36	98.82	99.58	12.69
	SD	3.57	4.42	5.71	8.50	4.82	7.94	4.35	2.27	1.20	5.26
$R^{(9)}_{BOOT}$	ME	2.30	3.29	5.25	13.05	40.00	80.90	95.98	98.64	99.49	11.95
	SD	3.87	4.75	6.03	8.71	4.74	8.00	4.62	2.49	1.35	5.43
$R^{(10)}_{BOOT}$	ME	2.59	3.65	5.72	13.74	40.35	80.28	95.62	98.46	99.40	11.26
	SD	4.18	5.05	6.33	8.88	4.67	8.05	4.88	2.69	1.50	5.60
$R^{(11)}_{BOOT}$	ME	2.88	4.01	6.19	14.39	40.65	79.72	95.26	98.28	99.31	10.61
	SD	4.47	5.33	6.60	9.03	4.61	8.08	5.12	2.90	1.65	5.75
$R^{(12)}_{BOOT}$	ME	3.17	4.36	6.64	15.00	40.93	79.18	94.91	98.10	99.22	10.00
	SD	4.73	5.59	6.85	9.17	4.55	8.11	5.33	3.08	1.80	5.89

Tabela B.28: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN2 e $n = 50$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

LOGN2 $n = 50$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	2.01	3.11	5.59	16.84	55.22	93.07	99.36	99.87	99.97	18.07
	SD	4.34	5.54	7.61	12.34	8.92	4.36	1.17	0.41	0.14	6.32
$R^{(2)}_{BOOT}$	ME	0.61	1.00	1.98	7.54	37.13	86.20	98.40	99.64	99.90	17.46
	SD	1.67	2.27	3.36	6.61	5.44	6.98	2.47	0.96	0.36	4.06
$R^{(3)}_{BOOT}$	ME	0.80	1.29	2.45	8.60	37.87	85.11	98.04	99.51	99.86	16.40
	SD	1.99	2.64	3.79	7.01	5.30	7.16	2.80	1.18	0.48	4.26
$R^{(4)}_{BOOT}$	ME	1.02	1.61	2.94	9.58	38.48	84.14	97.67	99.38	99.82	15.42
	SD	2.32	3.02	4.22	7.38	5.19	7.32	3.12	1.39	0.61	4.46
$R^{(5)}_{BOOT}$	ME	1.26	1.94	3.44	10.49	39.00	83.27	97.29	99.22	99.76	14.51
	SD	2.64	3.38	4.62	7.72	5.11	7.45	3.45	1.63	0.75	4.67
$R^{(6)}_{BOOT}$	ME	1.52	2.28	3.94	11.34	39.45	82.50	96.91	99.06	99.69	13.66
	SD	2.95	3.73	5.00	8.00	5.05	7.57	3.76	1.85	0.90	4.86
$R^{(7)}_{BOOT}$	ME	1.78	2.63	4.43	12.11	39.85	81.79	96.54	98.89	99.62	12.89
	SD	3.27	4.06	5.36	8.27	5.00	7.67	4.04	2.07	1.05	5.05
$R^{(8)}_{BOOT}$	ME	2.06	2.98	4.91	12.84	40.20	81.15	96.17	98.72	99.54	12.16
	SD	3.57	4.38	5.70	8.49	4.94	7.75	4.32	2.27	1.20	5.23
$R^{(9)}_{BOOT}$	ME	2.33	3.33	5.38	13.52	40.51	80.55	95.81	98.55	99.46	11.48
	SD	3.85	4.69	6.01	8.70	4.91	7.83	4.58	2.47	1.33	5.40
$R^{(10)}_{BOOT}$	ME	2.61	3.68	5.84	14.16	40.79	80.00	95.47	98.38	99.37	10.84
	SD	4.13	4.98	6.30	8.87	4.86	7.88	4.81	2.67	1.49	5.56
$R^{(11)}_{BOOT}$	ME	2.89	4.03	6.29	14.76	41.05	79.49	95.13	98.20	99.28	10.24
	SD	4.39	5.25	6.57	9.02	4.83	7.94	5.04	2.86	1.64	5.72
$R^{(12)}_{BOOT}$	ME	3.16	4.37	6.72	15.33	41.30	79.00	94.79	98.02	99.19	9.67
	SD	4.65	5.52	6.82	9.15	4.79	7.98	5.25	3.05	1.77	5.86

Tabela B.29: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN2 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

LOGN2 $n = 100$ $c = 1.5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	1.60	2.94	6.26	22.56	68.03	96.59	99.80	99.97	100.00	21.59
	SD	3.52	5.03	7.59	12.67	6.08	2.24	0.36	0.09	0.02	5.75
$R^{(2)}_{BOOT}$	ME	0.22	0.48	1.26	7.01	40.58	88.35	98.95	99.79	99.96	17.99
	SD	0.63	1.08	2.05	5.23	3.02	5.63	1.60	0.53	0.17	2.92
$R^{(3)}_{BOOT}$	ME	0.34	0.69	1.68	8.20	41.22	87.06	98.60	99.69	99.93	16.80
	SD	0.89	1.41	2.48	5.64	2.88	5.83	1.95	0.72	0.26	3.11
$R^{(4)}_{BOOT}$	ME	0.48	0.94	2.13	9.33	41.74	85.92	98.23	99.57	99.89	15.67
	SD	1.14	1.75	2.89	5.97	2.77	5.99	2.29	0.92	0.37	3.30
$R^{(5)}_{BOOT}$	ME	0.65	1.20	2.61	10.36	42.18	84.87	97.86	99.44	99.84	14.64
	SD	1.40	2.08	3.28	6.26	2.68	6.12	2.61	1.11	0.48	3.48
$R^{(6)}_{BOOT}$	ME	0.83	1.49	3.09	11.33	42.56	83.93	97.46	99.29	99.78	13.67
	SD	1.67	2.39	3.64	6.49	2.61	6.22	2.91	1.32	0.61	3.65
$R^{(7)}_{BOOT}$	ME	1.03	1.79	3.57	12.23	42.89	83.07	97.08	99.13	99.72	12.77
	SD	1.93	2.70	3.97	6.68	2.54	6.31	3.19	1.52	0.73	3.81
$R^{(8)}_{BOOT}$	ME	1.24	2.09	4.06	13.07	43.19	82.28	96.68	98.96	99.65	11.93
	SD	2.18	2.98	4.29	6.85	2.49	6.37	3.45	1.73	0.86	3.96
$R^{(9)}_{BOOT}$	ME	1.46	2.41	4.55	13.87	43.45	81.54	96.29	98.79	99.58	11.13
	SD	2.43	3.25	4.58	6.98	2.44	6.41	3.69	1.92	0.98	4.10
$R^{(10)}_{BOOT}$	ME	1.69	2.74	5.02	14.62	43.69	80.86	95.91	98.61	99.49	10.38
	SD	2.67	3.52	4.85	7.09	2.40	6.45	3.92	2.11	1.11	4.24
$R^{(11)}_{BOOT}$	ME	1.92	3.07	5.49	15.33	43.90	80.24	95.52	98.42	99.41	9.67
	SD	2.89	3.77	5.09	7.19	2.35	6.47	4.15	2.29	1.24	4.37
$R^{(12)}_{BOOT}$	ME	2.16	3.40	5.95	16.00	44.10	79.65	95.15	98.23	99.32	9.00
	SD	3.11	4.00	5.32	7.26	2.32	6.49	4.35	2.46	1.37	4.49

Tabela B.30: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\mu_{\sqrt{w}}(F_n) - \mu_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição LOGN2 e $n = 100$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

LOGN2 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(1)}_{BOOT}$	ME	1.74	3.27	6.93	23.98	67.87	96.13	99.75	99.96	99.99	21.13
	SD	3.56	5.16	7.74	12.48	6.53	2.23	0.41	0.10	0.03	5.80
$R^{(2)}_{BOOT}$	ME	0.24	0.52	1.38	7.77	41.32	87.46	98.75	99.76	99.95	17.23
	SD	0.65	1.09	2.03	5.14	2.77	5.43	1.67	0.54	0.17	2.84
$R^{(3)}_{BOOT}$	ME	0.36	0.75	1.85	9.00	41.88	86.21	98.37	99.64	99.91	16.00
	SD	0.88	1.42	2.47	5.54	2.69	5.63	2.02	0.74	0.27	3.04
$R^{(4)}_{BOOT}$	ME	0.51	1.01	2.34	10.13	42.33	85.10	97.97	99.50	99.87	14.87
	SD	1.14	1.75	2.90	5.89	2.63	5.80	2.35	0.95	0.38	3.24
$R^{(5)}_{BOOT}$	ME	0.69	1.30	2.86	11.17	42.73	84.11	97.56	99.35	99.81	13.83
	SD	1.41	2.08	3.30	6.17	2.58	5.94	2.67	1.17	0.50	3.43
$R^{(6)}_{BOOT}$	ME	0.88	1.61	3.38	12.14	43.07	83.21	97.15	99.19	99.74	12.86
	SD	1.67	2.41	3.67	6.41	2.54	6.05	2.96	1.37	0.63	3.61
$R^{(7)}_{BOOT}$	ME	1.09	1.93	3.89	13.04	43.37	82.39	96.73	99.02	99.67	11.96
	SD	1.93	2.71	4.00	6.61	2.50	6.14	3.25	1.58	0.77	3.78
$R^{(8)}_{BOOT}$	ME	1.32	2.27	4.40	13.88	43.64	81.64	96.32	98.83	99.59	11.12
	SD	2.19	3.01	4.32	6.78	2.47	6.21	3.51	1.79	0.91	3.94
$R^{(9)}_{BOOT}$	ME	1.56	2.61	4.91	14.66	43.88	80.94	95.91	98.65	99.51	10.34
	SD	2.45	3.30	4.62	6.92	2.43	6.27	3.76	1.98	1.04	4.09
$R^{(10)}_{BOOT}$	ME	1.80	2.96	5.41	15.40	44.09	80.30	95.51	98.45	99.41	9.60
	SD	2.68	3.57	4.88	7.05	2.40	6.33	3.99	2.17	1.17	4.23
$R^{(11)}_{BOOT}$	ME	2.05	3.31	5.90	16.09	44.29	79.69	95.11	98.25	99.31	8.91
	SD	2.92	3.82	5.12	7.15	2.38	6.37	4.20	2.36	1.31	4.37
$R^{(12)}_{BOOT}$	ME	2.31	3.65	6.38	16.74	44.47	79.12	94.72	98.04	99.21	8.26
	SD	3.15	4.05	5.35	7.24	2.35	6.39	4.40	2.53	1.45	4.49

B.2 Resultados para o coeficiente de correlação linear

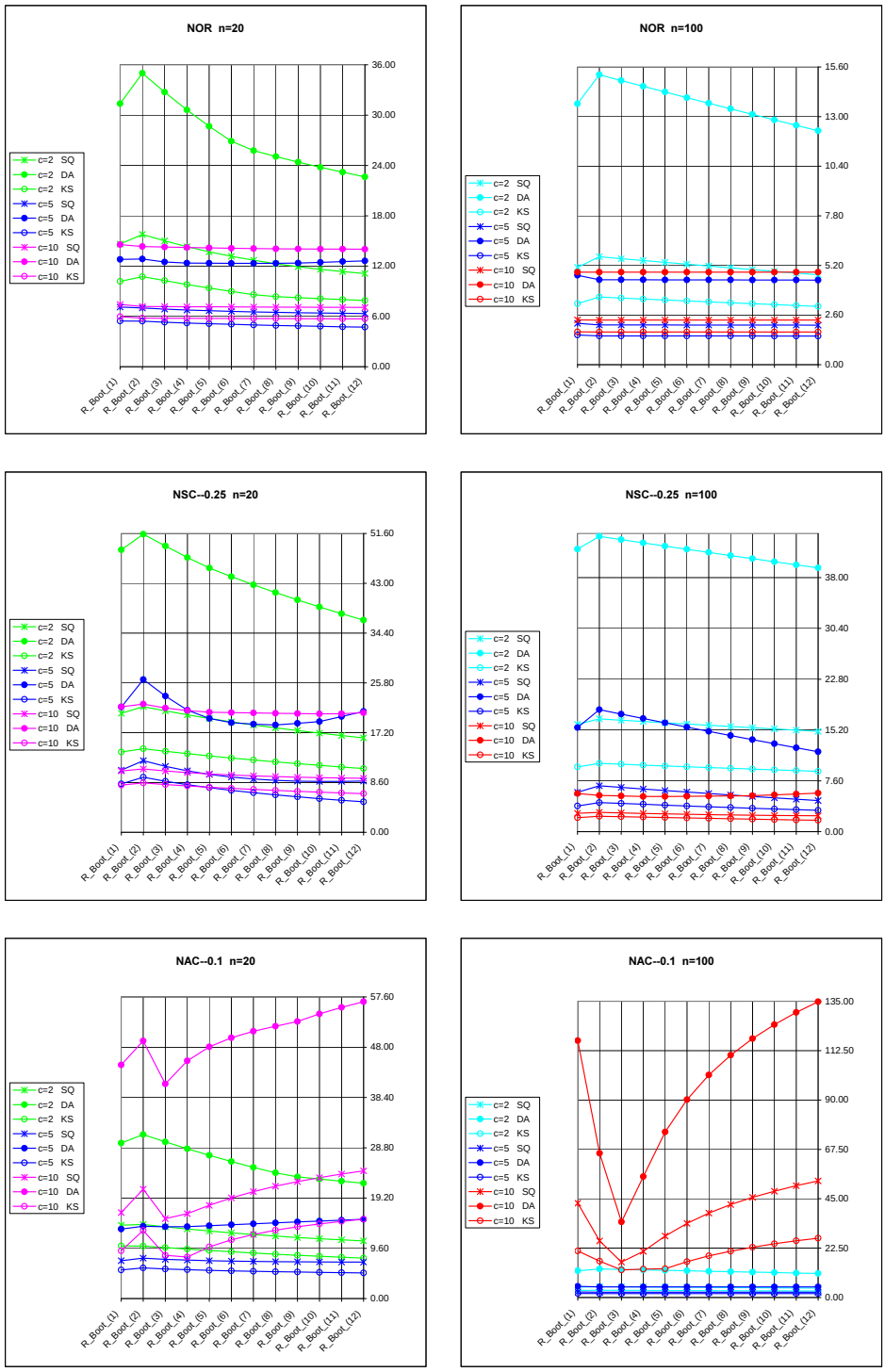


Figura B.4: Desempenho dos vários estimadores R_{Boot} na estimação da distribuição $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$ em termos dos indicadores globais, SQ, DA e KS, para as distribuições NOR, NSC–0.25 e NAC–0.1.

Tabela B.31: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NOR e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NOR $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	1.03	2.72	7.17	24.40	59.01	84.81	93.32	95.60	96.78	9.81
	SD	2.23	3.70	6.20	8.72	5.16	9.20	6.90	5.66	4.85	6.21
$R^{(5)}_{BOOT}$	ME	1.17	2.97	7.48	24.64	58.82	84.38	92.98	95.31	96.54	9.38
	SD	2.52	4.14	6.64	8.93	5.37	9.39	7.18	5.94	5.10	6.47
$R^{(6)}_{BOOT}$	ME	1.32	3.20	7.76	24.85	58.65	83.97	92.67	95.03	96.30	8.97
	SD	2.82	4.55	7.03	9.10	5.55	9.60	7.45	6.22	5.36	6.72
$R^{(7)}_{BOOT}$	ME	1.47	3.41	8.02	25.05	58.50	83.60	92.36	94.77	96.07	8.60
	SD	3.14	4.95	7.38	9.28	5.72	9.81	7.72	6.50	5.63	6.97
$R^{(8)}_{BOOT}$	ME	1.62	3.61	8.26	25.23	58.36	83.24	92.06	94.51	95.84	8.36
	SD	3.48	5.32	7.68	9.41	5.88	10.01	7.99	6.77	5.90	7.21
$R^{(9)}_{BOOT}$	ME	1.76	3.80	8.49	25.39	58.23	82.92	91.78	94.27	95.61	8.23
	SD	3.78	5.66	7.96	9.53	6.02	10.20	8.25	7.03	6.17	7.43
$R^{(10)}_{BOOT}$	ME	1.89	3.98	8.69	25.54	58.11	82.61	91.51	94.02	95.40	8.11
	SD	4.07	5.98	8.21	9.65	6.16	10.40	8.49	7.28	6.42	7.64
$R^{(11)}_{BOOT}$	ME	2.03	4.15	8.89	25.68	58.00	82.31	91.24	93.79	95.19	8.00
	SD	4.35	6.26	8.43	9.75	6.28	10.59	8.72	7.53	6.66	7.84
$R^{(12)}_{BOOT}$	ME	2.16	4.32	9.07	25.82	57.90	82.04	90.98	93.56	94.99	7.90
	SD	4.64	6.51	8.64	9.85	6.40	10.77	8.94	7.76	6.90	8.03
NOR $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	1.67	3.72	8.34	23.97	52.02	78.44	92.35	96.32	98.15	3.44
	SD	1.02	1.68	2.56	2.91	0.81	3.14	2.89	2.08	1.40	2.21
$R^{(5)}_{BOOT}$	ME	1.69	3.75	8.38	24.00	52.01	78.39	92.30	96.28	98.12	3.39
	SD	1.04	1.70	2.58	2.93	0.83	3.15	2.91	2.10	1.42	2.23
$R^{(6)}_{BOOT}$	ME	1.71	3.78	8.42	24.04	52.00	78.34	92.26	96.24	98.10	3.34
	SD	1.05	1.73	2.61	2.96	0.84	3.17	2.93	2.13	1.45	2.25
$R^{(7)}_{BOOT}$	ME	1.72	3.81	8.46	24.08	51.99	78.29	92.21	96.21	98.08	3.29
	SD	1.08	1.76	2.64	2.98	0.86	3.18	2.96	2.15	1.47	2.27
$R^{(8)}_{BOOT}$	ME	1.74	3.84	8.51	24.12	51.99	78.24	92.16	96.17	98.05	3.24
	SD	1.10	1.78	2.67	3.00	0.88	3.20	2.98	2.18	1.49	2.29
$R^{(9)}_{BOOT}$	ME	1.76	3.87	8.54	24.16	51.98	78.20	92.12	96.14	98.03	3.20
	SD	1.12	1.81	2.70	3.03	0.90	3.22	3.01	2.20	1.52	2.32
$R^{(10)}_{BOOT}$	ME	1.78	3.90	8.59	24.20	51.97	78.15	92.07	96.10	98.00	3.15
	SD	1.15	1.84	2.73	3.05	0.92	3.24	3.04	2.23	1.54	2.34
$R^{(11)}_{BOOT}$	ME	1.80	3.93	8.63	24.24	51.97	78.10	92.03	96.07	97.97	3.10
	SD	1.17	1.87	2.76	3.07	0.93	3.27	3.06	2.26	1.57	2.37
$R^{(12)}_{BOOT}$	ME	1.82	3.96	8.67	24.27	51.96	78.05	91.98	96.03	97.95	3.05
	SD	1.19	1.90	2.79	3.10	0.95	3.29	3.09	2.29	1.59	2.39

Tabela B.32: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NOR e $c = 5$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NOR $n = 20$ $c = 5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.61	5.09	9.75	24.73	53.87	80.22	91.67	94.95	96.66	5.22
	SD	2.94	4.32	5.92	6.96	2.79	8.10	6.57	5.29	4.32	5.52
$R^{(5)}_{BOOT}$	ME	2.66	5.15	9.81	24.77	53.85	80.13	91.57	94.86	96.58	5.13
	SD	3.07	4.46	6.02	7.02	2.86	8.21	6.74	5.46	4.49	5.64
$R^{(6)}_{BOOT}$	ME	2.71	5.21	9.86	24.81	53.83	80.06	91.48	94.78	96.51	5.06
	SD	3.22	4.58	6.13	7.07	2.92	8.31	6.90	5.65	4.69	5.75
$R^{(7)}_{BOOT}$	ME	2.76	5.25	9.91	24.84	53.81	79.99	91.40	94.70	96.44	4.99
	SD	3.36	4.70	6.23	7.11	2.97	8.40	7.04	5.82	4.88	5.86
$R^{(8)}_{BOOT}$	ME	2.80	5.30	9.97	24.88	53.80	79.93	91.33	94.63	96.37	4.93
	SD	3.49	4.81	6.32	7.16	3.02	8.48	7.18	5.98	5.05	5.97
$R^{(9)}_{BOOT}$	ME	2.85	5.35	10.01	24.91	53.79	79.87	91.26	94.56	96.30	4.87
	SD	3.61	4.93	6.41	7.19	3.06	8.56	7.30	6.13	5.22	6.06
$R^{(10)}_{BOOT}$	ME	2.88	5.39	10.06	24.94	53.77	79.81	91.20	94.49	96.24	4.81
	SD	3.73	5.03	6.49	7.23	3.10	8.63	7.42	6.27	5.38	6.15
$R^{(11)}_{BOOT}$	ME	2.92	5.44	10.10	24.96	53.76	79.76	91.13	94.43	96.19	4.76
	SD	3.84	5.13	6.57	7.26	3.13	8.70	7.53	6.40	5.52	6.24
$R^{(12)}_{BOOT}$	ME	2.96	5.48	10.14	24.99	53.75	79.72	91.07	94.37	96.13	4.72
	SD	3.94	5.22	6.64	7.30	3.16	8.76	7.64	6.52	5.65	6.32
NOR $n = 100$ $c = 5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.62	5.05	9.88	25.22	51.06	76.51	90.80	95.44	97.63	1.51
	SD	1.43	2.10	2.86	3.05	0.88	3.27	3.24	2.41	1.67	2.46
$R^{(5)}_{BOOT}$	ME	2.62	5.05	9.89	25.22	51.06	76.51	90.80	95.44	97.63	1.51
	SD	1.43	2.10	2.86	3.05	0.88	3.28	3.24	2.41	1.68	2.46
$R^{(6)}_{BOOT}$	ME	2.62	5.05	9.89	25.22	51.06	76.51	90.80	95.44	97.63	1.51
	SD	1.43	2.10	2.86	3.06	0.88	3.28	3.24	2.41	1.68	2.47
$R^{(7)}_{BOOT}$	ME	2.62	5.05	9.89	25.22	51.06	76.51	90.80	95.44	97.63	1.51
	SD	1.44	2.11	2.87	3.06	0.88	3.28	3.24	2.42	1.68	2.47
$R^{(8)}_{BOOT}$	ME	2.63	5.05	9.89	25.22	51.06	76.50	90.80	95.44	97.63	1.50
	SD	1.44	2.11	2.87	3.06	0.88	3.28	3.24	2.42	1.68	2.47
$R^{(9)}_{BOOT}$	ME	2.63	5.05	9.89	25.22	51.06	76.50	90.80	95.44	97.63	1.50
	SD	1.44	2.11	2.87	3.06	0.88	3.28	3.25	2.42	1.68	2.47
$R^{(10)}_{BOOT}$	ME	2.63	5.05	9.89	25.23	51.06	76.50	90.80	95.44	97.62	1.50
	SD	1.45	2.12	2.87	3.06	0.88	3.28	3.25	2.42	1.68	2.47
$R^{(11)}_{BOOT}$	ME	2.63	5.06	9.89	25.23	51.06	76.50	90.79	95.44	97.62	1.50
	SD	1.45	2.12	2.88	3.07	0.88	3.28	3.25	2.42	1.69	2.47
$R^{(12)}_{BOOT}$	ME	2.63	5.06	9.89	25.23	51.06	76.50	90.79	95.44	97.62	1.50
	SD	1.45	2.12	2.88	3.07	0.88	3.28	3.25	2.43	1.69	2.48

Tabela B.33: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NOR e $c = 10$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NOR $n = 20$ $c = 10$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.75	4.96	9.41	23.95	53.25	80.77	92.32	95.45	97.01	5.77
	SD	2.87	4.11	5.73	6.93	1.91	7.75	6.35	4.97	3.96	5.27
$R^{(5)}_{BOOT}$	ME	2.76	4.97	9.42	23.96	53.25	80.76	92.30	95.44	97.00	5.76
	SD	2.91	4.16	5.76	6.94	1.91	7.76	6.37	5.00	3.99	5.30
$R^{(6)}_{BOOT}$	ME	2.78	4.98	9.43	23.97	53.25	80.75	92.29	95.43	96.98	5.75
	SD	2.96	4.20	5.80	6.96	1.91	7.77	6.39	5.02	4.02	5.32
$R^{(7)}_{BOOT}$	ME	2.79	5.00	9.45	23.98	53.25	80.75	92.28	95.42	96.97	5.75
	SD	3.01	4.24	5.83	6.97	1.92	7.78	6.40	5.04	4.05	5.34
$R^{(8)}_{BOOT}$	ME	2.80	5.01	9.46	23.99	53.25	80.74	92.27	95.41	96.96	5.74
	SD	3.05	4.28	5.86	6.99	1.92	7.79	6.42	5.06	4.07	5.36
$R^{(9)}_{BOOT}$	ME	2.81	5.02	9.47	23.99	53.25	80.73	92.26	95.40	96.95	5.73
	SD	3.10	4.32	5.89	7.00	1.92	7.80	6.44	5.08	4.09	5.38
$R^{(10)}_{BOOT}$	ME	2.82	5.03	9.48	24.00	53.25	80.72	92.25	95.39	96.95	5.72
	SD	3.14	4.36	5.92	7.02	1.92	7.80	6.45	5.10	4.12	5.40
$R^{(11)}_{BOOT}$	ME	2.83	5.04	9.49	24.01	53.25	80.72	92.24	95.37	96.94	5.72
	SD	3.18	4.39	5.94	7.03	1.92	7.81	6.46	5.13	4.14	5.42
$R^{(12)}_{BOOT}$	ME	2.84	5.05	9.50	24.01	53.25	80.71	92.24	95.37	96.93	5.71
	SD	3.23	4.43	5.97	7.04	1.92	7.82	6.48	5.14	4.17	5.43
NOR $n = 100$ $c = 10$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.53	4.89	9.80	24.95	51.14	76.72	90.97	95.48	97.65	1.72
	SD	1.42	2.09	2.88	3.04	0.87	3.13	3.10	2.32	1.60	2.41
$R^{(5)}_{BOOT}$	ME	2.53	4.89	9.80	24.95	51.14	76.72	90.97	95.48	97.65	1.72
	SD	1.42	2.09	2.88	3.04	0.87	3.13	3.10	2.32	1.60	2.41
$R^{(6)}_{BOOT}$	ME	2.53	4.89	9.80	24.95	51.14	76.72	90.97	95.48	97.65	1.72
	SD	1.42	2.09	2.88	3.04	0.87	3.13	3.10	2.32	1.60	2.41
$R^{(7)}_{BOOT}$	ME	2.53	4.89	9.80	24.95	51.14	76.72	90.97	95.48	97.65	1.72
	SD	1.42	2.09	2.88	3.04	0.87	3.13	3.10	2.32	1.60	2.41
$R^{(8)}_{BOOT}$	ME	2.53	4.89	9.80	24.95	51.14	76.72	90.97	95.48	97.65	1.72
	SD	1.42	2.09	2.88	3.04	0.87	3.13	3.10	2.32	1.60	2.41
$R^{(9)}_{BOOT}$	ME	2.53	4.89	9.80	24.95	51.14	76.72	90.97	95.48	97.65	1.72
	SD	1.42	2.09	2.88	3.04	0.87	3.13	3.10	2.32	1.60	2.41
$R^{(10)}_{BOOT}$	ME	2.53	4.89	9.80	24.95	51.14	76.72	90.97	95.48	97.65	1.72
	SD	1.42	2.09	2.88	3.04	0.87	3.13	3.10	2.32	1.60	2.41
$R^{(11)}_{BOOT}$	ME	2.53	4.89	9.80	24.95	51.14	76.72	90.97	95.48	97.65	1.72
	SD	1.42	2.09	2.88	3.04	0.87	3.13	3.10	2.32	1.60	2.41
$R^{(12)}_{BOOT}$	ME	2.53	4.89	9.80	24.95	51.14	76.72	90.97	95.48	97.65	1.72
	SD	1.42	2.09	2.88	3.04	0.87	3.13	3.10	2.32	1.60	2.41

Tabela B.34: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NSC–0.25 e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NSC–0.25 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	0.69	1.69	4.73	20.14	61.46	88.58	95.21	96.83	97.62	13.58
	SD	1.27	2.28	4.32	8.28	7.40	9.20	6.06	4.83	4.09	5.87
$R^{(5)}_{BOOT}$	ME	0.76	1.80	4.93	20.37	61.20	88.19	94.93	96.63	97.45	13.19
	SD	1.41	2.45	4.53	8.39	7.48	9.42	6.29	5.02	4.27	6.02
$R^{(6)}_{BOOT}$	ME	0.83	1.92	5.11	20.58	60.97	87.82	94.67	96.41	97.28	12.82
	SD	1.55	2.63	4.72	8.49	7.55	9.66	6.53	5.24	4.47	6.18
$R^{(7)}_{BOOT}$	ME	0.90	2.03	5.29	20.78	60.75	87.48	94.40	96.20	97.10	12.48
	SD	1.69	2.80	4.90	8.58	7.61	9.88	6.79	5.47	4.68	6.34
$R^{(8)}_{BOOT}$	ME	0.97	2.14	5.46	20.96	60.56	87.16	94.14	95.99	96.92	12.16
	SD	1.83	2.97	5.09	8.66	7.68	10.10	7.06	5.71	4.90	6.50
$R^{(9)}_{BOOT}$	ME	1.04	2.25	5.62	21.13	60.38	86.86	93.90	95.78	96.73	11.86
	SD	1.97	3.15	5.26	8.74	7.75	10.31	7.33	5.95	5.14	6.67
$R^{(10)}_{BOOT}$	ME	1.11	2.36	5.78	21.29	60.21	86.56	93.66	95.57	96.55	11.56
	SD	2.11	3.32	5.42	8.82	7.81	10.52	7.59	6.21	5.37	6.83
$R^{(11)}_{BOOT}$	ME	1.18	2.46	5.92	21.45	60.05	86.28	93.43	95.37	96.37	11.28
	SD	2.25	3.47	5.58	8.89	7.87	10.72	7.84	6.47	5.61	6.98
$R^{(12)}_{BOOT}$	ME	1.25	2.57	6.07	21.60	59.91	86.02	93.22	95.18	96.19	11.02
	SD	2.38	3.64	5.72	8.96	7.92	10.90	8.08	6.72	5.85	7.13
NSC–0.25 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	0.43	1.27	3.89	17.54	51.96	84.96	96.37	98.62	99.41	9.96
	SD	0.47	1.04	2.21	4.31	1.38	4.57	2.72	1.53	0.86	2.54
$R^{(5)}_{BOOT}$	ME	0.44	1.29	3.95	17.63	51.94	84.83	96.29	98.58	99.39	9.83
	SD	0.49	1.06	2.24	4.32	1.43	4.58	2.76	1.56	0.89	2.56
$R^{(6)}_{BOOT}$	ME	0.46	1.32	4.00	17.73	51.91	84.71	96.22	98.54	99.37	9.71
	SD	0.50	1.09	2.27	4.34	1.48	4.61	2.80	1.60	0.92	2.58
$R^{(7)}_{BOOT}$	ME	0.47	1.35	4.06	17.83	51.89	84.59	96.15	98.50	99.35	9.59
	SD	0.52	1.11	2.30	4.36	1.53	4.64	2.85	1.64	0.95	2.61
$R^{(8)}_{BOOT}$	ME	0.49	1.38	4.12	17.91	51.87	84.47	96.08	98.46	99.33	9.47
	SD	0.54	1.14	2.34	4.38	1.58	4.67	2.90	1.68	0.99	2.64
$R^{(9)}_{BOOT}$	ME	0.50	1.42	4.18	18.00	51.85	84.36	96.01	98.42	99.30	9.36
	SD	0.56	1.17	2.38	4.40	1.64	4.72	2.95	1.74	1.03	2.68
$R^{(10)}_{BOOT}$	ME	0.52	1.45	4.24	18.09	51.83	84.24	95.94	98.37	99.28	9.24
	SD	0.58	1.21	2.41	4.43	1.70	4.76	3.00	1.78	1.07	2.71
$R^{(11)}_{BOOT}$	ME	0.53	1.48	4.30	18.18	51.82	84.13	95.86	98.33	99.25	9.13
	SD	0.60	1.24	2.46	4.46	1.75	4.81	3.06	1.84	1.11	2.75
$R^{(12)}_{BOOT}$	ME	0.55	1.51	4.35	18.26	51.80	84.02	95.79	98.29	99.23	9.02
	SD	0.62	1.27	2.49	4.49	1.81	4.86	3.12	1.90	1.16	2.79

Tabela B.35: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NSC–0.25 e $c = 5$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NSC–0.25 $n = 20$ $c = 5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.30	4.54	8.89	23.53	55.82	83.22	92.42	94.92	96.18	8.22
	SD	3.35	4.94	7.08	9.64	6.56	11.21	8.08	6.52	5.56	7.35
$R^{(5)}_{BOOT}$	ME	2.52	4.82	9.22	23.79	55.64	82.70	91.94	94.50	95.82	7.70
	SD	3.67	5.30	7.45	9.83	6.81	11.58	8.58	6.98	5.98	7.69
$R^{(6)}_{BOOT}$	ME	2.72	5.08	9.51	24.01	55.49	82.24	91.50	94.11	95.46	7.24
	SD	4.00	5.65	7.77	9.99	7.02	11.92	9.05	7.46	6.43	8.02
$R^{(7)}_{BOOT}$	ME	2.92	5.32	9.78	24.21	55.35	81.83	91.10	93.74	95.12	6.83
	SD	4.32	5.99	8.06	10.15	7.21	12.23	9.49	7.92	6.88	8.33
$R^{(8)}_{BOOT}$	ME	3.11	5.55	10.03	24.39	55.23	81.46	90.72	93.38	94.80	6.46
	SD	4.63	6.32	8.33	10.28	7.36	12.50	9.88	8.38	7.32	8.62
$R^{(9)}_{BOOT}$	ME	3.29	5.76	10.26	24.56	55.12	81.13	90.37	93.06	94.49	6.13
	SD	4.93	6.61	8.58	10.39	7.51	12.76	10.24	8.79	7.76	8.89
$R^{(10)}_{BOOT}$	ME	3.46	5.96	10.48	24.71	55.03	80.82	90.03	92.75	94.20	5.82
	SD	5.22	6.88	8.81	10.51	7.63	13.00	10.58	9.16	8.16	9.14
$R^{(11)}_{BOOT}$	ME	3.62	6.15	10.69	24.85	54.94	80.53	89.72	92.46	93.92	5.53
	SD	5.49	7.13	9.03	10.61	7.74	13.23	10.89	9.51	8.54	9.38
$R^{(12)}_{BOOT}$	ME	3.78	6.33	10.88	24.98	54.86	80.27	89.43	92.18	93.66	5.27
	SD	5.74	7.37	9.21	10.70	7.84	13.44	11.16	9.84	8.89	9.60
NSC–0.25 $n = 100$ $c = 5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	1.67	3.63	8.00	22.92	51.35	79.09	92.85	96.62	98.24	4.09
	SD	1.19	2.05	3.24	4.09	1.26	4.54	3.85	2.63	1.77	2.98
$R^{(5)}_{BOOT}$	ME	1.71	3.69	8.09	23.02	51.33	78.97	92.74	96.55	98.19	3.97
	SD	1.21	2.08	3.27	4.11	1.32	4.56	3.90	2.67	1.80	3.01
$R^{(6)}_{BOOT}$	ME	1.74	3.75	8.18	23.11	51.32	78.84	92.63	96.48	98.14	3.84
	SD	1.24	2.11	3.30	4.12	1.38	4.60	3.95	2.72	1.84	3.04
$R^{(7)}_{BOOT}$	ME	1.78	3.81	8.26	23.19	51.30	78.72	92.52	96.41	98.09	3.72
	SD	1.27	2.15	3.34	4.15	1.44	4.63	4.00	2.77	1.88	3.08
$R^{(8)}_{BOOT}$	ME	1.82	3.86	8.34	23.28	51.28	78.61	92.42	96.33	98.04	3.61
	SD	1.30	2.18	3.38	4.17	1.51	4.67	4.06	2.82	1.92	3.12
$R^{(9)}_{BOOT}$	ME	1.86	3.92	8.42	23.36	51.27	78.50	92.31	96.25	97.99	3.50
	SD	1.33	2.22	3.42	4.20	1.57	4.71	4.12	2.88	1.97	3.16
$R^{(10)}_{BOOT}$	ME	1.90	3.98	8.50	23.44	51.25	78.39	92.21	96.18	97.94	3.39
	SD	1.37	2.26	3.45	4.22	1.63	4.75	4.17	2.94	2.03	3.20
$R^{(11)}_{BOOT}$	ME	1.94	4.04	8.59	23.52	51.24	78.28	92.12	96.11	97.89	3.28
	SD	1.41	2.31	3.50	4.25	1.70	4.79	4.23	3.01	2.08	3.24
$R^{(12)}_{BOOT}$	ME	1.98	4.10	8.66	23.60	51.22	78.18	92.02	96.03	97.83	3.18
	SD	1.44	2.35	3.54	4.28	1.75	4.84	4.30	3.08	2.14	3.29

Tabela B.36: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NSC–0.25 e $c = 10$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NSC–0.25 $n = 20$ $c = 10$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.74	4.91	9.11	22.96	55.48	83.01	91.93	94.42	95.74	8.01
	SD	3.76	5.36	7.52	9.67	6.16	11.71	9.34	7.86	6.85	7.91
$R^{(5)}_{BOOT}$	ME	2.87	5.06	9.27	23.08	55.41	82.79	91.72	94.21	95.54	7.79
	SD	3.98	5.59	7.72	9.78	6.24	11.88	9.60	8.19	7.20	8.12
$R^{(6)}_{BOOT}$	ME	2.99	5.21	9.42	23.18	55.35	82.58	91.53	94.03	95.35	7.58
	SD	4.21	5.82	7.91	9.86	6.30	12.04	9.85	8.48	7.54	8.31
$R^{(7)}_{BOOT}$	ME	3.11	5.34	9.56	23.28	55.30	82.40	91.35	93.86	95.18	7.40
	SD	4.44	6.04	8.09	9.95	6.35	12.20	10.07	8.76	7.83	8.49
$R^{(8)}_{BOOT}$	ME	3.22	5.46	9.69	23.38	55.25	82.24	91.17	93.70	95.03	7.24
	SD	4.65	6.24	8.25	10.03	6.40	12.34	10.29	9.00	8.11	8.66
$R^{(9)}_{BOOT}$	ME	3.33	5.58	9.81	23.46	55.21	82.08	91.02	93.56	94.89	7.08
	SD	4.85	6.43	8.40	10.10	6.44	12.48	10.48	9.23	8.35	8.82
$R^{(10)}_{BOOT}$	ME	3.43	5.68	9.92	23.54	55.17	81.93	90.86	93.42	94.75	6.93
	SD	5.05	6.59	8.54	10.16	6.48	12.61	10.66	9.45	8.59	8.96
$R^{(11)}_{BOOT}$	ME	3.53	5.79	10.02	23.61	55.13	81.80	90.72	93.28	94.62	6.80
	SD	5.23	6.75	8.68	10.22	6.51	12.73	10.82	9.66	8.81	9.10
$R^{(12)}_{BOOT}$	ME	3.62	5.89	10.13	23.68	55.09	81.67	90.58	93.15	94.50	6.67
	SD	5.41	6.91	8.80	10.28	6.55	12.84	10.99	9.86	9.02	9.23
NSC–0.25 $n = 100$ $c = 10$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.66	5.14	9.99	25.15	51.27	77.17	90.98	95.25	97.37	2.17
	SD	1.86	2.83	3.95	4.37	1.54	4.79	4.40	3.25	2.30	3.45
$R^{(5)}_{BOOT}$	ME	2.69	5.18	10.04	25.20	51.26	77.11	90.92	95.20	97.33	2.11
	SD	1.89	2.85	3.98	4.39	1.57	4.82	4.44	3.29	2.34	3.47
$R^{(6)}_{BOOT}$	ME	2.72	5.22	10.09	25.24	51.26	77.05	90.86	95.15	97.29	2.05
	SD	1.91	2.88	4.00	4.40	1.60	4.85	4.47	3.33	2.37	3.50
$R^{(7)}_{BOOT}$	ME	2.74	5.26	10.15	25.29	51.25	76.99	90.79	95.10	97.25	1.99
	SD	1.93	2.90	4.03	4.42	1.63	4.87	4.51	3.37	2.41	3.52
$R^{(8)}_{BOOT}$	ME	2.77	5.30	10.20	25.34	51.24	76.93	90.73	95.05	97.21	1.93
	SD	1.95	2.93	4.06	4.43	1.65	4.90	4.55	3.41	2.45	3.55
$R^{(9)}_{BOOT}$	ME	2.80	5.34	10.25	25.38	51.24	76.87	90.68	95.00	97.18	1.87
	SD	1.98	2.96	4.08	4.45	1.68	4.93	4.58	3.45	2.49	3.58
$R^{(10)}_{BOOT}$	ME	2.83	5.38	10.30	25.43	51.23	76.81	90.61	94.95	97.14	1.81
	SD	2.01	2.99	4.12	4.48	1.71	4.96	4.62	3.50	2.53	3.61
$R^{(11)}_{BOOT}$	ME	2.86	5.42	10.34	25.47	51.23	76.75	90.56	94.90	97.10	1.75
	SD	2.04	3.02	4.15	4.49	1.74	4.99	4.67	3.54	2.58	3.64
$R^{(12)}_{BOOT}$	ME	2.89	5.46	10.39	25.51	51.22	76.70	90.50	94.85	97.06	1.70
	SD	2.07	3.05	4.18	4.52	1.77	5.02	4.71	3.59	2.62	3.68

Tabela B.37: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $c = 2$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NAC–0.1 $n = 20$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	1.17	2.86	7.47	24.32	57.87	84.46	93.34	95.78	97.02	9.46
	SD	1.85	3.26	5.65	7.98	4.07	8.61	6.50	5.24	4.38	5.67
$R^{(5)}_{BOOT}$	ME	1.27	3.01	7.68	24.49	57.77	84.19	93.13	95.60	96.87	9.19
	SD	2.09	3.54	5.92	8.13	4.18	8.73	6.68	5.43	4.56	5.84
$R^{(6)}_{BOOT}$	ME	1.38	3.16	7.88	24.65	57.68	83.94	92.93	95.42	96.73	8.94
	SD	2.34	3.82	6.18	8.27	4.29	8.84	6.86	5.62	4.74	6.00
$R^{(7)}_{BOOT}$	ME	1.48	3.31	8.06	24.79	57.60	83.70	92.73	95.25	96.58	8.70
	SD	2.59	4.10	6.44	8.41	4.39	8.97	7.03	5.81	4.94	6.17
$R^{(8)}_{BOOT}$	ME	1.57	3.45	8.23	24.92	57.52	83.48	92.55	95.09	96.44	8.48
	SD	2.82	4.34	6.68	8.53	4.48	9.10	7.21	6.00	5.12	6.34
$R^{(9)}_{BOOT}$	ME	1.67	3.58	8.40	25.04	57.45	83.27	92.36	94.93	96.30	8.27
	SD	3.04	4.59	6.91	8.63	4.57	9.24	7.39	6.17	5.30	6.49
$R^{(10)}_{BOOT}$	ME	1.77	3.72	8.56	25.16	57.39	83.08	92.18	94.78	96.16	8.08
	SD	3.26	4.82	7.12	8.74	4.65	9.38	7.55	6.34	5.49	6.65
$R^{(11)}_{BOOT}$	ME	1.87	3.85	8.70	25.26	57.32	82.90	92.00	94.63	96.02	7.90
	SD	3.47	5.04	7.32	8.83	4.74	9.51	7.71	6.51	5.65	6.80
$R^{(12)}_{BOOT}$	ME	1.97	3.97	8.84	25.36	57.26	82.72	91.83	94.48	95.89	7.72
	SD	3.67	5.25	7.52	8.91	4.81	9.64	7.86	6.69	5.82	6.94

NAC–0.1 $n = 100$ $c = 2$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	1.91	3.92	8.61	24.12	51.58	78.22	92.08	96.16	98.08	3.22
	SD	1.18	1.84	2.74	3.10	0.85	3.31	3.08	2.24	1.53	2.37
$R^{(5)}_{BOOT}$	ME	1.92	3.94	8.64	24.15	51.57	78.18	92.05	96.14	98.06	3.18
	SD	1.19	1.86	2.76	3.11	0.86	3.31	3.09	2.25	1.55	2.38
$R^{(6)}_{BOOT}$	ME	1.93	3.96	8.67	24.18	51.57	78.15	92.02	96.12	98.05	3.15
	SD	1.21	1.88	2.78	3.13	0.86	3.31	3.10	2.27	1.56	2.39
$R^{(7)}_{BOOT}$	ME	1.95	3.98	8.70	24.21	51.56	78.11	91.99	96.09	98.03	3.11
	SD	1.23	1.90	2.80	3.14	0.87	3.32	3.11	2.28	1.57	2.40
$R^{(8)}_{BOOT}$	ME	1.96	4.00	8.73	24.24	51.56	78.08	91.96	96.07	98.02	3.08
	SD	1.25	1.92	2.82	3.16	0.88	3.32	3.12	2.29	1.58	2.42
$R^{(9)}_{BOOT}$	ME	1.98	4.02	8.76	24.26	51.56	78.04	91.93	96.04	98.00	3.04
	SD	1.26	1.94	2.84	3.18	0.89	3.33	3.14	2.31	1.60	2.43
$R^{(10)}_{BOOT}$	ME	1.99	4.04	8.78	24.29	51.55	78.01	91.89	96.02	97.98	3.01
	SD	1.29	1.97	2.86	3.19	0.90	3.34	3.15	2.33	1.61	2.44
$R^{(11)}_{BOOT}$	ME	2.01	4.06	8.81	24.32	51.55	77.98	91.86	96.00	97.97	2.98
	SD	1.30	1.99	2.89	3.21	0.91	3.34	3.17	2.34	1.63	2.46
$R^{(12)}_{BOOT}$	ME	2.02	4.08	8.84	24.34	51.54	77.95	91.83	95.97	97.95	2.95
	SD	1.33	2.01	2.91	3.23	0.92	3.35	3.18	2.36	1.64	2.48

Tabela B.38: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $c = 5$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NAC–0.1 $n = 20$ $c = 5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.68	5.25	9.99	25.23	54.36	80.51	91.69	94.71	96.29	5.51
	SD	3.25	4.63	6.25	7.50	3.30	8.97	7.28	6.00	5.07	6.09
$R^{(5)}_{BOOT}$	ME	2.74	5.31	10.05	25.27	54.32	80.40	91.55	94.57	96.16	5.40
	SD	3.40	4.76	6.36	7.55	3.44	9.17	7.57	6.30	5.40	6.27
$R^{(6)}_{BOOT}$	ME	2.80	5.38	10.11	25.30	54.29	80.30	91.42	94.45	96.05	5.30
	SD	3.56	4.90	6.46	7.59	3.55	9.35	7.83	6.58	5.69	6.44
$R^{(7)}_{BOOT}$	ME	2.86	5.44	10.17	25.32	54.26	80.21	91.31	94.34	95.94	5.21
	SD	3.70	5.03	6.56	7.63	3.64	9.50	8.07	6.85	5.97	6.60
$R^{(8)}_{BOOT}$	ME	2.91	5.49	10.21	25.35	54.24	80.14	91.22	94.24	95.84	5.14
	SD	3.84	5.15	6.65	7.67	3.71	9.64	8.29	7.09	6.24	6.74
$R^{(9)}_{BOOT}$	ME	2.96	5.54	10.26	25.38	54.22	80.07	91.12	94.15	95.75	5.07
	SD	3.97	5.26	6.74	7.71	3.78	9.75	8.50	7.33	6.48	6.88
$R^{(10)}_{BOOT}$	ME	3.00	5.58	10.30	25.40	54.20	80.01	91.05	94.07	95.67	5.01
	SD	4.10	5.37	6.82	7.75	3.84	9.86	8.68	7.54	6.70	7.00
$R^{(11)}_{BOOT}$	ME	3.04	5.63	10.34	25.42	54.18	79.95	90.97	93.99	95.59	4.95
	SD	4.21	5.47	6.89	7.78	3.88	9.96	8.84	7.74	6.91	7.12
$R^{(12)}_{BOOT}$	ME	3.09	5.67	10.38	25.44	54.16	79.90	90.91	93.91	95.52	4.90
	SD	4.32	5.57	6.97	7.81	3.94	10.05	8.99	7.93	7.10	7.23

NAC–0.1 $n = 100$ $c = 5$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	2.52	5.06	9.89	25.09	51.16	76.81	91.02	95.50	97.58	1.81
	SD	1.39	2.09	2.89	3.10	0.92	3.39	3.24	2.42	1.73	2.49
$R^{(5)}_{BOOT}$	ME	2.52	5.06	9.89	25.09	51.16	76.81	91.02	95.50	97.58	1.81
	SD	1.39	2.09	2.89	3.10	0.92	3.39	3.24	2.42	1.73	2.49
$R^{(6)}_{BOOT}$	ME	2.52	5.06	9.89	25.10	51.16	76.81	91.02	95.50	97.58	1.81
	SD	1.39	2.09	2.89	3.10	0.92	3.39	3.24	2.42	1.73	2.49
$R^{(7)}_{BOOT}$	ME	2.52	5.06	9.89	25.10	51.16	76.81	91.02	95.50	97.58	1.81
	SD	1.39	2.09	2.89	3.11	0.92	3.39	3.24	2.42	1.73	2.50
$R^{(8)}_{BOOT}$	ME	2.53	5.06	9.89	25.10	51.16	76.81	91.02	95.50	97.57	1.81
	SD	1.39	2.09	2.89	3.11	0.92	3.39	3.24	2.42	1.73	2.50
$R^{(9)}_{BOOT}$	ME	2.53	5.06	9.89	25.10	51.16	76.80	91.01	95.49	97.57	1.80
	SD	1.39	2.10	2.89	3.11	0.92	3.39	3.25	2.42	1.73	2.50
$R^{(10)}_{BOOT}$	ME	2.53	5.06	9.90	25.10	51.16	76.80	91.01	95.49	97.57	1.80
	SD	1.39	2.10	2.89	3.11	0.92	3.39	3.25	2.42	1.73	2.50
$R^{(11)}_{BOOT}$	ME	2.53	5.06	9.90	25.10	51.16	76.80	91.01	95.49	97.57	1.80
	SD	1.39	2.10	2.90	3.11	0.92	3.39	3.25	2.42	1.73	2.50
$R^{(12)}_{BOOT}$	ME	2.53	5.06	9.90	25.10	51.16	76.80	91.01	95.49	97.57	1.80
	SD	1.39	2.10	2.90	3.11	0.92	3.40	3.25	2.43	1.73	2.50

Tabela B.39: Comparação dos estimadores *bootstrap* IF corrigidos na estimação da distribuição de $\rho_{\sqrt{w}}(F_n) - \rho_{\sqrt{w}}(F)$, quando o modelo gerador dos dados é a distribuição NAC–0.1 e $c = 10$. Na última coluna de cada tabela, os valores indicados correspondem respectivamente nas linhas ímpares a KS e nas linhas pares a SC.

NAC–0.1 $n = 20$ $c = 10$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	1.07	1.80	4.14	19.72	56.82	79.32	86.19	88.26	89.56	7.94
	SD	2.10	2.76	4.07	6.95	16.08	22.20	20.82	19.63	18.58	14.88
$R^{(5)}_{BOOT}$	ME	1.24	2.01	4.37	19.71	55.98	77.84	84.41	86.40	87.63	9.87
	SD	2.38	3.04	4.29	7.01	17.02	23.93	23.17	22.23	21.37	16.44
$R^{(6)}_{BOOT}$	ME	1.38	2.18	4.56	19.71	55.40	76.81	83.18	85.10	86.28	11.22
	SD	2.60	3.27	4.48	7.06	17.62	25.03	24.66	23.90	23.18	17.45
$R^{(7)}_{BOOT}$	ME	1.51	2.33	4.73	19.72	54.98	76.03	82.27	84.12	85.27	12.23
	SD	2.79	3.46	4.64	7.11	18.05	25.82	25.69	25.08	24.46	18.17
$R^{(8)}_{BOOT}$	ME	1.63	2.47	4.87	19.74	54.63	75.42	81.54	83.36	84.45	13.05
	SD	2.96	3.62	4.79	7.15	18.37	26.42	26.48	25.95	25.43	18.72
$R^{(9)}_{BOOT}$	ME	1.73	2.59	5.00	19.76	54.35	74.91	80.93	82.72	83.80	13.70
	SD	3.10	3.76	4.92	7.19	18.63	26.88	27.11	26.64	26.16	19.14
$R^{(10)}_{BOOT}$	ME	1.83	2.70	5.11	19.78	54.11	74.48	80.42	82.18	83.25	14.25
	SD	3.23	3.89	5.03	7.23	18.84	27.28	27.60	27.21	26.77	19.50
$R^{(11)}_{BOOT}$	ME	1.92	2.80	5.22	19.80	53.90	74.10	79.98	81.71	82.76	14.74
	SD	3.35	4.01	5.14	7.27	19.02	27.60	28.03	27.68	27.28	19.79
$R^{(12)}_{BOOT}$	ME	2.00	2.89	5.32	19.83	53.73	73.77	79.59	81.30	82.34	15.16
	SD	3.45	4.12	5.24	7.30	19.16	27.89	28.38	28.08	27.71	20.04
NAC–0.1 $n = 100$ $c = 10$											
p		2.5	5	10	25	50	75	90	95	97.5	$KS\|SC$
$R^{(4)}_{BOOT}$	ME	0.95	1.44	2.54	12.07	45.84	72.82	83.31	86.84	89.10	12.93
	SD	1.29	1.87	2.50	3.50	17.57	24.80	22.31	19.99	17.98	15.52
$R^{(5)}_{BOOT}$	ME	1.06	1.52	2.63	11.87	43.38	68.56	78.68	82.25	84.59	13.13
	SD	1.40	1.90	2.41	3.58	18.85	27.59	26.44	24.70	23.06	18.20
$R^{(6)}_{BOOT}$	ME	1.14	1.59	2.72	11.77	41.69	65.54	75.35	78.87	81.19	16.31
	SD	1.47	1.92	2.35	3.67	19.57	29.19	28.89	27.60	26.31	19.86
$R^{(7)}_{BOOT}$	ME	1.20	1.66	2.81	11.71	40.41	63.26	72.77	76.24	78.54	18.96
	SD	1.53	1.93	2.31	3.75	20.04	30.19	30.48	29.53	28.50	20.96
$R^{(8)}_{BOOT}$	ME	1.26	1.72	2.90	11.68	39.39	61.42	70.69	74.11	76.39	21.11
	SD	1.58	1.94	2.28	3.84	20.36	30.89	31.60	30.87	30.05	21.73
$R^{(9)}_{BOOT}$	ME	1.31	1.77	2.97	11.67	38.55	59.90	68.95	72.32	74.59	22.91
	SD	1.61	1.95	2.25	3.92	20.59	31.40	32.39	31.86	31.19	22.29
$R^{(10)}_{BOOT}$	ME	1.36	1.82	3.05	11.66	37.85	58.60	67.45	70.79	73.04	24.46
	SD	1.64	1.96	2.23	3.99	20.76	31.78	33.00	32.60	32.07	22.73
$R^{(11)}_{BOOT}$	ME	1.40	1.88	3.13	11.67	37.24	57.47	66.16	69.45	71.67	25.83
	SD	1.67	1.97	2.22	4.06	20.88	32.06	33.48	33.20	32.74	23.06
$R^{(12)}_{BOOT}$	ME	1.44	1.93	3.20	11.67	36.71	56.48	65.01	68.27	70.47	27.03
	SD	1.69	1.97	2.21	4.12	20.99	32.29	33.86	33.67	33.29	23.33

APÊNDICE C

Macros S

Este apêndice contém as macros relativas à implementação de alguns procedimentos referidos no Capítulo 4. Essas macros são apresentadas sem grandes comentários, pois assume-se que o leitor está familiarizado com os comandos básicos do *software* estatístico S-PLUS.

C.1 *Bootstrap* IF para a localização univariada

```
"Media.boot"<-function(data, B1 = 50000, nivelc=0.95, prej=F,
psi="infty",reamost = "if", estimador = "C", caf=sqrt(qchisq(0.99,1)), alfa=10)
{
#-------------------------------------------------------------------------------#
#  Realiza bootstrap para media univariada                                      #
#  usando o bootstrap do S+2000 usual e o que modifiquei(bootifgr)              #
#  data:       vector que corresponde aos dados de trabalho                     #
#  B1:         numero de reamostras bootstrap pretendidas                       #
#  reamost:    if                                                               #
#              unifo                                                            #
#  estimador:  C: media aritmetica:                                             #
#              R: LTS                                                           #
#  caf:        constante de afinacao (alto percentil da sif)                    #
#  alfa:       expoente da funcao peso                                          #
#  nivelc:     nivel pretendido para o intervalo de confianca                   #
#-------------------------------------------------------------------------------#
datat <- NULL
```

```
#-------------------------------------------------------------------------------#
# Funcao:mean.rob                                                               #
# funcao que determina a media robusta por lts                                  #
#-------------------------------------------------------------------------------#
mean.rob <-function(data)
{
mdr<-NA
mdr<-location.lts(data)$loc
return(mdr)
}


#-------------------------------------------------------------------------------#
# Funcao: peso.fun                                                              #
# funcao para construcao dos pesos (baseados em IF) para reamostragem da media  #
# univariada                                                                    #
peso.fun <- function(dat, n, caf, alfa,prej,psi)
{
#
# obtencao dos estimadores robustos de localizacao e escala
#
p <- NULL
d <- NULL
d2<-NULL
daux <- NULL
mau<-NULL
#
# estimadores robustos: mediana e MAD
#
ife <- dat-median(dat)
varif <- mad(dat)
daux <- ife/sqrt(varif)
d <- abs(daux)
#
# primeiro if para funcao psi geral com alfa fixo
#
w <- NULL
if ((psi=="falfa") & (prej==F))
{
  # expressao auxiliar para calculo dos primeiros wi
               exp1<-1/((1 + (((d - caf)^2)/(alfa * (caf^2))))^((alfa + 1)/2))
               w<-ifelse(d<=caf,1,exp1)
          }
#
# segundo if para funcao psi com alfa=infinito
#
```

```
   if ((psi=="infty") & (prej==F))
      {
# expressao auxiliar para calculo dos segundos wi
           exp2<-exp(-(((d - caf)^2)/(2* (caf^2))))
           w<-ifelse(d<=caf,1,exp2)
     }
#
# terceiro if para triming
#
   if (prej==T) # faz a rejeicao bruta (wi=1 ou 0) de observacoes e conta-as
      {
   contador<-0
      for(i in 1:n) {
if(d[i] <= caf)
w[i] <- 1
else{
contador<-contador+1
w[i]<-0
if (contador==1) mau<-c(i)
else mau<-c(mau,i)
 }
      }
     }
p <- w/sum(w)
list(p = p,mau=mau)
}

samp.boot.ifs2 <- function(n, B, pb = NULL)
{
# Funcao para reamostrar no bootstrap com pesos. Cada coluna é uma amostra de indices.
# para bootif do S+2000
xx<-sample(n, B * n, prob = pb, replace = T)
dim(xx) <- c(n, B)
xx
}
#-------------------------------------------------------------------------------#
#                                                                               #
#                            Programa principal                                 #
#                                                                               #
#-------------------------------------------------------------------------------#
#
# datat corresponde aos dados de trabalho que são transformados em matriz
#
    pindi<-NULL
    if(!(length(data) == 1)) {
if((is.vector(data)) | (is.data.frame(data)) | (is.matrix(data)))
```

```
datat <- as.matrix(data)
}
else (stop("ERRO:os dados nao sao matriz, vector ou data.frame"))
n <- nrow(datat)
if(estimador == "C") {
if(reamost == "if") {
saofp<-peso.fun(datat, n, caf, alfa,prej,psi)
pindi<-saofp$mau
if(!is.null(pindi))
{psai <- saofp$p[-pindi]
datat<-datat[-pindi,]
n<-nrow(datat)
}
else {psai<-saofp$p
  }
temp <- bootifgr(data = datat, B = B1, statistic = mean, sampler =
samp.boot.ifs2, args.sampler = psai, trace = F)}
else if(reamost == "unifo") {
temp <- bootstrap(datat, B = B1, statistic = mean, trace = F)
}
al<-1-nivelc
limites <- quantile(temp$replicates, probs = c(al/2, 1-(al/2)), na.rm = T)
    names(temp$observed) <- NULL
names(limites)<-NULL
return(temp$replicates,list(original.mean = temp$observed,
                   meanb.mean = temp$estimate$Mean, medianb.mean=median(tem$replicates),
                   trmb.mean=mean(temp$replicates,trim=0.05),
liminf = limites[1], limsup = limites[2]))
}

else if(estimador == "R") {
if(reamost == "if") {
psai <- peso.fun(datat, n, caf, alfa,prej,psi)$p
#print(psai*100)
temp <- bootifgr(data = datat, B = B1, statistic =mean.rob, sampler =
samp.boot.ifs2, args.sampler = psai, trace = F)
}
else if(reamost == "unifo") {
temp <- bootstrap(datat, B = B1, statistic = mean.rob, trace = F)
}
al<-1-nivelc
limites <- quantile(temp$replicates, probs = c(al/2, 1-(al/2)), na.rm = T)
names(temp$observed) <- NULL
return(list(original.mean = temp$observed, meanb.mean = temp$estimate$Mean,
medianb.mean=median(temp$replicates),trmb.mean=mean(temp$replicates,trim=0.05),
liminf = limites[1], limsup = limites[2]))
```

```
}
}
```

C.2 *Bootstrap* IF para coeficiente de correlação linear

```
"Compt.bootif" <- function(data, B1 = 100, prej=F,psi="falfa",reamost = "unifo",
                  estimador = "C", caf=10, alfa=10,nivelc=0.90)
{
#-------------------------------------------------------------------------------#
#  Realiza bootstrap para coeficiente de correlacao linear                      #
#  usando o bootstrap do S+ usual e o que modifiquei(bootpesos)                 #
#  data: matriz, ou data frame, corresponde aos dados de trabalho(nx2)          #
#  reamost:         if                                                          #
#                   unifo                                                       #
#  estimador:       C (usual, Pearson)                                          #
#                   R (MCD  com 50% ou 25% de ponto de rotura)                  #
#  caf:             constante de afinacao                                       #
#  alfa:            expoente da funcao peso                                     #
#  nivelc:          nivel pretendido para o intervalo de confianca              #
#  Ultima Modificacao: 30/10/2001                                               #
#-------------------------------------------------------------------------------#
   now<-proc.time()
   datat <- NULL

#-------------------------------------------------------------------------------#
# Funcao: corr.fun                                                              #
# funcao para determinar o coeficiente de correlação de Pearson                 #
# entram os dados (uma amostra bivariada) e sai valor da correlacao             #
#                                                                               #
#-------------------------------------------------------------------------------#
corr.fun <- function(dat)
{
vcorr <- cor(dat)[1, 2]
}
MCD.fun <- function(dat)
{
#-------------------------------------------------------------------------------#
# Funcao: MCD.fun                                                               #
# funcao para determinar o coeficiente de correlação por MCD                    #
# entram os dados (uma amostra bivariada) e sai valor da correlacao             #
#                                                                               #
#-------------------------------------------------------------------------------#
#mcd25
x <- cov.mcd(dat, cor = T, print = F, quan=0.75*(nrow(dat)))
#mcd50
```

```
#x <- cov.mcd(dat, cor = T, print = F)
return((x$cor)[1, 2])
}
#-------------------------------------------------------------------------------#
# Funcao: peso.fun                                                              #
# funcao para construcao dos pesos para reamostragem de rho                     #
# entram os dados (uma amostra bivariada) e sai vector de pesos                 #
#                                                                               #
#-------------------------------------------------------------------------------#
peso.fun <- function(dat, n, caf, alfa,prej,psi)
{
#
# obtencao dos estimadores robustos de localizacao e escala
#
p <- NULL
rorobu <- NA
rb <- NULL
nc <- ncol(dat)


y <- matrix(NA, n, nc)
d2 <- NULL
d <- NULL
daux <- NULL
mau<-NULL
#
# estimacao por MCD-estimator 50%
#
rb <- cov.mcd(dat, cor = T, print = F)
#print("cov.mcd nos pesos")
#print(rb)
center <- rb$center
cov <- rb$cov
rorobu <- rb$cor[1, 2]
for(i in 1:nc) {
y[, i] <- (dat[, i] - center[i])/sqrt(cov[i, i])
}
ife <- (-0.5 * rorobu) * (y[, 1]^2 + y[, 2]^2) + y[, 1] * y[, 2]
varif <- (1 - rorobu^2)^2
daux <- ife/sqrt(varif)
d <- abs(daux)
#d2 <- (ife^2)/varif
#
# primeiro if para funcao psi geral com alfa fixo
#
w <- NULL
```

```
if ((psi=="falfa") & (prej==F))
{
      # expressao auxiliar para calculo dos primeiros wi
               exp1<-1/((1 + (((d - caf)^2)/(alfa * (caf^2))))^((alfa + 1)/2))
               w<-ifelse(d<=caf,1,exp1)
          }
      #
# segundo if para funcao psi com alfa=infinito
#
   if ((psi=="infty") & (prej==F))
      {
 # expressao auxiliar para calculo dos segundos wi
           exp2<-exp(-(((d - caf)^2)/(2* (caf^2))))
           w<-ifelse(d<=caf,1,exp2)
     }
    #
# terceiro if para hard rejection
#
   if (prej==T) # faz a rejeicao bruta (wi=1 ou 0) de observacoes e conta-as
      {
contador<-0
      for(i in 1:n) {
if(d[i] <= caf)
w[i] <- 1
else{
contador<-contador+1
w[i]<-0
if (contador==1) mau<-c(i)
else mau<-c(mau,i)
 }
      }
     }
p <- w/sum(w)
list(p = p, w=w,rorobu = rorobu,mau=mau)
}

samp.boot.if <- function(inds, B, pb = NULL)
{
# Funcao para reamostrar no bootstrap com pesos.Cada coluna é uma amostra dos indices.
# versao para bootstrap do Splus4.5
n <- length(inds)
if(!is.null(pb)) matrix(sample(inds, B * n, prob = pb, replace = T), nrow = n)
else print("pesos nulos verificar o que esta a ocorrer")
}

#-------------------------------------------------------------------------------#
```

```
#                                                                          #
#                            Programa principal                            #
#                                                                          #
#--------------------------------------------------------------------------#
#
# datat corresponde aos dados de trabalho que são transformados em matriz
#
    pindi<-NULL
if(!(length(data) == 1)) {
if((is.vector(data)) | (is.data.frame(data)) | (is.matrix(data)))
datat <- as.matrix(data)
}
else (stop("ERRO:os dados nao sao matriz, vector ou data.frame"))
n <- nrow(datat)
if(estimador == "C") {
if(reamost == "if") {
saofp<-peso.fun(datat, n, caf, alfa,prej,psi)
pindi<-saofp$mau
if(!is.null(pindi))
{psai <- saofp$w[-pindi]
datat<-datat[-pindi,]
# n<-nrow(datat)
}
else {psai<-saofp$p }
temp <- bootpesos(data = datat, B = B1, statistic = corr.fun, sampler =
samp.boot.new, args.sampler = psai, trace = F)
}
else if(reamost == "unifo") {
temp <- bootstrap(datat, B = B1, statistic = corr.fun, trace = F)
}
            al<-1-nivelc
limites <- quantile(temp$replicates, probs =  c(al/2, 1-(al/2)), na.rm = T)
names(temp$observed) <- NULL
names(limites)<-NULL
return(list(rhooriginal = temp$observed, rhoestmean = temp$estimate$Mean,
rhoestmedian=median(temp$replicates),rhoesttrm=mean(temp$replicates,trim=0.05),
liminf = limites[1], limsup = limites[2]))
}
else if(estimador == "R") {
rhoclassico <- cor(datat)[1, 2]
if(reamost == "if") {
psai <- peso.fun(datat, n, caf, alfa,prej,psi)$p
temp <- bootpesos(data = datat, B = B1, statistic = MCD.fun, sampler =
samp.boot.if, args.sampler = psai, trace = F)
}
else if(reamost == "unifo") {
```

```
temp <- bootstrap(datat, B = B1, statistic = MCD.fun, trace = F)
}
            al<-1-nivelc
limites <- quantile(temp$replicates, probs =c(al/2, 1-(al/2)), na.rm = T)
names(temp$observed) <- NULL
#print(proc.time()-now)
return(list(rhooriginal = rhoclassico, rhorobusto = temp$observed,
       rhoestimate= temp$estimate$Mean,rhoestmedian=median(temp$replicates),
    rhoesttrm=mean(temp$replicates,trim=0.05),liminf = limites[1],
                     limsup = limites[2]))
}
}
```

C.3 *Bootstrap* IF para selecção de variáveis em ADL

```
"SelVallBoot"<-function(dat,k,B1=100,estalfa=minha.ld,pbif=NULL)
{
#-------------------------------------------------------------------------------#
# Funcao: SelVallBoot                                                           #
# funcao para teste bootstrap para dois grupos e m variáveis                    #
# calcula a matriz dos alfas por boot (nao entra o alfa zero)                   #
#  usa o bootstrap do S+ e o bootifgr(alteracao do boot s+2000 para grupos com  #
#  pesos                                                                        #
#                                                                               #
# INPUT: Dat: matriz de m variaveis (colunas) e n1+n2 linhas. As linhas         #
# estao ordenadas por grupo. Entra tambem vector de dimensao  k=(n1 n2)         #
# OUTPUT: p-values de todos os subconjuntos possiveis                           #
#                                                                               #
# Nota: facilmente modificavel para testar apenas uma hipotese (hipo)           #
#-------------------------------------------------------------------------------#
#now<-proc.time()
aa.gr<-NULL
m<-ncol(dat)
temp<-matrix(NA,B1,m)
prop<-NULL
n1<-k[1]
n2<-k[2]
n<-n1+n2
#prior1<-n1/n
#prior2<-n2/n
if (n!=nrow(dat)) stop("matriz dos dados não coincide com vector de dimensoes")
aa.gr<-rep(c(1:2),k)
if(is.null(pbif))
```

```
{temp<-bootstrap(data.frame(dat,aa.gr),B=B1,statistic=estalfa,
                        args.stat=list(k=k),group=aa.gr,trace=F)}
     else
    {temp<-bootifgr(data.frame(dat,aa.gr),B=B1,statistic=estalfa,
                                 args.stat=list(k=k),sampler=samp.boot.ifs2,
                                 args.sampler=pbif,group=aa.gr,trace=F,save.indices=F)}
pv <- NA
sbset <- Gerasubset(m)
lsb <- length(sbset)
for(i in 1:(lsb - 1)) {
sai <- Bootest(temp$replicate,B1, hipo = sbset[[i]])
pv[i] <- sai
}
mpv <- cbind(seq(1:(lsb - 1)), pv)
#ordmpv <- sort.col(mpv,c(1,2),columns.to.sort.by = c(2), F)
return(mpv)
}

Bootest<-function(tboo,B1,hipo)
{
pvalue<-NA
if (length(hipo)==1)
{
 Ahipo<-tboo[,hipo]
 mealfa<-mean(Ahipo)
 aux<-abs(Ahipo-mealfa)

 prop<-ifelse(aux>abs(mealfa),1,0)
 pvalue<-sum(prop)/B1
    }
else
{
 Ahipo<-as.matrix(tboo[,hipo])
 mealfa<-as.matrix(apply(Ahipo,2,mean))
 covalfa<-var(Ahipo)
 dh<-mahalanobis(Ahipo, center=mealfa, cov=covalfa)
 invcovalfa<-solve(covalfa)
 d0<-t(mealfa)%*%invcovalfa%*%mealfa
 prop<-ifelse(dh>d0,1,0)
 pvalue<-sum(prop)/B1
 }
 return(pvalue)
}

"Gerasubset"<-function(n)
{
```

```
# Funcao para gerar todos os subconjuntos de um conjunto de n elementos
# recorrendo a ordem lexicografica e conversao em binario

xx<-matrix(NA,(2^n-1),n)
for(i in 1:((2^n)-1))
{
aux<-bitpal(i)
if(length(aux)==n) xx[i,]<-aux
else xx[i,]<-c(rep(0,(n-length(aux))),aux)
}
yy<-seq(n,1)
xx<-xx[,yy]
return(apply(xx,1,saida))
}

"bitpal"<-function(n)
{
# funcao que converte um inteiro num binario
pal <- numeric(32)
i <- 0
while(n > 0) {
pal[32 - i] <- n %% 2
n <- n %/% 2
i <- i + 1
}
primeira <- match(1, pal)
pal[primeira:32]
}

saida<-function(xa)
{
# funcao para converter os 1 nos numeros inteiros do conjunto
xa<-as.vector(xa)
sa<-NULL
for(i in 1:length(xa))
{if(xa[i]==1) sa<-c(sa,i)}
      return(sa)
}


"minha.ld" <- function(dat, k, prior1 = 0.5, prior2=(1-prior1), boot = T)
{
#-------------------------------------------------------------------------------#
# Funcao: minha.ld                                                              #
# funcao para construcao da linear discriminante                                #
# para dois grupos e m variáveis                                                #
```

```
# INPUT: Dat: matriz de m variaveis (colunas) e n1+n2 linhas. As linhas        #
# estao ordenadas por grupo.                                                    #
#        k=(n1 n2): vector de dimensao de cada grupo                            #
#        boot:indicador logico para verificar se e para bootstrap ou nao        #
#             (=T nao calcula alfa0)                                            #
#                                                                               #
# OUTPUT:  vector de direccao alfa e constante de separação                     #
#-------------------------------------------------------------------------------#
#
# obtencao dos estimadores classicos de localizacao e escala para cada grupo
#
# tranformacao para entrar no boot que entra como dataframe sendo a ultima coluna
# a coluna com a distincao dos grupos
dat <- as.matrix(dat)
m <- ncol(dat)
if (boot) dat <- dat[,-m]
n1 <- k[1]
n2 <- k[2]
n <- n1 + n2
if(n != nrow(dat)) stop("matriz dos dados nao coincide com vector de dimensoes")
#
# estimacao da localizacao e escala de cada grupo
#
medg1 <- apply(dat[1:n1,  ], 2, mean)
covg1 <- var(dat[1:n1,  ])
medg2 <- apply(dat[(n1 + 1):(n),  ], 2, mean)
covg2 <- var(dat[(n1 + 1):(n),  ])
poolcov <- ((n1 - 1) * covg1 + (n2 - 1) * covg2)/(n -2)
poolinv <- solve(poolcov)
difmed <- (medg1 - medg2)
alfa <- poolinv %*% difmed
# normalizacao dos alfas
norma <- sqrt(sum(alfa^2))
alfa <- as.vector(alfa/norma)
if(!boot) {
somed <- (medg1 + medg2)
alfa0 <- as.numeric((t(alfa) %*% somed))
alfa0 <- (-0.5) * alfa0 - log(prior2/prior1)/norma
alfa <- c(alfa0, alfa)
}
    if (boot) return(alfa) else structure(list(alfa = alfa, prior = prior1), class = "ld")
}
```

C.4 Cálculo do estimador S de localização e dispersão

```
"SEST"<-function(data,bdp=0.5,Nsamp=600*(ncol(data)))
{
#-------------------------------------------------------------------------------#
# Calcula estimador S para localizacao e dispersao usando o algoritmo SURREAL     #
# (funcao biponderada de Tukey)                                                   #
#                                                                                 #
#     data:  matriz onde linhas=observacoes (n) e columas=variaveis (m)           #
#     bdp:   ponto de rotura pretendido para o estimador                          #
#     Nsamp: numero de subconjuntos aleatorios de dimensao (m+1)                  #
#            a considerar para a subamostragem                                    #
#-------------------------------------------------------------------------------#
now <- proc.time()
data<-as.matrix(data)
n<-nrow(data)
m<-ncol(data)
ma<-m+1
mb<-1/m
tol<-1e-5
Stil<-10e10
v1<-vcTukey(bdp,m)
c1<-v1$c1
kp<-v1$kp
l<-1
laux<-1
Nsing<-0
for (l in 1:Nsamp)
{
    J<-randomset(n,ma)
    xsetJ<-data[J,]
    muJ<-apply(xsetJ,2,mean)
    covJ<-var(xsetJ)
    if (!(IsSingular(covJ)))
      {
       covJ<-((meu.det(covJ))^(-mb))*covJ
       if (!(IsSingular(covJ)))
          {
           if (l > ceiling(Nsamp*0.2))
             {
              if (l==ceiling(Nsamp*(0.5))) laux<-2
              if (l==ceiling(Nsamp*(0.8))) laux<-4
           epsilon<-runif(1)
           epsilon<-epsilon^laux
           muJ<-epsilon*muJ+(1-epsilon)*mutil
```

```
           covJ<-epsilon*covJ+(1-epsilon)*covtil
}
                 covJ<-((meu.det(covJ))^(-mb))*covJ
            md<-sqrt(mahalanobis(data,muJ,covJ))
                if (mean(rhobiweight(md/Stil,c1))<kp)
             {
                if (Stil<5e10)  Stil<-scaleS(md,Stil,c1,kp,tol)
                else Stil<-scaleS(md,0,c1,kp,tol)
                mutil<-muJ
                covtil<-covJ
                psi<-psibiweight(md,c1*Stil)
                u<-psi/md
                ux<-diag(u)%*%data
                muJ<-apply(ux,2,mean)/mean(u)
                xcenter<-scale(data, center=muJ,scale=FALSE)
                covJ<-t(xcenter)%*%diag(u)%*%xcenter
                covJ<-((meu.det(covJ))^(-mb))*covJ
                 control<-0
                 cont<-1
                 while ((cont<3)&(control!=1))
                  {
                     cont<-cont+1
                       md<-sqrt(mahalanobis(data,muJ,covJ))
                       if ((mean(rhobiweight(md/Stil,c1)))<kp)
 {
                        mutil<-muJ
                        covtil<-covJ
                              control<-1
                              if (Stil==5e10) Stil<-scaleS(md,Stil,c1,kp,tol)
                              else Stil<-scaleS(md,0,c1,kp,tol)
                              }
                         else
   {
                            muJ<-(muJ+mutil)/2
                            covJ<-(covJ+covtil)/2
                            covJ<-((meu.det(covJ))^(-pb))*covJ
                          }
}
}
}
else {Nsing<-Nsing+1}
}
else {Nsing<-Nsing+1}
}
if(Nsing>0) cat("S estimator with",Nsamp,"subamostras das quais",Nsing," foram singulares \n")
covtil<-(Stil^2)*covtil
```

```
dmahaS<-mahalanobis(data,center=mutil,cov=covtil)
# hard-rejection
S.wt<-as.vector(ifelse(dmahaS < qchisq(0.975, m),1,0))
# If we want the reweighted cov and center then the results
# are obtain by cov.wt:
Sw<-cov.wt(data, wt=S.wt, cor=F, center=T)
Sw$cov <- sum(S.wt)/(sum(S.wt) - 1) * Sw$cov
Sw$cov <- as.matrix(Sw$cov)
if(IsSingular(Sw$cov)) {cat("A matriz de dispersao S é singular.\n") }
mah<-mahalanobis(data,center=Sw$center,cov=Sw$cov)
S.wt <-as.vector(ifelse(mah < qchisq(0.975, m),1,0))
print( proc.time() - now)
return(raw.center=mutil,raw.cov=covtil,S.wt=S.wt,center=Sw$center,
 cov=Sw$cov,Nsamp=Nsamp,Nsing=Nsing)
}


#-------------------------------------------------------------------------#
# Funcao: randomset                                                       #
# selecciona uma subamostra com dimensao ma e retorna os seus subindices #
#                                                                         #
#-------------------------------------------------------------------------#

randomset<-function(n,ma)
{
 aux<-NA
 xran<-runif(n)
 aux<-order(xran)
 J<-aux[1:ma]
}


#-----------------------------------------------------------------------#
# Funcao: Rhobiweight                                                   #
# Calcula a funcao rho biponderada de Tukey                             #
#                                                                       #
#-----------------------------------------------------------------------#

rhobiweight<-function(xx,c1)
{
vrho<-NULL
lessc1<-(xx^2)/2-(xx^4)/(2*(c1^2))+(xx^6)/(6*c1^4)
greatc1<-(c1^2)/6
vrho<-lessc1*abs(xx<c1)+greatc1*abs(xx>=c1)
}
```

```
#-----------------------------------------------------------------------#
# Funcao: psibiweight                                                   #
# funcao psi biponderada                                                #
#                                                                       #
#-----------------------------------------------------------------------#

psibiweight<-function(xx,c1)
{
vpsi<-NULL
vpsi<-(xx-(2*(xx^3))/(c1^2)+(xx^5)/(c1^4))*(abs(xx)<c1)
}



#-----------------------------------------------------------------------#
# Funcao: vcTukey                                                       #
# Calcula a constante de afinacao c1, para estimador S com fc.          #
# biponderada de Tukey dado bdp e m                                     #
#-----------------------------------------------------------------------#

vcTukey<-function(bdp,m)
{
cnew<-sqrt(qchisq(1-bdp, m))
maxit<-1000
epsilon<-10^(-8)
error<-10^6
contit<-1
while ((error>epsilon) & (contit<maxit))
{
cold<-cnew
kp<-vkpTukey(cold,m)
cnew<-sqrt(6*kp/bdp)
error<-abs(cold-cnew)
contit<-contit+1
}
if (contit==maxit) print("Maximo iteracoes")
return(c1=cnew,kp=kp)
}


#-----------------------------------------------------------------------#
# Funcao: vkpTukey                                                      #
# Calcula a constante kp para para estimador S com fc.                  #
# biponderada de Tukey segundo Proposicao 2.1 Cap 2                     #
#-----------------------------------------------------------------------#
```

```
vkpTukey<-function(c0,m)
{
   c2<-c0^2
   intg1<-(m/2)*pchisq(c2, (m+2))
   intg2<-(m+2)*m/(2*c2)*pchisq(c2, (m+4))
   intg3<-((m+4)*(m+2)*m)/(6*c2^2)*pchisq(c2, (m+6))
   intg4<-(c2/6)*(1-pchisq(c2, m))
   kp<-intg1-intg2+intg3+intg4
   return(kp)
}


#-----------------------------------------------------------------------#
#  Funcao: scaleS                                                       #
#  Calcula escala entrando as distancias dx                             #
#                                                                       #
#-----------------------------------------------------------------------#

scaleS<-function(dx,S0,c1,kp,tol)
{
if (S0>0) S<-S0
else S<-median(dx)/qnorm(3/4)
test<-2*tol
rhonew<-NA
rhoold<-NA
rhoold<-mean(rhobiweight(dx/S,c1))-kp
while (test>=tol)
 {
      delta<-rhoold/mean(psibiweight(dx/S,c1)*(dx/(S^2)))
      mmin<-1
      control<-0
      while ((mmin<10) & (control!=1))
          {
        rhonew<-mean(rhobiweight(dx/(S+delta),c1))-kp
           if (abs(rhonew) < abs(rhoold))
               {
                S<-S+delta
                control<-1
               }
            else
               {
         delta<-delta/2
                mmin<-mmin+1
                }
}
```

```
      if (mmin==10)
         test<-0
      else
         test<-(abs(rhoold)-abs(rhonew))/abs(rhonew)
      rhoold<-rhonew
     }
return(abs(S))
}

"IsSingular"<-function(matriz)
{
#-----------------------------------------------------------#
# Verifica se uma matriz pode ser singular ou nao           #
# Usada em S estimator                                      #
#-----------------------------------------------------------#

cont<-F
cont<-as.logical(cont)
m<-ncol(matriz)
if(!is.qr(matriz)) matriz <- qr(matriz)
if(matriz$rank < m) cont<-T
return(cont)
}

"meu.det"<-function(yy)
{
#-----------------------------------------------------------#
# Calcula o determinante de uma matriz                      #
# Usada em S estimator                                      #
#-----------------------------------------------------------#

library(Matrix)
mdet<-NA
if(!(length(yy) == 1)) {
if((is.data.frame(yy)) | (is.matrix(yy)))
yy <- as.Matrix(yy)
}
else (stop("ERRO:os dados nao sao matriz ou data.frame"))
aux<-det(yy,logarithm = F)
return((aux$sign*aux$modulus))
}
```

C.5 *Bootstrap* IFG para localização univariada

```
distbootw<-function(data,c1,d,B=5000,seed=1)
{
#-------------------------------------------------------------#
#  boot IF generalizado                                       #
#-------------------------------------------------------------#
   n<-length(data)
   aux<-correcmean(data,c1,d)
   cbias<-aux$cbias
   estw<-aux$est
   print(cbias)
   size<-aux$size
   estwo<-aux$estoriginal
   proboriginal<-aux$proboriginal
   prob<-aux$prob
   qua<- c(2.5,5,10,25,50,75,90,95,97.5)
   qua<-qua/100
   distbwo<-rep(0,B)
   distbw<-rep(0,B)
   distefron<-rep(0,B)
   for (i in 1:B){
   xbefron<-sample(data,n,replace=TRUE)
         xboriginal<-sample(data,n,replace=TRUE,prob=proboriginal)
         xb<-sample(data,floor(size),replace=TRUE,prob=prob)
         distefron[i]<-mean(xbefron)
         distbwo[i]<-mean(xboriginal)
         distbw[i]<-mean(xb)
   }
   estefron<-mean(distefron)
   estbootwo<-mean(distbwo)
   estbootw<-mean(distbw)
   varefron<-var(distefron)
   varbootwo<-var(distbwo)
   varbootw<-var(distbw)
   qdefron<-quantile(distefron, probs =qua ,na.rm = F)
   qdefronc<-quantile(distefron-estefron, probs =qua ,na.rm = F)
   qdo<-quantile(distbwo, probs =qua ,na.rm = F)
   qdocentrado<-quantile(distbwo-estw, probs =qua ,na.rm = F)
   qdbw<-quantile(distbw, probs =qua ,na.rm = F)
   print(qdbw)
   distbw1<-distbw-estw
   qdbw1<-quantile(distbw1, probs =qua ,na.rm = F)
   distbw2<-distbw1+cbias
   qdbw2<-quantile(distbw2, probs =qua ,na.rm = F)
```

```
   vv<-var(distbw1)
   cc<- sqrt((vv+25*cbias**2/n)/vv)
   distbw3<-(distbw1-mean(distbw1))*cc + mean(distbw1)
   qdbw3<-quantile(distbw3, probs =qua ,na.rm = F)
   qbmatriz<-cbind(qdefron,qdefronc,qdo,qdocentrado,qdbw,qdbw1,qdbw2,qdbw3)
   estpontual<-c(estefron,estwo,estw,estbootwo,estbootw)
   varboot<-c(varefron,varbootwo,varbootw)
   boxplot(distefron,distbwo,distbw,style.bxp="old")
   boxplot((distefron-estefron),(distbwo-estw),distbw1, distbw2,distbw3,style.bxp="old")
   return(qbmatriz,estpontual,varboot)
}

correcmean<- function(dat,c1,d)
{
#-------------------------------------------------#
# boot IF generalizado, correcao para a media     #
#-------------------------------------------------#
   n<-length(x)
   m0<-median(dat)
   s0<-mad(dat)
   u<-abs(dat-m0)/s0
   pesoorig<-sqrt(ifelse(u<c1,1,dnorm(u,c1,d)*d*sqrt(2*pi)))
   somapeso<-sum(pesoorig)
   proborig<-pesoorig/somapeso
   estorig<- sum(pesoorig*dat)/somapeso
   rpesos<-sqrt(ifelse(u<c1,1,dnorm(u,c1,d)*d*sqrt(2*pi)))
   m1<-sum(rpesos*dat)/sum(rpesos)
   u<-abs(dat-m1)/s0
   rpesos<-sqrt(ifelse(u<c1,1,dnorm(u,c1,d)*d*sqrt(2*pi)))
   m2<-sum(rpesos*dat)/sum(rpesos)
   u<-abs(dat-m2)/s0
   rpesos<-sqrt(ifelse(u<c1,1,dnorm(u,c1,d)*d*sqrt(2*pi)))
   m3<-sum(rpesos*dat)/sum(rpesos)
   u<-(dat-m3)/s0
   rpesos<-sqrt(ifelse(abs(u)<c1,1,dnorm(abs(u),c1,d)*d*sqrt(2*pi)))
   pesos<-rpesos**2
   drpesos<-ifelse(abs(u)<c1,0,-sign(u)*(abs(u)-c1)/(2*d**2)*rpesos)
   est<- sum(rpesos*dat)/sum(rpesos)
   cbias<- est-sum(pesos*dat)/sum(pesos)
   size<- (sum(rpesos)+sum(drpesos*u))**2/sum(pesos)
   prob<-pesos/sum(pesos)
   list(est=est,cbias=cbias,size=size,prob=prob,proboriginal=proborig,estoriginal=estorig)
}
```

BIBLIOGRAFIA

Agulló, J. (1996). Exact iterative computation of multivariate Minimum Volume Elipsoid estimator with a Branch and Bound Algorithm. In A. Prat (Ed.), *Proceedings in Computational Statistics*, Heidelberg, pp. 175–180. Physica-Verlag.

Ahmed, S. W. e Lachenbruch, P. A. (1977). Discriminant analysis when scale contamination is present in the initial sample. In J. V. Rysin (Ed.), *Classification and Clustering*, New York, pp. 331–353. Academic Press.

Alemayeha, D. (1987). Bootstrap method for multivariate analysis. In *Proceedings of Statistical Computations*, pp. 321–324. American Statistical Association.

Amado, C. e Pires, A. M. (1998). Uso de técnicas de reamostragem para selecção de variáveis em análise discriminante. In M. S. de Miranda e I. Pereira (Eds.), *Estatística: a Diversidade na Unidade*, Novas Tecnologias/Estatística, Lisboa, pp. 247–253. Edições Salamandra.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (2 ed.). New York: Wiley.

Ankembrand, T. e Tomassini, M. (1996). Predicting multivariate financial time series: the Swiss Bond Case. In *Proceedings of the IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering*, New York, pp. 27–33. IEEE.

Athreya, K. B. (1987). Bootstrap of the mean in the infinite variance case. *The Annals of Statistics* **15**, 724–731.

Babu, G. J. e Singh, K. (1983). Inference on means using the bootstrap. *The Annals of Statistics* **11**, 999–1003.

Babu, G. J. (1989). Applications of edgeworth expansions to bootstrap. In Y. Dodge (Ed.), *Statistical Data Analysis and Inference*, Amsterdam, pp. 223–237. Elsevier.

Banga, C. e Ghorbel, F. (1993). Optimal bootstrap sampling for fast image segmentation: application to retina image. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP-93)*, Minneapolis, USA.

Barnett, V. e Lewis, T. (1994). *Outliers in Statistical Data* (3 ed.). New York: Wiley.

Becker, C. e Gather, U. (2001). The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules. *Computational Statistics and Data Analysis* **36**, 119–127.

Beckman, R. J. e Cook, R. D. (1983). Outlier.....s (with comments and rejoinder). *Technometrics* **25**, 119–163.

Beran, R. e Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of covariance matrix. *The Annals of Statistics* **13**, 95–115.

Bhar, R. e Chiarella, C. (1996). Bootstrap algorithms and applications. In *Proceedings of the IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering*, New York, pp. 168–182. IEEE.

Bickel, P. J. e Freedman, D. A. (1981). Some asymptotics on the bootstrap. *The Annals of Statistics* **9**, 1196–1217.

Boente, G., Pires, A. M. e Rodrigues, I. M. (2002). Influence functions and outlier detection under the common principal components model: a robust approach. *Biometrika* **89**, 861–875.

Booth, J. G. e Hall, P. (1993). An improvement of the jackknife distribution function estimator. *The Annals of Statistics* **21**, 1467–1485.

Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions. *The Annals of Statistics* **16**, 1709–1722.

Box, G. E. P. (1953). Non–normality and tests on variances. *Biometrika* **40**, 318–335.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review* **78**, 1–3.

Burns, P. (1992). A genetic algorithm for robust regression estimation. StatSci Technical Note, Seattle, WA.

Butler, R. W., Davies, P. L. e Jhun, M. (1993). Asymptotics for the Minimum Covariance Determinant estimator. *The Annals of Statistics* **21**, 1385–1400.

Cadima, J., Cerdeira, J. O. e Minhoto, M. (2002). Algoritmos genético e de melhoramento local para a selecção de subconjuntos de variáveis. In L. Carvalho, F. Brilhante, e F. Rosado (Eds.), *Novos Rumos em Estatística*, Actas do IX Congresso Anual da Sociedade Portuguesa de Estatística, Lisboa, pp. 147–156. Edições SPE.

Campbell, N. A. (1978). The influence function as an aid in outlier detection in discriminant analysis. *Applied Statistics* **27**, 251–258.

Campbell, N. A. (1982). Robust procedures in Multivariate Analysis: robust Canonical Variate Analysis. *Applied Statistics* **31**, 1–8.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics* **14**, 1171–1179.

Chaloner, K. e Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika* **75**, 651–659.

Chatterjee, S. K. e Samanta, S. K. (1997). Variable selection for discrimination among several populations. *Communications in Statistics–Theory and Methods* **26**, 2565–2582.

Chatterjee, S. K. (1983). Estimation of misclassification probabilities by bootstrap methods. *Communications in Statistics B* **12**, 645–656.

Chatterjee, S. K. (1984). Variance estimation in factor analysis, an application of bootstrap. *British Journal of Mathematical and Statistical Psychology* **37**, 252–262.

Chen, Z. e Tyler, D. E. (2002). The influence function of Tukey's deepest depth median. *The Annals of Statistics* **30**, 1737–1759.

Ciarlini, P. (1997). Bootstrap Algorithms and Applications. In P. Ciarlini, M. Cox, F. Pavese, e D. Richter (Eds.), *Advanced Mathematical Tools in Metrology III*, Singapore, pp. 12–23. World Scientific.

Conover, W. J. (1980). *Practical Nonparametric Statistics.* New York: Wiley.

Constable, C. e Tauxe, L. (1990). The bootstrap for magnetic susceptibility tensors. *Journal of Geophysical Research* **90**, 8383–8395.

Constable, C. e Tauxe, L. (1992). Integrating a high-breakdown option into discriminant analysis in exploration geochemistry. *Journal of Geochemical Exploration* **43**, 191–203.

Cook, R. D. e Weisberg, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

Cox, M. G., Harris, P. M., Milton, M. J. T. e Woods, P. T. (1997). A method for evaluating trends in ozone-concentration data and its applications to data from the UK Rural Ozone Monitoring network. In P. Ciarlini, M. Cox, F. Pavese, e D. Richter (Eds.), *Advanced Mathematical Tools in Metrology III*, Singapore, pp. 171–177. World Scientific.

Critchley, F. e Vitiello, C. (1991). The influence of observations on misclassification probability estimates in linear discriminant analysis. *Biometrika* **78**, 677–690.

Critchley, F. (1985). Influence in principal components analysis. *Biometrika* **72**, 627–636.

Croux, C. e Dehon, C. (2001). Robust linear discriminant analysis using S-estimators. *The Canadian Journal of Statistics* **29**, 473–492.

Croux, C. e Haesbroeck, G. (1999). Influence function and efficiency of the Minimum Covariance Determinant scatter matrix estimator. *Journal of Multivariate Analysis* **71**, 161–190.

Croux, C., Rousseeuw, P. J. e Van Bael, A. (1996). Positive-breakdown regression by minimizing nested scale estimators. *Journal of Statistical Planning and Inference* **53**, 197–235.

Davies, L. e Gather, U. (1993). The identification of multiple outliers (with discussion). *Journal of the American Statistical Association* **88**, 782–801.

Davies, P. L. (1987). Asymptotic behaviour of S–estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics* **15**, 1269–1292.

Davies, P. L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid. *The Annals of Statistics* **20**, 1828–1843.

Davison, A. C. e Hinkley, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika* **75**, 417–431.

Davison, A. C. e Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge: Cambridge University Press.

Davison, A. C. (1988). Discussão dos artigos de D. V. Hinkley, T. J. Diciccio e J. P. Romano. *Journal of the Royal Statistical Society B* **50**, 356–357.

DeAngelis, D., Hall, P. e Young, G. A. (1993). Analytical and bootstrap approximations to estimator distributions in L1 regression. *Journal of the American Statistical Association* **88**, 1310–1316.

Devlin, S. J., Gnanadesikan, R. e Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62**, 531–545.

Diaconis, P. e Efron, B. (1983). Statistics on the computer. The bootstrap. *Scientific American* **5**, 116–130.

DiCiccio, T. J. e Romano, J. P. (1988). A review of bootstrap confidence intervals (with discussion). *Journal of the Royal Statistical Society B* **50**, 338–354.

Dolker, M., Halperin, S. e Divgi, D. R. (1982). Problems with bootstrapping Pearson correlations in very small bivariate samples. *Psychometrika* **47**, 529–530.

Donoho, D. L. e Huber, P. J. (1983). The notion of breakdown point. In P. Bickel, K. Doksum, e J. Hodges (Eds.), *A Festschrift for Erich Lehmann*, Belmont, California, pp. 157–184. Wadsworth.

Duarte Silva, A. P. (1998). A "Leaps and Bounds" algorithm for variable selection in two-group discriminant analysis. In A. Rizzi, M. Vichi, e H. Bock (Eds.), *Advances in Data Science and Classification*, Number 26 in Lecture Notes in Statistics, Rome, pp. 227–232. Springer–Verlag.

Duarte Silva, A. P. (2001). Efficient variable screening for multivariate analysis. *Journal of Multivariate Analysis* **76**, 35–62.

Dupuis, D. J. e Hamilton, D. C. (2000). Regression residuals and test statistics: assessing naive outlier deletion. *The Canadian Journal of Statistics* **28**, 259–275.

Efron, B. e Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross validation. *The American Statistician* **37**, 36–48.

Efron, B. e Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**, 54–77.

Efron, B. e Tibshirani, R. (1993). *An Introduction to the Bootstrap.* London: Chapman and Hall.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics* **7**, 1–26.

Efron, B. (1981a). Censored data and the bootstrap. *Journal of the American Statistical Association* **76**, 312–319.

Efron, B. (1981b). Nonparametric standard errors and confidence intervals (with discussion). *The Canadian Journal of Statistics* **9**, 139–172.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* Philadelphia: SIAM (Cbms-Nsf Regional Conference Series in Applied Mathematics : 38).

Efron, B. (1983). Estimating the error rate of a prediction rule: improvements on cross-validation. *Journal of the American Statistical Association* **78**, 316–331.

Efron, B. (1987). Better bootstrap confidence intervals (with comments and rejoinder). *Journal of the American Statistical Association* **82**, 171–200.

Efron, B. (1988). Computer-intensive methods in statistical regression. *SIAM Review* **30**, 421–449.

Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society B* **54**, 83–127.

Efron, B. (1994). Missing data, imputation, and the bootstrap (with discussion). *Journal of the American Statistical Association* **89**, 463–479.

Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*, Volume 19 of *Lecture Notes in Statistics.* New York: Springer–Verlag.

Fisher, N. I. e Hall, P. (1989). Bootstrap confidence regions for directional data. *Journal of the American Statistical Association* **89**, 996–1002.

Fisher, N. I. e Hall, P. (1991). Bootstrap algorithms for small samples. *Journal of Statistical Planning and Inference* **27**, 157–169.

Fisher, R. A. (1935). *The Design of Experiments, 1 ed.* Edinburgh: Oliver and Boyd.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

Fox, T., Hinkley, D. V. e Larntz, K. (1980). Jackknifing in nonlinear regression. *Technometrics* **22**, 123–129.

Fraiman, R., Yohai, V. e Zamar, R. (2001). Optimal robust M-estimates of location. *The Annals of Statistics* **29**, 194–223.

Freedman, D. A. e Peters, S. C. (1984). Bootstrapping a regression equation: some empirical results. *Journal of the American Statistical Association* **79**, 97–106.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics* **9**, 1218–1228.

Freedman, D. A. (1984). On bootstrapping two-stages least-squares estimates in stationary linear models. *The Annals of Statistics* **12**, 827–842.

Fung, W. K. (1996). The influence of observations on misclassification probability in multiple discriminant analysis. *Communications in Statistics-Theory and Methods* **25**, 1917–1930.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association* **70**, 320–328.

Gervini, D. (2002). The influence function of the Stahel–Donoho estimator of multivariate location and scatter. *Statistics and Probability Letters* **60**, 425–435.

Ghosh, M., Parr, W. C., Singh, K. e Babu, G. J. (1984). A note on bootstrapping the sample median. *The Annals of Statistics* **12**, 1130–1135.

Gnanadesikan, R. e Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **28**, 81–124.

Gong, G. (1986). Cross-validation, the jackknife, and the bootstrap. Excess error estimation in forward logistic regression. *Journal of the American Statistical Association* **81**, 108–113.

Gross, A. M. (1976). Confidence interval robustness with long-tailed symmetric distributions. *Journal of the American Statistical Association* **71**, 409–416.

Hall, P., Horowitz, J. L. e Jing, B. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika* **82**, 561–574.

Hall, P. (1985). Resampling a coverage process. *Stochastic Processes and their Applications* **20**, 231–246.

Hall, P. (1988). Theorical comparison of the bootstrap confidence intervals (with discussion). *The Annals of Statistics* **16**, 927–985.

Hall, P. (1992a). *The Bootstrap and Edgeworth Expansion.* New York: Springer–Verlag.

Hall, P. (1992b). Efficient bootstrap simulation. In R. LePage e L. Billard (Eds.), *Exploring the Limits of Bootstrap*, New York, pp. 127–143. Wiley.

Hammersley, J. M. e Handscomb, D. C. (1964). *Monte Carlo Methods.* London: Methuen.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. e Stahel, W. J. (1986). *Robust Statistics: the Approach based on Influence Functions.* New York: Wiley.

Hampel, F. R. (1968). *Contributions to the Theory of Robust Estimation.* Ph. D. thesis, University of California, Berkley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393.

Hampel, F. R. (1985). The breakdown points of the mean combined with some rejection rules. *Technometrics* **27**, 95–107.

Hampel, F. R. (2000). Robust inference. Research Report 93, ETH, Seminar für Statistik, ETH, Zürich. A aparecer em Encyclopedia of Environmetrics, Wiley.

Hand, D. J. (1981). *Discrimination and Classification.* New York: Wiley.

Hand, D. J. (1997). *Construction and Assessment of Classification rules.* New York: Wiley.

Hanna, S. R. (1989). Confidence limits for air quality model evaluations, as estimated by bootstrap and jackknife resampling methods. *Atmospheric Environment* **23**, 1385–1398.

Haynor, D. R. e Woods, S. D. (1989). Resampling estimates of precision in emission tomography. *IEEE Transactions on Medical Imaging* **8**, 337–343.

Herman, R. e Heunis, S. W. (1999). Assessment of the Monte Carlo simulation method for probabilistic feeder voltage drop calculations. *Transactions on South African Institute of Electrical Engineers (SAIEE))* **90**, 126–130.

He, X. e Fung, W. K. (2000). High breakdown estimation for multivariate populations with applications to discriminant analysis. *Journal of Multivariate Analysis* **72**, 151–162.

He, X., Simpson, D. G. e Portnoy, L. (1990). Breakdown robustness of tests. *The Annals of Statistics* **85**, 446–452.

He, X. e Simpson, D. G. (2000). Robust direction estimation. *The Annals of Statistics* **20**, 351–369.

He, X. e Wang, G. (1996). A qualitative robustness of S* estimators of multivariate location and dispersion. *Statistica Neerlandica* **51**, 257–268.

Hinkley, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics* **19**, 285–292.

Hjorth, J. S. U. (1994). *Computer Intensive Statistical Methods: Validation, Model Selection, and Bootstrap.* London: Chapman and Hall.

Hjort, N. L. (1992). On inference in parametric survival data models. *International Statistical Review* **60**, 355–387.

Hoaglin, D. C., Mosteller, F. e Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis.* New York: Wiley.

Hoaglin, D. C., Mosteller, F. e Tukey, J. W. (1985). *Exploring Data Tables, Trends and Shapes.* New York: Wiley.

Hoaglin, D. C., Mosteller, F. e Tukey, J. W. (1992). *Análise Exploratória de Dados. Técnicas Robustas. Um Guia.* Lisboa: Edições Salamandra. Tradução portuguesa de: Hoaglin, D. C., Mosteller, F. e Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis* . New York: Wiley.

Hogg, R. V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory (with comments and rejoinder). *Journal of the American Statistical Association* **69**, 909–927.

Huberty, C. J. (1994). *Applied Discriminant Analysis.* New York: Wiley.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73–101.

Huber, P. J. (1981). *Robust Statistics.* New York: Wiley.

Johns, M. V. (1988). Importance sampling for bootstrap confidence intervals. *Journal of the American Statistical Association* **83**, 709–714.

Jorgensen, M. A. (1993). Influence functions for iteratively defined statistics. *Biometrika* **80**, 253–265.

Katatani, K. e Ohta, N. (1998). Optimal robot self-localization and reliability evaluation. In *Proceedings of the 5th European Conference on Computer Vision*, Berlin, pp. 796–808. Springer–Verlag.

Kempthorne, O. (1954). *The Design and Analysis of Experiments.* New York: Wiley.

Krolik, J., Niezgoda, G. e Swingler, D. N. (1991). A bootstrap approach for evaluating source localization performance on real sensor array data. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, Toronto, pp. 1281–1284.

Kuhle, J. L. e Moorehead, J. D. (1990). Applying the bootstrap technique to real estate appraisal: An empirical analysis. *Journal of Real Estate Research* **5**, 33–40.

Lachenbruch, P. A. e Mickey, M. R. (1968). Estimation of error rate in discriminant analysis. *Technometrics* **10**, 1–11.

Lachenbruch, P. A. (1975). *Discriminant Analysis.* New York: Hafner Press.

Lambert, D. (1982). Qualitative robustness of tests. *Journal of the American Statistical Association* **77**, 352–357.

Lange, J. M., Voigt, H.-M., Burkhardt, S. e Göbel, R. (1998). The intelligent inspection engine-a real-time real-world visual classifier system. In *IEEE International Joint Conference on Neural Networks Proceedings*, New York, pp. 1810–1815.

Lele, S. (1991). Jackknifing linear estimating equations: asymptotic theory and applications in stochastic processes. *Journal of the Royal Statistical Society B* **52**, 253–267.

Liu, R. Y. e Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In R. LePage e L. Billard (Eds.), *Exploring the Limits of Bootstrap*, New York, pp. 225–248. Wiley.

Liu, R. Y. e Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association* **88**, 252–260.

Lopuhaä, H. P. e Rousseeuw, P. J. (1991). Breakdown properties of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics* **19**, 229–248.

Lopuhaä, H. P. (1989). On the relation between S–estimators and M–estimators of multivariate location and covariance. *The Annals of Statistics* **17**, 1662–1683.

Lopuhaä, H. P. (1998). Asymptotics of reweighted estimators of multivariate location and scatter. Technical report, Department of Technical Mathematics and Informatics, TUDelft.

Lopuhaä, H. P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *The Annals of Statistics* **5**, 1638–1665.

Lu, J., Ko, D. e Chang, T. (1997). The standardized influence matrix and its applications. *Journal of the American Statistical Association* **92**, 1572–1580.

Maitra, R. (1997). Estimating precision in functional images. *Journal of Computational and Graphical Statistics* **6**, 132–142.

Maiwald, D. e Böhme, J. F. (1994). Multiple testing for seismic data using bootstrap. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP-94)*, Adelaide.

Manly, B. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology.* London: Chapman and Hall.

Markatou, M. e Hettmansperger, T. P. (1990). Robust bounded influence tests in linear models. *Journal of the American Statistical Association* **85**, 187–190.

Markatou, M., Stahel, W. A. e Ronchetti, E. (1991). Robust M–type testing procedures for linear models. In W. Stahel e S. Weisberg (Eds.), *Directions in Robust Statistics and Diagnostics (Part I)*, New York, pp. 201–220. Springer–Verlag.

Maronna, R. A., Bustos, O. H. e Yohai, V. J. (1979). Bias and efficiency robustness of general M–estimators for regression with random carriers. In T. Gasser e M. Rosenblatt (Eds.), *Smoothing Techniques for Curve Estimation*, Lecture Notes in Mathematics 757, New-York, pp. 91–116. Springer–Verlag.

Maronna, R. A. e Yohai, V. J. (1995). The behaviour of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association* **90**, 330–341.

Maronna, R. A. e Zamar, R. H. (2002). Robust estimates of location and dispersion for high dimensional data sets. *Technometrics* **44**, 307–317.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics* **4**, 51–67.

Martin, R. D., Yohai, V. J. e Zamar, R. H. (1989). Minimax bias robust regression. *The Annals of Statistics* **17**, 1608–1630.

McKay, R. J. e Campbell, N. A. (1982b). Variable selection techniques in discriminant analysis II. Allocation. *British Journal of Mathematical and Statistical Psychology* **35**, 30–41.

McKay, R. J. e Campbell, N. A. (1982a). Variable selection techniques in discriminant analysis I. Description. *British Journal of Mathematical and Statistical Psychology* **35**, 1–29.

McLachlan, G. J. (1980). The efficiency of Efron's bootstrap approach applied to error rate estimation in discriminant analysis. *Journal of Statistical Computation and Simulation* **10**, 273–279.

McLachlan, G. J. (1987). Error rate estimation in discriminant analysis: recent advances. In A. K. Gupta (Ed.), *Advances in Multivariate Statistical Analysis*, Reidel, pp. 233–252. Dordrecht.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley.

Miller, R. G. (1974). An unbalanced jackknife. *The Annals of Statistics* **2**, 880–891.

Peck, R., Fisher, L. e Van Ness, J. (1989). Bootstrap confidence intervals for the numbers of clusters in the cluster analysis. *Journal of the American Statistical Association* **84**, 184–191.

Peterson, I. (1991). Pick a sample. *Science News* **140**, 56–58.

Pires, A. M. e Branco, J. A. (2002). Partial influence functions. *Journal of Multivariate Analysis* **83**, 451–468.

Pires, A. M. (1990). Estimação Robusta e a sua Aplicação a Componentes Principais. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

Pires, A. M. (1995). *Análise Discriminante: Novos Métodos Robustos de Estimação.* Ph. D. thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

Pison, G., Rousseeuw, P. J., Filzmoser, P. e Croux, C. (2000). A robust version of principal factor analysis. In J. Bethlehem e P. Van der Heijden (Eds.), *Proceedings in Computational Statistics, Compstat 2000*, Heidelberg, pp. 385–390. Physica-Verlag.

Politis, D. N. e Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics* **22**, 2031–2050.

Posse, C. (1992). Projection pursuit discriminant analysis for two groups. *Communications in Statistics - Theory Methods* **21**, 1–19.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society B* **11**, 18–84.

Quenouille, M. (1956). Notes on bias in estimation. *Biometrika* **43**, 353–360.

Radhakrishnan, R. e Kshirsagar, M. (1981). Influence functions for certain parameters in multivariate analysis. *Communications in Statistics - Theory and Methods* **10**, 515–529.

Randles, R. H., Broffit, J. D., Ramberg, J. S. e Hogg, R. V. (1978). Generalized linear and quadratic discriminant functions using robust estimates. *Journal of the American Statistical Association* **73**, 564–568.

Reid, D. C., Zoubir, A. M. e Boashash, B. (1996). The bootstrap applied to passive acoustic aircraft parameter estimation. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, Atlanta.

Rieder, H. (1978). A robust asymptotic testing model. *The Annals of Statistics* **6**, 1080–1094.

Rocke, D. M. e Woodruff, D. L. (1997). Robust estimation of multivariate location and shape. *Journal of Statistical Planning and Inference* **57**, 245–255.

Ronchetti, E. (1982). *Robust Testing in Linear Models: the Infinitesimal Approach.* Ph. D. thesis, ETH, Zurich, Switzerland.

Ronchetti, E. (1997). Robust inference by influence functions. *Journal of Statistical Planning and Inference* **57**, 59–72.

Rosner, B. (1977). Percentage points for the RST many outlier procedure. *Technometrics* **19**, 307–312.

Rousseeuw, P. J. e Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* New York: Wiley.

Rousseeuw, P. J. e Ronchetti, E. (1981). Influence curve for general statistics. *Journal of Computational and Applied Mathematics* **7**, 161–166.

Rousseeuw, P. J. e Van Driessen, K. (1999). A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics* **41**, 212–223.

Rousseeuw, P. J. e Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points (including three invited comments and the rejoinder). *Journal of the American Statistical Association* **85**, 633–651.

Rousseeuw, P. J. e Yohai, V. J. (1984). Robust regression by means of S–estimators. In J. Franke (Ed.), *Robust and Nonlinear Time Series Analysis*, Number 26 in Lecture Notes in Statistics, New York, pp. 256–272. Springer–Verlag.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association* **79**, 871–881.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, e W. Wertz (Eds.), *Mathematical Statistics and Applications*, Reidel, pp. 283–297. Dordrecht.

Ruppert, D. (1992). Computing S estimators for regression and multivariate location/dispersion. *Journal of Computational and Graphical Statistics* **1**, 253–270.

Salibian-Barrera, M. e Zamar, R. H. (2002a). Bootstrapping robust estimates of regression. *The Annals of Statistics* **30**, 556–582.

Salibian-Barrera, M. e Zamar, R. H. (2002b). Fast and stable bootstrap methods for estimates defined by estimating equations. Technical report, `http://mathstat.carleton.ca/~matias/pubs.html`.

Salibian-Barrera, M. (2000). *Contributions to the Theory of Robust Inference.* Ph. D. thesis, University of British Columbia, Vancouver, Canadá.

Salibian-Barrera, M. (2002). Fast and stable bootstrap methods for robust estimates. In *Proceedings of the of the 2002 Annual Meeting of the Interface Foundation*, Montreal, QC.

Schenker, N. (1985). Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association* **80**, 360–361.

Seber, G. A. F. (1984). *Multivariate Observations.* New York: Wiley.

Sen, P. K. (1982). On M–tests in linear models. *Biometrika* **69**, 245–248.

Sequeira, A. S. (1997). Tratamento estatístico de imagens de satélite. IST, Departamento de Matemática. Trabalho Final de Curso.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

Shao, J. e Tu, D. (1995). *The Jackknife and Bootstrap.* New York: Springer–Verlag.

Shao, J. (1989). The efficiency and consistency of approximations to the jackknife variance estimators. *Journal of the American Statistical Association* **84**, 114–119.

Shao, J. (1990). Bootstrap estimation of the asymptotic variances of statistical functionals. *Annals of the Institute of Statistical Mathematics* **42**, 737–752.

Shao, J. (1992). Bootstrap variance estimators with truncation. *Statistics and Probability Letters* **15**, 95–101.

Shi, X., Wu, C. F. J. e Chen, J. (1990). Weak and strong representations for quantile processes from finite populations with application to simulation size in resampling inference. *The Canadian Journal of Statistics* **18**, 141–148.

Shi, X. (1984). The approximate independence of jackknife pseudo-values and the bootstrap methods. *Journal Wuhan Inst. Hydra. Elect. Engin.* **2**, 83–90.

Shi, X. (1991). Some asymptotic results for jackknifing the sample quantile. *The Annals of Statistics* **19**, 496–503.

Shorack, G. R. (1982). Bootstrapping robust regression. *Communications in Statistics–Theory and Methods* **11**, 961–972.

Silverman, B. e Young, G. A. (1987). The bootstrap: to smooth or not to smooth? *Biometrika* **74**, 469–479.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics* **9**, 1187–1195.

Singh, K. (1998). Breakdown theory for bootstrap quantiles. *The Annals of Statistics* **26**, 1719–1732.

Srivastava, M. S. e Chan, Y. M. (1989). A comparison of bootstrap method and Edgeworth expansion in approximating the distribution of sample variance-one sample and two sample cases. *Communications in Statistics B* **18**, 339–361.

Staudte, R. G. e Sheather, S. J. (1990). *Robust Estimation and Testing.* New York: Wiley.

Stone, M. (1974). Cross validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B* **36**, 111–147.

Stromberg, A. J. (1997). Robust covariance estimates based on resampling. *Journal of Statistical Planning and Inference* **57**, 321–334.

S-PLUS (1999). *S-PLUS 2000 Guide to Statistics, Volume 2.* Data Analysis Products Division. Seattle, WA: MathSoft.

Takeuchi, K. (1971). A uniformly asymptotically efficient estimator of a location parameter. *Journal of the American Statistical Association* **66**, 292–301.

Tauxe, L., Kylstra, N. e Constable, C. (1991). Bootstrap statistics for paleomagnetic data. *Journal of Geophysical Research* **91**, 11723–11740.

Thorburn, D. E. (1977). On the asymptotic normality of the jackknife. *The Scandinavian Journal of Statistics* **4**, 113–118.

Todorov, V. K., Neykov, N. M. e Neytchev, P. N. (1990). Robust selection of variables in discrimination analysis based on Minimum Volume Ellipsoid and MCD estimators. In *Proceedings of the Conference CompStat 1990*, Heidelberg, pp. 193–198. Physica-Verlag.

Todorov, V. K., Neykov, N. M. e Neytchev, P. N. (1994). Robust two-group discrimination by bounded influence regression. A Monte-Carlo simulation. *Computational Statistics and Data Analysis* **17**, 289–302.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematics and Statistics* **29**, 614.

Tukey, J. W. (1970). *Exploratory Data Analysis* (1970-71: preliminary ed.). Reading, Massachusetts: Addison-Wesley.

Weiss, R. E. (1996). An approach to Bayesian sensitivity analysis. *Journal of the Royal Statistical Society B* **58**, 739–750.

Willems, G., Pison, G., Rousseeuw, P. e Van Aelst, S. (2002). A robust Hotelling test. *Metrika* **55**, 125–138.

Woodruff, D. L. e Rocke, D. M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association* **89**, 888–896.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *The Annals of Statistics* **14**, 1261–1350.

Wu, C. F. J. (1990). On the asymptotic properties of the jackknife histogram. *The Annals of Statistics* **18**, 1438–1452.

Yohai, V., Stahel, W. A. e Zamar, R. H. (1991). A procedure for robust estimation and inference in linear regression. In W. Stahel e S. Weisberg (Eds.), *Directions in Robust statistics and diagnostics, Part II*, pp. 365–374. Springer–Verlag.

Young, G. A. (1988). A note on bootstrapping the correlation coefficient. *Biometrika* **75**, 370–373.

Young, G. A. (1994). Bootstrap: more than a stab in the dark? (with discussion and rejoinder). *Statistical Science* **9**, 382–415.

Zhang, L. e Boos, D. D. (1992). Bootstrap critical values for testing homogeneity of covariance matrices. *Journal of the American Statistical Association* **87**, 425–429.

Zuo, Y., Cui, H. e He, X. (2002). On the Stahel-Donoho estimator and depth-weighted means of multivariate data. *The Annals of Statistics (to appear)*.