Robust Inference using the Extended Bregman Divergence and Optimal Tuning Parameter Selection

A thesis submitted in fulfillment of the requirements

for the degree of Ph.D. in Statistics

Thesis Author : Sancharee Basak

Thesis Supervisor : Dr. Ayanendranath Basu

INTERDISCIPLINARY STATISTICAL RESEARCH UNIT

INDIAN STATISTICAL INSTITUTE, KOLKATA

July 2022

Certificate

It is certified that the work contained in this thesis entitled 'Robust Inference using the Extended Bregman Divergence and Optimal Tuning Parameter Selection' by 'Sancharee Basak' has been carried out under my supervision and that it has not been submitted elsewhere for a degree.

Dr. Ayanendranath Basu

July 2022

Professor

Interdisciplinary Statistical Research Unit

Indian Statistical Institute, Kolkata

Acknowledgements

First and foremost, I extend my sincerest gratitude to my esteemed Ph.D. thesis supervisor and senior professor at the Indian Statistical Institute (ISI), Prof. Ayanendranath Basu. I would like to sincerely thank him for his indispensable advice, unwavering support and immense patience that he showed towards me during my Ph.D. journey. Without his able guidance and profound knowledge, it would have been impossible for me to carry out my doctoral research on robust statistical inference. His attention to every detail during the composition of my thesis and publications and his recommended references (books, published papers, etc.) have proven to be important components in the development of my writings, which are free from flaws. I am really indebted to him for letting me work as a research scholar under his tutelage for the past six years. Throughout this odyssey, I have picked up many valuable points, which will definitely help me in the future as a researcher and a tenured professor.

Additionally, I would like to thank the faculty of the Applied Statistics Unit (ASU) and the Theoretical Statistics and Mathematics Unit (TSMU), namely, Prof. Anil Kumar Ghosh, Prof. Tapas Samanta, Prof. Atanu Biswas and Prof. Arijit Chakrabarti, who taught me during my coursework period and their classes really helped me strengthen my base in Statistics, before embarking on my doctoral journey. I would also like to thank the administrative authorities at ISI, especially the Director and the Dean of Studies, for providing research scholars with state-of-the-art research facilities such as our library, which contains copious number of books, and our computer laboratory, which houses advanced computing systems. I would further like to thank the Research Fellow Advisory Committee (RFAC) and the Ph.D./D.Sc. committee for providing their valuable insights whenever I presented during the annual doctoral evaluations and during my thesis submission.

Furthermore, special thanks are reserved for Prof. Abhik Ghosh and Dr. Adhidev Biswas, who were senior research scholars when I joined the Interdisciplinary Statistical Research Unit (ISRU) at ISI. On numerous occasions, I have sought tips and suggestions from them regarding the composition of my thesis and they have never failed to oblige. My research interests and that of Prof. Ghosh's are somewhat aligned and his valuable counsel has helped me immensely to expand on his research work, thereby widening the scope of the research on robust inference.

Last but not the least, I dedicate this work to my parents, Mr. (Late) Ganga Charan Basak and Mrs. Sumita Basak, for standing by me in all of my endeavours since birth and supporting my ambitions to eventually become an academic in Statistics. They were/are truly my Rock of Gibraltar. By virtue of their determination and quality upbringing, I have become the person who I am today. Their constant encouragement and psychological guidance have emboldened me in my relentless pursuit to achieve academic excellence since my schooldays. Lastly, I thank my sister, Ms. Sayantanee Basak, for always being there for me as a friend and providing me with both emotional and psychological support.

Contents

Certificate i

Acknowledgements ii

Contents iv

List of Figures viii

List of Tables xiii

List of Abbreviations xv

List of Symbols xvii

1 Prelude 1

1.1 Introduction . . . 1

1.2 General Notation . . . 5

1.3 Some Well-Known Concepts . . . 7

1.3.1 Fisher Information . . . 7

1.3.2 First Order Efficiency . . . 9

1.3.3 Statistical Functionals . . . 10

1.3.4 M-Estimation . . . 13

1.4 Parametric Inference under Classical Approach . . . 16

1.4.1 Parametric Estimation by Likelihood Method . . . 17

1.4.2 Parametric Hypothesis Testing by Likelihood Method . . . 18

1.5 Robust Parametric Inference . . . 20

1.5.1 Minimum Disparity Estimation . . . 22

1.5.1.1 Disparities . . . 22

1.5.1.1.1 Specific Cases of Disparities: . . . 25

1.5.1.2 Minimum Disparity Estimation under Discrete Models . . . 26

1.5.1.3 Minimum Distance Estimation under Continuous Model . . . 30

1.5.1.3.1 Beran's Approach . . . 31

1.5.1.3.2 Basu-Lindsay Approach 32
1.5.2 Hypothesis Testing using Disparities 35
1.5.2.1 Testing of Hypothesis under the Discrete Model . . 35
1.5.2.2 Testing of Hypothesis under the Continuous Model 37
1.6 The Need for the Optimal Parameter Selection 38
1.7 Aim and Layout of the Thesis 40

2 A Useful Divergence : The Bregman Divergence 44
2.1 Definition 44
2.1.1 Use 45
2.1.2 Properties 45
2.1.3 Some Useful Divergences as Special Cases of the Bregman Divergence 46
2.1.3.1 Likelihood Disparity (LD) 47
2.1.3.2 Squared L_2 Distance 48
2.1.3.3 Density Power Divergence (DPD) 49
2.1.3.4 B-Exponential Divergence (BED) 51
2.1.4 Concluding Remarks 53

3 Choosing the ‘Optimal’ Tuning Parameter 54
3.1 Introduction 54
3.1.1 The Basic Idea in Tuning Parameter Selection 56
3.1.2 Warwick-Jones and Hong-Kim Algorithms 57
3.1.3 Our Proposal 59
3.2 The Three Algorithms : Some Comparisons 60
3.3 Applications 66
3.3.1 I.I.D. data examples 70
3.3.2 No outlier performance 83
3.4 Simulation Study 87
3.5 Computational Cost 92
3.6 Concluding Remarks 94

4 The Extended Bregman Divergence 96
4.1 Rationale behind this Extension 96
4.2 Proposal of the Extension 97
4.3 Some Special Cases of the Extended Bregman Divergence 99
4.3.1 Power Divergence (PD) 99
4.3.2 S-Divergence (SD) 100
4.3.3 S-Hellinger Divergence (SHD) 101
4.4 Concluding Remarks 102

5 A New Extended Bregman Super Family 103
5.1 Introduction 103
5.2 Generalized S-Bregman (GSB) Divergence 104
5.3 Special Cases 106
5.4 Discrete Setup 106

5.4.1 The Minimum GSB Divergence Estimator 107
5.4.2 Estimating Equation 107
5.4.3 Asymptotic Properties 109
5.4.4 Influence Function 124
5.4.5 Simulation Result 127
5.4.6 Selection of Optimal Tuning Parameters through Real Data Analysis 131
5.5 Conclusion 138

6 The Extended Bregman Divergence and Parametric Estimation in Continuous Models 143
6.1 Introduction 143
6.2 The Generalized S-Bregman (GSB) Divergence 143
6.3 The Estimation Scheme under Continuous Models 144
6.4 Estimating Equation 145
6.5 Influence Function under the Basu-Lindsay Approach 146
6.6 Asymptotic Distribution of the Minimum GSB* Divergence Estimator 151
6.7 Derivation of Transparent Kernel for the Minimum GSB* Divergence Estimator 157
6.8 Simulation Results 162
6.9 Real Data Analysis 171
6.10 Conclusion 174

7 Hypotheses Testing using the Extended Bregman Divergence 176
7.1 Introduction 176
7.2 GSD as a Special Form of Bregman Divergence 178
7.3 Testing Parametric Hypothesis using GSD (Simple Null Hypotheses) 181
7.3.1 One Sample Problem 182
7.3.1.1 Some Theorems 184
7.3.2 Two Sample Problem 190
7.3.2.1 Some Theorems 191
7.3.3 Robustness Properties of the GSDT (Simple NULL Hypothesis) 195
7.3.3.1 Influence Function of the Test 196
7.3.3.2 Level and Power Influence Function 197
7.3.4 Some Observations 203
7.4 Testing Parametric Hypothesis using GSD (Composite Null Hypotheses) 205
7.4.1 Influence Function of Restricted MGSDE 207
7.4.2 Discrete Setup 210
7.4.3 GSDT for Composite Hypotheses 214
7.4.4 Robustness Properties of the GSDT (Composite Hypotheses) 220
7.4.4.1 Influence Function of the Test 220
7.4.4.2 Influence Function of the Level and the Power . . . 223
7.5 Simulation Study 230
7.5.1 Some Observations 232

7.5.2 Comparison Among Some ‘Best’ DPDTs–SDTs–GSDTs and the LRT 234
7.6 Real Life Data Examples 271
7.6.1 One Sample Telephone Line Fault Data 272
7.6.2 One Sample Darwin’s Plant Fertilization Data 274
7.6.3 Two Sample Drosophila Data 275
7.7 Concluding Remarks 283

8 A review on the performance of the ‘optimal’ tuning parameter selection algorithm 284

9 Epilogue 291
9.1 Selection of Tuning Parameters 291
9.2 Extended Bregman Divergence 292
9.3 Robust Estimation 292
9.4 Robust Testing 293

A List of Published and Ongoing Papers 294

Bibliography 295

List of Figures

3.1 Asymptotic variance plots corresponding to Case 1. 62

3.2 Asymptotic variance plots corresponding to Case 2. 63

3.3 Asymptotic variance plots corresponding to Case 3. 64

3.4 Asymptotic variance of the MDPDEs against α. 64

3.5 A few different linear regression fits for Life Expectancy Data. . . . 69

3.6 A few different fits for Short's Data under the two-parameter normal model. 75

3.7 A few different fits for Newcomb's Data under the two-parameter normal model. 76

3.8 A few different fits for Telephone-line Fault data under the two-parameter normal model. 77

3.9 A few different fits for Insulating Fluid Data under the exponential model. 79

3.10 A few different regression fits for Star Cluster Data. 81

3.11 A few different regression fits for Gesell Adaptive Score Data. . . . 83

3.12 Residual plots for OLS and optimal IWJ fits, respectively, on Salinity Data. 85

3.13 Comparison among optimal tuning parameters of the three algorithms for samples from pure $N(0,1)$ (top panel) and $P(2)$ (bottom panel) models. 88

3.14 Comparison among optimal tuning parameters of the three algorithms for samples from $.9N(0,1)+.1N(8,1)$ (top panel) and $.9P(2)+.1P(15)$ (bottom panel) models. 89

3.15 The histogram and the asymptotic variance curve of a particular dataset which leads to an optimal $\alpha = 0$ for HK algorithm and an optimal $\alpha = 1$ for IWJ algorithm. 92

3.16 Number of iterations needed to obtain IWJ estimator under pure and contaminated normal model. 94

5.1 Examples of unbounded influence functions (left panel) and bounded influence functions (right panel) corresponding to $(\alpha, \lambda, \beta) \in$ each disjoint subsets contained in $\mathbb{S}$. 139

5.2 The first figure shows the region needed for bounded IF. Here, the grey, the orange, the green and the blue planes represent the boundaries of the sets $\mathbb{S}_1$, $\mathbb{S}_2$, $\mathbb{S}_3$ and $\mathbb{S}_4$, respectively. The 'best' solutions are given in red dots in the second figure. 140

5.3 Some significant fits for the Stolen Bases Data under the Poisson model. Here “clean data” refer to the modified data after removing all 8 outliers. 142

6.1 Examples of unbounded influence functions (left panel) and bounded influence functions (right panel). Only the right panel consists of subsets of $\mathbb{S}$. Those in the left panel are outside!. Here $n = 50$, $f_\theta = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$ with $(\mu, \sigma) = (0, 3)$ and $h = 1.06 * n^{-1/5} * \sigma$ 166

6.2 Some significant fits for the Alkalinity Data under the Normal model. 175

7.1 Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0, 1)$ model) 237

7.2 (Continued) Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0, 1)$ model) 238

7.3 Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0.5, 1)$ model) 239

7.4 (Continued) Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0.5, 1)$ model) 240

7.5 Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(1, 1)$ model) 241

7.6 (Continued) Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(1, 1)$ model) 242

7.7 Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(0, 1) + 0.1\ N(5, 1)$ model) 243

7.8 (Continued) Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(0, 1) + 0.1\ N(5, 1)$ model) 244

7.9 Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(0.5, 1) + 0.1\ N(-5, 1)$ model) 245

7.10 (Continued) Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(0.5, 1) + 0.1\ N(-5, 1)$ model) . . . 246

7.11 Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(1, 1) + 0.1\ N(-9, 1)$ model) 247

7.12 (Continued) Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(1, 1) + 0.1\ N(-9, 1)$ model)3 . . . 248

7.13 Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0, 1)$ model) 249
7.14 (Continued) Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0, 1)$ model) 250
7.15 Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0.5, 1)$ model) 251
7.16 (Continued) Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0.5, 1)$ model) 252
7.17 Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(1, 1)$ model) 253
7.18 (Continued) Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(1, 1)$ model) 254
7.19 Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of 0.9 $N(0, 1) + 0.1$ $N(5, 1)$ model) 255
7.20 (Continued) Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of 0.9 $N(0, 1) + 0.1$ $N(5, 1)$ model) 256
7.21 Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of 0.9 $N(0.5, 1) + 0.1$ $N(-5, 1)$ model) 257
7.22 (Continued) Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of 0.9 $N(0.5, 1) + 0.1$ $N(-5, 1)$ model) 258
7.23 Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of 0.9 $N(1, 1) + 0.1$ $N(-9, 1)$ model) 259
7.24 (Continued) Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of 0.9 $N(1, 1) + 0.1$ $N(-9, 1)$ model)260
7.25 Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(\frac{\sqrt{5}}{\sqrt{n}}, 1)$ models) 261
7.26 (Continued) Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(\frac{\sqrt{5}}{\sqrt{n}}, 1)$ models) 262
7.27 Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of 0.9 $N(\frac{\sqrt{5}}{\sqrt{n}}, 1) + 0.1$ $N(-5, 1)$ models) 263

7.28 (Continued) Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(\frac{\sqrt{5}}{\sqrt{n}}, 1) + 0.1\ N(-5, 1)$ models) . . 264

7.29 Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(\frac{\sqrt{5}}{\sqrt{n}}, 1)$ models) 265

7.30 (Continued) Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(\frac{\sqrt{5}}{\sqrt{n}}, 1)$ models) 266

7.31 Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(\frac{\sqrt{5}}{\sqrt{n}}, 1) + 0.1\ N(-5, 1)$ models) 267

7.32 (Continued) Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(\frac{\sqrt{5}}{\sqrt{n}}, 1) + 0.1\ N(-5, 1)$ models) 268

7.33 Simulated size (plots (A) and (C)) for testing $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ for pure data generated from $N(0, 1)$. Plots (B) and (D) represent the powers of the same test at $\mu = 0.5$ (data generated from $N(0.5, 1)$). In the above plots the GSDT is represented by the QDT. 269

7.34 Simulated size (plots (A) and (C)) for testing $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ for contaminated data generated from $0.9N(0, 1) + 0.1N(-2, 1)$. Plots (B) and (D) represent the powers of the same test at $\mu = 0.5$ (data generated from $0.9N(0.5, 1) + 0.1N(-6, 1)$). In the above plots the GSDT is represented by the QDT. 270

7.35 Plot of p-values of the GSDT (corresponding to two different α's) varying over different pairs of (τ, γ) in case of the Telephone Line Fault data (with outlier) 279

7.36 Plot of p-values of the GSDT (corresponding to two different α's) varying over different pairs of (τ, γ) in case of the Telephone Line Fault data (without outlier) 280

7.37 Plot of p-values of the GSDT (corresponding to two different α's) varying over different pairs of (τ, γ) in case of the Telephone Line Fault data (both with and without outlier) 281

7.38 Plot of p-values of the GSDT (corresponding to two different α's) varying over different pairs of (τ, γ) in case of the Darwin's Plant Fertilization data (both with and without outlier) 282

7.39 Plot of p-values of the GSDT (corresponding to two different α's) varying over different pairs of (τ, γ) in case of the Drosophila data (both with and without outlier) 282

8.1 Histogram of sequence of optimal α values corresponding to three algorithms under 1000 simulations for pure model. 287

8.2 Histogram of sequence of optimal α values corresponding to three algorithms under 1000 simulations for contaminated model. 287

List of Tables

3.1 Life expectancy at birth and health spending per capita (USD PPP), 2015 . . . 69
3.2 Optimal estimates for the Life Expectancy Data . . . 70
3.3 Tuning parameter sequence (Life Expectancy Data) . . . 70
3.4 Recessive lethal count . . . 71
3.5 Tuning parameter sequence (Drosophila Data, day 28) . . . 71
3.6 Tuning parameter sequence (Drosophila Data, day 177) . . . 72
3.7 Peritonitis Data . . . 73
3.8 Tuning parameter sequence (Peritonitis Data) . . . 73
3.9 Short's Data . . . 74
3.10 Tuning parameter sequence (Short's Data) . . . 74
3.11 Newcomb's Data . . . 75
3.12 Tuning parameter sequence (Newcomb's Data) . . . 76
3.13 Telephone-line Fault Data . . . 77
3.14 Tuning parameter sequence (Telephone-line Fault Data) . . . 77
3.15 Insulating Fluid Data . . . 78
3.16 Tuning parameter sequence (Insulating Fluid Data) . . . 78
3.17 Hertzsprung-Russell Star Cluster Data . . . 80
3.18 Tuning parameter sequence (Star Cluster Data) . . . 81
3.19 Optimal estimates for the Hertzsprung-Russell Star Cluster Data . 81
3.20 Gesell Adaptive Score Data . . . 82
3.21 Tuning parameter sequence (Gesell Adaptive Score Data) . . . 82
3.22 Optimal estimates for the Gesell Adaptive Score Data . . . 82
3.23 Salinity Data . . . 84
3.24 Tuning parameter sequence (Salinity Data) . . . 85
3.25 Optimal estimates for the Salinity Data . . . 85
3.26 The three optimal values of α for the full data as well as the outlier removed data corresponding to all datasets. . . . 86

5.1 Different divergences as special cases of GSB divergence . . . 106
5.2 MSEs of the minimum divergence estimators within the S-divergence family for pure and contaminated data . . . 132
5.3 MSEs of the minimum GSB divergence estimators under pure and contaminated data . . . 133
5.4 Optimal estimates in different cases for the Drosophila Data (unrestricted β) . . . 135

5.5 Optimal estimates in different cases for the Drosophila Data (β restricted to -4) . . . 135
5.6 Top 40 Home Run (HR) scorers of the AL in the 2019 MLB Regular Season . . . 141
5.7 Optimal estimates in different cases for the Stolen Bases Data . . . 142
5.8 Optimal estimates in different cases for the Stolen Bases Data (with $\beta = -4$) . . . 142

6.1 Comparison of MSEs of MSDE* and MGSBE* under the contaminated model $(1-\epsilon)N(0,9) + \epsilon N(15,9)$ . . . 167
6.2 Comparison of MSEs of MSDE* and MGSBE* under the contaminated model $(1-\epsilon)N(0,9) + \epsilon N(0,100)$ . . . 168
6.3 Comparison of MSEs of MSDE* and MGSBE* under the contaminated model $(1-\epsilon)N(0,9) + \epsilon\chi^2_{10}$ . . . 169
6.4 MSEs along with triplets (α, λ, β) of some MGSBE*s belonging to the 'best' region . . . 170
6.5 Short's Data . . . 172
6.6 Newcomb's Data . . . 173
6.7 Alkalinity Data . . . 174
6.8 Optimal estimates for the Alkalinity Data . . . 174

7.1 Telephone Line Fault Data . . . 274
7.2 Darwin's Plant Fertilization Data . . . 275
7.3 Frequencies of the number of recessive lethal daughters for the Drosophila data . . . 276
7.4 Value of the test statistic $S^*_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right)$ and associated p-values for small α in case of both the full and the clean Drosophila data . . . 278

8.1 Mean Square Error under pure and contaminated Poisson data . . . 287

List of Abbreviations

BD	Bregman Divergence
BED	Bregman Exponential Divergence
CAN	Consistent and Asymptotically Normal
CDF	Cumulative Distribution Function
CUAN	Consistent and Uniformly Asymptotically Normal
DCT	Dominated Convergence Theorem
DPD	Density Power Divergence
EBD	Extended Bregman Divergence
GSB	Generalised S-Bregman Divergence
GSD	Generalised Super Divergence family
HD	Hellinger Distance
HK	Hong and Kim
i.i.d.	Independent and Identically Distributed
IF	Influence Function
IFT	Implicit Function Theorem
IWJ	Iterated Warwick-Jones
KDE	Kernel Density Estimation
KLD	Kullback-Leibler Divergence
LD	Likelihood Disparity
LIF	Level Influence Function
LRT	Likelihood Ratio Test
MDPDE	Minimum Density Power Divergence Estimator
MGSBE	Minimum Generalised S-Bregman Divergence Estimator
MLE	Maximum Likelihood Estimation

MS*DE	Minimum S*-Divergence Estimator
MSDE	Minimum S-Divergence Estimator
MSE	Mean Squared Error
NCS	Neyman's Chi-Square
OLS	Ordinary Least Squares
OWJ	One-Step Warwick-Jones
PCS	Pearson's Chi-Square
PD	Power Divergence
PDF	Probability Density Function
PIF	Power Influence Function
PMF	Probability Mass Function
QDT	Q-Divergence Test
RMGSD*E	Restricted Minimum Generalised Super Divergence* Estimator
RMGSDE	Restricted Minimum Generalised Super Divergence Estimator
SDT	S-Divergence Test
SHD	S-Hellinger Distance
SLLN	Strong Law of Large Numbers

List of Symbols

θ	:	parameter of interest
Θ	:	parameter space
$\mathbb{R}$	:	set of real numbers
$\mathbb{R}^p$	:	set of p-dimensional real vectors
$\mathscr{F}$	:	family of probability distributions
f	:	model density
g	:	data density
F	:	cumulative distribution function of model density
G	:	cumulative distribution function of data density
G_n	:	empirical distribution depending on n
n	:	sample size
χ	:	support of a distribution
$\overset{\mathrm{D}}{\to}$	:	convergence in distribution
$\overset{\mathrm{P}}{\to}$	:	convergence in probability
H_0	:	null hypothesis
H_1	:	alternative hypothesis
$H_{1:n}$	:	contiguous alternative hypothesis depending on n
Λ_y	:	degenerate distribution function at y
$\phi(.)$	:	standard normal density function
$\Phi(.)$	:	standard normal distribution function
$u_\theta(x)$	:	score function
$i_\theta(x)$	:	information function
∇	:	derivative with respect to θ
∇_j	:	derivative with respect to jth component of θ

∇_{jk}	: second derivative with respect to jth and kth component of θ
∇_{jkl}	: third derivative with respect to jth, kth and lth component of θ
∇^2	: second derivative with respect to θ
$\frac{\partial f}{\partial x}$	: partial derivative of f with respect to x

Dedicated to my father, who currently resides in his heavenly abode, and my mother

Chapter 1

Prelude

1.1 Introduction

We need to do appropriate statistical analysis to make inference about the characteristics of a population of interest in a real life situation. This generally involves the selection of a suitable sample from that population and the extrapolation of findings from that sample to the whole population. In classical parametric inference it has been the aim of the statistician to model the true random variable of interest through appropriate parametric forms indexed by a finite number of parameters which provide a good fit to the observed data pattern and can be useful in further analysis and prediction. Till about the middle of the last century, this statistical analysis focused almost entirely on the efficiency aspect of the problem. Around the 1950s and 1960s, however, the need for the stability of the procedure under non-ideal conditions also started gaining recognition. In practice, all statistical methods rely on a bunch of assumptions, either implicitly or explicitly, and in classical procedures the failure of the assumptions to hold may put the validity of the analysis in question. Yet, small deviations from assumed model conditions are never really completely unexpected in real life. In addition, legitimately occurring outliers may also cause stability problems for classical (but efficient) methods of statistical analysis. But life is dynamic, so researchers have gladly welcomed the concept of "robustness" to cover for these deficiencies. Robustness is generally (but not always) associated with a loss in asymptotic efficiency, which is viewed by many authors as the cost of achieving robustness. In our research it will be our endeavor to search for robust techniques with minimal loss in efficiency.

In this thesis we will focus on the minimum distance approach to robust inference. This concept was initially pioneered by Wolfowitz (1952, 1953, 1957), who described the desirable properties of this method under suitable conditions. This idea, based on the quantification of a measure of discrepancy between the data and the model, is a natural one, and later research has shown that it may have an important role in generating inference procedures with some natural robustness properties. See, e.g., Donoho and Liu (1988).

Two broad types of distances are usually used in the literature for minimum distance inference. We describe them below.

1. The distances between the distribution functions of the data and the model; these include, for example, the Kolmogorov-Smirnov distance, the Cramér-von Mises distance (von Mises (1936, 1937, 1947)), the Anderson-Darling distance (Anderson and Darling (1952)), etc.

2. The distances between probability density functions (PDFs). More specifically, the distance between some non-parametric

density estimate of the data density and the model density; for example, the Pearson's χ^2 (Pearson (1900)), the Hellinger distance (Hellinger (1909)), the Kullback-Leibler divergence (Kullback and Leibler (1951)), the Bregman divergence (Bregman (1967)), etc.

The robustness concept started to catch up during the 1950s and 60s, and since then the research in this area has grown at a furious pace. The term 'robust' was coined by Box (1953), and profound early contributions were made by Tukey (1960), Huber (1964, 1965, 1967, 1968, 1970, 1972, 1973, 1975, 1981) and Hampel (1968, 1971,1974) in establishing the general theory of robustness in the presence of outliers. Hampel (1974) introduced the concept of the 'Influence Curve', which has become one of the most important and exclusive heuristic evalutators of robustness and currently goes by the modified name of 'Influence Function'. Following the works of Huber, Hampel and the other pioneers, the statistical community has witnessed a vast growth in this literature from a preliminary problem to a rich and complicated problem. The early literature includes, among many other prominent contributions, the works of Andrews et al. (1972), Maronna (1976), Yohai and Maronna (1979), Ronchetti (1982), Rousseeuw and Yohai (1984), Rousseeuw (1985), Maronna and Yohai (1995, 2004), Robinson et al. (2003), etc.

Under the minimum distance approach, the works of Parr and Schucany (1980, 1982), Boos (1981, 1982), Parr and De Wet (1981) and Wiens (1987) among others, made significant contributions using distance measures based on distribution functions. On the other hand, very significant contributions in the approach based on density-based divergences were made during this period by Beran (1977), Cressie and Read (1984), Tamura and Boos (1986), Simpson (1987, 1989), Donoho and Liu (1988), Lindsay (1994), etc. Since the information content of the model is linked to the score function, it appears that the approach based on density functions is best suited for retaining full efficiency under any robust approach.

Another important use of the minimum distance technique is in testing the goodness-of-fit of a statistical model. This part also relates to the closeness between the data and the model, but in an opposite sense compared to robust parametric estimation. In the latter case, the basic target is to down-weight the effect of outliers and small deviations; on the contrary, the aim of goodness-of-fit testing is to magnify the small deviations from the hypothesized model for achieving high power for the test. Evidently the distances which make the most useful contribution in robust parametric inference need not be the best ones for the goodness-of-fit testing problem. The contribution of Cressie and Read (1984), through the introduction of the 'Power Divergence' family, is meaningful in this regard. According to them, this family of statistics provides an innovative way to unify and extend the literature by linking the traditional goodness-of-fit test statistics through a single, real-valued parameter. However, as expected, the divergences which provide powerful goodness-of-fit tests are not necessarily the optimal ones in the context of robust inference.

Another significant part of robust inference whose omission will render our research incomplete is the selection of optimal tuning parameter(s). Most robust estimation procedures, including practically all minimum distance procedures, depend on the choice of a tuning parameter which normally controls the trade-off between robustness and efficiency. Hence, there will always be a risk of inappropriate analysis if the chosen tuning parameter(s) is/are inappropriate. We will follow up on some existing data driven choices for determining the value(s) of the tuning parameter(s) so that the full potential of these estimators might be realized. One major drawback of some of these methods is the dependency on an initial pilot estimator. With inappropriate pilots, the robustness issue of the estimator may itself be in jeopardy. Hence, our endeavour is to develop a general method which will optimize their performance and remove this pilot dependency to the extent possible.

1.2 General Notation

In this section, we will present some mathematical notation which we will encounter throughout this thesis.

(i) Unless mentioned otherwise, the term ‘log’ will represent the natural logarithm.

(ii) Unless mentioned otherwise, the term ‘density function’ will represent both the probability mass function for discrete models and the probability density function for continuous models.

(iii) The uppercase letters G and F_θ will denote the cumulative distribution functions, whereas the lowercase letters g and f_θ will denote the corresponding density functions.

(iv) Given an i.i.d. sample $X_1, X_2, \ldots, X_n$, the empirical version of the true distribution G will be denoted by G_n, which has the form

$$G_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x),$$

where $I(A)$ is the indicator function of the event A.

(v) $\Lambda_y(x)$ will denote the distribution function of the degenerate distribution at y and $\lambda_y(x)$ will denote the corresponding probability density function.

(vi) The uppercase letters J, V, etc. will denote the matrices under the model f_θ, whereas J_g, V_g, etc. will denote the corresponding matrices needed to express the asymptotic variance of the estimators under g.

(vii) Unless mentioned otherwise, we will assume that θ is a p-dimensional parameter and we will denote the parametric model family by $\mathcal{F} = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$.

(viii) The symbol ∇ will represent the gradient with respect to the parameter. More specifically, the jth component of the gradient will be denoted by ∇_j. Similarly, the second order derivative with respect to the parameter θ will be denoted by ∇^2, where ∇_{jk} will represent the joint partial derivative with respect to θ_j and θ_k.

(ix) Unless mentioned otherwise, χ will denote the support of a distribution.

1.3 Some Well-Known Concepts

1.3.1 Fisher Information

In the field of parametric estimation, the 'Fisher Information' is one of the most important concepts in the study of the parameter θ indexing the parametric model under study. Let us suppose that X has a density function f_θ with respect to a σ-finite measure υ on $\mathbb{R}$. Let $\Theta \subseteq \mathbb{R}$ be the parameter space, i.e., θ is a scalar parameter. Furthermore, for any measurable set $B \subset \mathbb{R}$, we assume that the relation.

$$\nabla \int_B f_\theta(x) d\upsilon(x) \;=\; \int_B \nabla f_\theta(x) d\upsilon(x), \tag{1.1}$$

is satisfied where ∇ is the gradient with respect to θ. Moreover the score function $u_\theta(x)$ equals the derivative of the log of the density, i.e.,

$$u_\theta(x) = \nabla \log(f_\theta(x)) = \frac{\nabla f_\theta(x)}{f_\theta(x)}.$$

From Equation (1.1), it can be shown that

$$E_\theta\left[u_\theta(X)\right] = 0.$$

The Fisher Information $I(\theta)$ turns out to be the variance of the score function, that is,

$$I(\theta) = V_\theta\left[u_\theta(X)\right] = E_\theta\left[u_\theta^2(X)\right].$$

Here $V_\theta(\cdot)$ represents variance with respect to f_θ. Evidently, a high value of $I(\theta)$ indicates a substantially high score function on the average, that is, the rate of change of the density function with respect to

θ is quite significant in such a scenario. For a family of distributions with highly variable u_θ, we intuitively expect that the estimation of the parameter θ using the sample to be easier – different values of θ change the behaviour of the score function u_θ. If this u_θ is close to zero, then we can conclude that the corresponding random variable does not provide much information about θ. On the contrary, if $|u_\theta|$ or u_θ^2 is large, then we can assume that the random variable plays a significant role in giving information about θ; thus the score is quite sensitive towards changes in θ. This Fisher Information basically measures the overall sensitivity of the functional relationship of f_θ to the changes in θ through imposing a weight $f_\theta\,(x)$ to this sensitivity at each potential outcome x.

Our discussion, so far, has been in the context of a scalar parameter. If we consider the p-dimensional case, $p > 1$, then, $\theta = (\theta_1, \ldots, \theta_p)^T$, and the score function becomes

$$u_\theta(x) \;\; = \;\; \left(u_{1\theta}(x), u_{2\theta}(x), \ldots, u_{p\theta(x)}\right)^T \tag{1.2}$$

where, $u_{j\theta}(x) = \frac{\nabla_j f_\theta(x)}{f_\theta(x)}$. Similarly, the Fisher Information (matrix) is defined in this case as

$$I(\theta) \;\; = \;\; E\left(u_\theta(X) u_\theta(X)^T\right)$$

where $I_{jk}(\theta) = E_\theta\left[u_{j\theta}(X) u_{k\theta}(X)\right]$. By definition, this matrix is non-negative definite.

In case of n i.i.d. sample observations from the density f_θ with a unidimensional $\theta \in \Theta$, an open subset of $\mathbb{R}$, suppose we want to estimate a continuously differentiable real-valued function $m(\theta)$. If T_n is an unbiased estimator of $m(\theta)$, then under certain regularity conditions, the variance of T_n can be bounded by the relation

$$V_\theta(T_n) \;\geq\; \frac{(m^{'}(\theta))^2}{nI(\theta)}, \tag{1.3}$$

where $I(\theta)$ denotes the Fisher Information and $m^{'}(\cdot)$ represents the derivative of $m(\cdot)$. The bound given in Equation (1.3) is called the Cramér-Rao lower bound; see, e.g., Rao (1973). For a p-dimensional θ, this lower bound becomes

$$V_\theta(T_n) \geq \frac{1}{n} h^T I^{-1}(\theta) h,$$

where $h = (h_1, h_2, \ldots, h_p)^T = (\nabla_1 m(\theta), \nabla_2 m(\theta), \ldots, \nabla_p m(\theta))^T$.

1.3.2 First Order Efficiency

In the field of parametric estimation, efficiency is a measure of the quality of an estimator – more specifically, it helps us to quantify the relative degree of undesirability of estimation errors of different magnitudes, through some particular choice of the loss function. Consider an i.i.d. random sample of size n from the true data generating distribution, and let f_θ be the model density. In looking for the most appropriate estimator, it is often sufficient to restrict ourselves to only consistent and asymptotically normal (CAN) estimators, T_n, for estimating $m(\theta)$, for which there exists a positive, asymptotic variance $\nu(\theta)$ such that

$$\frac{\sqrt{n}(T_n - m(\theta))}{\sqrt{\nu(\theta)}} \xrightarrow{\text{D}} Z \sim N(0,1)$$

as $n \to \infty$. The first order efficiency of these CAN estimators is checked through a comparison between $\nu(\theta)$ and the Cramér-Rao lower bound given in Equation (1.3) and an estimator can be called first order efficient, if its asymptotic variance coincides with the lower bound given by Equation (1.3).

Sometimes one may come across such estimators whose asymptotic variance is less than this lower bound (see Le Cam (1953)). These 'super-efficient estimators', which are not statistically meaningful, are to be eliminated from the set of CAN estimators to restrict our attention to consistent and uniformly asymptotically normal (CUAN) estimators only. Let us restrict our attention to the case $m(\theta) = \theta$, and let θ be a scalar. Under this setup, we will consider T_n to be first order efficient if

$$\sqrt{n}(T_n - \theta) \xrightarrow{\text{D}} N(0, \frac{1}{I(\theta)}), \text{ under } f_\theta.$$

For the multiparameter case, the asymptotic variance of $\sqrt{n}(T_n - \theta)$, where T_n is a first order efficient estimator, is given by $I^{-1}(\theta)$.

1.3.3 Statistical Functionals

Often a statistic can be viewed as a functional defined on an appropriate space of distribution functions. Let $T_n(X_1, \ldots, X_n)$ be a p-dimensional statistic based on a random sample $X_1, \ldots, X_n$ drawn from G. The empirical distribution G_n is as defined in Section 1.2. Suppose we can express T_n as $T(G_n)$, where $T : \mathcal{G} \to \mathbb{R}^p$ is a functional independent of n. Then this $T(\cdot)$ is called a statistical functional. Here, $\mathcal{G}$ refers to a suitable collection of the distribution functions, including all empirical distribution functions. Moreover, a statistical functional is considered to be linear if for any $G, F \in \mathcal{G}$,

$$T(\epsilon F + (1-\epsilon)G) \;=\; \epsilon T(F) + (1-\epsilon)T(G), \epsilon \in [0,1].$$

Evidently, any functional T with the form $T(G) = \int \phi(x)dG(x)$ is linear. In fact, a linear statistical functional must be of this form. Again, another significant property of any estimator in the form of statistical functional is given by the following definition.

Definition 1.1. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a distribution modeled by the parametric family $\mathcal{F} = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ and let $\hat{\theta} = T(G_n)$ be the estimator of the parameter θ where G_n is the empirical distribution function. This estimator will be called Fisher consistent, if the functional T satisfies $T\left(F_\theta\right) = \theta$.

According to von Mises (1936, 1939, 1947), through the use of the central limit theorem, it can be claimed that the asymptotic distribution of linear statistical functionals converge to normal distributions under certain suitable differentiability conditions. This specific formulation of statistics, based on distribution functions, is crucial in the context of the most heuristic tool for measuring the robustness of statistics, called the influence function (IF). It describes the effect of an additional observation at any point x on a statistic T, given a sample with distribution F. Roughly speaking, the influence function is the first derivative of a statistic T at an underlying distribution G, where the point x plays the role of the coordinate in the infinite dimensional space of probability distributions.

For any two distributions G and F, the von Mises derivative of a functional T at G, denoted by $T_G^{'}$, is defined as

$$T_G^{'}(F - G) \;=\; \frac{\partial}{\partial \epsilon} T\left(\epsilon F + (1-\epsilon)G\right)\Bigg|_{\epsilon=0} = \frac{\partial}{\partial \epsilon} T\left(G + \epsilon(F - G)\right)\Bigg|_{\epsilon=0},$$

if there exists a real-valued function $\phi_G(x)$, independent of F, such that

$$T_G^{'}(F - G) \;=\; \int \phi_G(x) d(F - G)(x),$$

and the additional restriction

$$\int \phi_G(x) dG(x) \;=\; 0$$

is required for the uniqueness of ϕ_G. This ϕ_G is called the IF of the functional T at G and will be denoted as $IF(y, T, G)$ in the upcoming portion of this thesis. Evidently, its mean is 0. However, the influence function can also be computed directly without relying on the existence of the von Mises derivative. Under appropriate conditions, it can be represented directly as

$$\phi_G(y) \;=\; \frac{\partial}{\partial \epsilon} T\left(\epsilon \Lambda_y + (1 - \epsilon)G\right)\bigg|_{\epsilon=0} \;=\; \frac{\partial}{\partial \epsilon} T\left(G + \epsilon(\Lambda_y - G)\right)\bigg|_{\epsilon=0},$$

where Λ_y is defined in Section 1.2. For illustrations, consider the mean functional

$$T_{mean}(G) \;=\; \int x dG.$$

It is a linear functional and a simple calculation leads to the fact that $IF(y, T_{mean}, G) = y - T_{mean}(G)$. Evidently, it is unbounded in y, which indicates the non-robust nature of the sample mean.

Again, the influence function of the statistical functional $T(G)$ has an important connection with the asymptotic distribution of the corresponding estimator $T(G_n)$. For this, we need an approximation of the form $T(G_n)-T(G) \approx T_G^{'}(G_n-G)$ in large samples. Let us define

$$A(\epsilon) \;=\; T\left(\epsilon F+(1-\epsilon)G\right), \epsilon \in [0,1].$$

A Taylor series expansion for $A(\epsilon)$ around $\epsilon = 0$ would be

$$A(\epsilon) = A(0) + \epsilon A^{'}(0) + \text{higher order terms.} \tag{1.4}$$

Replacing F by G_n, we evaluate the above expansion at $\epsilon = 1$ to get,

$$\begin{aligned} T(G_n) &= T(G) + T_G^{'}(G_n - G) + R_n \\ &= T(G) + \int \phi_G(x) dG_n(x) + R_n, \end{aligned}$$

where R_n = higher order terms.

Therefore,

$$\begin{aligned} \sqrt{n}\left(T(G_n) - T(G)\right) &= \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\phi_G(X_i) + \sqrt{n}R_n \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_G(X_i) + \sqrt{n}R_n. \end{aligned}$$

If $\sqrt{n}R_n \xrightarrow{\text{P}} 0$ as $n \to \infty$, we have

$$\sqrt{n}\left(T(G_n) - T(G)\right) \xrightarrow{\text{D}} N(0, V(\phi_G(X))).$$

1.3.4 M-Estimation

Huber introduced a flexible class of estimators – see Huber (1981) for an extended discussion – which later became quite useful in the field of robust estimation. It is a slight generalization of the MLE (maximum likelihood estimator, discussed later). Consider a scalar parameter θ. Now instead of solving $\sum u_\theta(X_i) = 0$ (the estimating equation of the MLE) based on an i.i.d. sample $X_1, X_2, \ldots, X_n$, consider the solution of $\sum \psi_\theta(X_i) = 0$ without restricting $\psi_\theta : \Theta \times \chi \to \mathbb{R}$ to the form of the score function. This class of estimators are called M-estimators. Huber calls them "maximum likelihood estimates under non-standard conditions".

Let the functional $T(G) = \theta$ be the solution of

$$\int \psi_\theta(x) dG(x) \;=\; 0. \tag{1.5}$$

Now, for a scalar parameter θ, if one replaces G by G_n, we obtain the M-estimator $T_n = T(G_n)$ as a solution of the estimating equation

$$\sum \psi_\theta(X_i) \;=\; 0. \tag{1.6}$$

Let $G_\epsilon = (1-\epsilon)G + \epsilon\Lambda_y$ be the contaminated distribution and let θ be a scalar parameter. After taking the derivative of the corresponding estimating equation under contamination

$$\int \psi_{T(G_\epsilon)}(x)\, dG_\epsilon(x) \;=\; 0$$

with respect to ϵ and evaluating at $\epsilon = 0$, the IF of the M-estimator is found to be

$$IF(y, T, G) \;=\; \frac{\psi_{T(G)}(y)}{\int \psi'_{T(G)}(x)\, dG(x)}.$$

Since the MLE is a special case of the M-estimator class, we can derive that the IF of the maximum likelihood functional T_{ML} is

$$IF(y, MLE, F_\theta) \;=\; \frac{u_\theta(y)}{I(\theta)}$$

under the parametric model family $\{F_\theta : \theta \in \Theta \subset \mathbb{R}\}$. On the other hand, consider the sample median $T_n = T(G_n)$ as an estimator of location, which is also an M-estimator solving the equation

$$\int \psi_{T_n}(x)\, dG(x) \;=\; 0$$

where $\psi_T(x) = [I(x > T) - I(x < T)]$, $I(\cdot)$ being the indicator function. The IF in this case becomes

$$IF(y, T, G) = \begin{cases} \dfrac{1}{2g(t)}, & \text{for} \;\; y > t \qquad (1.7) \\ -\dfrac{1}{2g(t)}, & \text{for} \;\; y < t. \qquad (1.8) \end{cases}$$

for the median function T, given by the equation $\int \psi_T(x)\, dG(x) = 0$. Clearly, these two IFs indicate the influence of small departures from the assumed distributions on the MLE and the sample median, respectively.

Since $u_\theta(y)$ is usually unbounded, it shows non-robust characteristics in case of the MLE, while the IF of the sample median is bounded and hence, it has better robustness and stability properties compared to both the sample mean and the MLE.

If we concentrate on its large sample properties, we can derive that,

$$\sqrt{n}\,(T_n - T(G)) \xrightarrow{\text{D}} N(0, \sigma_G^2)$$

for an M-estimator where,

$$\sigma_G^2 \;=\; \frac{\int \psi_{T(G)}^2(x)\, dG(x)}{\left(\int \psi_{T(G)}'(x)\, dG(x)\right)^2}.$$

In the multiparameter case, the expression of IF will become

$$IF(y,T,G) \;=\; \left[\int \psi_{T(G)}^{'}(x)\, dG(x)\right]^{-1} \psi_{T(G)}(y)$$

and the corresponding asymptotic variance will be

$$\begin{aligned}\Sigma_G &= J^{-1}KJ^{-1}\\ \text{where, } J &= E_G\left[\psi_{T(G)}^{'}(X)\right]\\ K &= E_G\left[\psi_{T(G)}(X)\psi_{T(G)}^T(x)\right].\end{aligned}$$

For more details on M-estimation, Huber (1981), Hampel et al. (1986) and Maronna et al. (2006) may be consulted.

1.4 Parametric Inference under Classical Approach

Parametric statistical inference is used in real life scenarios when we have a reasonable idea about the model describing the available data/the performed experiment except for a few numerical values labelling the model, called parameters. Consider the sample values of the observable random variables $\{X_1, X_2, \ldots, X_n\} = \boldsymbol{X}$, and the parametric family $\mathcal{F}$ of distribution function $F_\theta(x)$, $\theta \in \Theta$ describing it. In relation to the family $\mathcal{F}$, a function $l_\theta(\boldsymbol{x})$, which is assigned a value proportional to the probability density function of $\boldsymbol{X}$ at $\boldsymbol{X} = \boldsymbol{x}$ over θ for each specific choice of θ, is called the likelihood function.

On the basis of this likelihood function, in the early part of the twentieth century, Sir Ronald A. Fisher had initiated the development of the concrete mathematical formulation of the theory of maximum likelihood. At present, this technique is the default choice of most researchers in case of parametric inferential problems.

1.4.1 Parametric Estimation by Likelihood Method

Although we have already introduced the likelihood function, the key to the maximum likelihood estimation procedure, we are interested in viewing this technique within the framework of statistical functionals. Let us consider $X_1, X_2, \ldots, X_n$ to be an i.i.d. sample from G and let $\mathcal{F} = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ be a collection of distribution functions modelling the data generating distribution. As mentioned earlier, the likelihood function is defined as

$$l_\theta(\boldsymbol{X}) \;=\; \prod_i f_\theta(X_i). \tag{1.9}$$

According to Fisher (1922), this likelihood function is nothing but the 'frequency of occurrence of a particular value of a parameter of interest'. Fisher's idea concerns 'the value of the parameter having the highest frequency' which, in the probabilistic approach, coincides with the value of the parameter having the highest probability of occurrence. We introduce this technique here. The maximum likelihood estimation technique basically evaluates that value of θ which corresponds to the maximum of the likelihood. In practice, we often obtain the MLE by solving the system of likelihood equations obtained by equating the derivative of the log-likelihood to zero. Thus, the likelihood equations are

$$\frac{\partial \log\, l_\theta(\boldsymbol{X})}{\partial\theta} \;=\; \sum_{i=1}^{n} u_\theta(X_i) = 0,$$

where u_θ is the score function as mentioned earlier. As we have already shown that the MLE is an M-estimator also, the maximum likelihood functional $T_{ML}(G)$ will be defined as

$$\int u_{T_{ML}(G)}\,(x)\; dG(x) = 0.$$

Moreover, $T_{ML}(F_\theta) = \theta$, which shows that the MLE is Fisher consistent. In Section 1.3.4, we have already observed the expression of the IF of the MLE, which indicates the well-known lack of the robustness property of the MLE.

1.4.2 Parametric Hypothesis Testing by Likelihood Method

Due to its desirable asymptotic properties, Neyman and Pearson (1928) and Wilks (1938) developed the theory of testing of hypothesis based on likelihood methods. The details of the theory and the asymptotic properties of the likelihood-based tests are given in many standard texts of statistical inference and asymptotic theory; see, for example, Serfling (1980). Here we consider all likelihood-based classical tests under the same parametric setup.

To describe the tests of hypothesis based on likelihood methods, let us denote the average of the score functions as

$$Z_n(\theta) \;=\; \frac{1}{n}\sum_{i} u_\theta(X_i).$$

In such a scenario, three types of tests are popular and they are described below. As this introduction is being given just for a general flavour of likelihood based tests, in the following we describe the tests for a simple null hypothesis only. Consider the null hypothesis $H_0 : \theta = \theta_0$ to be tested against $H_1 : \theta \neq \theta_0$.

(i) Wald's test (Wald (1943)): This test, named after Abraham Wald, mainly deals with the weighted distance between the unrestricted estimate and its null hypothesized value. Here, $\hat{\theta}_n$ is obtained by maximizing the likelihood over the whole parameter space Θ, that is,

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} \log(l_\theta(\boldsymbol{X})).$$

The corresponding test statistic

$$W_n = n\left(\hat{\theta}_n - \theta_0\right)^T I(\theta_0)\left(\hat{\theta}_n - \theta_0\right)$$

has an asymptotic χ^2_p distribution under H_0.

(ii) Score test: The main component of this test, introduced by Rao (1948), is the score function evaluated under the null hypothesis. Since, under the null, θ is fixed, the required test statistic will be

$$S_n = nZ_n(\theta_0)^T I^{-1}(\theta_0) Z_n(\theta_0)$$

which will asymptotically follow a χ^2 distribution with p degrees of freedom under H_0. Moreover, under the null hypothesis, S_n is separated from W_n by $o_p(1)$ term only. Under the simple null hypothesis, the score test does not require any parameter estimation.

(iii) Likelihood Ratio test: Let $\hat{\theta}_n$ be the unrestricted maximum likelihood estimator obtained by maximizing the likelihood under the unrestricted parameter space Θ. A test statistic is constructed by evaluating the deviation of log-likelihood of $\hat{\theta}_n$ from the log-likelihood of θ_0. The likelihood ratio test statistic is

$$\lambda_n \;=\; 2\left[\log\left(l_{\hat{\theta}_n}(\boldsymbol{x})\right) - \log\left(l_{\theta_0}(\boldsymbol{x})\right)\right]$$

and in this case, this statistic asymptotically follows a χ^2 distribution with p degrees of freedom under H_0.

For all these three cases, a right-tailed test based on W_n, S_n and λ_n would be appropriate for the rejection of H_0. The three test statistics are asymptotically equivalent under H_0. Their asymptotic equivalence continues to hold in case of local alternatives converging sufficiently fast, but the statistics may behave differently for fixed non-local alternatives under certain regularity conditions.

1.5 Robust Parametric Inference

Statistical inference is mainly based on two things – one is the set of sample observations and the other is the set of assumptions about the underlying situation. The classical approach based on the maximum likelihood, which is the cornerstone of parametric inference, performs best when all these assumptions hold in the given scenario. However these assumptions only approximate the reality, and small deviations can never be entirely eliminated. The data analyst may also have to face gross errors and outliers which are incompatible with the model. In a crude sense, an outlier is an observation that lies outside the overall pattern of a distribution (Moore and McCabe (1999)), i.e, it is distant from the majority of the other observations (Grubbs (1969)). This is a geometrical view of an outlier. They are generally bad data points that can lead to highly inefficient and unstable performance of the classical technique, along with dangerous consequences. This problem has been observed for a long time and the initial attempts involved rejection of these 'outliers'. At the present time, researchers believe that outliers may contain valuable information about a system, and should be further scrutinized, rather than being subjectively deleted from the data set. Besides, in the present age of big and high-dimensional data, identifying outliers may be a very difficult task. The inference, therefore, should be dealt with suitable robust procedures, which automatically decide if and by how much to discount an observation suspected to be an outlier. In the next subsection we will expand the idea of outliers to probabilistic outliers, going beyond geometric outliers.

The word 'robust' is loaded with many inconsistent connotations, but we will proceed with the idea of 'robustness' given in Huber (1981), that is, 'insensitivity to small deviations from the assumptions'.

The robust approach to statistical modelling and data analysis aims at introducing techniques which produce stable statistics leading to reliable parameter estimates, associated tests and confidence intervals, not only when the data follow a given distribution exactly, but also when there are mild violations to the parametric assumptions or contamination is present in the data. This approach provides a very reliable method of detecting outliers, even in high-dimensional, multivariate scenarios. While the problem of robustness is quite old in the history of statistics, it has been formalized only in the later part of the twentieth century. Some approaches consider the general and abstract notions of stability, whereas others have taken different topological and geometrical aspects related to robustness into account. One such popular method adopted in the field of robust inference is the method based on disparities.

1.5.1 Minimum Disparity Estimation

In the field of robust inference, the main component of statistical modelling is to minimize the amount of discrepancy between the model and the data through some robust measuring tool. We have already mentioned disparities as one such class of discrepancy measuring tools, many of which are quite insensitive to the presence of outliers. A brief description of the minimum disparity estimation procedure is given below.

1.5.1.1 Disparities

The class of 'disparity' measures is essentially the family of chi-square type distances, also called the ϕ-divergences, or the f-divergences in the literature (see Csiszár (1963, 1967), Ali and Silvey (1966), Lindsay (1994), Pardo (2006) or Basu et al. (2011)). The class of chi-square type distances between two densities g and f includes, for example, the likelihood disparity (LD), the Kullback-Leibler divergence (KLD) and the (squared) Hellinger distance (HD), which are discussed in the next subsection. The main advantage of focusing on this family is that it allows us to comprehensively study a common class of estimators which includes the MLE as well as many remarkably strong robust estimators.

In the parametric estimation scheme that we are considering, the estimator corresponds to the parameter of the model density which is nearest to the observed data density in terms of the given divergence, the observed data density being a non-parametric representative of the true, unknown density, based on the given sample.

Pardo (2006) provides a nice description of minimum disparity methods in discrete models with finite support, based on the multinomial distribution. On the other hand, in case of continuous models, the construction of the data density inevitably requires the use of an appropriate smoothing technique, like kernel density estimation for chi-square type distances. To be more specific, we will follow the approach of Lindsay (1994) to describe this methodology through the residual adjustment function and the Pearson residual discussed later in this chapter.

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a distribution G, having density g with respect to the counting measure. The support of the distribution is taken to be, without loss of generality, $\chi = \{0, 1, 2, \ldots\}$. Let $r_n(x)$, relative frequency at x, be the data based estimate of the probability of occurrence of x. Moreover, we consider the parametric model family $\mathcal{F}$, which models G; both G and $\mathcal{F}$ belong to $\mathcal{G}$, the convex class of all distributions having densities with respect to the counting measure. We quantify the separation between the vectors $r_n = (r_n(0), r_n(1), \ldots)^T$ and $f_\theta = (f_\theta(0), f_\theta(1), \ldots)^T$, where, both vectors satisfy

$$\sum_{x=0}^{\infty} r_n(x) \;=\; \sum_{x=0}^{\infty} f_\theta(x) = 1.$$

Definition 1.2. Let C be a thrice differentiable, strictly convex function on $[-1, \infty)$, satisfying

$$C(0) \;=\; 0. \tag{1.10}$$

Let the *Pearson residual* at the value x be defined by

$$\delta(x) \;=\; \frac{r_n(x)}{f_\theta(x)} - 1. \tag{1.11}$$

(We will denote it by $\delta_n(x)$ whenever the dependence on n has to be made explicit). Then the disparity between the observed relative frequency vector r_n and the model probability vector f_θ generated by C is given by

$$\rho_C(r_n, f_\theta) \;=\; \sum_{x=0}^{\infty} C(\delta(x)) f_\theta(x). \tag{1.12}$$

The conditions mentioned in the above definition are called the disparity conditions. The function C is defined as the disparity generating function. Applying Jensen's theorem to the convex function C, we get

$$\begin{aligned} \sum C(\delta(x)) f_\theta(x) \;&\geq\; C\left(\sum \delta(x) f_\theta(x)\right) \\ &=\; C\left(E_\theta(\delta(X))\right) \\ &=\; C(0) = 0. \end{aligned}$$

This is the non-negativity property of the disparity function. Furthermore, by strict convexity, the equality holds if and only if when $f_\theta(x) = r_n(x) \; \forall \, x \in \chi$.

1.5.1.1.1 Specific Cases of Disparities: Specific forms of C lead to several well-known disparities, some of which have been mentioned earlier.

(i) If we consider $C(\delta) = (\delta+1)\log(\delta+1) - \delta$, then we get the likelihood disparity (LD) as

$$\begin{aligned} \mathrm{LD}(r_n, f_\theta) &= \sum \left[r_n \log\left(\frac{r_n}{f_\theta}\right) + (f_\theta - r_n) \right] \\ &= \sum r_n \log\left(\frac{r_n}{f_\theta}\right). \end{aligned}$$

(ii) The symmetric opposite of LD is the Kullback-Leibler divergence (KLD).

$$\begin{aligned} \mathrm{KLD}(r_n, f_\theta) &= \sum \left[f_\theta \log\left(\frac{f_\theta}{r_n}\right) + (r_n - f_\theta) \right] \\ &= \sum f_\theta \log\left(\frac{f_\theta}{r_n}\right). \end{aligned}$$

This corresponds to $C(\delta) = \delta - \log(\delta+1)$.

(iii) The (twice, squared) Hellinger distance (HD) will be generated for $C(\delta) = 2\left(\sqrt{\delta+1} - 1\right)^2$.

$$\mathrm{HD}(r_n, f_\theta) = 2\sum \left(\sqrt{r_n} - \sqrt{f_\theta}\right)^2.$$

(iv) If we take $C(\delta) = \frac{\delta^2}{2}$, then the generated divergence is Pearson's chi-square, which has the form

$$\mathrm{PCS}(r_n, f_\theta) = \sum \frac{(r_n - f_\theta)^2}{2 f_\theta}.$$

(v) If we take $C(\delta) = \frac{\delta^2}{2(\delta+1)}$, then the generated divergence is Neyman's chi-square (NCS), which has the form

$$\mathrm{NCS}(r_n, f_\theta) \;=\; \sum \frac{(r_n - f_\theta)^2}{2r_n}.$$

(vi) Another very important sub-family of disparities is the PD family, introduced by Cressie and Read (1984), through the expression

$$\mathrm{PD}_\lambda(r_n, f_\theta) \;=\; \frac{1}{\lambda(\lambda+1)} \sum r_n \left[\left(\frac{r_n}{f_\theta} \right)^{\lambda} - 1 \right].$$

This family, indexed by $\lambda \in \mathbb{R}$, has the disparity generating function

$$C(\delta) \;=\; \frac{(\delta+1)^{\lambda+1} - (\delta+1)}{\lambda(\lambda+1)} - \frac{\delta}{\lambda+1}.$$

For $\lambda = 1, -\frac{1}{2}$ and -2, this family coincides with the PCS, the HD and the NCS respectively, whereas, it generates the LD and the KLD whenever $\lambda \to 0$ and $\lambda \to -1$, respectively.

Some of these families will be discussed in details in the next chapter. For ease of presentation, we introduce this technique for discrete models and then for continuous models in the subsequent sections.

1.5.1.2 Minimum Disparity Estimation under Discrete Models

Under the discrete setup, the minimum distance estimator $\hat{\theta}$ of θ, based on the disparity ρ_c, will be defined as

$$\rho_c(r_n, f_{\hat{\theta}}) \;=\; \min_{\theta \in \Theta} \rho_c(r_n, f_\theta)$$

provided the minimum exists. Under certain regularity conditions, this estimator can be obtained by solving the estimating equation

$$-\nabla \rho_c(r_n, f_\theta) \;=\; \sum \Big(C^{'}(\delta)(\delta+1) - C(\delta)\Big) \nabla f_\theta = 0 \quad (1.13)$$

where ∇ represents the gradient with respect to θ and $C^{'}(\cdot)$ is the derivative of $C(\cdot)$.

Denoting $C^{'}(\delta)(\delta+1) - C(\delta)$ as $A(\delta)$, the estimating equation for θ would be of the form

$$-\nabla \rho_c(r_n, f_\theta) \;=\; \sum A(\delta) \nabla f_\theta = 0.$$

The function $A(\delta)$ may be standardized, without changing the estimating properties of the disparity, so that $A(0) = 0$ and $A'(0) = 1$. This standardized function is called the residual adjustment function (RAF) of the disparity. For the PD family, this function, indexed by the tuning parameter λ, equals

$$A_\lambda(\delta) \;=\; \frac{(\delta+1)^{\lambda+1} - 1}{\lambda+1}$$

with $A_\lambda(0) = 0$ and $A^{'}_\lambda(0) = 1 \ \forall \lambda \in \mathbb{R}$.

According to Basu et al. (2011), if a residual adjustment function satisfies the following property, then it can be considered as regular.

Definition 1.3. The residual adjustment function $A(\delta)$ will be called regular, if it is twice differentiable and $A^{'}(\delta)$ and $A^{''}(\delta)(1+\delta)$ are bounded on $[-1, \infty)$, where $A^{'}(\cdot)$ and $A^{''}(\cdot)$ represent the first and second order derivatives of $A(\cdot)$ with respect to its argument.

This RAF plays an important role in determining the robustness properties of the estimators. For an outlying observation, the Pearson residual (δ) has a high value, indicating that the observed proportion of that value is quite high compared to the proportion predicted by the model. This may be viewed as a probablistic outlier, where the mismatch between observed and predicted probabilities are highlighted in determining what is an outlier. This requires the specification of a parametric model, and this description contrasts the description of geometric outliers introduced in the previous subsection. Of course, the geometric and probabilistic concepts often coincide, but they need not. Now, the basic key of constructing a robust tool is to downweight outliers, which is made possible using only those estimators whose RAFs exhibit a dampened response to increasing δ. In this regard, the likelihood disparity has been considered as the benchmark. Its RAF is $A(\delta) = \delta$, which makes a 45° angle with the x-axis and increases linearly. The deviation of other estimators from this linearity will show us how stable an estimator is. If we consider the six common disparities mentioned earlier, then we can see that the NCS, the KLD and the HD possess concave RAFs dominated by the RAF of LD – indicating that the estimators based on these disparities have strong robustness properties; on the other hand, the PCS has a convex $A(\delta)$ – dominating the RAF of the LD and magnifying the effect of large Pearson residuals, thereby indicating that estimators based on the PCS are expected to be even worse than the MLE in terms of robustness.

The RAF, while being indicative of the general stability of the estimator, can also be used to study another robust tool of minimum disparity estimation, the influence function.

Theorem 1.4 (Lindsay (1994, Proposition 1))**.** *Under standard regularity conditions, the influence function of any minimum distance* *functional T with estimating equation $\sum A(\delta(x))\nabla f_\theta(x) = 0$ has the following expression*

$$IF(y, T, G) \;=\; J_g^{-1}\left\{A'(\delta(y))u_{\theta^g}(y) - E_g\left[A'(\delta(x))u_{\theta^g}(x)\right]\right\}$$

where, $\delta(x) = \frac{g(x)}{f_{\theta^g}(x)} - 1$. Also, we have

$$J_g \;=\; E_g\left[u_{\theta^g}(X)u_{\theta^g}^T(X)A'(\delta(X))\right] - \sum_x A(\delta(x))\nabla_2 f_{\theta^g}(x)$$

with θ^g being the best-fitting parameter.

In fact, the asymptotic distribution of the minimum distance estimators (MDE) also involves the RAF, $A(\delta)$. Under certain regularity conditions, the MDEs have the following asymptotic properties:

(i) There exists a sequence of consistent roots, $\hat{\theta}_n$, of the estimating Equation (1.13).

(ii) Moreover, at $\theta = \theta^g$,

$$\sqrt{n}\left(\hat{\theta}_n - \theta^g\right) \xrightarrow{a} N\left(0, J_g^{-1}V_gJ_g^{-1}\right),$$

where, V_g is defined as $V_g = Var_g\left[A'(\delta(X))u_{\theta^g}(X)\right]$,

When $G = F_\theta$, $\delta(x)$ is identically zero, and the asymptotic variance will be $I^{-1}(\theta)$, which is the asymptotic variance of the most efficient estimator, the MLE. Hence, when the model is true, the MDEs have full asymptotic efficiency. So when the model holds and the sample size is large, we would get robust estimators having efficiencies that are converging to that of the MLE, whereas in small to moderate samples, our mission is to find stable estimators compromising efficiency as less as possible. Moreover, under $G = F_\theta$, we get $\theta^g = \theta$,

$\delta(x) = 0$ for all x and hence the MDE corresponding to the estimating equation $\sum A(\delta(x))\nabla f_\theta(x) = 0$ has influence function given by $IF(y, T, G) = I^{-1}(\theta)u_\theta(y)$.

1.5.1.3 Minimum Distance Estimation under Continuous Model

Here, we will describe the technique under continuous models. Let $\mathcal{G}$ represent the class of all distributions having densities with respect to the Lebesgue measure. Moreover, the true data distribution G (with density g) and model family $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ (with density f_θ) belong to $\mathcal{G}$.

Suppose $X_1, \ldots, X_n$ are n i.i.d. observations from G and we want to find the minimum disparity estimate for the parameter θ. In case of discrete models, the conventional estimate of g is the vector of relative frequencies, but in case of continuous models, one needs to construct a continuous density estimate through kernel density estimation or other smoothing techniques so that there is no incompatibility of measures when constructing the disparity.

This branch of robust minimum distance estimation probably originated with the seminal work of Beran (1977), and the approach has been widely used in the subsequent literature. With the addition of this smoothing component the procedure has an additional level of theoretical complexity and the bandwidth selection becomes an issue. In this context, we propose to use the suggestion of Basu and Lindsay (1994), which helps to overcome the problems related to the slow convergence of the kernel and makes the bandwidth selection problem a less critical one.

For estimating the unknown parameter θ through a minimum divergence procedure in this setup, we now describe the following two approaches available to us.

1.5.1.3.1 Beran's Approach In this approach, one employs the kernel density estimation procedure for estimating the data density g by

$$g_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} W(x, X_i, h) = \int W(x, y, h)\, dG_n(y)$$

where W is some smooth kernel function, h is the bandwidth, G_n is the empirical distribution of G and X_i's are the given sample observations. The kernel function W is usually a symmetric density like the Epanechnikov or the Gaussian. The minimum distance estimate of θ is obtained through the minimization of $\rho(g_n^*, f_\theta)$ over $\theta \in \Theta$, where $\rho(\cdot, \cdot)$ is a generic divergence and the Pearson residual is defined as $\delta(x) = \frac{g_n^*(x)}{f(x)} - 1$.

Under suitable assumptions and differentiability of the model, the MDE will be obtained through the equation

$$-\nabla\rho_c(g_n^*, f_\theta) = \int_x A(\delta(x))\nabla f_\theta(x)dx = 0,$$

where the RAF is as defined earlier. The estimation procedure then proceeds as in the discrete case – the interpretation of the RAF $(A(\delta))$ and the disparity generating function $(C(\delta))$ will remain unaltered.

However, the kernel smoothing introduces a bias in the density estimate, which has to be asymptotically corrected by choosing the bandwidth $h = h_n$ to be a function of the sample size and letting it slide to zero at the appropriate rate with increasing n. This smoothing component, which is an intermediate step in our estimation scheme, adds an additional layer of theoretical complexity to the procedure, as the choice of the bandwidth now becomes crucial. Park and Basu (2004) have proved the existence and consistency of the MDE under assumptions on the model and $C(\delta)$, and they have further shown the asymptotic distribution of the MDE, which is,

$$\sqrt{n}\,(T(G_n) - \theta_0) \xrightarrow{\text{D}} N(0, I^{-1}(\theta_0)),$$

when $G = F_{\theta_0}$. See Park and Basu (2004) for further details.

1.5.1.3.2 Basu-Lindsay Approach Due to the complexity involved in Beran's approach, some appropriate modification is needed to simplify the estimation process. Here, we are going to refer to one such method, which follows from the work of Basu (1991) and Basu and Lindsay (1994).

This method differs from Beran's approach in that while Beran only took a non-parametric kernel density estimate g_n^* of g from the data, Basu and Lindsay also convoluted the model density with the same kernel. It basically suggests smoothing the model density and the true density with the same kernel function and the same bandwidth. The rationale behind this proposal is that the convolution of the model with the same kernel compensates for the bias due to the use of the kernel on the data, by imposing the same distortion on the model. Hence, the importance of the kernel has been diminished in this estimation procedure as compared to Beran's approach.

For a suitable kernel $W(x, y, h)$, suppose f_θ^* is the kernel-smoothed version of model density f_θ, which equals

$$f_\theta^*(x) \;=\; \int W(x, y, h)\, dF_\theta(y) = \int W(x, y, h)\, f_\theta(y)\, dy.$$

Here, we will find the estimator by minimizing the distance between g_n^* and f_θ^*, namely $\rho(g_n^*, f_\theta^*)$, which is defined as

$$\rho(g_n^*, f_\theta^*) \;=\; \int C(\delta^*(x)) f_\theta^*(x)\, dx$$

with the modified Pearson residual $\delta^*(x) = \frac{g_n^*(x)}{f_\theta^*(x)} - 1$, that is, a residual between the smoothed data and the smoothed model. In this case, the minimum distance estimator is obtained as the minimizer of $\rho(g_n^*, f_\theta^*)$. In our technical conditions, it will be assumed that $f_\theta(x) > 0$ for all x in the sample space, so the Pearson residual $\delta^*(x)$ is well defined. Since the kernel g_n^* converges to the smoothed model f_θ^* pointwise for any fixed bandwidth h, we have fixed h consistency in this approach. Thus, while a bandwidth selection will still be necessary in this case, it is certainly not as critical as it is under Beran's approach. Consistency, in any case, is not dependent on the choice of the bandwidth. See Basu and Lindsay (1994) for more details on this estimation scheme.

Under the assumption of differentiability of the model, the smoothed MDE can be obtained by solving

$$-\nabla \rho_c(g_n^*, f_\theta^*) \;=\; \int_x A(\delta^*(x)) \nabla f_\theta^*(x) dx = 0.$$

To find this element under this approach, we first define the first and second order derivatives of the log-likelihood of the smoothed version of the model density as

$$\begin{aligned} \widetilde{u}_\theta(x) &= \nabla \log f_\theta^*(x), \\ \nabla \widetilde{u}_\theta(x) &= \nabla^2 \log f_\theta^*(x), \end{aligned}$$

where the j-th element of the first order derivative and the (j,k)-th element of the second order derivative are denoted by $\widetilde{u}_{j\theta}(x)$ and $\widetilde{u}_{jk\theta}(x)$ respectively. The corresponding smoothed score function is defined as $u_\theta^*(y) = \int \widetilde{u}_\theta(x) W(x,y,h)dx$ with zero expectation with respect to f_θ.

We will denote the minimizer of the likelihood disparity between g_n^* and f_θ^* as the MLE*. This MLE* is, in general, distinct from the ordinary MLE, and is not automatically first order efficient. Through the imposition of some specific conditions on the kernel, the estimating equations of these two estimators, MLE and MLE*, would become equal and hence the estimators, too. More specifically, under the smoothing technique through such kind of a kernel, all MDEs that are asymptotically equivalent to MLE* will become first order efficient. In fact, when model becomes true, the IF as well as the asymptotic distribution of these MDEs obtained through the Basu-Lindsay approach (along with the implementation of such kernel) become identical with those of the MDEs obtained without smoothing at $g = f_\theta$ in case of the discrete model. These type of kernels are called transparent kernels (Definition 1.5) relative to the model. See Basu and Lindsay (1994) and Basu et al. (2011) for further details.

Definition 1.5. A kernel $W(x,y,h)$ is called a transparent kernel for the parametric model family $\mathcal{F} = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ if

$$A u_\theta(x) + B = u_\theta^*(x) \tag{1.14}$$

where A is a non-singular $p \times p$ matrix which may depend on θ and B is a p-dimensional vector. Since $E(u_\theta(X)) = E(u^*_\theta(X)) = 0$ under f_θ, we get the simplified form of Equation (1.14) and that is

$$Au_\theta(x) \;=\; u^*_\theta(x). \tag{1.15}$$

1.5.2 Hypothesis Testing using Disparities

In the field of hypothesis testing, one obvious tool is the likelihood ratio test mentioned earlier, which is one of the oldest techniques in the statistical literature. But, this test has shown acute sensitivity towards model mis-specification and presence of outliers. This is true for the Wald and score tests (based on the maximum likelihood estimator) also. An alternative could be robust tests based on disparities. The likelihood ratio test (LRT) can also be seen to belong to a larger class of 'disparity difference'-type tests – as a result, on one hand, this class contains the LRT which has several asymptotic optimality results and on the other hand, many members of this class possess strong robustness properties.

1.5.2.1 Testing of Hypothesis under the Discrete Model

Suppose $X_1, \ldots, X_n$ be n i.i.d. sample observations from a true distribution G where the model family is $\mathcal{F} = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$. Consider the hypotheses

$$H_0 : \theta \in \Theta_0 \subset \Theta \text{ vs. } H_1 : \theta \in \Theta \setminus \Theta_0.$$

The test is constructed using $\hat{\theta}$, the unconstrained MDE under $H_0 \cup H_1$ and θ_0, provided Θ_0 is a singleton test, otherwise $\hat{\theta}_0$, the constrained MDE under H_0 is used. Therefore, the test statistic is given by

$$DDT_{\rho_c}(g, f) = \begin{cases} 2n \left[\rho_c(g, f_{\hat{\theta}_0}) - \rho_c(g, f_{\hat{\theta}})\right], & \text{if } \Theta_0 \text{ is composite.} \\ 2n \left[\rho_c(g, f_{\theta_0}) - \rho_c(g, f_{\hat{\theta}})\right], & \text{if } \Theta_0 = \{\theta_0\}. \end{cases}$$

Furthermore, since the model is discrete and g is unknown we are going to use the conventional estimate of $g(x)$, that is, relative frequency at x, $r_n(x)$ in place of $g(x)$ to perform this test.

Under suitable assumptions (C1 - C4 of Basu et al. (2011), Chapter 5),

$$DDT_{\rho_c}(r_n, f_\theta) \overset{\text{a}}{\sim} \chi^2_r,$$

where r is the number of restrictions imposed by H_0 and $\overset{\text{a}}{\sim}$ denotes asymptotic distribution. Next, we explore the structure of the disparity difference test with respect to its robustness and hence, we will consider the contaminated model from now onward. Suppose $T(G)$ be the minimum distance functional corresponding to ρ_c and hence, $\hat{\theta}_n$ is the MDE after considering $\hat{g}(x) = r_n(x)$. For the true distribution G, let $H_0 : T(G) = \theta_0$.

Let g be the true density. In this case, the test statistic is given by

$$DDT_{\rho_c}(r_n, f_\theta) \;=\; 2n \left[\rho_c(r_n, f_{\theta_0}) - \rho_c(r_n, f_{\hat{\theta}_n})\right].$$

According to Lindsay (1994), when θ is scalar, under given conditions,

$$DDT_{\rho_c}(r_n, f_\theta) \overset{\text{a}}{\sim} C(g)\chi^2_1, \text{ under } H_0,$$

where $C(g) = Var_g\left[T^{'}(X)\right] \nabla_2 \rho_c(g, f_\theta)\Big|_{\theta=\theta_0}$. $T^{'}(X)$ is the IF of the MDE obtained through ρ_c. This $C(g)$ is called the chi-square inflation factor. For further details, see Lindsay (1994) and Basu et al. (2011).

1.5.2.2 Testing of Hypothesis under the Continuous Model

In the continuous case, a general approach for minimum disparity estimation and testing of hypothesis is much more difficult than the discrete case; however, see Park and Basu (2004) and Kuchibhotla and Basu (2015, 2017). Several authors have constructed robust tests on the basis of some specific distance and in this regard mention should be made of the work of Simpson (1989) with the Hellinger distance under continuous models.

We consider the setup of the continuous model given in Subsection 1.5.1.3. Let $H_0 : \theta \in \Theta_0$ be the null hypothesis of interest and we want to test it against $H_1 : \theta \in \Theta \backslash \Theta_0$. Since the model is continuous, we are to use the non-parametric density estimate g_n^* based on the kernel density estimate. If we proceed in the same way as mentioned earlier, we define the disparity difference test statistic as

$$DDT_{\rho_c}(r_n, f_\theta) \;=\; 2n\left[\rho_c(g_n^*, f_{\hat{\theta}_0}) - \rho_c(g_n^*, f_{\hat{\theta}})\right].$$

Here also, under suitable conditions, this statistic asymptotically follows χ^2_r under H_0. Here the estimators and the notation are as described in Section 1.5.2.1. See Simpson (1989).

For the Basu and Lindsay approach the distribution of the test statistic constructed in a similar spirit is a little more complicated. However, see Basu (1993), Agostinelli and Markatou (2001) and Basu et al. (2011).

1.6 The Need for the Optimal Parameter Selection

Almost all robust procedures including M-estimators and minimum distance estimators are dependent on the choice of one or more tuning parameters. These tuning parameters have a vital role in determining the trade-off between efficiency and robustness of the procedure. For example, when estimating the location parameter of the model using the Huber loss function in M-estimation, the choice of the tuning parameter c which determines the boundary between the quadratic and linear parts of the objective function, has a major role in the description of the M-estimator. As c runs away to infinity, the estimator settles on the mean; on the other hand, as c gets smaller and smaller, the estimator approaches the median. See Maronna et al. (2006), Section 2.2.2, for more details. Thus by varying the choice of the tuning parameter one can choose between a highly efficient but sensitive estimator and a strongly resistant but inefficient estimator. As we do not know apriori how much anomaly is present in the data and how much downweighting is necessary, a data based estimate of the tuning parameter which can moderate estimator depending on the necessity may be of great value.

While data-based estimates are not absent in the literature, see, e.g., Wang et al. (2007), the more general approach has been to choose tuning parameters specific to the model which retain a fixed proportion of the efficiency of the classical methods. The default value for c for Huber's ψ function in R-packages is 1.345 (rlm function), which achieves about 95% efficiency when the data are normally distributed. Other specific choices have been suggested by other researchers. These choices generally give good compromises between efficiency and robustness, but are never optimal as they do not make an attempt to determine the amount of anomaly in the data and choose the tuning parameter accordingly.

In the area of robust minimum distance estimation, some attempts have been made to select tuning parameters in an optimal manner, most notably with respect to the density power divergence. See, e.g., Hong and Kim (2001) and Warwick and Jones (2005). In case of minimum distance estimation, many variants of these methods have been tried in the literature, but the basic approach revolves around the Warwick and Jones (2005) idea of constructing an empirical estimate of the mean square error as a function of the tuning parameter (and a suitable "pilot" estimator). Subsequently one minimizes this mean square error over the tuning parameter to obtain a data-based "optimal" estimate of the tuning parameter specific to the data set. The overall technique has worked reasonably in many different situations, but, unfortunately, it remains the function of a suitable robust "pilot" estimator, and sometimes this dependence is quite acute. Therefore, in the present dissertation, It will also be our endeavor to select an optimal robust tuning parameter, which provides the best compromise in the efficiency-robustness trade off and removes the dependence on any particular pilot estimator.

1.7 Aim and Layout of the Thesis

The minimization of suitable statistical distances (between the data and the model densities) has proved to be a very useful technique in the field of robust inference. Here, the estimation of the parameter is derived in the context of closeness between the data and the model. This is the parameter of the model density which minimizes the discrepancy between the data density and the class of model densities. Several authors produced different robust estimators using several divergences through this approach. At the same time, some density-based divergences lead to estimators with high asymptotic efficiency, sometimes full asymptotic efficiency. Emphasizing the pattern of down-weighting the score function with the power of the density in the presence of outliers, Basu et al. (1998) developed the density power divergence (DPD) class which is obtained through a generalization of the estimating equation of the MLE and the minimum L_2 distance estimator. We will have a look at all these popular divergences in Chapter 2 of this thesis. All these divergences can be expressed as special cases of the Bregman divergence and hence, their characteristics can be simply proved from the properties of the ordinary Bregman divergence. However, since the data density must have a linear presence in the cross-product term of the ordinary Bregman form, several useful divergences cannot be captured through it. In this regard, we must mention one of the most prominent class of disparities, namely the Power Divergence (PD) family, introduced by Cressie and Read (1984). So, to bring most of the popular divergence families under one larger class, a modification of the Bregman divergence is required. Although it is not the first attempt in this direction, our belief is that this extension explores and adds something new in the field of robust analysis through a very simple and intuitive trick.

Another important goal of this research is to develop some method for selecting the tuning parameter optimally. The DPD, indexed by a single tuning parameter α, has been used as the basic tool for introduction and demonstration purposes. Larger values of α necessarily lead to a drop in the model efficiency and gain in stability, while the opposite scenario can be observed for small values of α. In Chapter 3, a refinement of an existing technique with the aim of eliminating the pilot dependency and the scope of discovering the optimal value of α to provide the best compromise between model efficiency and stability against data contamination, is proposed. Moreover, to extend the scope of its application, it is our target to apply this procedure to different scenarios – for example, i.i.d. sample, non-homogeneous independent samples, multiple linear regression models, etc. Our intention is to rigorously setup the theoretical result, along with its substantiation, through numerical applications. Finally we hope to study and show its successful use within the framework of the proposed extension.

In Chapter 4, we provide an extension of the ordinary Bregman divergence by considering an exponent of the density as the argument rather than the density function itself. It has made the class of the Bregman divergence wider and thus, it contains most of the popular density-based divergences as well as the class of disparities. A detailed description of this extension is discussed throughout this chapter.

We already know that two prominent classes of divergences in this area of statistics are the PD and the DPD families, but only the likelihood disparity is the common element between them. To connect them with each other, Ghosh et al. (2017) introduced a new divergence family, called the S-divergence family. This family also could not be expressed through the ordinary Bregman divergence, but now it becomes possible through this extension. To create a broader class of divergence families, our target is to join this S-divergence family (and hence, the PD and the DPD families) with another important member of the Bregman class, namely the B-Exponential divergence (BED) family and hence, a specific form of our proposed extension is required. In Chapter 5, through this specific form, a new super-family is introduced, namely the GSB divergence family. Another significant discovery through this family is that we can generate robust and remarkably efficient estimates which lie outside the PD, the DPD and even the S-divergence family. Along with this introductory part, we have explored its performance in the field of robust estimation under the discrete model, while in the next chapter we have extended our journey through the same kind of inference under the continuous model.

Having described the applications of our method in the field of estimation, the obvious/immediate next step is to explore the usage of this extension in the field of hypotheses testing. For demonstration purposes, another important large family of divergences, the Generalized Super divergence (GSD), introduced by Ghosh and Basu (2018), will play the same role in the field of testing, just like the GSB divergence in the field of estimation. With a slight modification, this divergence can also be brought under the umbrella of the extension – through the convex combination of the two extended Bregman forms, with specific choices of the convex combination. In Chapter 7, the disparity difference test based on this divergence and its performance under several scenarios have been analyzed through simulation studies and real life data examples. So, all these chapters have been framed to explore the usefulness of the extended Bregman divergence in the field of parametric inference.

Chapter 2

A Useful Divergence : The Bregman Divergence

2.1 Definition

Being motivated by the problem of convex programming, L. M. Bregman (1967) introduced the Bregman divergence, a measure of dissimilarity between any two vectors in the Euclidean space. In $\mathbb{R}^p$, it has the form

$$D_\psi\left(\boldsymbol{x}, \boldsymbol{y}\right) = \left\{ \psi\left(\boldsymbol{x}\right) - \psi\left(\boldsymbol{y}\right) - \left\langle \nabla\psi\left(\boldsymbol{y}\right), \boldsymbol{x} - \boldsymbol{y} \right\rangle \right\}, \qquad (2.1)$$

for any strictly convex function $\psi : \mathcal{S} \to \mathbb{R}$ and for any two p-dimensional vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$, where $\mathcal{S}$ is a convex subset of $\mathbb{R}^p$. Here, $\nabla\psi\left(\boldsymbol{y}\right)$ denotes the gradient of ψ with respect to its argument at $\boldsymbol{y} = \left(y_1, y_2, \ldots, y_p\right)^T$.

2.1.1 Use

Although this divergence has been introduced with the motivation of applying it in convex analysis, more specifically, to find the common points of convex sets and their application in convex programming, later it has been widely used for several mathematical as well as statistical purposes. Both quantization and clustering problems have been developed through the application of this divergence. In both hard and soft clustering, this divergence have been used by Banerjee et al. (2005). Moreover the bijection between this divergence and regular exponential families helps a lot in this regard. Further its mathematical implementation includes its connection with rank aggregation, web ranking, matrix nearness problems, etc. Bregman metric is one of the useful metrics we generally use in case of consideration of metric spaces. On the other hand, if we look at its use in statistics, the first thing that appears in our mind is its application in minimum distance estimation in different fields including robust inference. For minimization of suitable statistical distances, apart from the class of ϕ-divergences of Csiszár (1963) and Ali and Silvey (1966), the Bregman divergence has been extensively used. Its power to retain the robustness in the presence of outliers leads us to generate M-estimators, mentioned earlier. And, last but not the least, it has successful applications in the field of Bayesian inference also.

2.1.2 Properties

There are several advantageous properties of this well-known divergence. Some of them are mentioned below:

1. The convexity criterion of the function $\psi(\cdot)$ evidently leads to the non-negativity of this divergence.

2. It is convex in $\boldsymbol{y}$.

3. It is linear in the convex function ψ.

4. It is invariant under affine transformations.

5. Another important property of this divergence is centering. The mean vector is the minimizer of the expected Bregman divergence from any random vector. It reminds us of the fact that the mean of a set is the minimizer of squared error of elements of that set.

2.1.3 Some Useful Divergences as Special Cases of the Bregman Divergence

The Bregman divergence has significant applications in the domain of statistical inference for both discrete and continuous models. Given two densities g and f, the Bregman divergence between these densities (associated with the convex function ψ) is given by

$$D_{\psi}(g,f) = \int \left\{ \psi(g(x)) - \psi(f(x)) - (g(x) - f(x)) \nabla\psi(f(x)) \right\} dx. \tag{2.2}$$

It is easy to see that the function $\psi(y)$ and $\psi(y) + ay + b$ generate the same divergence, where a and b are finite real numbers. With specific choices of the convex function $\psi(\cdot)$ we can express several useful divergences as a part of this divergence class. Some examples are shown below.

2.1.3.1 Likelihood Disparity (LD)

This popular divergence can be expressed as a subfamily of the class of Bregman divergences by choosing $\psi\left(x\right) = x \log\left(x\right)$. Given a class of model densities $\{f_\theta\}$ and the data density g, this divergence has the form

$$\mathrm{LD}(g, f_\theta) = d_0\left(g, f_\theta\right) = \int g\left(x\right) \log\left(\frac{g\left(x\right)}{f_\theta\left(x\right)}\right) dx. \tag{2.3}$$

The minimum LD functional, $T_0\left(G\right)$, can be defined through the relation

$$d_0\left(g, f_{T_0(G)}\right) = \min_{\theta\in\Theta} d_0\left(g, f_\theta\right), \tag{2.4}$$

provided the minimum exists. Through the consideration of the empirical version of Equation (2.3) based on the random sample $X_1, X_2, \ldots, X_n$, we can obtain the minimum LD estimator of θ by minimizing $\sum_{i=1}^{n}\left(-1\right)\log f_\theta\left(X_i\right)$, i.e., by maximizing $\sum_{i=1}^{n}\log f_\theta\left(X_i\right)$ over $\theta \in \Theta$. But this objective function is just the log-likelihood given the data. Moreover, if we proceed one step ahead, we get the corresponding estimating equation.

$$\sum_{i=1}^{n} u_\theta\left(X_i\right) = 0, \tag{2.5}$$

where $u_\theta\left(x\right)$ denotes the score function at x. Thus the divergence in (2.3) leads to the maximum likelihood functional (discussed in Chapter 1) whenever minimized over the whole parameter space, Θ. This divergence is a version of the Kullback-Leibler divergence, as already observed earlier.

2.1.3.2 Squared L_2 Distance

The (squared) L_2 distance between the densities g and f_θ can be generated from this Bregman divergence by considering $\psi(x) = x^2$ which generates the form

$$L_2(g, f_\theta) = d_1(g, f_\theta) = \int (g(x) - f_\theta(x))^2 \, dx. \tag{2.6}$$

The minimum L_2 distance functional, $T_1(G)$, will be defined as

$$d_1\left(g, f_{T_1(G)}\right) = \min_{\theta \in \Theta} d_1(g, f_\theta), \tag{2.7}$$

provided the minimizer exists. After expanding Equation (2.6) and omitting the term involving g only (since that term has no role in the optimization), we get the revised objective function

$$\int f_\theta^2(x) \, dx - 2 \int f_\theta(x) \, g(x) \, dx \tag{2.8}$$

Next, replacing $g(\cdot)$ by $dG_n(\cdot)$ (G_n being the empirical version of G), we get the equation,

$$\int f_\theta^2(x) \, dx - 2 \int f_\theta(x) \, dG_n(x) = \int f_\theta^2(x) \, dx - 2\frac{1}{n} \sum_{i=1}^{n} f_\theta(X_i), \tag{2.9}$$

which we can actually minimize to obtain the minimum L_2 distance estimator of θ. Furthermore, an immediate calculation leads us to the estimating equation having the form

$$\frac{1}{n} \sum_{i=1}^{n} f_\theta(X_i) \, u_\theta(X_i) - \int f_\theta^2(x) \, u_\theta(x) \, dx = 0. \tag{2.10}$$

Through the presence of the weight $f_\theta(x)$ in Equation (2.10), the strong robustness property of the minimum L_2 distance estimator is quite evident.

2.1.3.3 Density Power Divergence (DPD)

If we observe the estimating equations (2.5) and (2.10) of the two previous cases, it may be easily seen that they both represent special cases of an extended general case. If we consider a general weighted likelihood equation having the form

$$\frac{1}{n}\sum_{i=1}^{n} f_\theta^\alpha(X_i)\, u_\theta(X_i) - \int f_\theta^{1+\alpha}(x)\, u_\theta(x)\, dx = 0, \tag{2.11}$$

indexed by tuning parameter α, then the choice of $\alpha = 0$ generates Equation (2.5), while $\alpha = 1$ recovers Equation (2.10).

Equation (2.11) is an unbiased estimating equation at the model corresponding to the well-known 'Density Power Divergence' (DPD) family. This family is generated by the function $\psi(x) = \frac{x^{\alpha+1}-x}{\alpha}$, applied to the general Bregman form in Equation (2.2) and is indexed by a non-negative tuning parameter α. As a function of α, the density power divergence may be expressed as

$$d_\alpha(g, f_\theta) = \int \left\{ f_\theta^{\alpha+1}(x) - \left(1 + \frac{1}{\alpha}\right) g(x)\, f_\theta^\alpha(x) + \frac{1}{\alpha} g^{\alpha+1}(x) \right\} dx \tag{2.12}$$

Therefore the required objective function for finding the minimum

DPD estimator (MDPDE), obtained by ignoring the term independent of θ and by replacing dG with dG_n, becomes

$$\begin{aligned}
&\int f_\theta^{\alpha+1}(x)\,dx - \left(1+\frac{1}{\alpha}\right)\int f_\theta^{\alpha}(x)\,dG_n(x) \\
= &\int f_\theta^{\alpha+1}(x)\,dx - \left(1+\frac{1}{\alpha}\right)\frac{1}{n}\sum_{i=1}^{n} f_\theta^{\alpha}(X_i) \\
= &\frac{1}{n}\sum_{i=1}^{n} V_\theta(X_i) \qquad (2.13)
\end{aligned}$$

where,

$$V_\theta(X_i) = \int f_\theta^{\alpha+1}(x)\,dx - \left(1+\frac{1}{\alpha}\right) f_\theta^{\alpha}(X_i)\,.$$

Since the true density g shows up linearly in the objective function, it has been replaced by dG_n, and thus this minimization procedure does not need any non-parametric smoothing. Moreover, it is quite clear that the derived MDPDE can be treated as an M-estimator (with a model-dependent ψ function) and hence its asymptotic properties can be developed from the theory of M-estimators too. Here, at the end of this subsection, we will mention its asymptotic properties for future reference.

The minimum DPD functional, $T_\alpha(G)$, can be defined as

$$d_\alpha\left(g, f_{T_\alpha(G)}\right) = \min_{\theta\in\Theta} d_\alpha(g, f_\theta)\,, \qquad (2.14)$$

provided the minimum exists. Moreover, this DPD family with $\alpha = 0$ coincides with the likelihood disparity (in the limiting sense) whereas, at $\alpha = 1$, it generates the L_2 distance between two densities. When α assumes its minimum value zero, the estimating equation has no density power downweighting, so that the corresponding estimator has weak stability and poor robustness properties. On the other hand, for large positive α, the power of the density helps to diminish the influence of the score function in the presence of contamination. From practical considerations, the value of α is restricted to lie within $[0, 1]$, since even larger values of α make the efficiency unacceptably low, and with a judicious choice we expect to get reasonable trade-offs between efficiency and robustness in this range.

Under certain regularity conditions given in Basu et al. (1998),

$$\sqrt{n}\left(\hat{\theta}_\alpha - \theta_\alpha^g\right) \to Z \sim N_p\left(0, J_\alpha^{-1}\left(\theta_\alpha^g\right) K_\alpha\left(\theta_\alpha^g\right) J_\alpha^{-1}\left(\theta_\alpha^g\right)\right), \qquad (2.15)$$

where $\hat{\theta}_\alpha$ and θ_α^g denote the MDPDE and the best fitting parameter corresponding to a pre-fixed α, and,

$$\begin{aligned} J_\alpha(\theta) &= \int u_\theta(x) u_\theta^T(x) f_\theta^{1+\alpha}(x)\, dx \\ &+ \int \{i_\theta(x) - \alpha u_\theta(x) u_\theta^T(x)\}\{g(x) - f_\theta(x)\} f_\theta^\alpha(x)\, dx, \end{aligned} \qquad (2.16)$$

$$K_\alpha(\theta) = \int u_\theta(x) u_\theta^T(x) f_\theta^{2\alpha}(x) g(x)\, dx - \psi_\theta \psi_\theta^T, \qquad (2.17)$$

where $\psi_\theta = \int u_\theta(x) f_\theta^\alpha(x) g(x)\, dx$, $i_\theta(x) = -\frac{d}{d\theta}(u_\theta(x))$, and the superscript T represents ‘transpose’.

2.1.3.4 B-Exponential Divergence (BED)

This divergence, introduced by Mukherjee et al. (2019), provides another subfamily of Bregman divergences through the function $\psi(x) = \frac{2\left(e^{\beta x} - \beta x - 1\right)}{\beta^2}$ for $\beta \in \mathbb{R}$. The form of this divergence is given by

$$d_\beta(g, f_\theta) = \frac{2}{\beta} \int \left\{ e^{\beta f_\theta(x)} \left(f_\theta(x) - \frac{1}{\beta} \right) - e^{\beta f_\theta(x)} g(x) + \frac{e^{\beta g(x)}}{\beta} \right\} dx.$$

The minimum BED functional, $T_\beta\left(G\right)$, can be defined as

$$d_\beta\left(g, f_{T_\beta(G)}\right) = \min_{\theta\in\Theta} d_\beta\left(g, f_\theta\right), \tag{2.18}$$

provided the minimum exists. The required objective function to derive the minimum BED estimator, obtained by ignoring the term independent of θ and by replacing dG with dG_n, becomes

$$\begin{aligned} & \int e^{\beta f_\theta(x)}\left(f_\theta\left(x\right)dx - \frac{1}{\beta}\right)dx - \int e^{\beta f_\theta(x)} dG_n\left(x\right) \\ = & \int e^{\beta f_\theta(x)}\left(f_\theta\left(x\right) - \frac{1}{\beta}\right)dx - \frac{1}{n}\sum_{i=1}^{n} e^{\beta f_\theta(X_i)} \\ = & \frac{1}{n}\sum_{i=1}^{n}\left\{\int e^{\beta f_\theta(x)}\left(f_\theta\left(x\right) - \frac{1}{\beta}\right)dx - e^{\beta f_\theta(X_i)}\right\}, \end{aligned} \tag{2.19}$$

over $\theta \in \Theta$. Evidently the expression in Equation (2.19) is obtained without any non-parametric smoothing component. Now minimization of Equation (2.19) produces an estimating equation of the form

$$\frac{1}{n}\sum_{i=1}^{n} f_\theta\left(X_i\right)u_\theta\left(X_i\right)e^{\beta f_\theta(X_i)} - \int f_\theta^2\left(x\right)u_\theta\left(x\right)e^{\beta f_\theta(x)}dx = 0. \tag{2.20}$$

Clearly this is also a weighted likelihood estimating equation, with the weight being $f_\theta\left(x\right)e^{\beta f_\theta(x)}$. From this weight, the outlier stability characteristic of the MBEDE is quite evident.

Moreover, the estimating equation is of the form $\sum_{i=1}^{n}\phi_\theta(X_i) = 0$, where,

$$\phi_\theta(X_i) = f_\theta\left(X_i\right)u_\theta\left(X_i\right)e^{\beta f_\theta(X_i)} - \int f_\theta^2\left(x\right)u_\theta\left(x\right)e^{\beta f_\theta(x)}dx. \tag{2.21}$$

Hence the MBEDE is also an M-estimator. Therefore, as in the case of the MDPDE, the asymptotic properties of this MBEDE also can be directly derived from the properties of M-estimators.

2.1.4 Concluding Remarks

Inappropriate selection of tuning parameter(s) can mislead our analysis and lead to incorrect insight generation. While some data-based techniques lead to reasonable selection of the tuning parameter, pilot dependency is an inherent feature of these methods which lead to different optimals for different pilots (for the same data/method of estimation). Hence, in the next chapter, we start our journey through giving our best effort to solve this issue. At first we will introduce a refined algorithm, more specifically, modify an existing algorithm so that it will lead us to find pilot-independent tuning parameter(s) generating robust estimators. The DPD family has been used as the basic illustrative tool for this purpose. Later this modification will be employed over the extension of the Bregman divergence for better analysis.

Chapter 3

Choosing the 'Optimal' Tuning Parameter

3.1 Introduction

In statistical inference, the two concepts of efficiency and robustness are often at odds, and it is a delicate task for the statistician to balance them both in a suitable manner. While the issue of robustness is a real concern in the present age of big data, this robustness should not come at the cost of a high efficiency loss at the model. Most robust procedures require the choice of a tuning parameter, which determines the trade-off between robustness and efficiency. In a real problem, selecting this tuning parameter 'optimally' based on the given data is an issue of great practical interest, which can protect the experimenter/statistician in both eventualities.

While our method is general, for the purpose of illustration we will concentrate on the density power divergence (DPD) family discussed earlier, which has had a major impact in the area of minimum distance estimation over the last two decades. There have been a few attempts to select the robustness tuning parameter of this family; in particular, the method proposed by Warwick and Jones (2005) has seen substantial application in subsequent data analysis problems, as have others; see, for example, Ghosh and Basu (2013), Park and Sriram (2017) and Kang and Lee (2014). In fact Jane Warwick wrote an entire PhD thesis on this topic, although the part relevant for us is more or less covered in the Warwick and Jones (2005) paper. The method described in Hong and Kim (2001) is also useful in this respect. We will refer to the method in Warwick and Jones (2005) as the Warwick and Jones method (the WJ method for short), and the method in Hong and Kim (2001) as the Hong and Kim method (the HK method for short). However, in some sense, these methods are not completely satisfactory, as the first one depends, sometimes quite heavily, on the choice of a pilot estimator and the second one sometimes leads to very non-robust estimators.

Here, we make a proposal which, by refining the approach in Warwick and Jones (2005), attempts to remove both of these deficiencies. This may eliminate the most important drawback in the application of the density power divergence estimator and make it more universally acceptable. However, as already mentioned, while we illustrate it with the DPD, the applicability of our method is not limited to this divergence alone. It can be applied to all methods which depend upon the choice of one or more variable tuning parameter(s). We will, in fact, apply it on the extended Bregman divergence in the later chapters.

3.1.1 The Basic Idea in Tuning Parameter Selection

As α tends to 0, the DPD $d_\alpha(g, f)$ converges to the Kullback-Leibler divergence $d_0(g, f)$, and given a sequence of i.i.d. observations $X_1, X_2, \ldots, X_n$, the corresponding empirical divergence measure equals the negative of the log-likelihood (plus a constant). Thus, the maximum likelihood estimator, asymptotically the most efficient estimator at the model under standard regularity conditions, belongs to the class of minimum DPD estimators. Larger values of α provide greater robustness and outlier stability, although the efficiency decreases with increasing α. Since robustness is a prime concern for us, we do not necessarily assume that the true distribution G belongs to the model; rather, we acknowledge that in reality, small deviations from the model are expected. At the same time, we hope to develop a procedure where these small deviations would not seriously degrade the statistical utility of the method. Large values of α protect the procedure against instability due to small deviations, but at the cost of a drop in model efficiency. We therefore wish to choose a data driven value of α in an 'optimal' way which balances the concerns of robustness and efficiency. We wish to choose a large value of α only when it is necessary.

It is already known to us that the minimum DPD procedure is Fisher consistent and when g belongs to the model, so that $g = f_{\theta^*}$ for some particular value θ^* of θ, simplified expressions for J and K in the variance expression may be obtained by replacing g with the model density in the expressions (2.16) and (2.17). In this case, $\theta_\alpha^g = \theta^*$ where θ_α^g is the best fitting parameter when the tuning parameter is α. As the robustness issue is a matter of concern for us, we will allow g to be the contaminated density $g(x) = (1 -$

$\epsilon) f_{\theta^*}(x) + \epsilon\delta(y-x)$, where δ is the Dirac delta function; this was also the approach taken by Warwick and Jones (2005). Here, θ^* is the true target parameter and estimators will be judged by their mean square error around θ^*. In general, of course, g may not involve f_θ per se; but, to keep a clear focus in our presentations, we will present almost all of our results with this contamination formulation in mind. (In the simulation study, we replace the delta function contamination with alternative contaminations.) We are to assess the performance of the estimator MDPDE $\hat{\theta}_\alpha$ through its summed mean square error $E\left\{\left(\hat{\theta}_\alpha-\theta^*\right)^T\left(\hat{\theta}_\alpha-\theta^*\right)\right\}$ which has the asymptotic formula

$$\begin{aligned} E\left\{\left(\hat{\theta}_\alpha-\theta^*\right)^T\left(\hat{\theta}_\alpha-\theta^*\right)\right\} &= n^{-1}\mathrm{tr}\left\{J_\alpha^{-1}\left(\theta_\alpha^g\right)K_\alpha\left(\theta_\alpha^g\right)J_\alpha^{-1}\left(\theta_\alpha^g\right)\right\} \\ &+ \left(\theta_\alpha^g-\theta^*\right)^T\left(\theta_\alpha^g-\theta^*\right), \qquad (3.1) \end{aligned}$$

where tr{.} denotes trace of a matrix.

3.1.2 Warwick-Jones and Hong-Kim Algorithms

Warwick and Jones (2005) proposed a useful method for choosing the tuning parameter associated with a family of robust estimators. The method was originally explored with the help of the family of minimum density power divergence estimators. The original Warwick and Jones (2005) suggestion for the selection of the optimal α consists of the following steps.

1. First, the asymptotic variance is estimated by substituting θ_α^g with $\hat{\theta}_\alpha$ and by substituting the true distribution G with the empirical distribution G_n in the forms of J and K to get $\hat{J}_\alpha\left(\hat{\theta}_\alpha\right)$,

$\hat{K}_\alpha\left(\hat{\theta}_\alpha\right)$ and $\hat{\psi}_{\hat{\theta}_\alpha}$ where

$$\begin{aligned}
\hat{J}_\alpha(\theta) &= \int\{(\alpha+1)\, u_\theta(x)\, u_\theta^T(x) - i_\theta(x)\} f_\theta^{1+\alpha}(x)\, dx \\
&+ \frac{1}{n}\sum\{i_\theta(X_i) - \alpha u_\theta(X_i)\, u_\theta^T(X_i)\} f_\theta^\alpha(X_i) \qquad (3.2) \\
\hat{K}_\alpha(\theta) &= \frac{1}{n}\sum u_\theta(X_i)\, u_\theta^T(X_i)\, f_\theta^{2\alpha}(X_i) - \hat{\psi}_\theta \hat{\psi}_\theta^T, \text{where,} \\
\hat{\psi}_\theta &= \frac{1}{n}\sum u_\theta(X_i)\, f_\theta^\alpha(X_i). \qquad (3.3)
\end{aligned}$$

Thus, the contribution of the variances to the summed MSE is estimated as

$$n^{-1}\mathrm{tr}\left\{\hat{J}_\alpha^{-1}\left(\hat{\theta}_\alpha\right)\hat{K}_\alpha\left(\hat{\theta}_\alpha\right)\hat{J}_\alpha^{-1}\left(\hat{\theta}_\alpha\right)\right\}. \qquad (3.4)$$

2. To estimate the asymptotic bias, θ_α^g is substituted with $\hat{\theta}_\alpha$ in the bias part of Equation (3.1). However, for the unknown θ^*, some suitable pilot estimator θ^p has to be used.

3. The estimates of variance and (squared) bias are added to get the summed empirical MSE (as a function of the tuning parameter α and the pilot estimator θ^p).

4. The summed empirical MSE is minimized over a fine grid of α values to obtain the optimal value of α (as a function of the pilot estimator θ^p).

On the other hand, in the Hong and Kim (2001) approach, the relevant objective function is the estimated asymptotic variance of the estimator. For each sample, they suggest the following.

1. Fix each $\alpha \in [0, 1]$ and evaluate the MDPDE.

2. Estimate the asymptotic variance empirically by substituting θ_α^g by $\hat{\theta}_\alpha$, i.e, calculate $V\left(\hat{\theta}_\alpha\right)$.

3. Choose that α which corresponds to the smallest estimated asymptotic variance.

These authors, therefore, drop the (squared) bias component in the objective function considered in Equation (3.1). We will discuss the pros and cons of these methods in the following sections.

3.1.3 Our Proposal

In the following, we will refer to the original Warwick and Jones approach as the 'one-step Warwick-Jones algorithm' or, in short, the OWJ algorithm, as opposed to the approach (which we will shortly describe) that uses the parameter estimate at a given step as the pilot estimate in the next step. Note that in the OWJ algorithm, the choice of the pilot can have a significant impact on the optimal tuning parameter, as the pilot invariably draws the final estimator towards itself. On the basis of repeated simulations, Warwick and Jones (2005) suggested the minimum L_2 distance estimator ($\hat{\theta}_1$) as the pilot estimator. Ghosh and Basu (2015) preferred a one-step algorithm with $\hat{\theta}_{0.5}$ as the pilot of their choice. But the essential issue of pilot-dependence is not bypassed in either case.

We take the view that if we are ready to accept the estimate obtained after the one-step algorithm as the 'optimal' estimate for the true unknown θ^*, we should also be prepared to view it as an updated pilot estimate for the continuation of the process. Thus, we propose to start the process with a suitable robust pilot estimator, but instead of terminating the algorithm after one step, the estimator obtained at the end of the step should be used as the updated pilot for the next step. The process should be continued until there is no further change in the estimate of θ^* (or, correspondingly, the estimate of the tuning parameter). If it can be demonstrated that the final converged estimate is independent of the initial choice of the pilot, it will provide us an 'optimal', pilot-independent estimate. In the following, we will refer to our proposed algorithm as the 'iterated WJ algorithm', or, in short, the IWJ algorithm.

3.2 The Three Algorithms : Some Comparisons

We now have three algorithms at our disposal to arrive at an estimate of the optimal tuning parameter. Through the present commentary, we want to setup the proper context where these methods may be compared in terms of their usefulness.

First we consider the Hong and Kim (HK) algorithm. It is clear that this algorithm will perform well in case of 'pure', outlier-free data, since the asymptotic variance is the only relevant quantity here. In case of contaminated data also, it often (but not always) works well. This may be explained as follows. A good robust estimator may often be expected to be closer to the true parameter under contamination compared to a non-robust estimator which is likely to show more variability. Thus, a robust estimator may be expected to have a smaller variance compared to the non-robust estimator, and, in many cases, the minimization of the asymptotic variance will recover a reasonably robust solution.

However, most robust estimators are devised to primarily control the bias under contamination; since the HK objective function has no bias component, there is no absolute guarantee that the criterion of low estimated asymptotic variance will necessarily lead to the desired optimal solution or even a robust solution. So although it is not a frequent phenomenon, the HK algorithm will, occasionally, fail and produce a highly non-robust solution.

Lack of robustness, on the other hand, is not a problem for the OWJ algorithm. In this case, any robust initial pilot – such as an estimator within the DPD family which may work as a robust estimator of θ in its own right – will produce a robust solution. The disadvantage here is that different robust pilot estimators can lead us to distinct, sometimes fairly disparate, solutions. In addition, the OWJ solution is more conservative than the HK solution under pure data and frequently produces a larger value of α as the optimal solution, compared to the HK method in order to give adequate importance to the robustness provision.

The IWJ method, as we will see in the following sections, appears to overcome this pilot-dependence. We will loosely consider all MDPDEs with $\alpha \geq 0.5$ as potential robust pilots. Our numerical illustrations, in each of the considered cases, will demonstrate that all robust pilots lead to the same IWJ optimal solution. The same has been observed in large scale simulation studies. In all of these cases, the final optimal estimator is the same for any robust initial pilot, making its choice unimportant in the final optimal solution.

Yet, the IWJ algorithm produces the same solution as the HK algorithm in a large number of cases. We discuss this in detail later in this section, but this imposes the responsibility on us to justify why the iterated method would still give a superior solution. We will provide some glimpses with real examples to show how the IWJ method

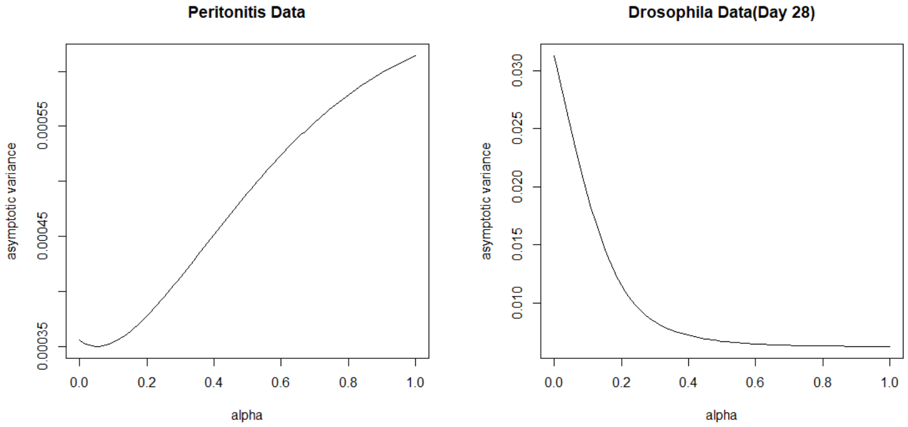


FIGURE 3.1: Asymptotic variance plots corresponding to Case 1.

produces a good compromise over the different scenarios based on estimated asymptotic variances. All these real data examples will be analyzed over different initial pilots in later sections. In the present section we will look at the estimated asymptotic variance curve over $\alpha \in [0, 1]$; more generally, we will consider the trace of the asymptotic covariance matrix in multi-parameter situations, as given in Equation (3.4). We consider the following cases; the IWJ optimal values in the following correspond to robust pilots, over which they are invariant. In the description below the example numbers refer to the examples considered in Section 3.3 later in this chapter to illustrate the performance of the three algorithms on real data.

1. **Case 1:** Here the curve of the estimated asymptotic variance has a single, global minimum. In this case, the IWJ algorithm and the HK algorithm lead to identical solutions. This happens, for example, in the cases of Drosophila data (Day 28) and Peritonitis data (Examples 3.2 and 3.3) with the common optimals being $\alpha = 0.99$ and 0.06, respectively: see Figure 3.1.

2. **Case 2:** Here the estimated asymptotic variance curve has more than one minimum with the global minimum at $\alpha = 0$ or at

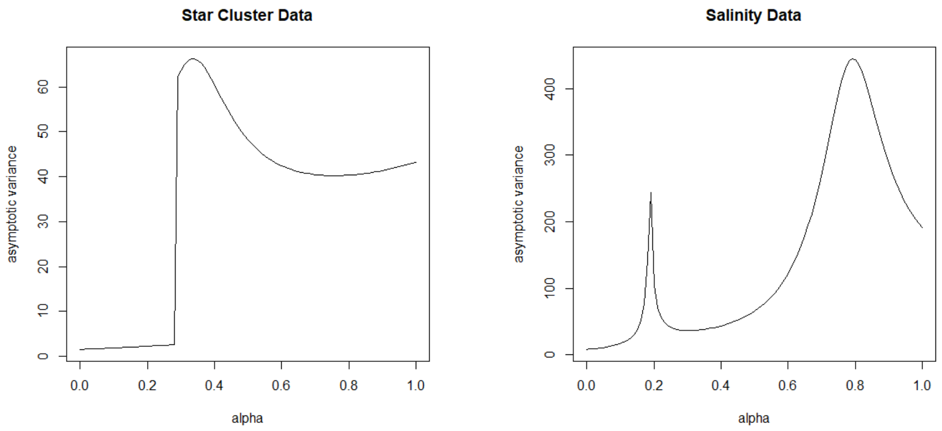


FIGURE 3.2: Asymptotic variance plots corresponding to Case 2.

some α close to 0 with no other local minimum to the left of it. In such situations, the HK solution corresponding to the global minimum of the asymptotic variance generally provides a non-robust solution. Starting from a robust pilot, the IWJ method converges to a larger value of α corresponding to a local minimum with a robust solution. Here the HK optimal value and the IWJ optimal value are distinct. This happens, for example, in the cases of Star Cluster data and Salinity data (Examples 3.8 and 3.10) with the IWJ optimals being $\alpha = 0.76$ and 0.30, respectively: see Figure 3.2.

3. **Case 3:** Here the estimated asymptotic variance curve has more than one minimum with the global minimum at $\alpha = 1$ or some other non-zero value of α with no other local minimum to the right of it. Here the IWJ as well as the HK algorithms correspond to the global minimum and thus the solutions are identical. This happens, for example, in the cases of Short's data and Telephone-line Fault data (Examples 3.4 and 3.6) with the common optimals being $\alpha = 0.98$ and 0.2, respectively: see Figure 3.3.

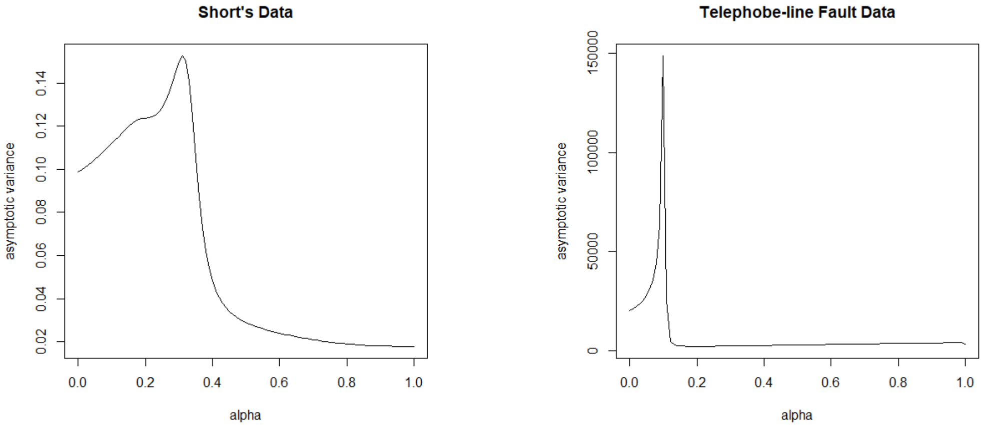


FIGURE 3.3: Asymptotic variance plots corresponding to Case 3.

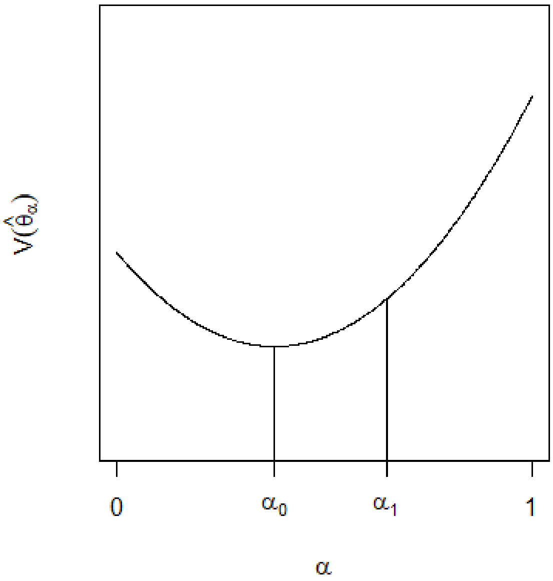


FIGURE 3.4: Asymptotic variance of the MDPDEs against α.

Theorem 3.1. *Suppose that the variance function $g(\alpha) = V(\hat{\theta}_\alpha)$ has a unique minimum at $\alpha = \alpha_0$. Then if the current pilot α_1 is distinct from α_0, the IWJ algorithm must take a step in the direction of α_0 and not remain stuck at α_1.*

Proof. Consider the graph in Figure 3.4. According to the IWJ algorithm, the pilot at the i-th step, namely, θ_p^i will be the MDPDE corresponding to the argmin (over α) of

$$h\left(\alpha\right) = V(\hat{\theta}_{\alpha}) + \left(\hat{\theta}_{\alpha} - \theta_p^{i-1}\right)^2 \tag{3.5}$$

Whenever $\theta_p^i = \theta_p^{i-1}$, the corresponding tuning parameter will be the optimal solution. Suppose that $\alpha = \alpha_1$ is the current solution (and hence the pilot for the next step) of the IWJ process; see Figure 3.4. At the next step, notice that, at any $\alpha > \alpha_1$, $V(\hat{\theta}_{\alpha})$ is greater than $V(\hat{\theta}_{\alpha_1})$ and, since α_1 is now the pilot, the bias at any such α is non-zero. Thus $h\left(\alpha\right) > h\left(\alpha_1\right)$ at any such α, and hence, the algorithm cannot take a step to the right of α_1.

We will show that the algorithm must take a (positive) step in the direction of α_0 and will not stay put at α_1. Let us choose some α close but to the left of α_1 and between α_1 and α_0 such that,

$$\hat{\theta}_{\alpha} = \hat{\theta}_{\alpha_1} \pm \epsilon. \tag{3.6}$$

Here $V(\hat{\theta}_{\alpha}) < V(\hat{\theta}_{\alpha_1})$. Now, considering a Taylor series expansion (up to first order) of $V(\hat{\theta}_{\alpha})$ around $V(\hat{\theta}_{\alpha_1})$, we get

$$V(\hat{\theta}_{\alpha}) = V(\hat{\theta}_{\alpha_1}) + \epsilon V'(\hat{\theta}_{\alpha})|_{\alpha=\alpha_1} + o\left(\epsilon\right) = V(\hat{\theta}_{\alpha_1}) + O\left(\epsilon\right). \tag{3.7}$$

Evidently, the $O\left(\epsilon\right)$ term is negative. From expression (3.7), we can say that, for small enough ϵ,

$$h\left(\alpha\right) = V(\hat{\theta}_{\alpha}) + \left(\hat{\theta}_{\alpha} - \hat{\theta}_{\alpha_1}\right)^2 = V(\hat{\theta}_{\alpha_1}) + O\left(\epsilon\right) + \epsilon^2 < V(\hat{\theta}_{\alpha_1}) = h\left(\alpha_1\right). \tag{3.8}$$

Thus the estimated MSE function $h(\cdot)$ is strictly smaller to the left of α_1 for small enough ϵ. Hence, the algorithm must take a step to the left rather than staying put at $\hat{\theta}_{\alpha_1}$. On the other hand, if the pilot is $\hat{\theta}_{\alpha_0}$, it is obvious that the algorithm cannot make any move in either direction any more.

One can similarly prove that if the current pilot is on the left of α_0, the IWJ algorithm must take a step to the right at the next stage. This theorem gives some justification of why the HK optimal and the IWJ optimal are identical when the estimated asymptotic variance function has a single, global, minimum. □

3.3 Applications

To further study the IWJ algorithm, we provide an extensive numerical study involving real data that conform to many different models and generally contain one or more outliers. In each example, we use (at least) eleven initial pilot estimates which will be the MDPDEs with $\alpha = 0.01, 0.1, 0.2, \ldots, 1$. As all the pilot estimates are from the minimum DPD class, we will let the expression 'pilot $\alpha = \alpha_0$' indicate that the pilot estimate is the MDPDE with $\alpha = \alpha_0$. Here, for each pilot, we have considered a fine grid of 101 values of α – 0, 0.01, 0.02, . . . , 1.0 – over which we are to find the optimal α by minimizing the empirical version of the objective function in Equation (3.1). In the tables, we will present the sequence over which the estimates of the tuning parameter progress in the iterated algorithm for each initial choice of the pilot. The background of each of the datasets along with the resulting parameter estimates are also described here. Our consistent observation in these examples is that the IWJ algorithm provides the same optimal estimate over a large range of initial pilot values of α, always containing the range $[0.5, 1]$.

In our real data examples, we will deal with both i.i.d. data models and linear regression models with normal errors. The approach to handling the i.i.d. data case has been described in the previous sections. In the linear regression model, our observations Y_i are conditionally independent and, given x_i, $Y_i \sim N\left(x_i^T\beta, \sigma^2\right)$, $i = 1, 2, \ldots, n$. Hence the observations are not identically distributed. Here we are interested in the MDPDEs of $\theta = \left(\beta^T, \sigma\right)$. The corresponding estimating equations as well as the asymptotic covariance matrices are given in Ghosh and Basu (2013) based on which the criterion in Equation (3.1) can be constructed. See Ghosh and Basu (2013) for an extended discussion of the linear regression case.

To motivate the proposal, we begin with the following recent example where the relevant problem will be posed and the difficulties with the classical analysis will be pointed out.

Example 3.1. *(Life Expectancy Data): This example deals with the relationship between life expectancy at birth (in years) and health spending per capita (in USD PPP, where purchasing power parity was used to convert the costs in local currency units to international dollar) in seventeen developed countries (fourteen European countries, USA, Australia and New Zealand). The data are obtained from the 'Health at a Glance 2017: OECD (Organisation for Economic Cooperation and Development) Indicators' publication (although the actual data refer to 2015). Normally, higher health spending per capita is expected to lead to higher life expectancy at birth. For these data, if we fit a linear relationship between the logarithm of health spending per capita and life expectancy at birth by the method of least squares, the fitted relationship, surprisingly, has a negative slope.*

A scrutiny of the data indicates that the observation corresponding to the United States represents a strong outlier in relation to the rest of the observations. It is clear that a better understanding of the general relationship between health spending and life expectancy at birth would be provided by a robustly fitted regression line which respects the expected positive relationship between health spending per capita and life expectancy at birth in most developed countries. The DPD method provides one principled method of doing so but, like most robust methods, depends on an unspecified tuning parameter. For objective analysis of the data, we also require the tuning parameter to be automatically, and reliably, specified. We have taken up this example to demonstrate the selection of the optimal value of the tuning parameter which gives robust estimates of the regression coefficients through refitting a linear regression model to these data based on the DPD using our proposed algorithm.

The application of all the three algorithms lead to an optimal tuning parameter of $\alpha = 0.98$ *for the calculation of the robust coefficient estimates of the linear regression model. On the other hand, all the three algorithms lead to the same robust estimates of intercept, slope as well as the variance corresponding to* $\alpha = 0.92$ *in case of data with the outlier removed. As expected, all these robust estimates (for full data) are quite different from the full data OLS estimates but close to the outlier deleted OLS estimates. The fitted curves corresponding to the optimal estimates for the full data along with the OLS estimates for full and outlier-deleted data are given in Figure 3.5. As the figure shows, the robust linear regression fit based on the optimal DPD tuning parameter is quite different from the least squares fit to the full dataset (but very similar to the least squares fit to the data when the outlier is removed), displaying a much more reasonable and* *informative positive slope. The iterative steps in the optimal tuning parameter selection rule are sequenced in Table 3.3.*

TABLE 3.1: Life expectancy at birth and health spending per capita (USD PPP), 2015

Country	Health spending per capita (USD PPP)	Life expectancy in years
Australia	4492.55	82.5
Austria	5100.02	81.3
Belgium	4778.45	81.1
Finland	3993.19	81.6
France	4529.59	82.4
Greece	2210.07	81.1
Iceland	4105.67	82.5
Ireland	5275.77	81.5
Luxembourg	6817.90	82.4
Netherlands	5296.71	81.6
New Zealand	3544.56	81.7
Norway	6190.14	82.4
Portugal	2663.70	81.2
Slovenia	2730.80	80.9
Sweden	5266.33	82.3
United Kingdom	4125.26	81.0
United States	9507.20	78.8

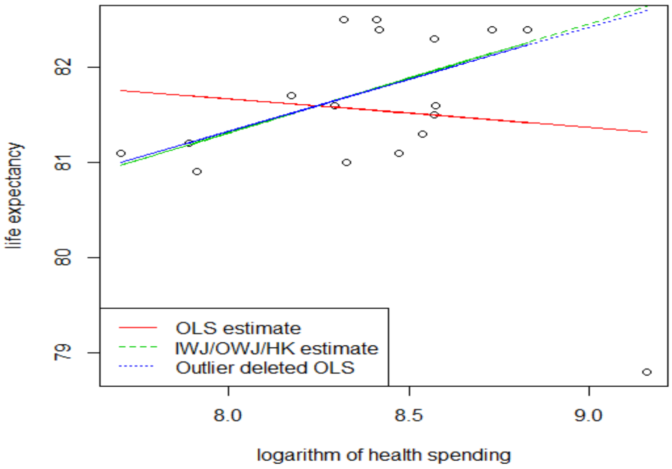


FIGURE 3.5: A few different linear regression fits for Life Expectancy Data.

TABLE 3.2: Optimal estimates for the Life Expectancy Data

methods	β_0	β_1	σ
IWJ and OWJ and HK	72.08283	1.153166	0.5547354
OLS	84.1027300	-0.3040737	0.9165954
Outlier-deleted OLS	72.54185	1.098033	0.4800876

TABLE 3.3: Tuning parameter sequence (Life Expectancy Data)

pilot α	iteration 1	iteration 2	iteration 3
0.01	**0**	**0**	**0**
0.1	.06	**.98**	**.98**
0.2	.95	**.98**	**.98**
$0.3 - .4$	.96	**.98**	**.98**
$0.5 - .7$	.97	**.98**	**.98**
$0.8 - 1$	**.98**	**.98**	**.98**

The results show that except for the $\alpha = 0.01$ as the starting pilot, all other pilots lead to the same final tuning parameter (and hence the same estimator). All the values from the point where the tuning parameter shows no further change are given in bold fonts. We now further illustrate the performance of the tuning parameter selection method through a host of popular real data examples, where the same observations as in our motivating example (Example 3.1) will be noticed.

3.3.1 I.I.D. data examples

Example 3.2. *(Drosophila Data): We consider a segment of data on drosophila (a type of fruit fly). The experimental protocol is described in Woodruff et al. (1984). The data were previously analyzed by Simpson (1987), and are presented in Table 3.4. These data contain information regarding chemical mutagenicity of Drosophila flies. This sex-linked recessive lethal test was conducted on these fruit flies*

Table 3.4: Recessive lethal count

No. of daughters (X)	0	1	2	3	4	≥ 5
No. of males (Day 28)	23	3	0	1	1	0
No. of males (Day 177)	23	7	3	0	0	1 (91)

where groups of male flies were exposed to different doses of a chemical. Each male was then mated with unexposed females. Here the variable of interest is the number of daughter flies (for each male) carrying a recessive lethal mutation on the X-chromosome. Having noted the frequencies in each cell, we are interested in modelling these data with a Poisson distribution and estimating its mean. In this context, we consider two specific experimental runs, on day 28 and day 177, respectively. For these data, the results for the optimal tuning parameter selection through the IWJ algorithm are given in Tables 3.5 and 3.6.

Table 3.5: Tuning parameter sequence (Drosophila Data, day 28)

pilot	iteration						
α	1	2	3	4	5	6	7
0.01	.08	.16	1	**.99**	**.99**	**.99**	**.99**
0.1	.18	1	**.99**	**.99**	**.99**	**.99**	**.99**
0.2 – 0.3	1	**.99**	**.99**	**.99**	**.99**	**.99**	**.99**
0.4	.83	.89	.93	.95	.96	**.99**	**.99**
0.5	.78	.86	.91	.94	.96	**.99**	**.99**
0.6	.79	.87	.91	.94	.96	**.99**	**.99**
0.7	.83	.89	.93	.95	.96	**.99**	**.99**
0.8	.87	.91	.94	.96	**.99**	**.99**	**.99**
0.9	.93	.95	.96	**.99**	**.99**	**.99**	**.99**
1	**.99**	**.99**	**.99**	**.99**	**.99**	**.99**	**.99**

In Tables 3.5 and 3.6, we observe that the final converged value of the tuning parameter is the same for every initial pilot under the IWJ algorithm. The HK algorithm produces the same optimal solution as

Table 3.6: Tuning parameter sequence (Drosophila Data, day 177)

pilot α	iteration 1	iteration 2	iteration 3
0.01	.02	**.03**	**.03**
$0.1 - 1$	**.03**	**.03**	**.03**

the IWJ in these two cases. In the case of the full data, the common optimal estimates of the Poisson mean parameter corresponding to day 28 and day 177 are 0.16311 and 0.3935, respectively. These estimates are seen to be substantially closer to 0.115385 and 0.393939, the outlier-deleted MLEs for these datasets, compared to 0.357143 and 3.058824, the corresponding full data MLEs.

Example 3.3. *(Peritonitis Data): This example involves the incidence of peritonitis in* 390 *kidney patients. The data are available in Table 2.4 in Basu et al. (2011). A geometric model with success probability θ has been fitted to these frequency data. The two largest observations of this dataset are mild to moderate outliers. In this case, the final optimal solutions for the tuning parameter under the IWJ algorithm are all the same except for the most non-robust initial pilot. The IWJ, OWJ and HK optimal parameter estimates are* 0.498394, 0.502111 *and* 0.498056*, respectively. In this example, the HK solution corresponds to $\alpha = 0.05$, being different from the IWJ optimal value. This is a numerical difference caused by the discreteness of the α-grid. If we consider a grid of 100,000 values over $\alpha \in [0, 1]$ instead of the grid of 100 values, then the iterated WJ optimal corresponds to $\alpha = 0.05213$, which is the HK optimal value also.*

On the contrary, if we consider MLEs for comparison, then the MLEs for the full data and the (two) outlier deleted data are 0.496183 *and* 0.509186*, respectively. In this example, the robust estimates are all* *closer to the MLE than to the outlier deleted estimate, but the differences between all estimates are small.*

Table 3.7: Peritonitis Data

cases (X)	0	1	2	3	4	5	6	7	8	9	10	11	≥ 12
frequencies	199	94	46	23	17	4	4	1	0	0	1	0	1

Table 3.8: Tuning parameter sequence (Peritonitis Data)

pilot α	iteration 1	iteration 2	iteration 3	iteration 4	iteration 5
.01	.03	.04	**.05**	**.05**	**.05**
.1	.07	**.06**	**.06**	**.06**	**.06**
.2	.1	.07	**.06**	**.06**	**.06**
.3	.12	.08	**.06**	**.06**	**.06**
.4	.14	.08	**.06**	**.06**	**.06**
.5 − .6	.16	.09	.07	**.06**	**.06**
.7	.18	.09	.07	**.06**	**.06**
.8 − .9	.19	.1	.07	**.06**	**.06**
1	.2	.1	.07	**.06**	**.06**

Example 3.4. *(Short's Data): In 1761, to determine the parallax of the sun, the angle subtended by the earth's radius as if viewed and measured from the surface of the sun, James Short made an analysis of observations of the 'transit of Venus', the apparent passage of the planet Venus across the face of the sun, as viewed from the earth. The raw data corresponding to 17 observations in one of his datasets are given in Table 3.9. Under the normal model, we are interested in estimating the mean μ and the standard deviation σ for these data. Here the (common) optimal parameter estimates obtained through the three algorithms are $\hat{\mu} = 8.419890$ and $\hat{\sigma} = 0.274061$. The associated optimal α is 0.98. For comparison, we note that the corresponding full data MLEs are 8.377647 and 0.845539 while the (five) outlier deleted MLEs are 8.464167 and 0.189361. The different fits are graphically presented in Figure 3.6.*

In Table 3.10, the tuning parameter sequences for the IWJ algorithm for the simultaneous estimation of the two normal parameters are presented.

The estimated asymptotic summed variance curve is given in the left panel of Figure 3.3. Here the HK, OWJ and IWJ algorithms all lead to the same optimal solution provided one uses a robust pilot in case of the IWJ algorithm.

Table 3.9: Short's Data

8.65	8.35	8.71	8.31	8.36	8.58	7.8	7.71	8.30
9.71	8.50	8.28	9.87	8.86	5.76	8.84	8.23	

Table 3.10: Tuning parameter sequence (Short's Data)

pilot α	iteration 1	iteration 2	iteration 3	iteration 4	iteration 5	iteration 6	iteration 7
.01	**0**	**0**	**0**	**0**	**0**	**0**	**0**
.1	.04	**0**	**0**	**0**	**0**	**0**	**0**
.2	.18	.16	.13	.09	.04	**0**	**0**
.3	.92	**.98**	**.98**	**.98**	**.98**	**.98**	**.98**
$.4-1$	**.98**	**.98**	**.98**	**.98**	**.98**	**.98**	**.98**

Example 3.5. *(Newcomb's Data): Newcomb's measurements of the velocity of light, given in Table 3.11, are based on observations, in the U.S. in 1882, of the passage time taken by light to travel over a distance of 3721 meters and back; the observations are given in Table 3.11. Here also, we assume normality and our purpose is to estimate the mean and the standard deviation using the three algorithms. The IWJ and HK optimal estimates are the same, at* $\hat{\mu} = 27.64296$ *and* $\hat{\sigma} = 5.053583$*, while the OWJ optimal estimates are* 27.56197 *and* 4.939927*. These values may be compared with 26.21212 and 10.66361 (the full data MLEs), and with 27.75000 and 5.04356 (the outlier deleted MLEs). In this case, the IWJ and OWJ optimal values, although distinct, appear to be equally effective at giving a good robust*

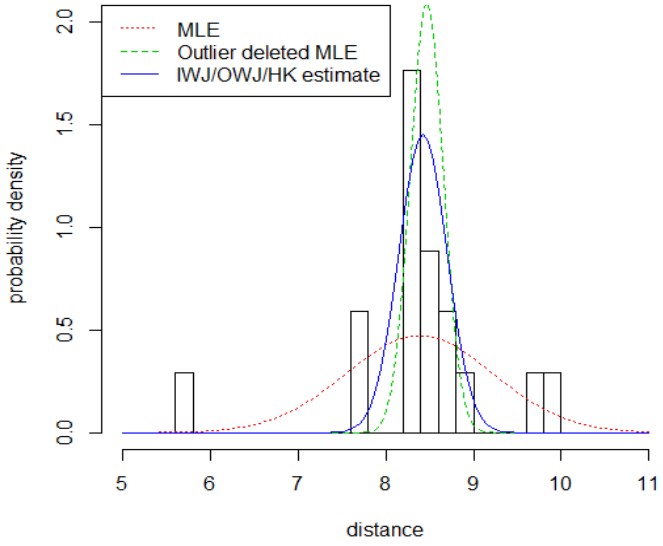


FIGURE 3.6: A few different fits for Short's Data under the two-parameter normal model.

solution. The relevant fits are given in Figure 3.7. In this case, each pilot with $\alpha \geq 0.1$ *leads to the same eventual optimal tuning parameter* $\alpha = 0.23$ *under the IWJ algorithm (which is also the HK optimal).*

TABLE 3.11: Newcomb's Data

28	26	33	24	34	-44	27	16	40	-2
29	22	24	21	25	30	23	29	31	19
24	20	36	32	36	28	25	21	28	29
37	25	28	26	30	32	36	26	30	22
36	23	27	27	28	27	31	27	26	33
26	32	32	24	39	28	24	25	32	25
29	27	28	29	16	23				

Example 3.6. *(Telephone-line Fault Data): These data, analysed by Welch (1987), involve results of an experiment where a method of reducing faults on telephone lines had been tested. Fourteen matched pairs of areas were considered, where the observations are differences between reciprocals of numbers of test and control fault rates, among*

Table 3.12: Tuning parameter sequence (Newcomb's Data)

pilot α	iteration 1	iteration 2	iteration 3	iteration 4	iteration 5
.01	.01	.01	0	**0**	**0**
.1	.16	.2	.22	**.23**	**.23**
.2	.22	**.23**	**.23**	**.23**	**.23**
.3	.24	**.23**	**.23**	**.23**	**.23**
.4	.26	.24	**.23**	**.23**	**.23**
.5	.28	.24	**.23**	**.23**	**.23**
.6	.3	.24	**.23**	**.23**	**.23**
.7	.33	.25	**.23**	**.23**	**.23**
.8	.35	.25	**.23**	**.23**	**.23**
.9	.39	.26	.24	**.23**	**.23**
1	.42	.26	.24	**.23**	**.23**

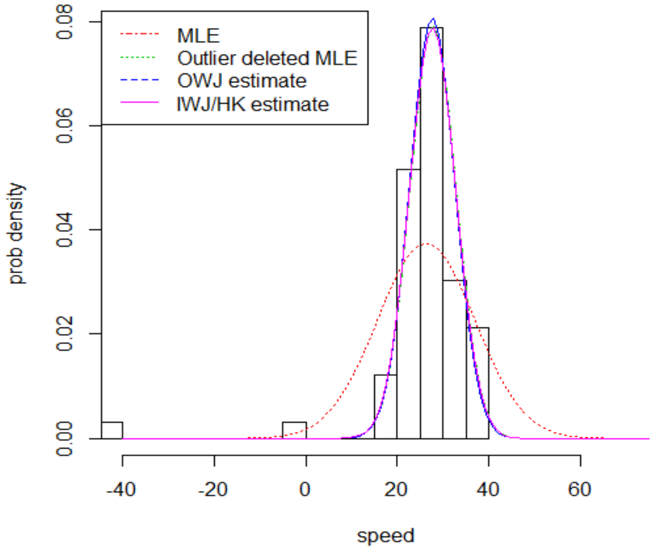


Figure 3.7: A few different fits for Newcomb's Data under the two-parameter normal model.

which −988 *is a large outlier. The data are given in Table* 3.13*. The HK/IWJ optimal estimates of* μ *and* σ *are* 123.29752 *and* 132.8642 *but the OWJ optimal estimates are* 123.89689 *and* 133.0281*, respectively. The full data MLEs, on the other hand, are* 38.92857 *and* 310.2318 *and the outlier deleted MLEs are* 117.9231 *and* 127.6139*.*

Once again, the IWJ and OWJ optimal values provide stable solutions, although there is a slight difference between them. The optimal IWJ solution corresponds to $\alpha = 0.2$ for all robust pilots, which equals the HK solution and slightly differs from the OWJ optimal (corresponds to $\alpha = .22$ in case of the L_2 pilot). Relevant fits are given in Figure 3.8.

Table 3.13: Telephone-line Fault Data

−988	−135	−78	3	59	83	93	110	189	197	204	229	269	310

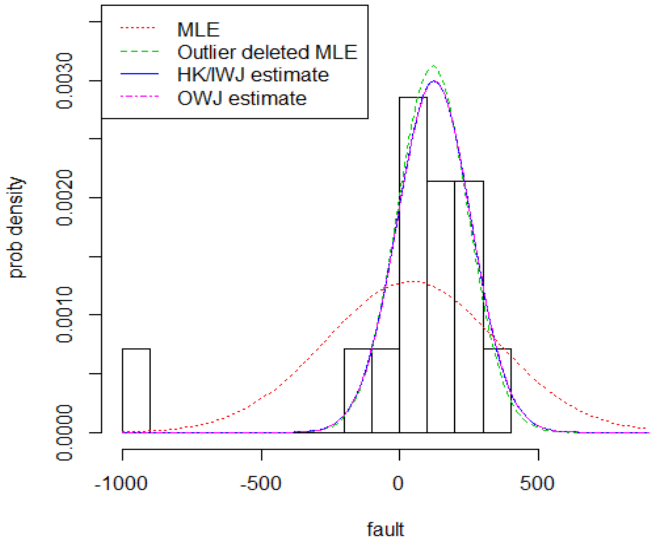


Figure 3.8: A few different fits for Telephone-line Fault data under the two-parameter normal model.

Table 3.14: Tuning parameter sequence (Telephone-line Fault Data)

pilot α	iteration 1	iteration 2	iteration 3
.01	**0**	**0**	**0**
.1	.13	**.2**	**.2**
.2 − .4	**.2**	**.2**	**.2**
.5 − .7	.21	**.2**	**.2**
.8 − 1	.22	**.2**	**.2**

Example 3.7. *(Insulating Fluid Data): Here we provide a non-normal example in continuous models. The data contain breakdown*

times of an insulating fluid between electrodes, and they are recorded at seven different voltages. The data are presented in Nelson (1982). Here we have taken the times associated with insulation corresponding to 34 *kV, which are assumed to follow an exponential distribution. The data contain one extreme outlier and four moderately severe outliers. If the initial pilot is non-robust, the optimal MDPDE of the mean parameter is the maximum likelihood estimate (MLE; corresponding to* $\alpha = 0$*). On the other hand, initial pilots corresponding to moderate to large* α *lead us to the minimum* L_2 *distance estimate as the optimal one, which is the HK optimal also. Each of the three algorithms leads to the minimum* L_2 *distance estimate as the optimal solution, which is 8.175565 in this case. For the full data, the MLE is* 14.35895*, whereas the outlier deleted MLE is* 4.645714*.*

Fits of selected estimates are given in Figure 3.9. Here both one (the last value in Table 3.15) outlier deleted and five (the last five values in Table 3.15) outlier deleted fits are also presented in the figure.

Table 3.15: Insulating Fluid Data

0.19	0.78	0.96	1.31	2.78	3.16	4.15	4.67	4.85	6.50
7.35	8.01	8.27	12.06	31.75	32.52	33.91	36.71	72.89	

Table 3.16: Tuning parameter sequence (Insulating Fluid Data)

pilot α	iteration 1	iteration 2	iteration 3	iteration 4	iteration 5
0.01	**0**	**0**	**0**	**0**	**0**
.1	.04	**0**	**0**	**0**	**0**
.2	.14	.09	.03	**0**	**0**
$.3 - 1$	**1**	**1**	**1**	**1**	**1**

Example 3.8. *(Hertzsprung-Russell Star Cluster Data): This example involves astronomical data: the observations form the Hertzsprung-Russell diagram of the star cluster CYG OB1, for which the number of observations,* $n = 47$*. The data, given in Rousseeuw and*

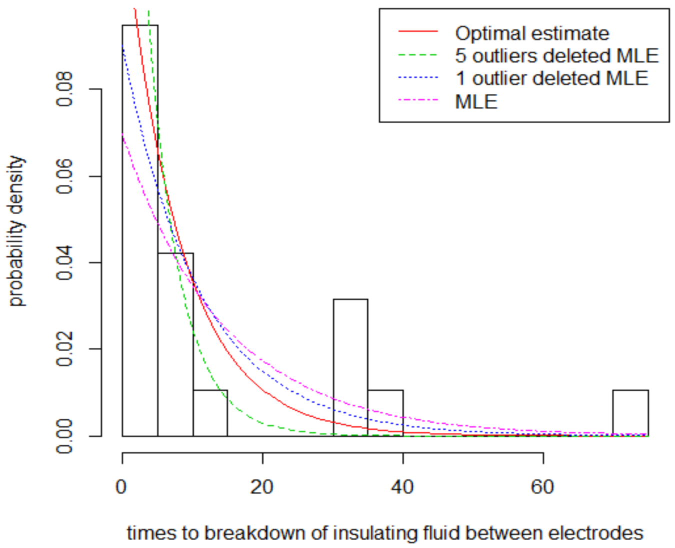


FIGURE 3.9: A few different fits for Insulating Fluid Data under the exponential model.

Leroy (1987), are presented in Table 3.17. The data on the logarithm of the surface temperature of the star (x)*, and the logarithm of its light intensity* (y) *are considered to follow the linear regression model* $y = \alpha + \beta x + \epsilon$*, where* $\epsilon \sim N\left(0, \sigma^2\right)$*. The data contain four large outliers (the 11-th, 20-th, 30-th and 34-th observations). The ordinary least squares (OLS) estimates and the robust estimates of the regression coefficients as well as of the scale of the error are presented in Table 3.19. Also see Figure 3.10 for a graphical representation of the fits of some of these estimates. Note that the HK and IWJ optimal values are distinct in this case; in fact the IWJ and OWJ optimal values are identical while the HK method leads to the OLS estimates, i.e., the IWJ optimal value of* α *is* $\alpha = 0.76$*, whereas the HK optimal value is* $\alpha = 0$*. Moreover, if we delete one more observation along with the four outliers, i.e., the 7-th observation considering it to be a moderate outlier, then the OLS estimate will become quite close to the optimal robust solution. The star cluster* *data asymptotic variance plot is given in the left panel of Figure 3.2.*

Table 3.17: Hertzsprung-Russell Star Cluster Data

index	log of	log of	index	log of	log of
	temperature	intensity		temperature	intensity
	(x)	(y)		(x)	(y)
1	4.37	5.23	25	4.38	5.02
2	4.56	5.74	26	4.42	4.66
3	4.26	4.93	27	4.29	4.66
4	4.56	5.74	28	4.38	4.90
5	4.30	5.19	29	4.22	4.39
6	4.46	5.46	30	3.48	6.05
7	3.84	4.65	31	4.38	4.42
8	4.57	5.27	32	4.56	5.10
9	4.26	5.57	33	4.45	5.22
10	4.37	5.12	34	3.49	6.29
11	3.49	5.73	35	4.23	4.34
12	4.43	5.45	36	4.62	5.62
13	4.48	5.42	37	4.53	5.10
14	4.01	4.05	38	4.45	5.22
15	4.29	4.26	39	4.53	5.18
16	4.42	4.58	40	4.43	5.57
17	4.23	3.94	41	4.38	4.62
18	4.42	4.18	42	4.45	5.06
19	4.23	4.18	43	4.50	5.34
20	3.49	5.89	44	4.45	5.34
21	4.29	4.38	45	4.55	5.54
22	4.29	4.22	46	4.45	4.98
23	4.42	4.42	47	4.42	4.50
24	4.49	4.85			

Example 3.9. *(Gesell Adaptive Score Data): This two-dimensional dataset involves the age (in months) at which a child utters its first word (x), and the corresponding Gesell adaptive score (y). The Gesell adaptive score test is given to children to measure their level of cognitive development. The data for 21 children, analysed by Mickey et al. (1967) are given in Table 3.20. Here also we consider the simple*

Table 3.18: Tuning parameter sequence (Star Cluster Data)

pilot α	iteration 1	iteration 2	iteration 3	iteration 4
0.01 – 0.2	**0**	**0**	**0**	**0**
0.3	.72	.76	**.76**	**.76**
0.4	.74	**.76**	**.76**	**.76**
0.5 – 0.7	.75	**.76**	**.76**	**.76**
0.8 – 1	**.76**	**.76**	**.76**	**.76**

Table 3.19: Optimal estimates for the Hertzsprung-Russell Star Cluster Data

methods	β_0	β_1	σ
IWJ and OWJ	−8.572644	3.065783	0.402574
HK and OLS	6.793469	−0.413304	0.552488
4 Outlier-deleted OLS	−4.056524	2.046657	0.396256
5 Outlier-deleted OLS	−7.403531	2.802837	0.3670406

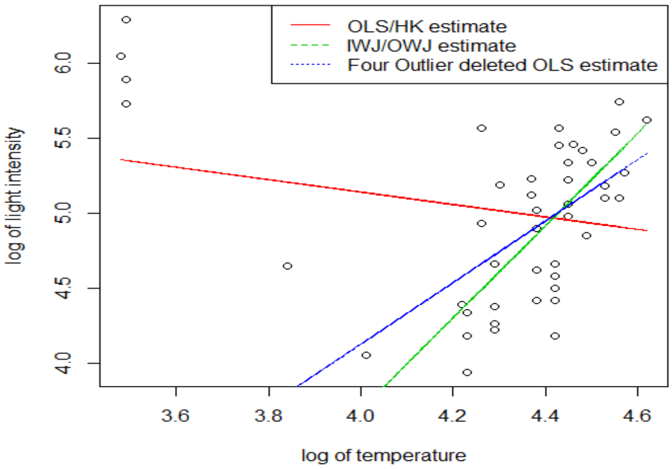


Figure 3.10: A few different regression fits for Star Cluster Data.

linear regression model. The estimated parameters are given in Table 3.22. The IWJ and HK optimal values are identical in this case and they correspond to $\alpha = 0.33$*; the OWJ optimal value is distinct. However the difference between the estimators is not of a very high order in this example; see Figure 3.11.*

Table 3.20: Gesell Adaptive Score Data

index	1	2	3	4	5	6	7	8	9	10	11
age (x)	15	26	10	9	15	20	18	11	8	20	7
score (y)	95	71	83	91	102	87	93	100	104	94	113
index	12	13	14	15	16	17	18	19	20	21	
age (x)	9	10	11	11	10	12	42	17	11	10	
score (y)	96	83	84	102	100	105	57	121	86	100	

Table 3.21: Tuning parameter sequence (Gesell Adaptive Score Data)

pilot α	iteration 1	iteration 2	iteration 3	iteration 4	iteration 5	iteration 6
0.1	.26	.32	**.33**	**.33**	**.33**	**.33**
0.2	.28	.32	**.33**	**.33**	**.33**	**.33**
0.2	.3	.32	**.33**	**.33**	**.33**	**.33**
0.3	.32	**.33**	**.33**	**.33**	**.33**	**.33**
0.4	.35	.34	**.33**	**.33**	**.33**	**.33**
0.5	.38	.34	**.33**	**.33**	**.33**	**.33**
0.6	.41	.35	.34	**.33**	**.33**	**.33**
0.7	.46	.36	.34	**.33**	**.33**	**.33**
0.8	.53	.39	.35	.34	**.33**	**.33**
0.9	.62	.42	.35	.34	**.33**	**.33**
1	.73	.48	.37	.34	**.33**	**.33**

Table 3.22: Optimal estimates for the Gesell Adaptive Score Data

methods	β_0	β_1	σ
IWJ and HK	110.557559	-1.219669	9.456469
OWJ	112.405253	-1.293444	8.982978
OLS	109.873840	-1.126989	10.484878
Outlier-deleted OLS	109.304679	-1.193311	8.185425

Example 3.10. *(Salinity Data): This example deals with a set of values measuring salt concentration of water and river discharge taken in the Pamlico Sound of North Carolina. Here, the salinity* (y) *together with three explanatory variables, namely, salinity lagged by two weeks* (x_1)*, the number of biweekly periods elapsed since the beginning of the spring season* (x_2) *and the volume of river discharge*

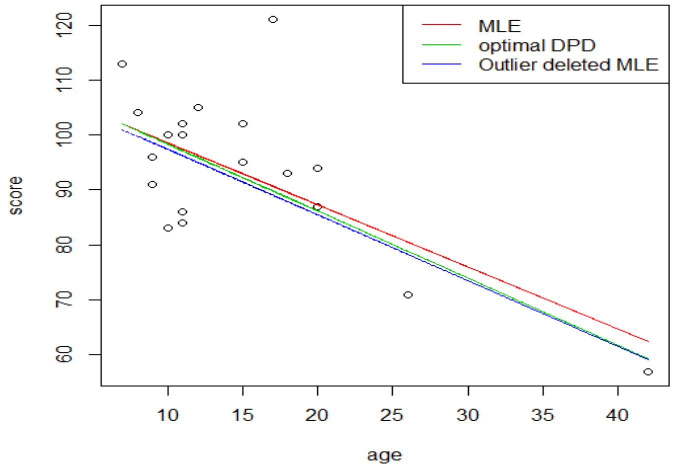


FIGURE 3.11: A few different regression fits for Gesell Adaptive Score Data.

into the sound (x_3) *are taken into consideration. The 28 observations are given in Table 3.23. These data have been originally presented in Ruppert and Carroll (1980); Rousseeuw and Leroy (1987) modeled these data using a multiple linear regression model. The different estimates under the multiple linear regression model* $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$, *where* $\epsilon \sim N\left(0, \sigma^2\right)$, *are given in Table 3.25. In this case also, the IWJ and HK algorithms lead to distinct optimal values of* α. *The IWJ and OWJ solutions, although also distinct, are very close. Both give robust solutions close to the outlier deleted OLS solution. However, the HK algorithm fails to give a robust solution. The residual plot in Figure 3.12 shows how the robust fit makes the big outlier stand out.*

3.3.2 No outlier performance

All the datasets that we have analyzed here can be strongly argued to contain one or more outliers. What would happen if these outliers were absent and the data exhibited much better model conformity? To what extent are the optimal α values pushed closer to zero? Take

Table 3.23: Salinity Data

Index	Lagged Salinity	Trend	Discharge	Salinity
	(x_1)	(x_2)	(x_3)	(y)
1	8.2	4	23.005	7.6
2	7.6	5	23.873	7.7
3	4.6	0	26.417	4.3
4	4.3	1	24.868	5.9
5	5.9	2	29.895	5.0
6	5.0	3	24.200	6.5
7	6.5	4	23.215	8.3
8	8.3	5	21.862	8.2
9	10.1	0	22.274	13.2
10	13.2	1	23.830	12.6
11	12.6	2	25.144	10.4
12	10.4	3	22.430	10.8
13	10.8	4	21.785	13.1
14	13.1	5	22.380	12.3
15	13.3	0	23.927	10.4
16	10.4	1	33.443	10.5
17	10.5	2	24.859	7.7
18	7.7	3	22.686	9.5
19	10.0	0	21.789	12.0
20	12.0	1	22.041	12.6
21	12.1	4	21.033	13.6
22	13.6	5	21.005	14.1
23	15.0	0	25.865	13.5
24	13.5	1	26.290	11.5
25	11.5	2	22.932	12.0
26	12.0	3	21.313	13.0
27	13.0	4	20.769	14.1
28	14.1	5	21.393	15.1

Newcomb's data, for example. The removal of the two largest outliers produces a nice bell shaped structure which is almost perfectly symmetric and exhibits no obvious aberrations from the assumed normal model. There is no apparent reason to use anything other than the maximum likelihood estimator in this case. However, does

Table 3.24: Tuning parameter sequence (Salinity Data)

pilot α	iteration 1	iteration 2	iteration 3
$0.01 - 1$	**0**	**0**	**0**
0.2	**.30**	**.30**	**.30**
$0.3 - 1$	.31	**.30**	**.30**

Table 3.25: Optimal estimates for the Salinity Data

methods	β_0	β_1	β_2	β_3	β_4
IWJ	18.264868	0.716406	−0.184652	−0.622322	0.943039
OWJ	18.288342	0.716755	−0.186029	−0.623212	0.939032
HK and OLS	9.590265	0.777105	−0.025512	−0.295036	1.231637
Outlier-deleted OLS	18.491419	0.697341	−0.157051	−0.630538	0.984163

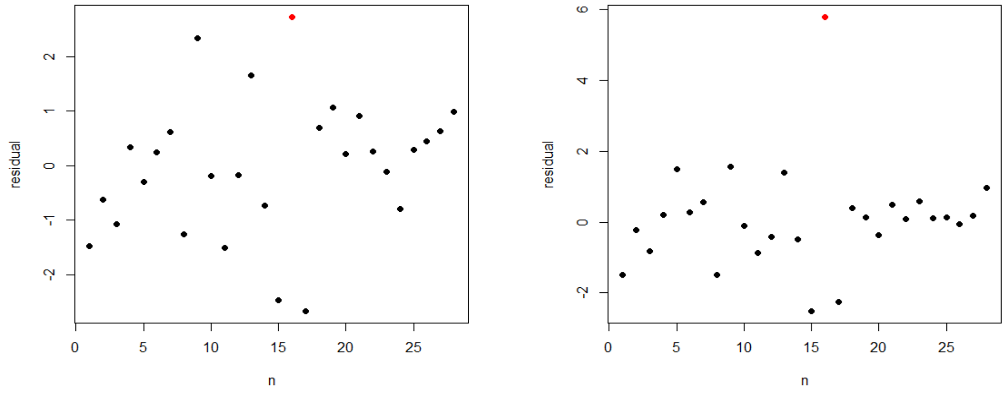


Figure 3.12: Residual plots for OLS and optimal IWJ fits, respectively, on Salinity Data.

our algorithm lead us to the maximum likelihood estimator in this case? In the following, we investigate such issues further.

Table 3.26 provides a comparison of the three algorithms by listing the full data optimal values of α and the outlier deleted optimal values of α for the three methods for all the examples studied by us. The numbers demonstrate that for the data involving outliers, the optimal tuning parameters obtained by the IWJ and OWJ algorithms

TABLE 3.26: The three optimal values of α for the full data as well as the outlier removed data corresponding to all datasets.

		Full data optimal αs			Outlier deleted optimal αs		
Dataset	Outlier deletion	IWJ	OWJ	HK	IWJ	OWJ	HK
Life Expectancy	one: index 17	0.98	0.98	0.98	0.92	0.92	0.92
Drosophila (1st run)	two: values $3, 4$	0.99	0.99	0.99	0	0	0
Drosophila (2nd run)	one: value 91	0.03	0.03	0.03	0	0	0
Peritonitis	two: values $10, 12$	0.06	0.2	0.05	0	0	0
Short	five: values $5.76, 9.87$ $9.71, 7.8, 7.71$	0.98	0.98	0.98	0	0.17	0
Newcomb	two: values $-44, -2$	0.23	0.42	0.23	0	0.54	0
Telephone-line Fault	one: value -988	0.2	0.22	0.2	0	0	0
Insulating Fluid	five: indices $15, 16, 17, 18, 19$	1	1	1	0	0.25	0
Hertzsprung-Russell Star Cluster	four: indices $11, 20, 30, 34$	0.76	0.76	0	0.68	0.70	0
Gesell Adaptive Score	one: index 19	0.33	0.73	0.33	0.03	0.77	0.03
Salinity	one: index 16	0.3	0.31	0	0.09	0.49	0

are often close. However, for the outlier deleted data, the IWJ algorithm is more successful in pushing the optimal tuning parameter closer to zero. In fact, in all of our i.i.d. data examples, the deletion of outliers leads to $\alpha = 0$ being the optimal tuning parameter for the IWJ algorithm. Even for the regression examples, the drop in the value of α due to outlier deletion is more considerable for the iterated algorithm. (The one-step method actually leads to an increase in the value of the optimal α in two cases). The life expectancy dataset is the only exception. On the whole it appears that for pure data, the iterated version provides a more suitable optimal value of α. HK provides even better choices of α for the pure data but at the expense of failing to be sufficiently robust for some datasets.

In Table 3.26, the IWJ and OWJ optimal values correspond to $\alpha = 1$ as the initial pilot. However, the IWJ optimal values are invariant for all pilot $\alpha \in [0.5, 1]$, unlike the OWJ algorithm.

3.4 Simulation Study

In this section, we present the results of a small simulation study comparing the values of the tuning parameter provided by HK, IWJ and OWJ methods in both pure and contaminated cases. In each case, the initial pilot value of α is taken to be 1. Our simulation scenarios are as follows.

- Case 1: We draw independent samples of size 50 from the standard normal distribution, $N(0, 1)$. The process is replicated 1000 times. For each sample, the optimal values of α under each of the three algorithms are determined. The process is then repeated, with the same sample size and number of replications, for the $P(2)$ – Poisson with mean 2 – distribution under the Poisson model with parameter θ.

- Case 2: Here the setup is exactly the same as that for Case 1 except that the normal data, in each sample, are contaminated with 10% of observations from $N(8, 1)$. Similarly the Poisson data, in each sample, are contaminated with 10% of observations from $P(15)$.

Our comparison comprises all pairwise scatterplots of HK, IWJ and OWJ optimal values. The results for the pure data cases are depicted in Figure 3.13 and for the contaminated data cases in Figure 3.14.

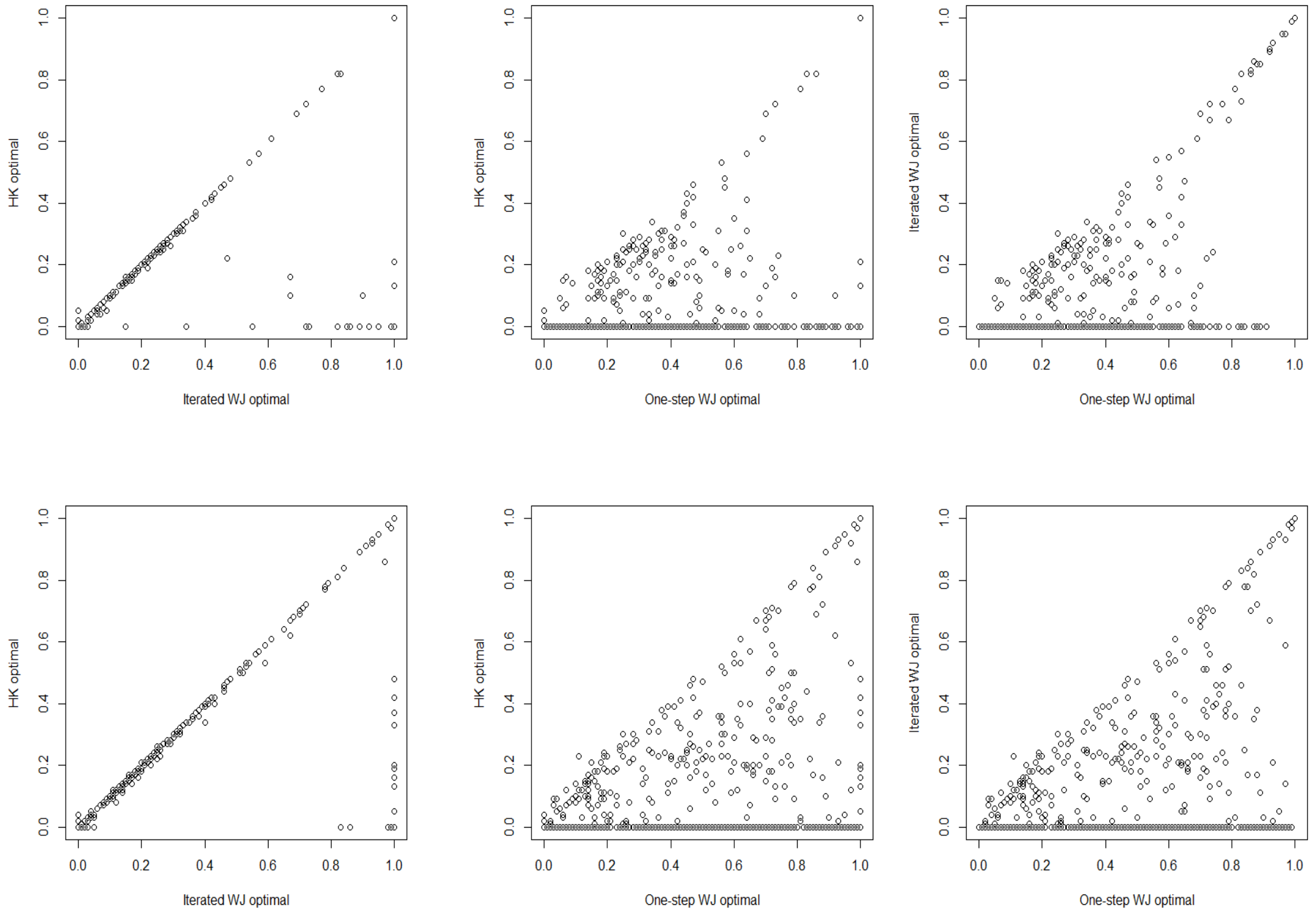


FIGURE 3.13: Comparison among optimal tuning parameters of the three algorithms for samples from pure $N(0,1)$ (top panel) and $P(2)$ (bottom panel) models.

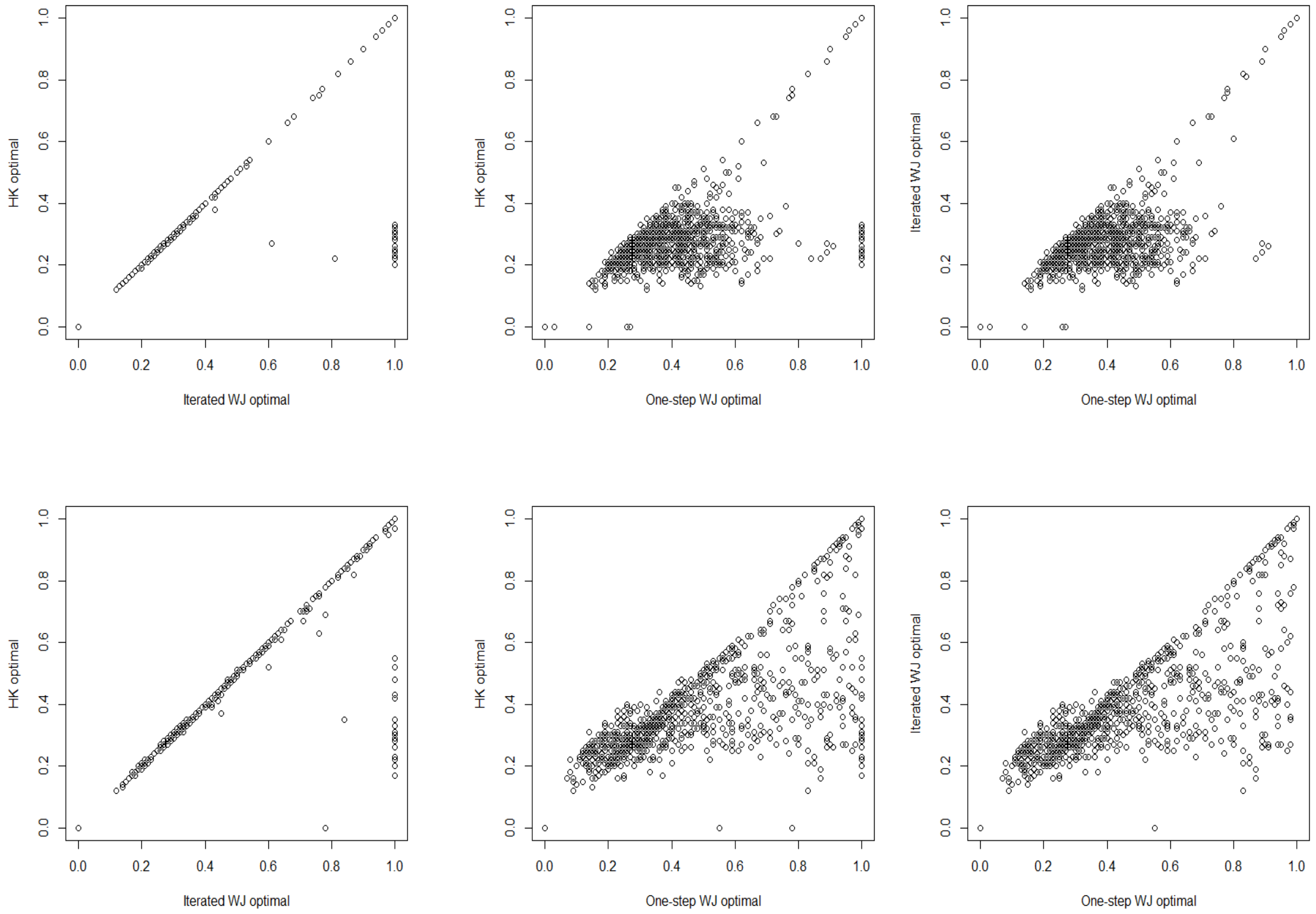


FIGURE 3.14: Comparison among optimal tuning parameters of the three algorithms for samples from $.9N(0,1) + .1N(8,1)$ (top panel) and $.9P(2) + .1P(15)$ (bottom panel) models.

From the graphs, we make the following observations.

1. In the pure data case, the HK and IWJ algorithms match for the vast majority of cases (Figure 3.13, left panel, top and bottom). For the normal data, all estimated optimal values of α are below 0.25 for either algorithm, except for one sample where HK yields $\alpha = 0$ and IWJ yields $\alpha = 1$. For the Poisson case also, the match between the two algorithms is near total. However in this case, some larger optimal values of α under the IWJ algorithm are observed, and there are at least eight cases where the HK solution equals a small value (≤ 0.3) but the IWJ solution equals 1.

2. Comparison with the OWJ algorithm demonstrates that the IWJ and HK algorithms lead to smaller optimal values in practically all the cases involving pure data. In particular, the latter two algorithms often lead to $\alpha = 0$ as the optimal value while OWJ leads to positive, sometimes fairly large positive, optimal values.

3. An inspection of the graphs in Figure 3.14 shows that under contamination, the optimal values of α, are, in general, higher for all three algorithms. There are extremely few samples with $\alpha = 0$ as the optimal solution, even for the HK algorithm.

4. For both the contaminated normal and Poisson models, HK and IWJ again show a high degree of match. But occasional cases where the HK solution is a low value and the IWJ solution (or the OWJ solution) equals $\alpha = 1$, indicate the occasional failure of the HK scheme.

5. In the case of the normal model, the OWJ optimal value is, in general, higher than the IWJ optimal value or the HK optimal value under contamination. There are quite a few cases where the OWJ optimal corresponds to the minimum L_2 estimate, whereas the HK as well as the IWJ optimals are substantially smaller than the OWJ optimal.

6. For the contaminated Poisson model, certain very small OWJ optimal values are associated with comparatively larger HK (or IWJ) optimal values. But for larger OWJ optimal values, the corresponding HK or IWJ solutions are usually smaller.

A point to be noted here is that occasionally we will come across cases where the HK optimal solution will be $\alpha = 0$ but the IWJ algorithm leads to an optimal value of $\alpha = 1$. Consider the normal distribution part of our simulation setup. For pure data, we observe that between 4-5% of the time we encounter the phenomenon that the IWJ optimal is 1 and the HK optimal is 0. On the other hand, this never happens in our simulations in case of contaminated data. It may be worthwhile, though, to scrutinize one of the cases (under pure data) where the HK optimal is $\alpha = 0$ and the IWJ optimal is $\alpha = 1$. The histogram of the data for this particular sample and the corresponding asymptotic variance curve over α are given in Figure 3.15. Although the data are generated from the pure normal model, there is clearly a substantially longer tail on the left. The overlaid normal density curves show that the HK optimal value of $\alpha = 0$ (corresponding to the minimum asymptotic variance) tries to accommodate the entire data whereas the IWJ optimal at $\alpha = 1$ (corresponding to a local minimum) robustly fits the majority of the data ignoring the tail on the left.

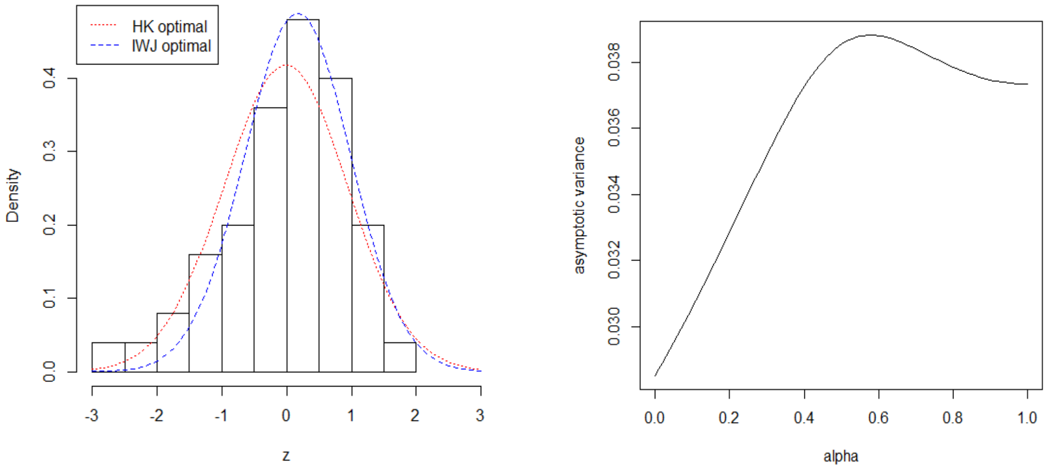


FIGURE 3.15: The histogram and the asymptotic variance curve of a particular dataset which leads to an optimal $\alpha = 0$ for HK algorithm and an optimal $\alpha = 1$ for IWJ algorithm.

3.5 Computational Cost

We have already demonstrated the advantages of the IWJ proposal in the previous sections of this chapter. However as the process is an iterative one, the experimenter would like to know about the computational cost involved in this procedure. To study this, we consider the number of iterations needed for the convergence of the procedure. Clearly a smaller number of iterations will indicate that the algorithm is more time-efficient. However, if the number of iterations is n, it does not mean that the computational complexity of the IWJ algorithm is n-times that of the OWJ algorithm, as all the ground work is done in the first step of the iteration including calculation of the estimates over a fine grid of α-values and the evaluation of the asymptotic variance curve. Thus, the subsequent iterations require only a very small fraction of the computational effort of the first iteration.

If we consider our real life data examples then in most of the cases, the IWJ algorithm converges in five iterations or fewer when starting from a robust pilot. For example, if the pilot is the MDPDE at $\alpha = 1$, the number of iterations in our real data examples (including the two cases of Example 3.2) are, in the order of the examples, 2, 2, 2, 5, 2, 5, 3, 2, 2, 6, 3. The worst case observed in these examples is in the Gesell Adaptive Score Data where the process takes six iterations to converge when starting from the MDPDE at $\alpha = 1$ as the robust pilot. However the computational time needed in this case is only 1.02 times the computational time required for the OWJ algorithm starting with the same pilot.

For our simulated data also, the process converges, most of the time (80% of the time or more), within 2-5 iterations, both for pure and contaminated data. In Figure 3.16, we present the frequency distributions of the number of iterations required for the IWJ algorithm to converge over 1000 replications for both pure and contaminated normal data. For pure data, the IWJ algorithm converges in just two iterations in more than 20% of the time although the mode of this frequency distribution is at four. In case of contaminated data also, the mode is at four but now the algorithm rarely converges in just two steps. On the other hand, in some rare cases, the algorithm may take a large number of iterations to converge in case of pure data. For example, in one stray pure data sample, the IWJ algorithm took 32 iterations to converge starting with the robust pilot at $\alpha = 1$. Further scrutiny shows that in this case, the asymptotic variance curve is very flat and the optimal solution is $\alpha = 0$; the passage of the algorithm from $\alpha = 1$ to $\alpha = 0$ over a flat asymptotic variance curve takes a while. Such large values are absent in the frequency distribution for contaminated data because in this case, the

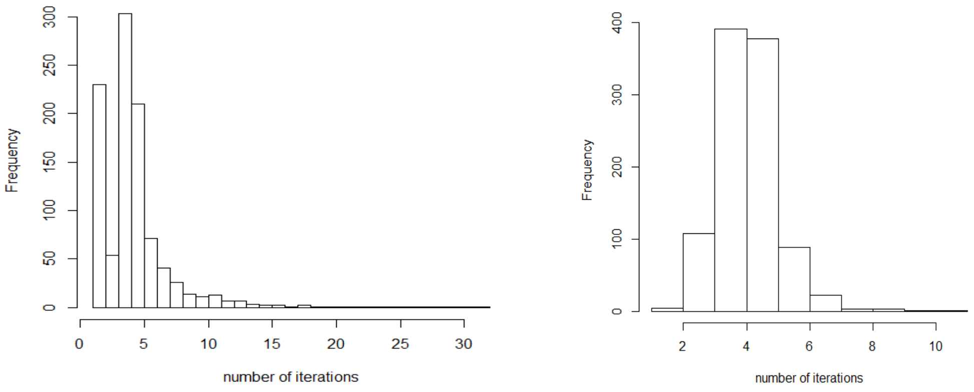


Figure 3.16: Number of iterations needed to obtain IWJ estimator under pure and contaminated normal model.

optimal solutions are generally substantially higher than zero which is reached in fewer steps starting from $\alpha = 1$. On the whole, the mean numbers of iterations for convergence are approximately the same for pure and contaminated data but the variance is larger in case of pure data.

3.6 Concluding Remarks

Here, we have proposed an iterated WJ algorithm for the selection of the optimal tuning parameter in the class of MDPDEs. Our findings show that when the pilot estimators are within the MDPDE class, all robust pilots lead to the same iterated optimal. In this sense, the iterated algorithm eliminates the dependence on the pilot estimator. The IWJ optimal value of α is frequently (but not always) equal to the HK optimal value. However, the advantage of the IWJ procedure is that it picks up a robust solution when it is appropriate but the HK algorithm fails to do so.

Our findings also indicate that for clean data the IWJ algorithm provides more suitable optimal values than the OWJ algorithm. For contaminated data, the one step and iterated algorithms give closer results.

On the whole, we feel that the IWJ optimal solution is successful in eliminating dependence on the pilot estimator and provides a good robust outcome where necessary. It also provides more efficient optimal solutions under pure data compared to the OWJ algorithm. It is, therefore, without doubt the best of the three algorithms for choosing the tuning parameter in minimum DPD estimation with which we have been concerned in this chapter.

When outliers are present in the dataset, the IWJ algorithm, on an average, exhibits superior performance relative to the HK algorithm, while, for outlier-free datasets, the IWJ and HK algorithms generally behave similarly (in fact are frequently identical). On the other hand, when we compare the IWJ and OWJ algorithms, both perform similarly under contaminated data and and generally exhibit similar degrees of robustness. For pure data, however, the OWJ algorithm often gets caught up in the neighborhood of the robust pilot under consideration, while the IWJ leads to a much more efficient choice through repeated iterations. Thus the IWJ algorithm enjoys the best of both worlds – it behaves like the HK solution under pure data, and like the OWJ solution under contamination.

Chapter 4

The Extended Bregman Divergence

We have already discussed the usefulness of the Bregman divergence. To construct a wider class of divergences as well as to use its nice properties in case of generating better estimates, we are going to extend this divergence and generate the "Extended Bregman divergence".

4.1 Rationale behind this Extension

We have already shown that several important divergence families (mentioned in second chapter) can be represented as subfamilies of the class of Bregman divergences. Yet, there are several other important divergences, e.g., the PD family, the S-divergence family, etc., which cannot be represented in the Bregman form. We will try to expand the structure of the Bregman divergence so that the above mentioned divergences can be accommodated within the Bregman form with this expanded definition. This we will do by utilizing powers of densities as arguments, rather than the arguments themselves; this leads to the generalized class of the extended Bregman divergences which is one step ahead through the modification of existing popular tools for minimum distance approach used extensively in this literature. This extension allows us to express several existing divergence families as special cases of it, which is not possible through the ordinary Bregman divergence, together with generating larger super-families of divergences as special cases.

4.2 Proposal of the Extension

It is evident that only the convexity criterion of the function $\psi(\cdot)$ in Equation (2.1) is necessary for the non-negativity property of the divergence $D_\psi(\boldsymbol{x}, \boldsymbol{y})$ to hold. One could, therefore, consider other quantities as the arguments rather than the points themselves in this measure. Hence, as long as ψ remains convex, any set of arguments whose equivalence translates to the equivalence of $\boldsymbol{x}$ and $\boldsymbol{y}$ can be used in the distance expression. This observation may be used to extend the Bregman divergence to have the form

$$D_\psi(\boldsymbol{x}, \boldsymbol{y}) = \left\{ \psi\left(\boldsymbol{x}^k\right) - \psi\left(\boldsymbol{y}^k\right) - \left\langle \nabla\psi\left(\boldsymbol{y}^k\right), \boldsymbol{x}^k - \boldsymbol{y}^k \right\rangle \right\}. \quad (4.1)$$

$\nabla\psi\left(\boldsymbol{y}^k\right)$ be the gradient of ψ with respect to its argument, evaluated at $\boldsymbol{y}^k = \left(y_1^k, y_2^k, \ldots, y_d^k\right)^T$ and ψ is a strictly convex function, mapping $\mathcal{S}$ to $\mathbb{R}$, $\mathcal{S}$ being a convex subset of $\mathbb{R}^{+p}$. Since our main purpose is to utilize this extension in the field of statistics where the arguments, being probability density functions, are inherently non-negative, restricting the domain of ψ to $\mathbb{R}^{+p}$ does not cause any difficulty. It is also not difficult to see that many of the properties of the Bregman divergence in Equation (2.1), are retained by the extended version in Equation (4.1). However, we will not make use of these properties in the present research, so we do not discuss them here any further.

Consider the standard setup of parametric estimation where G is the true data generating distribution modeled by the parametric family $\mathcal{F} = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$. Let g and f_θ be the corresponding densities. Further we assume that both G and F_θ belong to $\mathcal{G}$, the class of all cumulative distribution functions having densities with respect to some appropriate dominating measure. Our aim is to estimate the unknown parameter θ by choosing the model density closest to the true density in the Bregman sense. The definition of ordinary Bregman divergences as given in Equation (2.2), useful as it is, does not include many well-known and popular divergences which are extensively used in the literature for different purposes including parameter estimation. As already mentioned, the PD family, the S-divergence family are prominent examples of this. An inspection of the Bregman form in Equation (2.2) indicates that the term which involves both densities g and f is of the form

$$\int g\,(x)\,\nabla\psi\,(f\,(x))\,dx. \tag{4.2}$$

Here, the density g is present only as a linear term having exponent one. Given a random sample $X_1, X_2, \ldots, X_n$ from the true distribution G, the term in Equation (4.2) can be empirically estimated by $\frac{1}{n}\sum \nabla\psi(f_\theta(X_i))$ (with $f = f_\theta$ under the parametric model) so that one can construct an empirical version of the divergence without any non-parametric density estimation. On the other hand, this restricts the class of divergences that are expressible in the Bregman form.

Using an extension in the spirit of Equation (4.1) may allow the construction of richer classes of divergences. With this aim, we define the *extended Bregman divergence* between two densities g and f as

$$D_{\psi}^{(k)}(g,f)=\int\left\{\psi\left(g^{k}(x)\right)-\psi\left(f^{k}(x)\right)-\left(g^{k}(x)-f^{k}(x)\right)\nabla\psi\left(f^{k}(x)\right)\right\}dx. \quad (4.3)$$

Apart from the requirement of strict convexity of the function ψ, this formulation also depends on a positive index k with which the density is exponentiated. For the rest of the thesis, the notation $D_{\psi}^{(k)}(\cdot,\cdot)$ will refer to this general form in Equation (4.3), of which the divergence in Equation (2.2) is a special case for $k=1$. Evidently, $D_{\psi}^{(k)}(g,f)\geq 0$ for any choices of densities f and g with respect to the same measure. Moreover, the fact that $D_{\psi}^{(k)}(g,f)=0$ if and only if $g=f$, holds true in this case due to non-negativity property of a density as well as the consideration of strict convexity of the function $\psi(\cdot)$.

4.3 Some Special Cases of the Extended Bregman Divergence

4.3.1 Power Divergence (PD)

One of the most important subfamilies of the class of disparities is the Cressie–Read family (Cressie and Read, 1984) of Power Divergences. If we take $\psi(x)=\frac{x^{1+\frac{B}{A}}}{B}$, $A=1+\lambda$, $B=-\lambda$ and $\lambda\in\mathbb{R}$, with $k=A$ in Equation (4.1), we get the PD family expressed through the equation

$$PD_{\lambda}(g,f)=\frac{1}{\lambda(\lambda+1)}\int\left\{g(x)\left(\frac{g(x)}{f(x)}\right)^{\lambda}-1\right\}dx, \lambda\in\mathbb{R}. \quad (4.4)$$

The PD class is a subfamily of chi-square type distances. The latter class of divergences has the form

$$\rho\left(g,f\right)=\int C\left(\delta\left(x\right)\right)f\left(x\right)dx, \tag{4.5}$$

where C is a strictly convex function and $\delta\left(x\right)=\frac{g(x)}{f(x)}-1$. The power divergence corresponds to the specific convex function

$$C\left(\delta\right)=\frac{\left(\delta+1\right)^{\lambda+1}-\left(\delta+1\right)}{\lambda\left(\lambda+1\right)}-\frac{\delta}{\lambda+1}. \tag{4.6}$$

This family belong to the class of disparities and hence is totally disjoint with the DPD family except the only common significant member the likelihood disparity, which occurs at the limiting case $\lambda\rightarrow 0$.

4.3.2 S-Divergence (SD)

To create a bridge between the PD and the DPD families, Ghosh et al. (2017) introduced this divergence with the form

$$SD_{(\alpha,\lambda)}\left(g,f_\theta\right)=\int\left\{\frac{1}{B}\left(g^{A+B}\left(x\right)-f_\theta^{A+B}\left(x\right)\right)-\left(g^{A}\left(x\right)-f_\theta^{A}\left(x\right)\right)\frac{A+B}{AB}f_\theta^{B}\left(x\right)\right\}dx. \tag{4.7}$$

If we take $\psi\left(x\right)=\frac{x^{1+\frac{B}{A}}}{B}$, $A=1+\lambda\left(1-\alpha\right)$, $B=\alpha-\lambda\left(1-\alpha\right)$, $A+B=1+\alpha$, $\alpha\geq -1$, $\lambda\in\mathbb{R}$ and $k=A$ in Equation (4.3), we get Equation (4.7). The minimum S-divergence functional, $T_{(\alpha,\lambda)}\left(G\right)$, can be defined as

$$SD_{(\alpha,\lambda)}\left(g,f_{T_{(\alpha,\lambda)}(G)}\right)=\min_{\theta\in\Theta}SD_{(\alpha,\lambda)}\left(g,f_\theta\right), \tag{4.8}$$

provided the minimum exists. The essential estimating equation required for obtaining this functional is

$$\int K\left(\delta\left(x\right)\right) f_{\theta}^{1+\alpha}\left(x\right) u_{\theta}\left(x\right) dx = 0, \tag{4.9}$$

where, $\delta\left(x\right) = \frac{g(x)}{f_{\theta}(x)} - 1$ and $K\left(\delta\right) = \frac{(\delta+1)^{A}-1}{A}$. This is one of the most useful divergence families in the domain of robust minimum distance inference due to its capacity to generate much more robust estimator(s) than the DPD and PD families. Through this extension of Bregman divergence, it is now possible to express the S-divergence as a special case of this extended Bregman divergence family.

4.3.3 S-Hellinger Divergence (SHD)

If we take $\psi\left(x\right) = \frac{2e^{\beta x}}{\beta^{2}}$ with $k \geq 0$ in Equation (4.3), it will generate an extension of the BED family having the form

$$BED_{\beta}^{(k)}\left(g, f_{\theta}\right) = \frac{2}{\beta}\int \left\{ e^{\beta f_{\theta}^{k}(x)}\left(f_{\theta}^{k}\left(x\right) - \frac{1}{\beta}\right) - e^{\beta f_{\theta}^{k}(x)} g^{k}\left(x\right) + \frac{e^{\beta g^{k}(x)}}{\beta}\right\} dx. \tag{4.10}$$

It can be easily shown that, as $\beta \to 0$ and $k = \frac{1+\alpha}{2}$, $\alpha \in [0,1]$, the application of L'Hospital's rule leads to (constant time) the S-Hellinger Distance (SHD) family with the form

$$SHD_{\alpha}\left(g, f_{\theta}\right) = \frac{2}{1+\alpha}\int \left(g^{\frac{1+\alpha}{2}}\left(x\right) - f_{\theta}^{\frac{1+\alpha}{2}}\left(x\right)\right)^{2} dx. \tag{4.11}$$

This was introduced as a special case of the S-divergence family. This family cannot be expressed through the normal expression of the Bregman divergence, but through this extension, we can express this S-Hellinger divergence family as a (limiting) member of the extended BED class which becomes a subclass of the extended Bregman class.

4.4 Concluding Remarks

Through this extension, the scope of bringing all divergences under one umbrella has been evidently extended. The next step is to use this extension for discovery of some broader class of divergences which enables us to proceed toward further refined analysis through more and more robust inferential procedures.

Chapter 5

A New Extended Bregman Super Family

5.1 Introduction

In the previous chapter, our aim was to extend the scope of the Bregman divergence by utilizing the powers of densities (rather than the densities themselves) as arguments; this leads to the generalized class of the extended Bregman divergences that can then be used to generate new divergences which could provide more refined tools for minimum divergence inference compared to the current state of the art. Note that the use of the Bregman divergence in statistics is relatively recent; the class of density power divergences defined earlier, is a prominent example of Bregman divergences having significant applications in statistical inference. Many minimum divergence procedures have natural robustness properties against data contamination and outliers. As our class of divergences become more and more rich and refined, we expect that better options for statistical data analysis involving parametric inference will be available.

The extended Bregman divergence allows us to express several existing divergence families as special cases of it, which is not possible through the ordinary Bregman divergence. Consequently, the extended Bregman idea can be used to generate large super-families of divergences containing, together with the existing divergences, many new and useful divergence families as special cases.

5.2 Generalized S-Bregman (GSB) Divergence

We have already seen that one of the most popular subclasses of Bregman divergences, the DPD family, and the most popular subclass of disparities, the PD family, have only one common member, i.e., the likelihood disparity. To connect these two families, Ghosh et al. (2017) developed a new class of divergences, namely, the S-divergence class, which contains both families as special cases. Apart from acting as a bridge between these two families, this class also serves the purpose of getting more stable minimum distance estimators based on divergences lying outside both the PD and the DPD families (but within the S-divergence class).

The S-divergences were developed by Ghosh et al. (2017) from first principles balancing different divergence considerations. We have, in the previous chapter of this thesis, demonstrated that this S-divergence class is embedded within the extended Bregman family, and thus can be directly recovered from the latter. In the present section we will consider a refinement of the S-divergence class, within the framework of extended Bregman divergences, which is a rich class of divergences and combines the PD, DPD and BED families

(Section 2.1.3.4) in a single coherent class. This is a three tuning parameter family, each of which can be linked to one of the constituent families (α for the DPD, λ for the PD and β for the BED) and can serve as the source of new divergences which can provide better compromises between efficiency and robustness compared to the S-divergence class. With this motivation we are now going to generate this new super-divergence family, and refer to it as the class of generalized S-Bregman (GSB for short) divergences. In constructing the GSB divergence, we use the convex function $\psi(x) = e^{\beta x} + \frac{x^{1+\frac{B}{A}}}{B}$, $A = 1 + \lambda(1-\alpha)$, $B = \alpha - \lambda(1-\alpha)$, $A + B = 1 + \alpha$, $\alpha \geq -1$, $\beta, \lambda \in \mathbb{R}$, which, together with the exponent $k = A$, generates a divergence with the form

$$D^*(g,f) = \int \left\{ e^{\beta f^A}\left(\beta f^A - \beta g^A - 1\right) + e^{\beta g^A} + \frac{1}{B}\left(g^{A+B} - f^{A+B}\right) - \left(g^A - f^A\right)\frac{A+B}{AB} f^B \right\} dx. \tag{5.1}$$

The divergence measure D^* is our GSB divergence. It is a function of α, λ and β, which we suppress for brevity on the left hand side of Equation (5.1).

If we put $A+B = 0$ in the above expression with $A \neq 0$ and $B \neq 0$, we will get the extended BED family with parameter β and exponent $k = A$. Moreover, if $A = 1$, i.e., $\lambda = 0$, then it will lead to the ordinary BED family with parameter β. However, if we put $\beta = 0$, it will lead to the S-divergence family with parameters α and λ (in terms of A and B). More specifically, when $\alpha = 0$ and $\beta = 0$, it leads to the PD family with parameter λ. On the other hand, $\beta = 0$ and $\lambda = 0$ recovers the DPD family with parameter α. Thus, it acts as a connector between the BED and the S-divergence family.

5.3 Special Cases

We will get several well-known divergences or divergence families from the general form of the GSB divergence for particular choices of the three tuning parameters α, λ and β. Some such choices are given in Table 5.1.

TABLE 5.1: Different divergences as special cases of GSB divergence

α	λ	β	Divergences
$\alpha = -1$	$\lambda = 0$	$\beta \in \mathbb{R}$	Bregman Exponential Divergence[a]
$\alpha = 0$	$\lambda \in \mathbb{R}$	$\beta = 0$	Power Divergence
$\alpha \geq 0$	$\lambda = 0$	$\beta = 0$	Density Power Divergence
$\alpha \geq 0$	$\lambda \in \mathbb{R}$	$\beta = 0$	S-Divergence
$\alpha = 0$	$\lambda = -1$	$\beta = 0$	Kullback-Liebler Divergence
$\alpha = 0$	$\lambda = 0$	$\beta = 0$	Likelihood Disparity
$\alpha = 0$	$\lambda = -.5$	$\beta = 0$	Hellinger Distance
$\alpha \in \mathbb{R}$	$\lambda = -.5$	$\beta = 0$	S-Hellinger Distance
$\alpha = 0$	$\lambda = 1$	$\beta = 0$	Pearson's Chi-square Divergence
$\alpha = 0$	$\lambda = -2$	$\beta = 0$	Neyman's Chi-square Divergence
$\alpha = 1$	$\lambda \in \mathbb{R}$	$\beta = 0$	(squared) L_2 Distance

[a] This is a constant times the B-exponential divergence. It basically generates all the members of the BED family corresponding to the same β except the (squared) L_2 distance, which occurs when $\beta \to 0$. However, as seen above, the (squared) L_2 distance remains a member of the GSB class for other choices of the tuning parameters.

5.4 Discrete Setup

In this section, we will focus on the discrete setup for parametric estimation based on the GSB divergence. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed observations from an unknown distribution G where the support is taken, without loss of generality, to be $\chi = \{0, 1, 2, 3, \ldots, \}$. On the other hand, we consider a parametric family of distributions $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$, also supported on χ, to model the true data generating distribution G. In this setup, we assume both G and $\mathcal{F}$ to have densities g and f_θ with respect to the counting measure. Let the best fitting parameter be $\theta^g = T_{\alpha,\beta,\lambda}(G)$, and we are interested in estimating the parameter θ.

5.4.1 The Minimum GSB Divergence Estimator

Under the parametric setup described in this section, we would like to identify the best fitting parameter θ^g by choosing the element of the model family of distributions which provides the closest match to the true density g in terms of the given divergence. The minimum GSB divergence functional $T_{\alpha,\lambda,\beta} : \mathcal{G} \to \Theta$ is defined by the relation

$$D^* \left(g, f_{T_{\alpha,\lambda,\beta}}\right) = \min\{D^* \left(g, f_\theta\right) : \theta \in \Theta\},$$

provided the minimum exists. If the parametric model family is identifiable, it follows from the definition of the divergence that $D^* \left(g, f_\theta\right) = 0$, if and only if, $g = f_\theta$. Thus, $T_{\alpha,\lambda,\beta} \left(F_\theta\right) = \theta$, uniquely. Hence, we can conclude that the functional $T_{\alpha,\lambda,\beta}$ is Fisher consistent.

5.4.2 Estimating Equation

To find the best fitting parameter, a straightforward differentiation of the GSB divergence of Equation (5.1) (given the density g) leads to the estimating equation

$$\int \left\{ A\beta^2 e^{\beta f_\theta^A(x)} f_\theta^A\left(x\right) + \frac{(A+B)}{A} f_\theta^B\left(x\right) \right\} \left(f_\theta^A\left(x\right) - g^A\left(x\right) \right) u_\theta\left(x\right) dx = 0. \tag{5.2}$$

In practice, the true density g is unknown, so one has to use a suitable non-parametric density estimator $\hat{g}$ for g, depending on the situation.

Since we concentrate on the discrete parametric setup only, the natural choice for $\hat{g}$ is the vector of relative frequencies as obtained from the sample data. Thus the estimating equation becomes

$$\begin{aligned} &\sum_{x=0}^{\infty} A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x)\, u_\theta(x) + \sum_{x=0}^{\infty} (A+B)\, f_\theta^{A+B}(x)\, u_\theta(x) \\ =\ &\sum_{x=0}^{\infty} A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{A}(x)\, \hat{g}^A(x)\, u_\theta(x) + \sum_{x=0}^{\infty} (A+B)\, f_\theta^{B}(x)\, \hat{g}^A(x)\, u_\theta(x)\,. \end{aligned} \tag{5.3}$$

For $A = 1$, Equation (5.3) reduces to

$$\begin{aligned} &\sum_{x=0}^{\infty} \beta^2 e^{\beta f_\theta(x)} f_\theta^{2}(x)\, u_\theta(x) + \sum_{x=0}^{\infty} (1+B)\, f_\theta^{1+B}(x)\, u_\theta(x) \\ =\ &\frac{1}{n}\sum_{i=1}^{n} \beta^2 e^{\beta f_\theta(X_i)} f_\theta(X_i)\, u_\theta(X_i) + \frac{1}{n}\sum_{i=1}^{n} (1+B)\, f_\theta^{B}(X_i)\, u_\theta(X_i)\,. \end{aligned} \tag{5.4}$$

Since the left hand side of the above equation is non-random and the right hand side is a sum of i.i.d. terms, it is of the form $\sum_{i=1}^{n} \psi_\theta(X_i) = 0$ and the corresponding estimator belongs to the M-estimator class.

In accordance with the information on the first three rows of Table 5.1, we will refer to the parameters α, λ and β as the DPD parameter, the PD parameter and the BED parameter, respectively.

5.4.3 Asymptotic Properties

Under the discrete setup employed in this section, the minimum GSB divergence estimator is obtained as a solution of the estimating equation

$$\begin{aligned}
&\sum_{x=0}^{\infty}\left\{A\beta^2 e^{\beta f_\theta^A(x)} f_\theta^A(x)+\frac{(A+B)}{A} f_\theta^B(x)\right\}\left(f_\theta^A(x)-\hat{g}^A(x)\right)u_\theta(x)=0\\
\Rightarrow\ &\sum_{x=0}^{\infty}\left\{A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x)+(A+B)\, f_\theta^{A+B}(x)\right\}\frac{\left(1-\frac{\hat{g}^A(x)}{f_\theta^A(x)}\right)}{A}u_\theta(x)=0\\
\Rightarrow\ &\sum_{x=0}^{\infty}K(\delta(x))\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x)+(A+B)\, f_\theta^{A+B}(x)\right)u_\theta(x)=0, \qquad (5.5)
\end{aligned}$$

where, $\delta(x)=\delta_n(x)=\frac{\hat{g}(x)}{f_\theta(x)}-1=\frac{r_n(x)}{f_\theta(x)}-1$, $K(\delta)=\frac{(\delta+1)^A-1}{A}$ and $u_\theta(x)$ is the score function at x. We denote the minimum GSB divergence estimator, obtained as a solution of the above equation, as $\hat{\theta}$. Let

$$\begin{aligned}
J_g &= E_g\left(u_{\theta^g}(X)\,u_{\theta^g}^T(X)\,K'\left(\delta_g^g(X)\right)\left((A+B)\, f_{\theta^g}^{\alpha}(X)+A^2\beta^2 e^{\beta f_{\theta^g}^A(X)} f_{\theta^g}^{2A-1}(X)\right)\right)\\
&+ \sum_{x=0}^{\infty}K\left(\delta_g^g(x)\right)\left((A+B)\, f_{\theta^g}^{A+B}(x)+A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x)\right)i_{\theta^g}(x)\\
&- \sum_{x=0}^{\infty}K\left(\delta_g^g(x)\right)\left((A+B)^2 f_{\theta^g}^{A+B}(x)+A^3\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x)\left(2+\beta f_{\theta^g}^A(x)\right)\right)u_{\theta^g}(x)\,u_{\theta^g}^T(x)\\
V_g &= Var_g\left(u_{\theta^g}(X)\,K'\left(\delta_g^g(X)\right)\left((A+B)\, f_{\theta^g}^{\alpha}(X)+A^2\beta^2 e^{\beta f_{\theta^g}^A(X)} f_{\theta^g}^{2A-1}(X)\right)\right), \qquad (5.6)
\end{aligned}$$

where X is a random variable having density g, Var_g represents variance under the density g, $\theta=\theta^g$, $\delta_g(x)=\frac{g(x)}{f_\theta(x)}-1$, $K'(\cdot)$ is the derivative of $K(\cdot)$ with respect to its argument, $\delta_g^g(x)=\frac{g(x)}{f_{\theta^g}(x)}-1$ and $i_\theta(x)=-u'_\theta(x)$, the negative of the derivative of the score function with respect to the parameter.

First we set up some regularity conditions:

A1. The model family $\mathcal{F}$ is identifiable.

A2. The model distribution as well as the true distribution have the same support χ, which is independent of the parameter θ.

A3. There exists an open subset $\omega \subset \Theta$, of which θ^g is an interior point. For almost all x, $f_\theta(x)$ possesses third partial derivative of the type $\nabla_{jkl} f_\theta(x)$ for all $\theta \in \omega$.

A4. The matrix J_g is positive definite.

A5. $\sum_{x=0}^{\infty} g^{\frac{1}{2}}(x)\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A-1}(x) + (A+B) f_\theta^{A+B-1}(x)\right)|u_\theta(x)|$,
$\sum_{x=0}^{\infty} g^{\frac{1}{2}}(x)\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A-1}(x) + (A+B) f_\theta^{A+B-1}(x)\right)|u_{j\theta}(x)||u_{k\theta}(x)|$
and
$\sum_{x=0}^{\infty} g^{\frac{1}{2}}(x)\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A-1}(x) + (A+B) f_\theta^{A+B-1}(x)\right)|u_{jk\theta}(x)|$
are bounded for all j, k and for all $\theta \in \omega$. Here, $u_{j\theta}(x)$ and $u_{jk\theta}(x)$ denote the j-th element of $\nabla \log f_\theta(x)$ and (j,k)-th element of $\nabla^2 \log f_\theta(x)$, respectively.

A6. For almost all x, there exists $M_{j,k,l}(x)$, $M_{jk,l}(x)$, $M_{jkl}(x)$, $M^{(1)}_{j,k,l}(x)$ and $M^{(2)}_{j,k,l}(x)$ such that they dominate
$\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A-1}(x) + (A+B) f_\theta^{A+B-1}(x)\right) u_{j\theta}(x)\, u_{k\theta}(x)\, u_{l\theta}(x)$,
$\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A-1}(x) + (A+B) f_\theta^{A+B-1}(x)\right) u_{jk\theta}(x)\, u_{l\theta}(x)$,
$\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A-1}(x) + (A+B) f_\theta^{A+B-1}(x)\right) u_{jkl\theta}(x)$,
$\left\{(A+B)^2 f_\theta^{A+B-1}(x) + A^3\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A-1}(x)\left(2+\beta f_\theta^A(x)\right)\right\} u_{j\theta}(x)\, u_{k\theta}(x)\, u_{l\theta}(x)$,
$(A+B)^3 f_\theta^{A+B-1}(x)\, u_{j\theta}(x)\, u_{k\theta}(x)\, u_{l\theta}(x)$
$+\left\{A^4\beta^2 e^{\beta f_\theta^A(x)}\left(2 f_\theta^{2A-1}(x) + \beta\left(f_\theta^{3A-1}(x) + f_\theta^{A-1}(x) + 4 f_\theta^{2A-1}(x)\right)\right)\right\} u_{j\theta}(x)\, u_{k\theta}(x)\, u_{l\theta}(x)$
in absolute value, respectively for all j, k and l and they are uniformly bounded in expectation with respect to g and f_θ for all $\theta \in \omega$.

A7. Suppose, C_1 and C_2 represent the bounds of $K'(\delta)$ and $K''(\delta)(1+\delta)$, respectively, where $K'(\cdot)$ and $K''(\cdot)$ represent the first and second order derivatives of $K(\cdot)$ with respect to its argument δ.

Next, we are going to state (and prove) some lemmas required for establishing the main theoretical result (Theorem 5.5). At first we define the *Hellinger Residuals* as

$$\triangle_n(x) = \frac{r_n^{1/2}(x)}{f_\theta^{1/2}(x)} - 1; \triangle_g(x) = \frac{g^{1/2}(x)}{f_\theta^{1/2}(x)} - 1. \tag{5.7}$$

Lemma 5.1. *Define,* $\eta_n(x) = \sqrt{n}\left(\triangle_n(x) - \triangle_g(x)\right)^2$. *For any* $k \in [0,2]$ *and for any* $x \in \chi$*, we have,*

1. $E_g\left\{\eta_n^k(x)\right\} \leq n^{\frac{k}{2}} E_g\left\{|\delta_n(x) - \delta_g(x)|^k\right\} \leq \left[\frac{g(x)(1-g(x))}{f_\theta^2(x)}\right]^{\frac{k}{2}}$.
2. $E_g\left\{|\delta_n(x) - \delta_g(x)|\right\} \leq \left[2\frac{g(x)(1-g(x))}{f_\theta(x)}\right]$.

Proof. The proof follows the arguments of Lemma 2.13 of Basu et al. (2011). For non-negative quantities a, b, we have $\left(\sqrt{a} - \sqrt{b}\right)^2 \leq |a-b|$. Again, under g, $nr_n(x) \sim Bin(n, g(x))$. Using these facts, we find

$$
\begin{aligned}
E_g\left\{\eta_n^k(x)\right\} &= n^{\frac{k}{2}} E_g \left(\frac{r_n^{1/2}(x)}{f_\theta^{1/2}(x)} - \frac{g^{1/2}(x)}{f_\theta^{1/2}(x)} \right)^{2k} \\
&\leq n^{\frac{k}{2}} E_g \left\{ \left| \delta_n(x) - \delta_g(x) \right| \right\}^k \\
&\leq n^{\frac{k}{2}} E_g \left\{ \left(\delta_n(x) - \delta_g(x) \right)^2 \right\}^{\frac{k}{2}} \\
&\quad \text{for} \ \ k \in [0, 2], \quad \text{(by Lyapunov's inequality)} \\
&= \frac{n^{\frac{k}{2}}}{f_\theta^k(x)} E_g \left\{ \left(r_n(x) - g(x) \right)^2 \right\}^{\frac{k}{2}} \\
&= \frac{n^{\frac{k}{2}}}{f_\theta^k(x)} \left\{ \frac{g(x)(1 - g(x))}{n} \right\}^{\frac{k}{2}} \\
&= \left\{ \frac{g(x)(1 - g(x))}{f_\theta^2(x)} \right\}^{\frac{k}{2}}. \qquad (5.8)
\end{aligned}
$$

Hence the first part is proved. For second part, we can write,

$$
\begin{aligned}
E_g\left\{ \left| \delta_n(x) - \delta_g(x) \right| \right\} &= \frac{1}{f_\theta(x)} E_g \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} I(X_i = x) - g(x) \right| \right\} \\
&\leq \frac{1}{n f_\theta(x)} \sum_{i=1}^{n} E_g \left\{ \left| I(X_i = x) - g(x) \right| \right\} \\
&= \frac{2g(x)(1 - g(x))}{f_\theta(x)}, \qquad (5.9)
\end{aligned}
$$

where the last relation holds from the results regarding the mean deviation of a Binomial random variable. □

Lemma 5.2. $E_g\left\{\eta_n^k(x)\right\} \to 0$, *as* $n \to \infty$, *for* $k \in [0, 2)$ *and for* $x \in \chi$.

Proof. By Lemma 2.9 of Basu et al (2011), we have, as $n \to \infty$,

$$
n^{\frac{1}{4}} \left(\sqrt{r_n(x)} - \sqrt{g(x)} \right) \to 0, \qquad (5.10)
$$

with probability one for each x belonging to the given support. Since $f(x) = x^2$ is a continuous function, by Continuous Mapping Theorem, we can further claim that, with probability one,

$$\begin{aligned} & n^{\frac{1}{2}}\left(\sqrt{r_n(x)} - \sqrt{g(x)}\right)^2 \to 0 \\ \Rightarrow \;\; & \eta_n(x) \to 0. \end{aligned}$$

Moreover, from the previous lemma, we have got that $\sup_n E_g\left\{\eta_n^k(x)\right\}$ is bounded for $k \in [0,2)$. Hence the remaining part immediately follows from the Theorem 4.5.2 of Chung (1974). □

Let us now define,

$$\begin{aligned} a_n(x) &= K(\delta_n(x)) - K(\delta_g(x)) \\ b_n(x) &= (\delta_n(x) - \delta_g(x))\, K'(\delta_g(x)) \\ \tau_n(x) &= \sqrt{n}|a_n(x) - b_n(x)|. \end{aligned}$$

Next, we are going to find the asymptotic distribution of

$$S_{1n} = \sqrt{n}\sum_{x=0}^{\infty} a_n(x)\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x) + (A+B)\, f_\theta^{A+B}(x)\right) u_\theta(x)$$

and,

$$S_{2n} = \sqrt{n}\sum_{x=0}^{\infty} b_n(x)\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x) + (A+B)\, f_\theta^{A+B}(x)\right) u_\theta(x).$$

Lemma 5.3. *Under assumption (A5), as $n \to \infty$, $E_g|S_{1n} - S_{2n}| \to 0$ and, hence, $S_{1n} - S_{2n} \xrightarrow{p} 0$.*

Proof. Let us consider,

$$c = \sqrt{\delta_n(x) + 1}; \quad d = \sqrt{\delta_g(x) + 1}, \tag{5.11}$$

then, using Lemma 2.15 of Basu et al. (2011), we can claim that, for some positive constant γ, we can write,

$$\begin{aligned}
& |K(c^2-1)-K(d^2-1)-(c^2-d^2)K(d^2-1)| \leq \gamma(c-d)^2 \\
\Rightarrow\quad & |K(\delta_n(x))-K(\delta_g(x))-(\delta_n(x)-\delta_g(x))K'(\delta_g(x))| \leq \gamma\left\{\frac{r_n^{1/2}(x)}{f_\theta^{1/2}(x)}-\frac{g^{1/2}(x)}{f_\theta^{1/2}(x)}\right\}^2 \\
\Rightarrow\quad & \tau_n(x) \leq \gamma\sqrt{n}(\triangle_n(x)-\triangle_g(x))^2 = \gamma\eta_n(x)\,. \qquad (5.12)
\end{aligned}$$

By Lemma 5.1,
$E_g\left\{\tau_n(x)\right\} \leq \gamma E_g\left(\eta_n(x)\right) \leq \gamma\frac{(g(x)(1-g(x))^{\frac{1}{2}}}{f_\theta(x)} \leq \gamma\frac{g^{\frac{1}{2}}(x)}{f_\theta(x)}$.
By Lemma 5.2,
$E_g\left\{\tau_n(x)\right\} \leq \gamma E_g\left\{\eta_n(x)\right\} \to 0$ as $n\to\infty$.
Therefore, for finite α, β and λ,

$$\begin{aligned}
E_g|S_{1n}-S_{2n}| &\leq \sum_{x=0}^{\infty} E_g\{\tau_n(x)\}\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x) + (A+B)\, f_\theta^{A+B}(x)\right)|u_\theta(x)| \\
&\leq \gamma\sum_{x=0}^{\infty} g^{\frac{1}{2}}(x)\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A-1}(x) + (A+B)\, f_\theta^{\alpha}(x)\right)|u_\theta(x)| < \infty,
\end{aligned}$$

by assumption (A5). Hence, by the dominated convergence theorem (DCT), $E_g|S_{1n}-S_{2n}| \to 0$ as $n\to\infty$. The desired result then follows from Markov's inequality. □

Lemma 5.4. *Let all the relevant expressions be evaluated at $\theta = \theta^g$, and let V_g be as defined in Equation (5.6). Then, under g, S_{1n} converges in distribution to $N_p(0, V_g)$, whenever V_g is finite.*

Proof. From Lemma 5.3, we can say that the asymptotic distributions of S_{1n} and S_{2n} are same. Under $\theta = \theta^g$,

$$\begin{aligned} S_{2n} &= \sqrt{n}\sum_{x=0}^{\infty}\left(\delta_n\left(x\right)-\delta_g\left(x\right)\right)K'\left(\delta_g\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^{A}(x)} f_{\theta^g}^{2A}\left(x\right)+\left(A+B\right)f_{\theta^g}^{A+B}\left(x\right)\right)u_{\theta^g}\left(x\right) \\ &= \sqrt{n}\sum_{x=0}^{\infty}\left(r_n\left(x\right)-g\left(x\right)\right)K'\left(\delta_g\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^{A}(x)} f_{\theta^g}^{2A-1}\left(x\right)+\left(A+B\right)f_{\theta^g}^{\alpha}\left(x\right)\right)u_{\theta^g}\left(x\right) \\ &= \sqrt{n}\Bigg[\frac{1}{n}\sum_{i=1}^{n}K'\left(\delta_g\left(X_i\right)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^{A}(X_i)} f_{\theta^g}^{2A-1}\left(X_i\right)+\left(A+B\right)f_{\theta^g}^{\alpha}\left(X_i\right)\right)u_{\theta^g}\left(X_i\right) \\ &- E_g\left\{K'\left(\delta_g\left(X\right)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^{A}(X)} f_{\theta^g}^{2A-1}\left(X\right)+\left(A+B\right)f_{\theta^g}^{\alpha}\left(X\right)\right)u_{\theta^g}\left(X\right)\right\}\Bigg] \\ &\rightarrow Z\sim N_p\left(0,V_g\right), \end{aligned} \tag{5.13}$$

in distribution by the central limit theorem. Hence the proof. □

Now, we will prove the main asymptotic result regarding the minimum GSB divergence estimator using the given conditions and the results established so far.

Theorem 5.5. *Under the setup described in this section and regularity conditions (A1)-(A7), there exists a consistent sequence of roots $\hat{\theta}_n$ of the estimating equation (5.3). Moreover, the asymptotic distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta^g\right)$ is p-dimensional normal with (vector) mean* 0 *and covariance matrix $J_g^{-1}V_gJ_g^{-1}$.*

Proof. First we are going to establish the consistency part and then the asymptotic normality.

Proof of Consistency: Consider the behaviour of $D^*\left(r_n, f_\theta\right)$ on a sphere Q_a which has radius a and centre at θ^g. We will show that for a sufficiently small, $P\left(D^*\left(r_n, f_\theta\right) > D^*\left(r_n, f_{\theta^g}\right)\right) \rightarrow 1$ for all θ on the surface of Q_a, so that asymptotically there will be a local minimum of the GSB divergence with respect to θ in the interior of Q_a. Therefore, for any $a > 0$ sufficiently small, the minimum GSB divergence estimating equation has a solution θ_n within Q_a with probability tending to one where the estimating equation must be satisfied. Now, considering a Taylor series expansion of $D^*\left(r_n, f_\theta\right)$ around $\theta = \theta^g$, we get

$$\begin{aligned}
D^*\left(r_n, f_{\theta^g}\right) - D^*\left(r_n, f_\theta\right) = & - \sum_{j=1}^{p} \left(\theta_j - \theta_j^g\right) \nabla_j D^*\left(r_n, f_\theta\right)|_{\theta=\theta^g} \\
& - \frac{1}{2}\sum_{j=1}^{p} \left(\theta_j - \theta_j^g\right)\left(\theta_k - \theta_k^g\right) \nabla_{jk} D^*\left(r_n, f_\theta\right)|_{\theta=\theta^g} \\
& - \frac{1}{6}\sum_{j=1}^{p} \left(\theta_j - \theta_j^g\right)\left(\theta_k - \theta_k^g\right)\left(\theta_l - \theta_l^g\right) \nabla_{jkl} D^*\left(r_n, f_\theta\right)|_{\theta=\theta^*} \\
= & \ S_1 + S_2 + S_3, say \qquad (5.14)
\end{aligned}$$

where, θ^* lies between θ and θ^g. For the first term S_1,

$$\begin{aligned}
& \nabla_j D^*\left(r_n, f_\theta\right)|_{\theta=\theta^g} \\
= & -\sum_{x=0}^{\infty} K\left(\delta_n^g(x)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x) + (A+B)\, f_{\theta^g}^{A+B}(x)\right) u_{j\theta^g}(x),
\end{aligned} \qquad (5.15)$$

where, $\delta_n^g(x)$ is the $\delta_n(x)$ evaluated at $\theta = \theta^g$. We will now show that,

$$\begin{aligned}
& \sum_{x=0}^{\infty} K\left(\delta_n^g(x)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x) + (A+B)\, f_{\theta^g}^{A+B}(x)\right) u_{j\theta^g}(x) \\
\to & \sum_{x=0}^{\infty} K\left(\delta_g^g(x)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x) + (A+B)\, f_{\theta^g}^{A+B}(x)\right) u_{j\theta^g}(x),
\end{aligned} \qquad (5.16)$$

in probability as $n \to \infty$. Note that, the right hand side of the above expression is zero by definition of the minimum GSB divergence estimator. Moreover, by assumption (A7) and considering the one-term

Taylor series expansion, we have,

$$\begin{aligned} & |\sum_{x=0}^{\infty} K\left(\delta_n^g(x)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x)+(A+B)\, f_{\theta^g}^{A+B}(x)\right) u_{j\theta^g}(x) \\ - & \sum_{x=0}^{\infty} K\left(\delta_g^g(x)\right) u_{j\theta^g}(x)\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x)+(A+B)\, f_{\theta^g}^{A+B}(x)\right)| \\ \leq & \; C_1 \sum_{x=0}^{\infty} |\delta_n^g(x)-\delta_g^g(x)||u_{j\theta^g}(x)|\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x)+(A+B)\, f_{\theta^g}^{A+B}(x)\right). \end{aligned} \tag{5.17}$$

Moreover, for finite α, β and λ, using the lemmas proved earlier in this chapter, we have

$$E|\delta_n^g(x)-\delta_g^g(x)| \leq \frac{|g(x)(1-g(x))|^{\frac{1}{2}}}{f_{\theta^g}(x)\sqrt{n}} \leq \frac{1}{2f_{\theta^g}(x)\sqrt{n}} \to 0 \text{ as } n \to \infty. \tag{5.18}$$

Since $0 \leq g(x) \leq 1$ for all x, by assumption (A5) and Lemma 5.1(2), it follows that,

$$\begin{aligned} & E\left[C_1 \sum_{x=0}^{\infty} |\delta_n^g(x)-\delta_g^g(x)||u_{j\theta^g}(x)|\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x)+(A+B)\, f_{\theta^g}^{A+B}(x)\right)\right] \\ \leq & \; 2C_1 \sum_{x=0}^{\infty} \sqrt{g(x)}|u_{j\theta^g}(x)|\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A-1}(x)+(A+B)\, f_{\theta^g}^{\alpha}(x)\right) < \infty. \end{aligned}$$

Once again we have the desired result by applying the dominated convergence theorem and Markov's inequality. As a consequence, we can say that

$$\nabla_j D^*(r_n, f_\theta)|_{\theta=\theta^g} \xrightarrow{p} 0$$

as $n \to \infty$. Since this is an $o_p(1)$ term, for sufficiently small a, we can say that $P\left(|S_1| < pa^3\right) \to 1$ as $n \to \infty$, where a is the radius of the sphere and p is its dimension.

Next, we come to the quadratic term S_2,

$$\begin{aligned}
& \nabla_{jk} D^*\left(r_n, f_\theta\right)|_{\theta=\theta^g} \\
= \ & \nabla_k \left(-\sum_{x=0}^{\infty} K\left(\delta_n^g(x)\right) \left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x) + (A+B) f_{\theta^g}^{A+B}(x)\right) u_{j\theta^g}(x)|_{\theta=\theta^g} \right) \\
= \ & -\left\{ \sum_{x=0}^{\infty} K'\left(\delta_n^g(x)\right)\left(-\left(1+\delta_n^g(x)\right)\right) \left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x) + (A+B) f_{\theta^g}^{A+B}(x)\right) u_{j\theta^g}(x)\, u_{k\theta^g}(x) \right. \\
+ \ & \sum_{x=0}^{\infty} K\left(\delta_n^g(x)\right) \left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x) + (A+B) f_{\theta^g}^{A+B}(x)\right) u_{jk\theta^g}(x) \\
+ \ & \left. \sum_{x=0}^{\infty} K\left(\delta_n^g(x)\right) \left((A+B)^2 f_{\theta^g}^{A+B}(x) + A^3\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x)\left(2+\beta f_{\theta^g}^A(x)\right)\right) u_{j\theta^g}(x)\, u_{k\theta^g}(x) \right\}.
\end{aligned}$$

Next, we are going to show,

$$\begin{aligned}
& - \sum_{x=0}^{\infty} K'\left(\delta_n^g(x)\right)\left(\left(1+\delta_n^g(x)\right)\right) \left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x) + (A+B) f_{\theta^g}^{A+B}(x)\right) u_{j\theta^g}(x)\, u_{k\theta^g}(x) \\
\rightarrow \ & - \sum_{x=0}^{\infty} K'\left(\delta_g^g(x)\right)\left(\left(1+\delta_g^g(x)\right)\right) \left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x) + (A+B) f_{\theta^g}^{A+B}(x)\right) u_{j\theta^g}(x)\, u_{k\theta^g}(x),
\end{aligned}$$

in probability as $n \to \infty$. Again, by assumption (A7) and one-term Taylor series expansion, we get,

$$\begin{aligned}
& |K'\left(\delta_n^g\right)\left(\left(\delta_n^g+1\right)\right) - K'\left(\delta_g^g\right)\left(\left(\delta_g^g+1\right)\right)| \\
\leq \ & |\delta_n^g - \delta_g^g||K''\left(\delta^*\right)\left(\delta^*+1\right) + K'\left(\delta^*\right)| \\
\leq \ & |\delta_n^g - \delta_g^g|\left(C_2 + C_1\right). \qquad (5.19)
\end{aligned}$$

Thus, we get,

$$\begin{aligned}
& |\sum_{x=0}^{\infty} K'\left(\delta_n^g(x)\right)\left(\left(1+\delta_n^g(x)\right)\right) \left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x) + (A+B) f_{\theta^g}^{A+B}(x)\right) u_{j\theta^g}(x)\, u_{k\theta^g}(x) \\
- \ & \sum_{x=0}^{\infty} K'\left(\delta_g^g(x)\right)\left(\left(1+\delta_g^g(x)\right)\right) \left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x) + (A+B) f_{\theta^g}^{A+B}(x)\right) u_{j\theta^g}(x)\, u_{k\theta^g}(x)| \\
\leq \ & \left(C_1 + C_2\right) \sum_{x=0}^{\infty} |\delta_n^g(x) - \delta_g^g(x)| \left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x) + (A+B) f_{\theta^g}^{A+B}(x)\right) u_{j\theta^g}(x)\, u_{k\theta^g}(x).
\end{aligned}$$

(5.20)

Again, by assumption (A5), Lemma 5.1(2) and an application of the DCT, we can prove our desired result. Similarly, we can prove

$$\begin{aligned}
&\sum_{x=0}^{\infty} K\left(\delta_n^g\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}\left(x\right)+\left(A+B\right) f_{\theta^g}^{A+B}\left(x\right)\right) u_{jk\theta^g}\left(x\right)\\
\rightarrow\ &\sum_{x=0}^{\infty} K\left(\delta_g^g\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}\left(x\right)+\left(A+B\right) f_{\theta^g}^{A+B}\left(x\right)\right) u_{jk\theta^g}\left(x\right),
\end{aligned}$$

and,

$$\begin{aligned}
&\sum_{x=0}^{\infty} K\left(\delta_n^g\left(x\right)\right)\left(\left(A+B\right)^2 f_{\theta^g}^{A+B}\left(x\right)+A^3\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}\left(x\right)\left(2+\beta f_{\theta^g}^{A}\left(x\right)\right)\right) u_{j\theta^g}\left(x\right) u_{k\theta^g}\left(x\right)\\
\rightarrow\ &\sum_{x=0}^{\infty} K\left(\delta_g^g\left(x\right)\right)\left(\left(A+B\right)^2 f_{\theta^g}^{A+B}\left(x\right)+A^3\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}\left(x\right)\left(2+\beta f_{\theta^g}^{A}\left(x\right)\right)\right) u_{j\theta^g}\left(x\right) u_{k\theta^g}\left(x\right).
\end{aligned}$$

Thus, combining all these three parts, we get,

$$\nabla_{jk} D^*\left(r_n, f_\theta\right)|_{\theta=\theta^g} \xrightarrow{p} J_g^{j,k},$$

where $J_g^{j,k}$ represents the (j,k)-th element of J_g. Now, we can write

$$2S_2 = \sum_{j,k=1}^{p}\left(\theta_j-\theta_j^g\right)\left(\theta_k-\theta_k^g\right)\left\{-\nabla_{jk} D^*\left(r_n, f_\theta\right)|_{\theta=\theta^g}-\left(-J_g^{j,k}\right)\right\}+\sum_{j,k=1}^{p}\left(\theta_j-\theta_j^g\right)\left(\theta_k-\theta_k^g\right)\left(-J_g^{j,k}\right).$$

Since the first expression of the right hand side of the above equation is an $o_p\left(1\right)$ term, we can say that this term is $< p^2a^3$ with probability tending to one. Letting μ_1 be the largest eigenvalue of $J_g^{j,k}$, the quadratic term given in the right hand side of the equation is $< \mu_1 a^2$. Combining these two, we can say that there exists a and c such that whenever, $a < a_0$ and $c = \frac{\left(ap^2+\mu_1\right)}{2}$, we have $S_2 < -ca^2$ with probability tending to one. Lastly, considering the third term S_3, we have,

$$
\begin{aligned}
&- \quad \nabla_{jkl} D^{*}\left(r_n, f_\theta\right)|_{\theta=\theta^*} \\
&= \quad \sum_{x=0}^{\infty} K''\left(\delta_n^*\left(x\right)\right)\left(1+\delta_n^*\left(x\right)\right)^2 \left\{A^2\beta^2 e^{\beta f_{\theta^*}^{A}(x)} f_{\theta^*}^{2A}\left(x\right) + \left(A+B\right) f_{\theta^*}^{A+B}\left(x\right)\right\} u_{j\theta^*}\left(x\right) u_{k\theta^*}\left(x\right) u_{l\theta^*}\left(x\right) \\
&- \quad \sum_{x=0}^{\infty} K'\left(\delta_n^*\left(x\right)\right)\left(1+\delta_n^*\left(x\right)\right) \left\{A^2\beta^2 e^{\beta f_{\theta^*}^{A}(x)} f_{\theta^*}^{2A}\left(x\right) + \left(A+B\right) f_{\theta^*}^{A+B}\left(x\right)\right\} u_{j\theta^*}\left(x\right) u_{k\theta^*}\left(x\right) u_{l\theta^*}\left(x\right) \\
&- \quad \sum_{x=0}^{\infty} K'\left(\delta_n^*\left(x\right)\right)\left(1+\delta_n^*\left(x\right)\right) \left\{A^2\beta^2 e^{\beta f_{\theta^*}^{A}(x)} f_{\theta^*}^{2A}\left(x\right) + \left(A+B\right) f_{\theta^*}^{A+B}\left(x\right)\right\} u_{jl\theta^*}\left(x\right) u_{k\theta^*}\left(x\right) \\
&- \quad \sum_{x=0}^{\infty} K'\left(\delta_n^*\left(x\right)\right)\left(1+\delta_n^*\left(x\right)\right) \left\{A^2\beta^2 e^{\beta f_{\theta^*}^{A}(x)} f_{\theta^*}^{2A}\left(x\right) + \left(A+B\right) f_{\theta^*}^{A+B}\left(x\right)\right\} u_{j\theta^*}\left(x\right) u_{kl\theta^*}\left(x\right) \\
&- \quad 2\sum_{x=0}^{\infty} K'\left(\delta_n^*\left(x\right)\right)\left(1+\delta_n^*\left(x\right)\right) \left\{\left(A+B\right)^2 f_{\theta^*}^{A+B}\left(x\right)\right\} u_{j\theta^*}\left(x\right) u_{k\theta^*}\left(x\right) u_{l\theta^*}\left(x\right) \\
&- \quad 2\sum_{x=0}^{\infty} K'\left(\delta_n^*\left(x\right)\right)\left(1+\delta_n^*\left(x\right)\right) \left\{A^3\beta^2 e^{\beta f_{\theta^*}^{A}(x)} f_{\theta^*}^{2A}\left(x\right)\left(2+\beta f_{\theta^*}^{A}\left(x\right)\right)\right\} u_{j\theta^*}\left(x\right) u_{k\theta^*}\left(x\right) u_{l\theta^*}\left(x\right) \\
&- \quad \sum_{x=0}^{\infty} K'\left(\delta_n^*\left(x\right)\right)\left(1+\delta_n^*\left(x\right)\right) \left\{A^2\beta^2 e^{\beta f_{\theta^*}^{A}(x)} f_{\theta^*}^{2A}\left(x\right) + \left(A+B\right) f_{\theta^*}^{A+B}\left(x\right)\right\} u_{jk\theta^*}\left(x\right) u_{l\theta^*}\left(x\right) \\
&+ \quad \sum_{x=0}^{\infty} K\left(\delta_n^*\left(x\right)\right) \left\{A^2\beta^2 e^{\beta f_{\theta^*}^{A}(x)} f_{\theta^*}^{2A}\left(x\right) + \left(A+B\right) f_{\theta^*}^{A+B}\left(x\right)\right\} u_{jkl\theta^*}\left(x\right) \\
&+ \quad \sum_{x=0}^{\infty} K\left(\delta_n^*\left(x\right)\right) \left\{\left(A+B\right)^2 f_{\theta^*}^{A+B}\left(x\right) + A^3\beta^2 e^{\beta f_{\theta^*}^{A}(x)} f_{\theta^*}^{2A}\left(x\right)\left(2+\beta f_{\theta^*}^{A}\left(x\right)\right)\right\} u_{jk\theta^*}\left(x\right) u_{l\theta^*}\left(x\right) \\
&+ \quad \sum_{x=0}^{\infty} K\left(\delta_n^*\left(x\right)\right) \left\{\left(A+B\right)^2 f_{\theta^*}^{A+B}\left(x\right) + A^3\beta^2 e^{\beta f_{\theta^*}^{A}(x)} f_{\theta^*}^{2A}\left(x\right)\left(2+\beta f_{\theta^*}^{A}\left(x\right)\right)\right\} u_{jl\theta^*}\left(x\right) u_{k\theta^*}\left(x\right) \\
&+ \quad \sum_{x=0}^{\infty} K\left(\delta_n^*\left(x\right)\right) \left\{\left(A+B\right)^2 f_{\theta^*}^{A+B}\left(x\right) + A^3\beta^2 e^{\beta f_{\theta^*}^{A}(x)} f_{\theta^*}^{2A}\left(x\right)\left(2+\beta f_{\theta^*}^{A}\left(x\right)\right)\right\} u_{j\theta^*}\left(x\right) u_{kl\theta^*}\left(x\right) \\
&+ \quad \sum_{x=0}^{\infty} K\left(\delta_n^*\left(x\right)\right) \left(A+B\right)^3 f_{\theta^*}^{A+B}\left(x\right) u_{j\theta^*}\left(x\right) u_{k\theta^*}\left(x\right) u_{l\theta^*}\left(x\right) \\
&+ \quad \sum_{x=0}^{\infty} K\left(\delta_n^*\left(x\right)\right) \left\{A^4\beta^2 e^{\beta f_{\theta^*}^{A}(x)} \left(2f_{\theta^*}^{2A}\left(x\right) + \beta\left(f_{\theta^*}^{3A}\left(x\right) + f_{\theta^*}^{A}\left(x\right) + 4f_{\theta^*}^{2A}\left(x\right)\right)\right)\right\} u_{j\theta^*}\left(x\right) u_{k\theta^*}\left(x\right) u_{l\theta^*}\left(x\right).
\end{aligned}
\tag{5.21}
$$

Now we are to show that the terms in right hand side of the above equation is bounded. At first, let them name as $t_1,\ t_2,\ \ldots,\ t_{13}$, respectively.

$$\begin{aligned}
|t_1| &\leq C_2 \sum_{x=0}^{\infty} |1 + \delta_n^* (x) | M_{j,k,l} (x) f_\theta^* (x) \\
&= C_2 \sum_{x=0}^{\infty} r_n (x) M_{j,k,l} (x) \\
&\rightarrow C_2 E_g M_{j,k,l} (X) < \infty,
\end{aligned}$$

by the central limit theorem. Hence, t_1 is bounded. Similarly, by assumption (A6),

$$\begin{aligned}
|t_2| &\leq C_1 \sum_{x=0}^{\infty} |1 + \delta_n^* (x) | M_{j,k,l} (x) f_\theta^* (x) \\
&= C_1 \sum_{x=0}^{\infty} r_n (x) M_{j,k,l} (x) \\
&\rightarrow C_1 E_g M_{j,k,l} (X) < \infty.
\end{aligned}$$

Hence, t_2 is bounded. Similarly, by assumption (A6), we can show that t_3, t_4, t_5, t_6 and t_7 are bounded. Now, it is to be noted that, $|K(\delta)| = |\int_0^\delta K'(\delta) d\delta| \leq C_1 |\delta|$ which implies $|K(\delta_n^*(x))| \leq C_1 \frac{r_n(x)}{f_{\theta^*}(x)}$. So,

$$\begin{aligned}
|t_8| &\leq C_1 \sum_{x=0}^{\infty} \frac{r_n(x)}{f_{\theta^*}(x)} \left(A^2 \beta^2 e^{\beta f_{\theta^*}^A (x)} f_{\theta^*}^{2A} (x) + (A+B) f_{\theta^*}^{A+B} (x) \right) u_{jkl\theta^*} (x) \\
&= C_1 \sum_{x=0}^{\infty} r_n (x) \left(A^2 \beta^2 e^{\beta f_{\theta^*}^A (x)} f_{\theta^*}^{2A-1} (x) + (A+B) f_{\theta^*}^{A+B-1} (x) \right) u_{jkl\theta^*} (x) \\
&\leq C_1 \sum_{x=0}^{\infty} r_n (x) M_{jkl} (x) \\
&\rightarrow C_1 E_g (M_{jkl} (X)) < \infty. \qquad (5.22)
\end{aligned}$$

Similarly, by assumption (A6), rest of the terms can be proved to be bounded. Hence, on the sphere Q_a, $P\left(|S_3| < ba^3\right) \rightarrow 1$, b being sufficiently small. Combining all, we get, $\max(S_1 + S_2 + S_3) < pa^3 - ca^2 + ba^3 < a^2((b+p)a - c)$. Hence, it will be < 0 whenever $a < \frac{c}{b+p}$. Thus if a is sufficiently small, there exists a sequence of roots θ_n

depending on a, such that,

$$\begin{aligned} P & \quad \left(||\theta_n - \theta^g||_2 < a\right) \\ = & \quad P\left(D^*\left(r_n, f_{\theta^g}\right) - D^*\left(r_n, f_\theta\right) < 0\right) \\ \to & \quad 1, \end{aligned}$$

where $||.||_2$ denotes the L_2 norm. The only part that remains necessary to establish is that ultimately θ_n is independent of a. This is evident from the fact that, by different choices of a, we get several sequence of roots and by the continuity of the GSB divergence, the limit exists and it will be again a root of our proposed divergence. Hence, the **consistency** part is proved.

Proof of Asymptotic Normality: Considering the Taylor series expansion of $\sum\limits_{x=0}^{\infty} K\left(\delta\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}\left(x\right) + \left(A+B\right) f_\theta^{A+B}\left(x\right)\right) u_\theta\left(x\right)$ about $\theta = \theta^g$, we get

$$\begin{aligned} & \sum_{x=0}^{\infty} K\left(\delta_n\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}\left(x\right) + \left(A+B\right) f_\theta^{A+B}\left(x\right)\right) u_\theta\left(x\right) \\ = & \sum_{x=0}^{\infty} K\left(\delta_n^g\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}\left(x\right) + \left(A+B\right) f_{\theta^g}^{A+B}\left(x\right)\right) u_{\theta^g}\left(x\right) \\ + & \sum_{k=1}^{p}\left(\theta_k - \theta_k^g\right)\nabla_k\left(\sum_{x=0}^{\infty} K\left(\delta_n\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}\left(x\right) + \left(A+B\right) f_\theta^{A+B}\left(x\right)\right) u_\theta\left(x\right)\right)|_{\theta=\theta^g} \\ + & \frac{1}{2}\sum_{k,l=1}^{p}\left(\theta_k - \theta_k^g\right)\left(\theta_l - \theta_l^g\right)\nabla_{kl}\left(\sum_{x=0}^{\infty} K\left(\delta_n\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}\left(x\right) + \left(A+B\right) f_\theta^{A+B}\left(x\right)\right) u_\theta\left(x\right)\right)|_{\theta=\theta^*}, \end{aligned}$$

where θ^* lies between θ and θ^g. Now, replacing θ by θ_n, θ_n being a root of estimating equation (5.2), the left hand side of the above expression comes out to be zero and hence we get

$$\begin{aligned}
&\sqrt{n}\sum_{x=0}^{\infty} K\left(\delta_n^g\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}\left(x\right) + \left(A+B\right) f_{\theta^g}^{A+B}\left(x\right)\right) u_{\theta^g}\left(x\right) \\
= \;& \sqrt{n}\sum_{k=1}^{p}\left(\theta_{n,k}-\theta_k^g\right)\Bigg\{ -\nabla_k\left(\sum_{x=0}^{\infty} K\left(\delta_n\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_{\theta}^A(x)} f_{\theta}^{2A}\left(x\right) + \left(A+B\right) f_{\theta}^{A+B}\left(x\right)\right) u_{\theta}\left(x\right)\right)|_{\theta=\theta^g} \\
- \;& \frac{1}{2}\sum_{l=1}^{p}\left(\theta_{n,l}-\theta_l^g\right)\nabla_{kl}\left(\sum_{x=0}^{\infty} K\left(\delta_n\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_{\theta}^A(x)} f_{\theta}^{2A}\left(x\right) + \left(A+B\right) f_{\theta}^{A+B}\left(x\right)\right) u_{\theta}\left(x\right)\right)|_{\theta=\theta^*}\Bigg\}.
\end{aligned} \tag{5.23}$$

Clearly, the first term within the braces in the right hand side converges to J_g in probability and the second term is an $o_p\left(1\right)$ term which we have already proved. Moreover, we can rewrite as,

$$\begin{aligned}
&\sqrt{n}\sum_{x=0}^{\infty} K\left(\delta_n^g\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}\left(x\right) + \left(A+B\right) f_{\theta^g}^{A+B}\left(x\right)\right) u_{\theta^g}\left(x\right) \\
= \;& \sqrt{n}\sum_{x=0}^{\infty}\left\{K\left(\delta_n^g\left(x\right)\right) - K\left(\delta_g^g\left(x\right)\right)\right\}\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}\left(x\right) + \left(A+B\right) f_{\theta^g}^{A+B}\left(x\right)\right) u_{\theta^g}\left(x\right) \\
= \;& S_{1n}|_{\theta=\theta^g},
\end{aligned} \tag{5.24}$$

which, by Lemma 5.4, goes to $N_p\left(0, V_g\right)$ in distribution as $n \to \infty$. Using this result, the representation in Equation (5.23) and applying Lemma 4.1 from Lehmann (1983), we can conclude that $\sqrt{n}\left(\theta_n - \theta^g\right)$ follows $N\left(0, J_g^{-1}V_gJ_g^{-1}\right)$ asymptotically. □

Corollary 5.6. *When $g = f_\theta$ for some $\theta \in \Theta$, then $\sqrt{n}\left(\theta_n - \theta\right) \sim N\left(0, J^{-1}VJ^{-1}\right)$ asymptotically, where,*

$$\begin{aligned}
J &= E_{f_\theta}\left\{u_\theta\left(X\right)u_\theta^T\left(X\right)\left(\left(A+B\right)f_\theta^\alpha\left(x\right) + A^2\beta^2 e^{\beta f_\theta^A(X)} f_\theta^{2A-1}\left(X\right)\right)\right\} \\
&= \sum_{x=0}^{\infty}\left\{u_\theta\left(x\right)u_\theta^T\left(x\right)\left(\left(A+B\right)f_\theta^\alpha\left(x\right) + A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A-1}\left(x\right)\right)\right\} f_\theta\left(x\right). \\
V &= V_{f_\theta}\left\{u_\theta\left(X\right)\left(\left(A+B\right)f_\theta^\alpha\left(X\right) + A^2\beta^2 e^{\beta f_\theta^A(X)} f_\theta^{2A-1}\left(X\right)\right)\right\} \\
&= \left(A+B\right)^2 \sum_{x=0}^{\infty} u_\theta\left(x\right)u_\theta^T\left(x\right)f_\theta^{1+2\alpha}\left(x\right) \\
&+ A^4\beta^4 \sum_{x=0}^{\infty} e^{2\beta f_\theta^A(x)} f_\theta^{4A-1}\left(x\right)u_\theta\left(x\right)u_\theta^T\left(x\right) \\
&+ 2\left(A+B\right)A^2\beta^2 \sum_{x=0}^{\infty} e^{\beta f_\theta^A(x)} f_\theta^{2A+\alpha}\left(x\right)u_\theta\left(x\right)u_\theta^T\left(x\right) - \zeta\zeta',
\end{aligned} \tag{5.25}$$

with, $\zeta = \sum_{x=0}^{\infty} u_\theta\left(x\right)\left(\left(A+B\right)f_\theta^{A+B}\left(x\right) + A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}\left(x\right)\right).$

5.4.4 Influence Function

Here we study the stability of our proposed class of estimators by exploiting the influence function (IF), which measures the effect of adding an infinitesimal mass to the distribution and is one of the most important heuristic tools of robustness. A simple differentiation of a contaminated version of the estimating equation (5.2) leads to the expression

$$IF\left(y, G, T_{\alpha,\lambda,\beta}\right) = J_g^{-1}N_g\left(y\right), \quad \text{where,} \tag{5.26}$$

$$\begin{aligned}
N_g\left(y\right) &= \left(A^2\beta^2 e^{\beta f_{\theta^g}^A(y)} f_{\theta^g}^A\left(y\right) + (A+B)f_{\theta^g}^B\left(y\right)\right) g^{A-1}\left(y\right)u_{\theta^g}\left(y\right) \\
&- \sum_{x=0}^{\infty}\left(A^2\beta^2 e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^A\left(x\right) + (A+B)f_{\theta^g}^B\left(x\right)\right) g^A\left(x\right)u_{\theta^g}\left(x\right),
\end{aligned}$$

$$\begin{aligned}
J_g &= A^2\beta^2 \sum_{x=0}^{\infty} e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^A(x) \left(2f_{\theta^g}^A(x) - g^A(x)\right) u_{\theta^g}(x) u_{\theta^g}^T(x) \\
&+ A\beta^3 \sum_{x=0}^{\infty} e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^{2A}(x) \left(f_{\theta^g}^A(x) - g^A(x)\right) u_{\theta^g}(x) u_{\theta^g}^T(x) \\
&+ \frac{(A+B)}{A} \sum_{x=0}^{\infty} f_{\theta^g}^B(x) \left((A+B) f_{\theta^g}^A(x) - B g^A(x)\right) u_{\theta^g}(x) u_{\theta^g}^T(x) \\
&+ A\beta^2 \sum_{x=0}^{\infty} e^{\beta f_{\theta^g}^A(x)} f_{\theta^g}^A(x) \left(g^A(x) - f_{\theta^g}^A(x)\right) i_{\theta^g}(x) \\
&- \frac{(A+B)}{A} \sum_{x=0}^{\infty} f_{\theta^g}^B \left(f_{\theta^g}^A(x) - g^A(x)\right) i_{\theta^g}(x),
\end{aligned}$$

where, g is the density of G. Note that, this J_g is similar with the matrix given in Equation (5.6). If the distribution G belongs to the model family $\mathcal{F}$ with $g = f_\theta$, then the influence function reduces to,

$$IF(y, F_\theta, T_{\alpha,\lambda,\beta}) = J^{-1} N(y), \text{where,} \tag{5.27}$$

$$\begin{aligned}
J &= \sum_{x=0}^{\infty} \left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x) + (A+B) f_\theta^{A+B}(x)\right) u_\theta(x) u_\theta^T(x), \\
N(y) &= A^2\beta^2 e^{\beta f_\theta^A(y)} f_\theta^{2A-1}(y) u_\theta(y) + (A+B) f_\theta^{A+B-1}(y) u_\theta(y) \\
&- \sum_{x=0}^{\infty} A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x) u_\theta(x) - \sum_{x=0}^{\infty} (A+B) f_\theta^{A+B}(x) u_\theta(x).
\end{aligned}$$

Again, this J is identical with the J given in Equation (5.25) of Corollary 5.6. Evidently, the influence function is dependent on all the three tuning parameters. Whenever the matrix J is non singular, the boundedness of the influence function depends on the ability of the coefficients to control the score function $u_\theta(y)$ in the first two terms of the numerator. In most parametric models including all exponential family models, $f_\theta^\tau(y) u_\theta(y)$ remains bounded for any $\tau > 0$; in the case $\tau = 0$, however the expression equals $u_\theta(y)$ and there is no control over it to keep it bounded. For the second term of the numerator in Equation (5.28), this is achieved when $A+B > 1$, i.e., when $\alpha > 0$. The first term of the numerator contains an additional exponential term. However, given that $f_\theta(y) \leq 1$ for any value y in the support of a discrete random variable, the first term of the numerator is easily seen to be bounded for any fixed non-zero real β when $2A - 1 > 0$, i.e., $A > 1/2$. We now list the different possible cases for boundedness of the influence function as follows:

1. $\beta = 0$; here the first and third terms of the numerator vanish, and the only other condition necessary is $A + B > 1$, i.e., $\alpha > 0$. This is essentially the S-divergence case, and shows that all minimum S-divergence functionals with $\alpha > 0$ have bounded influence (irrespective of the value of λ). In this case the allowable region for the triplet (α, λ, β) for bounded influence is $\mathbb{S}_1 = (\alpha > 0, \lambda \in \mathbb{R}, \beta = 0)$.

2. $\beta \neq 0$, $A = 0$. In this case also the first and third terms of the numerator drop out and the additional required condition is $\alpha > 0$. However, since $A = 1 + \lambda(1 - \alpha) = 0$, this implies $\lambda = -\frac{1}{1-\alpha}$. In this case the influence function is independent of β. Now the relevant region for the triplet is $\mathbb{S}_2 = \Big(\alpha > 0, \lambda = -\frac{1}{1-\alpha}, \beta \neq 0\Big)$.

3. Now suppose $A + B = 0$, without the components being individually zero. In this case the second and fourth terms get eliminated and we have $\alpha = -1$. In this case the condition $2A - 1 > 0$ translates to $\lambda > -\frac{1}{4}$. Here the corresponding region for the triplet is $\mathbb{S}_3 = \big(\alpha = -1, \lambda \geq -\frac{1}{4}, \beta \neq 0\big)$.

4. Now we allow all the terms β, A and $A + B$ to be non-zero. In this case all the four terms of the numerator are non-vanishing. Then, beyond the condition on β, the required conditions are

$\alpha > 0$ and $\lambda\,(1-\alpha) > -\frac{1}{2}$. The region here is
$\mathbb{S}_4 = \big(\alpha > 0, \lambda(1-\alpha) > -\frac{1}{2}, \beta \neq 0\big)$.

Combining all the cases, we see that the IF will be bounded if the triplet $(\alpha, \lambda, \beta) \in \mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3 \cup \mathbb{S}_4$.

It is easily seen that the four constituent subregions are disjoint. For illustration, we present some plots for bounded and unbounded influence functions for the minimum GSB functional under the Poisson(θ) model in Figure 5.1, where the true data distribution is Poisson(3). In the four rows of the right panel we give examples of triplets belonging to the four disjoint components of $\mathbb{S}$. In the first two rows of the right panel, the α value alone determines the shape of the curve. On the i-th row of the left panel, on the other hand, the triplets are slightly different from the triplets of i-th row on the right, but far enough to be pushed out of $\mathbb{S}_i$. Accordingly, all the plots on the left correspond to unbounded influence functions. Generally, it may also be observed that for increasing β the curves get flatter in each plot, where IF varies over different β. We will provide further illustration of the bounded influence region of the triplet through three-dimensional graphs at the end of the simulation section.

5.4.5 Simulation Result

In the simulation section our aim is to demonstrate that by choosing non-zero values of the parameter β, we may be able to generate procedures which, in a suitable sense, improve upon the estimators that are provided by the existing standard, the class of S-divergences. We consider the Poisson (θ) model for illustration, and choose samples of size 50 from the $(1-\epsilon)$Poisson$(3) + \epsilon$Poisson(10) mixture, where the second component is the contaminant and $\epsilon \in [0, 1)$ is the contaminating proportion. The values 0, 0.05, 0.1 and 0.2 are considered for ϵ, and at each contamination level, the samples are replicated 1000 times. The Poisson parameter is estimated in each of the 1000 replications, for each contamination level, and at each of several (α, λ, β) triplets considered in our study. Subsequently we construct the empirical mean square error (MSE) against the target value of 3, for each tuning parameter triplet and each contamination level over the 1000 replications.

In case of the minimum S-divergence estimator, Ghosh et al. (2017) have empirically identified a subset of (α, λ) collections which represent good choices. According to them, the zone of 'best' estimators correspond to an elliptical subset of the tuning parameter space, with $\alpha \in [0.1, 0.6]$ and $\lambda \in [-1, -0.3]$. We hope to show that for most of the (α, λ) combinations (including the best ones) there is a corresponding better or competitive (α, λ, β) combination with a non-zero β, thus providing an option which appears to perform better, at least to the extent of the findings in these simulations.

We begin with an exploration of the S-divergence, since this is the basis for comparison. The MSEs are presented in Table 5.3 over a cross-classified grid with α values in $\{0.1, 0.25, 0.4, 0.5, 0.6, 0.8, 1\}$ and λ values in $\{-1, -0.7, -0.5, -0.3, 0, 0.2, 0.5, 0.8, 1\}$, a total of 63 cells. In each cell the empirical MSEs for $\epsilon = 0, 0.05, 0.1$ and 0.2 are presented in a column of four elements, in that order, followed by the corresponding combination of tuning parameters $(\alpha, \lambda, \beta = 0)$. We have carried the $\beta = 0$ parameter in each triplet of parameters, to indicate that the S-divergence is indeed a special case of the GSB divergence. It may be noted that between all the cells, there is no unique (α, λ) combination which produces an overall best result (in terms of smallest MSE) over all the four columns (levels of contamination).

We now expand the exploration by considering, in addition, a grid of possible non-zero β values at each (α, λ) combination to see if the results can be improved. To be conservative about our definition of improvement, we declare the existence of a 'better' triplet in the GSB sense if all the four mean square errors corresponding to a (α, λ) combination within the S-divergence family in Table 5.2 are improved (reduced) by a suitable member of the GSB divergence class which is strictly outside the S-divergence family (corresponding to a non-zero β).

Our exploration indicates that in a large majority of the 63 cells there is a member of the GSB divergence with a non-zero β parameter which improves (over all the four cells) the performance of the corresponding S-divergence estimator with the same (α, λ) combination. Interestingly it turns out that in practically all the cases where an improvement is observed it happens for a negative value of β (it is observed to be zero in rare cases, but is never positive). A more detailed inspection indicates that in many of these cases, the improvement occurs at the value $\beta = -4$.

In order to summarize the findings of this rather large exploration (presented in Table 5.3) in a meaningful manner, we first note the following different cases,

1. (First Case) These are the cells where all the four mean square errors for the S-divergence case are reduced by the minimum GSB divergence estimator with the same values of (α, λ) and

$\beta = -4$. These cells are highlighted with the blue colour in Table 5.3. (There are 18 such cells).

2. (Second Case) These are the cells where all the four MSEs for the S-divergence case are reduced by a minimum GSB divergence estimator with $\beta = -4$ but with a different (α, λ) combination than that for the corresponding cell. These cells are highlighted in red in Table 5.3. (There are 39 such cells).

3. (Third Case) These are the cells where all the four MSEs are reduced by a minimum GSB divergence estimator outside the S-divergence family, but with $\beta \neq -4$, and not necessarily the same (α, λ). These cells are highlighted in orange in Table 5.3. (There is one such cell).

4. (Fourth Case) These are the cells where some triplet within the minimum GSB divergence class can improve upon the three MSEs under contamination ($\epsilon = 0.05, 0.1, 0.2$) but not all the four MSEs simultaneously. While these are not 'better' triplets in the sense described earlier in the section, the pure data MSEs (not reported here) for these triplets are close to those of the S-Divergence MSEs for these cells; in this sense these triplets are at least competitive. These cells are highlighted in green in Table 5.3. (There are three such cells).

5. (Fifth Case) These are the cells where no (α, λ, β) provides an improvement over the S-divergence results in the sense of any of the previous four cases (although there are competitive alternatives). These cells remain in black in Table 5.3. There are 2 such cells.

On the whole, therefore, it turns out that we observe improvements in 57 out of the 63 cells in all four rows of the column of MSEs in that cell by choosing $\beta = -4$ together with the S-divergence parameters. Even in the handful of cases (cells) where we do not have an improvement in all the rows of the column, there generally are competitive (although not strictly better) options within the minimum GSB divergence class with a negative value of β. In Table 5.3, in each cell, we also present the particular (α, λ, β) combination which generates the mean square errors (improved over Table 5.2 in most cases, as we have seen) reported in that cell.

In Figure 5.2, we provide a three-dimensional plot (as described in that section) in the three-dimensional (α, λ, β) plane, where the region $\mathbb{S}$ has been expressed as a union of several colour-coded subregions representing the individual components. The triplets corresponding to the improved MSE solutions reported in the cells of Table 5.3 all belong to the blue subregion of this figure, indicating that all improved solutions are provided by bounded influence estimators.

5.4.6 Selection of Optimal Tuning Parameters through Real Data Analysis

Our simulations in the previous section seem to suggest that the minimum divergence estimators within the GSB class with $\beta = -4$ often provides good options for data analysis. To take full advantage of this observation, this subclass of the GSB family should be explored further. However, we want to fully exploit the flexibility of the three parameter system, and noting that in some cases the optimal is outside the $\beta = -4$ subclass, including some which generate the most competitive solutions in the full system, we want to use an

Table 5.2: MSEs of the minimum divergence estimators within the S-divergence family for pure and contaminated data

0.1968	0.0836	0.0704	0.0708	0.0733	0.0802	0.0876
0.1974	0.0981	0.0855	0.0852	0.0869	0.0926	0.0994
0.1753	0.1063	0.1012	0.1028	0.1054	0.1116	0.1118
0.3099	0.2245	0.2119	0.2113	0.2130	0.2200	0.2298
(0.1, −1, 0)	(0.25, -1, 0)	(0.4, -1, 0)	(0.5, -1, 0)	(0.6, -1, 0)	(0.8, -1, 0)	(1, -1, 0)
0.0751	0.0666	0.0673	0.0698	0.0729	0.0800	0.0876
0.0893	0.0830	0.0831	0.0847	0.0869	0.0927	0.0994
0.1081	0.1044	0.1045	0.1056	0.1073	0.1121	0.1118
0.2830	0.2505	0.2328	0.2264	0.2231	0.2233	0.2298
(0.1, -0.7, 0)	(0.25, -0.7, 0)	(0.4, -0.7, 0)	(0.5, -0.7, 0)	(0.6, -0.7, 0)	(0.8, -0.7, 0)	(1, -0.7, 0)
0.0638	0.0635	0.0665	0.0694	0.0727	0.0799	0.0876
0.0836	0.0821	0.0832	0.0849	0.0871	0.0927	0.0994
0.1203	0.1120	0.1087	0.1083	0.1089	0.1125	0.1118
0.3715	0.2958	0.2559	0.2408	0.2319	0.2258	0.2298
(0.1, -0.5, 0)	(0.25, -0.5, 0)	(0.4, -0.5, 0)	(0.5, -0.5, 0)	(0.6, -0.5, 0)	(0.8, -0.5, 0)	(1, -0.5, 0)
0.0600	0.0622	0.0660	0.0691	0.0725	0.0798	0.0876
0.0895	0.0846	0.0843	0.0856	0.0875	0.0928	0.0994
0.1554	0.1264	0.1149	0.1118	0.1109	0.1129	0.1118
0.5669	0.3709	0.2904	0.2605	0.2424	0.2286	0.2298
(0.1, -0.3, 0)	(0.25, -0.3, 0)	(0.4, -0.3, 0)	(0.5, -0.3, 0)	(0.6, -0.3, 0)	(0.8, -0.3, 0)	(1, -0.3, 0)
0.0592	0.0617	0.0657	0.0688	0.0721	0.0796	0.0876
0.1415	0.0971	0.0880	0.0873	0.0883	0.0929	0.0994
0.3491	0.1774	0.1308	0.1196	0.1147	0.1136	0.1118
1.3860	0.6555	0.3832	0.3061	0.2655	0.2333	0.2298
(0.1, 0, 0)	(0.25, 0, 0)	(0.4, 0, 0)	(0.5, 0, 0)	(0.6, 0, 0)	(0.8, 0, 0)	(1, 0, 0)
0.0608	0.0621	0.0657	0.0687	0.0721	0.0794	0.0876
0.3302	0.1231	0.0930	0.0892	0.0890	0.0930	0.0994
0.8565	0.2745	0.1508	0.1276	0.1181	0.1141	0.1118
2.5938	1.0853	0.5550	0.3553	0.2867	0.2370	0.2298
(0.1, 0.2, 0)	(0.25, 0.2, 0)	(0.4, 0.2, 0)	(0.5, 0.2, 0)	(0.6, 0.2, 0)	(0.8, 0.2, 0)	(1, 0.2, 0)
0.0671	0.0638	0.0658	0.0685	0.0718	0.0792	0.0876
1.1434	0.3251	0.1115	0.0943	0.0907	0.0931	0.0994
2.3829	0.8165	0.2234	0.1489	0.1255	0.1151	0.1118
4.8817	2.4261	0.8641	0.4847	0.3338	0.2434	0.2298
(0.1, 0.5, 0)	(0.25, 0.5, 0)	(0.4, 0.5, 0)	(0.5, 0.5, 0)	(0.6, 0.5, 0)	(0.8, 0.5, 0)	(1, 0.5, 0)
0.0778	0.0676	0.0665	0.0685	0.0716	0.0790	0.0876
1.9928	0.9339	0.1951	0.1068	0.0936	0.0933	0.0994
3.7890	1.9909	0.4869	0.1994	0.1378	0.1162	0.1118
6.6731	4.2592	1.6784	0.7520	0.4130	0.2511	0.2298
(0.1, 0.8, 0)	(0.25, 0.8, 0)	(0.4, 0.8, 0)	(0.5, 0.8, 0)	(0.6, 0.8, 0)	(0.8, 0.8, 0)	(1, 0.8, 0)
0.0863	0.0717	0.0673	0.0686	0.0714	0.0789	0.0876
2.4449	1.3987	0.3803	0.1283	0.0967	0.0934	0.0994
4.5000	2.7992	0.9117	0.2793	0.1514	0.1171	0.1118
7.5320	5.3554	2.5215	1.0745	0.4969	0.2572	0.2298
(0.1, 1, 0)	(0.25, 1, 0)	(0.4, 1, 0)	(0.5, 1, 0)	(0.6, 1, 0)	(0.8, 1, 0)	(1, 1, 0)

overall data-based tuning parameter selection rule in which all the three parameters are allowed to vary over reasonable supports. The aim is to select the 'best' tuning parameter combination depending

Table 5.3: MSEs of the minimum GSB divergence estimators under pure and contaminated data

0.0623	0.0696	0.0704	0.0708	0.0687	0.0720	0.0763
0.0816	0.0843	0.0855	0.0852	0.0833	0.0859	0.0892
0.1115	0.1056	0.1012	0.1028	0.1042	0.1060	0.1077
0.2831	0.2207	0.2119	0.2113	0.2110	0.2162	0.2115
(0.4, −0.4, -4)	(0.8, -0.5, -4)	(0.4, -1, 0)	(0.5, -1, 0)	(0.8, 0, -7.5)	(0.8, -0.3, -4)	(1, -1, -4)
0.0642	0.0681	0.0681	0.0696	0.0696	0.0720	0.0763
0.0816	0.0826	0.0826	0.0843	0.0843	0.0859	0.0892
0.1076	0.1043	0.1043	0.1055	0.1055	0.1060	0.1077
0.2514	0.2135	0.2135	0.2207	0.2207	0.2162	0.2115
(0.6, -0.5, -4)	(0.8, 0, -8)	(0.8, 0, -8)	(0.8, -0.5, -4)	(0.8, -0.5, -4)	(0.8, -0.3, -4)	(1, -0.7, -4)
0.0623	0.0623	0.0659	0.0678	0.0678	0.0696	0.0763
0.0816	0.0816	0.0825	0.0834	0.0834	0.0843	0.0892
0.1115	0.1115	0.1071	0.1061	0.1061	0.1055	0.1077
0.2831	0.2831	0.2417	0.2295	0.2295	0.2207	0.2115
(0.4, -0.4, -4)	(0.4, -0.4, -4)	(0.5, -0.3, -4)	(0.6, -0.3, -4)	(0.6, -0.3, -4)	(0.8, -0.5, -4)	(1, -0.5, -4)
0.0600	0.0619	0.0642	0.0659	0.0678	0.0720	0.0763
0.0845	0.0822	0.0819	0.0825	0.0834	0.0859	0.0892
0.1294	0.1154	0.1091	0.1071	0.1061	0.1060	0.1077
0.4049	0.3080	0.2602	0.2417	0.2295	0.2162	0.2115
(0.1, -0.3, -4)	(0.25, -0.3, -4)	(0.4, -0.3, -4)	(0.5, -0.3, -4)	(0.6, -0.3, -4)	(0.8, -0.3, -4)	(1, -0.3, -4)
0.0600	0.0600	0.0644	0.0644	0.0644	0.0755	0.0763
0.0845	0.0845	0.0817	0.0817	0.0817	0.0887	0.0892
0.1294	0.1294	0.1073	0.1073	0.1073	0.1077	0.1077
0.4049	0.4049	0.2492	0.2492	0.2492	0.2139	0.2115
(0.1, -0.3, -4)	(0.1, -0.3, -4)	(0.8, -1, -4)	(0.8, -1, -4)	(0.8, -1, -4)	(0.8, 0, -4)	(1, 0, -4)
0.0600	0.0600	0.0644	0.0644	0.0644	0.0779	0.0763
0.0845	0.0845	0.0817	0.0817	0.0817	0.0907	0.0892
0.1294	0.1294	0.1073	0.1073	0.1073	0.1092	0.1077
0.4049	0.4049	0.2492	0.2492	0.2492	0.2146	0.2115
(0.1, -0.3, -4)	(0.1, -0.3, -4)	(0.8, -1, -4)	(0.8, -1, -4)	(0.8, -1, -4)	(0.8, 0.2, -4)	(1, 0.2, -4)
0.0600	0.0600	0.0644	0.0644	0.0644	0.0696	0.0763
0.0845	0.0845	0.0817	0.0817	0.0817	0.0843	0.0892
0.1294	0.1294	0.1073	0.1073	0.1073	0.1055	0.1077
0.4049	0.4049	0.2492	0.2492	0.2492	0.2207	0.2115
(0.1, -0.3, -4)	(0.1, -0.3, -4)	(0.8, -1, -4)	(0.8, -1, -4)	(0.8, -1, -4)	(0.8, -0.5, -4)	(1, 0.5, -4)
0.0600	0.0600	0.0644	0.0644	0.0644	0.0696	0.0763
0.0845	0.0845	0.0817	0.0817	0.0817	0.0843	0.0892
0.1294	0.1294	0.1073	0.1073	0.1073	0.1055	0.1077
0.4049	0.4049	0.2492	0.2492	0.2492	0.2207	0.2115
(0.1, -0.3, -4)	(0.1, -0.3, -4)	(0.8, -1, -4)	(0.8, -1, -4)	(0.8, -1, -4)	(0.8, -0.5, -4)	(1, 0.8, -4)
0.0600	0.0600	0.0644	0.0644	0.0644	0.0696	0.0763
0.0845	0.0845	0.0817	0.0817	0.0817	0.0843	0.0892
0.1294	0.1294	0.1073	0.1073	0.1073	0.1055	0.1077
0.4049	0.4049	0.2492	0.2492	0.2492	0.2207	0.2115
(0.1, -0.3, -4)	(0.1, -0.3, -4)	(0.8, -1, -4)	(0.8, -1, -4)	(0.8, -1, -4)	(0.8, -0.5, -4)	(1, 1, -4)

on data contamination. Thus datasets which show very close compatibility to the model should be analyzed by a triplet providing an efficient solution, while a more anomalous one should have a more robust member of the GSB class to deal with it.

In Chapter 3 we have described in detail the existing tuning parameter selection algorithms in the literature – such as Hong and Kim (2001) and Warwick and Jones (2005) – and also presented a refinement which we believe provides an improvement over the existing techniques. In the following we have taken up two real data examples and considered the problem of selecting the 'optimal' tuning parameters in each case, under the same nomenclature and notation as in Chapter 3. The OWJ algorithm considered here uses the minimum L_2 distance estimator as the pilot. Although the IWJ algorithm is pilot independent, for computational purposes it needs to commence from some suitable robust pilot for which also we utilize the minimum L_2 distance estimator. While the IWJ algorithm is our preferred method, we demonstrate the use of all the three algorithms in the following data sets. Along with unrestricted β, we have also given the estimates corresponding to pre-fixed $\beta = -4$.

Example 5.1. *(Drosophila Data): In Chapter 3, Table 3.4, we have presented two sets of data, which are based on a chemical mutagenicity experiment and may be robustly modeled using the Poisson(θ) distribution. These data were previously analyzed by Simpson (1987), and the whole experimental protocol is given in Woodruff et al. (1984). For the sake of completeness we briefly describe the experiment again. The experimenter exposed groups of male flies to different doses of the chemical and then mated each male with an unexposed female fly. Finally 100 daughter flies from each male were sampled to count the number of daughters carrying the sign of mutation. Corresponding to the variable denoting the number of recessive lethal daughters, we are interested to observe the frequencies denoting the number of exposed male flies. Two experimental run are considered for our analysis— one on Day 28 and other on Day 177. The data of Day 28 consist* *of two mild outliers with observed frequencies* $\boldsymbol{r} = (23, 3, 1, 1)$ *at* $\boldsymbol{x} = (0, 1, 3, 4)$*, where, the dataset of Day 177 consists of a single large outlier with observed frequencies* $\boldsymbol{r} = (23, 7, 3, 1)$ *at* $\boldsymbol{x} = (0, 1, 2, 91)$*.*

Poisson models are fitted to the datasets by estimating the Poisson parameter using the minimum GSB divergence estimation technique. Moreover, we have applied the three algorithms for finding the optimal tuning parameter triplets with the optimal estimates for both cases – one for unrestricted β *and another with restricted* $\beta = -4$*. The details are given in Table 5.4 and 5.5. It is clearly observed that the optimal solutions provide excellent robust fits.*

Table 5.4: Optimal estimates in different cases for the Drosophila Data (unrestricted β)

day	data	method	optimal $\hat{\theta}$	optimal (α, λ, β)
28	Full data	IWJ/OWJ/HK MLE	0.1181 0.357	$(0.37, -0.50, -8)$ $(0, 0, 0)$
	Clean data	IWJ/OWJ/HK MLE	0.1159 0.115	$(0.53, -1, -8)$ $(0, 0, 0)$
177	Full data	IWJ/OWJ/HK MLE	0.3591 3.05	$(0.51, -1, -8)$ $(0, 0, 0)$
	Clean data	IWJ/OWJ/HK MLE	0.3909 0.3939	$(0.41, -1, -8)$ $(0, 0, 0)$

Table 5.5: Optimal estimates in different cases for the Drosophila Data (β restricted to -4)

day	data	method	optimal $\hat{\theta}$	optimal (α, λ, β)
28	Full data	IWJ/OWJ/HK MLE	0.1209 0.357	$(0.37, -0.64, -4)$ $(0, 0, 0)$
	Clean data	IWJ/OWJ/HK MLE	0.1199 0.115	$(0.11, 1, -4)$ $(0, 0, 0)$
177	Full data	IWJ/OWJ/HK MLE	0.3539 3.05	$(0.49, -0.98, -4)$ $(0, 0, 0)$
	Clean data	IWJ/OWJ/HK MLE	0.3907 0.3939	$(0, 1, -4)$ $(0, 0, 0)$

Example 5.2. *(Peritonitis Data): This example involves the incidence of peritonitis in* 390 *kidney patients. This dataset was provided by Professor Peter W. M. John (personal communication) of the Department of Mathematics, University of Texas at Austin, USA. These data have been presented in Table 3.7 of this thesis. A thorough scrutiny leads us to the consideration of a geometric model and here we are interested to estimate the 'success' probability (probability of contracting peritonitis for a kidney patient). The values at 10 and 12 may be regarded as mild outliers. Here, the IWJ solution coincides with the HK solution where the estimate of success probability is 0.5110 corresponding to* $(\alpha, \lambda, \beta) = (0.41, -0.84, -3.5)$*. The OWJ solution gives a slightly different success probability of 0.5105 corresponding to* $(\alpha, \lambda, \beta) = (0.17, -0.60, -3)$*. In case of clean data these IWJ, OWJ and HK estimates will be 0.5044, 0.5061 and 0.5029 corresponding to* $(\alpha, \lambda, \beta) = (0.47, -1, -2)$*,* $(0.29, -1, -1)$ *and* $(0.55, -1, -3)$*, respectively, being slightly different from each other. On the contrary, the MLEs for the full dataset and the (two) outlier deleted dataset are* 0.4962 *and* 0.5092*, respectively.*

If, however, we prefix β *at* -4*, then, for full data the IWJ/HK solution will be 0.5115 corresponding to* $(\alpha, \lambda, \beta) = (0.41, -0.84, -4)$*. The OWJ solution gives a slightly different success probability of 0.5092 corresponding to* $(\alpha, \lambda, \beta) = (0.45, -0.84, -4)$*. In case of clean data, IWJ solution coincides with the OWJ solution, i.e.,* $\hat{\theta} =$ 0.5056 *corresponding to* $(\alpha, \lambda, \beta) = (0.01, -1, -4)$ *and the HK estimates will be 0.5017 with* $(\alpha, \lambda, \beta) = (0.59, -1, -4)$*. It may be noted that as the outliers are mild/moderate, the full and the clean data estimates do not show a very wide departure. In fact the same phenomenon is observed in the general case considered in the previous paragraph, where the range of the tuning parameters are unrestricted.*

Now we consider a more recent dataset for the implementation of our new proposal.

Example 5.3. *(Stolen Bases Data): In the 'Major League Baseball (MLB) Player Batting Stats' data for the 2019 MLB Regular Season, obtained from the ESPN.com website, one variable of interest is the number of Stolen Bases (SB) awarded to the top 40 Home Run (HR) scorers of the American League (AL). This dataset, containing three extreme and five moderate outliers, can be well-modelled by the Poisson distribution if not for the outliers. We are interested in estimating θ, the average number of Stolen Bases (SB) awarded to the MLB batters of the AL throughout the whole regular season. The 'optimal' estimates, derived from the implementation of the three algorithms under the Poisson model, are presented in Table 5.7. The fitted polygons corresponding to some of these optimal estimates are given in Figure 5.3. At first, for each fixed $\hat{\theta}$, we have evaluated $f_{\hat{\theta}}(x)$ with x ranging from 0 to the maximum number of awarded stolen bases in the data, and then joined them using line-segments to get a closed polygon of estimated probabilities for each estimation technique. It is clear that except for the full data MLE, all the other estimators primarily describe the main model conforming part of the data and sacrifice the outliers.*

In the simulation section, we have obtained most of the optimal MSEs corresponding to $\beta = -4$. Here, in the above-mentioned real datasets, we observe that whether the value of β be pre-fixed or not, the GSB estimators obtained in the former case is not significantly different than in the latter one. Although it will certainly take more research to determine why $\beta = -4$ works well in many cases, we have demonstrated, and propose the use of the optimal tuning parameter search over the unrestricted space for β, while acknowledging that the $\beta = -4$ case will work quite well in most cases.

5.5 Conclusion

Earlier we have provided an extension of the ordinary Bregman divergence and in this chapter, we have made use of the suggested approach in generating a particular super-family of divergences which seems to work very well in practice and provides new minimum divergence techniques that appear to improve the performance of the S-divergence based procedures in many cases. Since the results presented here are based on a single study, more research will be necessary to decide to what extent the observed advantages of the procedures considered here can be generalized, but clearly there appears to be enough evidence to suggest such explorations are warranted. An obvious follow up step is to suitably handle the case of continuous models, where the construction of the density and the divergence are more difficult and this will be explored in a detailed manner in the next chapter.

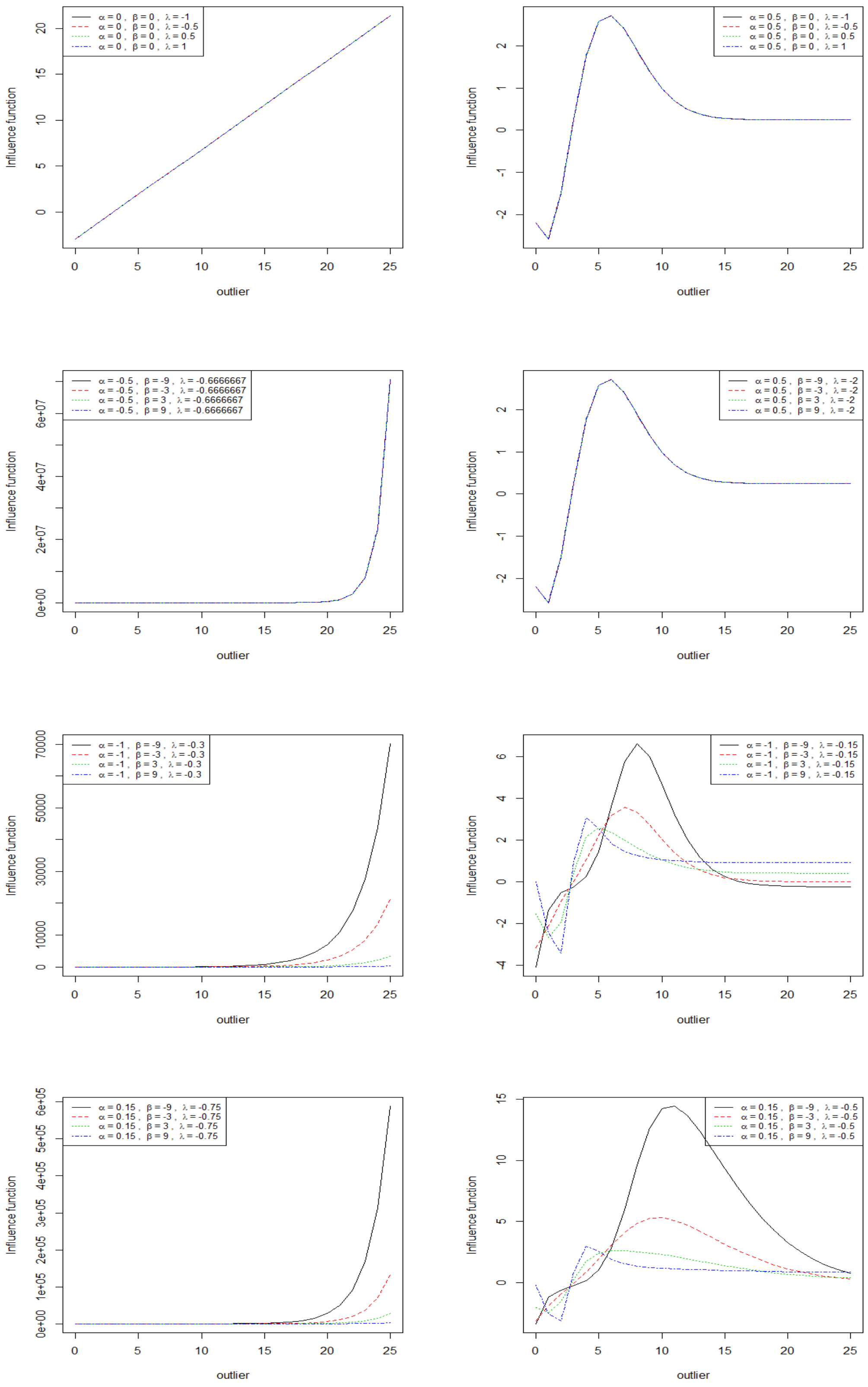


FIGURE 5.1: Examples of unbounded influence functions (left panel) and bounded influence functions (right panel) corresponding to $(\alpha, \lambda, \beta) \in$ each disjoint subsets contained in $\mathbb{S}$.

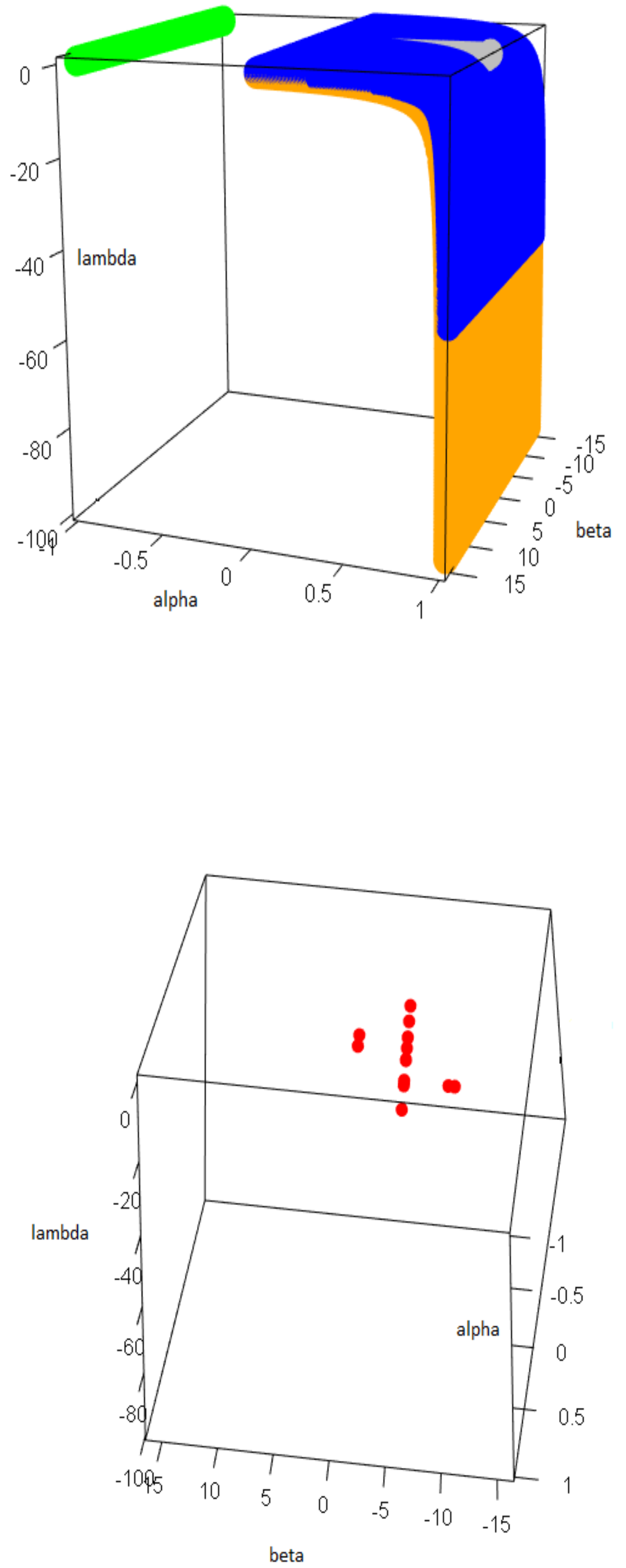


FIGURE 5.2: The first figure shows the region needed for bounded IF. Here, the grey, the orange, the green and the blue planes represent the boundaries of the sets $\mathbb{S}_1$, $\mathbb{S}_2$, $\mathbb{S}_3$ and $\mathbb{S}_4$, respectively. The 'best' solutions are given in red dots in the second figure.

Table 5.6: Top 40 Home Run (HR) scorers of the AL in the 2019 MLB Regular Season

Player	Team	HR	SB
Jorge Soler	KC	48	3
Mike Trout	LAA	45	11
Nelson Cruz	MIN	41	0
Alex Bregman	HOU	41	5
George Springer	HOU	39	6
Gleyber Torres	NYY	38	5
J. D. Martinez	BOS	36	2
Max Kepler	MIN	36	1
Matt Olson	OAK	36	0
Matt Chapman	OAK	36	1
Trey Mancini	BAL	35	1
Edwin Encarnacion	NYY/SEA	34	0
Carlos Santana	CLE	34	4
Gary Sanchez	NYY	34	0
Miguel Sano	MIN	34	0
Kole Calhoun	LAA	33	4
Xander Bogaerts	BOS	33	4
Marcus Semien	OAK	33	10
Jose Abreu	CHW	33	2
Austin Meadows	TB	33	12
Eddie Rosario	MIN	32	3
Francisco Lindor	CLE	32	22
Rafael Devers	BOS	32	8
Randal Grichuk	TOR	31	2
Jose Altuve	HOU	31	6
Renato Nunez	BAL	31	1
Mitch Garver	MIN	31	0
Eloy Jimenez	CHW	31	0
Yuli Gurriel	HOU	31	5
Rougned Odor	TEX	30	11
Daniel Vogelbach	SEA	30	0
Mookie Betts	BOS	29	16
Brett Gardner	NYY	28	10
Danny Santana	TEX	28	21
Aaron Judge	NYY	27	3
Yordan Alvarez	HOU	27	0
DJ LeMahieu	NYY	26	5
Mark Canha	OAK	26	3
Hunter Dozier	KC	26	2
Teoscar Hernandez	TOR	26	6

Table 5.7: Optimal estimates in different cases for the Stolen Bases Data

data	method	optimal $\hat{\theta}$	optimal (α, λ, β)
Full data (with outliers)	IWJ	2.6270	$(0.65, -0.98, -8)$
	OWJ	2.5086	$(0.73, -1, -8)$
	HK	2.6409	$(0.65, -1, -8)$
	MLE	4.875	$(0, 0, 0)$
excluding 8 outliers	IWJ	2.6426	$(0.53, -0.98, -8)$
	OWJ	2.5633	$(0.65, -0.98, -8)$
	HK	2.6426	$(0.53, -0.98, -8)$
	MLE	2.5625	$(0, 0, 0)$

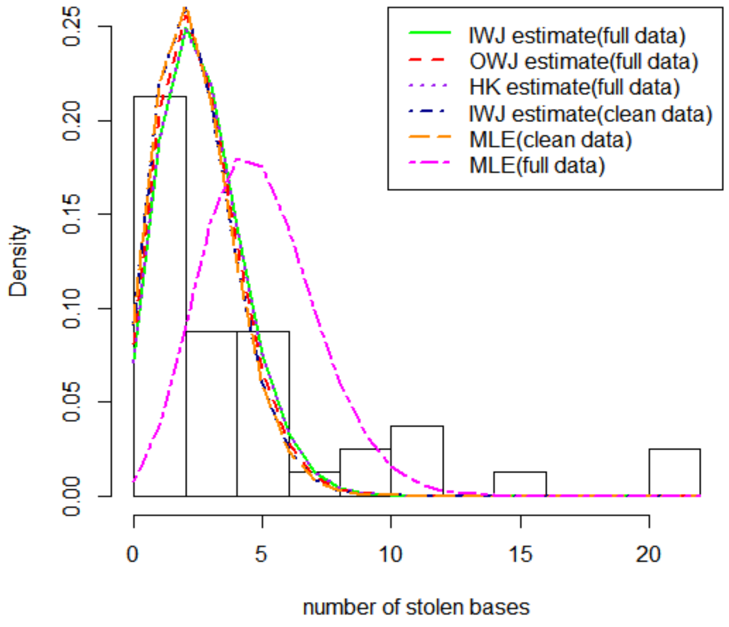


Figure 5.3: Some significant fits for the Stolen Bases Data under the Poisson model. Here "clean data" refer to the modified data after removing all 8 outliers.

Table 5.8: Optimal estimates in different cases for the Stolen Bases Data (with $\beta = -4$)

data	method	optimal $\hat{\theta}$	optimal (α, λ, β)
Full data (with outliers)	IWJ	2.7392	$(0.09, -0.52, -4)$
	OWJ	2.6083	$(0.25, -0.52, -4)$
	HK	2.7608	$(0.07, -0.52, -4)$
	MLE	4.875	$(0, 0, 0)$
excluding 8 outliers	IWJ/HK	2.5774	$(0.49, -0.98, -4)$
	OWJ	2.4749	$(0.25, 1, -4)$
	MLE	2.5625	$(0, 0, 0)$

Chapter 6

The Extended Bregman Divergence and Parametric Estimation in Continuous Models

6.1 Introduction

We have already proposed an extension of the ordinary Bregman divergence and with a special form, this proposal allows a new super divergence family – the GSB family. Its performance under the discrete model has already been explored in the previous chapter. In this chapter, we are going to do the same under the continuous setup.

6.2 The Generalized S-Bregman (GSB) Divergence

By using different ψ functions and different exponents k, one can generate different classes of previously unexplored divergences which might provide improved choices in parametric estimation. In this section, we will consider the divergence class mentioned in the last chapter, which is a generalized form containing both the S-divergence and the Bregman Exponential divergence (BED) proposed by Mukherjee et al. (2019). This divergence class has the form

$$D^*(g,f) = \int \left\{ e^{\beta f^A}\left(\beta f^A - \beta g^A - 1\right) + e^{\beta g^A} + \frac{1}{B}\left(g^{A+B} - f^{A+B}\right) - \left(g^A - f^A\right)\frac{A+B}{AB} f^B \right\} dx, \tag{6.1}$$

where $A+B = 1+\alpha$, $A = 1+\lambda(1-\alpha)$, $B = \alpha - \lambda(1-\alpha)$, $\alpha \geq -1$, $\beta, \lambda \in \mathbb{R}$. The above divergence corresponds to $\psi(x) = e^{\beta x} + \frac{x^{1+\frac{B}{A}}}{B}$ with $k = A$. The divergence in Equation (6.1) is referred to as the GSB (Generalized S-Bregman) divergence.

6.3 The Estimation Scheme under Continuous Models

We assume that both the true data generating distribution G as well as the model family $\mathcal{F} = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^p, p \geq 1\}$ belong to the class of all probability distributions having densities with respect to the Lebesgue measure.

Suppose $X_1, X_2, \ldots, X_n$ be independently and identically distributed observations from an unknown distribution G having density g with respect to the Lebesgue measure. To find the minimum divergence estimator of θ, we wish to minimize a suitable distance/divergence between the data density and the model density f_θ. Since the data density g is unknown, we need an empirical, non-parametric estimate of the data density to construct the divergence. In case of discrete models, one can simply consider the relative frequencies r_n of the support points in the random sample to generate the estimate $\hat{g}$ of g

but this is not applicable here. Whenever the model is continuous, the data generated by a sample are still discrete, so that there is an obvious mismatch of measures in using relative frequencies for the estimate of the data density. This necessitates the construction of a continuous, non-parametric density estimate from the data through a suitable smoothing method like kernel density estimation. For estimating the unknown parameter θ through a minimum divergence procedure in this setup, the two approaches that we have mentioned in the first chapter are available to us. Between the two approaches, we have discussed the advantage of using the Basu-Lindsay approach over Beran's approach. Here we are going to use the former approach from now on.

6.4 Estimating Equation

A routine differentiation of the Expression (6.1), with g replaced by g_n^* and f replaced by the smoothed model density f_θ^* gives the reduced estimating equation

$$\int K(\delta_n^*\left(x\right))\left(A^2\beta^2 e^{\beta {f_\theta^*}^A(x)} {f_\theta^*}^{2A}\left(x\right) + \left(A+B\right) {f_\theta^*}^{A+B}\left(x\right)\right)\tilde{u}_\theta\left(x\right)dx = 0 \quad (6.2)$$

with $\delta_n^*\left(x\right) = \frac{(g_n^*(x))}{(f_\theta^*(x))} - 1$. Here g_n^* is some suitable non-parametric estimate of g under this setup which has already been introduced in Chapter 1; $\tilde{u}_\theta$ is as defined in Section 1.5.1.3.2. Also $K(\delta^*) = \frac{(\delta^*+1)^A - 1}{A}$, here. The estimator obtained through solving Equation (6.2) will be denoted as the minimum GSB* divergence estimator (in the same spirit as the minimum S^*-divergence estimator, introduced by Ghosh and Basu (2017)). It is generally different from the minimum divergence estimator that would be derived under the Beran approach. The asymptotic distributions are also normally different;

however note the exceptions under transparent kernels described in Section 6.7.

6.5 Influence Function under the Basu-Lindsay Approach

For calculating the influence function, we consider the kernel smoothed version of the true density g given by

$$g^*(x) \;=\; \int W(x,y,h)\,dG(y) = \int W(x,y,h)\,g(y)\,dy. \quad (6.3)$$

The minimum GSB* divergence functional, denoted by $T^*_{\alpha,\lambda,\beta}(G)$, will be defined by the relation

$$D^*\left(g^*, f^*_{T^*_{\alpha,\lambda,\beta}(G)}\right) = \min\{D^*(g^*, f^*_\theta) : \theta \in \Theta\}, \quad (6.4)$$

where D^* is as defined in Equation (6.1), provided the minimum exists. We will call to it the best fitting parameter. Consider the contaminated distribution $G_\epsilon(x) = (1-\epsilon)G(x) + \epsilon\Lambda_y(x)$, with Λ_y being the distribution degenerate at y. Let g_ϵ denote the corresponding contaminated density with g^*_ϵ being its kernel-smoothed version. Evidently, $g^*_\epsilon = (1-\epsilon)g^* + \epsilon W(x,y,h)$. To derive the IF, we take the derivative of both sides of the equation

$$\int K(\delta^*_\epsilon(x))\left(A^2\beta^2 e^{\beta {f^*_\theta}^{A}(x)} {f^*_\theta}^{2A}(x) + (A+B)\,{f^*_\theta}^{A+B}(x)\right)\tilde{u}_\theta(x)\,dx = 0, \quad (6.5)$$

where $\delta^*_\epsilon(x) = \frac{g^*_\epsilon(x)}{f^*_{\theta_\epsilon}(x)} - 1$ and we get,

$$IF(y; G, T^*_{\alpha,\lambda,\beta}) = [J^*_g]^{-1} N^*_g(y), \text{ where}$$

$$\begin{aligned}
N_g^*(y) &= \int \left[A^2\beta^2 e^{\beta {f^*_{\theta^g}}^A(x)} {f^*_{\theta^g}}^A(x)\, \tilde{u}_{\theta^g}(x) {g^*}^{A-1}(x)(W(x,y,h)-g^*(x))\right] dx \\
&+ \int \left[(A+B){f^*_{\theta^g}}^B(x) {g^*}^{A-1}(x)\, \tilde{u}_{\theta^g}(x)(W(x,y,h)-g^*(x))\right] dx \\
&= \int \left(A^2\beta^2 e^{\beta {f^*_{\theta^g}}^A(x)} {f^*_{\theta^g}}^A(x) + (A+B){f^*_{\theta^g}}^B(x)\right) W(x,y,h) {g^*}^{A-1}(x)\, \tilde{u}_{\theta^g}(x)dx \\
&- \int \left(A^2\beta^2 e^{\beta {f^*_{\theta^g}}^A(x)} {f^*_{\theta^g}}^A(x) + (A+B){f^*_{\theta^g}}^B(x)\right) {g^*}^A(x)\, \tilde{u}_{\theta^g}(x)dx. \\
J_g^* &= \int \left(A^3\beta^2 e^{\beta {f^*_{\theta^g}}^A(x)} {f^*_{\theta^g}}^{2A}(x)(2+\beta {f^*_{\theta^g}}^A(x)) + (A+B)^2 {f^*_{\theta^g}}^{A+B}(x)\right) \tilde{u}_{\theta^g}(x)\tilde{u}_{\theta^g}^T(x)dx \\
&-\int \left(A^2\beta^2 e^{\beta {f^*_{\theta^g}}^A(x)} {f^*_{\theta^g}}^{2A}(x) + (A+B){f^*_{\theta^g}}^{A+B}(x)\right) \tilde{i}_{\theta^g}(x)dx \\
&-\int \left(A^3\beta^2 e^{\beta {f^*_{\theta^g}}^A(x)} {g^*}^A(x){f^*_{\theta^g}}^A(x)(1+\beta {f^*_{\theta^g}}^A(x)) + (A+B)B{f^*_{\theta^g}}^B(x){g^*}^A(x)\right) \tilde{u}_{\theta^g}(x)\, \tilde{u}_{\theta^g}^T(x)\, dx \\
&+\int \left(A^2\beta^2 e^{\beta {f^*_{\theta^g}}^A(x)} {f^*_{\theta^g}}^A(x){g^*}^A(x) + (A+B){f^*_{\theta^g}}^B(x){g^*}^A(x)\right) \tilde{i}_{\theta^g}(x)dx, \qquad (6.6)
\end{aligned}$$

with θ^g being the best fitting parameter under G. Moreover, when $g = f_\theta$, the influence function becomes

$$\begin{aligned}
J^* &= \int \left((A+B){f^*_\theta}^{A+B}(x)\tilde{u}_\theta(x)\, \tilde{u}_\theta^T(x) + A^2\beta^2 e^{\beta {f^*_\theta}^A(x)} {f^*_\theta}^{2A}(x)\tilde{u}_\theta(x)\, \tilde{u}_\theta^T(x)\right) dx, \\
N^* &= \int \left(A^2\beta^2 e^{\beta {f^*_\theta}^A(x)} {f^*_\theta}^{2A-1}(x) + (A+B){f^*_\theta}^{A+B-1}(x)\right) \tilde{u}_\theta(x) W(x,y,h)dx \\
&- \int \left(A^2\beta^2 e^{\beta {f^*_\theta}^A(x)} {f^*_\theta}^{2A}(x) + (A+B){f^*_\theta}^{A+B}(x)\right) \tilde{u}_\theta(x)dx. \qquad (6.7)
\end{aligned}$$

For this influence function to be bounded, we need to control the two terms (Term I and Term II) in the first integral of N^*. For most parametric model densities f_θ with score function u_θ, the integral $\int f_\theta^\tau(x) u_\theta(x)dx$ is bounded for $\tau > 0$, as the model density f_θ downweights the score functions of unlikely values. With a bounded kernel, therefore, Term II will be controlled whenever $A + B > 1$. If the density $f_\theta^*(x)$ is bounded (alternatively, if $f_\theta(x)$ and the kernel are bounded), Term I is also bounded for $2A > 1$. On the other hand, for $\beta \leq 0$, we have $e^{\beta {f^*_\theta}^A(x)} \leq 1$ hence boundedness of the kernel is sufficient for controlling Term I along with the condition $2A > 1$. In particular, under the normal model, the influence function is bounded for $A + B > 1$ and $2A > 1$.

For unbounded model densities, the influence functions may be bounded for specific models and specific tuning parameter combinations. But the generalized arguments in such cases may not be so obvious.

For bounded densities, with bounded kernel functions, therefore, the subregion of $\mathbb{R}^3$ in terms of tuning parameter combinations which lead to bounded influence may be analysed by the conditions described above. These calculations describe the collection $\mathbb{S}$ of triplets (α, λ, β) for which the influence function is bounded, where

$$\mathbb{S} \;=\; \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3 \cup \mathbb{S}_4, \tag{6.8}$$

where,

$$\mathbb{S}_1 = \left\{\alpha > 0, \lambda \in \mathbb{R}, \beta = 0\right\} \quad , \quad \mathbb{S}_2 = \left\{\alpha > 0, \lambda = -\frac{1}{1-\alpha}, \beta \neq 0\right\},$$
$$\mathbb{S}_3 = \left\{\alpha = -1, \lambda \geq -\frac{1}{4}, \beta \neq 0\right\} \quad , \quad \mathbb{S}_4 = \left\{\alpha > 0, \lambda(1-\alpha) > -\frac{1}{2}, \beta \neq 0\right\}.$$

Some plots of the bounded and the unbounded IFs are given in Figure 6.1.

Lemma 6.1. *In case of $g = f_\theta$ for some $\theta \in \Theta$, matrix J^* can be expressed further as*

$$J^* \;=\; J^*_{\alpha,\lambda,\beta}(F_\theta) = E_\theta\left[-\nabla u_\theta^{\alpha,\lambda,\beta^*}(X))\right], \text{ where} \tag{6.9}$$

$$u_\theta^{\alpha,\lambda,\beta^*}(y) \;=\; \int \left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A-1}(x) + (A+B)\, f_\theta^{*A+B-1}(x)\right) \tilde{u}_\theta(x)\, W(x,y,h)\, dx. \tag{6.10}$$

Proof. To prove this lemma, we note that from the expression of $u_\theta^{\alpha,\lambda,\beta^*}(y)$ given in Equation (6.10), we can write

$$\begin{aligned}
\nabla u_\theta^{\alpha,\lambda,\beta^*}(X) &= \int \nabla \tilde{u}_\theta(x)\left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A-1}(x) + (A+B) f_\theta^{*A+B-1}(x)\right) W(x,X,h)dx \\
&+ \int A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A-1}(x)\left(\beta A f_\theta^{*A}(x) + 2A - 1\right) W(x,X,h)\tilde{u}_\theta(x)\tilde{u}_\theta^T(x)dx \\
&+ \int (A+B)(A+B-1) f_\theta^{*A+B-1}(x) W(x,X,h)\tilde{u}_\theta(x)\tilde{u}_\theta^T(x)dx. \qquad (6.11)
\end{aligned}$$

Taking expectation of the first term with respect to X, we get,

$$\begin{aligned}
&E_\theta\left[\int \nabla \tilde{u}_\theta(x)\left(A^2\beta^2 e^{\beta f_\theta^{*A}}(x) f_\theta^{*2A-1}(x) + (A+B) f_\theta^{*A+B-1}(x)\right) W(x,X,h)dx\right] \\
&= \int\left[\int \nabla \tilde{u}_\theta(x)\left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A-1}(x) + (A+B) f_\theta^{*A+B-1}(x)\right) W(x,y,h)dx\right] f_\theta(y)dy \\
&= \int \nabla \tilde{u}_\theta(x)\left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A-1}(x) + (A+B) f_\theta^{*A+B-1}(x)\right)\left[\int W(x,y,h) f_\theta(y)dy\right] dx \\
&= \int \nabla \tilde{u}_\theta(x)\left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A-1}(x) + (A+B) f_\theta^{*A+B-1}(x)\right) f_\theta^*(x)dx \\
&= \int \nabla \tilde{u}_\theta(x)\left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A} + (A+B) f_\theta^{*A+B}(x)\right) dx. \qquad (6.12)
\end{aligned}$$

Manipulating the other terms similarly, the other terms on the right hand side of Equation (6.11) becomes

$$\begin{aligned}
&E_\theta\left[\int A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A-1}(x)\left(\beta A f_\theta^{*A}(x) + 2A - 1\right) W(x,X,h)\tilde{u}_\theta(x)\tilde{u}_\theta^T(x)dx\right] \\
+\ &E_\theta\left[\int (A+B)(A+B-1) f_\theta^{*A+B-1}(x) W(x,X,h)\tilde{u}_\theta(x)\tilde{u}_\theta^T(x)dx\right] \\
=\ &\int\left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A}(x)\left(\beta A f_\theta^{*A}(x) + 2A - 1\right) + (A+B)(A+B-1) f_\theta^{*A+B}(x)\right)\tilde{u}_\theta(x)\tilde{u}_\theta^T(x)dx. \\
&\qquad (6.13)
\end{aligned}$$

Combining Equation (6.12) and Equation (6.13), we can say

$$\begin{aligned}
&E_\theta[\nabla u_\theta^{\alpha,\lambda,\beta^*}(X)] \\
=\ & \int \nabla \tilde{u}_\theta(x) \left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A}(x) + (A+B) f_\theta^{*A+B}(x) \right) dx \\
+\ & \int \left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A}(x) \left(\beta A f_\theta^{*A}(x) + 2A - 1 \right) + (A+B)(A+B-1) f_\theta^{*A+B}(x) \right) \tilde{u}_\theta(x) \tilde{u}_\theta^T(x) dx \\
=\ & \int \left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A}(x) + (A+B) f_\theta^{*A+B}(x) \right) \tilde{u}_\theta(x) dx \\
-\ & \int \left(A^3\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A}(x) \left(\beta f_\theta^{*A}(x) + 2 \right) + (A+B)^2 f_\theta^{*A+B}(x) \right) \tilde{u}_\theta(x) \tilde{u}_\theta^T(x) dx \\
+\ & \int \left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A}(x) \left(\beta A f_\theta^{*A}(x) + 2A - 1 \right) + (A+B)(A+B-1) f_\theta^{*A+B}(x) \right) \tilde{u}_\theta(x) \tilde{u}_\theta^T(x) dx \\
& \hspace{30em} (6.14) \\
=\ & -\int \left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A}(x) \left(\beta A f_\theta^{*A}(x) + 2A \right) + (A+B)^2 f_\theta^{*A+B}(x) \right) \tilde{u}_\theta(x) \tilde{u}_\theta^T(x) dx \\
+\ & \int \left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A}(x) \left(\beta A f_\theta^{*A}(x) + 2A - 1 \right) + (A+B)(A+B-1) f_\theta^{*A+B}(x) \right) \tilde{u}_\theta(x) \tilde{u}_\theta^T(x) dx \\
=\ & -\int \left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A}(x) + (A+B) f_\theta^{*A+B}(x) \right) \tilde{u}_\theta(x) \tilde{u}_\theta^T(x) dx = -J^*.
\end{aligned}$$

In Equation (6.14), the first two terms are derived using integration by parts of Equation (6.12). Evidently, this proves the lemma. □

Corollary 6.2. *Under the model, with the help of the above lemma, the influence function of minimum GSB* divergence estimator derived earlier in this section, simplifies to*

$$IF(y, F_\theta, T^*_{\alpha,\lambda,\beta}) = [J^*]^{-1} \left\{ u_\theta^{\alpha,\lambda,\beta^*}(y) - E_\theta(u_\theta^{\alpha,\lambda,\beta^*}(X)) \right\}, \quad (6.15)$$

where,

$$J^* \ = \ E_\theta \left[-\nabla u_\theta^{\alpha,\lambda,\beta^*}(X)) \right]. \quad (6.16)$$

6.6 Asymptotic Distribution of the Minimum GSB* Divergence Estimator

Suppose that $X_1, X_2, \ldots X_n$ are n i.i.d. observations from the true density g with g^* being the smoothed version of g, given in Equation (6.3). To find the closest match between g and the elements of the model family $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, the divergence between g_n^* and f_θ^* is minimized over $\theta \in \Theta$. First we list the assumptions required to prove the asymptotic results

1. $\mathcal{F}$ is identifiable, i.e., for any θ_1 and θ_2, $\theta_1 = \theta_2 \Rightarrow f_{\theta_1}(x) = f_{\theta_2}(x)$ for almost all x.

2. The densities within the model family have a common a support χ, which is independent of the parameter θ.

3. The kernel-integrated model family of densities is smooth, i.e., each $f_\theta^*(x)$ satisfies the conditions of Lehmann (1983, p. 409, p. 429).

4. The matrix J_g^* as defined in Equation (6.6) is positive definite.

5. The quantities
$$\int (g^*)^{\frac{1}{2}}(x)\left(A^2\beta^2 e^{\beta {f_\theta^*}^A(x)} {f_\theta^*}^{2A-1}(x) + (A+B)\,{f_\theta^*}^{A+B-1}(x)\right)|\tilde{u}_{j\theta}(x)|dx,$$
$$\int (g^*)^{\frac{1}{2}}(x)\left(A^2\beta^2 e^{\beta {f_\theta^*}^A(x)} {f_\theta^*}^{2A-1}(x) + (A+B)\,{f_\theta^*}^{A+B-1}(x)\right)|\tilde{u}_{j\theta}(x)||\tilde{u}_{k\theta}(x)|dx \text{ and}$$
$$\int (g^*)^{\frac{1}{2}}(x)\left(A^2\beta^2 e^{\beta {f_\theta^*}^A(x)} {f_\theta^*}^{2A-1}(x) + (A+B)\,{f_\theta^*}^{A+B-1}(x)\right)|\tilde{u}_{jk\theta}(x)|dx$$
are bounded for all j, k and for all $\theta \in \omega$, an open neighbourhood of the best fitting parameter θ^g.

6. There exists functions $M_{jkl}(x)$, $M_{jk,l}(x)$ and $M_{j,k,l}(x)$, $M_{j,k,l}^{(1)}(x)$ and $M_{j,k,l}^{(2)}(x)$ such that
$$\left(A^2\beta^2 e^{\beta {f_\theta^*}^A(x)} {f_\theta^*}^{2A-1}(x) + (A+B)\,{f_\theta^*}^{A+B-1}(x)\right)\tilde{u}_{jkl\theta}(x),$$

$\left(A^2\beta^2 e^{\beta f_\theta^{*^A}(x)} f_\theta^{*^{2A-1}}(x) + (A+B)\, f_\theta^{*^{A+B-1}}(x)\right) \tilde{u}_{jk\theta}(x)\, \tilde{u}_{l\theta}(x)$ and
$\left(A^2\beta^2 e^{\beta f_\theta^{*^A}(x)} f_\theta^{*^{2A-1}}(x) + (A+B)\, f_\theta^{*^{A+B-1}}(x)\right) \tilde{u}_{j\theta}(x)\, \tilde{u}_{k\theta}(x)\, \tilde{u}_{l\theta}(x)$,
$\left\{(A+B)^2 f_\theta^{*A+B-1}(x) + A^3\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A-1}(x)\left(2+\beta f_\theta^{*A}(x)\right)\right\} \tilde{u}_{j\theta}(x)\, \tilde{u}_{k\theta}(x)\, \tilde{u}_{l\theta}(x)$,
$(A+B)^3 f_\theta^{*A+B-1}(x)\, \tilde{u}_{j\theta}(x)\, \tilde{u}_{k\theta}(x)\, \tilde{u}_{l\theta}(x)$
$+ \left\{A^4\beta^2 e^{\beta f_\theta^{*A}(x)}\left(2 f_\theta^{*2A-1}(x) + \beta\left(f_\theta^{*3A-1}(x) + f_\theta^{*A-1}(x) + 4 f_\theta^{*2A-1}(x)\right)\right)\right\} \tilde{u}_{j\theta}(x)\, \tilde{u}_{k\theta}(x)\, \tilde{u}_{l\theta}(x)$
are dominated by these functions and the expectations are uniformly bounded with respect to g^* and f_θ^* for all x and all $\theta \in \omega$.

7. Suppose, C_1 and C_2 represent the bounds of $K'(\delta^*)$ and $K''(\delta^*)(1+\delta^*)$, respectively, where $K'(\cdot)$ and $K''(\cdot)$ represent the first and second order derivatives of $K(\cdot)$ with respect to its argument δ^*, where $\delta^*(x) = \frac{g^*(x)}{f_\theta^*(x)} - 1$.

To derive the asymptotic result of MGSBDE, we will assume, from now on, the above-mentioned conditions hold. Under these assumptions, we are going to state and prove, some set of lemmas necessary to establish our asymptotic results.

Lemma 6.3 (Basu and Lindsay (1994, Lemma 6.1))**.** *For each fixed x in the support, $V_g(g_n^*(x)) = \frac{\nu(x)}{n}$, provided it exists, where*

$$\nu(x) = \int W^2(x,y,h)\, g(y)\, dy - (g^*(x))^2 . \tag{6.17}$$

We further assume the kernel to be bounded, i.e.,

$$W(x,y,h) \le N(h) < \infty. \tag{6.18}$$

Therefore, further calculations lead us to the following

$$\begin{aligned} \nu(x) &\leq \int W^2(x,y,h)\, g(y)\, dy \\ &\leq N(h)\int W(x,y,h)\, g(y)\, dy \\ &\leq N(h) g^*(x). \end{aligned} \tag{6.19}$$

Lemma 6.4. *In the above-mentioned setup, with probability* 1*,*

$$n^{\frac{1}{4}}\{(g_n^*(x))^{\frac{1}{2}} - (g^*(x))^{\frac{1}{2}}\} \to 0. \tag{6.20}$$

provided $\nu(x) < \infty$.

Proof. With the finiteness assumption of $\nu(x)$, the proof mimics that of Lemma 5.2, except the fact that $r_n(x)$ has been used as estimate of g under discrete model, whereas, $g_n^*(x)$ is used as estimate of g under continuous model. □

Next, under continuous setup, we define *Hellinger residuals* as

$$\triangle_n^*(x) = \frac{g_n^{*1/2}(x)}{f_\theta^{*1/2}(x)} - 1; \triangle_g^*(x) = \frac{g^{*1/2}(x)}{f_\theta^{*1/2}(x)} - 1. \tag{6.21}$$

Furthermore, we define,

$$\delta_n^*(x) = \frac{g_n^*(x)}{f_\theta^*(x)} - 1; \delta_g^*(x) = \frac{g^*(x)}{f_\theta^*(x)} - 1. \tag{6.22}$$

Lemma 6.5. *Define* $\eta_n^*(x) = \sqrt{n}\left(\triangle_n^*(x) - \triangle_g^*(x)\right)^2$. *For any* $k \in [0,2]$ *and any* $x \in \chi$*, we have,*

1. $E_g\{\eta_n^{*k}(x)\} \leq n^{\frac{k}{2}} E_g\{|\delta_n^*(x) - \delta_g^*(x)|^k\} \leq \left\{\frac{\nu(x)}{(f_\theta^*(x))^2}\right\}^{\frac{k}{2}}$.

2. $E_g\{|\delta_n^*(x) - \delta_g^*(x)|\} \leq \left(\frac{\sqrt{\nu(x)}}{f_\theta^*(x)}\right)$.

Proof. The proof of this lemma is follows from Lemma 5.1. □

Lemma 6.6. *$E_g\{\eta_n^{*k}(x)\} \to 0$, as $n \to \infty$, for $k \in [0,2)$ and any $x \in \chi$.*

Proof. The proof is exactly similar to the proof of Lemma 5.2. □

Let us now define

$$\begin{aligned} a_n^*(x) &= K(\delta_n^*(x)) - K\left(\delta_g^*(x)\right) \\ b_n^*(x) &= \left(\delta_n^*(x) - \delta_g^*(x)\right) K'\left(\delta_g^*(x)\right) \\ \text{and} \quad \tau_n^*(x) &= \sqrt{n}|a_n^*(x) - b_n^*(x)|. \end{aligned} \tag{6.23}$$

Now, we will find the limiting distributions of the following

$$\begin{aligned} S_{1n}^*(x) &= \sqrt{n}\int_x a_n^*(x)\left(A^2\beta^2 e^{\beta {f_\theta^*}^A(x)} {f_\theta^*}^{2A}(x) + (A+B){f_\theta^*}^{A+B}(x)\right)\tilde{u}_\theta(x)\,dx, \\ S_{2n}^*(x) &= \sqrt{n}\int_x b_n^*(x)\left(A^2\beta^2 e^{\beta {f_\theta^*}^A(x)} {f_\theta^*}^{2A}(x) + (A+B){f_\theta^*}^{A+B}(x)\right)\tilde{u}_\theta(x)\,dx. \end{aligned} \tag{6.24}$$

Lemma 6.7. *Under assumption (5), $E_g|S_{1n}^* - S_{2n}^*| \to 0$ as $n \to \infty$ and as $n \to \infty$*

$$S_{1n}^* - S_{2n}^* \xrightarrow{p} 0.$$

Proof. Using Lemma 5.3, Lemma 6.5 and Equation (6.19), we have

$$\begin{aligned} E|S_{1n}^* - S_{2n}^*| &\leq \int E(\tau_n^*(x))\left(A^2\beta^2 e^{\beta {f_\theta^*}^A(x)} {f_\theta^*}^{2A}(x) + (A+B){f_\theta^*}^{A+B}(x)\right)|\tilde{u}_\theta(x)|dx \\ &\leq \gamma\int E(\eta_n^*(x))\left(A^2\beta^2 e^{\beta {f_\theta^*}^A(x)} {f_\theta^*}^{2A}(x) + (A+B){f_\theta^*}^{A+B}(x)\right)|\tilde{u}_\theta(x)|dx \\ &\leq \gamma\int \frac{\nu^{1/2}(x)}{f_\theta^*(x)}\left(A^2\beta^2 e^{\beta {f_\theta^*}^A(x)} {f_\theta^*}^{2A}(x) + (A+B){f_\theta^*}^{A+B}(x)\right)|\tilde{u}_\theta(x)|dx \\ &\leq \gamma N^{\frac{1}{2}}(h)\int_x (g^*(x))^{\frac{1}{2}}\left(A^2\beta^2 e^{\beta {f_\theta^*}^A(x)} {f_\theta^*}^{2A-1}(x) + (A+B){f_\theta^*}^{A+B-1}(x)\right)|\tilde{u}_\theta(x)|dx \\ &< \infty. \end{aligned} \tag{6.25}$$

Then, by DCT, we have $E_g|S_{1n}^* - S_{2n}^*| \to 0$ as $n \to \infty$, and hence, by Markov's Inequality, it follows that

$$S_{1n}^* - S_{2n}^* \xrightarrow{p} 0$$

as $n \to \infty$. □

Lemma 6.8. *Under g, S_{1n}^* converges in distribution to $N_p\left(0, V_g^*\right)$, whenever*

$$V_g^* = Var_g\left\{\int W\left(x, X, h\right) K'\left(\delta_g^*\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A-1}\left(x\right) + \left(A+B\right) f_\theta^{*A+B-1}\left(x\right)\right)\tilde{u}_\theta\left(x\right)dx\right\} \tag{6.26}$$

is finite.

Proof. By the previous lemma, the asymptotic distributions of S_{1n}^* and S_{2n}^* are the same, which has helped us to write, under g,

$$\begin{aligned} & S_{2n}^* \\ = \; & \sqrt{n}\int_x b_n^*\left(x\right)\left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A}\left(x\right) + \left(A+B\right) f_\theta^{*A+B}\left(x\right)\right)\tilde{u}_\theta\left(x\right)dx \\ = \; & \sqrt{n}\int_x \left(\delta_n^*\left(x\right) - \delta_g^*\left(x\right)\right) K'\left(\delta_g^*\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A}\left(x\right) + \left(A+B\right) f_\theta^{*A+B}\left(x\right)\right)\tilde{u}_\theta\left(x\right)dx \\ = \; & \sqrt{n}\int_x \left(g_n^*\left(x\right) - g^*\left(x\right)\right) K'\left(\delta_g^*\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A-1}\left(x\right) + \left(A+B\right) f_\theta^{*A+B-1}\left(x\right)\right)\tilde{u}_\theta\left(x\right)dx \\ = \; & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_x \left(W(x, X_i, h) - E_g(W(x, X_i, h))\right) K'\left(\delta_g^*\left(x\right)\right)\left(A^2\beta^2 e^{\beta f_\theta^{*A}(x)} f_\theta^{*2A-1}\left(x\right) + \left(A+B\right) f_\theta^{*A+B-1}\left(x\right)\right) \\ & \tilde{u}_\theta\left(x\right)dx. \end{aligned}$$

The remaining part immediately follows from the above through an application of the Central Limit Theorem. □

Theorem 6.9. *Under the above-mentioned assumptions, there exists a consistent sequence of roots θ_n^* of the estimating equation (6.2). Moreover, $\sqrt{n}\left(\theta_n^* - \theta^g\right)$ asymptotically follows a p-dimensional normal with mean 0 and $[J_g^*]^{-1}V_g^*[J_g^*]^{-1}$, where J_g^* and V_g^* are as defined in Equation (6.6) and Equation (6.26) (after replacing θ by θ^g in Equation (6.26)).*

Proof. We can prove the main theorem of consistency and normality with slight modifications of the proof of Theorem 5.5 based on continuity and all the lemmas mentioned above. Hence we are just representing here the minor differences.

For the linear term (an analogy of S_1 in proof of Theorem 5.5), we have

$$\begin{aligned}
& \left| \int_x K\left(\delta_n^{*g}(x)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^{*A}(x)} f_{\theta^g}^{*2A}(x) + (A+B)\, f_{\theta^g}^{*A+B}(x)\right)\tilde{u}_{j\theta^g}(x)\,dx \right. \\
- & \left. \int_x K\left(\delta_g^{*g}(x)\right)\left(A^2\beta^2 e^{\beta f_{\theta^g}^{*A}(x)} f_{\theta^g}^{*2A}(x) + (A+B)\, f_{\theta^g}^{*A+B}(x)\right)\tilde{u}_{j\theta^g}(x)\,dx \right| \\
\leq & \ C_1 \int_x \left|\delta_n^{*g}(x) - \delta_g^{*g}(x)\right| \left|u_{j\theta^g}(x)\right| \left(A^2\beta^2 e^{\beta f_{\theta^g}^{*A}(x)} f_{\theta^g}^{*2A}(x) + (A+B)\, f_{\theta^g}^{*A+B}(x)\right) dx.
\end{aligned}$$

and,

$$\begin{aligned}
& E(C_1 \int_x \left|\delta_n^{*g}(x) - \delta_g^{*g}(x)\right| \left|u_{j\theta^g}(x)\right| \left(A^2\beta^2 e^{\beta f_{\theta^g}^{*A}(x)} f_{\theta^g}^{*2A}(x) + (A+B)\, f_{\theta^g}^{*A+B}(x)\right) dx \\
\leq & \ C_1 \int_x \nu^{1/2}(x) \left|u_{j\theta^g}(x)\right| \left(A^2\beta^2 e^{\beta f_{\theta^g}^{*A}(x)} f_{\theta^g}^{*2A-1}(x) + (A+B)\, f_{\theta^g}^{*A+B-1}(x)\right) dx \\
\leq & \ C_1 N^{1/2}(h) \int_x \left|u_{j\theta^g}(x)\right| g^{*1/2}(x) \left(A^2\beta^2 e^{\beta f_{\theta^g}^{*A}(x)} f_{\theta^g}^{*2A-1}(x) + (A+B)\, f_{\theta^g}^{*A+B-1}(x)\right) dx \\
< & \ \infty. \qquad (6.27)
\end{aligned}$$

Hence the term converges, as expected. For the quadratic term (an analogy of S_2 in proof of Theorem 5.5), we have

$$\begin{aligned}
& \left| K'\left(\delta_n^{*g}\right)\left(1+\delta_n^{*g}\right)\left((A+B)^2 f_{\theta^g}^{A+B}(x) + A^3\beta^2 e^{\beta f_{\theta^g}^{A}(x)} f_{\theta^g}^{2A}(x)\left(2+\beta f_{\theta^g}^{A}(x)\right)\right) u_{j\theta^g}(x)\, u_{k\theta^g}(x) \right. \\
- & \left. \ K'\left(\delta_g^{*g}\right)\left(1+\delta_g^{*g}\right)\left((A+B)^2 f_{\theta^g}^{A+B}(x) + A^3\beta^2 e^{\beta f_{\theta^g}^{A}(x)} f_{\theta^g}^{2A}(x)\left(2+\beta f_{\theta^g}^{A}(x)\right)\right) u_{j\theta^g}(x)\, u_{k\theta^g}(x) \right| \\
\xrightarrow{\mathrm{P}} & \ 0,
\end{aligned}$$

and

$$\left| K\left(\delta_n^{*g}(x)\right) - K\left(\delta_g^{*g}(x)\right) \right| \left(A^2\beta^2 e^{\beta f_{\theta^g}^{A}(x)} f_{\theta^g}^{2A}(x) + (A+B)\, f_{\theta^g}^{A+B}(x)\right) u_{jk\theta^g}(x) \xrightarrow{\mathrm{P}} 0. \qquad (6.28)$$

Thus, combining all these,

$$\nabla_{jk} D^* \left(g^*, f_\theta^*\right)|_{\theta=\theta^g} \xrightarrow{p} J_g^{j,k}.$$

The rest is exactly the same as the proof of Theorem 5.5. □

Corollary 6.10. *When the true density g belongs to the model family $\{f_\theta : \theta \in \Theta\}$, i.e., $g = f_\theta$ for some $\theta \in \Theta$, the asymptotic distribution of $\sqrt{n}\left(\theta_n^* - \theta\right)$ is normal with mean 0 and covariance matrix $\left(J^*\right)^{-1} V^* \left(J^*\right)^{-1}$, where $V^* = V_{f_\theta}\left(u_\theta^{\alpha,\lambda,\beta^*}(X)\right)$ and J^* is defined in Equation (6.7).*

6.7 Derivation of Transparent Kernel for the Minimum GSB* Divergence Estimator

Under the continuous model, we have already derived several properties of of our proposed estimators through the implementation of Basu-Lindsay approach. In 1994, Basu and Lindsay have proposed the concept of "Transparent Kernel" and proved that through its imposition, any minimum disparity estimator obtained under the Basu-Lindsay approach have the same asymptotic distribution as the maximum likelihood estimator. In the same spirit, in order to obtain same features on asymptotic distribution of the minimum GSB* divergence estimators, we are going to discuss and develop the required conditions on kernels throughout this section. For this purpose, let us first assume, $G = F_\theta$ for some $\theta \in \Theta$, i.e., the true distribution belongs to the model.

Lemma 6.11. *Suppose the kernel function $W(x, y, h)$ used in smoothing the densities is such that*

$$u_\theta^{\alpha,\lambda,\beta*}(y) = M\left(A^2\beta^2 e^{\beta f_\theta^A(y)} f_\theta^{2A-1}(y) + (A+B) f_\theta^{A+B-1}(y)\right) u_\theta(y) + L \tag{6.29}$$

for a p-vector L depending possibly on α, λ, β and h but not on θ, and a $p \times p$ non-singular matrix M possibly depending on θ, α, λ, β and h, where for each component θ_j of θ, we have either

$$\int \left(A^2\beta^2 e^{\beta f_\theta^A(y)} f_\theta^{2A}(y) + (A+B) f_\theta^{A+B}(y)\right) u_{\theta_j}(y)\, dy = 0 \tag{6.30}$$

or the j-th column of M is independent of θ. Then the influence function for the minimum GSB divergence estimator will be functionally the same as that of the minimum GSB divergence estimator as given in Equation (5.27).*

Proof. We first need to show that conditions (6.29) and (6.30) together imply that

$$E_\theta[-\nabla u_\theta^{\alpha,\lambda,\beta^*}(X)] = M \int \left((A+B) f_\theta^{A+B}(x) + A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x)\right) u_\theta(x)\, u_\theta^T(x) dx. \tag{6.31}$$

In order to prove that, first we differentiate both sides of Equation (6.29) with respect to θ and get,

$$\begin{aligned}
& \nabla u_\theta^{\alpha,\lambda,\beta^*}(x) \\
= \;& M \nabla u_\theta(x) \left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A-1}(x) + (A+B) f_\theta^{A+B-1}(x)\right) \\
+ \;& M \left\{A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A-1}(x) \left(2A - 1 + \beta A f_\theta^A(x)\right) + (A+B)(A+B-1) f_\theta^{A+B-1}(x)\right\} u_\theta(x) u_\theta^T(x) \\
+ \;& [(\nabla_1 M) u_\theta(x)\ (\nabla_2 M) u_\theta(x) \ldots\ (\nabla_p M) u_\theta(x)] \left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A-1}(x) + (A+B) f_\theta^{A+B-1}(x)\right).
\end{aligned} \tag{6.32}$$

Here, ∇_j represents the derivative with respect to the j-th component of θ. Taking expectation with respect to f_θ on both sides of the above equation, we get

$$
\begin{aligned}
E[\nabla u_\theta^{\alpha,\lambda,\beta^*}(X)] &= M \int \nabla u_\theta(x) \left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x) + (A+B) f_\theta^{A+B}(x) \right) dx \\
&+ M \int A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x) \left(2A - 1 + \beta A f_\theta^A(x)\right) u_\theta(x) u_\theta^T(x) dx \\
&+ M \int (A+B)(A+B-1) f_\theta^{A+B}(x) u_\theta(x) u_\theta^T(x) dx. \qquad (6.33)
\end{aligned}
$$

The remaining terms turn out to be zero due to condition (6.30). Again, integrating by parts the first integral of the above equation, we get

$$
\begin{aligned}
& M \int \nabla u_\theta(x) \left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x) + (A+B) f_\theta^{A+B}(x) \right) dx \\
= \; & -M \int \left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x) \left(2A + \beta A f_\theta^A(x)\right) + (A+B)^2 f_\theta^{A+B}(x) \right) u_\theta(x) u_\theta^T(x) dx, \\
& \qquad (6.34)
\end{aligned}
$$

since, $f_\theta^\tau(x) u_\theta(x)$, $\tau > 0$, goes to zero as x tends to its most extreme value in either tail. Now, combining Equation (6.33) and Equation (6.34), we get the desired result given in Equation (6.31).

Then it follows that

$$
\begin{aligned}
\mathrm{IF}(y, F_\theta, T^*_{\alpha,\lambda,\beta}) &= [J_\theta^*]^{-1} \left\{ u_\theta^{\alpha,\lambda,\beta^*}(y) - E_\theta[u_\theta^{\alpha,\lambda,\beta^*}(X)] \right\} \\
&= \left[M \int u_\theta(x) u_\theta^T(x) \left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x) + (A+B) f_\theta^{A+B}(x) \right) dx \right]^{-1} \\
& \quad \left(M \left(A^2\beta^2 e^{\beta f_\theta^A(y)} f_\theta^{2A-1}(y) + (A+B) f_\theta^{A+B-1}(y) \right) u_\theta(y) + L \right) \\
&- \left[M \int u_\theta(x) u_\theta^T(x) \left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x) + (A+B) f_\theta^{A+B}(x) \right) dx \right]^{-1} \\
& \quad \left(E_\theta \left[M u_\theta(X) \left(A^2\beta^2 e^{\beta f_\theta^A(X)} f_\theta^{2A-1}(X) + (A+B) f_\theta^{A+B-1}(X) \right) \right] + L \right) \\
&= \left[\int u_\theta(x) u_\theta^T(x) \left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x) + (A+B) f_\theta^{A+B}(x) \right) dx \right]^{-1} \\
& \quad \left\{ \left(A^2\beta^2 e^{\beta f_\theta^A(y)} f_\theta^{2A-1}(y) + (A+B) f_\theta^{A+B-1}(y) \right) u_\theta(y) \right\} \\
&- \left[\int u_\theta(x) u_\theta^T(x) \left(A^2\beta^2 e^{\beta f_\theta^A(x)} f_\theta^{2A}(x) + (A+B) f_\theta^{A+B}(x) \right) dx \right]^{-1} \\
& \quad E_\theta \left[u_\theta(X) \left(A^2\beta^2 e^{\beta f_\theta^A(X)} f_\theta^{2A-1}(X) + (A+B) f_\theta^{A+B-1}(X) \right) \right]. \qquad (6.35)
\end{aligned}
$$

Equation (6.35) is identical to the expression of the IF of the minimum GSB divergence estimator given in Equation (5.27). Hence the proof. □

Furthermore, in some particular cases of the GSB divergence, we can show that something extra can be achieved through such assumptions on kernel. We can not only prove the result regarding the influence function but also eventually show that the estimating equations of estimators obtained through Basu-Lindsay approach are same with the estimating equations of ordinary estimators obtained without smoothing under discrete model and hence, indeed the estimators are equal. In fact, in such scenario, the condition of $g = f_\theta$ is not required at all. The following corollary will give a clear picture of it.

Corollary 6.12. *Under the assumption of condition (6.29) on kernel function, if we consider, further,*

1. *$\beta = 0$, i.e. the condition becomes $u_\theta^{\alpha *}(y) = M(1+\alpha)u_\theta(y)\, f_\theta^\alpha(y) +$ L, then the estimating equation of minimum DPD* estimator (MDPDE*)*

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta^{\alpha *}(X_i) - E_\theta(u_\theta^{\alpha *}(X)) = 0 \tag{6.36}$$

where $u_\theta^{\alpha *}(y) = \int \tilde{u}_\theta(x)\,\{f_\theta^*(x)\}^\alpha W(x,y,h)dx$ will reduce to the estimating equation of minimum DPD estimator (MDPDE)

$$\frac{1}{n}\sum_{i=1}^{n} f_\theta^\alpha(X_i)\, u_\theta(X_i) - E_\theta(f_\theta^\alpha(X)\, u_\theta(X)) = 0. \tag{6.37}$$

2. *$\alpha = -1$ and $\lambda = 0$: then condition (6.29) becomes $u_\theta^{\beta*}(y) = M\beta^2 u_\theta(y) f_\theta(y) e^{\beta f_\theta(y)} + L$, then the estimating equation of minimum BED* estimator (MBEDE*)*

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta^{\beta*}(X_i) - E_\theta(u_\theta^{\beta*}(X)) = 0 \tag{6.38}$$

where, $u_\theta^{\beta*}(y) = \int \tilde{u}_\theta(x) e^{\beta f_\theta^*(x)} f_\theta^*(x) W(x, y, h) dx$ will reduce to the estimating equation of minimum BED estimator (MBEDE)

$$\frac{1}{n}\sum_{i=1}^{n} e^{\beta f_\theta(X_i)} f_\theta(X_i) u_\theta(X_i) - E_\theta(e^{\beta f_\theta(X)} f_\theta(X) u_\theta(X)) = 0. \tag{6.39}$$

Therefore, under these scenarios, MDPDE and MDPDE as well as MBEDE* and MBEDE will be the same. Trivially it shows their asymptotic equivalence.*

Corollary 6.13. *Under the assumptions of kernel given in Lemma 6.11, the minimum GSB* divergence estimators θ_n^* asymptotically follow normal distribution with mean zero and variance-covariance matrix mentioned in Expression (5.25). Furthermore all first order asymptotic properties of minimum GSB* divergence estimator will be similar to those of the original minimum GSB divergence estimator given in Chapter 5.*

We will refer to any kernel function satisfying condition (6.29) and (6.30) as α, λ, β-transparent kernel. Moreover, at $\alpha = \mathbf{0}$ and $\beta = \mathbf{0}$, this kernel coincides with the transparent kernel defined in Basu and Lindsay (1994). For the mean and the variance of the normal model, the Gaussian kernel is an example of a transparent kernel at $\alpha = \mathbf{0}$ and $\beta = \mathbf{0}$. Note that, in these cases, no restriction needs to be imposed on tuning parameter λ.

6.8 Simulation Results

Earlier, Ghosh et al. (2015), and to some extent Basu et al. (2013), have looked at the values of the tuning parameters that appear to provide the best compromise in terms of robustness and efficiency in divergence tests based on the S-divergence and the DPD, respectively. Eventually, the minimum S^*-divergence estimators belong to the family of minimum GSB* divergence estimators, too, and hence they can be listed as the 'best' MGSBE*s with specific choices of the triplet (α, λ, β). Here, in this section, our main aim is to expand this list with a significant modification, that is, to extend the region of tuning parameter(s) which can generate such 'best' minimum GSB* divergence estimators which are in turn better than the existing S^*-divergence estimators in terms of robustness and/or efficiency, but they will essentially lie outside the family of the minimum DPD and the minimum S^*-divergence estimators.

For efficiency calculations, we consider samples from the pure model, while for illustrations of robustness, the data are generated from contaminated model densities. We choose samples of size 50 from three different setups. For the first case, we choose samples from the $(1-\epsilon)N(0, 9) + \epsilon N(15, 9)$ mixed distribution. For the second case, samples of the same size are drawn from the $(1-\epsilon)N(0, 9)+\epsilon N(0, 100)$ mixture. Lastly, samples of the same size are drawn from the $(1 - \epsilon)N(0, 9) + \epsilon\chi^2_{10}$ mixture. The second component is the contaminant and $\epsilon \in [0, 1)$ is the contaminating proportion and our intention is to estimate the parameters of the main, larger component. The values 0, 0.05, 0.1 and 0.2 are considered for ϵ, and at each contamination level, the whole procedure is replicated 1000 times. In each of the 1000 replications, the normal parameters are estimated corresponding to each contamination level and each (α, λ, β) triplet is considered in our study. We then construct the empirical mean square error (MSE) against the target value of μ $(= 0)$ and σ $(= 3)$, for each tuning parameter combination over the 1000 replications.

For our simulation purpose, we will follow the Basu-Lindsay approach. Our parametric model is the $N(\mu, \sigma^2)$ family, and we will employ the Gaussian kernel with the bandwidth being the well-known Silverman's bandwidth. Through the convolution property of the normal distribution, $f^*_\theta(x)$ will become the density of $N\left(\mu, \sigma^2 + h^2\right)$, where h is the employed bandwidth.

We will first compute the minimum S^*-divergence estimators for each sample over several choices of (α, λ) belonging to the set $\mathbb{A} = \{(\alpha, \lambda) : \alpha \in (0, 1), \lambda \in (-1, 1)\}$ with $\beta = 0$, and then find the minimum GSB* divergence estimators over a grid of non-zero β for each selected pair $(\alpha, \lambda) \in \mathbb{A}$; since β lies over the whole real line, we have restricted its range between $(-8, 8)$ in our explorations. We then compare the MSEs under four scenarios of the MSDE*s with those of the MGSBE*s, and, surprisingly, here also, we have ended up with the 'best' dominating GSB* divergence estimators with $\beta = -4$ corresponding to many of the choices $(\alpha, \lambda) \in \mathbb{A}$ – all of these have a better performance, at least to the extent of the findings in these simulations in either sense of robustness and efficiency.

Here, we present our derived result in a slightly different way than the presentation given in case of the discrete setup. Generally, we have observed four scenarios–

1. First we consider those examples of estimators, where MSDE*s

are not uniformly dominated by any other member of MGDBE* class. There exists some MGSBE*s which may beat these MSDE* in three of the four cases, but not in all of them.

2. This case is similar to the previous one except the thing that in order to get further improvement for cases $\epsilon = 0.05, 0.10, 0.20$, we have to become liberal with respect to the MSE for the case $\epsilon = 0$.

3. Next we have presented the cases where the MGSBE*s beat the MSDE*s in terms of the contaminated data mean square errors, while being competitive in terms of the MSE for $\epsilon = 0$.

4. Lastly, we give examples of a zone of tuning parameters, where some MGSBE* provides better (reduced) measures at each of the four entries compared to the corresponding MSDE*.

For each contaminated model, we are providing different small tables consisting of at least two examples corresponding to each of these four cases rather than representing them in a single table. For each table, each cell consists of a block of five numbers – the four MSEs corresponding to four contaminating proportions $\epsilon = 0, 0.05, 0.1, 0.2$ along with the corresponding triplet (α, λ, β). For comparison purpose, generally, each table consists of even numbers of columns – for each pair of columns, the left one represents the MSEs of MSDE*s estimators whereas the same thing in the right column corresponds to the minimum GSB* divergence estimators. For further ease of understanding, the non-coloured numbers correspond to the MSDE*s, whereas the coloured values correspond to the MGSBE*s. Moreover, examples of the MGSBE*s corresponding to the second, the third and the fourth cases are given through red, green and blue coloured

MSE values in Tables (6.1(B),6.3(B)), Tables (6.1(C),6.3(C)) and Tables (6.1(D),6.2(B),6.3(D)), respectively.

Here we have shown some particular cases only, but actually, we have searched MGSBE*s over the set $\mathbb{A}$, as mentioned earlier. Based on this search, we can conclude the MGSBE*s corresponding to any $\alpha \in [0.1, 0.8]$ and $\lambda \in [-0.7, -0.3]$, with $\beta = -4$, as the competitive alternatives of the existing standard MSDE*s (as suggested by Ghosh et al. (2017)). Some cases belonging to this preferred region are given in Table 6.4.

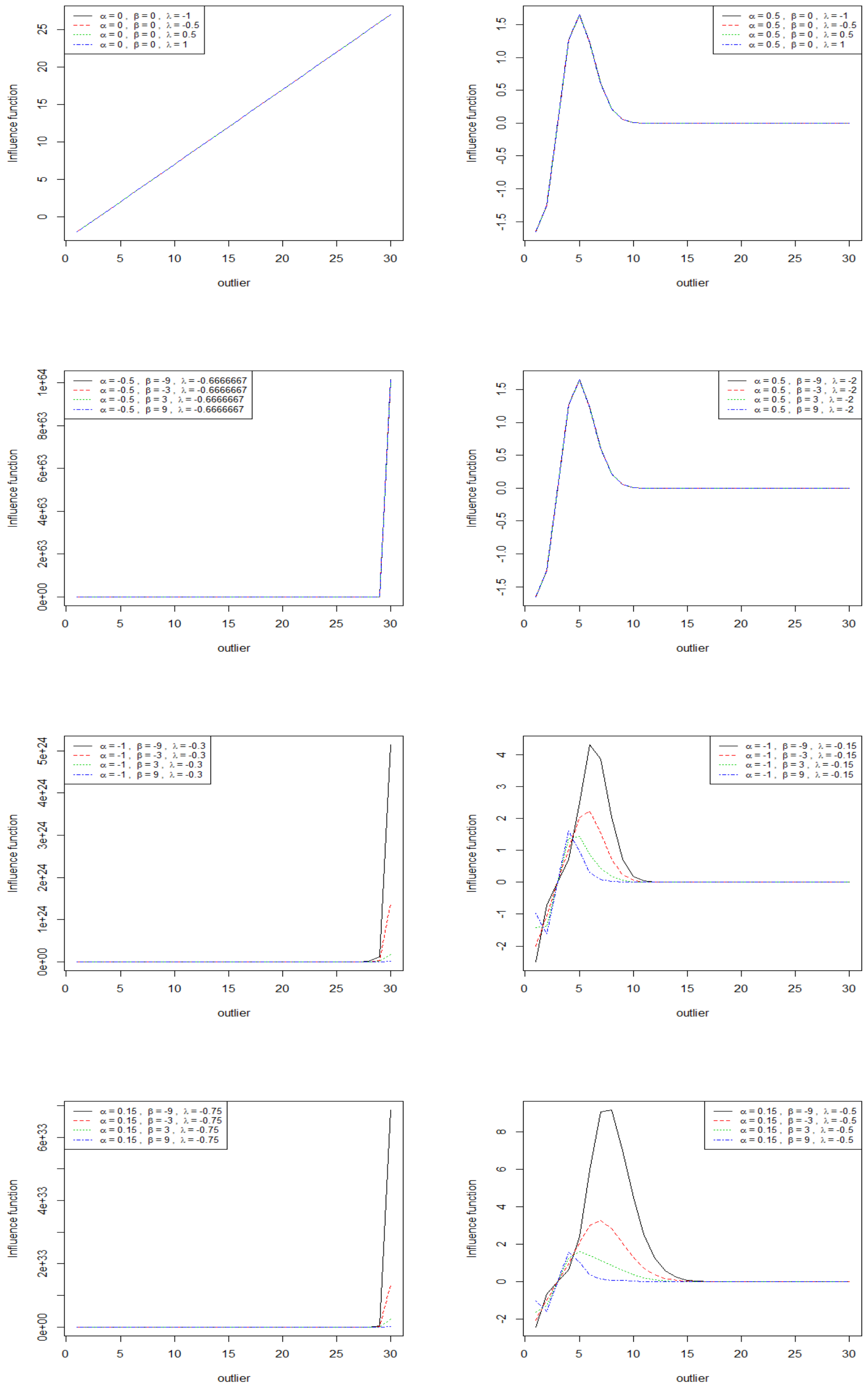


FIGURE 6.1: Examples of unbounded influence functions (left panel) and bounded influence functions (right panel). Only the right panel consists of subsets of $\mathbb{S}$. Those in the left panel are outside!. Here $n = 50$, $f_\theta = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$ with $(\mu, \sigma) = (0, 3)$ and $h = 1.06 * n^{-1/5} * \sigma$

Table 6.1: Comparison of MSEs of MSDE* and MGSBE* under the contaminated model $(1-\epsilon)N(0,9)+\epsilon N(15,9)$

(a)

S^*-divergence	GSB* divergence	S^*-divergence	GSB* divergence
0.3341	0.3222	0.3106	0.3328
0.3991	0.4040	0.3944	0.3995
0.6321	0.6262	0.6815	0.6380
5.2920	5.0263	6.9180	5.7525
(0.6, -0.7, 0)	(0.8, -0.4, -4)	(0.4, -0.7, 0)	(0.8, -0.6, -4)

(b)

S^*-divergence	GSB* divergence
0.2955	0.3103
0.4022	0.3967
0.7689	0.7083
8.3008	7.8682
(0.25, -0.7, 0)	(0.6, -0.5, -4)
0.2813	0.2950
0.5069	0.4093
1.5619	0.8835
13.9369	10.5650
(0.1, -0.5, 0)	(0.6, -0.7, -4)

(c)

S^*-divergence	GSB* divergence
0.3208	0.3208
0.4025	0.3972
0.7079	0.6732
7.0816	6.9023
(0.5, -0.5, 0)	(0.6, -0.4, -4)
0.2904	0.2904
0.4979	0.4199
1.5257	1.0108
13.8318	11.6017
(0.25, -0.3, 0)	(0.4, -0.5, -4)

(d)

S^*-divergence	GSB* divergence	S^*-divergence	GSB* divergence
0.3072	0.3006	0.3321	0.3279
0.4285	0.4026	0.4087	0.3978
0.9598	0.8038	0.7088	0.6479
9.9708	9.4896	6.7164	6.1973
(0.4, -0.3, 0)	(0.5, -0.5, -4)	(0.6, -0.3, 0)	(0.8, -0.7, -4)
0.3047	0.2976	0.3564	0.3467
0.4984	0.4240	0.4167	0.4065
1.4900	0.9696	0.6600	0.6236
13.0312	10.6559	4.9637	4.7685
(0.4, 0, 0)	(0.25, -0.3, -4)	(0.8, 0, 0)	(0.8, -0.3, -4)

Table 6.2: Comparison of MSEs of MSDE* and MGSBE* under the contaminated model $(1-\epsilon)N(0,9)+\epsilon N(0,100)$

(a)

S^*-divergence	GSB* divergence	S^*-divergence	GSB* divergence
0.3106	0.3175	0.2854	0.2904
0.3767	0.3827	0.3711	0.3722
0.4935	0.4912	0.5393	0.5149
1.0387	0.9826	1.3916	1.2132
(0.4, -0.7, 0)	(0.74, -0.7, -4)	(0.1, -0.7, 0)	(0.4, -0.5, -4)

(b)

S^*-divergence	GSB* divergence	S^*-divergence	GSB* divergence
0.3801	0.3626	0.3341	0.3279
0.4233	0.4126	0.3912	0.3892
0.5190	0.5082	0.4947	0.4923
0.9223	0.9125	0.9556	0.9496
(1, -1, 0)	(1, -0.5, -4)	(0.6, -0.7, 0)	(0.8, -0.7, -4)
0.2813	0.2793	0.2904	0.2828
0.3833	0.3800	0.3822	0.3748
0.5891	0.5722	0.5600	0.5437
1.6507	1.5538	1.4593	1.3933
(0.1, -0.5, 0)	(0.25, -0.5, -4)	(0.25, -0.3, 0)	(0.5, -0.7, -4)
0.3175	0.3006	0.3031	0.2904
0.3875	0.3744	0.3960	0.3722
0.5171	0.4986	0.5831	0.5149
1.1346	1.0863	1.5140	1.2132
(0.5, 0, 0)	(0.5, -0.5, -4)	(0.4, 0.2, 0)	(0.4, -0.5, -4)

Table 6.3: Comparison of MSEs of MSDE* and MGSBE* under the contaminated model $(1-\epsilon)N(0,9)+\epsilon\chi^2_{10}$

(a)

S^*-divergence	GSB* divergence	S^*-divergence	GSB* divergence
0.3584	0.3103	0.3222	0.3328
0.4782	0.4714	0.4684	0.4701
0.7903	0.9292	0.8814	0.8437
2.9494	3.8506	3.5611	3.3292
(0.8, -1, 0)	(0.6, -0.5, -4)	(0.5, -0.7, 0)	(0.8, -0.6, -4)

(b)

S^*-divergence	GSB* divergence
0.2854	0.2904
0.5122	0.4980
1.2033	1.1270
5.1364	4.8735
(0.1, -0.7, 0)	(0.4, -0.5, -4)
0.2745	0.2767
0.9164	0.6111
2.3973	1.6196
8.1453	6.8306
(0.1, 0, 0)	(0.4, -0.7, -4)

(c)

S^*-divergence	GSB* divergence
0.2904	0.2904
0.5315	0.4980
1.2500	1.1270
5.3144	4.8735
(0.25, -0.3, 0)	(0.4, -0.5, -4)
0.3296	0.3296
0.4894	0.4707
0.9433	0.8589
3.8557	3.4199
(0.6, 0.2, 0)	(0.6, -0.3, -4)

(d)

S^*-divergence	GSB* divergence	S^*-divergence	GSB* divergence
0.3331	0.3295	0.3072	0.3006
0.4735	0.4707	0.4916	0.4799
0.8623	0.8589	1.0338	1.0112
3.4300	3.4199	4.3622	4.3011
(0.6, -0.5, 0)	(0.6, -0.3, -4)	(0.4, -0.3, 0)	(0.5, -0.5, -4)
0.3175	0.3113	0.2855	0.2793
0.4942	0.4754	0.7409	0.5546
1.0068	0.9489	1.8903	1.3954
4.1993	3.9459	6.9269	6.0041
(0.5, 0, 0)	(0.4, -0.3, -4)	(0.25, 0.2, 0)	(0.25, -0.5, -4)

Table 6.4: MSEs along with triplets (α, λ, β) of some MGSBE*s belonging to the 'best' region

(a) Simulated results generated from the $(1-\epsilon)N(0,9)+\epsilon N(15,9)$ model

GSB* divergence	GSB* divergence	GSB* divergence	GSB* divergence
0.2904	0.2976	0.2767	0.2793
0.4199	0.4240	0.5802	0.4908
1.0108	0.9696	2.3223	1.6413
11.6017	10.6559	18.2625	15.6017
(0.4, -0.5, -4)	(0.25, -0.3, -4)	(0.4, -0.7, -4)	(0.25, -0.5, -4)

(b) Simulated results generated from the $(1-\epsilon)N(0,9)+\epsilon N(0,100)$ model

GSB* divergence	GSB* divergence	GSB* divergence	GSB* divergence
0.2904	0.2976	0.2767	0.2793
0.3722	0.3761	0.3947	0.3800
0.5149	0.5138	0.6324	0.5722
1.2132	1.1797	1.8803	1.5538
(0.4, -0.5, -4)	(0.25, -0.3, -4)	(0.4, -0.7, -4)	(0.25, -0.5, -4)

(c) Simulated results generated from the $(1-\epsilon)N(0,9)+\epsilon\chi^2_{10}$ model

GSB* divergence	GSB* divergence	GSB* divergence	GSB* divergence
0.2904	0.2976	0.2767	0.2793
0.4980	0.4924	0.6111	0.5546
1.1270	1.0714	1.6196	1.3954
4.8735	4.5668	6.8306	6.0041
(0.4, -0.5, -4)	(0.25, -0.3, -4)	(0.4, -0.7, -4)	(0.25, -0.5, -4)

6.9 Real Data Analysis

Here we consider the practical implementation of our proposal on real data and illustrate the strong outlier stability of the proposed method just like the discrete setup.

The natural question that will once again come up in this connection is which set of tuning parameters to use when analyzing a particular set of real data. As already observed in previous chapters, for data which follow the model very closely, the maximum likelihood estimator with the tuning parameter triplet $(\alpha, \lambda, \beta) = (0, 0, 0)$ or something close should work well. For data which involve some departure from the model, more stable divergences such as those with large values of α and/or large negative values of λ may be more desirable. In a particular situation, however, it is not known apriori what proportion of data are anomalous, and we must have an automatic data-based selection plan for tuning parameters. In this context we extend and use the tuning parameter selection strategy described in Chapter 3. Our procedure will construct an empirical mean square error as a function of the tuning parameters and an initial robust pilot estimator, which can then be optimized over the set of tuning parameters. We have already observed that the iterated Warwick Jones algorithm can lead to pilot independent estimates with good performance.

We apply this tuning parameter selection algorithm on our proposed minimum GSB* divergence estimators in our subsequent data analysis exercise. As the simulation study of the previous and the current chapters suggest that the desired results correspond to $\beta = -4$ in most of the cases and the optimals corresponding to restricted as well as unrestricted βs are quite close to each other, it appears to be a reasonable strategy to extend the search over a two-dimensional space with pre-fixed $\beta = -4$. Hence, here we consider the restricted β-case only. For comparison we will also present the tuning parameter(s) selected by the HK, OWJ and IWJ algorithms (as defined in Section 3).

Example 6.1. *(Short's Data): See Stigler (1977) for a description of the data. The raw observations, containing one extreme outlier, are presented in Table 6.5.*

Table 6.5: Short's Data

8.65	8.35	8.71	8.31	8.36	8.58	7.8	7.71	8.30
9.71	8.50	8.28	9.87	8.86	5.76	8.84	8.23	

For the full dataset, the MLE of (μ, σ) *equals* $(8.378, 0.846)$*, whereas the MLE after removing the extreme outlier is* $(8.541, 0.552)$*. After the implementation of the three algorithms on the full data, the OWJ, IWJ and HK estimators will be* $(8.410, 0.423)$*,* $(8.409, 0.537)$ *and* $(8.275, 0.874)$*, corresponding to the triplets* $(0.57, 0.96, -4)$*,* $(0.50, 0.96, -4)$ *and* $(0.08, 0.89, -4)$*, respectively. It is clear that the non-robust HK estimator is quite close to the full data MLE, whereas the opposite scenario is observed in case of the IWJ and the OWJ estimators. On the other hand, if we apply the three algorithms on the outlier-deleted data, then the HK and the IWJ estimators are identical, i.e,* $(8.515, 0.544)$*, which is not far from the outlier-deleted MLE, and the OWJ estimator, although not identical is close to these two, i.e.,* $(8.410, 0.388)$ *corresponding to* $(0.57, 0.96, -4)$*.*

Example 6.2. *(Newcomb's Data): These data include Newcomb's measurements of the velocity of light, which are based on observations, in the U.S. in 1882, of the passage time of light over a certain* *distance. These data can also be found in Stigler (1977). The observations, containing two outliers, are given in Table 6.6. The full data MLE of* (μ, σ) *under the normal model equals* $(26.212, 10.745)$*, whereas the outlier-deleted MLE is* $(27.750, 5.083)$*. For the full data, if we apply the three algorithms, we would get* $(27.472, 4.984)$ *and* $(26.472, 10.483)$ *as the OWJ/IWJ estimator and the HK estimator, corresponding to the triplets* $(0.56, 0.98, -4)$ *and* $(0.45, -0.98, -4)$*, respectively. Evidently, the OWJ/IWJ estimator is close to the outlier-deleted MLE, whereas the HK estimator is closer to the full data MLE. On the other hand, if we consider the outlier-deleted data, the OWJ/IWJ estimator will be* $(27.515, 4.821)$ *corresponding to* $(0.13, 0.98, -4)$ *and the HK estimator will be* $(27.783, 4.940)$ *corresponding to* $(0.01, 0.98, -4)$ *and, as expected, both the estimators are quite close to each other.*

Table 6.6: Newcomb's Data

28	26	33	24	34	−44	27	16	40	−2
29	22	24	21	25	30	23	29	31	19
24	20	36	32	36	28	25	21	28	29
37	25	28	26	30	32	36	26	30	22
36	23	27	27	28	27	31	27	26	33
26	32	32	24	39	28	24	25	32	25
29	27	28	29	16	23				

Example 6.3. *(Alkalinity Data): This dataset is based on the average alkalinity level of public water wells in Suffolk County, New York, USA in 1990. These data can be found in Thode Jr. (2002, p. 347). The sample contains observations of 58 wells, which can be well-modelled by the normal distribution. The data, containing three outliers, are represented in the table below.*

The implementation of these three algorithms and the maximum likelihood estimation on the full and the outlier-deleted data led us to the

TABLE 6.7: Alkalinity Data

29	34	30	36	48	32	42	36	48	38
42	31	48	48	35	46	32	27	45	45
23	35	31	27	34	41	39	36	72	38
39	63	35	31	21	26	41	29	38	60
41	29	44	50	33	33	38	39	28	34
26	26	30	26	37	34	31	33		

following results given in Table 6.8. Some significant fits based on these derived estimates are given in Figure 6.2.

TABLE 6.8: Optimal estimates for the Alkalinity Data

data	method	optimal $\hat{\theta}$	optimal (α, λ, β)
Full data	IWJ/OWJ	$(35.0381, 7.7976)$	$(0.57, -0.58, -4)$
	HK	$(37.1492, 10.5708)$	$(0.01, 0.96, -4)$
	MLE	$(36.9483, 9.6016)$	$(0, 0, 0)$
Clean data	IWJ	$(35.3194, 7.0145)$	$(0.15, -0.58, -4)$
	OWJ	$(35.1444, 7.3329)$	$(0.43, -0.58, -4)$
	HK	$(35.3308, 6.8655)$	$(0.01, 0.96, -4)$
	MLE	$(35.4182, 7.0545)$	$(0, 0, 0)$

6.10 Conclusion

The extension of the Bregman divergence has led us to the introduction of some new super family of divergences which, through further refinement, has helped us to generate highly robust estimators together with an insignificant compromise in efficiency. Under both continuous setup (as well as the discrete set up considered in the previous section), we have figured out some estimators whose performances are even better than the 'best' minimum S^*-divergence estimators. Our next target is to implement this extension in the field of testing of hypotheses. Through some further modification in this extension, we are hoping to achieve some good tests leading to better analyses.

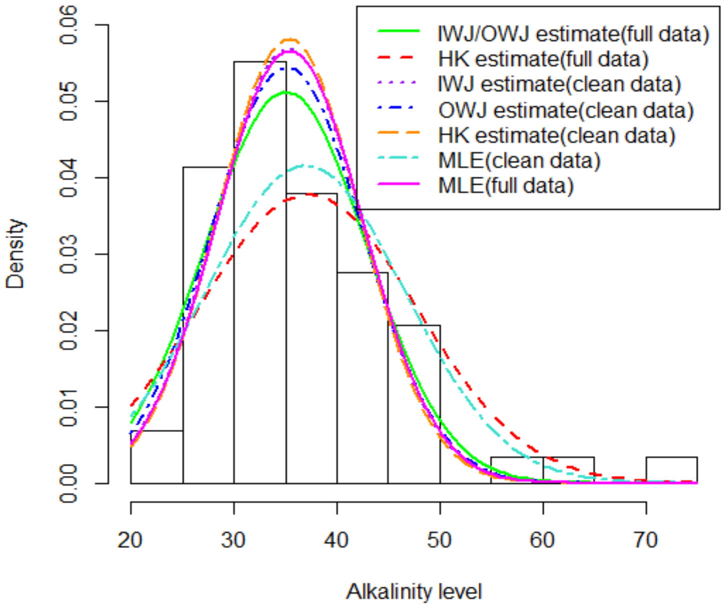


Figure 6.2: Some significant fits for the Alkalinity Data under the Normal model.

Chapter 7

Hypotheses Testing using the Extended Bregman Divergence

7.1 Introduction

Hypothesis testing is one of the two fundamental activities in the field of statistical inference. It helps us to judge the validity of an unsubstantiated claim on the basis of an available sample in any real-life scenario. Although the philosophy of testing procedures and the theory of optimal tests was initiated in the early decades of the twentieth century, new discoveries and modifications of testing tools are still useful and represent challenging research areas. The use of such developments in the statistical domain depends on several of its asymptotic and other optimality issues. One such major issue is the stability of a testing procedure, and the problem of keeping a balance between robustness and efficiency; in the last few chapters we have dealt with this issue in the context of parametric estimation. Here we take up the hypothesis testing case.

The classical likelihood ratio test (LRT) often represents the default application tool in the hypothesis testing context and has some asymptotic optimality properties, but its lack of robustness can lead to problems in real situations. Our target is to devise a test (or a system of tests) based on the extended Bregman divergence and its variants which provide a good compromise between efficiency and robustness.

Since the LRT is extremely non-robust, "disparity difference tests" and other tests in the similar spirit have been considered useful, robust alternatives to it in recent years. Several divergences have been used for this purpose which utilize the minimum possible value of the divergence between the data and a model density. See, for example, Simpson (1989), Lindsay (1994), Basu et al. (2011), Basu et al. (2013), Ghosh et al. (2015), etc. In our research, we will use the Bregman divergence and its extensions for testing purposes.

Proposing the extended Bregman divergence, and using it for statistical inference, has been a primary focus of this thesis. In the previous chapters we have used a particular member of this extended Bregman class (the GSB family) to demonstrate the advantages derived out of it in parametric estimation. In the present chapter we will perform tests of hypothesis using divergences that are based on the extended Bregman idea, but because of certain advantages which will be apparent from the future discussion, we will base all our demonstrations in case of hypothesis testing on yet another generalized family of divergences. This divergence is referred to as the generalized S-divergence (GSD) family. This family has been derived earlier by Ghosh and Basu (2018) through entirely different considerations of testing tubular hypothesis; also see Park and Basu (2003), who developed an earlier, simpler, one parameter version of this divergence (the generalized Kullback-Leibler divergence) based on fewer parameters (also through tubular hypothesis ideas).

The GSD family has the form

$$Q_{(\alpha,\tau,\gamma)}(g,f) = \frac{1}{\tau\bar{\tau}(\alpha-\gamma)} \int \left[\left\{ \tau \left(\frac{g}{f}\right)^{1+\alpha} + \bar{\tau} \right\} - \left\{ \tau \left(\frac{g}{f}\right)^{1+\gamma} + \bar{\tau} \right\}^{\frac{1+\alpha}{1+\gamma}} \right] f^{1+\alpha} d\mu, \tag{7.1}$$

where $\alpha \in [0,1)$, $\tau \in (0,1)$, $\bar{\tau} = 1-\tau$ and $\gamma \in \mathbb{R} - \{-1\}$ with $\gamma \neq \alpha$. Moreover, this divergence can be extended over $\alpha \in [0,1]$, $\tau \in [0,1]$ and $\gamma \in \mathbb{R}$ through their continuous limits. In the next subsection we will outline the development of the GSD following the extended Bregman principles. Tests of hypotheses based on the PD, DPD and the S-divergence have already been attempted before; see, e.g., Ghosh et al. (2017). Since the S-divergence is a special case of the GSD, studying the tests of hypotheses based on the GSD allows an exploration of this larger superfamily beyond the tests based on S-divergences, and the additional benefits that can be derived out of this in hypothesis testing.

7.2 GSD as a Special Form of Bregman Divergence

Due to several desirable properties of the Bregman divergence, it is always an advantage to show any other divergence as a special case of the Bregman divergence, if possible. In that case, the said desirable properties are automatically inherited by that divergence. Previously, the DPD family and the BED family could be expressed as special forms of this divergence, but the PD family and the S-divergence family could not. Now, through our extension, it is possible for us to move one step ahead – the last two above-mentioned divergence families have become a part of this extension. The GSD family, which is our primary tool in the demonstration of robust testing procedures, has been expressed as a special form of this extension through a slight modification – more specifically, a convex combination of two extended Bregman divergences with specific choices of arguments, ψ and k.

Considering Equation (4.1) with $\psi\left(x\right) = \frac{x^{\frac{A+B}{A}}}{B}$, $k = 1$ along with $f^{1+\gamma}$ and $\tau g^{1+\gamma} + \bar{\tau} f^{1+\gamma}$ as arguments, we would get

$$\begin{aligned} D_1\left(g, f\right) &= \int \left\{ \frac{1}{B} f^{1+\alpha} - \frac{1}{B}\left(\tau g^{1+\gamma} + \bar{\tau} f^{1+\gamma}\right)^{\frac{1+\alpha}{1+\gamma}} \right\} \\ &- \frac{A+B}{AB} \int \left(\tau f^{1+\gamma} - \tau g^{1+\gamma}\right)\left(\tau g^{1+\gamma} + \bar{\tau} f^{1+\gamma}\right)^{\frac{\alpha-\gamma}{1+\gamma}}. \end{aligned}$$

Here $A = 1 + \lambda(1-\alpha)$ and $B = \alpha - \lambda(1-\alpha)$. Again, if we consider the same equation associated with the same $\psi\left(\cdot\right)$ and k along with $g^{1+\gamma}$ and $\tau g^{1+\gamma} + \bar{\tau} f^{1+\gamma}$ as arguments, we then get

$$\begin{aligned} D_2\left(g, f\right) &= \int \left\{ \frac{1}{B} g^{1+\alpha} - \frac{1}{B}\left(\tau g^{1+\gamma} + \bar{\tau} f^{1+\gamma}\right)^{\frac{1+\alpha}{1+\gamma}} \right\} \\ &- \frac{A+B}{AB} \int \left(\bar{\tau} f^{1+\gamma} - \bar{\tau} g^{1+\gamma}\right)\left(\tau g^{1+\gamma} + \bar{\tau} f^{1+\gamma}\right)^{\frac{\alpha-\gamma}{1+\gamma}}. \end{aligned}$$

Now, consideration of the 'convex combination' of $D_1(g,f)$ and $D_2(g,f)$ would produce to the following

$$\begin{aligned}
& \bar{\tau} D_1(g,f) + \tau D_2(g,f) \\
= \;& \tau\bar{\tau} \int \left\{ \frac{1}{\tau B} f^{1+\alpha} + \frac{1}{\bar{\tau} B} g^{1+\alpha} - \frac{1}{\tau\bar{\tau} B} \left(\tau g^{1+\gamma} + \bar{\tau} f^{1+\gamma} \right)^{\frac{1+\alpha}{1+\gamma}} \right\} \\
= \;& \tau\bar{\tau} \int \left\{ \frac{1}{\tau B} + \frac{1}{\bar{\tau} B} \left(\frac{g}{f} \right)^{1+\alpha} - \frac{1}{\tau\bar{\tau} B} \left(\tau \left(\frac{g}{f} \right)^{1+\gamma} + \bar{\tau} \right)^{\frac{1+\alpha}{1+\gamma}} \right\} f^{1+\alpha} \\
= \;& \tau\bar{\tau} \times Q_{(\alpha,\tau,\gamma)}(g,f). \qquad (7.2)
\end{aligned}$$

Moreover, a convex combination of two divergences is also a divergence – this fact evidently proves that this GSD family is indeed a divergence family, satisfying all the criteria of being a divergence. Thus the application of this divergence will provide an illustration of the usefulness of divergences generated in the spirit of the extended Bregman principle in the context of hypothesis testing.

In classical inference, the LRT is the first choice in practically all applications, but due to its non-robust characteristics, robust tests have been in demand and the search for improved robust tests is still meaningful. In this journey, a remarkable step was taken by Basu et al. (2013) – by introducing tests based on the DPD. This is called the Density Power Divergence based Test (DPDT) with α as the tuning parameter. Later on, Ghosh et al. (2015) introduced another robust test based on the S-divergence. This is called the S-divergence based Test (SDT) with a pair of tuning parameters (α, λ), of which the DPDT is a special case; more specifically, the SDT with $\lambda = 0$ coincides with the DPDT having the same value of α as the SDT. Here, we take it one step further by considering the generalization of the SDT – we have constructed a test based on the Generalized S-Divergence mentioned earlier, with the hope of being successful in our journey of developing a more general class of robust tests for more refined analysis in real life scenarios. Although, this divergence (indexed by the triplet of tuning parameters (α, τ, γ)) is denoted by 'Q' with its arguments g and f, keeping similarity with the name of this divergence, we will refer the test based on it as the generalized S-divergence based test (GSDT) with the triplet (α, τ, γ).

Both the DPD and the S-divergence are special cases of the GSD. For fixed α and τ, the GSD in Equation (7.1) with $\gamma \to -1$ leads to the S-divergence with the same α and $\lambda = \frac{\alpha\tau-(1-\tau)}{1-\alpha}$. On the other hand, keeping α fixed, if we take a specific choice of τ depending on α, i.e., $\tau = \frac{1}{1+\alpha}$ and $\gamma \to -1$, then this GSD leads us to the DPD indexed by the same α. Here, using this GSD to define the GSDT, we are basically extending the path of availing robust tests through some theorems, simulations and real life data analysis.

7.3 Testing Parametric Hypothesis using GSD (Simple Null Hypotheses)

Consider the problem of testing a simple null hypothesis based on the available sample(s). Let $\mathcal{E} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ represent the parametric family of densities. Moreover, we assume that the true data generating density, denoted by g, belongs to this model family. Now, we will develop a testing procedure using the GSD family in this setup. For clear presentation, we consider the following cases separately.

(i) One Sample Problem: $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ where θ_0 is a fixed value in the parameter space Θ. In this case, a random sample of size n will be available from the population under study.

(ii) Two Sample Problem: $H_0 : \theta_1 = \theta_2$ vs. $H_1 : \theta_1 \neq \theta_2$ where θ_1 and θ_2 are fixed values of the model parameters describing two different populations. In this scenario, two independent random samples of sizes n and m are available from these two populations.

Ghosh and Basu (2018) have already established the first order influence function of the minimum GSD estimator (MGSDE) at the model, which is a function of α alone and shown it to be identical to the influence function of the MDPDE at the model with the same value of α. Therefore, their theoretical robustness properties are also similar. It also turns out that the asymptotic distribution of the MGSDE is the same as that of the MDPDE with the same value of α (irrespective of the values of τ and γ) at least in the case of discrete models. We will keep these desirable properties in mind when constructing the test procedures in the following sections.

7.3.1 One Sample Problem

For testing the null hypothesis under the one sample problem, the general test statistic based on the GSD with parameters α, τ and γ will be defined as

$$T_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha, \theta_0\right) = 2n\, Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\theta_0}\right), \tag{7.3}$$

where $\hat{\theta}_\alpha$ is the MDPDE of θ at α.

It is important to explain the motivation behind describing the test statistic as in Equation (7.3). Notice that this test statistic, although not directly in the disparity difference form, is developed in the same spirit. When the model is correctly specified, and θ_0 represents the true value of the parameter, any statistic of the form (7.3) with a consistent estimator of the model parameter in the first argument on the left hand side of Equation (7.3) is likely to assume a small value for moderate to large sample sizes. The most appropriate choice, of course, would be to use the estimator $\hat{\theta}_{(\alpha,\tau,\gamma)}$, which represents the model density closest to the data density for the choice of the given GSD. However, we prefer to use the estimator $\hat{\theta}_\alpha$ instead for two reasons.

Firstly, the estimator ($\hat{\theta}_\alpha$) used in the statistic (7.3) is a member of the minimum density power divergence estimator class, and thus the evaluation of this estimator involves no non-parametric density estimation, leading to huge theoretical and computational advantages; secondly, whenever $\theta = \theta_0$, Ghosh and Basu (2018) have already shown that, at the model, the influence function of the MGSDE is the same with that of the MDPDE for the same value of α. Their asymptotic distributions are also the same for discrete parametric models. Thus, although it is neither a theoretical or computational necessity, we expect that the asymptotic behavior of the statistic in (7.3) would be similar to what would have been obtained if one used the statistic $\hat{\theta}_{(\alpha,\tau,\gamma)}$ in its definition instead.

This is the reason why we use this divergence as our main platform for the testing part. Since the MDPDE is easy to use, we want to switch towards such a divergence, whose null distribution is the same as the distribution of the MDPDE. On the other hand, although our GSB divergence is able to generate such estimators which are strongly robust and which are more or similarly efficient compared to the existing standard estimators, its expression is complicated and the estimator will inherently require a non-parametric smoothing component in the computation of the divergence. This is why, in order to explore the advantage of the usage of the extended Bregman divergence, our initial choice is to proceed with the GSD divergence, although, undoubtedly, the usage of the GSB divergence in testing is an open path for future research, with an expectation of getting some desirable results as in the case of estimation.

7.3.1.1 Some Theorems

Here, we are going to present a combined set of conditions given in Lehmann (1983) and Basu et al. (2011), respectively. These conditions are necessary to establish the asymptotic properties of our proposed GSD test statistic.

B1. Each model density of the parametric model family, $\mathcal{E}$, as well as the true density g, must have a common support, χ, which is independent of θ.

B2. As the true density is assumed to belong to the model, we have $g = f_{\theta^g}$ for some $\theta^g \in \Theta$. There exists an open subset $\omega \subset \Theta$, of which the true parameter θ^g is an interior point. For almost all x, $f_\theta\left(x\right)$ possesses third partial derivatives of the type $\nabla_{jkl} f_\theta\left(x\right)$ for all $\theta \in \omega$. Moreover, the third order partial derivatives are continuous with respect to θ.

B3. The first and second order derivatives of the score function u_θ are such that
$E_\theta(u_\theta(X)) = 0, \quad I(\theta) = E_\theta(u_\theta(X)u_\theta(X)^T) = -E_\theta(\nabla u_\theta(X)).$

B4. The Fisher information matrix $I(\theta)$ is positive definite for all $\theta \in \omega$.

B5. The matrix J_α is positive definite (where J_α is as defined later in Equation (7.5)).

B6. The integrals $\int f_\theta^{1+\alpha}(x)dx$ and $\int f_\theta^\alpha(x)g(x)dx$ are thrice differentiable with respect to θ. Moreover, the derivatives and the integrals can be interchanged.

B7. For all $x \in \chi$, there exists a function $M_{jkl}(x)$ such that it dominates the third order partial derivative of

$$\int f_\theta^{1+\alpha}(x)dx - (1+\frac{1}{\alpha})f_\theta^\alpha(x) \tag{7.4}$$

in absolute value and $E_g(M_{jkl}(X)) < \infty$, for all j,k,l and for all $\theta \in \omega$.

From now onwards, we are going to refer these conditions as the *Lehmann and Basu et al. conditions (B1)-(B7).*

Lemma 7.1. *(Corollary 2.1 of Dik and de Gunst (1985)). For $X \sim N_q(0,\Sigma)$ and a q-dimensional, real-valued symmetric matrix B, the distribution of X^TBX is the same with the distribution of $\sum_{i=1}^r \lambda_i Z_i^2$ $\Big($where $Z_i \overset{i.i.d.}{\sim} N(0,1)\Big)$, $r = rank(\Sigma B \Sigma)$ $(r \geq 1)$ and $\lambda_1,\ldots,\lambda_r$ are the non-zero eigenvalues of $B\Sigma$.*

Lemma 7.2. *(Corollary 2.2 of Dik and de Gunst (1985)). For $X \sim N_q(\mu,\Sigma)$ and for Q being a real, symmetric, non-negative definite matrix of order q, the quadratic form X^TQX has the same distribution as that of the random variable $\sum_{i=1}^r \lambda_i(U_i + w_i)^2 + \zeta$, where $r = rank(\Sigma Q \Sigma)$, $\lambda_1,\ldots,\lambda_r$ are the positive eigenvalues of $Q\Sigma$, $U_1,\ldots,U_r \overset{i.i.d.}{\sim} N(0,1)$, $\boldsymbol{w} = \Lambda_r^{-1}P^TS^TQ\mu$ and $\zeta = \mu^TQ\mu - w^T\Lambda_p w$ with S being any $q \times s$ square root of Σ, $\Lambda_r = diag(\lambda_1,\ldots,\lambda_r)$ and P is the matrix of the corresponding orthonormal eigenvectors.*

Theorem 7.3. *Under the Lehmann and Basu et al. conditions (B1)-(B7), the asymptotic null distribution of the test statistic $T_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha,\theta_0\right)$*

is the same as the distribution of $\sum_{i=1}^{r} \lambda_i Z_i^2$*, where* $Z_i \sim N(0,1)$ *independently and* λ_i*'s are the non-zero eigenvalues of* $A_\alpha(\theta_0)\Sigma_\alpha(\theta_0)$ *with*

$$\begin{aligned}
A_\alpha(\theta_0) &= \nabla^2 Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)\big|_{\theta=\theta_0} \\
&= \left((1+\alpha)\int f_{\theta_0}^{\alpha-1}\frac{\partial f_{\theta_0}}{\partial\theta_i}\frac{\partial f_{\theta_0}}{\partial\theta_j}\right)_{i,j=1,\ldots,p}, \\
r &= rank\left(V_\alpha(\theta_0)\, J_\alpha^{-1}(\theta_0)\, A_\alpha(\theta_0)\, J_\alpha^{-1}(\theta_0)\, V_\alpha(\theta_0)\right), \\
\Sigma_\alpha(\theta_0) &= J_\alpha^{-1}(\theta_0)V_\alpha(\theta_0)J_\alpha^{-1}(\theta_0),
\end{aligned}$$

where

$$J_\alpha(\theta_0) = \int u_{\theta_0}u_{\theta_0}^T f_{\theta_0}^{1+\alpha}, \tag{7.5}$$

$$V_\alpha(\theta_0) = \int u_{\theta_0}u_{\theta_0}^T f_{\theta_0}^{1+2\alpha} - \left(\int u_{\theta_0} f_{\theta_0}^{1+\alpha}\right)\left(\int u_{\theta_0} f_{\theta_0}^{1+\alpha}\right)^T. \tag{7.6}$$

Proof. We consider the second order of Taylor series expansion of $Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)$ around $\theta = \theta_0$ at $\theta = \hat{\theta}_\alpha$ (MDPDE at fixed α) and we get

$$\begin{aligned}
Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\theta_0}\right) &= Q(\alpha,\tau,\gamma)\left(f_{\theta_0}, f_{\theta_0}\right) \\
&+ \sum_{i=1}^{p} \nabla_i Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)\Bigg|_{\theta=\theta_0}\left(\hat{\theta}_\alpha^i - \theta_0^i\right) \\
&+ \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \nabla_{ij} Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)\Bigg|_{\theta=\theta_0}\left(\hat{\theta}_\alpha^i - \theta_0^i\right)\left(\hat{\theta}_\alpha^j - \theta_0^j\right) \\
&+ o\left(||\hat{\theta}_\alpha - \theta_0||^2\right),
\end{aligned}$$

where ∇_i and ∇_{ij} are as defined in Chapter 1. Evidently, $Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_0}, f_{\theta_0}\right) = 0$. Also, $\nabla_i Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)\Bigg|_{\theta=\theta_0} = 0.$

Moreover, $A_\alpha(\theta_0) = \big(\big(a_{ij}^\alpha(\theta_0)\big)\big)$ with

$$\begin{aligned} a_{ij}^\alpha(\theta_0) &= \nabla_{ij} Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)\Big|_{\theta=\theta_0} \\ &= \left(\frac{\partial^2 Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)}{\partial\theta_i \partial\theta_j}\right)\Bigg|_{\theta=\theta_0} \\ &= (1+\alpha)\int f_{\theta_0}^{\alpha-1}\left(\frac{\partial f_\theta}{\partial\theta_i}\frac{\partial f_\theta}{\partial\theta_j}\right)\Bigg|_{\theta=\theta_0}. \end{aligned}$$

Therefore,

$$\begin{aligned} T_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha, \theta_0\right) &= 2n\, Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\theta_0}\right) \\ &= \sqrt{n}\left(\hat{\theta}_\alpha - \theta_0\right)^T A_\alpha(\theta_0)\sqrt{n}\left(\hat{\theta}_\alpha - \theta_0\right) + n \times o\left(||\hat{\theta}_\alpha - \theta_0||^2\right). \end{aligned}$$

Hence, $T_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha, \theta_0\right)$ and $\sqrt{n}\left(\hat{\theta}_\alpha - \theta_0\right)^T A_\alpha(\theta_0)\sqrt{n}\left(\hat{\theta}_\alpha - \theta_0\right)$ have the same asymptotic distribution. Again, we already know the asymptotic distribution of MDPDE $\hat{\theta}_\alpha$, which is given by,

$$\sqrt{n}\left(\hat{\theta}_\alpha - \theta_0\right) \overset{\text{a}}{\sim} N\left(0, J_\alpha^{-1}(\theta_0)\, V_\alpha(\theta_0)\, J_\alpha^{-1}(\theta_0)\right),$$

where J_α and V_α have the expressions given in Equation (7.5) and (7.6). Furthermore, according to Lemma 7.1, we can say,

$$\begin{aligned} T_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha, \theta_0\right) &\overset{\text{D}}{=} \sqrt{n}\left(\hat{\theta}_\alpha - \theta_0\right)^T A_\alpha(\theta_0)\sqrt{n}\left(\hat{\theta}_\alpha - \theta_0\right) \\ &\overset{\text{D}}{=} \sum_{i=1}^r \lambda_i Z_i^2,\ Z_i \overset{\text{i.i.d.}}{\sim} N(0,1), \end{aligned}$$

where, $\overset{\text{D}}{=}$ denotes the quantities on either side are distributionally equivalent in asymptotic sense. Here λ_i's are non-zero eigenvalues of $A_\alpha(\theta_0)\, J_\alpha^{-1}(\theta_0)\, V_\alpha(\theta_0)\, J_\alpha^{-1}(\theta_0)$ and
$r = rank\left(V_\alpha(\theta_0)\, J_\alpha^{-1}(\theta_0)\, A_\alpha(\theta_0)\, J_\alpha^{-1}(\theta_0) V_\alpha(\theta_0)\right)$. □

Next, we consider the test under a contiguous alternative hypothesis

$$H_{1,n} : \theta = \theta_n, \text{ where, } \quad \theta_n = \theta_0 + \frac{\Delta}{\sqrt{n}}, \tag{7.7}$$

where Δ is a fixed vector in $\mathbb{R}^p$ such that $\theta_n \in \Theta \subset \mathbb{R}^p$. Now, we are going to present the asymptotic distribution of our proposed test statistic under (7.19) in the following theorem.

Theorem 7.4. *Under the Lehmann and Basu et al. conditions (B1)-(B7), whenever $H_{1,n}$ is true, the asymptotic distribution of $T_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha, \theta_0\right)$ is the same as the distribution of $\sum_{i=1}^{r} \lambda_i (Z_i + w_i)^2 + \xi$, where $Z_i \sim$ N(0,1) independently, λ_i's are the positive eigenvalues of $A_\alpha(\theta_0) J_\alpha^{-1}(\theta_0)$ $V_\alpha(\theta_0)\ J_\alpha^{-1}(\theta_0)$, the values $\mathbf{w} = (w_1, \ldots, w_r)^T$ and ξ are given by*

$$\begin{aligned} \mathbf{w} &= \Lambda_r^{-1} P^T S^T A_\alpha(\theta_0)\Delta, \\ \xi &= \Delta^T A_\alpha(\theta_0)\Delta - w^T \Lambda_r w. \end{aligned}$$

Also, S is any square root of $\left[J_\alpha^{-1}(\theta_0) V_\alpha(\theta_0) J_\alpha^{-1}(\theta_0)\right]$, $\Lambda_r = diag(\lambda_1, \ldots, \lambda_r)$ and P is the matrix of corresponding orthonormal eigenvectors.

Proof. $\sqrt{n}\left(\hat{\theta}_\alpha - \theta_0\right)$ can be rewritten as

$$\sqrt{n}\left(\hat{\theta}_\alpha - \theta_0\right) = \sqrt{n}\left(\hat{\theta}_\alpha - \theta_n\right) + \sqrt{n}\left(\theta_n - \theta_0\right) = \sqrt{n}\left(\hat{\theta}_\alpha - \theta_n\right) + \Delta. \tag{7.8}$$

Under $H_{1,n}$ we get, from Equation (7.8) and the consistency of $\hat{\theta}_\alpha$

$$\sqrt{n}\left(\hat{\theta}_\alpha - \theta_0\right) \overset{a}{\sim} N\left(\Delta, J_\alpha^{-1}(\theta_0)\, V_\alpha(\theta_0)\, J_\alpha^{-1}(\theta_0)\right).$$

Again, we established that $T_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha, \theta_0\right)$ and $\sqrt{n}\left(\hat{\theta}_\alpha - \theta_0\right)^T A_\alpha(\theta_0)\sqrt{n}\left(\hat{\theta}_\alpha - \theta_0\right)$ have the same asymptotic distribution. Therefore, the rest of the proof can be readily established using Lemma 7.2. □

Theorem 7.5. *Under the Lehmann and Basu et al. conditions (B1)-(B7), the power function of the test statistic defined in Equation (7.3) at significance level β is given by*

$$\Pi_{n,\beta}^{(\alpha,\tau,\gamma)} \;=\; 1 - \Phi_n\left(\frac{\sqrt{n}}{\sigma_{(\alpha,\tau,\gamma)}(\theta^*)}\left(\frac{t_\beta^{(\alpha,\tau,\gamma)}}{2n} - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right)\right)\right)$$

where, Φ_n tends to Φ (the cdf of N(0, 1)) uniformly, $t_\beta^{(\alpha,\tau,\gamma)}$ is the quantile of the order $(1-\beta)$ of the asymptotic distribution of the test statistic given in Equation (7.3) and $\sigma^2_{(\alpha,\tau,\gamma)}(\theta^*) = M^T_{(\alpha,\tau,\gamma)}(\theta^*)J_\alpha^{-1}(\theta^*)V_\alpha(\theta^*)J_\alpha^{-1}(\theta^*)M_{(\alpha,\tau,\gamma)}(\theta^*)$.

Proof. A first order Taylor series expansion of $Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\theta_0}\right)$ around θ^*, $\theta^* \neq \theta_0$ at $\theta = \hat{\theta}_\alpha$, we get,

$$\begin{aligned} Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\theta_0}\right) &= Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right) + \sum_{i=1}^{p}\frac{\partial}{\partial\theta_i}Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)\Bigg|_{\theta=\theta^*}\left(\hat{\theta}_\alpha^i - \theta_i^*\right) \\ &+ o\left(||\hat{\theta}_\alpha - \theta^*||\right). \end{aligned}$$

From the asymptotic theory of MDPDE, under $\theta = \theta^*$,

$$\sqrt{n}\left(\hat{\theta}_\alpha - \theta^*\right) \overset{\text{a}}{\sim} N\left(0, J_\alpha^{-1}(\theta^*)\, V_\alpha(\theta^*)\, J_\alpha^{-1}(\theta^*)\right), \text{ as } n \to \infty.$$

Therefore, the asymptotic distributions of $\sqrt{n}\left(Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\theta_0}\right) - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right)\right)$ and $M^T_{(\alpha,\tau,\gamma)}(\theta^*)\sqrt{n}\left(\hat{\theta}_\alpha - \theta^*\right)$ are same due to the fact that $\sqrt{n} \times\ o\left(||\hat{\theta}_\alpha - \theta^*||\right) = o_p(1)$, where $M_{(\alpha,\tau,\gamma)}(\theta) = \nabla Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)$.

Hence, $\sqrt{n}\left(Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\theta_0}\right) - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right)\right) \overset{\text{a}}{\sim} N(0, \sigma^2_{(\alpha,\tau,\gamma)}(\theta^*))$, where $\sigma^2_{(\alpha,\tau,\gamma)}(\theta^*)$ has the expression as mentioned in the statement of the theorem. Therefore, the asymptotic power at θ^* can be evaluated as

$$\begin{aligned}
& P_{\theta=\theta^*}\left(T_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha, \theta_0\right) > t_\beta^{(\alpha,\tau,\gamma)}\right) \\
= \; & 1 - P_{\theta=\theta^*}\left(Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\theta_0}\right) \leq \frac{t_\beta^{(\alpha,\tau,\gamma)}}{2n}\right) \\
= \; & 1 - P_{\theta=\theta^*}\left(\frac{\sqrt{n}\left(Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\theta_0}\right) - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right)\right)}{\sigma_{(\alpha,\tau,\gamma)}(\theta^*)}\right. \\
& \qquad \left. \leq \frac{\sqrt{n}\left(\frac{t_\beta^{(\alpha,\tau,\gamma)}}{2n} - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right)\right)}{\sigma_{(\alpha,\tau,\gamma)}(\theta^*)}\right) \\
\to \; & 1 - \Phi\left(\left(\frac{t_\beta^{(\alpha,\tau,\gamma)}}{2n} - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right)\right)\frac{\sqrt{n}}{\sigma_{(\alpha,\tau,\gamma)}(\theta^*)}\right).
\end{aligned}$$

□

Moreover, at any significance level β, as $n \to \infty$, this rejection rule $T_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha, \theta_0\right) > t_\beta^{(\alpha,\tau,\gamma)}$ leads to a probability which tends to 1. Hence, this test is consistent.

7.3.2 Two Sample Problem

Here we have two independent populations modeled by the same parametric family, and we want to test whether the parameters for the two populations are equal. In this case, the test statistic will be modified as

$$S_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right) = \frac{2nm}{m+n} Q_{(\alpha,\tau,\gamma)}\left(f_{{}^{(1)}\hat{\theta}_\alpha}, f_{{}^{(2)}\hat{\theta}_\alpha}\right), \qquad (7.9)$$

where, ${}^{(1)}\hat{\theta}_\alpha$ and ${}^{(2)}\hat{\theta}_\alpha$ are the MDPDEs at α based on independent samples of sizes n and m, respectively, from the two populations.

7.3.2.1 Some Theorems

Theorem 7.6. *Assume that the Lehmann and Basu et al conditions (B1)-(B7) are satisfied by both populations, which are modeled by the same parametric family. The asymptotic distribution of $S_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right)$ under H_0 is the same as the distribution of $\sum_{i=1}^{r} \lambda_i Z_i^2$, where $Z_1, \ldots, Z_r$ are independent standard normal random variables and $\lambda_1, \ldots, \lambda_r$ are the non-zero eigenvalues of $A_\alpha(\theta_1)J_\alpha^{-1}(\theta_1)V_\alpha(\theta_1)J_\alpha^{-1}(\theta_1)$ and $r = rank\left(V_\alpha(\theta_1)J_\alpha^{-1}(\theta_1)A_\alpha(\theta_1)J_\alpha^{-1}(\theta_1)V_\alpha(\theta_1)\right)$.*

Proof. From the known results about the MDPDE, we have, under f_{θ_1} and f_{θ_2}, respectively,

$$\begin{aligned}
\sqrt{n}\left({}^{(1)}\hat{\theta}_\alpha - \theta_1\right) &\overset{\text{a}}{\sim} N\left(0, J_\alpha^{-1}(\theta_1)\, V_\alpha(\theta_1)\, J_\alpha^{-1}(\theta_1)\right) \quad \text{and} \\
\sqrt{m}\left({}^{(2)}\hat{\theta}_\alpha - \theta_2\right) &\overset{\text{a}}{\sim} N\left(0, J_\alpha^{-1}(\theta_2)\, V_\alpha(\theta_2)\, J_\alpha^{-1}(\theta_2)\right).
\end{aligned}$$

Suppose

$$\begin{aligned}
\lim_{n,m\to\infty} \frac{m}{m+n} &= \omega \\
\Rightarrow \lim_{n,m\to\infty} \frac{n}{m+n} &= \lim_{n,m\to\infty}\left(1 - \frac{m}{m+n}\right) \\
&= 1 - \lim_{n,m\to\infty} \frac{m}{m+n} \\
&= 1 - \omega,
\end{aligned}$$

where $\omega \in (0,1)$. Therefore, as $n, m \to \infty$, we have, under f_{θ_1} and f_{θ_2}, respectively,

$$\begin{aligned}
\sqrt{\frac{mn}{m+n}}\left({}^{(1)}\hat{\theta}_\alpha - \theta_1\right) &\overset{\text{a}}{\sim} N\left(0, \omega J_\alpha^{-1}(\theta_1)\, V_\alpha(\theta_1)\, J_\alpha^{-1}(\theta_1)\right), \\
\sqrt{\frac{mn}{m+n}}\left({}^{(2)}\hat{\theta}_\alpha - \theta_2\right) &\overset{\text{a}}{\sim} N\left(0, (1-\omega) J_\alpha^{-1}(\theta_2)\, V_\alpha(\theta_2)\, J_\alpha^{-1}(\theta_2)\right).
\end{aligned}$$

Under H_0, as $n, m \to \infty$,

$$\begin{aligned}
&\sqrt{\frac{mn}{m+n}}\left({}^{(1)}\hat{\theta}_\alpha - \theta_1\right) - \sqrt{\frac{mn}{m+n}}\left({}^{(2)}\hat{\theta}_\alpha - \theta_2\right)\\
= \; & \sqrt{\frac{mn}{m+n}}\left(\left({}^{(1)}\hat{\theta}_\alpha - {}^{(2)}\hat{\theta}_\alpha\right) - (\theta_1 - \theta_2)\right)\\
\overset{\text{a}}{\sim} \; & N\left(0, J_\alpha^{-1}(\theta_1)\, V_\alpha(\theta_1)\, J_\alpha^{-1}(\theta_1)\right).
\end{aligned}$$

A second order Taylor series expansion of $Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)$ around $\theta_1 = \theta_2$ at $\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right)$ leads us to the following

$$\begin{aligned}
Q_{(\alpha,\tau,\gamma)}\left(f_{{}^{(1)}\hat{\theta}_\alpha}, f_{{}^{(2)}\hat{\theta}_\alpha}\right) &= Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\\
&+ \sum_{i=1}^{p} \frac{\partial}{\partial\theta_{1i}} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\Bigg|_{\theta_1=\theta_2}\left(\hat{\theta}_\alpha^{1i} - \theta_{1i}\right)\\
&+ \sum_{i=1}^{p} \frac{\partial}{\partial\theta_{2i}} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\Bigg|_{\theta_1=\theta_2}\left(\hat{\theta}_\alpha^{2i} - \theta_{2i}\right)\\
&+ \frac{1}{2}\sum_{i,j=1}^{p} \frac{\partial^2}{\partial\theta_{1i}\partial\theta_{1j}} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\Bigg|_{\theta_1=\theta_2}\left(\hat{\theta}_\alpha^{1i} - \theta_{1i}\right)\left(\hat{\theta}_\alpha^{1j} - \theta_{1j}\right)\\
&+ \frac{1}{2}\sum_{i,j=1}^{p} \frac{\partial^2}{\partial\theta_{2i}\partial\theta_{2j}} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\Bigg|_{\theta_1=\theta_2}\left(\hat{\theta}_\alpha^{2i} - \theta_{2i}\right)\left(\hat{\theta}_\alpha^{2j} - \theta_{2j}\right)\\
&+ \sum_{i,j=1}^{p} \frac{\partial^2}{\partial\theta_{1i}\partial\theta_{2j}} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\Bigg|_{\theta_1=\theta_2}\left(\hat{\theta}_\alpha^{1i} - \theta_{1i}\right)\left(\hat{\theta}_\alpha^{2j} - \theta_{2j}\right)\\
&+ o(||{}^{(1)}\hat{\theta}_\alpha - \theta_1||^2) + o(||{}^{(2)}\hat{\theta}_\alpha - \theta_2||^2).
\end{aligned}$$

We have,

$$\begin{aligned}
&\frac{\partial}{\partial\theta_1} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\\
= \; & \frac{1+\alpha}{\tau\bar{\tau}(\alpha-\tau)}\int\left[\tau f_{\theta_1}^{\alpha}\frac{\partial f_{\theta_1}}{\partial\theta_1} - \tau f_{\theta_1}^{\gamma}\frac{\partial f_{\theta_1}}{\partial\theta_1}\left\{\tau f_{\theta_1}^{1+\gamma} + \bar{\tau} f_{\theta_2}^{1+\gamma}\right\}^{\frac{\alpha-\gamma}{1+\gamma}}\right] d\mu,\\
&\frac{\partial}{\partial\theta_2} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\\
= \; & \frac{1+\alpha}{\tau\bar{\tau}(\alpha-\tau)}\int\left[\bar{\tau} f_{\theta_2}^{\alpha}\frac{\partial f_{\theta_2}}{\partial\theta_2} - \bar{\tau} f_{\theta_2}^{\gamma}\frac{\partial f_{\theta_2}}{\partial\theta_2}\left\{\tau f_{\theta_1}^{1+\gamma} + \bar{\tau} f_{\theta_2}^{1+\gamma}\right\}^{\frac{\alpha-\gamma}{1+\gamma}}\right] d\mu.
\end{aligned}$$

Here, $\bar{\tau} = 1 - \tau$. Moreover, at $\theta_1 = \theta_2$,

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_1^2} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right) &= \frac{\partial^2}{\partial\theta_2^2} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right) \\
&= (1+\alpha) \int f_{\theta_1}^{\alpha-1} \left(\frac{\partial f_{\theta_1}}{\partial\theta_1}\right)^2 d\mu, \\
\frac{\partial^2}{\partial\theta_1\partial\theta_2} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right) &= \frac{\partial^2}{\partial\theta_2\partial\theta_1} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right) \\
&= -(1+\alpha) \int f_{\theta_1}^{\alpha-1} \left(\frac{\partial f_{\theta_1}}{\partial\theta_1}\right)^2 d\mu \\
&= -\frac{\partial^2}{\partial\theta_1^2} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right).
\end{aligned}$$

Therefore, combining all these, we get,

$$\begin{aligned}
2Q_{(\alpha,\tau,\gamma)}\left(f_{{}^{(1)}\hat{\theta}_\alpha}, f_{{}^{(2)}\hat{\theta}_\alpha}\right) &= \left({}^{(1)}\hat{\theta}_\alpha - \theta_1\right)^T A_\alpha(\theta_1) \left({}^{(1)}\hat{\theta}_\alpha - \theta_1\right) \\
&- 2\left({}^{(1)}\hat{\theta}_\alpha - \theta_1\right)^T A_\alpha(\theta_1) \left({}^{(2)}\hat{\theta}_\alpha - \theta_2\right) \\
&+ \left({}^{(2)}\hat{\theta}_\alpha - \theta_2\right)^T A_\alpha(\theta_1) \left({}^{(2)}\hat{\theta}_\alpha - \theta_2\right) \\
&+ o(||{}^{(1)}\hat{\theta}_\alpha - \theta_1||^2) + o(||{}^{(2)}\hat{\theta}_\alpha - \theta_2||^2) \\
&= \left({}^{(1)}\hat{\theta}_\alpha -{}^{(2)} \hat{\theta}_\alpha\right)^T A_\alpha(\theta_1) \left({}^{(1)}\hat{\theta}_\alpha -{}^{(2)} \hat{\theta}_\alpha\right) \\
&+ o(||{}^{(1)}\hat{\theta}_\alpha - \theta_1||^2) + o(||{}^{(2)}\hat{\theta}_\alpha - \theta_2||^2).
\end{aligned}$$

Hence, under H_0, when $n, m \to \infty$,

$$2Q_{(\alpha,\tau,\gamma)}\left(f_{{}^{(1)}\hat{\theta}_\alpha}, f_{{}^{(2)}\hat{\theta}_\alpha}\right) \overset{\mathrm{D}}{=} \left({}^{(1)}\hat{\theta}_\alpha -{}^{(2)} \hat{\theta}_\alpha\right)^T A_\alpha(\theta_1) \left({}^{(1)}\hat{\theta}_\alpha -{}^{(2)} \hat{\theta}_\alpha\right).$$

Therefore, using the fact that $n \times o(||{}^{(1)}\hat{\theta}_\alpha - \theta_1||^2) = m \times o(||{}^{(2)}\hat{\theta}_\alpha - \theta_2||^2) = o_p(1)$, we can say,

$$S_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right) \overset{\mathrm{D}}{=} \sum_{i=1}^{r} \lambda_i Z_i^2, \tag{7.10}$$

(by Lemma 7.1) where λ_i's, Z_i's and $A_\alpha(\theta_1)$ are as defined in the theorem. Hence, the proof. □

Theorem 7.7. *Under the Lehmann and Basu et al. conditions (B1)-(B7), an approximation to the power function of the test statistic in Equation (7.9), at significance level β is given by*

$$\Pi_{m,n,\beta}^{(\alpha,\tau,\gamma)}(\theta_1,\theta_2) \quad = \quad 1-\Phi\left(\frac{\sqrt{\frac{mn}{m+n}}}{\sigma_{(\alpha,\tau,\gamma)}(\theta_1,\theta_2)}\left(\frac{S_\beta^{(\alpha,\tau,\gamma)}}{2}\frac{m+n}{mn}-Q_{(\alpha,\tau,\gamma)}(f_{\theta_1},f_{\theta_2})\right)\right).$$

Moreover, the probability of rejecting the null hypothesis $H_0:\theta_1=\theta_2$, through the rejection rule $S_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right) > S_\beta^{(\alpha,\tau,\gamma)}$ with pre-fixed significance level β, tends to 1 as $n,m\to\infty$. Hence, this test is also consistent in the Fraser's sense.

Proof.

$$\begin{aligned} Q_{(\alpha,\tau,\gamma)}\left(f_{^{(1)}\hat{\theta}_\alpha}, f_{^{(2)}\hat{\theta}_\alpha}\right) &= Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1},f_{\theta_2}\right)+\sum_{i=1}^{p}\frac{\partial}{\partial\theta_{1i}}Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1},f_{\theta_2}\right)\left(\hat{\theta}^{1i}-\theta_{1i}\right) \\ &+ \sum_{i=1}^{p}\frac{\partial}{\partial\theta_{2i}}Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1},f_{\theta_2}\right)\left(\hat{\theta}^{2i}-\theta_{2i}\right)+o(||^{(1)}\hat{\theta}_\alpha-\theta_1||) \\ &+ o(||^{(2)}\hat{\theta}_\alpha-\theta_2||). \end{aligned}$$

Then,

$$\begin{aligned} Q_{(\alpha,\tau,\gamma)}\left(f_{^{(1)}\hat{\theta}_\alpha}, f_{^{(2)}\hat{\theta}_\alpha}\right)-Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1},f_{\theta_2}\right) &= G_\alpha^T\left({}^{(1)}\hat{\theta}_\alpha-\theta_1\right)+H_\alpha^T\left({}^{(2)}\hat{\theta}_\alpha-\theta_2\right) \\ &+ o(||^{(1)}\hat{\theta}_\alpha-\theta_1||)+o(||^{(2)}\hat{\theta}_\alpha-\theta_2||), \end{aligned}$$

where,

$$\begin{aligned} G_\alpha &= (g_1,\dots,g_p)^T=\left(\left(\frac{\partial}{\partial\theta_{1i}}Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1},f_{\theta_2}\right)\right)\right)_{i=1,2,\dots,p}, \\ H_\alpha &= (h_1,\dots,h_p)^T=\left(\left(\frac{\partial}{\partial\theta_{2i}}Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1},f_{\theta_2}\right)\right)\right)_{i=1,2,\dots,p}. \end{aligned}$$

Evidently, when $\theta_1\neq\theta_2$,

$$\begin{aligned} \sqrt{n}\,G_\alpha^T\left({}^{(1)}\hat{\theta}_\alpha-\theta_1\right) &\overset{\text{a}}{\sim} N\left(0, G_\alpha^T J_\alpha^{-1}(\theta_1)\,V_\alpha(\theta_1)\,J_\alpha^{-1}(\theta_1)G_\alpha\right), \\ \sqrt{n}\,H_\alpha^T\left({}^{(2)}\hat{\theta}_\alpha-\theta_2\right) &\overset{\text{a}}{\sim} N\left(0, H_\alpha^T J_\alpha^{-1}(\theta_2)\,V_\alpha(\theta_2)\,J_\alpha^{-1}(\theta_2)H_\alpha\right), \end{aligned}$$

and in that case, the random variable

$$\begin{aligned}
& \sqrt{\frac{mn}{m+n}}\left(Q_{(\alpha,\tau,\gamma)}\left(f_{{}^{(1)}\hat{\theta}_\alpha}, f_{{}^{(2)}\hat{\theta}_\alpha}\right) - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\right) \\
= & \sqrt{\frac{m}{m+n}}\sqrt{n}\, G_\alpha^T\left({}^{(1)}\hat{\theta}_\alpha - \theta_1\right) + \sqrt{\frac{n}{m+n}}\sqrt{m}\, H_\alpha^T\left({}^{(2)}\hat{\theta}_\alpha - \theta_2\right) \\
+ & \sqrt{\frac{mn}{m+n}} \times o(||{}^{(1)}\hat{\theta}_\alpha - \theta_1||) + \sqrt{\frac{mn}{m+n}} \times o(||{}^{(2)}\hat{\theta}_\alpha - \theta_2||) \\
\overset{a}{\sim} & N\left(0, \omega G_\alpha^T J_\alpha^{-1}(\theta_1)\, V_\alpha(\theta_1)\, J_\alpha^{-1}(\theta_1) G_\alpha + (1-\omega) H_\alpha^T J_\alpha^{-1}(\theta_2)\, V_\alpha(\theta_2)\, J_\alpha^{-1}(\theta_2) H_\alpha\right).
\end{aligned}$$

Therefore, for any fixed significance level β, if we start with the rejection rule, $S_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right) > S_\beta^{(\alpha,\tau,\gamma)}$, we then conclude that

$$\begin{aligned}
& \Pi_{m,n,\beta}^{(\alpha,\tau,\gamma)} = P_{H_1}\left(S_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right) > S_\beta^{(\alpha,\tau,\gamma)}\right) \\
= & P_{H_1}\left(Q_{(\alpha,\tau,\gamma)}\left(f_{{}^{(1)}\hat{\theta}_\alpha}, f_{{}^{(2)}\hat{\theta}_\alpha}\right) > \frac{m+n}{mn}\frac{S_\beta^{(\alpha,\tau,\gamma)}}{2}\right) \\
= & P_{H_1}\left[\frac{\left(Q_{(\alpha,\tau,\gamma)}\left(f_{{}^{(1)}\hat{\theta}_\alpha}, f_{{}^{(2)}\hat{\theta}_\alpha}\right) - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\right)\sqrt{\frac{mn}{m+n}}}{\sigma_{(\alpha,\tau,\gamma)}(\theta_1,\theta_2)}\right. \\
& \qquad \left. > \frac{\left(\frac{m+n}{mn}\frac{S_\beta^{(\alpha,\tau,\gamma)}}{2} - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\right)\sqrt{\frac{mn}{m+n}}}{\sigma_{(\alpha,\tau,\gamma)}(\theta_1,\theta_2)}\right] \\
\to & 1 - \Phi\left[\frac{\left(\frac{m+n}{mn}\frac{S_\beta^{(\alpha,\tau,\gamma)}}{2} - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\right)\sqrt{\frac{mn}{m+n}}}{\sigma_{(\alpha,\tau,\gamma)}(\theta_1,\theta_2)}\right],
\end{aligned}$$

where $\sigma_{(\alpha,\tau,\gamma)}(\theta_1,\theta_2)$ is the square root of $\omega G_\alpha^T J_\alpha^{-1}(\theta_1) V_\alpha(\theta_1) J_\alpha^{-1}(\theta_1) G_\alpha +$ $(1-\omega) H_\alpha^T J_\alpha^{-1}(\theta_2)\, V_\alpha(\theta_2)\, J_\alpha^{-1}(\theta_2) H_\alpha$. Evidently, this test is consistent in the Fraser's sense. □

7.3.3 Robustness Properties of the GSDT (Simple NULL Hypothesis)

Now, we cultivate the robustness of our proposed statistic. Evidently, it depends on the robustness of both the GSD family and the DPD family. To study this robustness, we will consider the contaminated distribution $G_\epsilon = (1-\epsilon)G + \epsilon\Lambda_y$ where Λ_y is the distribution degenerate at y and ϵ is the proportion of contamination. The true distribution G is still assumed to belong to the model family. Under its consideration, the following tools and certain theorems will be established.

7.3.3.1 Influence Function of the Test

Following to Hampel's IF of the test, we first define the GSDT functional as

$$T^{(1)}_{(\alpha,\tau,\gamma)}(G) \;=\; Q_{(\alpha,\tau,\gamma)}\left(f_{T_\alpha(G)}, f_{\theta_0}\right),$$

where, $T_\alpha(G)$ is the minimum DPD functional. The IF will be given by

$$\begin{aligned} IF(y, T^{(1)}_{(\alpha,\tau,\gamma)}, G) &= \frac{\partial}{\partial\epsilon} T^{(1)}_{(\alpha,\tau,\gamma)}(G_\epsilon)\big|_{\epsilon=0} \\ &= M_{(\alpha,\tau,\gamma)}(T_\alpha(G))^T IF(y, T_\alpha, G), \end{aligned}$$

where, $IF(y, T_\alpha, G)$ is the IF of $T_\alpha(G)$ and $M_{(\alpha,\tau,\gamma)}(T_\alpha(G)) = \frac{\partial}{\partial\theta} Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)\big|_{\theta=T_\alpha(G)}$. Since at the null hypothesis $G = F_{\theta_0}$, $T_\alpha(G) = \theta_0$ (by Fisher consistency property) and hence, $M_{(\alpha,\tau,\gamma)}(\theta_0) = 0$, the Hampel's first order IF of the test statistic is 0 under null hypothesis. For further analysis, we need to calculate the second order IF which is given in the following expression

$$\begin{aligned} IF_2\left(y, T^{(1)}_{(\alpha,\tau,\gamma)}, G\right) &= \frac{\partial^2}{\partial\epsilon^2} T^{(1)}_{(\alpha,\tau,\gamma)}(G_\epsilon)\big|_{\epsilon=0} \\ = \;& M_{(\alpha,\tau,\gamma)}(T_\alpha(G))^T \frac{\partial^2}{\partial\epsilon^2} T_\alpha(G_\epsilon)\big|_{\epsilon=0} \\ + \;& IF(y, T_\alpha, G)^T \nabla^2 Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)\big|_{\theta=T_\alpha(G)} IF(y, T_\alpha, G). \end{aligned}$$

Evidently at the null hypothesis, it will reduce to

$$IF_2\left(y, T^{(1)}_{(\alpha,\tau,\gamma)}, F_{\theta_0}\right) \;=\; IF\left(y, T_\alpha, F_{\theta_0}\right)^T A_\alpha(\theta_0)\, IF(y, T_\alpha, F_{\theta_0}).$$

Therefore, at the null, this test statistic possesses the same robustness property of the test statistic based on the S-divergence, introduced by Ghosh et al. (2017). At the null, the second order IF of our proposed test depends only on α. Hence, the IF will be bounded if and only if the IF of $T_\alpha(G)$ is bounded. Evidently, it will be bounded $\forall\, \alpha > 0$. An extremely non-robust case corresponds to $\alpha = 0$, since $T_{\alpha=0}(G)$ coincides with the MLE and it indeed leads to a non-robust test procedure as well.

7.3.3.2 Level and Power Influence Function

Because our proposed tests are consistent, we have also studied their behaviour under contiguous alternatives. Now, to study the robustness of the test, we are to consider contamination over these contiguous alternatives.

Due to the consistency property of the test, to set up the alternative hypotheses, we have taken $H_{1,n} : \theta = \theta_n$, where, $\theta_n = \theta_0 + \frac{\Delta}{\sqrt{n}}$, with Δ having the same dimension as θ. Now, we are to consider the contaminations such that their effects vanish as θ_n tends to θ_0 at the same rate to avoid confusion between the null and alternative hypotheses.

For the level, we consider the distribution

$$F^L_{n,\epsilon,y} \;=\; \left(1 - \frac{\epsilon}{\sqrt{n}}\right) F_{\theta_0} + \frac{\epsilon}{\sqrt{n}}\Lambda_y.$$

and for power, we consider

$$F^P_{n,\epsilon,y} \;=\; \left(1-\frac{\epsilon}{\sqrt{n}}\right)F_{\theta_n}+\frac{\epsilon}{\sqrt{n}}\Lambda_y.$$

Then, the level influence function (LIF) is given by

$$LIF\left(y, T^{(1)}_{(\alpha,\tau,\gamma)}, F_{\theta_0}\right) \;=\; \lim_{n\to\infty}\frac{\partial}{\partial\epsilon}P_{F^L_{n,\epsilon,y}}\left(T_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha,\theta_0\right) > t^{(\alpha,\tau,\gamma)}_\beta\right)\Bigg|_{\epsilon=0},$$

and the power influence function (PIF) is given by

$$PIF\left(y, T^{(1)}_{(\alpha,\tau,\gamma)}, F_{\theta_0}\right) \;=\; \lim_{n\to\infty}\frac{\partial}{\partial\epsilon}P_{F^P_{n,\epsilon,y}}\left(T_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha,\theta_0\right) > t^{(\alpha,\tau,\gamma)}_\beta\right)\Bigg|_{\epsilon=0}.$$

Theorem 7.8. *Under the Lehmann and Basu et al. conditions (B1)-(B7), the following results hold for any $\Delta \in \mathbb{R}^P$ and $\epsilon \geq 0$:*

(i) The asymptotic distribution of the GSDT under $F^P_{n,\epsilon,y}$ is the same as that of $W^T A_\alpha(\theta_0)W$, where W follows p-variate normal distribution with mean $\tilde{\Delta} = \Delta + \epsilon IF(y, T_\alpha, F_{\theta_0})$ and variance-covariance matrix $\Sigma_\alpha(\theta_0)$. Equivalently, this distribution is the same as that of $\sum_{i=1}^r \lambda_i\chi^2_{1,\delta_i}$ where, $\lambda_1,\ldots,\lambda_r$ are the r non-zero eigenvalues of $A_\alpha(\theta_0)\Sigma_\alpha(\theta_0)$ and $\chi^2_{1,\delta_1},\ldots,\chi^2_{1,\delta_r}$ are r independent non-central χ^2_1 variables with non-centrality parameter $\delta_i = \mu_i^2$ of $\mu = (\mu_1,\ldots,\mu_r)^T = P_\alpha(\theta_0)\Sigma_\alpha^{-1/2}(\theta_0)\tilde{\Delta}$, $P_\alpha(\theta_0)$ being the matrix of normalized eigenvectors of $A_\alpha(\theta_0)\Sigma_\alpha(\theta_0)$.

(ii) The asymptotic power will be

$$\begin{aligned} power(\Delta,\epsilon) &= \lim_{n\to\infty}P_{F^P_{n,\epsilon,y}}\left(T_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha,\theta_0\right) > t^{(\alpha,\tau,\gamma)}_\beta\right) \\ &= \sum_{\nu=0}^{\infty}C_\nu(\theta_0,\tilde{\Delta})P\left(\chi^2_{r+2\nu} > \frac{t^{(\alpha,\tau,\gamma)}_\beta}{\lambda_{(1)}}\right), \quad (7.11)\end{aligned}$$

where, $\lambda_{(1)} = min\{\lambda_1, \ldots, \lambda_r\}$ *and*

$$C_\nu(\theta_0, \tilde{\Delta}) = \frac{1}{\nu!}\left(\prod_{j=1}^{r} \frac{\lambda_{(1)}}{\lambda_j}\right)^{1/2} e^{-\delta/2} E(\hat{Q}^\nu)$$

with $\delta = \mu^T \mu$ *and*

$$\hat{Q} = \frac{1}{2}\sum_{j=1}^{r}\left[\left(1 - \frac{\lambda_{(1)}}{\lambda_j}\right)^{1/2} Z_j + \mu_j \left(\frac{\lambda_{(1)}}{\lambda_j}\right)^{1/2}\right]^2$$

for r *independent standard normal variables* $Z_1, Z_2, \ldots, Z_r$.

Proof. (i) Let, θ_n^* be the minimum DPD functional under the distribution $F_{n,\epsilon,y}^P$. A second order Taylor series expansion of $Q_{(\alpha,\tau,\gamma)}(f_\theta, f_{\theta_0})$ around $\theta = \theta_n^*$ at $\theta = \hat{\theta}_\alpha$ gives

$$\begin{aligned} Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\theta_0}\right) &= Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_n^*}, f_{\theta_0}\right) + \left(\hat{\theta}_\alpha - \theta_n^*\right)\frac{\partial}{\partial\theta} Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)\Big|_{\theta=\theta_n^*} \\ &+ \frac{1}{2}\left(\hat{\theta}_\alpha - \theta_n^*\right)^T \nabla^2 Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)\Big|_{\theta=\theta_n^*}\left(\hat{\theta}_\alpha - \theta_n^*\right) \\ &+ o\left(||\hat{\theta}_\alpha - \theta_n^*||^2\right) \\ &= Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_n^*}, f_{\theta_0}\right) + \left(\hat{\theta}_\alpha - \theta_n^*\right) M_{(\alpha,\tau,\gamma)}\left(\theta_n^*\right) \\ &+ \frac{1}{2}\left(\hat{\theta}_\alpha - \theta_n^*\right)^T A_\alpha\left(\theta_n^*\right)\left(\hat{\theta}_\alpha - \theta_n^*\right) + o\left(||\hat{\theta}_\alpha - \theta_n^*||^2\right). \end{aligned} \tag{7.12}$$

Under $F_{n,\epsilon,y}^P$, θ_n^* is the best fitting parameter for tuning parameter α, and by the consistency of $\hat{\theta}_\alpha$ we have

$$\sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right) \overset{\text{a}}{\sim} N\left(0, \Sigma_\alpha(\theta_0)\right).$$

Again, as $n \to \infty$, $\theta_n^* \to \theta_0$ and hence by continuity, $A_\alpha\left(\theta_n^*\right) \to A_\alpha\left(\theta_0\right)$ element-wise as $n \to \infty$.

From Theorem 3, Ghosh et al. (2016), we get

$$(\theta_n^* - \theta_0) = \frac{\Delta}{\sqrt{n}} + \frac{\epsilon}{\sqrt{n}} IF(y, T_\alpha, F_{\theta_0}) + o(\frac{1}{\sqrt{n}}).$$

Then, by a first order Taylor series expansion of $M_{(\alpha,\tau,\gamma)}(\theta)$ around $\theta = \theta_0$ at $\theta = \theta_n^*$ we get, using Ghosh et al. (2015, Theorem 3.1)

$$\begin{aligned} M_{(\alpha,\tau,\gamma)}(\theta_n^*) - M_{(\alpha,\tau,\gamma)}(\theta_0) &= A_\alpha(\theta_0)\frac{\Delta}{\sqrt{n}} + A_\alpha(\theta_0)\frac{\epsilon}{\sqrt{n}} IF(y, T_\alpha, F_{\theta_0}) \\ &+ o\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \tag{7.13}$$

However, $M_{(\alpha,\tau,\gamma)}(\theta_0) = 0$ and $\tilde{\Delta} = \Delta + \epsilon IF(y, T_\alpha, F_{\theta_0})$. Hence, Equation (7.13) becomes

$$\begin{aligned} M_{(\alpha,\tau,\gamma)}(\theta_n^*) &= A_\alpha(\theta_0)\frac{\tilde{\Delta}}{\sqrt{n}} + o\left(n^{-1/2}\right) \\ \Rightarrow \sqrt{n} M_{(\alpha,\tau,\gamma)}(\theta_n^*) &= A_\alpha(\theta_0)\tilde{\Delta} + o(1). \end{aligned} \tag{7.14}$$

Again, considering the second order Taylor series expansion of $S_{(\alpha,\tau,\gamma)}(f_\theta, f_{\theta_0})$ around $\theta = \theta_0$ at $\theta = \theta_n^*$, we get,

$$\begin{aligned} & Q_{(\alpha,\tau,\gamma)}(f_{\theta_n^*}, f_{\theta_0}) - Q_{(\alpha,\tau,\gamma)}(f_{\theta_0}, f_{\theta_0}) \\ = & (\theta_n^* - \theta_0)^T \frac{\partial}{\partial\theta} Q_{(\alpha,\tau,\gamma)}(f_\theta, f_{\theta_0}) \Big|_{\theta=\theta_0} \\ + & \frac{1}{2}(\theta_n^* - \theta_0)^T \nabla^2 Q_{(\alpha,\tau,\gamma)}(f_\theta, f_{\theta_0}) \Big|_{\theta=\theta_0} (\theta_n^* - \theta_0) + o\left(\frac{1}{n}\right) \\ = & (\theta_n^* - \theta_0)^T M_{(\alpha,\tau,\gamma)}(\theta_0) + \frac{1}{2}(\theta_n^* - \theta_0)^T A_\alpha(\theta_0)(\theta_n^* - \theta_0) + o\left(\frac{1}{n}\right) \\ = & \frac{\Delta^T}{\sqrt{n}} M_{(\alpha,\tau,\gamma)}(\theta_0) + \frac{\epsilon}{\sqrt{n}} IF(y, T_\alpha, F_{\theta_0})^T M_{(\alpha,\tau,\gamma)}(\theta_0) + o\left(\frac{1}{\sqrt{n}}\right) M_{(\alpha,\tau,\gamma)}(\theta_0) \\ + & \frac{1}{2n}\Delta^T A_\alpha(\theta_0)\Delta + \frac{\epsilon}{n}\Delta^T A_\alpha(\theta_0) IF(y, T_\alpha, F_{\theta_0}) \\ + & \frac{\epsilon^2}{2n} IF(y, T_\alpha, F_{\theta_0})^T A_\alpha(\theta_0) IF(y, T_\alpha, F_{\theta_0}) + o\left(\frac{1}{n}\right). \end{aligned} \tag{7.15}$$

But evidently $Q_{(\alpha,\tau,\gamma)}(f_{\theta_0}, f_{\theta_0}) = 0$, $M_{(\alpha,\tau,\gamma)}(\theta_0) = 0$ and

$IF(y, T_\alpha, F_{\theta_0})^T M_{(\alpha,\tau,\gamma)}(\theta_0) = 0$. Therefore, Equation (7.15) can be rewritten as

$$\begin{aligned}
& 2n\, Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_n^*}, f_{\theta_0}\right) \\
= \;& \Delta^T A_\alpha(\theta_0)\Delta + 2\epsilon\Delta^T A_\alpha(\theta_0) IF(y, T_\alpha, F_{\theta_0}) \\
+ \;& \epsilon^2 IF(y, T_\alpha, F_{\theta_0})^T A_\alpha(\theta_0) IF(y, T_\alpha, F_{\theta_0}) + o(1) \\
= \;& (\Delta + \epsilon IF(y, T_\alpha, F_{\theta_0}))^T A_\alpha(\theta_0) \left(\Delta + \epsilon IF(y, T_\alpha, F_{\theta_0})\right) + o(1) \\
= \;& \tilde{\Delta}^T A_\alpha(\theta_0)\tilde{\Delta} + o(1).
\end{aligned}$$

Now, $n \times o\left(||\hat{\theta}_\alpha - \theta_n^*||^2\right) = o_p(1)$, therefore combining all these, we get

$$\begin{aligned}
& 2n\, Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\theta_0}\right) \\
= \;& \tilde{\Delta}^T A_\alpha(\theta_0)\tilde{\Delta} + 2\sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right)^T A_\alpha(\theta_0)\tilde{\Delta} \\
+ \;& \sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right)^T A_\alpha(\theta_n^*)\sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right) + o_p(1) + o(1) \\
= \;& \left[\tilde{\Delta} + \sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right)\right]^T A_\alpha(\theta_0)\left[\tilde{\Delta} + \sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right)\right] + o_p(1) + o(1).
\end{aligned}$$

Thus, under $F_{n,\epsilon,y}^P$, the asymptotic distribution of the GSDT statistic is the same as that of $\left(\tilde{\Delta} + W_0\right)^T A_\alpha(\theta_0)\left(\tilde{\Delta} + W_0\right)$, where $W_0 = \sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right) \overset{\text{a}}{\sim} N(0, \Sigma_\alpha(\theta_0))$. Hence, the proof of the first part of (i) follows using $W = \tilde{\Delta} + W_0$. By spectral decomposition, we have

$$\Sigma_\alpha^{1/2}(\theta_0) A_\alpha(\theta_0) \Sigma_\alpha^{1/2}(\theta_0) \;=\; P_\alpha(\theta_0)\Gamma_r(\theta_0)P_\alpha(\theta_0),$$

where $\Gamma_r(\theta_0) = diag(\lambda_1, \lambda_2, \ldots, \lambda_r, 0, \ldots, 0)$, λ_i's being the eigenvalues of $A_\alpha(\theta_0)\Sigma_\alpha(\theta_0)$ and $P_\alpha(\theta_0)$ is as defined in the statement of the theorem. Now, considering $W = W_0 + \tilde{\Delta}$, we can rewrite,

$$\begin{aligned} W^T A_\alpha(\theta_0) W &= W^T \Sigma_\alpha^{-1/2}(\theta_0) \left[\Sigma_\alpha^{1/2}(\theta_0) A_\alpha(\theta_0) \Sigma_\alpha^{1/2}(\theta_0)\right] \Sigma_\alpha^{-1/2}(\theta_0) W \\ &= W^T \Sigma_\alpha^{-1/2}(\theta_0) P_\alpha(\theta_0) \Gamma_r(\theta_0) P_\alpha(\theta_0) \Sigma_\alpha^{-1/2}(\theta_0) W \\ &= (W^* + \mu)^T \Gamma_r(\theta_0) (W^* + \mu), \end{aligned}$$

where, $W^* = P_\alpha(\theta_0)\Sigma_\alpha^{-1/2}(\theta_0)W_0$ and evidently, $W^* \sim N(0, I_r)$ which follows from the definition of $P_\alpha(\theta_0)$. Thus,

$$W^T A_\alpha(\theta_0) W = \sum_{i=1}^{r} \lambda_i (W_i^* + \mu_i)^2. \tag{7.16}$$

It is immediately seen that the asymptotic distribution of the random variable in Equation (7.16) is the same as that of $\sum_{i=1}^r \lambda_i \chi^2_{1,\delta_i}$ where δ = vector of non-centrality parameters $= (\delta_1, \delta_2, \ldots, \delta_r)^T$ $= (\mu_1^2, \mu_2^2, \ldots, \mu_r^2)^T$. Hence, the proof of the second part of (i).

(ii) According to Kotz et al. (1967b), the distribution function $(F_n(\alpha;\xi;y))$ of the linear combination of non-central χ^2 random variables of the form $\sum_{i=1}^n \alpha_i (Z_i + \xi_i)^2$, where $Z_i \overset{i.i.d.}{\sim} N(0,1)$, can be represented by the series expansion of the distribution function $(G(n;y))$ of central χ^2_n random variables as follows

$$F_n(\alpha;\xi;y) = \sum_{k=0}^{\infty} a_k^C G\left(n + 2k; \frac{y}{\beta}\right) \tag{7.17}$$

for some $\beta > 0$ and $a_k^C = Ae^{-\frac{\xi^T\xi}{2}} \frac{E(\hat{Q}^k)}{k!}$. Here, $\hat{Q}$ is defined as

$$\hat{Q}(Z) = \frac{1}{2}\sum_{j=1}^{n} \left(\gamma_j^{\frac{1}{2}} Z_j + \xi_j (1-\gamma_j)^{\frac{1}{2}}\right)^2,$$

where,

$$\gamma_j = 1 - \frac{\beta}{\alpha_j} \quad , \quad A = \prod_{j=1}^{n} \left(\frac{\beta}{\alpha_j}\right)^{\frac{1}{2}},$$

and Z_i's are i.i.d. $N(0,1)$ random variables. Several authors suggested several choices for β, but the most popular one is Ruben's suggestion, that is, $\beta = \min\{\lambda_i\} = \lambda_{(1)}$. Now, considering $n = r$, $\alpha_i = \lambda_i \forall i$, $\xi_j = \mu_j \forall j$, we get

$$\begin{aligned} a_k^C &= \prod_{i=1}^{r} \left(\frac{\lambda_{(1)}}{\lambda_i}\right)^{\frac{1}{2}} e^{-\frac{\mu^T\mu}{2}} \frac{E(\hat{Q}^k)}{k!} \\ \text{and} \quad \hat{Q}(Z) &= \frac{1}{2}\sum_{j=1}^{r} \left(\left(1 - \frac{\lambda_{(1)}}{\lambda_j}\right)^{\frac{1}{2}} Z_j + \mu_j \left(\frac{\lambda_{(1)}}{\lambda_j}\right)^{\frac{1}{2}} \right)^2 . \end{aligned}$$

Using all these substitutions, we will get our desired result. Here, $\{a_k^C\}_{k=1,\ldots,n}$ will be replaced by $\{C_\nu(\theta_0, \tilde{\Delta})\}_{\nu=1,\ldots,r}$.

□

7.3.4 Some Observations

(i) If $\Delta = \epsilon = 0$, then $F_{n,\epsilon,y}^P$ coincides with the null distribution. In that case, $\tilde{\Delta} = \Delta + \epsilon IF(y, T_\alpha, F_{\theta_0}) = 0$ and hence, $\mu = 0$. It implies that the asymptotic distribution of the GSDT statistic has the same distribution as the random variable $Z = \sum_{i=1}^{r} \lambda_i \chi_1^2$, a linear combination of independent central χ_1^2 distributions. Evidently, it coincides with the null distribution.

(ii) If $\epsilon = 0$, then $\tilde{\Delta} = \Delta$, as there is no contamination. In this case, the asymptotic distribution of the GSDT under the contiguous alternative $H_{1,n} : \theta = \theta_n$, where, $\theta_n = \theta_0 + \frac{\Delta}{\sqrt{n}}$ is as given in part (i) of this theorem, but $\tilde{\Delta}$ is replaced by Δ. In such a case, the asymptotic power will be given by the same expression, with $\tilde{\Delta}$ being replaced by Δ in Equation (7.11).

(iii) If $\Delta = 0$, we would then get the asymptotic (null) distribution under $\theta = \theta_0$ through part (i) of Theorem 7.8 with $\tilde{\Delta} =$

$\epsilon IF(y, T_\alpha, F_{\theta_0})$. In that case, from part (ii) of the above-mentioned theorem, we can express the asymptotic level using the same expression, with the term $C_\nu\left(\theta_0, \tilde{\Delta}\right)$ being replaced by $C_\nu\left(\theta_0, \epsilon IF\left(y, T_\alpha, F_{\theta_0}\right)\right)$ in Equation (7.11).

Using part (ii) of Thoerem 7.8, the LIF and PIF can now be derived.

Theorem 7.9. *Under the Lehmann and Basu et al. conditions (B1)-(B7), provided $IF(y, T_\alpha, F_{\theta_0})$ of MDPD functional is bounded, the PIF and the LIF are given by*

$$\begin{aligned} & PIF\left(y, T^{(1)}_{(\alpha,\tau,\gamma)}, F_{\theta_0}\right) \\ = \; & IF(y, T_\alpha, F_{\theta_0})^T \left(\sum_{\nu=0}^{\infty} \left[\frac{\partial}{\partial t} C_\nu(\theta_0, t)\Big|_{t=\Delta}\right] P\left(\chi^2_{r+2\nu} > \frac{t_\beta^{(\alpha,\tau,\gamma)}}{\lambda_{(1)}}\right)\right), \end{aligned}$$

and

$$\begin{aligned} & LIF\left(y, T^{(1)}_{(\alpha,\tau,\gamma)}, F_{\theta_0}\right) \\ = \; & IF(y, T_\alpha, F_{\theta_0})^T \left(\sum_{\nu=0}^{\infty} \left[\frac{\partial}{\partial t} C_\nu(\theta_0, t)\Big|_{t=0}\right] P\left(\chi^2_{r+2\nu} > \frac{t_\beta^{(\alpha,\tau,\gamma)}}{\lambda_{(1)}}\right)\right). \end{aligned}$$

Proof.

$$\begin{aligned} PIF\left(y, T^{(1)}_{(\alpha,\tau,\gamma)}, F_{\theta_0}\right) &= \frac{\partial}{\partial \epsilon} power(\Delta, \epsilon)\Big|_{\epsilon=0} \\ &= \sum_{\nu=0}^{\infty} \frac{\partial}{\partial \epsilon} C_\nu(\theta_0, \tilde{\Delta})\Big|_{\epsilon=0} P\left(\chi^2_{r+2\nu} > \frac{t_\beta^{(\alpha,\tau,\gamma)}}{\lambda_{(1)}}\right). \end{aligned} \tag{7.18}$$

Differentiating the Taylor series expansion given in Equation (12) of Ghosh et al. (2015) with respect to ϵ and evaluating it at $\epsilon = 0$, we will get the expression

$$\frac{\partial}{\partial \epsilon} C_\nu(\theta_0, \tilde{\Delta})\bigg|_{\epsilon=0} = IF(y, T_\alpha, F_{\theta_0})^T \left[\frac{\partial}{\partial t} C_\nu(\theta_0, t)\bigg|_{t=\Delta} \right],$$

provided $IF(y, T_\alpha, F_{\theta_0})$ is bounded.

Therefore, putting the above-mentioned expression in Equation (7.18), we get the desired form of the PIF as mentioned in the theorem.

To derive the LIF, we first note that if we take $\Delta = 0$, the PIF becomes the LIF. Hence, the reduced form of the PIF, which is considered to be the LIF, is totally the same with the expression stated in the theorem above, with the partial derivative being evaluated at $t = 0$ instead of $t = \Delta$. □

7.4 Testing Parametric Hypothesis using GSD (Composite Null Hypotheses)

In case of testing with composite hypotheses, the null parameter space generally consists of some pre-defined restrictions and under these restrictions, the parameters are needed to be estimated to perform the tests. The likelihood ratio test is the default classical test in any testing scenario and it uses the restricted MLE for conducting the test. But due to its demonstrated non-robustness in the presence of outliers, we want some robust alternatives with good asymptotic properties.

We have already developed a robust test using the GSD family for simple null hypotheses; now we are going to extend this for composite null hypotheses. Consider a random sample $X_1, X_2, \ldots, X_n$ from true density g and a parametric family of densities $\mathcal{E} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$

to model g. We are going to test the following set of hypotheses

$$H_0 : \theta \in \Theta_0, \text{ v}s\ H_1 : \theta \notin \Theta_0, \tag{7.19}$$

where, the restricted parameter space Θ_0 is defined by a set of $r < p$ restrictions $h(\theta) = 0$, such that, the $p \times r$ matrix $H(\theta) = \frac{\partial h(\theta)}{\partial \theta}$ exists with rank r and is a continuous function of θ. Here, the estimator can be obtained by minimizing $Q_{(\alpha,\tau,\gamma)}(\hat{g}, f_\theta)$ over Θ subject to the conditions $h(\theta) = 0$ or, equivalently, minimizing the same over Θ_0, where $\hat{g}$ is some suitable non-parametric estimate of g. The restricted minimum GSD functional is defined by the relation

$$Q_{(\alpha,\tau,\gamma)}(g, f_{\tilde{T}_{(\alpha,\tau,\gamma)}(G)}) = \min_{\theta \in \Theta_0} Q_{(\alpha,\tau,\gamma)}(g, f_\theta) = \min_{h(\theta)=0} Q_{(\alpha,\tau,\gamma)}(g, f_\theta), \tag{7.20}$$

provided the minimum exists. Therefore, using the method of Lagrange's multiplier, this constrained minimization problem can be solved through the estimating equation given below.

$$\begin{cases} \displaystyle\int K(\delta(x)) f_\theta^{1+\alpha}(x) u_\theta(x) dx + H(\theta)\lambda_n = 0, \\ h(\theta) = 0, \end{cases} \tag{7.21}$$

where,

$$K(\delta) = \frac{1}{\tau(\alpha - \gamma)} \left[\left(\tau(\delta + 1)^{1+\gamma} + (1 - \tau) \right)^{\frac{\alpha-\gamma}{1+\gamma}} - 1 \right]$$

with,

$$\delta(x) = \delta_n(x) = \frac{\hat{g}(x)}{f_\theta(x)} - 1,$$

and λ_n is the vector of Lagrange's multipliers.

7.4.1 Influence Function of Restricted MGSDE

For the unrestricted case, Ghosh and Basu (2018) have already shown that, whenever $g = f_{\theta_0}$ and the simple null hypothesis is true), the IF of MGSDE coincides with the IF of MDPDE. Considering this asymptotic equivalence, we have already used $\hat{\theta}_\alpha$ instead of $\hat{\theta}_{(\alpha,\tau,\gamma)}$ in case of constructing the test statistic under the simple null. Therefore, under the restrictions $h(\theta) = 0$, if we can prove the similar kind of statement, then, replacing the restricted estimator RMGSDE $\tilde{\theta}_{(\alpha,\tau,\gamma)}$ by the restricted estimator RMDPDE $\tilde{\theta}_\alpha$ in constructing the GSDT, will be more credible.

If we consider any statistical divergence in a more general form, i.e.,

$$\rho(g, f_\theta) = \int D(g, f_\theta) d\mu, \tag{7.22}$$

with $\tilde{\theta}^g = \tilde{T}_\rho(G)$ and $\tilde{\theta}_\epsilon = \tilde{T}_\rho(G_\epsilon)$ being the best fitting parameters under pure data distribution G and contaminated data distribution G_ϵ (where, $G_\epsilon(x) = (1-\epsilon)G(x) + \epsilon\Lambda_y(x)$), respectively, then from Theorem 3.1 of Ghosh (2015b), we will use here the general form of the IF of the restricted minimum divergence estimator to serve our purpose, which we state in the following lemma.

Lemma 7.10. *Assuming rank*$(H(\tilde{\theta}^g)) = r$*, the IF of* $\tilde{T}_\rho(G)$ *will be of the form,*

$$\begin{aligned} IF(y, \tilde{T}_\rho, G) &= [N_0(\tilde{\theta}^g)^T N_0(\tilde{\theta}^g) + H(\tilde{\theta}^g)H(\tilde{\theta}^g)^T]^{-1} N_0(\tilde{\theta}^g)^T \\ &\quad [\xi_0(\tilde{\theta}^g) - M_0(y; \tilde{\theta}^g)], \end{aligned} \tag{7.23}$$

where,

$$
\begin{aligned}
N_0(\theta) &= \int D^{(2)}(g(x), f_\theta(x))\{\nabla^2 f_\theta(x)\} d\mu(x) \\
&+ \int D^{(2,2)}(g(x), f_\theta(x))\{\nabla f_\theta(x)\}\{\nabla f_\theta(x)\}^T d\mu(x) \\
M_0(y;\theta) &= D^{(1,2)}(g(y), f_\theta(y))\{\nabla f_\theta(y)\} \\
\xi(\theta) &= \int D^{(1,2)}(g(x), f_\theta(x))\{\nabla f_\theta(x)\} g(x) d\mu(x), \qquad (7.24)
\end{aligned}
$$

with additional restrictions $h(\theta) = 0$. *Here,* $D^{(i)}(\cdot,\cdot)$ *denotes the first order partial derivative of* $D(\cdot,\cdot)$ *w.r.t. its* i*-th argument and* $D^{(i,j)}(\cdot,\cdot)$ *denotes the second order partial derivative of* $D(\cdot,\cdot)$ *w.r.t. its* i*-th and* j*-th arguments,* $i, j = 1, 2$.

In case of RMGSDE, $\tilde{\theta}^g = \tilde{T}_{(\alpha,\tau,\gamma)}(G)$ and $\tilde{\theta}_\epsilon = \tilde{T}_{(\alpha,\tau,\gamma)}(G_\epsilon)$. Here,

$$
\rho(g, f_\theta) = Q_{(\alpha,\tau,\gamma)}(g, f_\theta), \qquad (7.25)
$$

where,

$$
D(g, f_\theta) = \frac{1}{\tau\bar{\tau}(\alpha-\gamma)} \left[\left\{ \tau g^{1+\alpha} + \bar{\tau} f^{1+\alpha} \right\} - \left\{ \tau g^{1+\gamma} + \bar{\tau} f^{1+\gamma} \right\}^{\frac{1+\alpha}{1+\gamma}} \right].
$$

Next, we have the following

$$
\begin{aligned}
D^{(1)}(g,f) &= \frac{(1+\alpha)}{\bar{\tau}(\alpha-\gamma)} \left[g^\alpha - g^\gamma \left\{ \tau g^{1+\gamma} + \bar{\tau} f^{1+\gamma} \right\}^{\frac{\alpha-\gamma}{1+\gamma}} \right] \\
D^{(1,2)}(g,f) &= -(1+\alpha)(gf)^\gamma \left\{ \tau g^{1+\gamma} + \bar{\tau} f^{1+\gamma} \right\}^{\frac{\alpha-2\gamma-1}{1+\gamma}} \\
D^{(2)}(g,f) &= \frac{(1+\alpha)}{\tau(\alpha-\gamma)} \left[f^\alpha - f^\gamma \left\{ \tau g^{1+\gamma} + \bar{\tau} f^{1+\gamma} \right\}^{\frac{\alpha-\gamma}{1+\gamma}} \right] \\
D^{(2,2)}(g,f) &= \frac{(1+\alpha)}{\tau(\alpha-\gamma)} \left[\alpha f^{\alpha-1} - \gamma f^{\gamma-1} \left\{ \tau g^{1+\gamma} + \bar{\tau} f^{1+\gamma} \right\}^{\frac{\alpha-\gamma}{1+\gamma}} \right] \\
&- \frac{(1+\alpha)}{\tau} \left[\bar{\tau} f^{2\gamma} \left\{ \tau g^{1+\gamma} + \bar{\tau} f^{1+\gamma} \right\}^{\frac{\alpha-2\gamma-1}{1+\gamma}} \right]. \qquad (7.26)
\end{aligned}
$$

Therefore, from (7.24),

$$\begin{aligned}
N(\theta) &= \frac{(1+\alpha)}{\tau(\alpha-\gamma)}\int\left[(1+\alpha)f_\theta^{1+\alpha}-(1+\gamma)f_\theta^{1+\gamma}\left\{\tau g^{1+\gamma}+\bar{\tau}f_\theta^{1+\gamma}\right\}^{\frac{\alpha-\gamma}{1+\gamma}}\right]u_\theta u_\theta^T d\mu\\
&- \frac{(1+\alpha)\bar{\tau}}{\tau}\int f_\theta^{2(1+\gamma)}\left\{\tau g^{1+\gamma}+\bar{\tau}f_\theta^{1+\gamma}\right\}^{\frac{\alpha-2\gamma-1}{1+\gamma}}u_\theta u_\theta^T d\mu\\
&- \frac{(1+\alpha)}{\tau(\alpha-\gamma)}\int\left[f_\theta^{1+\alpha}-f_\theta^{1+\gamma}\left\{\tau g^{1+\gamma}+\bar{\tau}f_\theta^{1+\gamma}\right\}^{\frac{\alpha-\gamma}{1+\gamma}}\right]i_\theta d\mu\\
M(y;\theta) &= -(1+\alpha)g^\gamma(y)f_\theta^{1+\gamma}(y)\left\{\tau g^{1+\gamma}(y)+\bar{\tau}f_\theta^{1+\gamma}(y)\right\}^{\frac{\alpha-2\gamma-1}{1+\gamma}}u_\theta(y)\\
\xi(\theta) &= -(1+\alpha)\int(gf_\theta)^{1+\gamma}\left\{\tau g^{1+\gamma}+\bar{\tau}f_\theta^{1+\gamma}\right\}^{\frac{\alpha-2\gamma-1}{1+\gamma}}u_\theta d\mu = E_g(M(X;\theta))
\end{aligned} \tag{7.27}$$

Substituting all these expressions in Equation (7.23), we get the required form of influence function. In particular, when $g = f_{\theta_0}$ for some $\theta_0 \in \Theta_0$, expressions in (7.27) reduce to the following

$$\begin{aligned}
N(\theta_0) &= (1+\alpha)\int f_{\theta_0}^{1+\alpha}u_{\theta_0}u_{\theta_0}^T d\mu\\
M(y;\theta_0) &= -(1+\alpha)f_{\theta_0}^\alpha(y)u_{\theta_0}(y)\\
\xi(\theta_0) &= -(1+\alpha)\int f_{\theta_0}^{1+\alpha}(y)u_{\theta_0}(y) = (1+\alpha)E_{f_{\theta_0}}(f_{\theta_0}^\alpha(X)u_{\theta_0}(X)).
\end{aligned}$$

In that case, the IF of RMGSDE will be simplified to

$$\begin{aligned}
IF(y,\tilde{T}_{(\alpha,\tau,\gamma)},F_{\theta_0}) &= \left\{\left[\int f_{\theta_0}^{1+\alpha}u_{\theta_0}u_{\theta_0}^T d\mu\right]^2+\frac{1}{(1+\alpha)^2}H(\theta_0)H(\theta_0)^T\right\}^{-1}\\
&\times\left\{\int f_{\theta_0}^{1+\alpha}u_{\theta_0}u_{\theta_0}^T d\mu\right\}\\
&\times\left\{f_{\theta_0}^\alpha(y)u_{\theta_0}(y)-E_{f_{\theta_0}}(f_{\theta_0}^\alpha(X)u_{\theta_0}(X))\right\},
\end{aligned} \tag{7.28}$$

which is dependent on α only, and moreover, it is exactly identical with the influence function of minimum DPD estimator derived under this restricted setup.

7.4.2 Discrete Setup

Here, we will consider the set of non-negative integers as the support of the model density. Evidently, the empirical estimate of $g(x)$ will be $r_n(x)$, the relative frequency at x based on a given sample of size n. Hence, the estimates can be obtained through an estimating equation after replacing $\hat{g}(x)$ with $r_n(x)$ and the integral with sum over the support. Let us denote the restricted best fitting parameter obtained by minimizing $Q_{(\alpha,\tau,\gamma)}(g, f_\theta)$ over Θ_0 under g by $\tilde{\theta}^g$.

For the following theorem, we assume that the Lehmann and Basu et al. conditions hold, as do the assumptions made in Chapter 5 in connection with the existence and consistency of the MGSBE under the discrete setup, for the restricted set $\Theta_0 = \{\theta : h(\theta) = 0\}$.

Theorem 7.11. *Under this setup of a discrete model and under the null hypothesis,*

(i) there exists a consistent sequence of roots $\tilde{\theta}_n$ to the restricted minimum GSD estimating equation (7.21) (after replacing $\hat{g}(x)$ by $r_n(x)$),

(ii) $\sqrt{n}(\tilde{\theta}_n - \tilde{\theta}_g) \overset{a}{\sim} N_p(0, \tilde{P}_g \tilde{V}_g \tilde{P}_g)$, where,

$$\tilde{P}_g = \tilde{J}_g^{-1}\left[I_p - H(\tilde{\theta}^g)\left\{H(\tilde{\theta}^g)^T \tilde{J}_g^{-1} H(\tilde{\theta}^g)\right\}^{-1} H(\tilde{\theta}^g)^T \tilde{J}_g^{-1}\right]$$

with

$$\begin{aligned}\tilde{J}_g &= E_g\left[u_{\tilde{\theta}^g}(X)\, u_{\tilde{\theta}^g}^T(X)\, K'\left(\tilde{\delta}_g^g(X)\right) f_{\tilde{\theta}^g}^{\alpha}(X)\right] - \int K\left(\tilde{\delta}_g^g(x)\right) \nabla^2 f_{\tilde{\theta}^g}(x) dx,\\ \tilde{V}_g &= Var_g\left[K'\left(\tilde{\delta}_g^g(X)\right) f_{\tilde{\theta}^g}^{\alpha}(X)\, u_{\tilde{\theta}^g}(X)\right];\ \tilde{\delta}_g^g(X) = \frac{g(x)}{f_{\tilde{\theta}^g}(x)} - 1. \qquad (7.29)\end{aligned}$$

Proof. The proof of part (i) is exactly similar to the proof of part (i) of Theorem 5.5. Hence, we need to prove the asymptotic normality of restricted MGSDE only.

We know $\tilde{\theta}^g$ is the restricted best fitting parameter under g. Putting this $\tilde{\theta}^g$ in our essential estimating equation

$$Q_n(\theta) = \sum_{x=0}^{\infty} K\left(\delta_n(x)\right) f_\theta^{1+\alpha}(x) u_\theta(x) = 0, \tag{7.30}$$

where,

$$\begin{aligned} K(\delta) &= \frac{1}{\tau(\alpha-\gamma)}\left[\left(\tau(\delta+1)^{1+\gamma}+\bar{\tau}\right)^{\frac{\alpha-\gamma}{1+\gamma}}-1\right], \\ \delta_n(x) &= \frac{r_n(x)}{f_\theta(x)}-1, \end{aligned}$$

we can conclude that

$$\begin{aligned} \sqrt{n} Q_n(\tilde{\theta}^g) &= \sqrt{n}\sum_{x=0}^{\infty} K\left(\tilde{\delta}_n^g(x)\right) f_{\tilde{\theta}^g}^{1+\alpha}(x) u_{\tilde{\theta}^g}(x) \\ &\overset{a}{\sim} N_p(0, \tilde{V}_g) \end{aligned} \tag{7.31}$$

and

$$\nabla Q_n(\tilde{\theta}^g) \xrightarrow{P} \tilde{J}_g, \tag{7.32}$$

where

$$\tilde{\delta}_n^g(x) = \frac{r_n(x)}{f_{\tilde{\theta}^g}(x)} - 1.$$

by proceeding as in the proof of the asymptotic normality of the MGSBE in Theorem 5.5. The use of the first order Taylor series expansion of $Q_n(\theta)$ around $\tilde{\theta}^g$ at $\tilde{\theta}_n$ leads us to

$$\sqrt{n} Q_n(\tilde{\theta}_n) = \sqrt{n} Q_n(\tilde{\theta}^g) + \sqrt{n}\nabla Q_n(\tilde{\theta}^g)\left(\tilde{\theta}_n - \tilde{\theta}^g\right) + \sqrt{n} \times o\left(||\tilde{\theta}_n - \tilde{\theta}^g||\right).$$

Through consistency, it is clear that $\sqrt{n} \times o\left(||\tilde{\theta}_n - \tilde{\theta}^g||\right) = o_p(1)$. Similarly, we have

$$\sqrt{n}\, h(\tilde{\theta}_n) \;=\; \sqrt{n}\, h(\tilde{\theta}^g) + H^T(\tilde{\theta}^g)\sqrt{n}\left(\tilde{\theta}_n - \tilde{\theta}^g\right) + \sqrt{n} \times o\left(||\tilde{\theta}_n - \tilde{\theta}^g||\right).$$

Since $h(\tilde{\theta}^g) = 0$ and $\tilde{\theta}_n$ is the solution of the estimating equation (7.21), we can conclude that

$$H^T(\tilde{\theta}^g)\sqrt{n}\left(\tilde{\theta}_n - \tilde{\theta}^g\right) + o_p(1) \;=\; 0.$$

Furthermore, the reduced estimating equation will be

$$\begin{aligned}\sqrt{n}Q_n(\tilde{\theta}^g) + \nabla Q_n(\tilde{\theta}^g)\sqrt{n}\left(\tilde{\theta}_n - \tilde{\theta}^g\right) + \sqrt{n}H(\tilde{\theta}_n)\lambda_n + o_p(1) \;&=\; 0,\\ H^T(\tilde{\theta}^g)\sqrt{n}\left(\tilde{\theta}_n - \tilde{\theta}^g\right) + o_p(1) \;&=\; 0.\end{aligned} \tag{7.33}$$

If we rewrite Equation (7.33) in a matrix format, we then get

$$\begin{pmatrix} \nabla Q_n(\tilde{\theta}^g) & H(\tilde{\theta}^g) \\ H^T(\tilde{\theta}^g) & 0 \end{pmatrix} \begin{pmatrix} \sqrt{n}\left(\tilde{\theta}_n - \tilde{\theta}^g\right) \\ \sqrt{n}\lambda_n \end{pmatrix} \;=\; \begin{pmatrix} -\sqrt{n}Q_n(\tilde{\theta}^g) \\ 0 \end{pmatrix} + o_p(1).$$

From the above expression, we get by taking the inverse of the first matrix of the left-hand side,

$$\sqrt{n}\left(\tilde{\theta}_n - \tilde{\theta}^g\right) \;=\; -P_n(\tilde{\theta}^g)\sqrt{n}Q_n(\tilde{\theta}^g) + o_p(1),$$

where,

$$\begin{aligned}P_n(\theta) \;=\; [\nabla Q_n(\theta)]^{-1}\Big[I_p - H(\theta)\left\{H^T(\theta)\,[\nabla Q_n(\theta)]^{-1}\,H(\theta)\right\}^{-1}\\ H^T(\theta)\,[\nabla Q_n(\theta)]^{-1}\Big].\end{aligned}$$

Again, using the above expression and (7.32), we can conclude that $P_n(\tilde{\theta}^g) \xrightarrow{\mathrm{P}} \tilde{P}_g$ and hence, by (7.31), the theorem is proved. □

Corollary 7.12. *When $g = f_{\theta_0}$ for some $\theta_0 \in \Theta$ satisfying the restrictions $h(\theta) = 0$, then $\tilde{\theta}^g = \theta_0$ and expressions in (7.29) reduce to*

$$\begin{aligned} \tilde{J}_{f_{\theta_0}} &= E_{f_{\theta_0}}\left[u_{\theta_0}(X)u_{\theta_0}^T(X)f_{\theta_0}^{\alpha}(X)\right], \\ \tilde{V}_{f_{\theta_0}} &= Var_{f_{\theta_0}}\left[u_{\theta_0}(X)f_{\theta_0}^{\alpha}(X)\right], \end{aligned}$$

and hence, $\tilde{P}_{f_{\theta_0}} = \tilde{J}_{f_{\theta_0}}^{-1}\left[I_p - H(\theta_0)\left\{H(\theta_0)^T\tilde{J}_{f_{\theta_0}}^{-1}H(\theta_0)\right\}^{-1}H(\theta_0)^T\tilde{J}_{f_{\theta_0}}^{-1}\right]$. *Therefore, at the model f_{θ_0},*

$$\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right) \overset{a}{\sim} N_p\left(0, \tilde{P}_{f_{\theta_0}}\tilde{V}_{f_{\theta_0}}\tilde{P}_{f_{\theta_0}}\right). \tag{7.34}$$

Evidently, the distribution is independent of tuning parameters τ and γ and depends on α only. Therefore, just like the unrestricted case, here also, the asymptotic distribution of restricted MGSDE is identical with that of restricted MDPDE at f_{θ_0}, as may be verified through a comparison of the result in Equation (7.35) and the result obtained in Theorem 2 of Basu et al. (2018).

Unlike the discrete case, the derivation of the MSGDE estimator in case of the continuous distribution is much more complicated, as it inevitably needs some form of non-parametric smoothing to construct the divergence, and the estimator (as well as its distribution) is a function of the kernel and the bandwidth. Trying to get an asymptotic distribution in the nature of the discrete distribution case will require the use of the Beran approach, or the use of a transparent kernel in the spirit of the discussion in Section 6.7. We, however, do neither, as the estimation of minimum GSD estimator under the kernel smoothing protocol is not what we are looking for. Our aim is to find a powerful but simple test of the hypothesis under our consideration, and noting the equivalence of the influence functions of the minimum GSD estimator and the minimum DPD estimator, we will continue to use the minimum DPD estimators, both in the unrestricted setup and under the restrictions imposed by the null, as the parameter estimates in the GSDT.

7.4.3 GSDT for Composite Hypotheses

To test $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \notin \Theta_0$, if we consider a test statistic based on the GSD family, then ideally, it should be constructed as

$$\tilde{T}_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_{(\alpha,\tau,\gamma)}, \tilde{\theta}_{(\alpha,\tau,\gamma)}\right) = 2n\, Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_{(\alpha,\tau,\gamma)}}, f_{\tilde{\theta}_{(\alpha,\tau,\gamma)}}\right), \quad (7.35)$$

where $\hat{\theta}_{(\alpha,\tau,\gamma)}$ and $\tilde{\theta}_{(\alpha,\tau,\gamma)}$ are the unrestricted and restricted MGSDE, respectively, with fixed α, τ and γ. However, depending on the triplet of the tuning parameters, this may entail the use of an estimator which requires non-parametric smoothing in the construction of the corresponding divergence in continuous models. In the spirit of our previous discussion, the observed equivalence of the influence functions of the MDPDE and MGSDE, and the equivalence of the asymptotic distributions of these two estimators – at least for discrete models – we will continue to supplant the MGSDE with the MDPDE of the parameter in Equation (7.35), and use the modified test statistic

$$\tilde{T}_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_{\alpha}, \tilde{\theta}_{\alpha}\right) = 2n\, Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_{\alpha}}, f_{\tilde{\theta}_{\alpha}}\right), \quad (7.36)$$

and explore its behaviour in robust testing of hypotheses.

Theorem 7.13. *Under the Lehmann and Basu et al. conditions (B1)-(B7), the asymptotic null distribution of the GSDT statistic* $\tilde{T}_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha, \tilde{\theta}_\alpha\right)$ *coincides with the distribution of* $\sum_{i=1}^r \tilde{\eta}_i Z_i^2$*, where* $Z_i \sim N(0,1)$ *independently, and* $\tilde{\eta}_i$*'s are non-zero eigenvalues of* $A_\alpha(\theta_0)\tilde{\Sigma}_\alpha(\theta_0)$ *where*
$A_\alpha(\theta_0) = \nabla^2 Q_{(\alpha,\tau,\gamma)}(f_\theta, f_{\theta_0})\Big|_{\theta=\theta_0}$*,* $\theta_0 \in \Theta_0$ *is the true parameter value under* H_0*, and*

$$\begin{aligned}\tilde{\Sigma}_\alpha(\theta_0) &= \left[J_\alpha^{-1}(\theta_0) - P_\alpha(\theta_0)\right] V_\alpha(\theta_0) \left[J_\alpha^{-1}(\theta_0) - P_\alpha(\theta_0)\right], \qquad (7.37)\\ r &= rank\left(V_\alpha(\theta_0)\left[J_\alpha^{-1}(\theta_0) - P_\alpha(\theta_0)\right] A_\alpha(\theta_0)\left[J_\alpha^{-1}(\theta_0) - P_\alpha(\theta_0)\right] V_\alpha(\theta_0)\right).\end{aligned}$$

Proof. A second-order Taylor series expansion of $Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\tilde{\theta}_\alpha}\right)$ around $\theta = \tilde{\theta}_\alpha$ at $\theta = \hat{\theta}_\alpha$ will give us

$$\begin{aligned}Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\tilde{\theta}_\alpha}\right) &= Q_{(\alpha,\tau,\gamma)}\left(f_{\tilde{\theta}_\alpha}, f_{\tilde{\theta}_\alpha}\right) + \sum_{i=1}^p \frac{\partial}{\partial\theta_i} Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\tilde{\theta}_\alpha}\right)\Bigg|_{\theta=\tilde{\theta}_\alpha}\left(\hat{\theta}_\alpha^i - \tilde{\theta}_\alpha^i\right)\\ &+ \frac{1}{2}\sum_{i,j=1}^p \frac{\partial^2}{\partial\theta_i\partial\theta_j} Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\tilde{\theta}_\alpha}\right)\Bigg|_{\theta=\tilde{\theta}_\alpha}\left(\hat{\theta}_\alpha^i - \tilde{\theta}_\alpha^i\right)\left(\hat{\theta}_\alpha^j - \tilde{\theta}_\alpha^j\right)\\ &+ o\left(||\hat{\theta}_\alpha - \tilde{\theta}_\alpha||^2\right).\end{aligned}$$

Evidently,

$$\begin{aligned}Q_{(\alpha,\tau,\gamma)}\left(f_{\tilde{\theta}_\alpha}, f_{\tilde{\theta}_\alpha}\right) &= 0,\\ \frac{\partial}{\partial\theta_i} Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\tilde{\theta}_\alpha}\right)\Bigg|_{\theta=\tilde{\theta}_\alpha} &= 0, \text{ and}\\ \frac{\partial^2}{\partial\theta_i\partial\theta_j} Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\tilde{\theta}_\alpha}\right)\Bigg|_{\theta=\tilde{\theta}_\alpha} &= (1+\alpha)\int f_{\tilde{\theta}_\alpha}^{\alpha-1}\left(\frac{\partial f_{\tilde{\theta}_\alpha}}{\partial\theta_i}\right)\left(\frac{\partial f_{\tilde{\theta}_\alpha}}{\partial\theta_j}\right)\\ &= a_{ij}^\alpha(\tilde{\theta}_\alpha), \qquad (7.38)\end{aligned}$$

where $a_{ij}^{\alpha}(\tilde{\theta}_\alpha)$ is the (i,j)th element of $A_\alpha(\tilde{\theta}_\alpha)$. Again,

$$\begin{aligned}\tilde{T}_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha,\tilde{\theta}_\alpha\right) &= 2n\, Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\tilde{\theta}_\alpha}\right)\\ &= n\sum_{i,j=1}^{p}\frac{\partial^2}{\partial\theta_i\partial\theta_j}Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\tilde{\theta}_\alpha}\right)\Big|_{\theta=\tilde{\theta}_\alpha}\left(\hat{\theta}_\alpha^i-\tilde{\theta}_\alpha^i\right)\left(\hat{\theta}_\alpha^j-\tilde{\theta}_\alpha^j\right)\\ &+ n\times o\left(||\hat{\theta}_\alpha-\tilde{\theta}_\alpha||^2\right)\\ &= \sqrt{n}\left(\hat{\theta}_\alpha-\tilde{\theta}_\alpha\right)^T A_\alpha(\tilde{\theta}_\alpha)\sqrt{n}\left(\hat{\theta}_\alpha-\tilde{\theta}_\alpha\right)+n\times o\left(||\hat{\theta}_\alpha-\tilde{\theta}_\alpha||^2\right).\end{aligned}$$

Furthermore, from Theorem 3 in Basu et al. (2018), we have the result,

$$\sqrt{n}\left(\hat{\theta}_\alpha-\tilde{\theta}_\alpha\right)\overset{\text{a}}{\sim}N\left(0,\tilde{\Sigma}_\alpha(\theta_0)\right),$$

where $\tilde{\Sigma}_\alpha(\theta_0)$ is as defined in the statement of the theorem. Therefore, just like the other theorems in case of the simple hypotheses, here also, by Lemma 7.1, we can conclude that the asymptotic distribution of the GSDT statistic $\tilde{T}_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha,\tilde{\theta}_\alpha\right)$ under $\theta_0\in\Theta_0$ coincides with the distribution of the random variable $\sum_{i=1}^{r}\tilde{\eta}_i Z_i^2$, where $Z_i\overset{i.i.d.}{\sim}N(0,1)$ and where r and $\tilde{\eta}_i$'s are as defined in the theorem. □

Next, we are going to derive the expression of the power function of the above-mentioned GSDT statistic $\tilde{T}_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha,\tilde{\theta}_\alpha\right)$ at any point $\theta^*\notin\Theta_0$. Moreover, when $\theta^*\notin\Theta_0$ is the true parameter value then $\hat{\theta}_\alpha\xrightarrow{P}\theta^*$, but $\tilde{\theta}_\alpha\xrightarrow{P}\theta_0$ with $\theta_0\in\Theta_0$, with $\theta_0\neq\theta^*$. Hence, under Lehmann and Basu et al. conditions, $\sqrt{n}\left(\hat{\theta}_\alpha-\theta^*\right)\overset{\text{a}}{\sim}N(0,\Sigma_\alpha(\theta^*))$, $\Sigma_\alpha(\theta^*)=J_\alpha^{-1}(\theta^*)V_\alpha(\theta^*)J_\alpha^{-1}(\theta^*)$, and $\sqrt{n}\left(\tilde{\theta}_\alpha-\theta_0\right)\overset{\text{a}}{\sim}N(0,P_\alpha(\theta_0)V_\alpha(\theta_0)P_\alpha(\theta_0))$.

Theorem 7.14. *Under the Lehmann and Basu et al. conditions (B1)-(B7), for any $\theta^*\notin\Theta_0$, an approximate expression of the power* *function at $\theta = \theta^*$ $(\neq \theta_0)$ at significance level β is given by*

$$\tilde{\Pi}_{n,\beta}^{(\alpha,\tau,\gamma)}(\theta^*) \;=\; 1 - \Phi\left(\frac{\sqrt{n}}{\tilde{\sigma}_{(\alpha,\tau,\gamma)}(\theta^*,\theta_0)}\left(\frac{\tilde{t}^{\beta}_{(\alpha,\tau,\gamma)}}{2n} - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right)\right)\right),$$

where $\tilde{t}^{\beta}_{(\alpha,\tau,\gamma)}$ is the $(1-\alpha)$th quantile of the asymptotic null distribution of the GSDT and

$$\begin{aligned}\tilde{\sigma}^2_{(\alpha,\tau,\gamma)}(\theta^*,\theta_0) \;=\;& M^T_{1,(\alpha,\tau,\gamma)}(\theta^*,\theta_0)\Sigma_\alpha(\theta^*)M_{1,(\alpha,\tau,\gamma)}(\theta^*,\theta_0)\\ +\;& M^T_{1,(\alpha,\tau,\gamma)}(\theta^*,\theta_0)A_{12}M_{2,(\alpha,\tau,\gamma)}(\theta^*,\theta_0)\\ +\;& M^T_{2,(\alpha,\tau,\gamma)}(\theta^*,\theta_0)A^T_{12}M_{1,(\alpha,\tau,\gamma)}(\theta^*,\theta_0)\\ +\;& M^T_{2,(\alpha,\tau,\gamma)}(\theta^*,\theta_0)P_\alpha(\theta_0)V_\alpha(\theta_0)P_\alpha(\theta_0)M_{2,(\alpha,\tau,\gamma)}(\theta^*,\theta_0),\end{aligned}$$

where,

$$\begin{aligned}M_{1,(\alpha,\tau,\gamma)}(\theta^*,\theta_0) &= \nabla Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\theta_0}\right)\Bigg|_{\theta=\theta^*},\\ M_{2,(\alpha,\tau,\gamma)}(\theta^*,\theta_0) &= \nabla Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta}\right)\Bigg|_{\theta=\theta_0},\end{aligned} \tag{7.39}$$

and, $\quad Cov(\hat{\theta}_\alpha, \tilde{\theta}_\alpha) = A_{12}$ *for some $p \times p$ matrix A_{12}.*

Proof. A first order Taylor series expansion of $Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\tilde{\theta}_\alpha}\right)$ around f_{θ^*}, $\theta^* \neq \tilde{\theta}_\alpha$ at $\theta = \hat{\theta}_\alpha$, we get,

$$\begin{aligned}Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\tilde{\theta}_\alpha}\right) \;=\;& Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\tilde{\theta}_\alpha}\right) + \sum_{i=1}^{p}\frac{\partial}{\partial\theta_i}Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\tilde{\theta}_\alpha}\right)\Bigg|_{\theta=\theta^*}\left(\hat{\theta}^i_\alpha - \theta^{*i}\right)\\ +\;& o\left(||\hat{\theta}_\alpha - \theta^*||\right)\\ =\;& Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\tilde{\theta}_\alpha}\right) + M^T_{1,(\alpha,\tau,\gamma)}\left(\theta^*, \tilde{\theta}_\alpha\right)\left(\hat{\theta}_\alpha - \theta^*\right) + o\left(||\hat{\theta}_\alpha - \theta^*||\right).\end{aligned}$$

A similar step for $Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\tilde{\theta}_\alpha}\right)$ around f_{θ_0} at $\theta = \tilde{\theta}_\alpha$ will lead us to

$$\begin{aligned} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\tilde{\theta}_\alpha}\right) &= Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right) + \sum_{j=1}^{r} \frac{\partial}{\partial\theta_j} Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta}\right)\Bigg|_{\theta=\theta_0} \left(\tilde{\theta}_\alpha^j - \theta_0^j\right) \\ &+ o\left(||\tilde{\theta}_\alpha - \theta_0||\right) \\ &= Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right) + M^T_{2,(\alpha,\tau,\gamma)}\left(\theta^*, \theta_0\right)\left(\tilde{\theta}_\alpha - \theta_0\right) + o\left(||\tilde{\theta}_\alpha - \theta_0||\right). \end{aligned}$$

Note that,

$$\begin{aligned} &\tilde{\theta}_\alpha = \text{restricted MDPDE under } H_0 \\ .\ \Rightarrow\ &\tilde{\theta}_\alpha \to \theta_0, \quad \text{as} \quad n \to \infty, \\ \Rightarrow\ &M_{1,(\alpha,\tau,\gamma)}\left(\theta^*, \tilde{\theta}_\alpha\right) \to M_{1,(\alpha,\tau,\gamma)}\left(\theta^*, \theta_0\right), \quad \text{as} \quad n \to \infty. \end{aligned}$$

Hence, combining all the above, we get

$$\begin{aligned} &\sqrt{n}\left(Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\tilde{\theta}_\alpha}\right) - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right)\right) \\ = \ &M^T_{1,(\alpha,\tau,\gamma)}\left(\theta^*, \tilde{\theta}_\alpha\right)\sqrt{n}\left(\hat{\theta}_\alpha - \theta^*\right) + M^T_{2,(\alpha,\tau,\gamma)}\left(\theta^*, \theta_0\right)\sqrt{n}\left(\tilde{\theta}_\alpha - \theta_0\right) \\ + \ &o\left(||\hat{\theta}_\alpha - \theta^*||\right) + o\left(||\tilde{\theta}_\alpha - \theta_0||\right) \\ \to \ &M^T_{1,(\alpha,\tau,\gamma)}\left(\theta^*, \theta_0\right)\sqrt{n}\left(\hat{\theta}_\alpha - \theta^*\right) + M^T_{2,(\alpha,\tau,\gamma)}\left(\theta^*, \theta_0\right)\sqrt{n}\left(\tilde{\theta}_\alpha - \theta_0\right) \\ = \ &M^T_{(\alpha,\tau,\gamma)}\left(\theta^*, \theta_0\right)\begin{pmatrix} \sqrt{n}\left(\hat{\theta}_\alpha - \theta^*\right) \\ \sqrt{n}\left(\tilde{\theta}_\alpha - \theta_0\right) \end{pmatrix}, \quad \text{as} \quad n \to \infty, \end{aligned}$$

where,

$$M^T_{(\alpha,\tau,\gamma)}\left(\theta^*, \theta_0\right) = \left(M^T_{1,(\alpha,\tau,\gamma)}\left(\theta^*, \theta_0\right) \quad M^T_{2,(\alpha,\tau,\gamma)}\left(\theta^*, \theta_0\right)\right).$$

Again, we have already shown that

$$\begin{aligned} \sqrt{n}\left(\hat{\theta}_\alpha - \theta^*\right) &\overset{\text{a}}{\sim} N\left(0, J_\alpha^{-1}(\theta^*)V_\alpha(\theta^*)J_\alpha^{-1}(\theta^*)\right), \\ \sqrt{n}\left(\tilde{\theta}_\alpha - \theta_0\right) &\overset{\text{a}}{\sim} N\left(0, P_\alpha(\theta_0)V_\alpha(\theta_0)P_\alpha(\theta_0)\right). \end{aligned}$$

Hence, under Basu et al. conditions,

$$\begin{aligned}\begin{pmatrix}\sqrt{n}\left(\hat{\theta}_\alpha-\theta^*\right)\\ \sqrt{n}\left(\tilde{\theta}_\alpha-\theta_0\right)\end{pmatrix} &\overset{\text{a}}{\to} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}J_\alpha^{-1}(\theta^*)V_\alpha(\theta^*)J_\alpha^{-1}(\theta^*) & A_{12}\\ A_{21} & P_\alpha(\theta_0)V_\alpha(\theta_0)P_\alpha(\theta_0)\end{pmatrix}\right)\\ &= N\left(\begin{pmatrix}0\\0\end{pmatrix},\ \Sigma^*\right).\end{aligned}$$

It immediately follows that the asymptotic distribution of $\sqrt{n}\left(Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha},f_{\tilde{\theta}_\alpha}\right)-Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*},f_{\theta_0}\right)\right)$ and $M^T_{(\alpha,\tau,\gamma)}\left(\theta^*,\theta_0\right)\begin{pmatrix}\sqrt{n}\left(\hat{\theta}_\alpha-\theta^*\right)\\ \sqrt{n}\left(\tilde{\theta}_\alpha-\theta_0\right)\end{pmatrix}$ are the same. Therefore, as $n\to\infty$,

$$\begin{aligned}\sqrt{n}\left(Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha},f_{\tilde{\theta}_\alpha}\right)-Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*},f_{\theta_0}\right)\right) &\overset{\text{a}}{\sim} N\left(0, M^T_{(\alpha,\tau,\gamma)}\left(\theta^*,\theta_0\right)\Sigma^* M_{(\alpha,\tau,\gamma)}\left(\theta^*,\theta_0\right)\right)\\ &\equiv N\left(0,\tilde{\sigma}^2_{(\alpha,\tau,\gamma)}\left(\theta^*,\theta_0\right)\right),\end{aligned}$$

where,

$$\begin{aligned}&\tilde{\sigma}^2_{(\alpha,\tau,\gamma)}\left(\theta^*,\theta_0\right)\\ =\ & M_{1,(\alpha,\tau,\gamma)}\left(\theta^*,\theta_0\right)^T\ J_\alpha^{-1}(\theta^*)V_\alpha(\theta^*)J_\alpha^{-1}(\theta^*)\ M_{1,(\alpha,\tau,\gamma)}\left(\theta^*,\theta_0\right)\\ +\ & M_{1,(\alpha,\tau,\gamma)}\left(\theta^*,\theta_0\right)^T\ A_{12}\ M_{2,(\alpha,\tau,\gamma)}\left(\theta^*,\theta_0\right)\\ +\ & M_{2,(\alpha,\tau,\gamma)}\left(\theta^*,\theta_0\right)^T\ A_{21}\ M_{1,(\alpha,\tau,\gamma)}\left(\theta^*,\theta_0\right)\\ +\ & M_{2,(\alpha,\tau,\gamma)}\left(\theta^*,\theta_0\right)^T\ P_\alpha(\theta_0)V_\alpha(\theta_0)P_\alpha(\theta_0)\ M_{2,(\alpha,\tau,\gamma)}\left(\theta^*,\theta_0\right).\end{aligned}$$

We fix $\theta = \theta^* \neq \theta_0$. The power at $\theta = \theta^*$ for a fixed significance level β will be

$$\begin{aligned}
\tilde{\Pi}_{n,\beta}^{(\alpha,\tau,\gamma)} &= P_{\theta=\theta^*}\left(\tilde{T}_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha, \tilde{\theta}_\alpha\right) > \tilde{t}_\beta^{(\alpha,\tau,\gamma)}\right) \\
&= P_{\theta=\theta^*}\left(Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\tilde{\theta}_\alpha}\right) > \frac{\tilde{t}_\beta^{(\alpha,\tau,\gamma)}}{2n}\right) \\
&= 1 - P_{\theta=\theta^*}\Bigg(\frac{\sqrt{n}\left(Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\tilde{\theta}_\alpha}\right) - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right)\right)}{\tilde{\sigma}_{(\alpha,\tau,\gamma)}\left(\theta^*, \theta_0\right)} \\
&\qquad\qquad\qquad \leq \frac{\sqrt{n}\left(\frac{\tilde{t}_\beta^{(\alpha,\tau,\gamma)}}{2n} - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right)\right)}{\tilde{\sigma}_{(\alpha,\tau,\gamma)}\left(\theta^*, \theta_0\right)}\Bigg) \\
&\to 1 - \Phi\left(\frac{\sqrt{n}}{\tilde{\sigma}_{(\alpha,\tau,\gamma)}\left(\theta^*, \theta_0\right)}\left(\frac{\tilde{t}_\beta^{(\alpha,\tau,\gamma)}}{2n} - Q_{(\alpha,\tau,\gamma)}\left(f_{\theta^*}, f_{\theta_0}\right)\right)\right).
\end{aligned}$$

As $n \to \infty$, $\tilde{\Pi}_{n,\beta}^{(\alpha,\tau,\gamma)} \to 1$, that is, the test is consistent in the Fraser's sense. □

7.4.4 Robustness Properties of the GSDT (Composite Hypotheses)

7.4.4.1 Influence Function of the Test

We have already defined this term and its interpretation in the field of statistical hypotheses testing in Section 7.3.3.1 where the hypotheses were simple. Here, in case of composite hypotheses, the GSDT functional will be defined as

$$\tilde{T}_{(\alpha,\tau,\gamma)}^{(1)}(G) = Q_{(\alpha,\tau,\gamma)}\left(f_{T_\alpha(G)}, f_{\tilde{T}_\alpha(G)}\right),$$

where $T_\alpha(G)$ is the unrestricted minimum DPD functional derived over the whole parameter Θ and $\tilde{T}_\alpha(G)$ is the restricted minimum

DPD functional derived over Θ_0. Here, the contamination distribution remains the same as earlier.

Therefore, the first order IF of the GSDT functional is given by

$$\begin{aligned} IF\left(y, \tilde{T}^{(1)}_{(\alpha,\tau,\gamma)}, G\right) &= \frac{\partial}{\partial \epsilon}\tilde{T}^{(1)}_{(\alpha,\tau,\gamma)}(G_\epsilon)\Big|_{\epsilon=0} \\ &= \frac{\partial}{\partial \theta} Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\tilde{T}_\alpha(G)}\right)\Big|_{\theta=T_\alpha(G)} \frac{\partial}{\partial \epsilon} T_\alpha(G_\epsilon)\Big|_{\epsilon=0} \\ &+ \frac{\partial}{\partial \theta} Q_{(\alpha,\tau,\gamma)}\left(f_{T_\alpha(G)}, f_\theta\right)\Big|_{\theta=\tilde{T}_\alpha(G)} \frac{\partial}{\partial \epsilon} \tilde{T}_\alpha(G_\epsilon)\Big|_{\epsilon=0} \\ &= M_{1,(\alpha,\tau,\gamma)}\left(T_\alpha(G), \tilde{T}_\alpha(G)\right)^T IF(y, T_\alpha, G) \\ &+ M_{2,(\alpha,\tau,\gamma)}\left(T_\alpha(G), \tilde{T}_\alpha(G)\right)^T IF(y, \tilde{T}_\alpha, G). \end{aligned}$$

Under H_0, if we consider $\theta_0 \in \Theta_0$ is the true parameter value, then in case of $G = F_{\theta_0}$, $T_\alpha(F_{\theta_0}) = \tilde{T}_\alpha(F_{\theta_0}) = \theta_0$ and $M_{1,(\alpha,\tau,\gamma)}(\theta_0, \theta_0) =$

$M_{2,(\alpha,\tau,\gamma)}(\theta_0, \theta_0) = 0$. Therefore, the first order IF under the composite null is zero. Next, we consider the second order IF

$$\begin{aligned}
IF_2\left(y, \tilde{T}^{(1)}_{(\alpha,\tau,\gamma)}, G\right) &= \left.\frac{\partial^2}{\partial\epsilon^2}\tilde{T}^{(1)}_{(\alpha,\tau,\gamma)}(G_\epsilon)\right|_{\epsilon=0} \\
&= \left.\frac{\partial}{\partial\epsilon}\left[M_{1,(\alpha,\tau,\gamma)}\left(T_\alpha(G), \tilde{T}_\alpha(G)\right)^T IF(y, T_\alpha, G)\right]\right|_{\epsilon=0} \\
&+ \left.\frac{\partial}{\partial\epsilon}\left[M_{2,(\alpha,\tau,\gamma)}\left(T_\alpha(G), \tilde{T}_\alpha(G)\right)^T IF(y, \tilde{T}_\alpha, G)\right]\right|_{\epsilon=0} \\
&= M_{1,(\alpha,\tau,\gamma)}\left(T_\alpha(G), \tilde{T}_\alpha(G)\right)^T \left.\frac{\partial^2}{\partial\epsilon^2}T_\alpha(G_\epsilon)\right|_{\epsilon=0} \\
&+ M_{2,(\alpha,\tau,\gamma)}\left(T_\alpha(G), \tilde{T}_\alpha(G)\right)^T \left.\frac{\partial^2}{\partial\epsilon^2}\tilde{T}_\alpha(G_\epsilon)\right|_{\epsilon=0} \\
&+ IF\left(y, T_\alpha, G\right)^T \nabla_{\theta_1}\nabla_{\theta_1}Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\Big|_{\theta_1=T_\alpha(G),\theta_2=\tilde{T}_\alpha(G)} IF\left(y, T_\alpha, G\right) \\
&+ IF\left(y, T_\alpha, G\right)^T \nabla_{\theta_1}\nabla_{\theta_2}Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\Big|_{\theta_1=T_\alpha(G),\theta_2=\tilde{T}_\alpha(G)} IF\left(y, \tilde{T}_\alpha, G\right) \\
&+ IF\left(y, \tilde{T}_\alpha, G\right)^T \nabla_{\theta_2}\nabla_{\theta_1}Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\Big|_{\theta_1=T_\alpha(G),\theta_2=\tilde{T}_\alpha(G)} IF\left(y, T_\alpha, G\right) \\
&+ IF\left(y, \tilde{T}_\alpha, G\right)^T \nabla_{\theta_2}\nabla_{\theta_2}Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_1}, f_{\theta_2}\right)\Big|_{\theta_1=T_\alpha(G),\theta_2=\tilde{T}_\alpha(G)} IF\left(y, \tilde{T}_\alpha, G\right) \\
&= M_{1,(\alpha,\tau,\gamma)}\left(T_\alpha(G), \tilde{T}_\alpha(G)\right)^T \left.\frac{\partial^2}{\partial\epsilon^2}T_\alpha(G_\epsilon)\right|_{\epsilon=0} \\
&+ M_{2,(\alpha,\tau,\gamma)}\left(T_\alpha(G), \tilde{T}_\alpha(G)\right)^T \left.\frac{\partial^2}{\partial\epsilon^2}\tilde{T}_\alpha(G_\epsilon)\right|_{\epsilon=0} \\
&+ IF\left(y, T_\alpha, G\right)^T A_{1,1,(\alpha,\tau,\gamma)}\left(T_\alpha(G), \tilde{T}_\alpha(G)\right) IF\left(y, T_\alpha, G\right) \\
&+ IF\left(y, T_\alpha, G\right)^T A_{1,2,(\alpha,\tau,\gamma)}\left(T_\alpha(G), \tilde{T}_\alpha(G)\right) IF\left(y, \tilde{T}_\alpha, G\right) \\
&+ IF\left(y, \tilde{T}_\alpha, G\right)^T A_{2,1,(\alpha,\tau,\gamma)}\left(T_\alpha(G), \tilde{T}_\alpha(G)\right) IF\left(y, T_\alpha, G\right) \\
&+ IF\left(y, \tilde{T}_\alpha, G\right)^T A_{2,2,(\alpha,\tau,\gamma)}\left(T_\alpha(G), \tilde{T}_\alpha(G)\right) IF\left(y, \tilde{T}_\alpha, G\right),
\end{aligned} \tag{7.40}$$

where, $A_{i,j,(\alpha,\tau,\gamma)}(\theta_1^*, \theta_2^*) = \nabla_{\theta_i} \nabla_{\theta_j} Q_{(\alpha,\tau,\gamma)} \left(f_{\theta_1}, f_{\theta_2}\right)\Big|_{\theta_1=\theta_1^*, \theta_2=\theta_2^*}$, $i, j = 1, 2$. Moreover, when $G = F_{\theta_0}$, it reduces to the following

$$\begin{aligned} IF_2\left(y, \tilde{T}^{(1)}_{(\alpha,\tau,\gamma)}, F_{\theta_0}\right) &= \left(IF\left(y, T_\alpha, F_{\theta_0}\right) - IF\left(y, \tilde{T}_\alpha, F_{\theta_0}\right)\right)^T A_\alpha(\theta_0) \\ &\qquad \left(IF\left(y, T_\alpha, F_{\theta_0}\right) - IF\left(y, \tilde{T}_\alpha, F_{\theta_0}\right)\right). \end{aligned}$$

Evidently, when $G = F_{\theta_0}$, the second order influence function is dependent only on α, and, at the model, the use of the MDPDE in place of the MGSDE does not alter the influence function.

7.4.4.2 Influence Function of the Level and the Power

As in the simple null hypothesis case, here also we need to observe the IF of the level and power of the GSDT. We will analyze its power performance in case of fixed and contiguous alternatives $H_{1,n} : \theta = \theta_n,\ \theta_n = \theta_0 + \frac{\Delta}{\sqrt{n}}$ for non-negative Δ, $\theta_0 \in \Theta_0$; for existence of θ_0, we consider it as a limit point of Θ_0, a closed subset of Θ. Since we are mainly concerned with the stability of these tests in the presence of outliers, we consider the contaminated cases here. As a result, we want to derive the level influence function (LIF) and power influence function (PIF) as defined earlier.

Theorem 7.15. *Under the above-mentioned setup, assuming that the Lehmann and Basu et al. conditions (B1)-(B7) hold, for any $\Delta \in \mathbb{R}^p$ and $\epsilon > 0$,*

(i) the asymptotic distribution of $\tilde{T}_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha, \tilde{\theta}_\alpha\right)$ under $F^P_{n,\epsilon,y}$ is the same as that of the distribution of $W^T A_\alpha(\theta_0) W$, where, $W \sim N_p(\tilde{\Delta}^*, \tilde{\Sigma}_\alpha(\theta_0))$,

$$\tilde{\Delta}^* = \Delta + \epsilon\left(IF(y, T_\alpha, F_{\theta_0}) - IF(y, \tilde{T}_\alpha, F_{\theta_0})\right)$$

and, $\tilde{\Sigma}_\alpha(\theta_0)$ *is same as defined in Theorem 7.13. Equivalently, this distribution is the same as that of* $\sum_{i=1}^r \tilde{\eta}_i \chi^2_{1,\tilde{\delta}_i}$, *where* $\tilde{\eta}_1, \ldots, \tilde{\eta}_r$ *are the non-zero eigenvalues of* $A_\alpha(\theta_0)\tilde{\Sigma}_\alpha(\theta_0)$ *and* $\left(\sqrt{\tilde{\delta}_1}, \sqrt{\tilde{\delta}_2}, \ldots, \sqrt{\tilde{\delta}_r}\right)^T = \tilde{V}_\alpha(\theta_0)\tilde{\Sigma}_\alpha^{-1/2}(\theta_0)\tilde{\Delta}^*$ *with* $\tilde{V}_\alpha(\theta_0)$ *being the matrix of the normalized eigenvectors of* $A_\alpha(\theta_0)\tilde{\Sigma}_\alpha(\theta_0)$*.*

(ii) the asymptotic power will be

$$\tilde{P}(\Delta, \epsilon) = \sum_{\nu=0}^{\infty} \tilde{C}_\nu^\alpha(\theta_0, \tilde{\Delta}^*) P\left(\chi^2_{r+2\nu} > \frac{\tilde{t}_\beta^{(\alpha,\tau,\gamma)}}{\tilde{\eta}_{(1)}}\right), \qquad (7.41)$$

where $\tilde{\eta}_{(1)} = \min_i \tilde{\eta}_i$,

$$C_\nu(\theta_0, \tilde{\Delta}^*) = \frac{1}{\nu!}\left(\prod_{j=1}^r \frac{\tilde{\eta}_{(1)}}{\tilde{\eta}_j}\right)^{1/2} e^{-\tilde{\delta}/2} E(\tilde{Q}^\nu)$$

and

$$\tilde{Q} = \frac{1}{2}\sum_{j=1}^r \left[\left(1 - \frac{\tilde{\eta}_{(1)}}{\tilde{\eta}_j}\right)^{1/2} Z_j + \sqrt{\tilde{\delta}_j}\left(\frac{\tilde{\eta}_{(1)}}{\tilde{\eta}_j}\right)^{1/2}\right]^2$$

for r *independent standard normal variables* $Z_1, Z_2, \ldots, Z_r$*.*

Proof. Let us consider $\theta_n^* =$ Restricted best fitting parameter under $F^P_{n,\epsilon,y}$ and $\tilde{\theta}_n^* =$ Unrestricted best fitting parameter under $F^P_{n,\epsilon,y}$ at first. A second order Taylor series expansion of $Q_{(\alpha,\tau,\gamma)}\left(f_\theta, f_{\tilde{\theta}_\alpha}\right)$ around $\theta = \theta_n^*$ at $\theta = \hat{\theta}_\alpha$ gives

$$\begin{aligned} Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\tilde{\theta}_\alpha}\right) &= Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_n^*}, f_{\tilde{\theta}_\alpha}\right) + \left(\hat{\theta}_\alpha - \theta_n^*\right)^T M_{1,(\alpha,\tau,\gamma)}\left(\theta_n^*, \tilde{\theta}_\alpha\right) \\ &+ \frac{1}{2}\left(\hat{\theta}_\alpha - \theta_n^*\right)^T A_{1,1,(\alpha,\tau,\gamma)}\left(\theta_n^*, \tilde{\theta}_\alpha\right)\left(\hat{\theta}_\alpha - \theta_n^*\right) \\ &+ o\left(||\hat{\theta}_\alpha - \theta_n^*||^2\right), \end{aligned}$$

where $M_{i,(\alpha,\tau,\gamma)}(\cdot,\cdot)$ and $A_{i,j,(\alpha,\tau,\gamma)}(\cdot,\cdot)$, $i,j = 1,2$ are as defined in Equations (7.39) and (7.40). Again, a second order Taylor series expansion of $Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_n^*}, f_\theta\right)$ around $\theta = \tilde{\theta}_n^*$ at $\theta = \tilde{\theta}_\alpha$ will lead us to

$$\begin{aligned}
Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_n^*}, f_{\tilde{\theta}_\alpha}\right) &= Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_n^*}, f_{\tilde{\theta}_n^*}\right) + \left(\tilde{\theta}_\alpha - \tilde{\theta}_n^*\right)^T M_{2,(\alpha,\tau,\gamma)}\left(\theta_n^*, \tilde{\theta}_n^*\right) \\
&+ \frac{1}{2}\left(\tilde{\theta}_\alpha - \tilde{\theta}_n^*\right)^T A_{2,2,(\alpha,\tau,\gamma)}\left(\theta_n^*, \tilde{\theta}_n^*\right)\left(\tilde{\theta}_\alpha - \tilde{\theta}_n^*\right) \\
&+ o\left(||\tilde{\theta}_\alpha - \tilde{\theta}_n^*||^2\right), \\
M_{1,(\alpha,\tau,\gamma)}\left(\theta_n^*, \tilde{\theta}_\alpha\right) &= M_{1,(\alpha,\tau,\gamma)}\left(\theta_n^*, \tilde{\theta}_n^*\right) + \left(\tilde{\theta}_\alpha - \tilde{\theta}_n^*\right) A_{2,1,(\alpha,\tau,\gamma)}\left(\theta_n^*, \tilde{\theta}_n^*\right) \\
&+ o\left(||\tilde{\theta}_\alpha - \tilde{\theta}_n^*||\right), \\
A_{1,1,(\alpha,\tau,\gamma)}\left(\theta_n^*, \tilde{\theta}_\alpha\right) &= A_{1,1,(\alpha,\tau,\gamma)}\left(\theta_n^*, \tilde{\theta}_n^*\right) + o_p(1).
\end{aligned}$$

Again, we consider the Taylor series expansion of $M_{j,(\alpha,\tau,\gamma)}\left(\theta, \tilde{\theta}_n^*\right)$ around $\theta = \theta_0$ at $\theta = \theta_n^*$ to get

$$\begin{aligned}
M_{j,(\alpha,\tau,\gamma)}\left(\theta_n^*, \tilde{\theta}_n^*\right) &= M_{j,(\alpha,\tau,\gamma)}\left(\theta_0, \tilde{\theta}_n^*\right) + A_{1,j,(\alpha,\tau,\gamma)}\left(\theta_0, \tilde{\theta}_n^*\right)\frac{\Delta}{\sqrt{n}} \\
&+ \frac{\epsilon}{\sqrt{n}} A_{1,j,(\alpha,\tau,\gamma)}\left(\theta_0, \tilde{\theta}_n^*\right) IF(y, T_\alpha, F_{\theta_0}) + o\left(\frac{1}{\sqrt{n}}\right) \\
&= M_{j,(\alpha,\tau,\gamma)}\left(\theta_0, \tilde{\theta}_n^*\right) + \frac{\Delta + \epsilon IF(y, T_\beta, F_{\theta_0})}{\sqrt{n}} A_{1,j,(\alpha,\tau,\gamma)}\left(\theta_0, \tilde{\theta}_n^*\right) + o\left(\frac{1}{\sqrt{n}}\right) \\
&= M_{j,(\alpha,\tau,\gamma)}\left(\theta_0, \tilde{\theta}_n^*\right) + A_{1,j,(\alpha,\tau,\gamma)}\left(\theta_0, \tilde{\theta}_n^*\right)\frac{\tilde{\Delta}}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

We have already used the following relation before,

$$\sqrt{n}(\theta_n^* - \theta_0) = \Delta + \epsilon IF(y, T_\alpha, F_{\theta_0}). \tag{7.42}$$

Similarly, considering $\tilde{\theta}_n^*$ as a function of ϵ_n, i.e., $f(\epsilon_n) = \frac{\epsilon}{\sqrt{n}}$, we get

$$\begin{aligned}
\tilde{\theta}_n^* = f(\epsilon_n) &= \sum_{k=0}^{\infty} \frac{1}{k!} \frac{\epsilon^k}{n^{k/2}} \frac{\partial^k f(\epsilon_n)}{\partial \epsilon} \bigg|_{\epsilon=0} \\
&= \theta_0 + \frac{\epsilon}{\sqrt{n}} IF(y, \tilde{T}_\alpha, F_{\theta_0}) + \sum_{k=2}^{\infty} \frac{1}{k!} \frac{\epsilon^k}{n^{k/2}} \frac{\partial^k f(\epsilon_n)}{\partial \epsilon} \bigg|_{\epsilon=0} \\
&= \theta_0 + \frac{\epsilon}{\sqrt{n}} IF(y, \tilde{T}_\alpha, F_{\theta_0}) + o(\frac{1}{\sqrt{n}}). \qquad (7.43)
\end{aligned}$$

Therefore, for each $j, k = 1, 2$, application of Taylor series expansion gives us

$$\begin{aligned}
M_{j,(\alpha,\tau,\gamma)}\left(\theta_0, \tilde{\theta}_n^*\right) &= M_{j,(\alpha,\tau,\gamma)}\left(\theta_0, \theta_0\right) + \frac{\epsilon}{\sqrt{n}} A_{2,j,(\alpha,\tau,\gamma)}\left(\theta_0, \theta_0\right) IF(y, \tilde{T}_\alpha, F_{\theta_0}) + o\left(\frac{1}{\sqrt{n}}\right), \\
M_{j,(\alpha,\tau,\gamma)}\left(\theta_n^*, \theta_0\right) &= M_{j,(\alpha,\tau,\gamma)}\left(\theta_0, \theta_0\right) + A_{1,j,(\alpha,\tau,\gamma)}\left(\theta_0, \theta_0\right) \frac{\Delta}{\sqrt{n}} + \frac{\epsilon}{\sqrt{n}} A_{1,j,(\alpha,\tau,\gamma)}\left(\theta_0, \theta_0\right) IF(y, T_\alpha, F_{\theta_0}) \\
&+ o\left(\frac{1}{\sqrt{n}}\right), \\
A_{j,k,(\alpha,\tau,\gamma)}\left(\theta_n^*, \tilde{\theta}_n^*\right) &= A_{j,k,(\alpha,\tau,\gamma)}\left(\theta_0, \tilde{\theta}_n^*\right) + o(1).
\end{aligned}$$

We know θ_n^* and $\tilde{\theta}_n^*$ are the unrestricted and restricted MDPD functionals obtained over Θ and Θ_0 respectively. Moreover, we have considered contiguous alternatives. Hence, both $\theta_n^*, \tilde{\theta}_n^* \to \theta_0$ as $n \to \infty$.

Hence, for $j, k = 1, 2$, $M_{j,(\alpha,\tau,\gamma)}\left(\theta_n^*, \theta_0\right) \to M_{j,(\alpha,\tau,\gamma)}\left(\theta_0, \theta_0\right)$, $M_{j,(\alpha,\tau,\gamma)}\left(\theta_0, \tilde{\theta}_n^*\right) \to M_{j,(\alpha,\tau,\gamma)}\left(\theta_0, \theta_0\right)$, $A_{j,k,(\alpha,\tau,\gamma)}\left(\theta_n^*, \theta_0\right) \to A_{j,k,(\alpha,\tau,\gamma)}\left(\theta_0, \theta_0\right)$ and $A_{j,k,(\alpha,\tau,\gamma)}\left(\theta_0, \tilde{\theta}_n^*\right) \to A_{j,k,(\alpha,\tau,\gamma)}\left(\theta_0, \theta_0\right)$ as $n \to \infty$.

Furthermore, $M_{j,(\alpha,\tau,\gamma)}\left(\theta_0, \theta_0\right) = 0$ and $A_{j,k,(\alpha,\tau,\gamma)}\left(\theta_0, \theta_0\right) = (-1)^{j+k} A_\alpha(\theta_0)$.

Thus, we get

$$\begin{aligned}
M_{j,(\alpha,\tau,\gamma)}\left(\theta_n^*,\tilde{\theta}_n^*\right) &= \frac{\epsilon}{\sqrt{n}} A_{2,j,(\alpha,\tau,\gamma)}\left(\theta_0,\theta_0\right) IF(y,\tilde{T}_\alpha,F_{\theta_0}) + o\left(\frac{1}{\sqrt{n}}\right) \\
&+ A_{1,j,(\alpha,\tau,\gamma)}\left(\theta_0,\tilde{\theta}_n^*\right) \frac{\tilde{\Delta}}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right) \\
&= \frac{\epsilon}{\sqrt{n}}(-1)^{2+j} A_\alpha(\theta_0) IF(y,\tilde{T}_\alpha,F_{\theta_0}) + \frac{\tilde{\Delta}}{\sqrt{n}}(-1)^{1+j} A_\alpha(\theta_0) \\
&+ o\left(\frac{1}{\sqrt{n}}\right). \\
\Rightarrow \sqrt{n} M_{j,(\alpha,\tau,\gamma)}\left(\theta_n^*,\tilde{\theta}_n^*\right) &= (-1)^{1+j} A_\alpha(\theta_0)\left[\tilde{\Delta} - \epsilon IF(y,\tilde{T}_\alpha,F_{\theta_0})\right] + o(1) \\
&= (-1)^{1+j} A_\alpha(\theta_0)\tilde{\Delta}^* + o(1).
\end{aligned}$$

A second order Taylor expansion of $Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_n^*}, f_\theta\right)$ around $\theta = \theta_0$ at $\theta = \tilde{\theta}_n^*$ gives

$$\begin{aligned}
Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_n^*}, f_{\tilde{\theta}_n^*}\right) &= Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_n^*}, f_{\theta_0}\right) + \frac{\epsilon}{\sqrt{n}} M_{2,(\alpha,\tau,\gamma)}\left(\theta_n^*,\theta_0\right) IF(y,\tilde{T}_\alpha,F_{\theta_0}) \\
&+ \frac{\epsilon^2}{n} IF(y,\tilde{T}_\alpha,F_{\theta_0})^T A_{2,2,(\alpha,\tau,\gamma)}\left(\theta_n^*,\theta_0\right) IF(y,\tilde{T}_\alpha,F_{\theta_0}) + o\left(\frac{1}{n}\right).
\end{aligned}$$

But we have already shown in Theorem 7.8 that,

$$2n\, Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_n^*}, f_{\theta_0}\right) = \tilde{\Delta}^T A_\alpha(\theta_0)\tilde{\Delta} + o(1).$$

Hence, we can conclude that

$$\begin{aligned}
2n\, Q_{(\alpha,\tau,\gamma)}\left(f_{\theta_n^*}, f_{\tilde{\theta}_n^*}\right) &= \tilde{\Delta}^T A_\alpha(\theta_0)\tilde{\Delta} + 2\sqrt{n}\epsilon M_{2,(\alpha,\tau,\gamma)}\left(\theta_n^*,\theta_0\right) IF(y,\tilde{T}_\alpha,F_{\theta_0}) \\
&+ \epsilon^2 IF(y,\tilde{T}_\alpha,F_{\theta_0})^T A_{2,2,(\alpha,\tau,\gamma)}\left(\theta_n^*,\theta_0\right) IF(y,\tilde{T}_\alpha,F_{\theta_0}) + o(1) \\
&= \tilde{\Delta}^T A_\alpha(\theta_0)\tilde{\Delta} + 2\epsilon(-1)^3 IF(y,\tilde{T}_\alpha,F_{\theta_0})^T A_\alpha(\theta_0)\tilde{\Delta}^* + o(1) \\
&+ \epsilon^2 IF(y,\tilde{T}_\alpha,F_{\theta_0})^T(-1)^4 A_\alpha(\theta_0) IF(y,\tilde{T}_\alpha,F_{\theta_0}) + o(1) \\
&= \left[\tilde{\Delta} - \epsilon IF(y,\tilde{T}_\alpha,F_{\theta_0})\right]^T A_\alpha(\theta_0)\left[\tilde{\Delta} - \epsilon IF(y,\tilde{T}_\alpha,F_{\theta_0})\right] \\
&= \tilde{\Delta}^{*T} A_\alpha(\theta_0)\tilde{\Delta}^* + o(1).
\end{aligned}$$

Finally, we have derived the following

$$\begin{aligned}
2nQ_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\tilde{\theta}_\alpha}\right) &= \tilde{\Delta}^{*T}A_\alpha(\theta_0)\tilde{\Delta}^* + 2\tilde{\Delta}^{*T}A_\alpha(\theta_0)\sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right) \\
&- 2\tilde{\Delta}^{*T}A_\alpha(\theta_0)\sqrt{n}\left(\tilde{\theta}_\alpha - \tilde{\theta}_n^*\right) + \left(\tilde{\theta}_\alpha - \tilde{\theta}_n^*\right)^T A_\alpha(\theta_0)\left(\tilde{\theta}_\alpha - \tilde{\theta}_n^*\right) \\
&- 2\sqrt{n}\left(\tilde{\theta}_\alpha - \tilde{\theta}_n^*\right)^T A_\alpha(\theta_0)\sqrt{n}\left(\hat{\theta}_\alpha - \tilde{\theta}_n^*\right) + \sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right)^T A_\alpha(\theta_0) \\
&\quad \sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right) + o_p(1) + o(1) \\
&= \left[\tilde{\Delta}^* + \sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right) - \sqrt{n}\left(\tilde{\theta}_\alpha - \tilde{\theta}_n^*\right)\right]^T A_\alpha(\theta_0) \\
&\quad \left[\tilde{\Delta}^* + \sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right) - \sqrt{n}\left(\tilde{\theta}_\alpha - \tilde{\theta}_n^*\right)\right] + o_p(1) + o(1),
\end{aligned}$$

because $n \times o\left(||\hat{\theta}_\alpha - \theta_n^*||^2\right) = o_p(1)$ and $n \times o\left(||\tilde{\theta}_\alpha - \tilde{\theta}_n^*||^2\right) = o_p(1)$, which follows from the asymptotic distribution of the MDPDE and the RMDPDE.

Hence, under $F^P_{n,\epsilon,y}$, as $n \to \infty$,

$$\tilde{T}_{(\alpha,\tau,\gamma)}\left(\hat{\theta}_\alpha, \tilde{\theta}_\alpha\right) \overset{\mathrm{D}}{=} W^T A_\alpha(\theta_0) W,$$

where,

$$W \overset{\mathrm{D}}{=} \left[\tilde{\Delta}^* + \sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right) - \sqrt{n}\left(\tilde{\theta}_\alpha - \tilde{\theta}_n^*\right)\right].$$

Just like the null case given in Theorem 7.13, here also, one can prove that,

$$\tilde{\Delta}^* + \sqrt{n}\left(\hat{\theta}_\alpha - \theta_n^*\right) - \sqrt{n}\left(\tilde{\theta}_\alpha - \tilde{\theta}_n^*\right) \overset{\mathrm{a}}{\sim} N\left(\tilde{\Delta}^*, \tilde{\Sigma}_\alpha(\theta_0)\right),$$

where, $\tilde{\Sigma}_\alpha(\theta_0)$ is given in Equation (7.37).

Hence, the proof of the first statement of part (i) is complete. The next part immediately follows from Lemma 7.2. This is quite similar to the proof of the last statement of part (i) of Theorem 7.8, where we were deriving the asymptotic distribution of our test statistic under $F^P_{n,\epsilon,y}$ in case of a simple null hypothesis.

The part (ii) follows from the infinite series expansion of a linear combination of non-central chi-squares, in terms of the central χ^2 distribution as derived in Kotz et al. (1967a). We have already followed the same thing in case of deriving the asymptotic power in case of a simple null hypothesis and hence omitted it here. □

- In case of $\epsilon = 0$, the expression of the asymptotic power will remain the same except $\tilde{C}^\alpha_\nu(\theta_0, \tilde{\Delta}^*)$ will be replaced by $\tilde{C}^\alpha_\nu(\theta_0, \Delta)$.
- In case of $\Delta = 0$, the asymptotic level under $F^L_{n,\epsilon,y}$ as

$$\begin{aligned}\tilde{P}(\Delta = 0, \epsilon) &= \sum_{\nu=0}^{\infty} \tilde{C}^\alpha_\nu \left(\theta_0, \epsilon \left(IF(y, T_\alpha, F_{\theta_0}) - IF(y, \tilde{T}_\alpha, F_{\theta_0})\right)\right) \\ &\qquad P\left(\chi^2_{r+2\nu} > \frac{\tilde{t}^{(\alpha,\tau,\gamma)}_\beta}{\tilde{\eta}_{(1)}}\right).\end{aligned}$$

- In case of $\Delta = 0$ and $\epsilon = 0$, the asymptotic distribution of the GSDT under $F^P_{n,\epsilon,y}$ coincides with the asymptotic distribution of the GSDT under the null hypothesis defined in Theorem 7.3.

Following the procedure adopted in case of the simple null hypotheses, we can now find the expressions of the PIF and the LIF as follows

$$\begin{aligned}& PIF\left(y, \tilde{T}^{(1)}_{(\alpha,\tau,\gamma)}, F_{\theta_0}\right) \\ =\ & \frac{\partial}{\partial\epsilon}\tilde{P}(\Delta, \epsilon)\bigg|_{\epsilon=0} = \sum_{\nu=0}^{\infty} \frac{\partial}{\partial\epsilon}\tilde{C}^\alpha_\nu(\theta_0, \tilde{\Delta}^*)\bigg|_{\epsilon=0} P\left(\chi^2_{r+2\nu} > \frac{\tilde{t}^{(\alpha,\tau,\gamma)}_\beta}{\tilde{\eta}_{(1)}}\right) \\ =\ & \left(IF(y, T_\alpha, F_{\theta_0}) - IF(y, \tilde{T}_\alpha, F_{\theta_0})\right)\left(\sum_{\nu=0}^{\infty} \frac{\partial}{\partial\epsilon}\tilde{C}^\alpha_\nu(\theta_0, t)\bigg|_{t=\Delta} P\left(\chi^2_{r+2\nu} > \frac{\tilde{t}^{(\alpha,\tau,\gamma)}_\beta}{\tilde{\eta}_{(1)}}\right)\right), \\ & LIF\left(y, \tilde{T}^{(1)}_{(\alpha,\tau,\gamma)}, F_{\theta_0}\right) \\ =\ & \left(IF(y, T_\alpha, F_{\theta_0}) - IF(y, \tilde{T}_\alpha, F_{\theta_0})\right)\left(\sum_{\nu=0}^{\infty} \frac{\partial}{\partial\epsilon}\tilde{C}^\alpha_\nu(\theta_0, t)\bigg|_{t=0} P\left(\chi^2_{r+2\nu} > \frac{\tilde{t}^{(\alpha,\tau,\gamma)}_\beta}{\tilde{\eta}_{(1)}}\right)\right),\end{aligned}$$

for any bounded IF of MDPDE under the Lehmann and Basu et al. conditions (B1)-(B7).

7.5 Simulation Study

Here, we provide some extensive numerical evidence of the performance of our proposed tests by demonstrating their strong robustness properties. The test statistic is dependent on the data through the MDPDE, $\hat{\theta}_\alpha$, and hence, the robustness of the test depends on the MDPDE as well as the three associated tuning parameters, α, τ and γ of the GSD. Throughout this section, we consider the univariate normal distribution with mean μ and variance σ^2 to explore the performance of our proposed test statistics for testing a hypothesis about μ. Based on a sample of size n from this population, we want to test $H_0 : \mu = 0$ against omnibus alternatives for both cases of sigma being known (simple hypotheses) and being unknown (composite hypotheses) respectively, and to study its power, we consider several alternative values of μ. In this scenario, whenever σ is known,

$$\frac{T_{(\alpha,\tau,\gamma)}(\hat{\mu}_\alpha, \mu_0)}{\lambda} \overset{a}{\sim} \chi^2_1$$

as $n \to \infty$, where, $\hat{\mu}_\alpha$ denotes the MDPDE of μ at a fixed α value, μ_0 denotes the null hypothesized value and λ = eigenvalue of $A(\mu_0)J^{-1}(\mu_0)K(\mu_0)J^{-1}(\mu_0) = \frac{(1+\alpha)^{5/2}}{(1+2\alpha)^{3/2}(2\pi)^{\alpha/2}\sigma^\alpha}$. For $\alpha = 0$, λ becomes 1. The observed level or power will be evaluated as the proportion of the derived test statistics, derived based on the samples of sizes n = 20, 50 and 100, being larger than $\chi^2_{0.05,1} = 3.84146$ in a number of 1000 replications.

To analyze a pure case scenario, we consider the model $N(0,1)$ under null hypothesis and the models $N(0.5,1)$ and $N(1,1)$ under alternative hypothesis and similarly for a contaminated case scenario, we consider the model $(1-\epsilon)\ N(0,1) + \epsilon\ N(5,1)$ under null hypothesis and the models $(1-\epsilon)\ N(0.5,1) + \epsilon\ N(-5,1)$ and $(1-\epsilon)\ N(1,1)$ $+\ \epsilon\ N(-9,1)$ under alternative hypothesis, with the value of ϵ being 0.10. The asymptotic level is calculated under several combinations of (α,τ,γ); more specifically, we have taken some standard choices of $\alpha = \{0.15, 0.25, 0.5, 0.75, 0.9, 1\}$ and τ is well spread over the $(0,1)$ interval, that is, $\{0.05, 0.15, 0.30, 0.5, 0.7, 0.85, 0.95\}$, but the result is symmetric with respect to $\tau = 0.5$, so we omit the case of $\tau > 0.5$. In case of γ, we already know that $\gamma \to -1$ generates the performance of the SDT, introduced by Ghosh et al. (2015). So, we have decided to consider $\gamma = \{-10, -5, -1, 0, 1, 5, 10\}$ to explore the performance of the GSDT.

Till now, we have fixed the model and the sample size to analyze its performance under fixed alternatives varying over tuning parameters, but now we are going to thoroughly illustrate the case of contiguous alternatives to study its performance depending on several models and different sample sizes. Here, we consider, under alternative $H_{1,n}$: $\mu = \mu_1 = \mu_0 + \frac{\Delta}{\sqrt{n}}$ for $\Delta = \sqrt{5}, \mu_0 = 0$ and different sample sizes $n = 20, 50, 100$. We have studied its performance under this setup for both σ being known and unknown. The graphical presentation of the p-values with respect to the given choices of the tuning parameters and sample size n are given in Figures 7.1–7.24 for fixed alternatives and in Figures 7.25–7.32 for contiguous alternatives.

7.5.1 Some Observations

(i) For given values of n and α, if τ is fixed at a very small value and we observe the variation w.r.t. γ then both level and power decreases as γ approaches 0. On the other hand, as γ goes far from 0 on the either side of the real line, there is an increment in both level and power.

(ii) For given values of n and α, if τ is fixed at a moderate to relatively large value and we observe the variation w.r.t. γ then both level and power increases as γ approaches 0. On the other hand, as γ goes far from 0 on the either side of the real line, there is a decrement in both level and power.

(iii) Keeping the values of n and α fixed and γ being far from 0 on either side of real line, if we concentrate on how τ varies, we will observe that both level and power diminishes as τ increases to 0.5, but then rises as τ further increases to 1. Since the pattern of increment and decrement of level/power are essentially the same on either side of $\tau = 0.5$, we have omitted the observation for the $\tau > 0.5$ case. But this change w.r.t. τ is almost negligible when γ approaches 0.

(iv) If the values of n, γ and τ are fixed and we look at the common parameter of the SDT, the DPDT and the GSDT, that is, α, we can then observe that small to moderately large α's are giving reasonable results.

(v) If the values of α, τ and γ are fixed, then as the sample size n increases, the empirical level and power exhibit opposing behaviour. On one hand, the level decreases, but on the other, the power increases and this is very much desirable to us.

(vi) If we consider a pure model where σ is known to us and $n = 100$, then the empirical level almost reaches the nominal level of 0.05 whenever $\gamma \in [-1, 1]$, irrespective of the value of τ.

(vii) Again, for a pure model with σ-known case, the empirical levels corresponding to $\tau < 0.1$ are significantly greater than the nominal level value of 0.05. Also, if we consider both the pure and contaminated models and when σ is either known or unknown to us, then corresponding to $\tau < 0.1$, the empirical power attains quite high values which must be a consequence of underestimating the true cut-off values.

(viii) Keeping α fixed, the surface plots, corresponding to various the pairs of (τ, γ), are gently sloping downwards with the increment in the sample size n.

(ix) For fixed α, τ, γ and n, the empirical level and power are larger in the σ-unknown case, as compared to the known case. If we consider the pure and contaminated model, then the contaminated one will give us larger level and smaller power values, as compared to the pure model. When we consider both fixed and contiguous alternatives, then in case of the contiguous one, the values of the power are less.

(x) In case of contiguous alternative, the power remains almost constant whenever $\gamma \in [-1, 1]$ and for varying values of the sample size n. This picture becomes clearer as τ increases in $[0.1, 0.5]$.

(xi) In case of contiguous alternative, the pattern of change of power with respect to varying α (keeping remaining parameters fixed) is slightly different from the case of fixed alternative (mentioned in (iii) above). Here, as α increases, the empirical power decreases uniformly. The scenarios with respect to varying γ values (mentioned in (i) and (ii)) and τ (mentioned in (iii)) are quite similar to the case of fixed alternative.

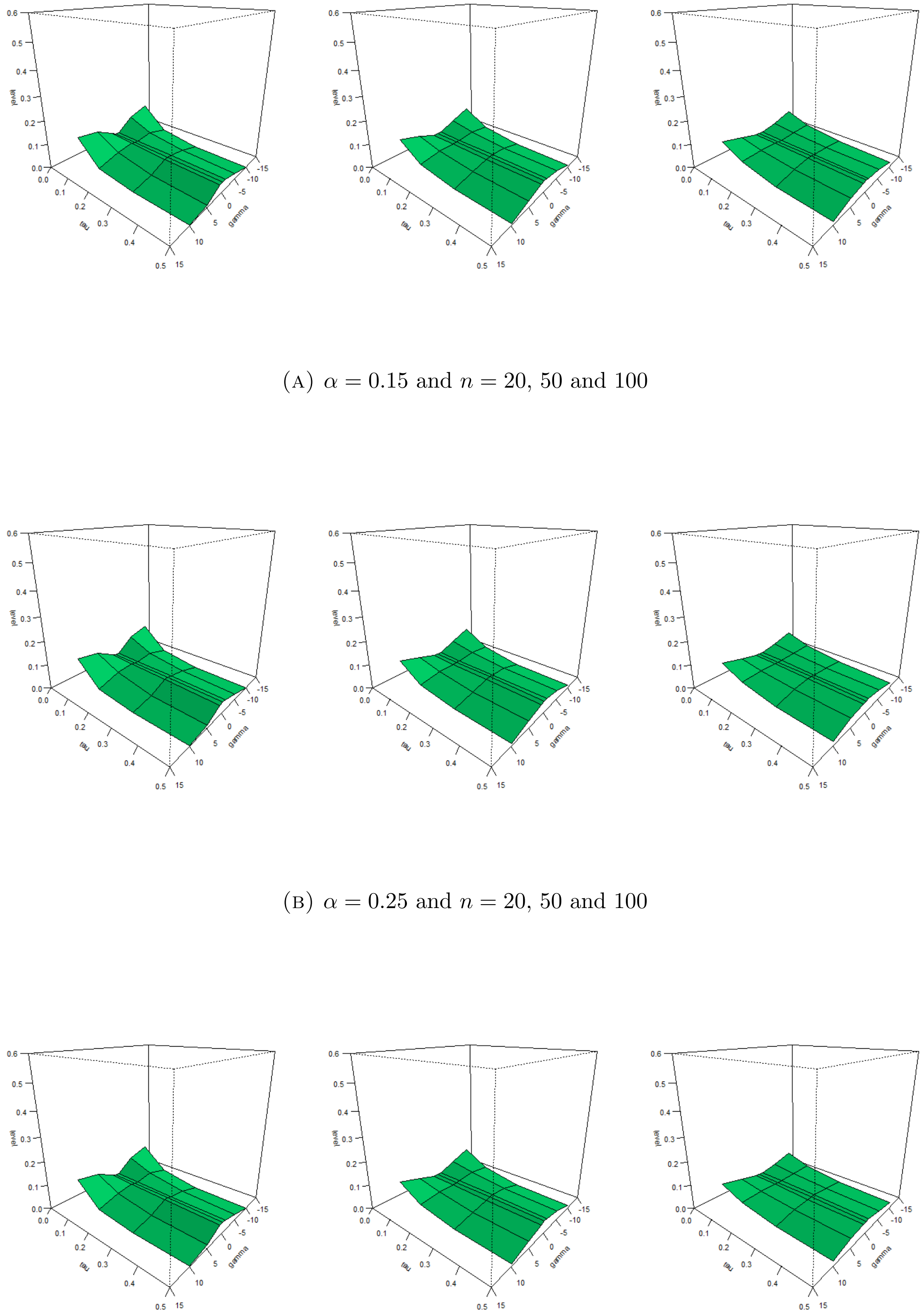
Based on these observations, we can conclude that for $\alpha \in [0.1, 0.5]$, $\tau \in [0.1, 0.5]$ and $\gamma \in [-1, 1]$, the empirical level and power are quite stable – for the pure model, the powers are found to be quite satisfactory and competitive with SDTs and DPDTs and the sizes are quite close to the nominal level 0.05, even more closer than the class of SDTs and DPDTs. On the other hand, for the contaminated model, there are many members of the GSDTs within this region which give more robust size along with almost the same power as generated by the 'best' SDTs or DPDTs. Therefore, our preferable best region is approximately a cuboid with $\alpha \in [0.1, 0.5]$, $\tau \in [0.1, 0.5]$ and $\gamma \in [-1, 1]$, in the sense that it expands the choice of robust tests lying outside the class of DPDTs or SDTs, which maintain levels close to the nominal one with high power (in pure model) and show robust characteristics in generating size along with satisfactory power in the presence of contamination.

7.5.2 Comparison Among Some 'Best' DPDTs–SDTs–GSDTs and the LRT

To follow up on the previous discussion, we have further arranged some specific plots to make a comparison of the DPDTs and the SDTs belonging to the 'best' region (as declared by Basu et al. (2011) and Ghosh et al.(2015) respectively) with some of the 'best' GSDTs in terms of level and power, by varying over sample sizes which range from small ($n = 5$) to large ($n = 100$). Moreover, we have taken the LRT into account for this purpose. We have made a setup of testing $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ with both cases of σ being known and unknown. For studying the observed level, the data have been generated from the $N(0, 1)$ distribution, whereas, to generate power, we have taken the sample from the $N(0.5, 1)$ population. In either case, the level of significance is 0.05. This pure model scenario has been constructed for efficiency purposes, but to illustrate the robustness, we need to consider the presence of contamination – the same tests are to be repeated under the mixture of contaminated sample observations taken from $N(-2, 1)$ for studying observed level, whereas, we are to sample from $N(-6, 1)$ for studying observed power. In both cases, the proportion of contamination (ϵ) is 0.1. In such scenarios, the derived results are given in Figure 7.33 and 7.34.

If we look at the pure model scenario, we can then observe that all the robust tests, along with the t-test, have almost the same power, but when we consider the level, it is not so. There are several combinations among our chosen one, which produce less size than the t-test; among them, the GSDT with $\tau = 0.65$ and $\gamma = 1$ has the least size for the σ-known case and on the other hand, the same thing is observed in case of the GSDT with $\tau = 0.15$ and $\gamma = 1$ for the σ-unknown case. It is to be noted that both the cases here belong to the best region that we have mentioned earlier. Overall, the proposed tests (within the preferred region) appear to be quite competitive with other tests.

To study the stability of the power, we now enlighten ourselves of the contaminated case. Here, the GSDTs are quite resistant. The scenario with respect to the LRT is totally different under contamination, since its non-robust nature has been revealed – its level is getting increased with n, whereas, the power is getting decreased. Here also, for moderate to large sample sizes, all robust tests perform quite better with almost the same power and their powers are significantly larger than that of the t-test, but if we concentrate on the size of the tests, the conclusion will be same as in the pure case.

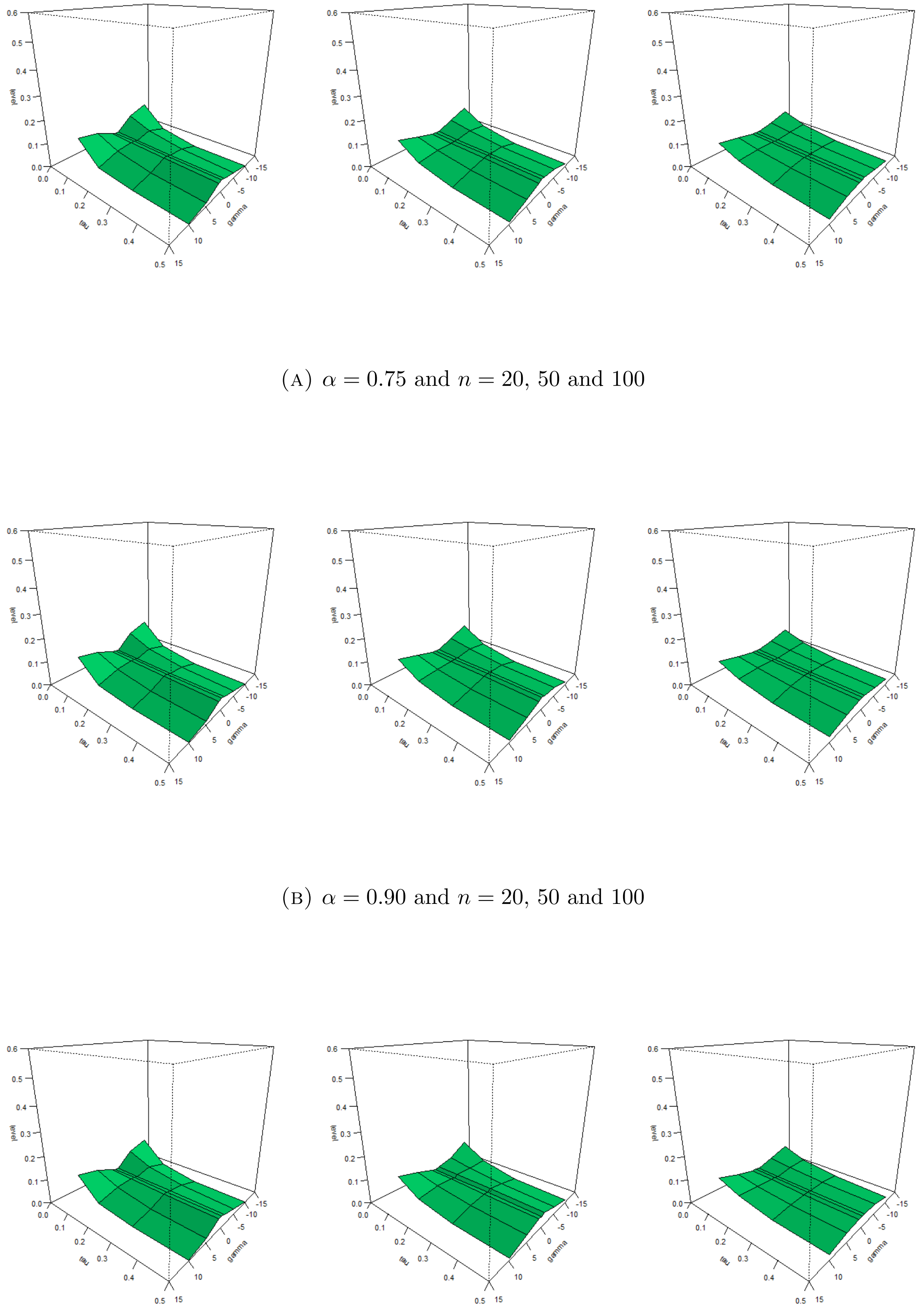


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.1: Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0, 1)$ model)

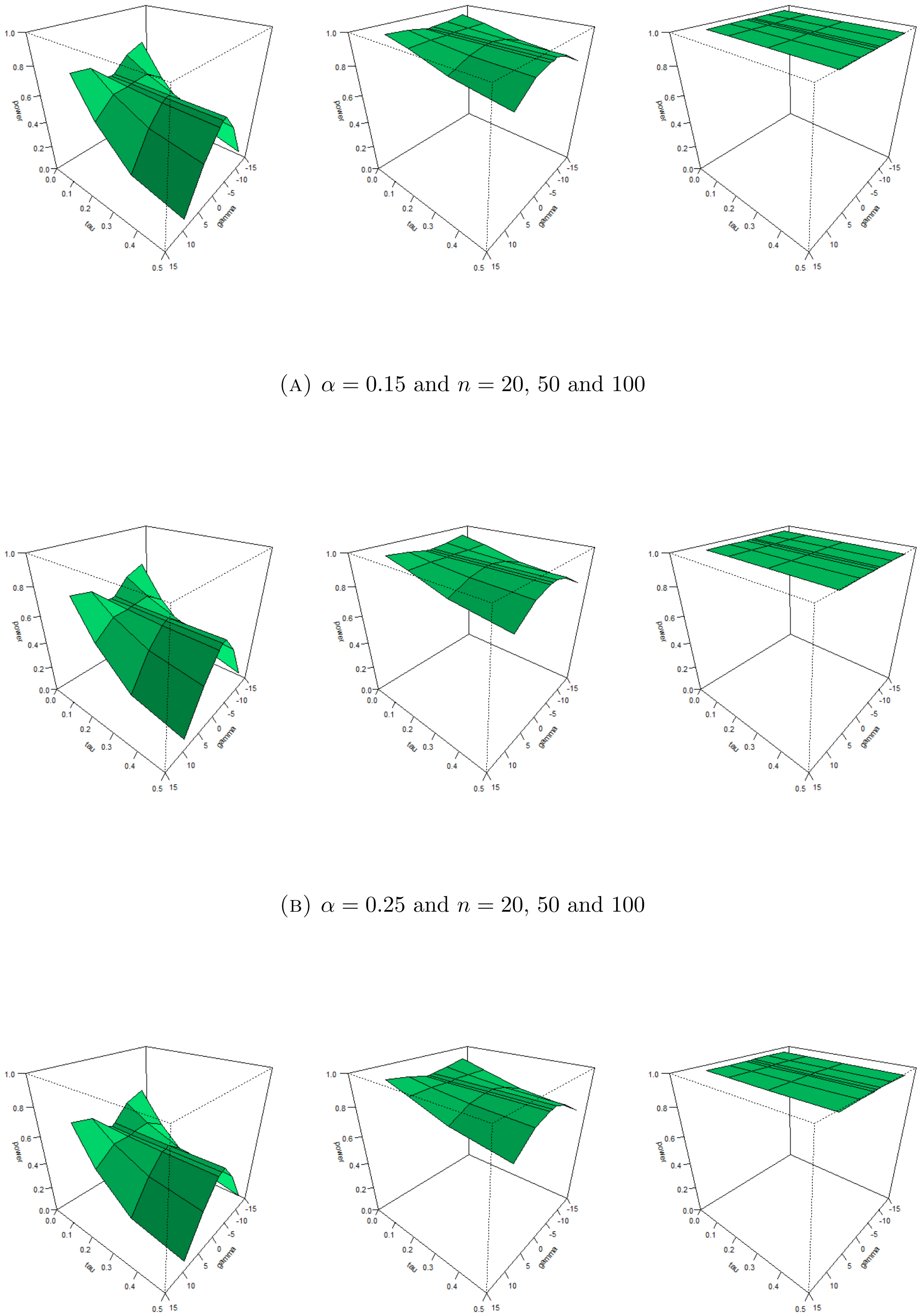


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.2: (Continued) Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0,1)$ model)

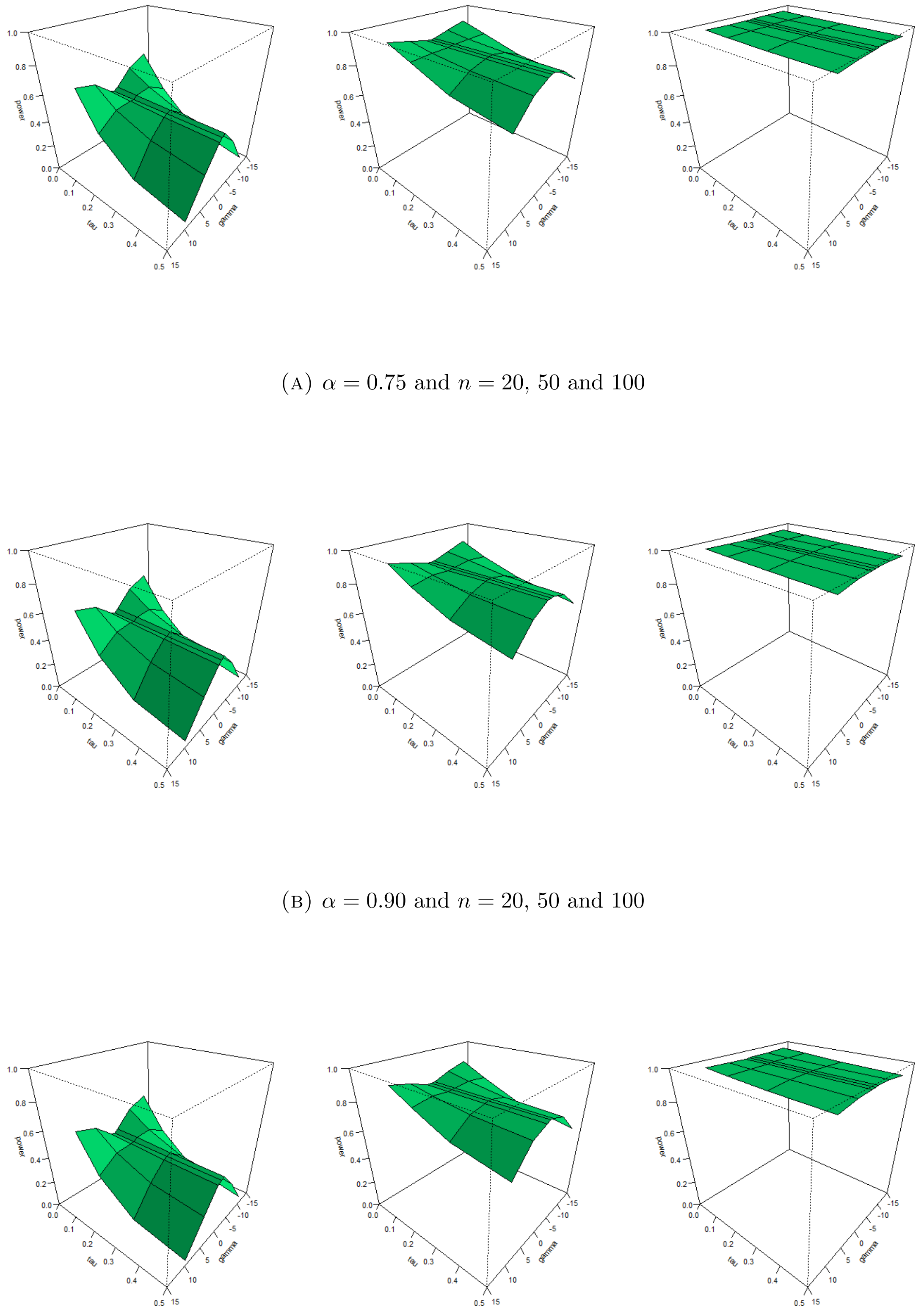


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.3: Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0.5, 1)$ model)

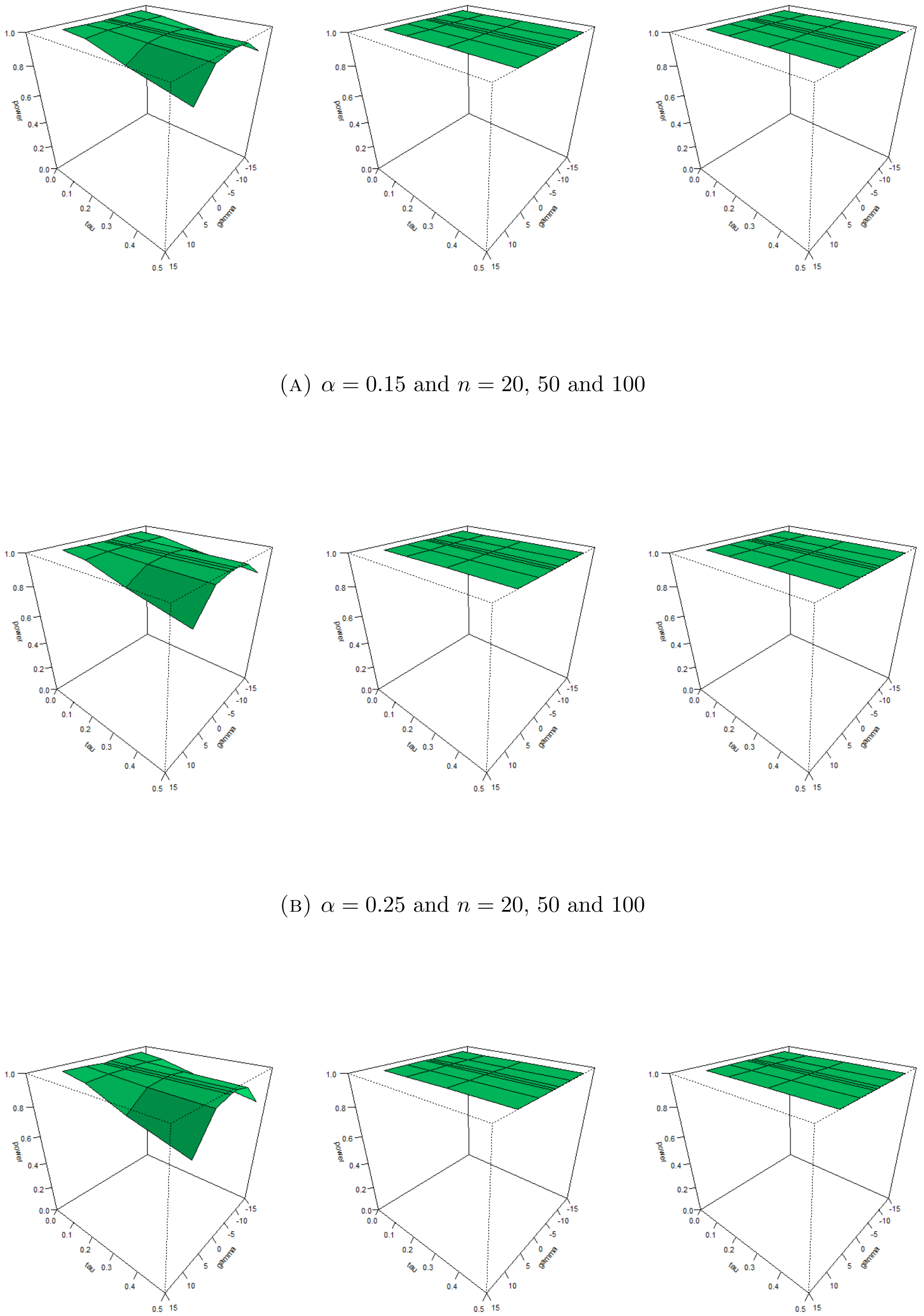


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.4: (Continued) Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0.5, 1)$ model)

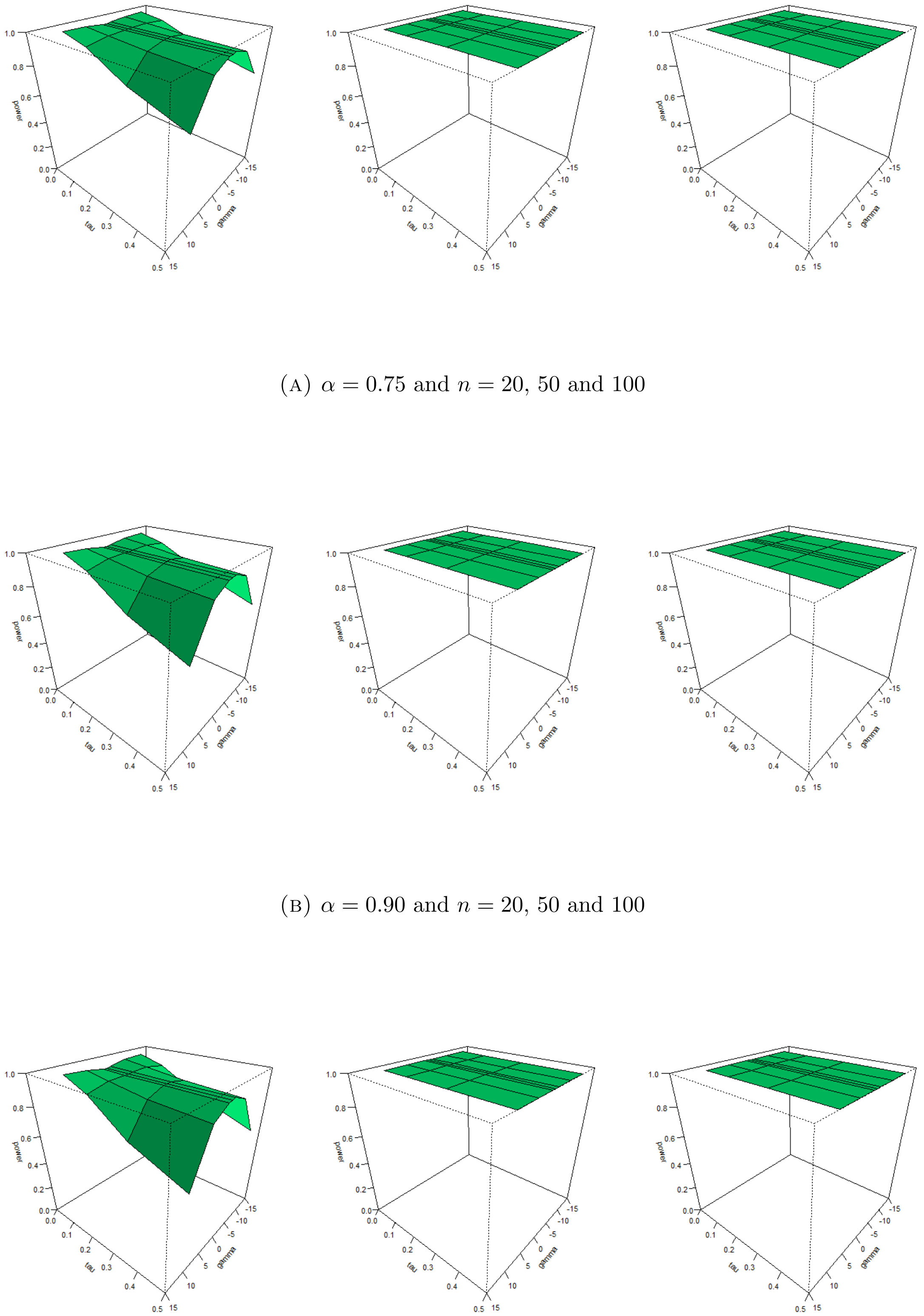


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.5: Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(1, 1)$ model)

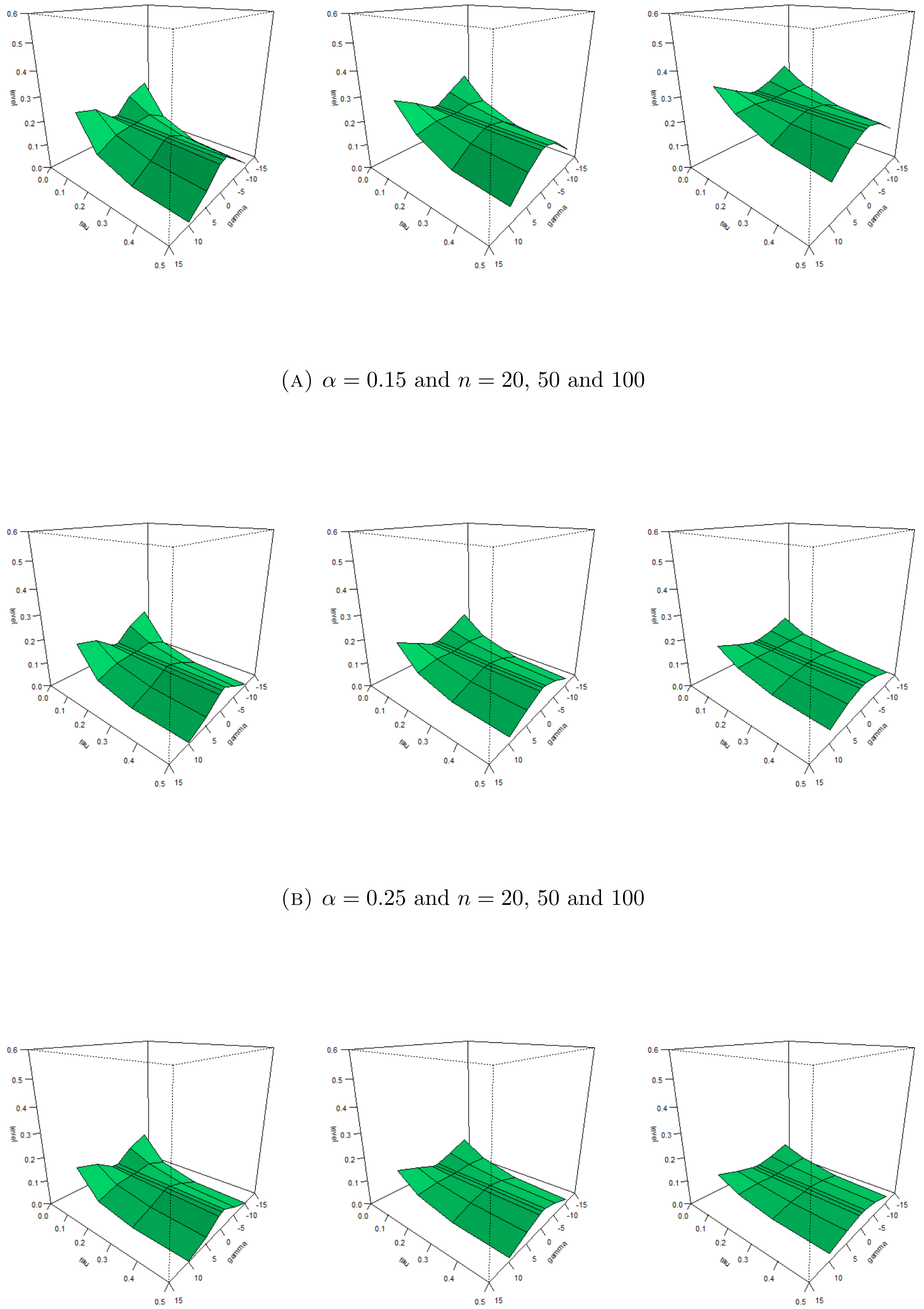


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.6: (Continued) Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(1, 1)$ model)

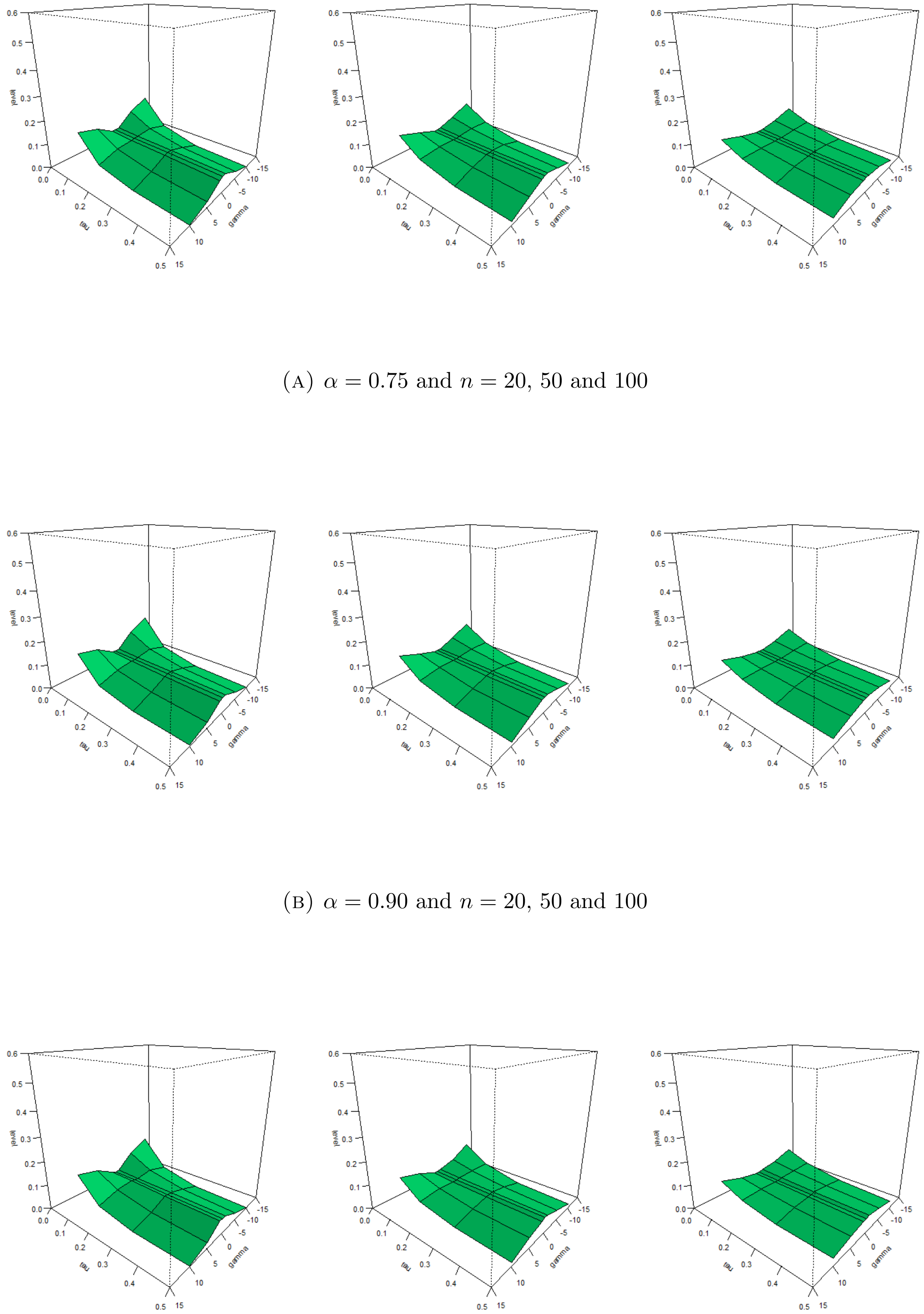


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.7: Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(0,1) + 0.1\ N(5,1)$ model)

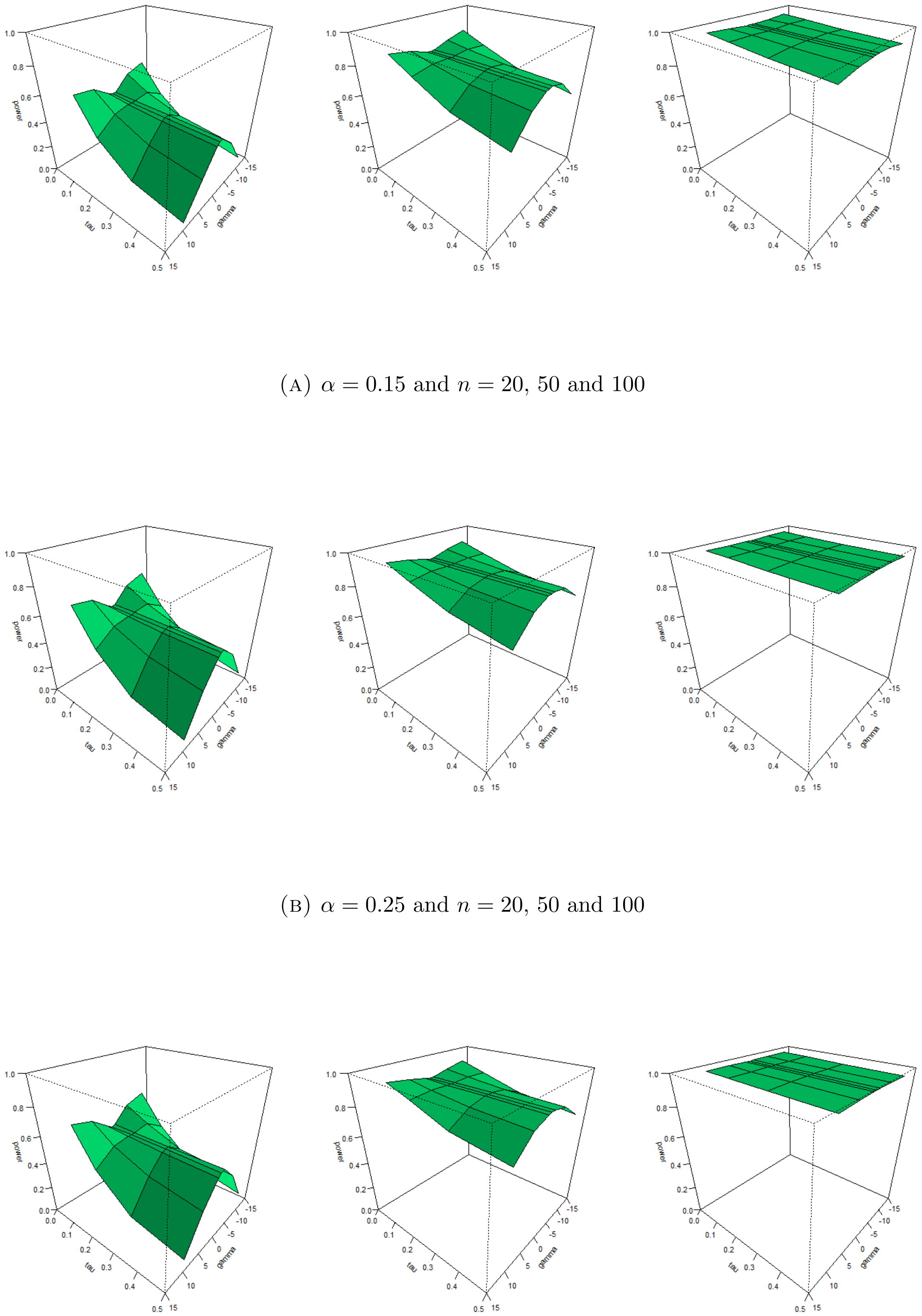


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.8: (Continued) Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(0,1) + 0.1\ N(5,1)$ model)

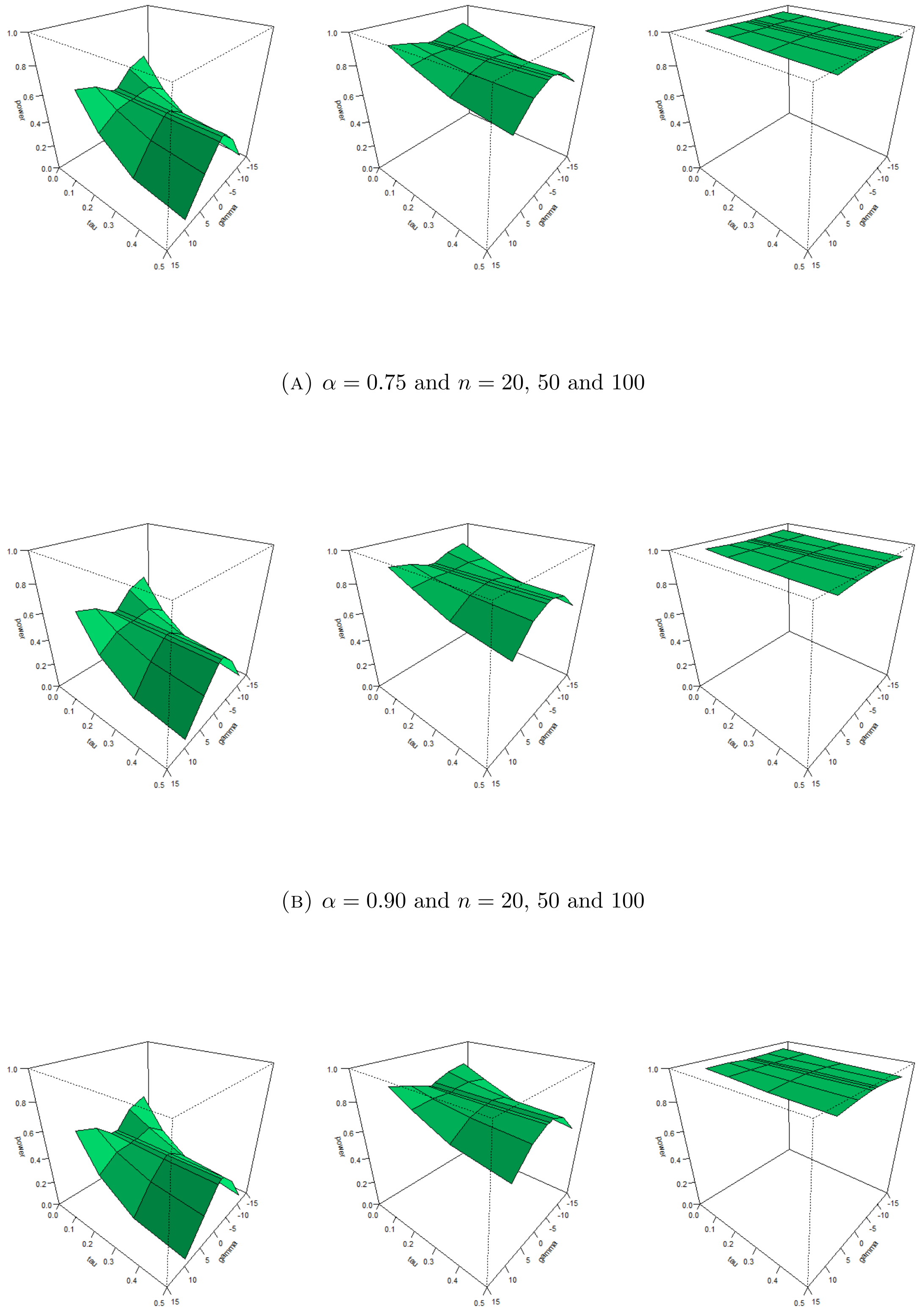


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.9: Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(0.5, 1) + 0.1\ N(-5, 1)$ model)

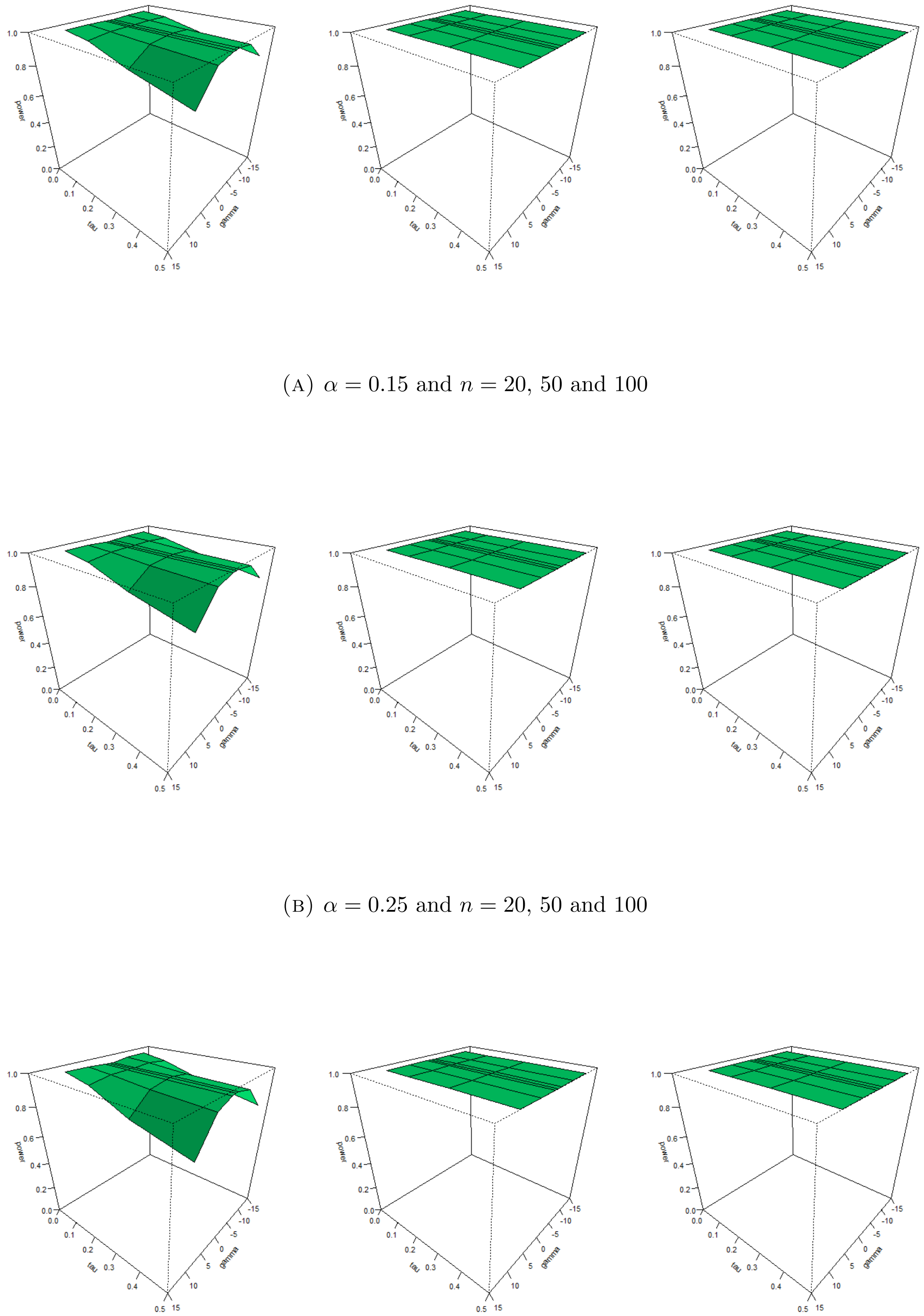


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.10: (Continued) Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(0.5, 1) + 0.1\ N(-5, 1)$ model)

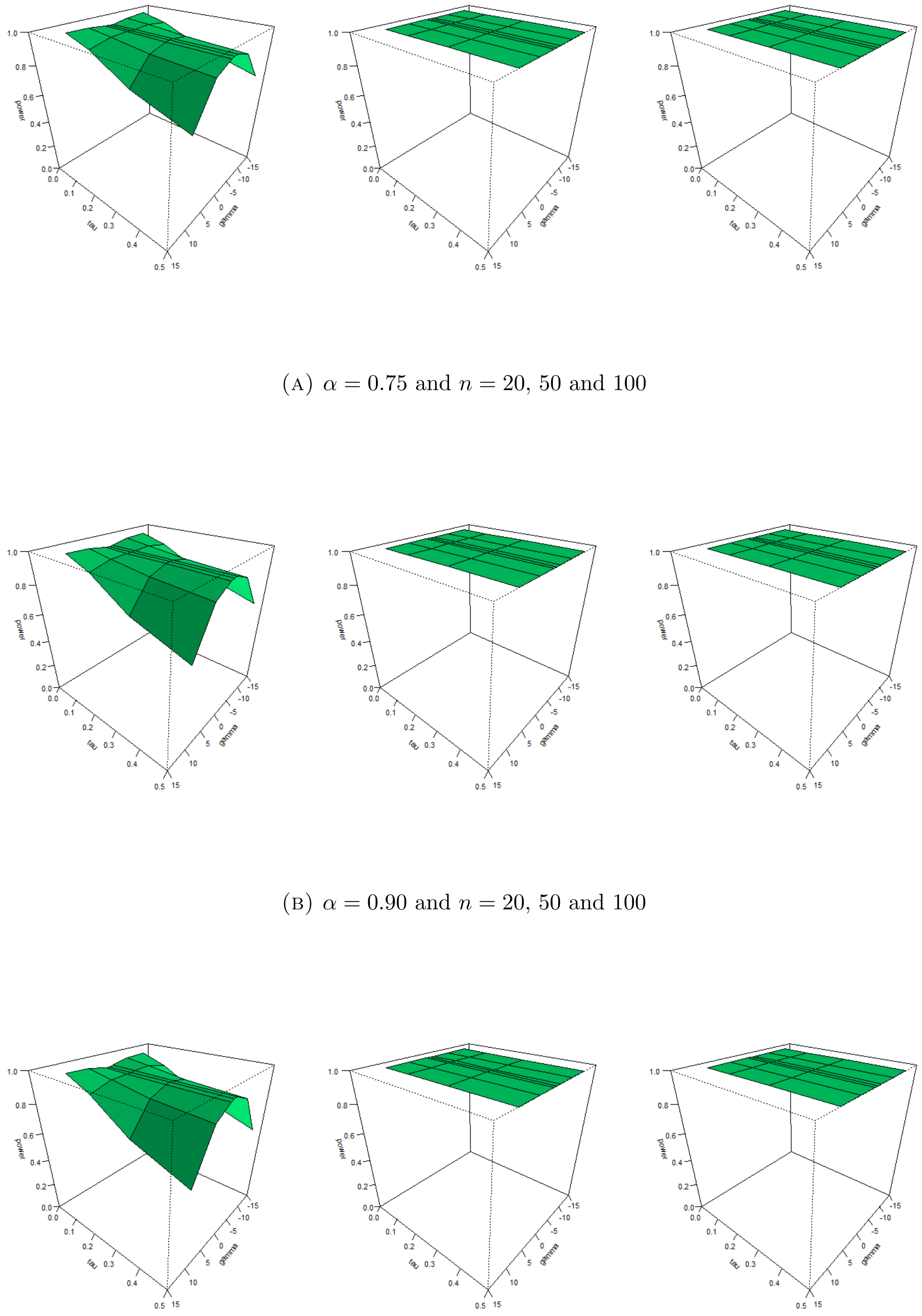


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.11: Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(1,1) + 0.1\ N(-9,1)$ model)

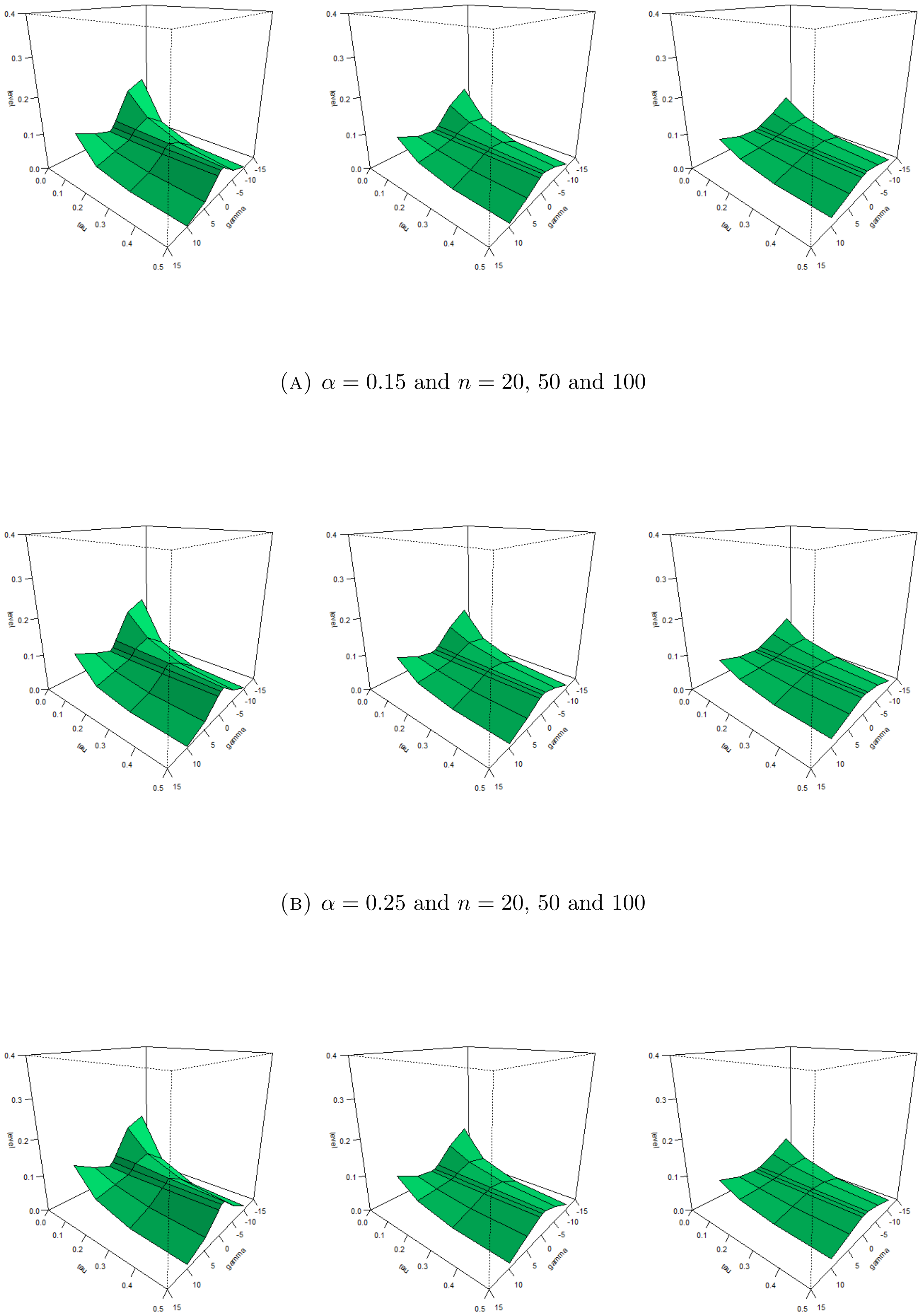


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.12: (Continued) Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(1, 1) + 0.1\ N(-9, 1)$ model)3

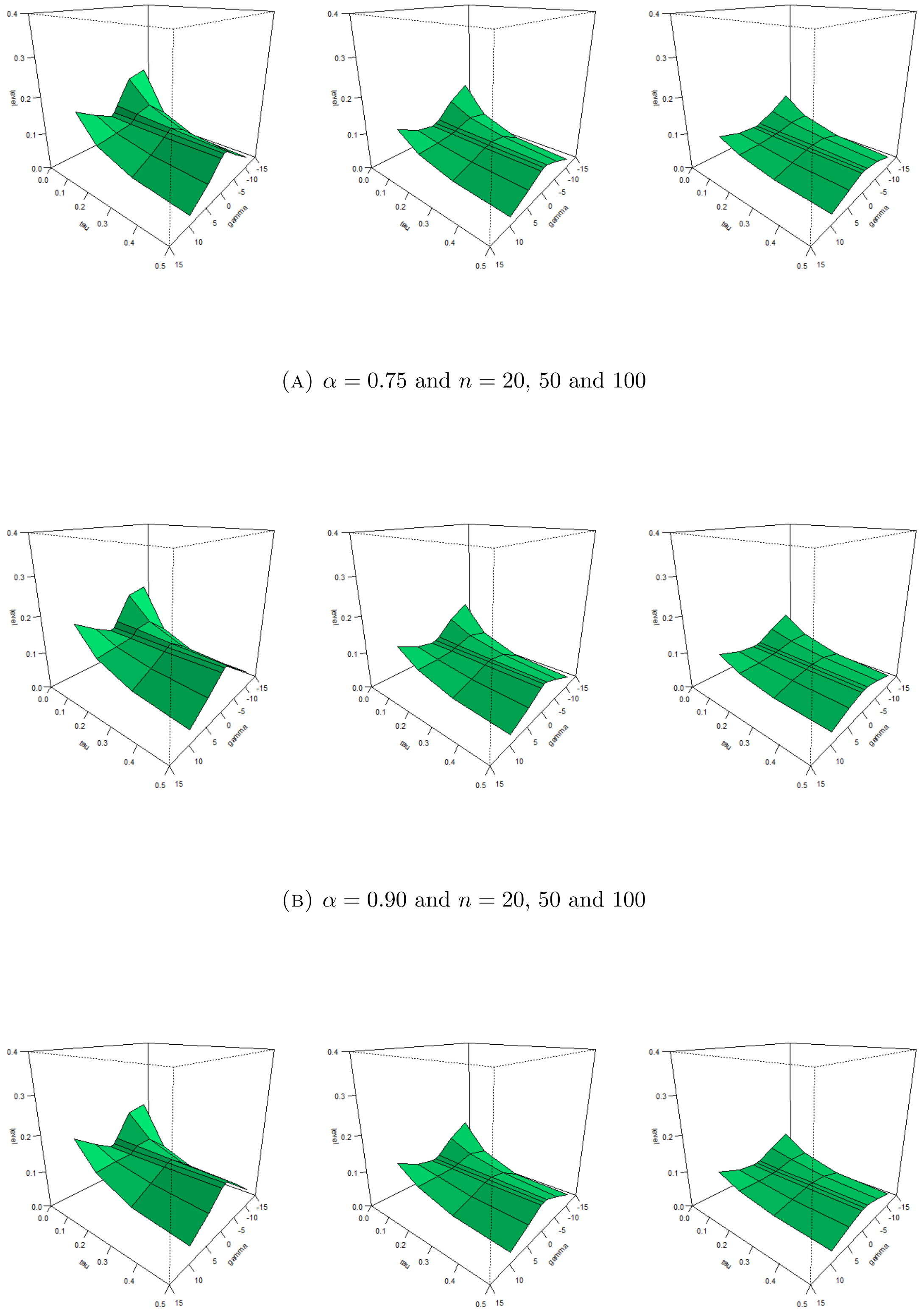


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.13: Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0, 1)$ model)

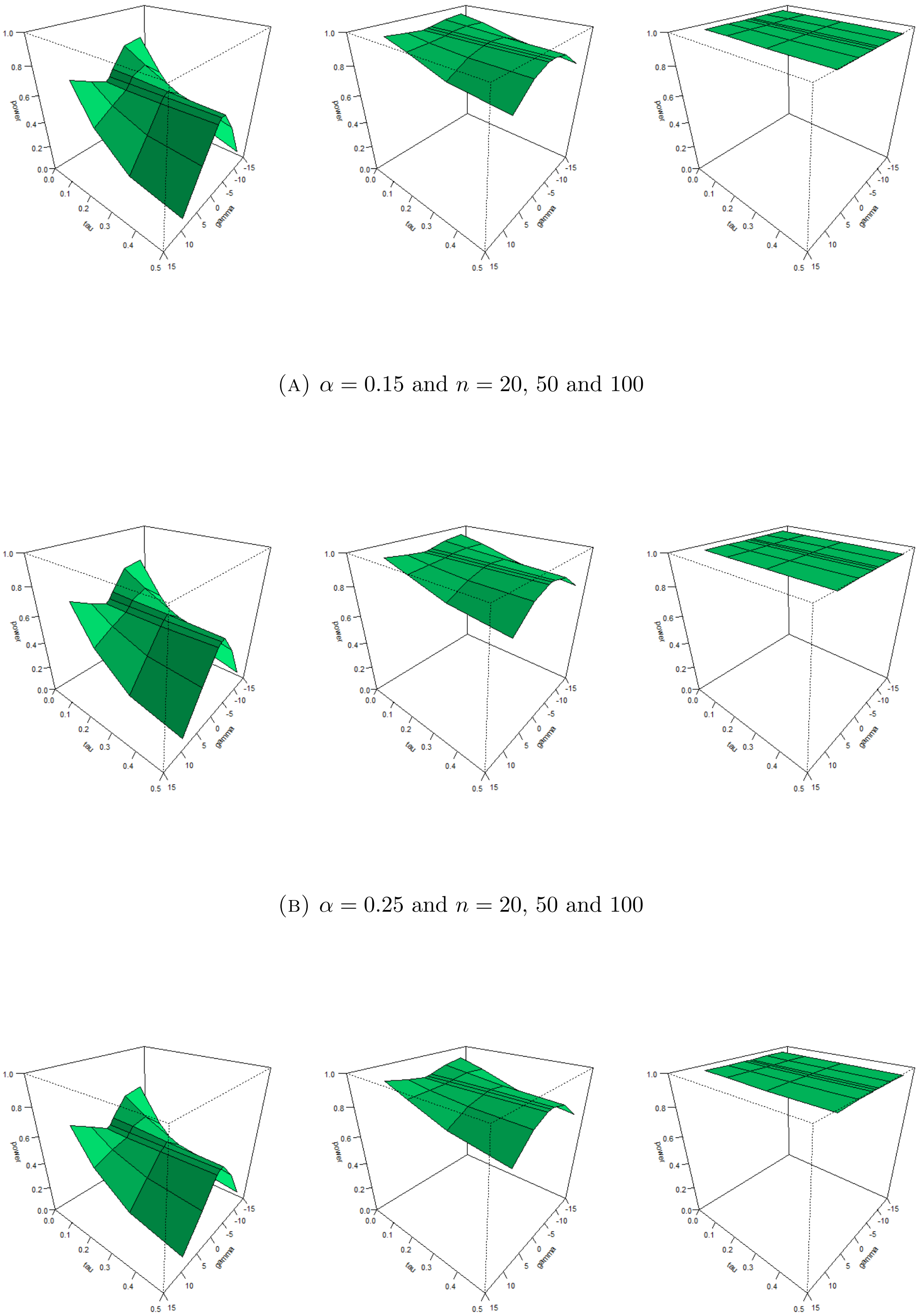


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.14: (Continued) Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0, 1)$ model)

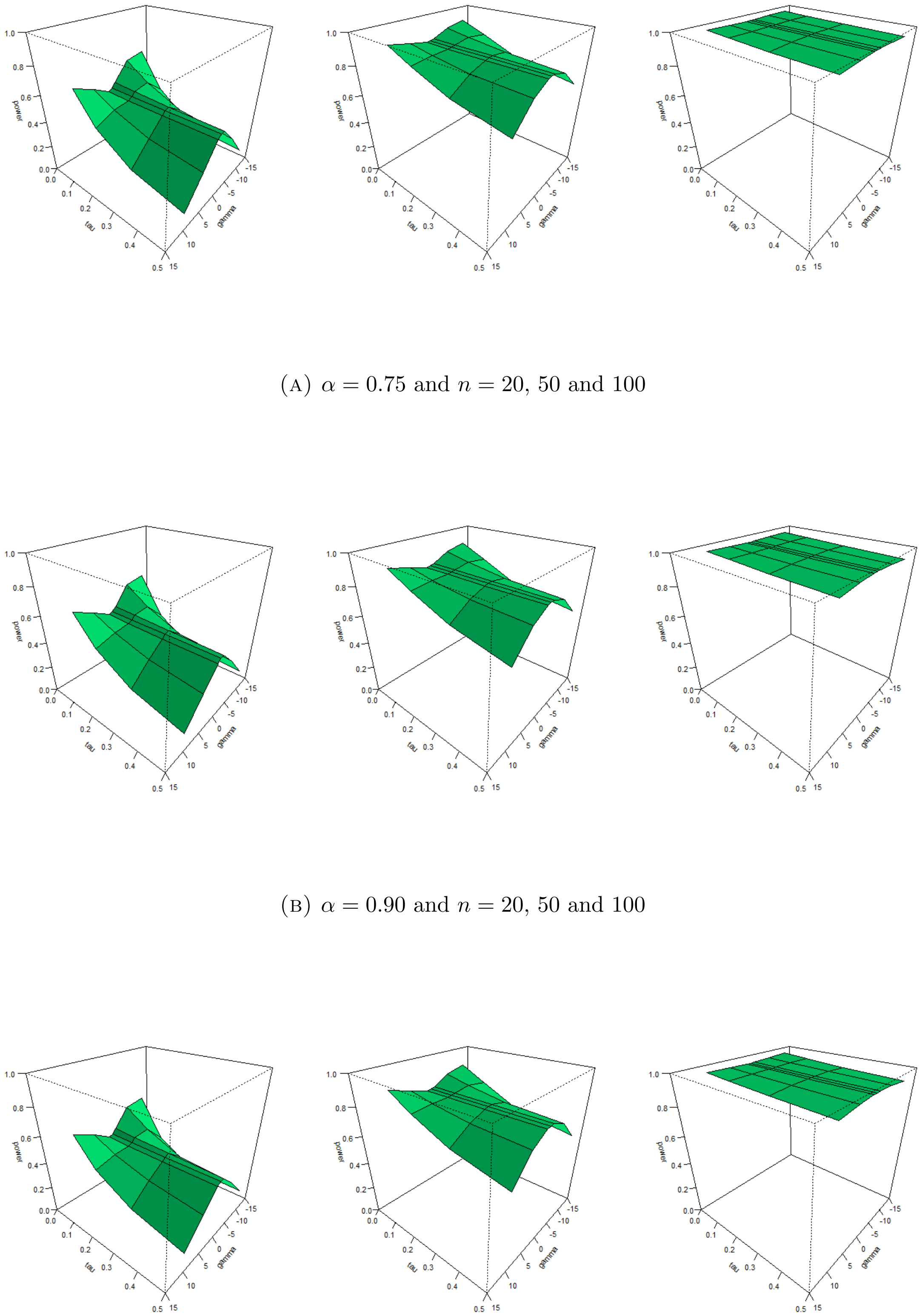


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.15: Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0.5, 1)$ model)

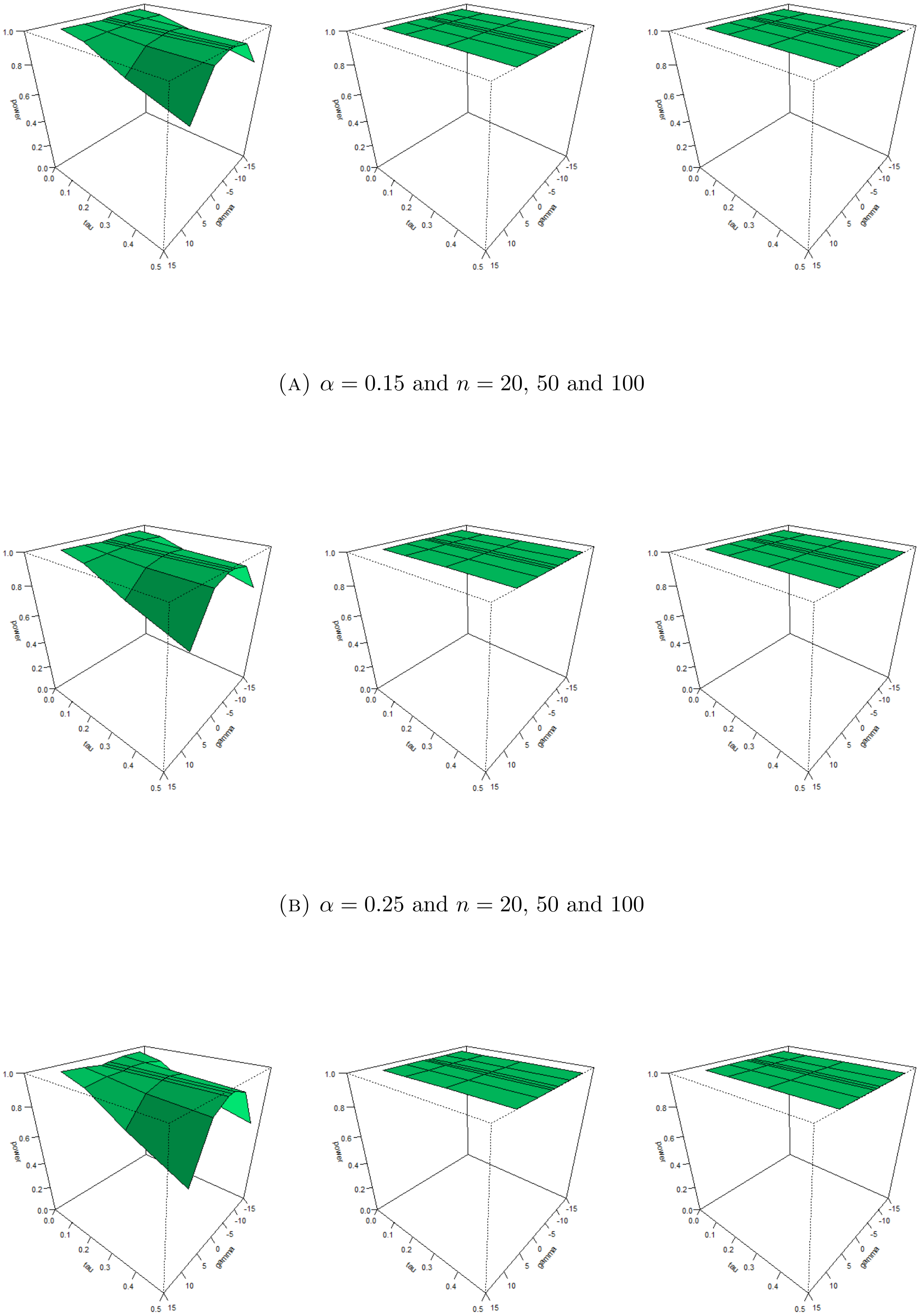


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.16: (Continued) Simulated power of the GSDT for testing H_1 : $\mu = 0.5$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(0.5, 1)$ model)

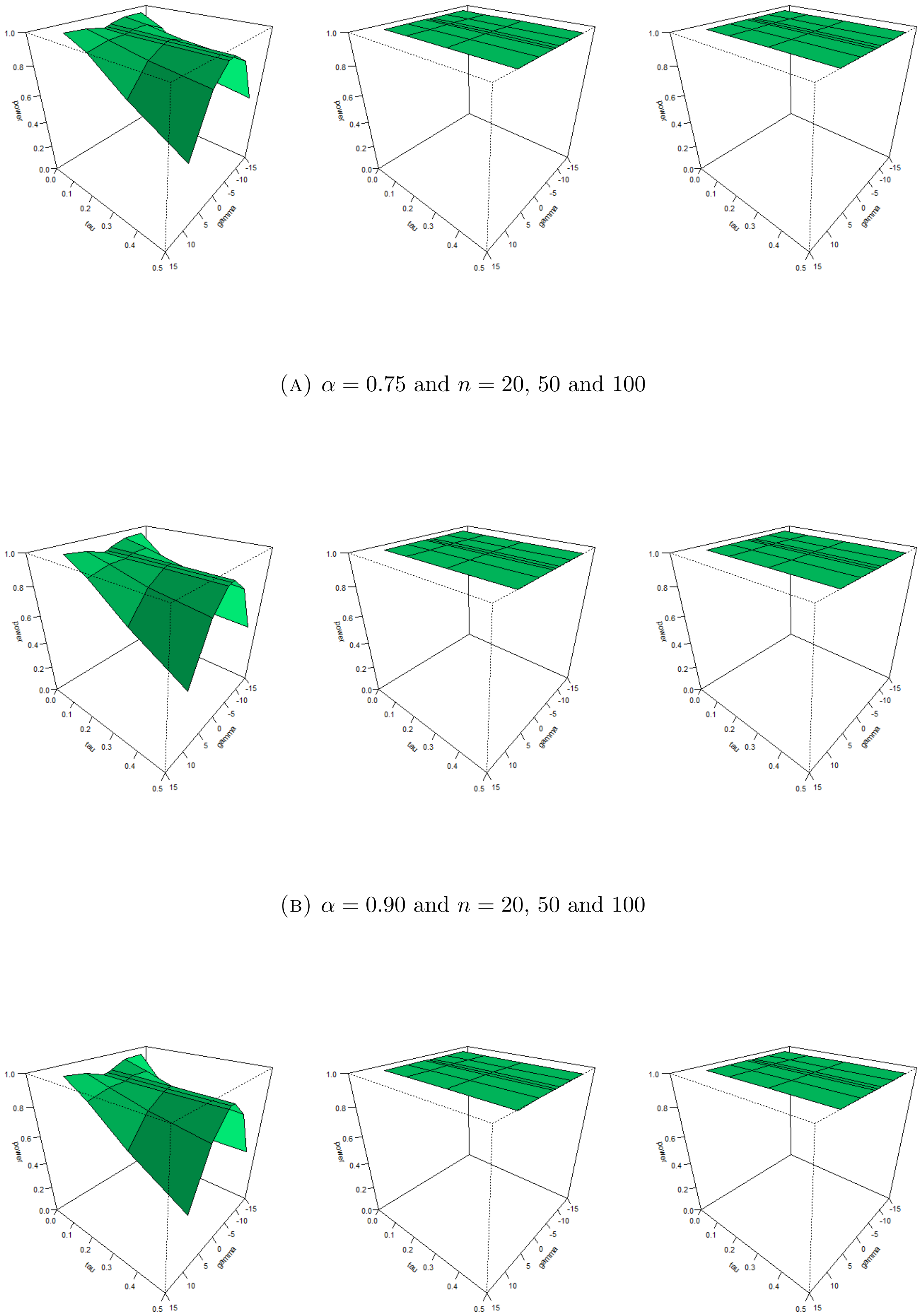


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.17: Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(1, 1)$ model)

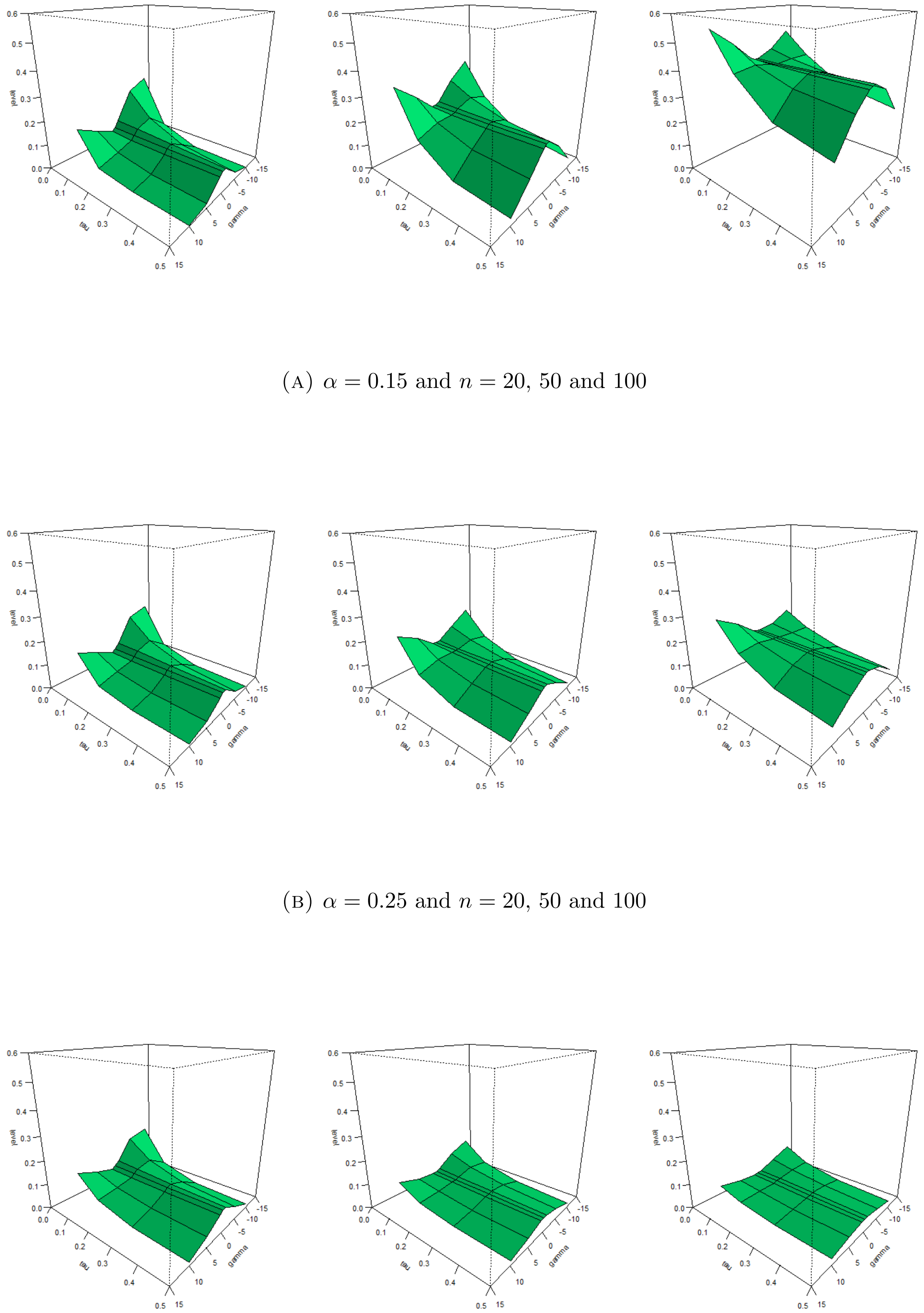


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.18: (Continued) Simulated power of the GSDT for testing H_1 : $\mu = 1.0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(1, 1)$ model)

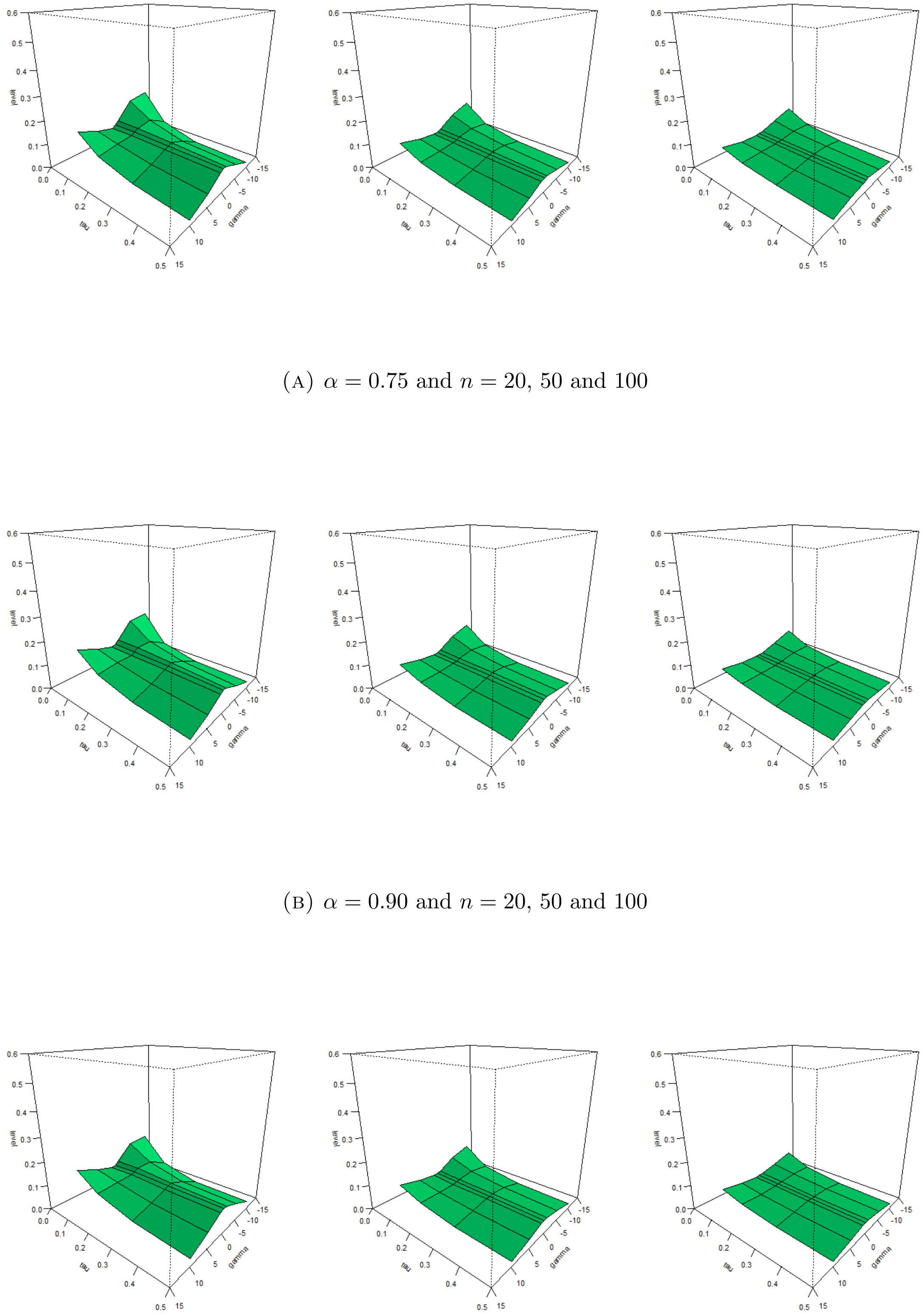


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.19: Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(0,1) + 0.1\ N(5,1)$ model)

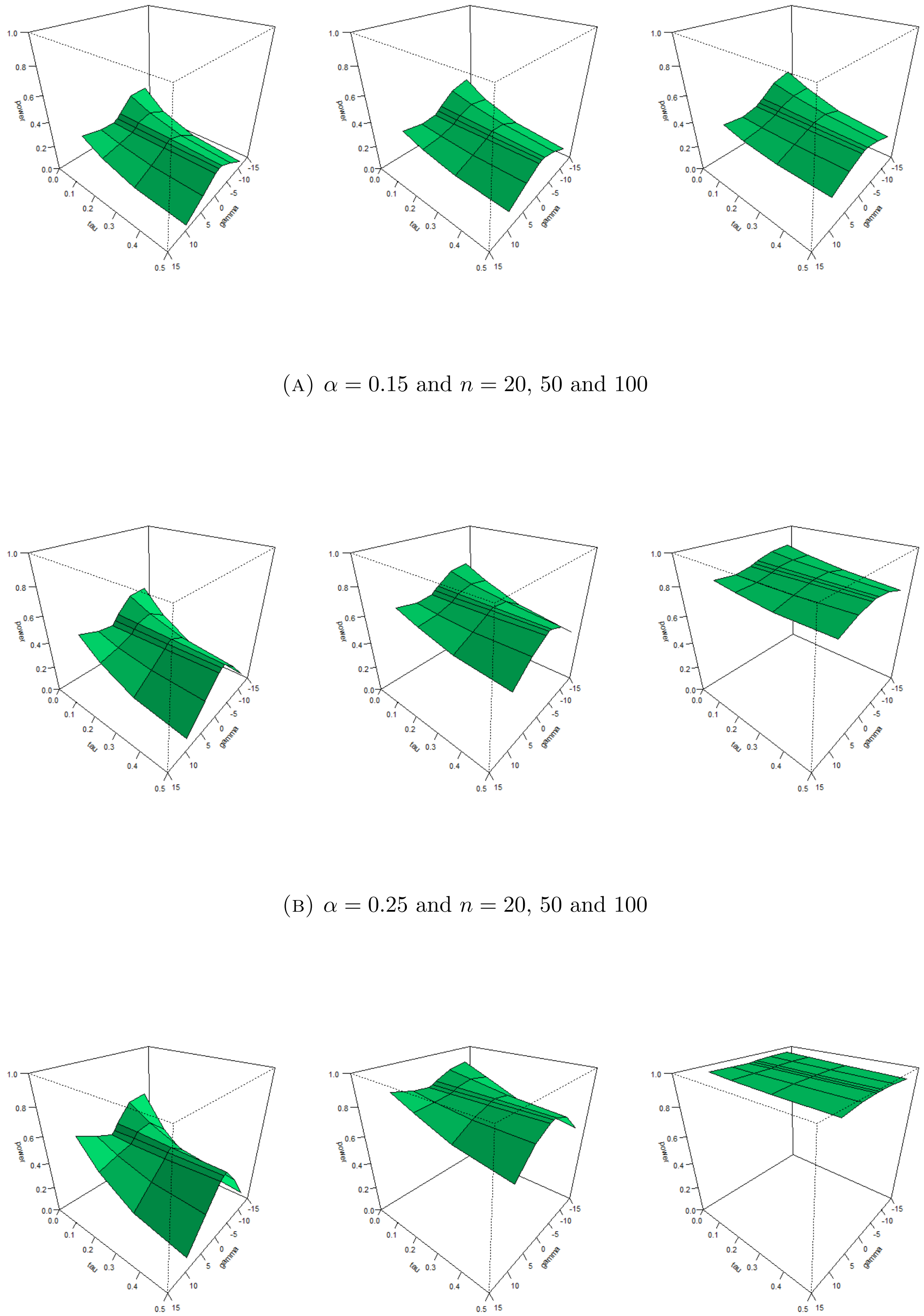


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.20: (Continued) Simulated level of the GSDT for testing $H_0 : \mu = 0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(0, 1) + 0.1\ N(5, 1)$ model)

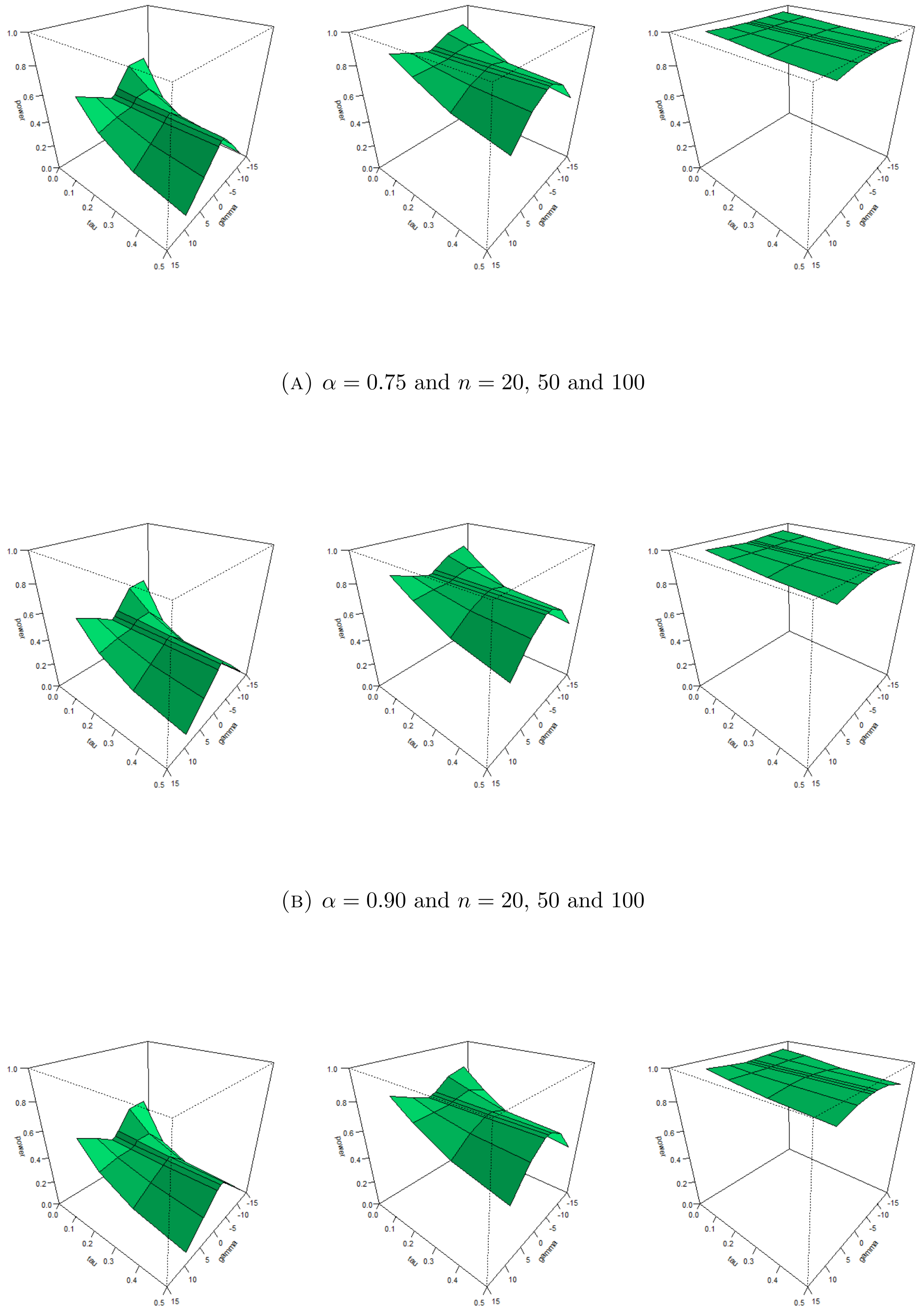


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.21: Simulated power of the GSDT for testing $H_1 : \mu = 0.5$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of 0.9 $N(0.5, 1) + 0.1$ $N(-5, 1)$ model)

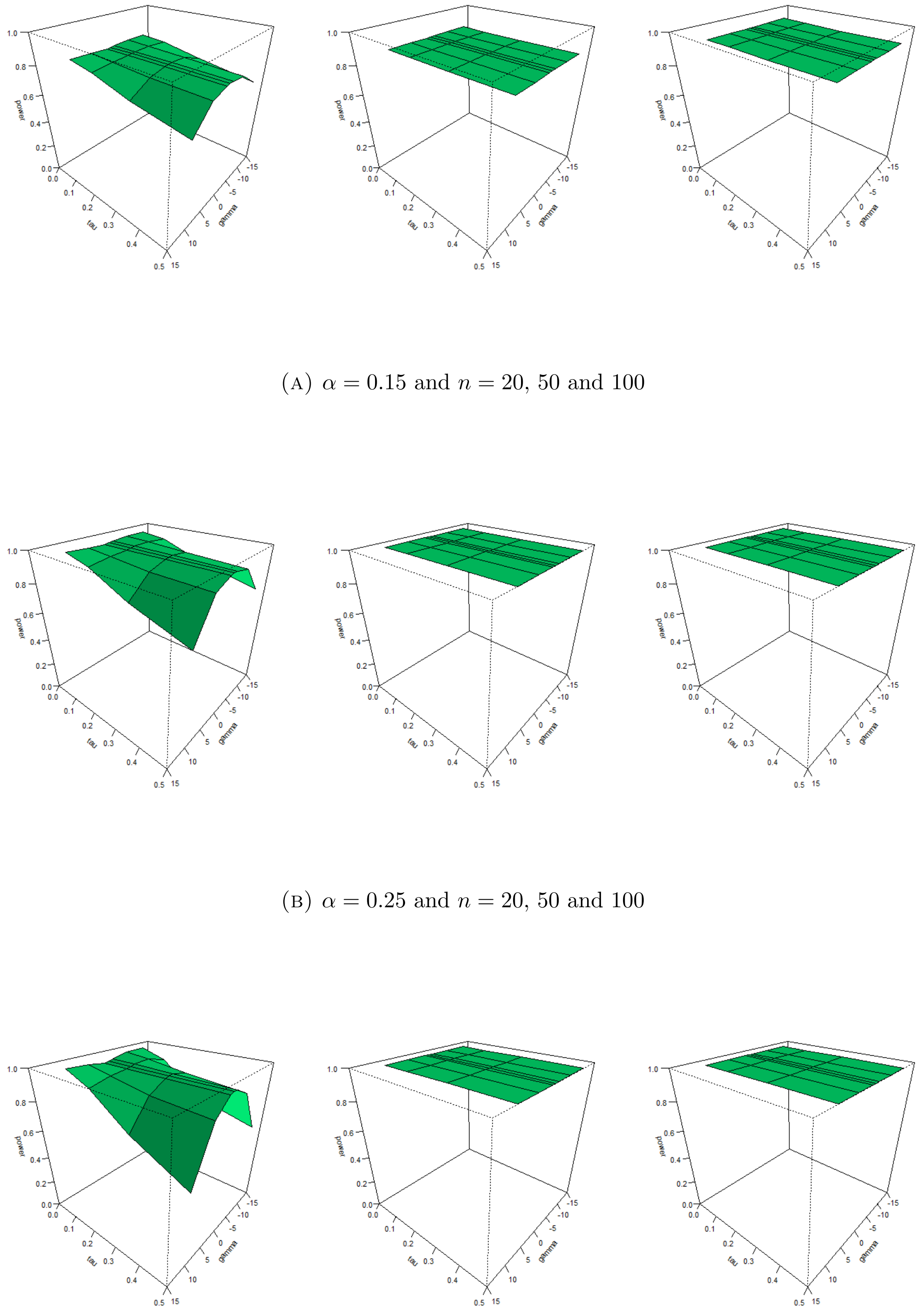


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.22: (Continued) Simulated power of the GSDT for testing H_1 : $\mu = 0.5$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(0.5, 1) + 0.1\ N(-5, 1)$ model)

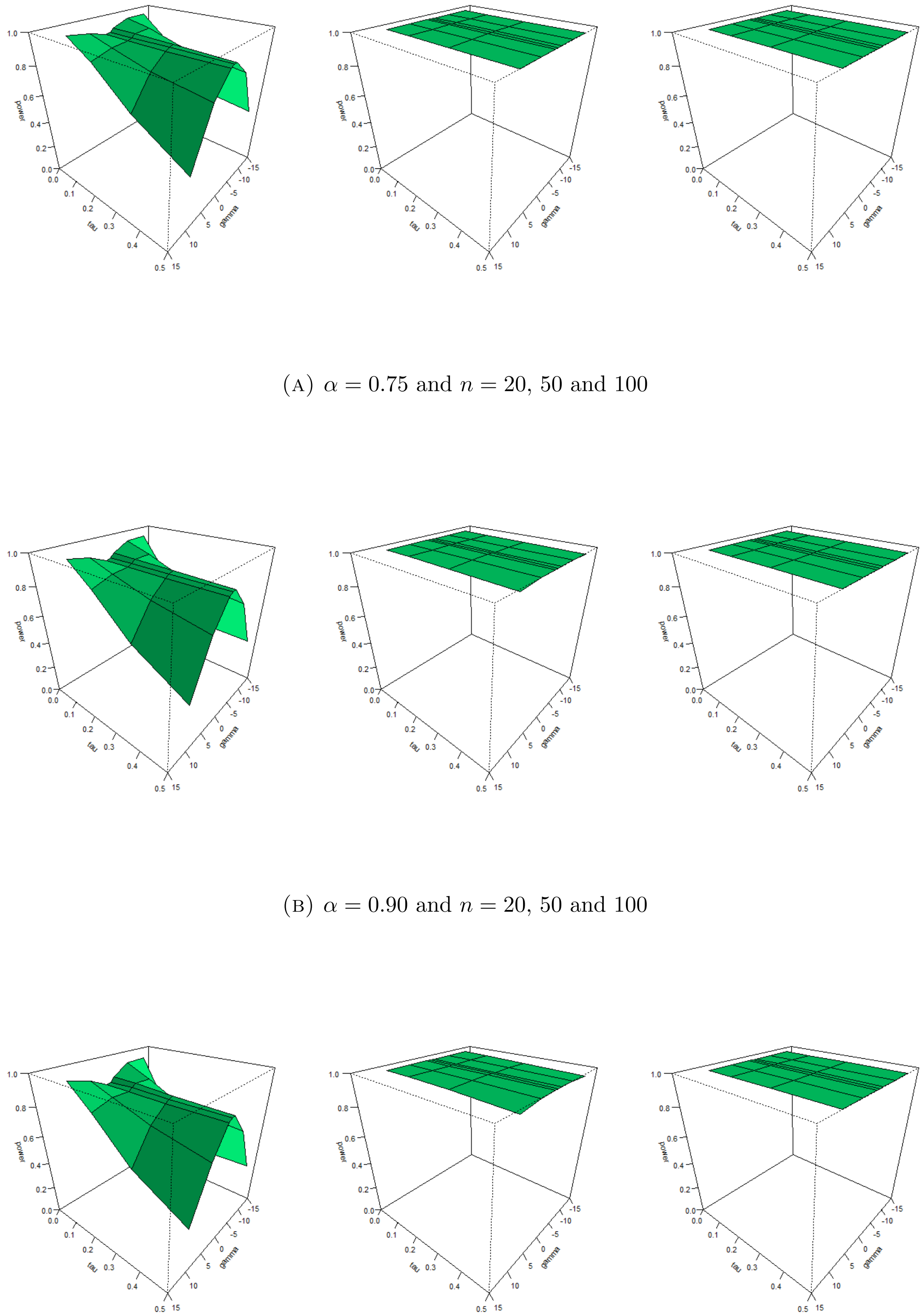


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.23: Simulated power of the GSDT for testing $H_1 : \mu = 1.0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(1, 1) + 0.1\ N(-9, 1)$ model)

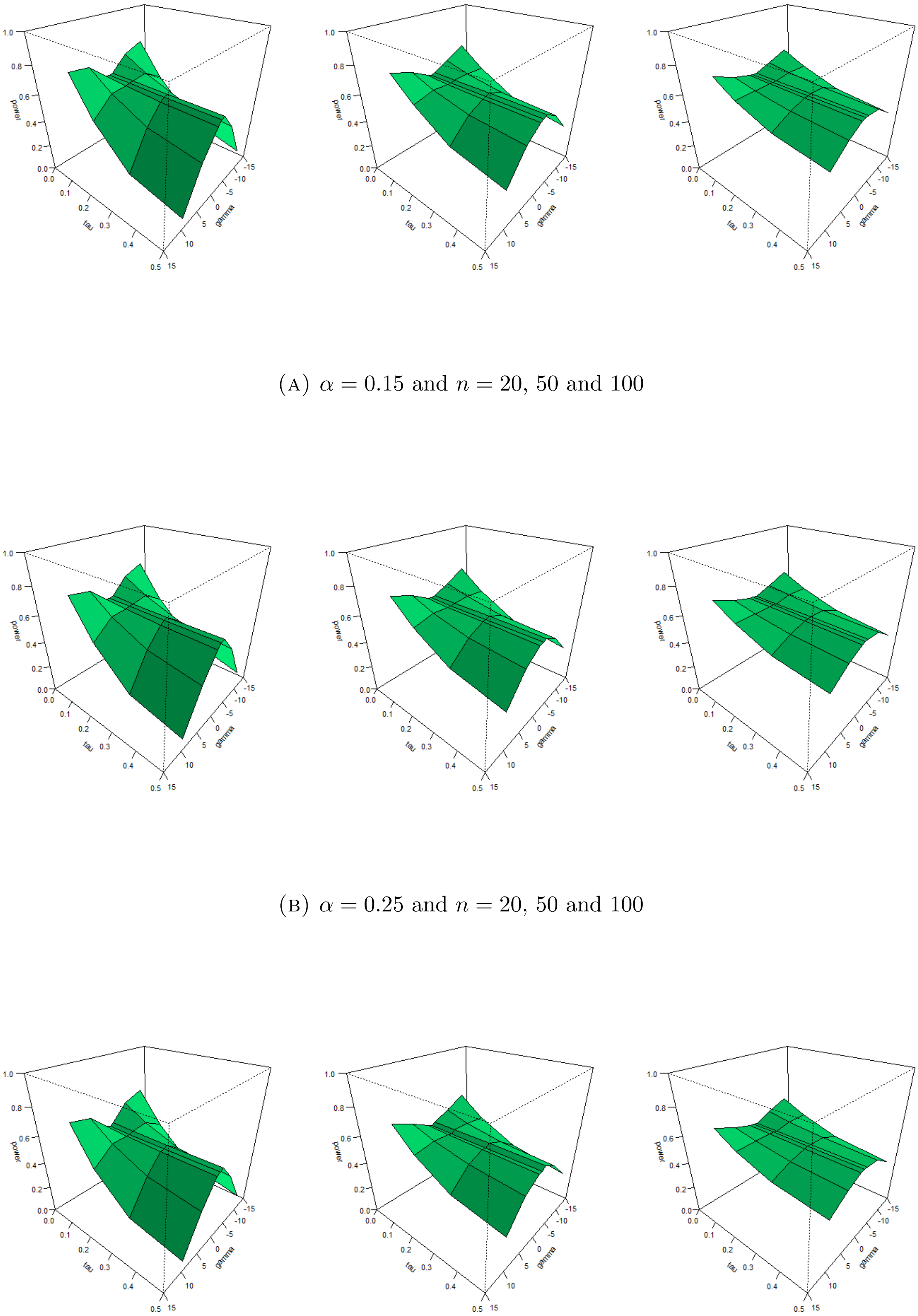


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.24: (Continued) Simulated power of the GSDT for testing H_1 : $\mu = 1.0$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(1, 1) + 0.1\ N(-9, 1)$ model)

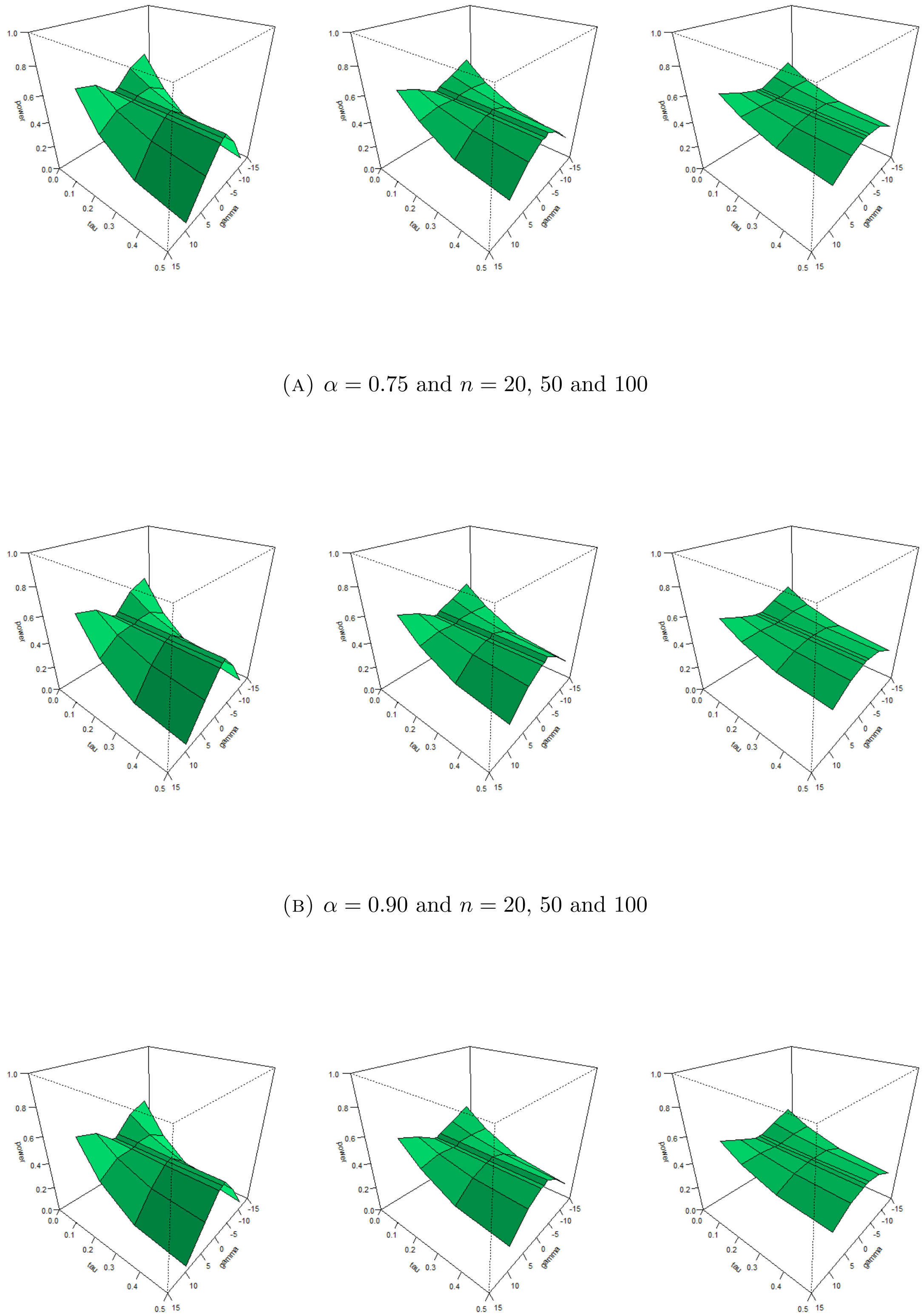


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.25: Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(\frac{\sqrt{5}}{\sqrt{n}}, 1)$ models)

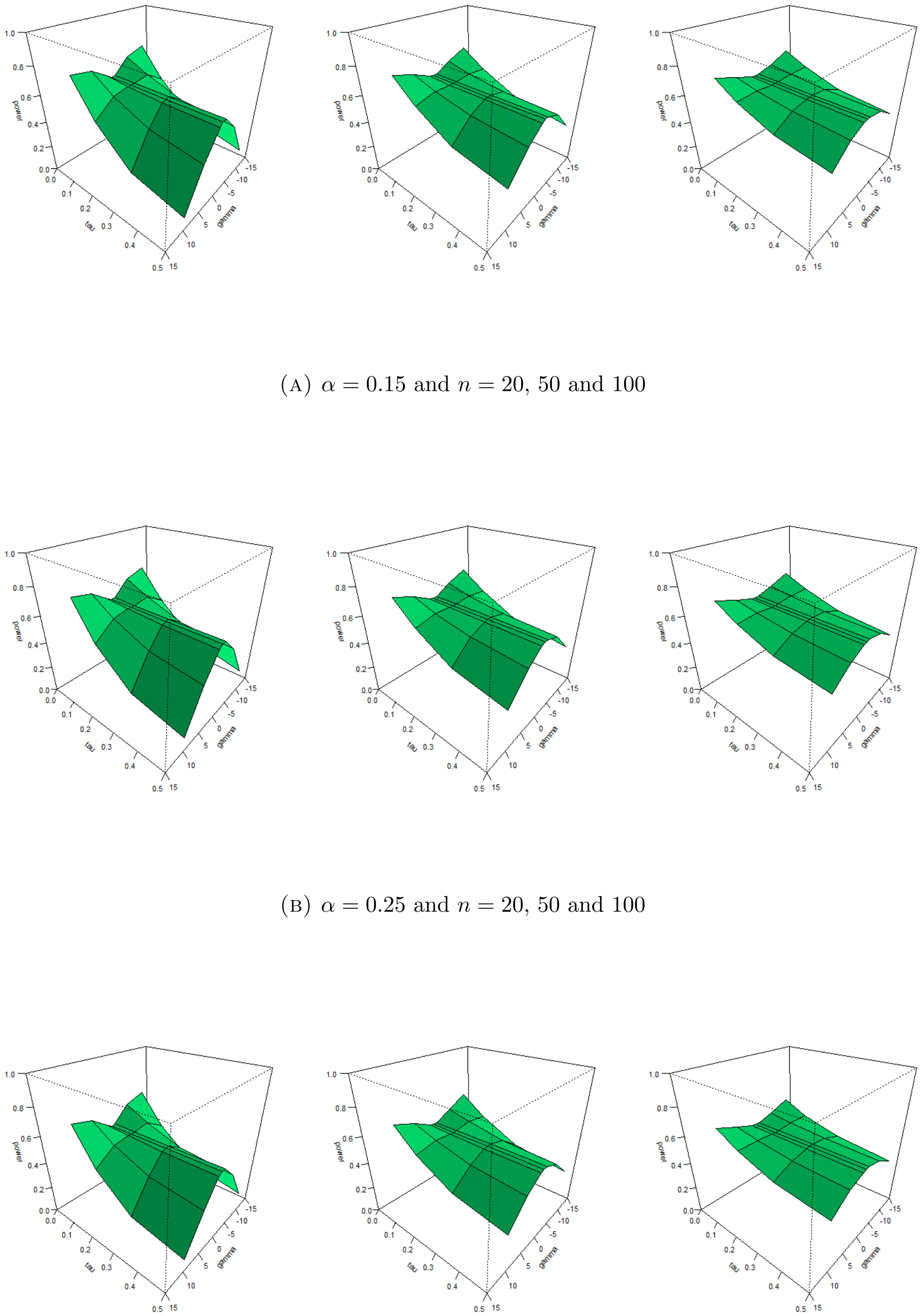


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.26: (Continued) Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(\frac{\sqrt{5}}{\sqrt{n}}, 1)$ models)

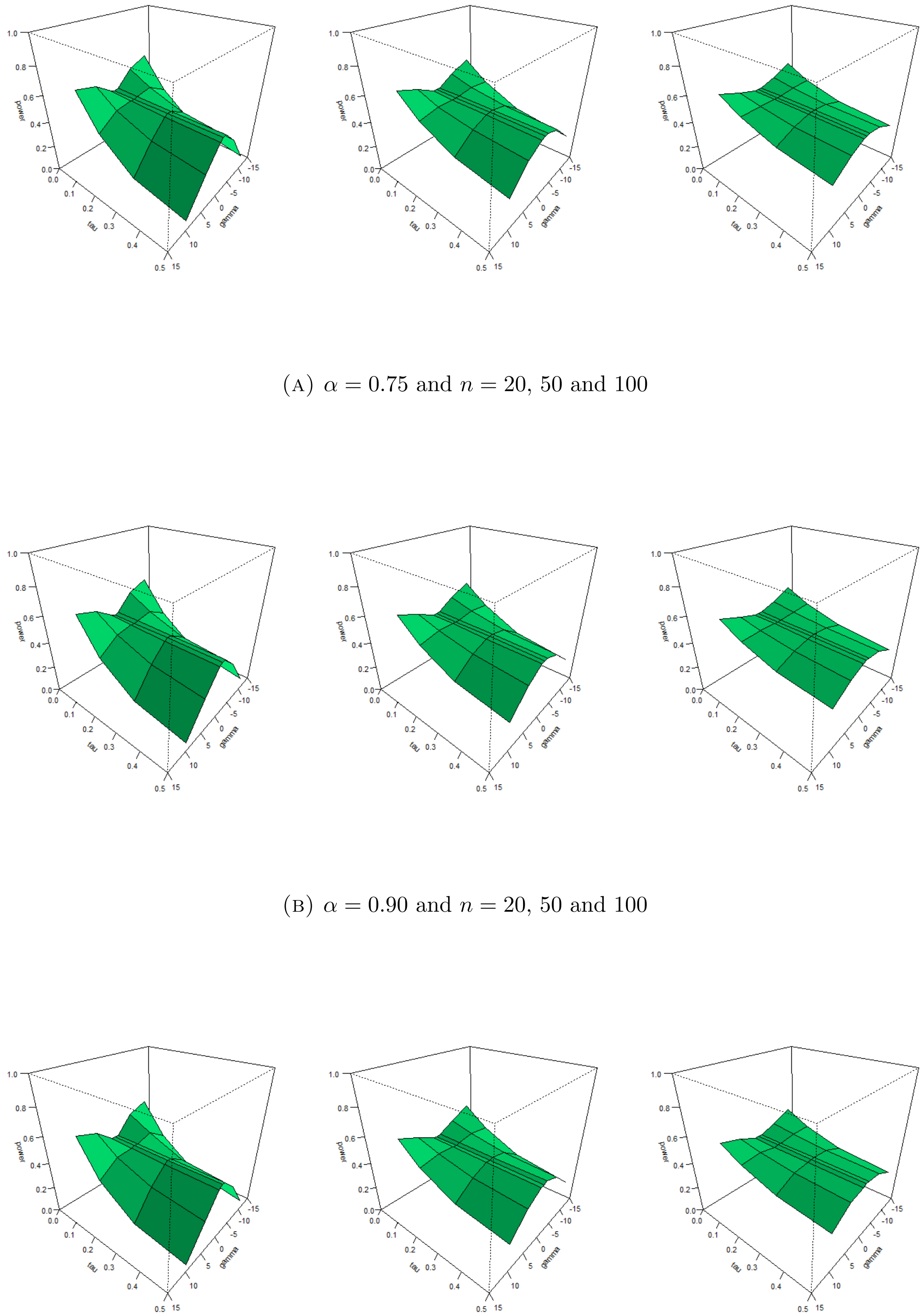


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.27: Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of 0.9 $N(\frac{\sqrt{5}}{\sqrt{n}}, 1) + 0.1\ N(-5, 1)$ models)

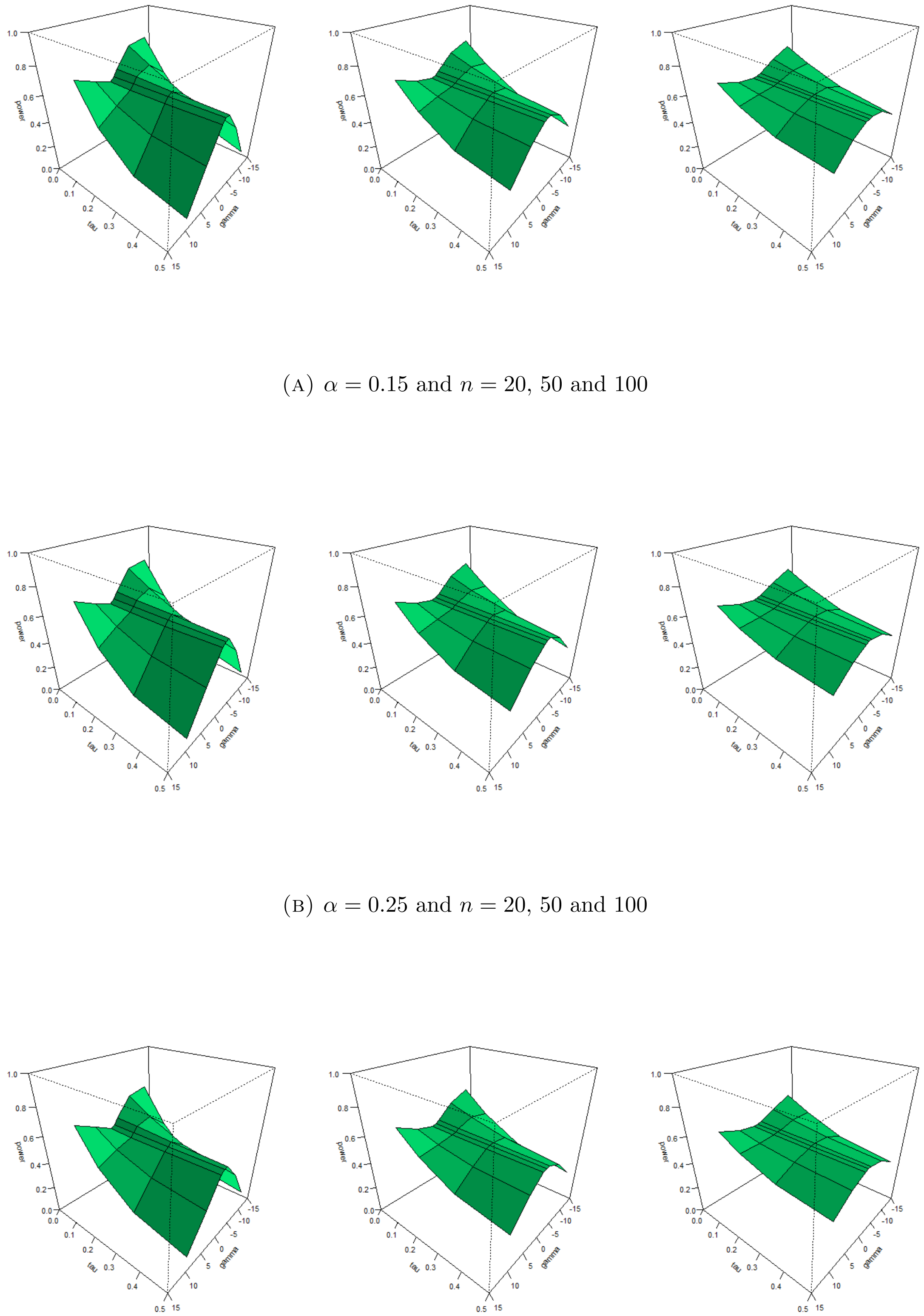


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.28: (Continued) Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ known) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(\frac{\sqrt{5}}{\sqrt{n}}, 1) + 0.1\ N(-5, 1)$ models)

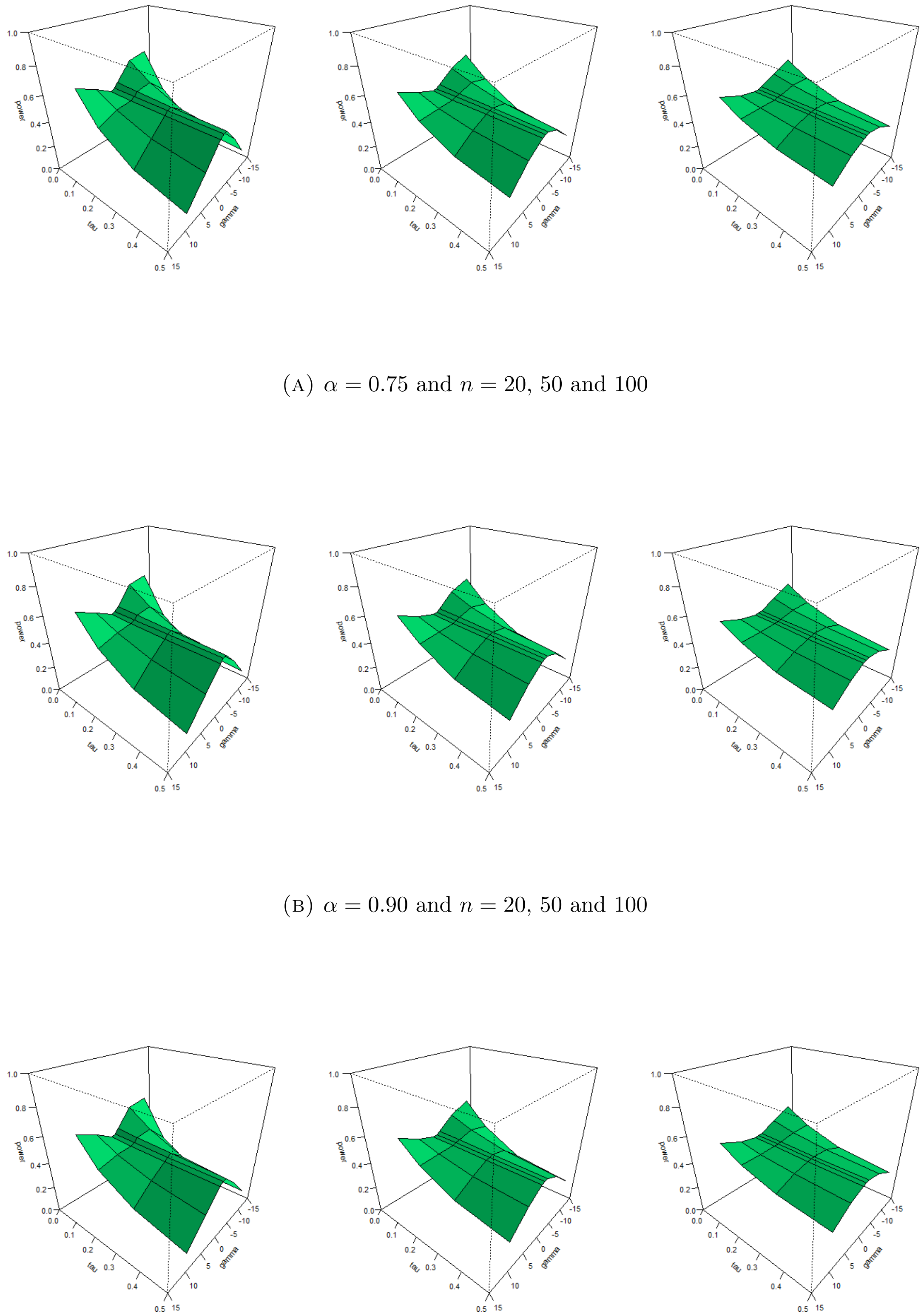


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.29: Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(\frac{\sqrt{5}}{\sqrt{n}}, 1)$ models)

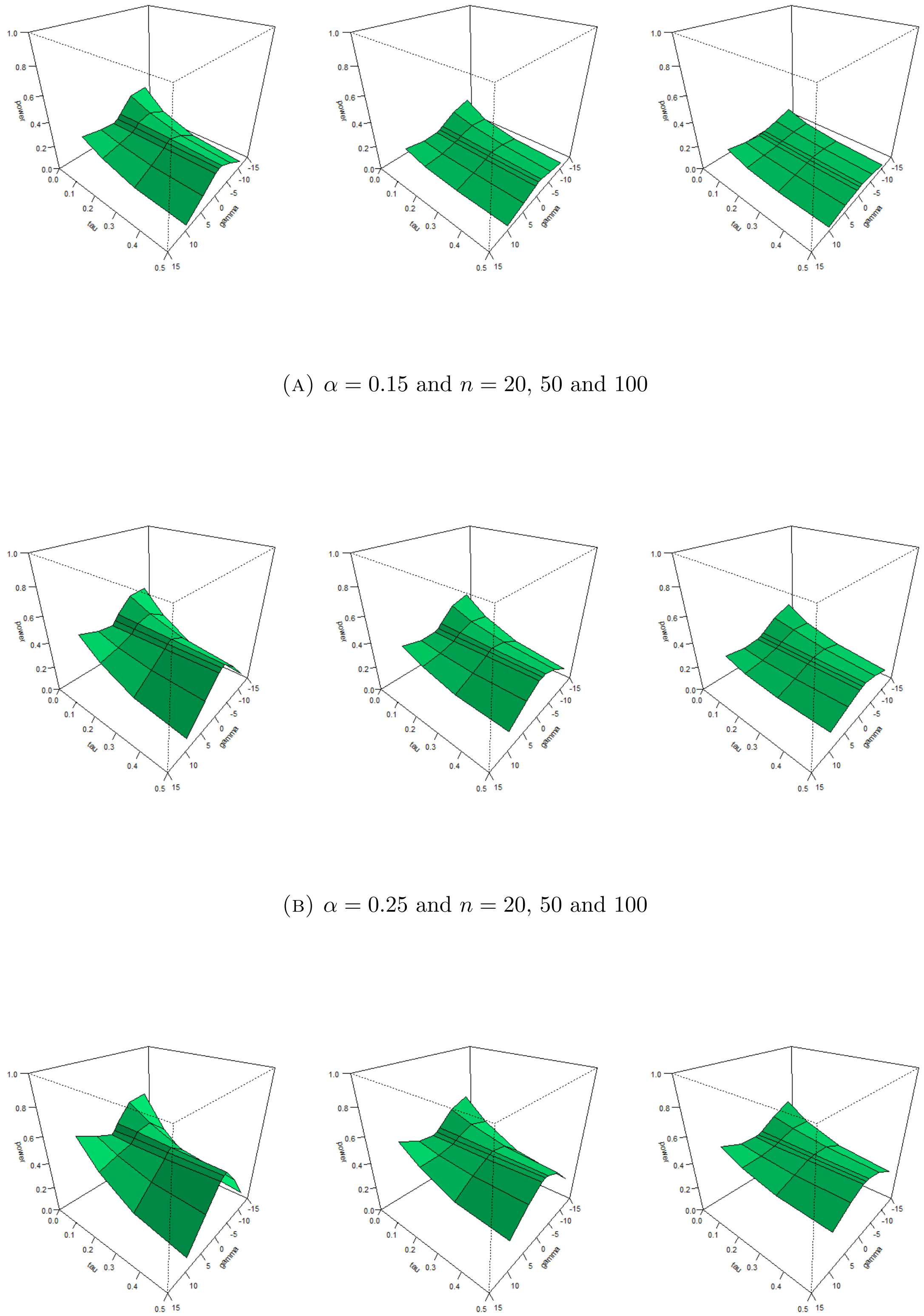


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.30: (Continued) Simulated power of the GSDT for testing H_1 : $\mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $N(\frac{\sqrt{5}}{\sqrt{n}}, 1)$ models)

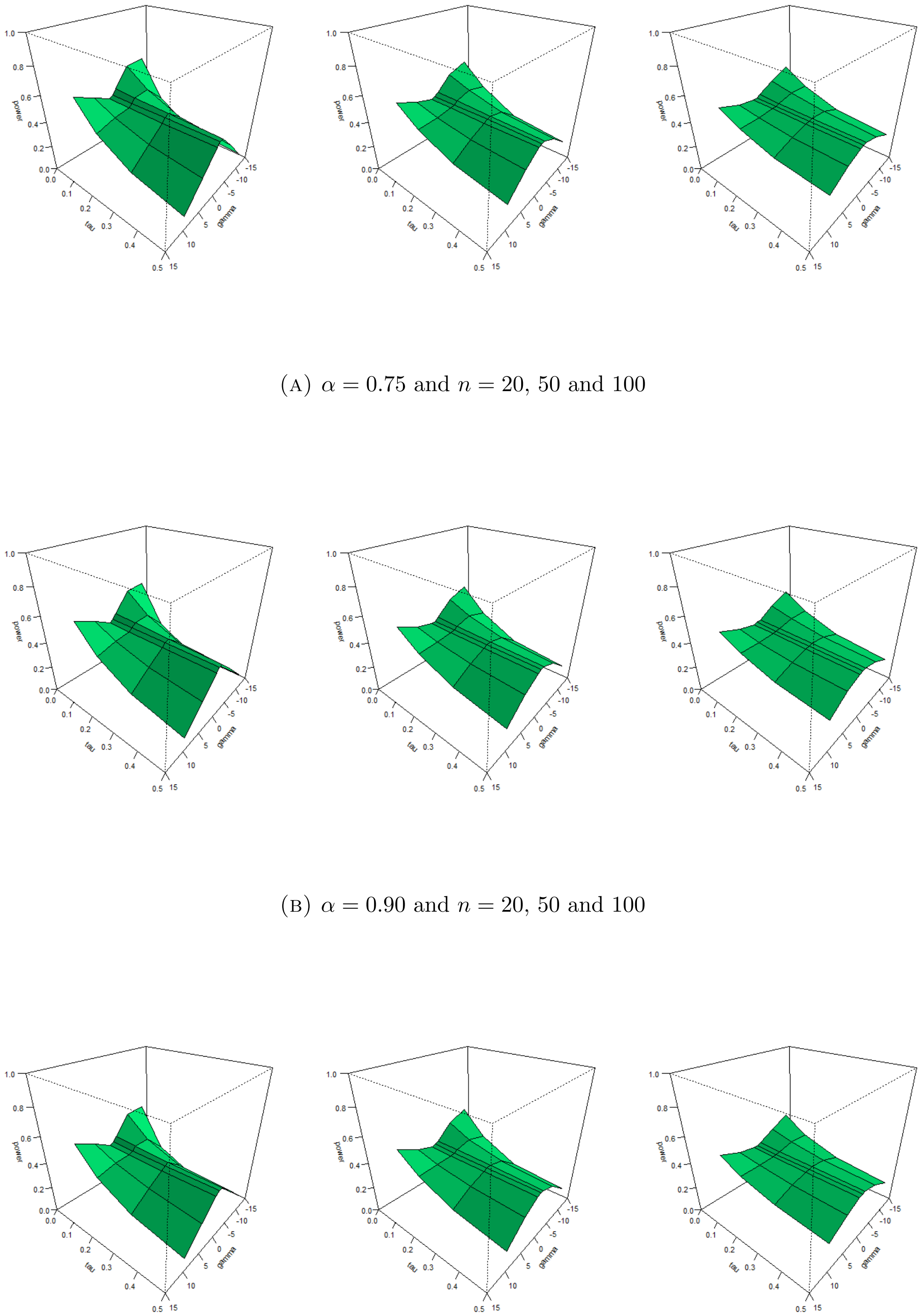


(A) $\alpha = 0.15$ and $n = 20$, 50 and 100

(B) $\alpha = 0.25$ and $n = 20$, 50 and 100

(C) $\alpha = 0.50$ and $n = 20$, 50 and 100

FIGURE 7.31: Simulated power of the GSDT for testing $H_1 : \mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of $0.9\ N(\frac{\sqrt{5}}{\sqrt{n}}, 1) + 0.1\ N(-5, 1)$ models)

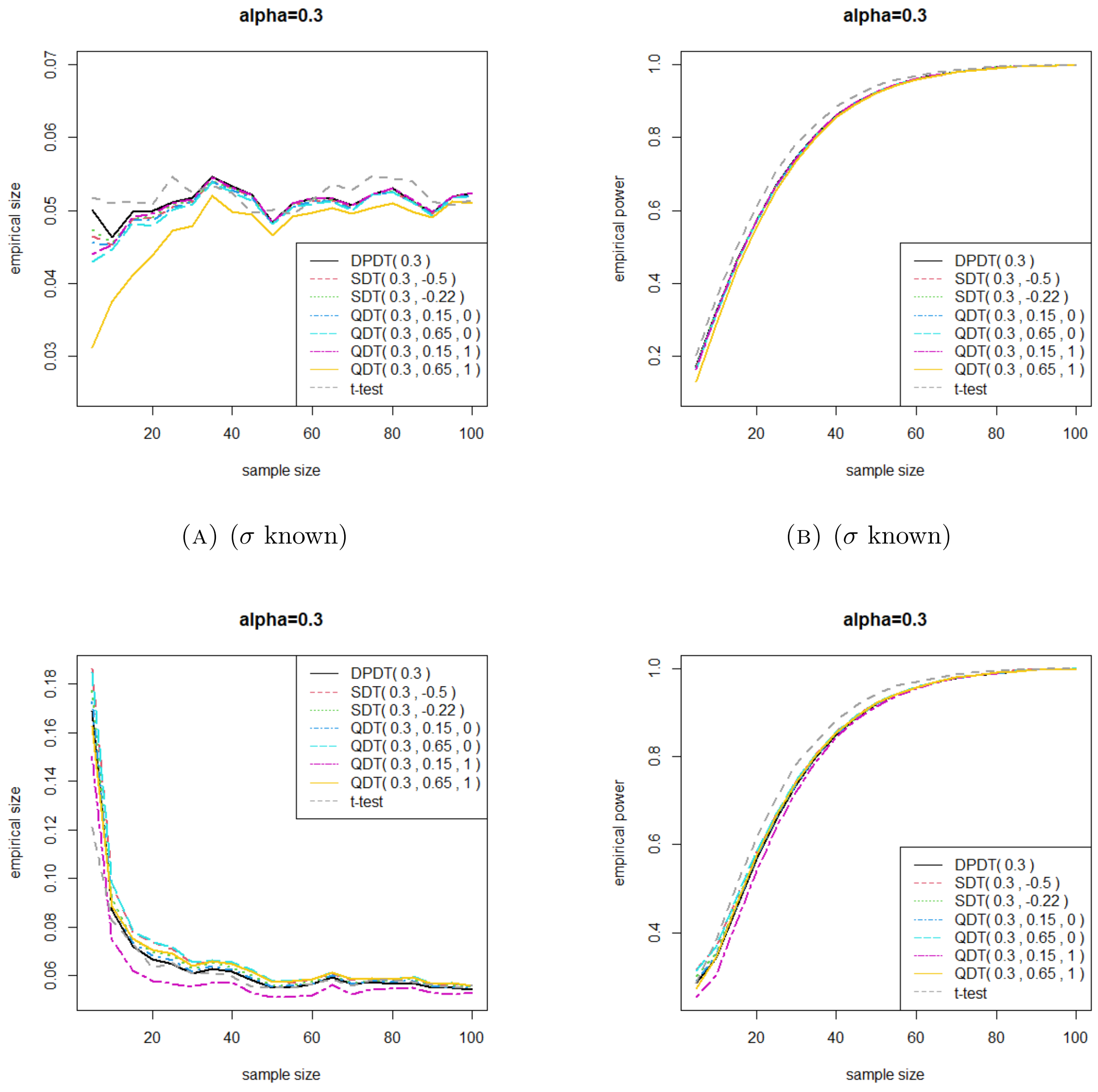


(A) $\alpha = 0.75$ and $n = 20$, 50 and 100

(B) $\alpha = 0.90$ and $n = 20$, 50 and 100

(C) $\alpha = 1.00$ and $n = 20$, 50 and 100

FIGURE 7.32: (Continued) Simulated power of the GSDT for testing H_1 : $\mu = \frac{\sqrt{5}}{\sqrt{n}}$ (with σ unknown) corresponding to different sample sizes and different triplets (α, τ, γ) (in case of 0.9 $N(\frac{\sqrt{5}}{\sqrt{n}}, 1) + 0.1\ N(-5, 1)$ models)



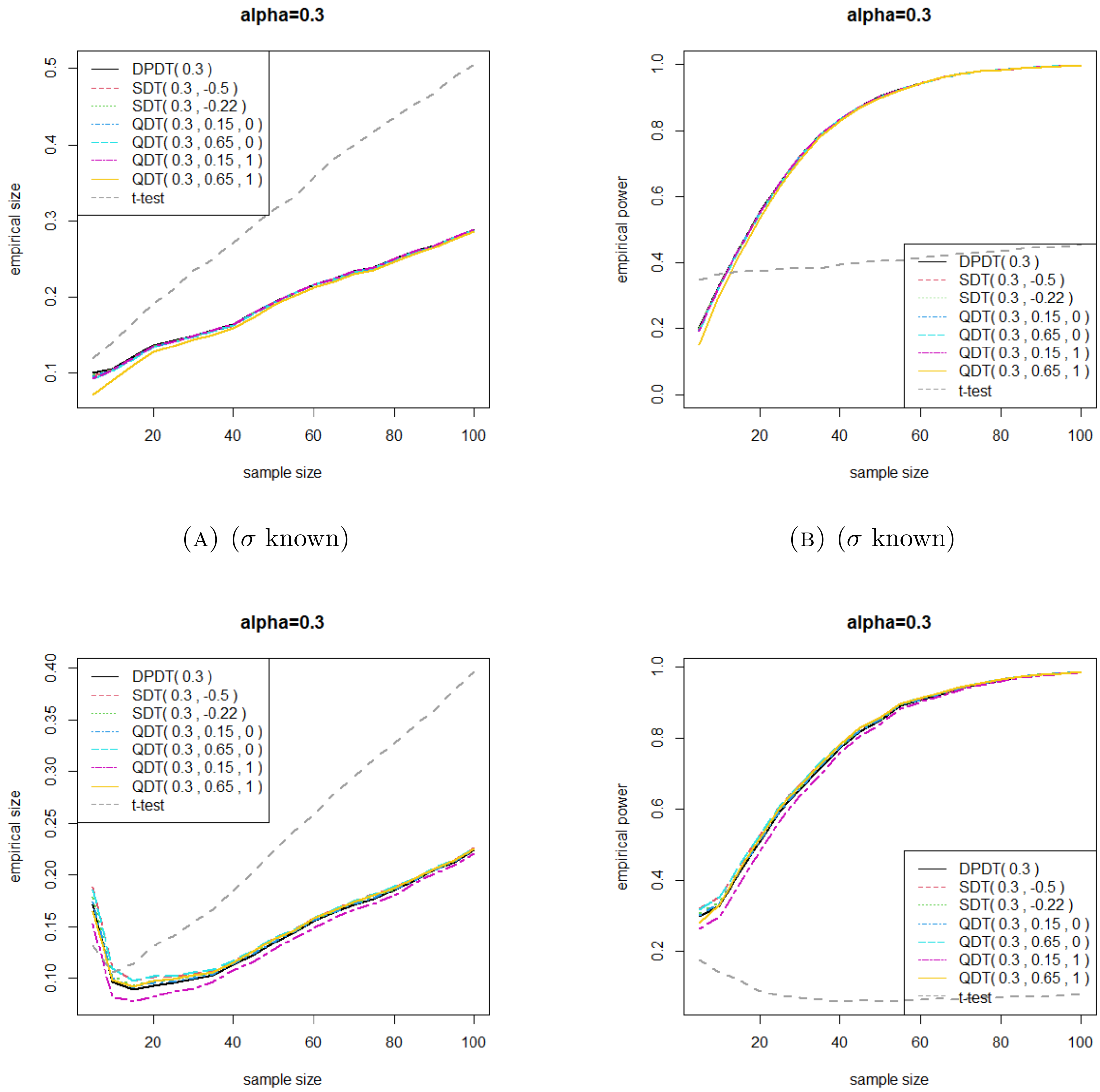
(A) (σ known)

(B) (σ known)

(C) (σ unknown)

(D) (σ unknown)

FIGURE 7.33: Simulated size (plots (A) and (C)) for testing $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ for pure data generated from $N(0, 1)$. Plots (B) and (D) represent the powers of the same test at $\mu = 0.5$ (data generated from $N(0.5, 1)$). In the above plots the GSDT is represented by the QDT.



(A) (σ known)

(B) (σ known)

(C) (σ unknown)

(D) (σ unknown)

FIGURE 7.34: Simulated size (plots (A) and (C)) for testing $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ for contaminated data generated from $0.9N(0, 1) + 0.1N(-2, 1)$. Plots (B) and (D) represent the powers of the same test at $\mu = 0.5$ (data generated from $0.9N(0.5, 1) + 0.1N(-6, 1)$). In the above plots the GSDT is represented by the QDT.

7.6 Real Life Data Examples

In this section, we proceed one step further by applying our proposed GSDTs in real life scenarios. Here, we have considered one discrete and two continuous datasets – the discrete one follows the Poisson distribution, whereas the two continuous datasets follow the normal model.

To study their performance under these models, we have taken the help of graphical representations, more specifically, 3D surface plots. Keeping the value of α fixed, these surface plots of p-values are constructed with respect to τ and γ. Here, we have chosen two discriminating α's corresponding to two opposing statistical decisions for testing any set of hypotheses. If the decision goes in favour of the rejection of H_0 at α_0 and against rejection of H_0 at α_1, where $\alpha_1 > \alpha_0$, then the same decision of the rejection of the null hypothesis holds true for any $\alpha < \alpha_0$ and the opposite decision holds true for any $\alpha > \alpha_1$. Similarly, if the decision goes against the rejection of H_0 at α_0 and in favour of the rejection of H_0 at α_1, where $\alpha_1 > \alpha_0$, then the same decision of the rejection of the null hypothesis holds true for any $\alpha > \alpha_1$ and the opposite decision holds true for any $\alpha < \alpha_0$. Here, we will show that the deviation between the p-values for these two α's is becoming less in case of the clean data rather than the full data. Sometimes, the two surface plots of p-values are getting overlapped with each other, either partially or fully. Now, through the following three examples, we are going to prove the satisfactory results of our proposed statistics belonging to the 'best' region, as derived in the previous section.

7.6.1 One Sample Telephone Line Fault Data

Next, we consider an interesting dataset containing the records on telephone line faults given in Table 7.1, to picturize the nature of the GSDT in robust inference. The data, analyzed earlier by Simpson (1989), consist of observations on the ordered differences between the inverse rates of test and control for 14 matched pairs. The data contain one large outlier and excluding this, the dataset can be well-modelled by the $\mathrm{N}(\mu, \sigma^2)$ distribution. First we have considered two sets of hypotheses for testing μ, namely,

$$H_0 : \mu = 0 \text{ vs. } H_1 : \mu \neq 0 \qquad \text{and} \qquad H_0^{'} : \mu = 115 \text{ vs. } H_1^{'} : \mu \neq 115. \tag{7.44}$$

Figures 7.35 and 7.36 represent the p-values for these two cases, in case of full data and clean data, respectively. For the σ-known case, σ is considered to be 132, as suggested by Basu et al. (2013), whereas, for the σ-unknown case, the MDPDE has been used for a fixed α. Due to the presence of the outlier, the MLEs are getting distorted and as a result, if we perform a t-test, then it would fail to reject the null hypothesis $H_0 : \mu = 0$ due to its non-robust nature, whereas, if we look at Figures 7.35(a) and 7.35(c), we can then see that for the larger α between the two cases considered, the decision reverses; more specifically, $\gamma \in [-1, 1]$ and $\tau < 0.3$ strongly reject the null hypothesis. For the σ-known case, rejection of the null hypothesis with low p-values is observed when $\alpha > 0.01$ and the same result is observed for $\alpha > 0.1$ in the σ-unknown case. Now, if we consider the second set of hypotheses, the outlier influenced the t-test to reject the null. Here also, the opposite picture is very much visible for the GSDTs in Figures 7.35(b) and 7.35(d). Larger α, along with smaller τ and γ values, makes the test resistant through the failure to reject

H_0. Next, if we illustrate the case of the clean data, we can then see that the LRT test or the t-test are giving the desired results for both the cases, whereas, in case of our proposed test also, the results coincide for those two discriminating α's (considered in case of the full data), which is quite clear from these two almost-overlapping surfaces. Overall, the robust nature of the GSDTs with a pair of (τ, γ) belonging to the 'best' region is quite clear in the scenarios corresponding to the two sets of hypotheses.

The presence of outlier in the normal model generally inflates the σ estimate and as a result, the LRT test does not fulfill the requirements of testing. Therefore, in this case, to check the performance of the GSDTs, we are going to test

$$H_0 : \sigma = 132 \quad \text{vs.} \quad H_1 : \sigma \neq 132, \tag{7.45}$$

with $\mu = 115$ (known), as suggested by Basu et al. (2013). Keeping the outlier intact, the data lead to $\hat{\sigma}$ = standard deviation of the full data = 321.94, while after removing this outlier, we get, $\hat{\sigma}$ = standard deviation of clean data = 132.82. Now, we are to check whether the GSDTs help us to get correct statistical decisions via testing or not. If we look at Figure 7.37(a), that is, the full data case, we can see that whenever $\alpha > 0.1$, the result is in favour of the failure of rejection of H_0 for each pair of (τ, γ) along with the 'best' region. On the other hand, from Figure 7.37(b), that is, the clean data case, the p-values being almost 1 coincide with the statistical decision of the LRT test, that is, leads to the failure of rejection of H_0.

Lastly, we consider the most practical case, that is, a test on $\theta = (\mu, \sigma)^T$, since both are unknown. Here, we consider the hypothesis

$H_0 : (\mu, \sigma)^T = (115, 132)^T$ against its omnibus alternative. Here, the dimension of the parameters is greater than 1 and hence, the proportion of the test statistic $Q_{(\alpha,\tau,\gamma)}\left(f_{\hat{\theta}_\alpha}, f_{\theta_0}\right)$ being greater than or equal to $\sum_i \lambda_i Z_i^2$, with λ_i's being the non-zero eigenvalues of matrix $A_\alpha(\theta_0) J_\alpha^{-1}(\theta_0) K_\alpha(\theta_0) J_\alpha^{-1}(\theta_0)$, has been considered to be the simulated p-value, for a pre-fixed (α, τ, γ). From Figure 7.37(c), it is clear that $\alpha \geq 0.103$ leads to the failure of rejection of H_0, irrespective of any choice of (τ, γ) considered in this study. On the other hand, if we consider the outlier-deleted case, both surfaces, corresponding to two different α's, generate p-values belonging to $[0.9, 1]$ leading us to the failure to reject H_0 and, in fact, these two surfaces are highly overlapped in Figure 7.37(d).

Table 7.1: Telephone Line Fault Data

Pair	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Difference	-988	-135	-78	3	59	83	93	110	189	197	204	229	289	310

7.6.2 One Sample Darwin's Plant Fertilization Data

Charles Darwin had performed an experiment on 15 pairs of a certain variety of plants, one self-fertilized and the other cross-fertilized, to justify the claim – whether these two types of plants have different growth rates. This data set, consisting of 15 pairs of differences in height of these two types of plants after a specific time period, is given in Table 7.2. There are two moderate outliers in these data. Under the assumption of the normal model, an obvious intuition is to test for

$$H_0 : \mu = 0 \quad \text{against} \quad H_1 : \mu \neq 0. \tag{7.46}$$

Since the two outliers are geometrically well-separated from the rest of the data, they influence the LRT to fail to reject H_0 and if we consider tests based on the GSD class, we can then observe that the two-sided p-values (in Figure 7.38) are quite different for the two surfaces corresponding to two α's in case of the full data. However, the opposite scenario can be observed in case of the clean data. Moreover, if we consider larger α values, then irrespective of the values of τ and γ, the two surfaces fully overlap each other and for each of these cases, the p-values being extremely low indicates strong rejection of H_0. So, here also, larger value of α helps the test statistic to retain its robust characteristics for both the cases of full and clean data.

TABLE 7.2: Darwin's Plant Fertilization Data

Pair	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Difference	-67	-48	6	8	14	16	23	24	28	29	41	49	56	60	75

7.6.3 Two Sample Drosophila Data

Let us end this section by considering a real life dataset (given in Table 7.3) consisting of two independent samples on the occasional spurious counts in Drosophila Assay. These data have been analyzed by Woodruff et al. (1984), Simpson (1989), Basu et al. (2013), etc. Some male flies were exposed to a certain degree of chemical and the remaining were not. The variable of interest is based on the number of daughter flies carrying a recessive lethal mutation. The first group is outlier-free, whereas, the second group has two large outliers. These variables are assumed to follow the Poisson distribution with means θ_1 (control group) and θ_2 (treatment group) respectively.

Here, our specific target is to check the validity of our apprehension regarding the use of the chemical reducing the lethal mutation and

Table 7.3: Frequencies of the number of recessive lethal daughters for the Drosophila data

	0	1	2	3	4	5	6	7
Observed (control)	159	15	3	0	0	0	0	0
Observed (treated)	110	11	5	0	0	0	1	1

hence, the obvious hypotheses will be –

$$H_0 : \theta_1 \geq \theta_2 \qquad \text{vs.} \qquad H_1 : \theta_1 < \theta_2. \tag{7.47}$$

Let ${}^{(1)}\hat{\theta}_\alpha$ and ${}^{(2)}\hat{\theta}_\alpha$ denote the MDPDEs at fixed α for the two groups respectively, whereas ${}^{(0)}\hat{\theta}_\alpha$ denote the common MDPDE at that fixed α under the null hypothesis. We know that the LRT is the most popular test in such a scenario. But due to the presence of outliers, the mean of the treatment group appeared to be larger, which is not consistent with the remaining part of the data. Hence, we need some robust test in such a scenario. Therefore, an obvious choice is to compare the GSDTs of several combinations of α, τ and γ with this LRT.

We have discussed earlier about two sample test statistics and their distributions for testing

$$H_0^{'} : \theta_1 = \theta_2 \qquad \text{vs.} \qquad H_1^{'} : \theta_1 \neq \theta_2. \tag{7.48}$$

Furthermore, at any fixed (α, τ, γ), the test statistic for testing the set of hypotheses given in (7.48) will be of the form

$$S_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right) = \frac{1}{\lambda\left(\hat{\theta}_\alpha\right)} \frac{2mn}{m+n} Q_{(\alpha,\tau,\gamma)}\left(f_{{}^{(1)}\hat{\theta}_\alpha}, f_{{}^{(2)}\hat{\theta}_\alpha}\right), \tag{7.49}$$

where, $\lambda\left({}^{(0)}\hat{\theta}_\alpha\right) = \frac{A_\alpha\left({}^{(0)}\hat{\theta}_\alpha\right) K_\alpha\left({}^{(0)}\hat{\theta}_\alpha\right)}{J_\alpha^2\left({}^{(0)}\hat{\theta}_\alpha\right)}$. Therefore, to test (H_0, H_1) instead of $(H_0^{'}, H_1^{'})$, we modify the test statistic given in (7.49) in the following way

$$S^*_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right) = \sqrt{S_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right)}\ \mathrm{sgn}\left({}^{(2)}\hat{\theta}_\alpha - {}^{(1)}\hat{\theta}_\alpha\right),$$

where,

$$\mathrm{sgn}(u) = \begin{cases} -1, \ \text{if } u < 0 \\ 0, \ \text{if } u = 0 \\ 1, \ \text{if } u > 0. \end{cases}$$

Since,

$$\begin{aligned} S_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right) &\overset{a}{\sim} \chi^2_1 \\ \Rightarrow S^*_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right) &\overset{a}{\sim} N(0,1). \end{aligned}$$

Therefore, for testing H_0 vs. H_1, a right-tailed test based on $S^*_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right)$ would be appropriate. Using this concept, we have represented p-values corresponding to two different α's along with the specified ranges of τ and γ for both the full and the clean data. Here, in spite of the presence of outliers, for every choice of (τ, γ), the larger α value will result in the failure to reject H_0, while the smaller α is unable to do so. But for the outlier deleted data, both kinds of α's lead to the failure to reject H_0 with partially overlapped surfaces in Figure 7.39. In fact, in case of the full data, it is clear from the figure that the GSDTs generate insignificant p-values for any $\alpha > 0.1$ and hence, they are quite resistant here also.

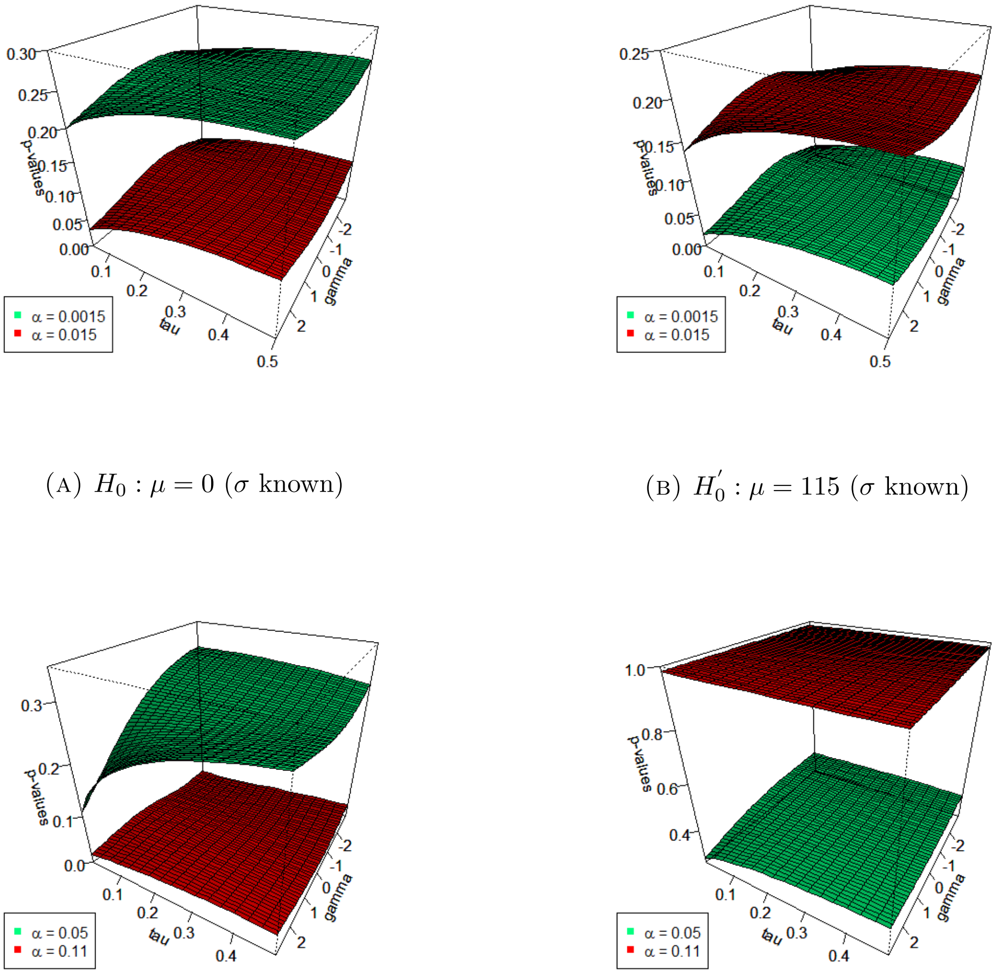
We have already seen the performance of the GSDTs with respect to the small α values, but here we will compare it with other popular tests in the same scenario. For this purpose, keeping α fixed at 0.07, we have analyzed the robustness among the GSDTs, the SDTs and the DPDTs by a tabular representation of the test statistics,

Table 7.4: Value of the test statistic $S^*_{(\alpha,\tau,\gamma)}\left({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha\right)$ and associated p-values for small α in case of both the full and the clean Drosophila data

Test	Full Data		Clean Data	
	Test Statistic	P-value	Test Statistic	P-value
LRT	2.9586	0.0015	1.0986	0.1359
$\text{SDT}_{0.07,-0.5}$	1.6854	0.0460	1.0013	0.1583
$\text{SDT}_{0.07,-0.33}$	1.6647	0.0479	0.9934	0.1602
$\text{GSDT}_{0.07,0.15,0}$	1.6421	0.0503	0.9850	0.1623
$\text{GSDT}_{0.07,0.15,1}$	1.5573	0.0597	0.9518	0.1706
$\text{DPD}_{0.07}$	1.6278	0.0518	0.9791	0.1638

along with the corresponding p-values for both the full and the clean data in Table 7.4. It is quite obvious that the LRT fails to give a satisfactory result and the SDT also fails to retain its robustness. The performance of the DPDT and the GSDTs, though, are satisfactory to some extent. In case of the outlier-deleted data, all the robust tests provide satisfactory results here. Lastly, in case of $\text{GSDT}_{(0.07,0.15,1)}$, we get the strongest evidence regarding the failure to reject H_0.

Thus, for each real data set, the outliers cannot affect the decision of our robust tests whenever the triplet of associated tuning parameters (α, τ, γ) belongs to the 'best' region, as mentioned in the previous section. In case of the outlier-deleted data, smaller values of α lying outside the 'best' region also lead to the satisfactory result, but the other tuning parameters have to remain within that region in most of the cases.



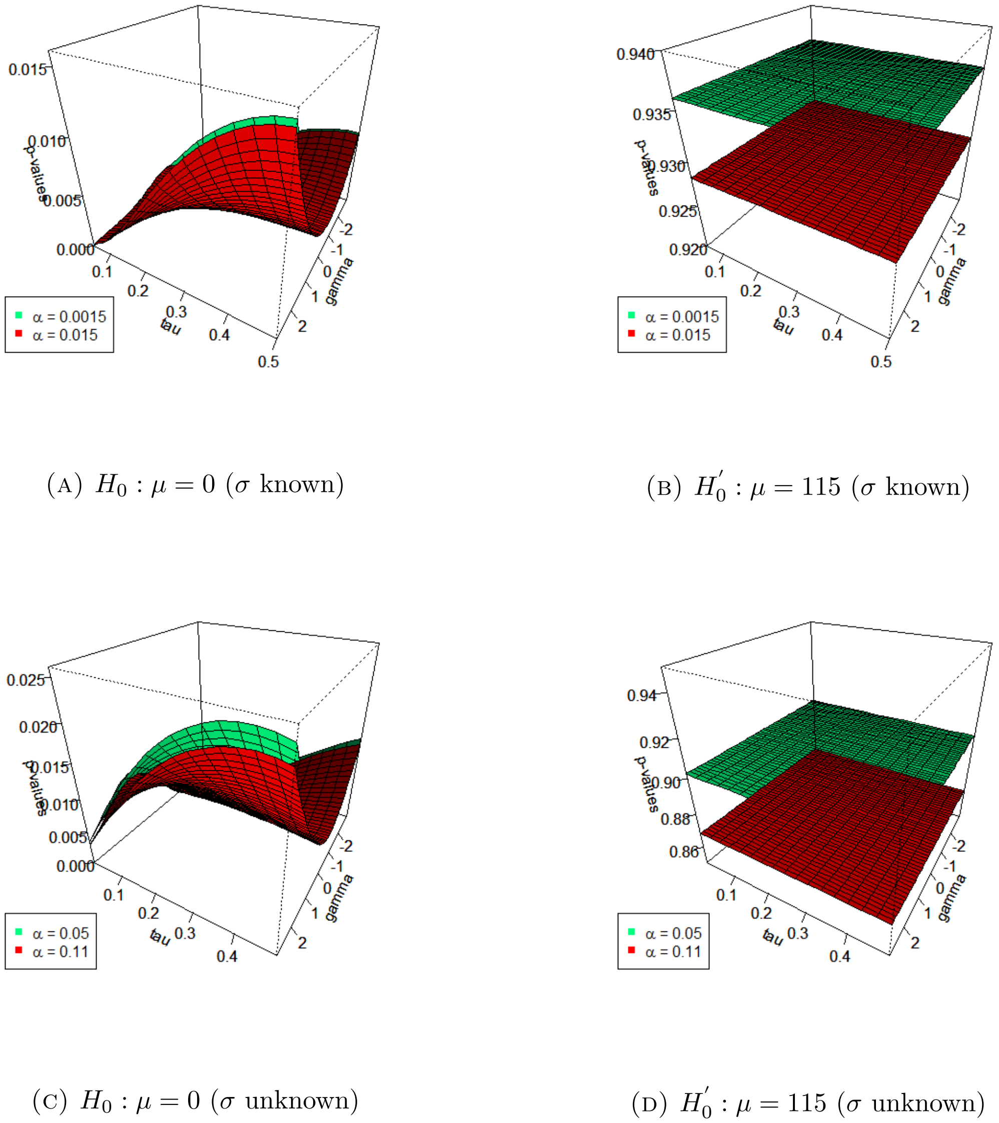
(A) $H_0 : \mu = 0$ (σ known)

(B) $H_0^{'} : \mu = 115$ (σ known)

(C) $H_0 : \mu = 0$ (σ unknown)

(D) $H_0^{'} : \mu = 115$ (σ unknown)

FIGURE 7.35: Plot of p-values of the GSDT (corresponding to two different α's) varying over different pairs of (τ, γ) in case of the Telephone Line Fault data (with outlier)



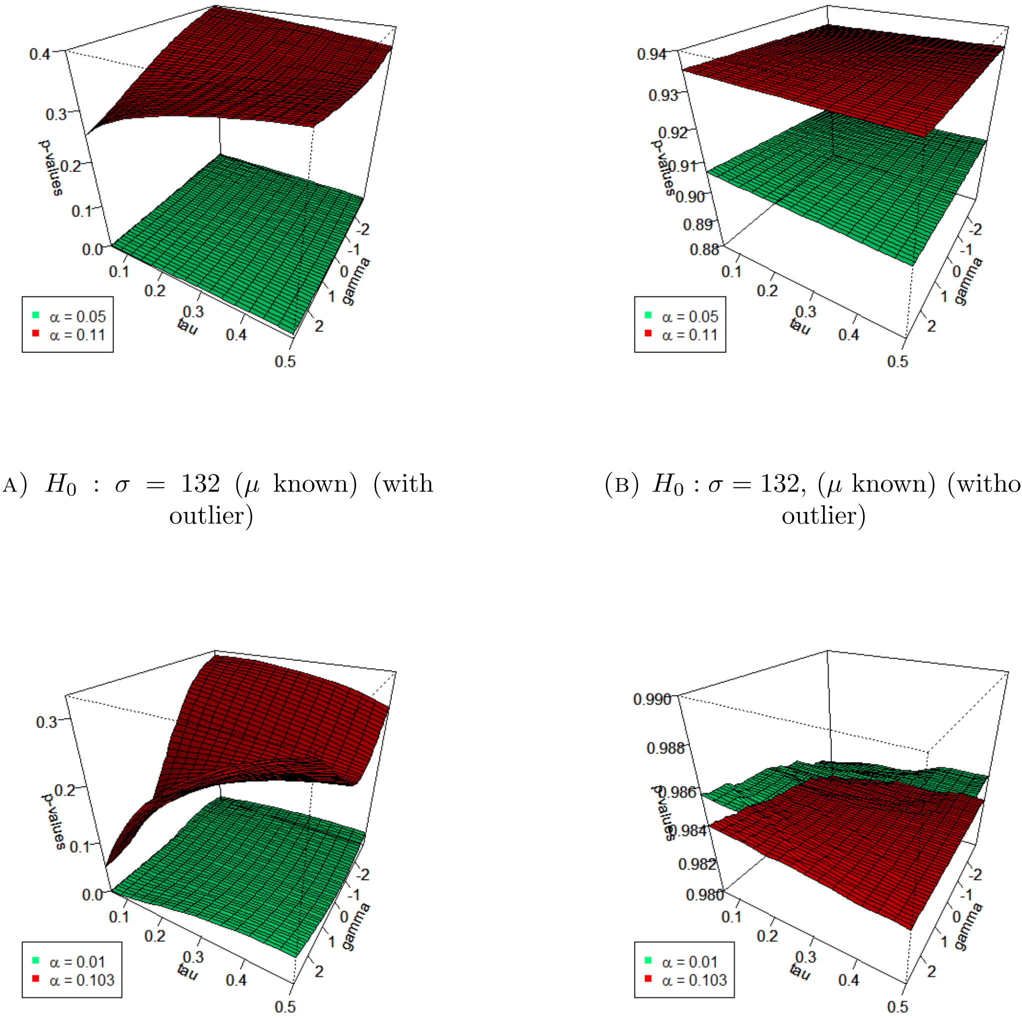
(A) $H_0 : \mu = 0$ (σ known)

(B) $H_0^{'} : \mu = 115$ (σ known)

(C) $H_0 : \mu = 0$ (σ unknown)

(D) $H_0^{'} : \mu = 115$ (σ unknown)

FIGURE 7.36: Plot of p-values of the GSDT (corresponding to two different α's) varying over different pairs of (τ, γ) in case of the Telephone Line Fault data (without outlier)



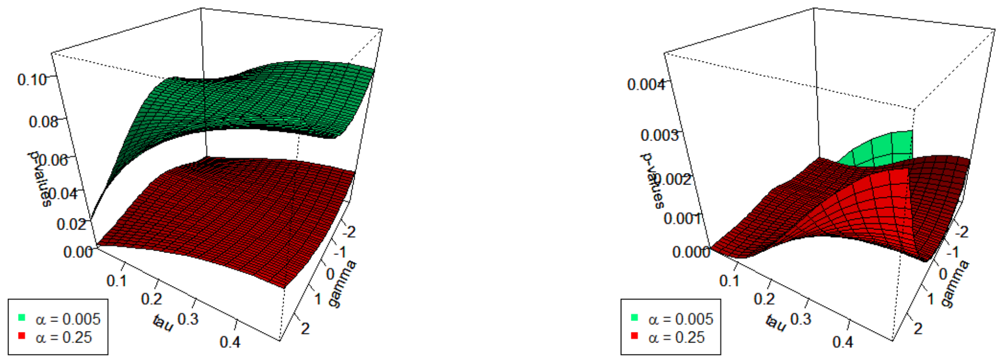
(A) $H_0 : \sigma = 132$ (μ known) (with outlier)

(B) $H_0 : \sigma = 132$, (μ known) (without outlier)

(C) $H_0 : (\mu, \sigma)^T = (115, 132)^T$ (with outlier)

(D) $H_0 : (\mu, \sigma)^T = (115, 132)^T$ (without outlier)

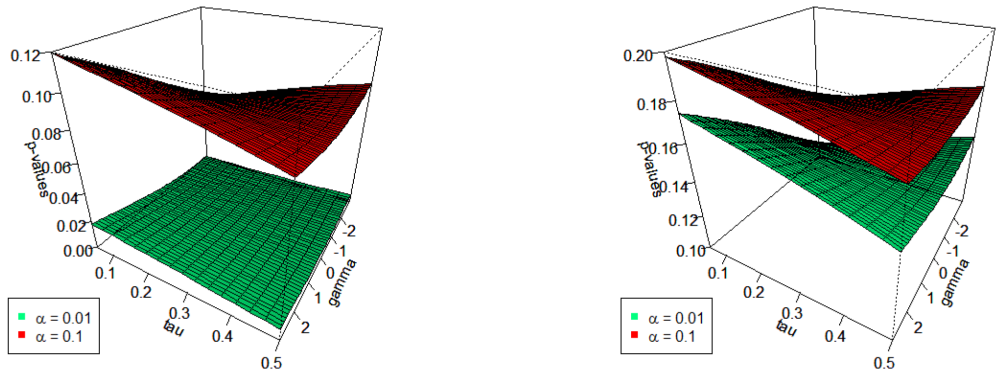
FIGURE 7.37: Plot of p-values of the GSDT (corresponding to two different α's) varying over different pairs of (τ, γ) in case of the Telephone Line Fault data (both with and without outlier)



(A) $H_0 : \mu = 0$ (σ unknown) (with outlier)

(B) $H_0 : \mu = 0$ (σ unknown) (without outlier)

Figure 7.38: Plot of p-values of the GSDT (corresponding to two different α's) varying over different pairs of (τ, γ) in case of the Darwin's Plant Fertilization data (both with and without outlier)



(A) $H_0 : \theta_1 = \theta_2$ (with outlier)

(B) $H_0 : \theta_1 = \theta_2$ (without outlier)

Figure 7.39: Plot of p-values of the GSDT (corresponding to two different α's) varying over different pairs of (τ, γ) in case of the Drosophila data (both with and without outlier)

7.7 Concluding Remarks

Ghosh and Basu (2018) has already demonstrated partially the worth of the GSD family through its application in the field of robust estimation. In this research, we have completed the remaining portion by illustrating its performance in the field of robust testing of hypotheses. We believe that we have already provided a complete and appropriate theoretical machinery regarding robust tests, based on the class of the Generalised Super Divergence (GSD); not only that, we have concluded it with extensive numerical studies and real data analyses. These help us to extend the path, generating robust tests with satisfactory results in terms of level and power, which, besides the classes of the SDT and the DPDT, can act as a very useful and robust alternative to our most conventional likelihood ratio test. We are wrapping up this chapter here with a hope that this GSDT will be extensively applied, which we will definitely endeavour towards in future.

Chapter 8

A review on the performance of the ‘optimal’ tuning parameter selection algorithm

No research can be an end in itself. Irrespective of the quality, depth and novelty of the research, there is always the possibility of refining, improving and extending it. This must also be the case with the tuning parameter selection issue, which is one of the most important highlights of this thesis. In this connection some additional words may be useful at this stage.

After the introduction of the iterated Warwick-Jones (IWJ) tuning parameter selection algorithm early in this thesis, all real data analysis based on the DPD or other extended Bregman divergences have been performed on the basis of this proposed algorithm. While we have all the theoretical and empirical indicators suggesting the superiority of our proposed algorithm in comparison to all other existing ones, some discussion is necessary about how good the performance of these “optimal” estimators is in relation to some reasonable fixed α procedures, and why we should always select the tuning parameter through this algorithm rather than use a fixed α. At the same time, a further case could be made to highlight the benefits of the iterative procedure (IWJ) compared to the one step procedure (OWJ).

While the following discussion will be a general one involving all divergence based estimators within the extended Bregman class (it will, in fact, be applicable to any robust parameter estimation method which depends on the choice of a tuning parameter), for streamlining the subsequent discussion we will use the DPD and the MDPDE as our platform of illustration. Let us recall that we have proposed the use of the tuning parameter α in the range $[0, 1]$. This is not on the basis of some objective criterion, but on the basis of common sense, as the choice of larger values of α lead to extremely poor model efficiencies. It is, of course, true that the there is a trade off between model efficiency and outlier stability which can be controlled by the tuning parameter α; however, this by itself, cannot lead to a unique choice of the tuning parameter. In fact, detailed analysis shows that in call cases (at least in all the models, sample sizes and contamination types studied by us) the mean square error (MSE) under contamination does not have a completely monotonic relation with the tuning parameter. The mean square initially drops with α under contamination, but at some point it picks up again, and shows an upward trend thereafter. This must be a consequence of the somewhat poor efficiency for large α estimators. Thus, while $\alpha = 1$ provides the greatest outlier down-weighting among the tuning parameters considered by us, it is not necessarily the estimator which provides the best performance under contamination, and, to some extent, leans towards partially down-weighting legitimate observations also. Moreover, in most data contamination examples that we have seen, the difference between the robust estimators belonging to the range $\alpha \in [0.5, 1]$ is not very high. Also notice that Ghosh and Basu (2013) had suggested that the estimator at $\alpha = 0.5$ (rather than at $\alpha = 1$) be selected as the pilot estimator when implementing the algorithm of Warwick and Jones (2005). There is also the issue that in relatively low sample sizes the choice of the empirical in place of the unknown truth leads to a certain amount of approximation.

To show that the algorithm we are proposing is doing well in both situations (under pure data and under contamination), it would, therefore, be meaningful to compare the performance of the estimator based on the "optimal tuning parameter" with the estimator based on the fixed tuning parameter $\alpha = 0.5$. We cannot realistically expect to beat the maximum likelihood estimator or estimators based on very low values of α under pure data, but our aim would be to make the estimator under study substantially close to the MLE under pure data compared to the estimator at $\alpha = 0.5$. In case of contaminated data, we hope to make the proposed estimator competitive to the MDPDE at $\alpha = 0.5$ and significantly improved over the MLE.

To this end, we performed a simulation study where the Poisson(θ) model is used. Data are first generated from the pure Poisson(3) distribution, where the sample size is 50, and the number of replications is 1000. Subsequently data are generated from the contaminated 0.9 Poisson(3) + 0.1 Poisson(15) distribution, but the estimator (i.e., θ) is calculated under the Poisson(θ) model. The mean square error of the estimators are computed against the target value of 3. The improvement due to the choice of the optimal tuning parameter is quite apparent (see Table 8.1). Under pure data, this optimal choice produces an estimator which is practically identical in performance to the estimator at $\alpha = 0.25$ (fixed), but is significantly superior in

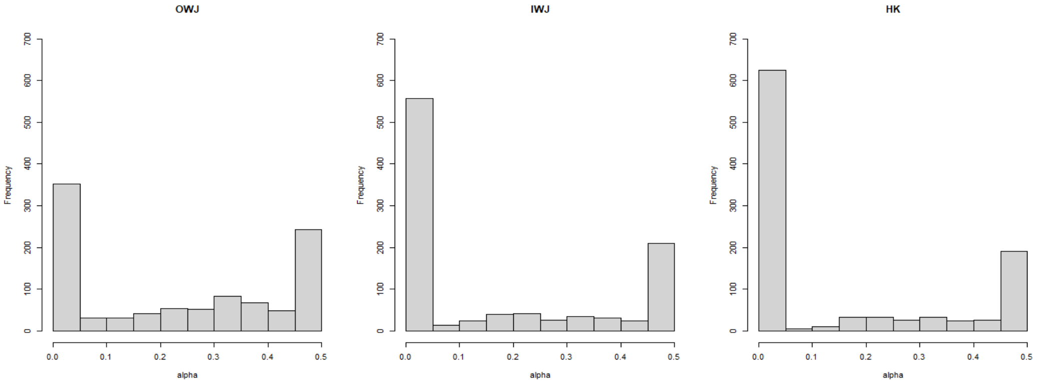


FIGURE 8.1: Histogram of sequence of optimal α values corresponding to three algorithms under 1000 simulations for pure model.

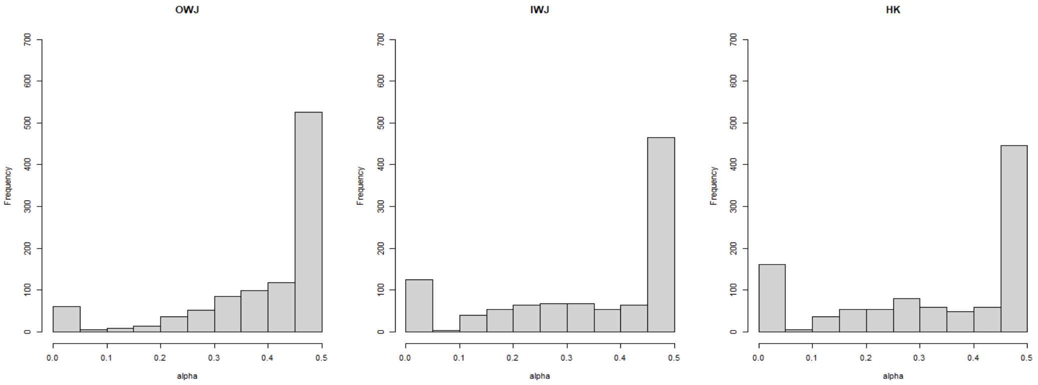


FIGURE 8.2: Histogram of sequence of optimal α values corresponding to three algorithms under 1000 simulations for contaminated model.

performance to the estimator at $\alpha = 0.5$ (fixed). On the other hand, under contamination, the optimal α estimator performs as well as the estimator at $\alpha = 0.5$, but is substantially better than the estimator at $\alpha = 0.25$ (and, of course, is far better than the MLE). Similar results are obtained when the results are repeated with different seed values.

TABLE 8.1: Mean Square Error under pure and contaminated Poisson data

	α			
	0	0.25	0.5	Optimum
Pure data	0.0607	0.0633	0.0693	0.0634
Contaminated Data	1.8238	0.1023	0.0838	0.0836

In the above calculations, we have searched for the optimal α is a fine grid on $[0, 0.5]$. Similar improvements will be possible for the MGSBDE with a judicious choice of the ranges of the tuning parameters. We have provided the estimators based on the optimal parameters for every data analysis example we have looked at. In the simulations we have not actually performed the optimal tuning parameter selection, since there are three tuning parameters for GSB, and if the complexity of the tuning parameter selection algorithm is of the order of m for the DPD, in the GSB case it will shoot up to m^3, and over 1000 or 10000 replications the computational burden would be very high.

Before we end this chapter, we provide a glimpse of the characteristics of the sequence of the optimal α's, in 1000 simulations under both pure and contaminated models, upon which the figures of Table 8.1 are based. Furthermore, to make a comparative study, the corresponding histograms are given in Figures 8.1 and 8.2, which we describe in the following.

Figure 8.1 provides the histograms for the optimal α values for the three algorithms under pure data over the 1000 replications. It may be observed, however, that the peak at α near zero is substantially shorter for the OWJ algorithm compared to the other two. As the OWJ algorithm starts from the most robust estimator within the class and only takes a single step, most of the time the corresponding optimal tuning parameter remains bounded away from $\alpha = 0$, even when the latter is the most desirable solution. Clearly the OWJ algorithm down-weights more than what is necessary in many (if not most) cases under pure data. On the other hand, the peaks of the histograms for the three algorithms under contaminated data (Figure

8.2) show that all the three algorithms have a very high mode around the largest value of α as one should expect. The peak of the OWJ algorithm is slightly higher than the other two with the additional cases, mostly representing the situations where high down-weighting was unnecessary.

The contrast between the IWJ and HK algorithms is less stark in these histograms compared to that between the two Warwick-Jones algorithms. The peaks (around zero and around the largest value of α) are approximately the same for both the algorithms, for both pure and contaminated data. However, even if the difference is slight, in both cases the peak of the HK algorithm for the largest α in the considered range is a little smaller than that of the IWJ. Further scrutiny reveals that each of these cases represent such situations where data down-weighting would have been appropriate, but the HK algorithm fails to provide that, indicating its occasional nonrobust behavior. Exactly the same situation was observed in Figure 3.14 earlier.

In summary, the range of the tuning parameters also need to be judiciously selected, part of which must be based on experience and empirical evidence. In practically all the situations we looked at in our simulations, including the pure data and the contaminated data cases, the optimally selected tuning parameter would lead to an estimator which would be among the best performing ones. From computing cost considerations we have refrained from actually doing so in all the simulations, but we trust that the simulations that we have presented help us in understanding how to choose the ranges of the individual tuning parameters, while the illustration given in this section provides a glimpse of how the optimal tuning parameter leads to a strong performance in all the scenarios.

Chapter 9

Epilogue

In the previous chapters, we have developed an extension to the popular Bregman divergence and explored its use in the field of robust estimation and testing - via theoretical approaches, simulation studies and real life scenarios. We have provided here a sufficient outline of this extension and the methodology of using it to construct several new divergences and to carry out their applications. In the subsequent portions of our epilogue, we will briefly sketch our future endeavours to conclude this journey.

9.1 Selection of Tuning Parameters

To introduce this extension, we have first started with the most pressing issue, that is, the choice of tuning parameter(s) in Chapter 3, because we have witnessed several times that despite using several robust tools (divergences), the analysis has become futile due to the dependence on tuning parameters and a trade-off between robustness and efficiency. So, keeping this in mind, we have refined the methodology of Warwick-Jones (2005) and combining its concept with the

Hong and Kim (2001) algorithm, we have developed the iterated one, which has been mentioned in this chapter. Later, we have implemented it on several i.i.d. discrete and continuous data examples and basic regression problems and in each case, we have got satisfactory results. In future, it would be worthwhile if one tries to explore its utility in more realistic problems, i.e, in case of generalized linear models. Lastly, we would like to mention that we have analyzed its performance through the DPD and the GSB divergence and one can also study its performance for other divergences as well.

9.2 Extended Bregman Divergence

Through the consideration of the exponent of arguments, we have extended the Bregman divergence in Chapter 4. Using this extension, we have shown that it becomes possible to bring the DPD (indexed by parameter α), the PD (indexed by parameter λ) and the S-divergence (indexed by parameters (α, λ)) families under one umbrella along with the BED (indexed by parameter β) family. In fact, we have developed a new super family of divergences through this extension – the GSB divergence (indexed by parameters (α, λ, β)) family. In future, through several choices of exponents and convex functions, new divergence families can be discovered and the scope of the usage of this extension will become much wider.

9.3 Robust Estimation

In Chapters 5 and 6 of this dissertation, we have explored its performance through the GSB divergence in the field of robust estimation.

We have come to the conclusion that within the GSB family, there are some choices of tuning parameters generating estimators which are quite competitive with respect to the DPD and S-divergence families, but all of them are lying outside these two sub-families. But most of the choices correspond to $\beta = -4$. Hence, an obvious future work is to find out whether there is any underlying theoretical reason for the best choice of β being equal to -4. Furthermore, we have applied our proposed IWJ algorithm on some i.i.d. examples through this GSB divergence for estimation purpose. In future, we shall apply the same for non-i.i.d. cases as well.

9.4 Robust Testing

In this section, we have considered the problem of constructing robust tests using the extended Bregman divergence. For this purpose, we have used another superfamily, the Generalized S-Divergence (GSD) family as our basic tool. Here also, through simulations and i.i.d. real data examples, we have explored the 'best' region providing several tests with robust size and satisfactory power, which should essentially lie outside the families of the DPDT and the SDT. We can explore the performance of the GSDT for non-i.i.d. cases also. Furthermore, as in the case of robust estimation, we believe that in the field of hypothesis testing also, the GSB divergence will contribute some significant tests through some specific choices of (α, λ, β), which we shall derive in future.

Appendix A

List of Published and Ongoing Papers

$\ast$ **Published Papers:**

- Basak, S., Basu, A. and Jones, M. C. (2021). On the 'optimal' density power divergence tuning parameter. *Journal of Applied Statistics*, 48:536-556.
- Basak, S. and Basu, A. (2022). The extended Bregman divergence and parametric estimation. *Statistics*, 56:699-718.

$\ast$ **Ongoing Papers:**

- Basak, S. and Basu, A.. The extended Bregman divergence and parametric estimation in continuous models.
- Basak, S. and Basu, A.. Hypotheses testing using the extended Bregman divergence.

Bibliography

[1] Agostinelli, C. and Markatou, M. (2001). Tests of hypotheses based on the weighted likelihood methodology. *Statistica Sinica*, 11:499–514.

[2] Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28:131–142.

[3] Amari, S. (2009). α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Transactions on Information Theory*, 55:4925–4931.

[4] Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23:193–212.

[5] Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton University Press, Princeton, New Jersey.

[6] Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005). Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749.

[7] Basu, A. (1991). *Minimum disparity estimation in the continuous case: efficiency, distributions, robustness and algorithms.* PhD thesis, The Pennsylvania State University, Pennsylvania, USA.

[8] Basu, A. (1993). Minimum disparity estimation: Application to robust tests of hypothesis. Technical report, Center for Statistical Sciences, The University of Texas at Austin, Texas, USA.

[9] Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85:549–559.

[10] Basu, A. and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46:683–705.

[11] Basu, A., Mandal, A., Martin, N., and Pardo, L. (2013). Testing statistical hypotheses based on the density power divergence. *Annals of the Institute of Statistical Mathematics*, 65:319–348.

[12] Basu, A., Mandal, A., Martin, N., and Pardo, L. (2018). Testing composite hypothesis based on the density power divergence. *Sankhya B*, 80:222–262.

[13] Basu, A., Shioya, H., and Park, C. (2011). *Statistical Inference: The Minimum Distance Approach.* CRC Press, Boca Raton.

[14] Beran, R. J. (1977). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5:445–463.

[15] Boos, D. D. (1981). Minimum distance estimators for location and goodness of fit. *Journal of the American Statistical Association*, 76:663–670.

[16] Boos, D. D. (1982). Minimum Anderson–Darling estimation. *Communications in Statistics: Theory and Methods*, 11:2747–2774.

[17] Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, 40:318–335.

[18] Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217.

[19] Chung, K. L. (1974). *A Course in Probability Theory.* Academic Press, Cambridge.

[20] Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46:440–464.

[21] Csiszár, I. (1963). Eine informations theoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von Markoffschen ketten. *Publ. Math. Inst. Hungar. Acad. Sci.*, 3:85–107.

[22] Csiszár, I. (1967a). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318.

[23] Csiszár, I. (1967b). On topological properties of f-divergences. *Studia Scientiarum Mathematicarum Hungarica*, 2:329–339.

[24] Dik, J. J. and de Gunst, M. C. M. (1985). The distribution of general quadratic forms in normal variables. *Statistica Neerlandica*, 39:14–26.

[25] Donoho, D. L. and Liu, R. C. (1988). The "automatic" robustness of minimum distance functionals. *The Annals of Statistics*, 16:552–586.

[26] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368.

[27] Ghosh, A. (2015a). Asymptotic properties of minimum S-divergence estimator for discrete models. *Sankhya A*, 77:380–407.

[28] Ghosh, A. (2015b). Influence function analysis of the restricted minimum divergence estimators: A general form. *Electronic Journal of Statistics*, 9:1017–1040.

[29] Ghosh, A. and Basu, A. (2013). Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. *Electronic Journal of Statistics*, 7:2420–2456.

[30] Ghosh, A. and Basu, A. (2015). Robust estimation for non-homogeneous data and the selection of the optimal tuning parameter: The density power divergence approach. *Journal of Applied Statistics*, 42:2056–2072.

[31] Ghosh, A. and Basu, A. (2017). The minimum S-divergence estimator under continuous models: The Basu-Lindsay approach. *Statistical Papers*, 58:341–372.

[32] Ghosh, A. and Basu, A. (2018). A new family of divergences originating from model adequacy tests and application to robust statistical inference. *IEEE Transactions on Information Theory*, 64:5581–5591.

[33] Ghosh, A., Basu, A., and Pardo, L. (2015). On the robustness of a divergence based test of simple statistical hypotheses. *Journal of Statistical Planning and Inference*, 161:91–108.

[34] Ghosh, A., Harris, I. R., Maji, A., Basu, A., and Pardo, L. (2017). A generalized divergence for statistical inference. *Bernoulli*, 23:2746–2783.

[35] Ghosh, A., Mandal, A., Martin, N., and Pardo, L. (2016). Influence analysis of robust Wald-type tests. *Journal of Multivariate Analysis*, 147:102–126.

[36] Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11:1–21.

[37] Gutmann, M. and Hirayama, J. (2012). Bregman divergence as general framework to estimate unnormalized statistical models. *arXiv preprint arXiv:1202.3727*.

[38] Hampel, F. R. (1968). *Contributions to the theory of robust estimation*. PhD thesis, The University of California, Berkeley, California, USA.

[39] Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42:1887–1896.

[40] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–393.

[41] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley and Sons, New York.

[42] Hellinger, E. D. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271.

[43] Hong, C. and Kim, Y. (2001). Automatic selection of the tuning parameter in the minimum density power divergence estimation. *Journal of the Korean Statistical Society*, 30:453–465.

[44] Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35:73–101.

[45] Huber, P. J. (1965). A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36:1753–1758.

[46] Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–233.

[47] Huber, P. J. (1968). Robust confidence limits. *Z. Wahr. Verw. Geb.*, 10:269–278.

[48] Huber, P. J. (1970). Studentizing robust estimates. *In: Puri M. L. (ed), Nonparametric Techniques in Statistical Inference.* Cambridge University Press, Cambridge.

[49] Huber, P. J. (1972). The 1972 Wald Lecture Robust Statistics: A Review. *The Annals of Mathematical Statistics*, 43:1041–1067.

[50] Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1:799–821.

[51] Huber, P. J. (1975). Robustness and designs. *In: Srivastava J. N. (ed), A Survey of Statistical Design and Linear Models.* North Holland Publishing Company, Amsterdam.

[52] Huber, P. J. (1981). *Robust Statistics.* John Wiley and Sons, New York.

[53] Jana, S. and Basu, A. (2019). A characterization of all single-integral, non-kernel divergence estimators. *IEEE Transactions on Information Theory*, 65:7976–7984.

[54] Kang, J. and Lee, S. (2014). Minimum density power divergence estimator for Poisson autoregressive models. *Computational Statistics & Data Analysis*, 80:44–56.

[55] Kotz, S., Johnson, N. L., and Boyd, D. W. (1967a). Series representations of distributions of quadratic forms in normal variables. I. Central case. *The Annals of Mathematical Statistics*, 38:823–837.

[56] Kotz, S., Johnson, N. L., and Boyd, D. W. (1967b). Series representations of distributions of quadratic forms in normal variables. II. Non-central case. *The Annals of Mathematical Statistics*, 38:838–848.

[57] Kuchibhotla, A. K. and Basu, A. (2015). A general set up for minimum disparity estimation. *Statistics & Probability Letters*, 96:68–74.

[58] Kuchibhotla, A. K. and Basu, A. (2017). On the asymptotics of minimum disparity estimation. *TEST*, 26:481–502.

[59] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.

[60] Le Cam, L. M. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *University of California Publications in Statistics*, 1:277–330.

[61] Lehmann, E. L. (1983). *Theory of Point Estimation.* John Wiley and Sons, New York.

[62] Liese, F. and Vajda, I. (1987). *Convex Statistical Distances.* Teubner, Leipzig.

[63] Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics*, 22:1081–1114.

[64] Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4:51–67.

[65] Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics.* John Wiley and Sons, Chichester.

[66] Maronna, R. A. and Yohai, V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90:330–341.

[67] Maronna, R. A. and Yohai, V. J. (2004). Robust estimation of multivariate location and scatter. *Encyclopedia of Statistical Sciences*, 11.

[68] Mickey, M. R., Dunn, O. J., and Clark, V. (1967). Note on the use of stepwise regression in detecting outliers. *Computers and Biomedical Research*, 1:105–111.

[69] Moore, D. S. and McCabe, G. P. (1999). *Introduction to the Practice of Statistics.* W. H. Freeman, New York.

[70] Mukherjee, T., Mandal, A., and Basu, A. (2019). The B-exponential divergence and its generalizations with applications to parametric estimation. *Statistical Methods & Applications*, 28:241–257.

[71] Nelson, W. (1982). *Applied Life Data Analysis.* John Wiley and Sons, New York.

[72] Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 20:175–240.

[73] OECD (2017). *Health at a Glance 2017: OECD Indicators.* OECD Publishing, Paris. `http://dx.doi.org/10.1787/health_glance-2017-en`.

[74] Pardo, L. (2006). *Statistical Inference Based on Divergence Measures.* Chapman & Hall/CRC, New York.

[75] Park, C. and Basu, A. (2003). The generalized Kullback-Leibler divergence and robust inference. *Journal of Statistical Computation and Simulation*, 73:311–332.

[76] Park, C. and Basu, A. (2004). Minimum disparity estimation: Asymptotic normality and breakdown point results. *Bulletin of Informatics and Cybernetics*, 36:19–33.

[77] Park, J. H. and Sriram, T. N. (2017). Robust estimation of conditional variance of time series using density power divergences. *Journal of Forecasting*, 36:703–717.

[78] Parr, W. C. and De Wet, T. (1981). Minimum CVM-norm parameter estimation. *Communications in Statistics: Theory and Methods*, 10:1149–1166.

[79] Parr, W. C. and Schucany, W. R. (1980). Minimum distance and robust estimation. *Journal of the American Statistical Association*, 75:616–624.

[80] Parr, W. C. and Schucany, W. R. (1982). Minimum distance estimation and components of goodness-of-fit statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44:178–189.

[81] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50:157–175.

[82] Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 44:50–57.

[83] Rao, C. R. (1973). *Linear Statistical Inference and its Applications.* John Wiley and Sons, New York.

[84] Robinson, J., Ronchetti, E. M., and Young, G. A. (2003). Saddlepoint approximations and tests based on multivariate M-estimates. *The Annals of Statistics*, 31:1154–1169.

[85] Ronchetti, E. M. (1982). *Robust testing in linear models: the infinitesimal approach.* PhD thesis, ETH, Zurich.

[86] Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *In: Grossmann W., Pflug G., Vincze I., Wertz W. (eds), Mathematical Statistics and Applications*, 8:283–297. Reidel Publishing Company, Dordrecht.

[87] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* John Wiley and Sons, New York.

[88] Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. *In: Franke J., Härdle W., Martin D. (eds), Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics*, 26:256–272. Springer, New York.

[89] Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, 75:828–838.

[90] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley and Sons, New York.

[91] Simpson, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *Journal of the American Statistical Association*, 82:802–807.

[92] Simpson, D. G. (1989). Hellinger deviance tests: Efficiency, breakdown points, and examples. *Journal of the American Statistical Association*, 84:107–113.

[93] Stigler, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics*, 5:1055–1098.

[94] Tamura, R. N. and Boos, D. D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, 81:223–229.

[95] Thode Jr., H. C. (2002). *Testing For Normality.* Marcel Dekker, New York.

[96] Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *In: Olkin I., Ghurye S. G., Hoeffding W., Madow W. G., Mann H. B. (eds), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling.* Stanford University Press, Redwood City.

[97] von Mises, R. (1936). Les lois de probabilité pour les fonctions statistiques. *Annales de l'institut Henri Poincaré*, 6:185–212.

[98] von Mises, R. (1937). Sur les fonctions statistiques. *Soc. Math. de France, Conference de la Reunion Internat. de Mathem, Paris, France.*

[99] von Mises, R. (1939). Sur les fonctions statistiques. *Bulletin de la Société Mathématique de France*, 67:177–184.

[100] von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18:309–348.

[101] Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54:426–482.

[102] Wang, Y. G., Lin, X., Zhu, M., and Bai, Z. (2007). Robust estimation using the Huber function with a data-dependent tuning constant. *Journal of Computational and Graphical Statistics*, 16:468–481.

[103] Warwick, J. and Jones, M. C. (2005). Choosing a robustness tuning parameter. *Journal of Statistical Computation and Simulation*, 75:581–588.

[104] Wiens, D. P. (1987). Robust weighted Cramér-von Mises estimators of location, with minimax variance in ϵ-contamination neighbourhoods. *The Canadian Journal of Statistics*, 15:269–278.

[105] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9:60–62.

[106] Wolfowitz, J. (1952). Consistent estimators of the parameters of a linear structural relation. *Scandinavian Actuarial Journal*, 3-4:132–151.

[107] Wolfowitz, J. (1953). Estimation by the minimum distance method. *Annals of the Institute of Statistical Mathematics*, 5:9–23.

[108] Wolfowitz, J. (1957). The minimum distance method. *The Annals of Mathematical Statistics*, 28:75–88.

[109] Woodruff, R. C., Mason, J. M., Valencia, R., and Zimmering, S. (1984). Chemical mutagenesis testing in Drosophila: I. Comparison of positive and negative control data for sex-linked recessive lethal mutations and reciprocal translocations in three laboratories. *Environmental Mutagenesis*, 6:189–202.

[110] Yohai, V. J. and Maronna, R. A. (1979). Asymptotic behaviour of M-estimators for the linear model. *The Annals of Statistics*, 7:258–268.