DOCTORAL THESIS

Robust Matrix Factorization using the Density Power Divergence and its Applications

Author
Subhrajyoty Roy

Supervisors
Ayanendranath Basu
&
Abhik Ghosh

A thesis submitted to the Indian Statistical Institute in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Statistics

Interdisciplinary Statistical Research Unit
Applied Statistics Division

Declaration

I declare that

1. The research work presented in this thesis titled "Robust Matrix Factorization using the Density Power Divergence and its Applications" is original and has been done by myself under the general guidance of my advisors, Prof. Ayanendranath Basu and Dr. Abhik Ghosh.

2. This thesis represents my ideas in my own words. Whenever I have used ideas or materials (text, data, theoretical analysis) from other sources, I have adequately cited and referenced the original source.

3. I have adhered to all principles of academic honesty and integrity in my submission.

4. This work has not been submitted to any other Institute or University for a degree.

5. I have followed the guidelines provided by the Institute in writing this thesis.

Place: Indian Statistical Institute, Kolkata
Date: June 5, 2025

Subhrajyoty Roy
Research Fellow (Statistics)
Interdisciplinary Statistical Research Unit
Applied Statistics Division
Indian Statistical Institute
Kolkata, 700108, India.

Abstract

In the modern era of big data, high-dimensional datasets are becoming increasingly common across a range of disciplines, including machine learning, natural language processing, finance, and genomics. Extracting meaningful information from these datasets often requires uncovering low-dimensional structures hidden within the data. Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) are widely used matrix factorization techniques for this purpose. However, the traditional methods to compute these are extremely sensitive to outliers, with even a single aberrant observation potentially leading to highly imprecise results. This issue is exacerbated in high-dimensional datasets, where outliers are difficult to detect. Classical robust inference techniques, such as M-estimators, struggle due to their diminishing breakdown points as the data dimension becomes extremely large.

This thesis addresses these challenges by proposing a novel class of robust matrix factorization techniques based on the minimum density power divergence estimator (MDPDE). The MDPDE, a member of the broader class of minimum divergence estimators, is well-known for its robustness and efficiency across diverse applications. Crucially, it offers a dimension-free asymptotic breakdown point, making it particularly well-suited for high-dimensional settings. In this work, we leverage this estimator to develop robust versions of SVD and PCA, referred to as rSVDdpd and rPCAdpd, respectively.

The thesis is structured as follows: In Chapter 1, we provide the necessary background on classical matrix factorization techniques, introduce key concepts related to minimum divergence estimators, particularly the MDPDE, and the notations to be used throughout the thesis. Chapter 2 presents the novel rSVDdpd algorithm, detailing its theoretical properties, including different equivariance properties, algorithmic convergence and consistency. Through simulation studies, we demonstrate the algorithm's superior robustness compared to existing methods, particularly in high-dimensional settings. We also apply the rSVDdpd algorithm to the problem of video surveillance background modelling, showcasing its real-world applicability.

Chapter 3 extends this methodology to robust PCA, resulting in the rPCAdpd algorithm. We establish its theoretical properties such as orthogonal equivariance, consistency and asymptotic normality. We also demonstrate that its influence function remains bounded, ensuring its robustness to outliers. Comparative studies with benchmark datasets reveal that rPCAdpd outperforms existing robust PCA algorithms, particularly in scenarios with high-dimensional data with a low signal-to-noise ratio.

The robust SVD and the PCA algorithms introduced in Chapters 2 and 3 require a robust estimate of the rank of the low-dimensional component of the data matrix. To this end, we propose a new penalized criterion, DICMR, in Chapter 4. Theoretical results on selection consistency and B-robustness are established, and extensive simulation studies show that DICMR is the best-performing among penalized methods, and also provides competitive performance relative to cross-validation methods while being computationally efficient.

A key contribution of this thesis, explored in Chapter 5, is the demonstration that the MDPDE has a dimension-free lower bound to its asymptotic breakdown point. This property makes it uniquely robust in high-dimensional settings, a significant improvement over classical M-estimators. We further generalize this result in Chapter 6, showing that the dimension-free breakdown point holds for a broader class of estimators known as minimum generalized Alpha-Beta divergence estimators. We derive the necessary and sufficient conditions under which the corresponding divergence measures are well-defined and nonnegative, contributing to the theoretical understanding of generating novel statistical divergence measures that may lead to robust estimation in high-dimensional data.

Chapter 7 concludes the thesis, summarizing the key findings and outlining directions for future research. This includes potential extensions of the proposed algorithms to other matrix factorization problems and the exploration of more practical applications beyond those demonstrated in the thesis.

Overall, this thesis aims to contribute to the field of robust statistics by developing scalable, robust matrix factorization techniques with strong theoretical guarantees and practical relevance in high-dimensional data analysis.

Acknowledgements

"We are smart, but not because we stand on the shoulders of giants or are giants ourselves. We stand on the shoulders of a very large pyramid of hobbits." — Joseph Henrich

As I reach near the end of my doctoral journey, I am very much aware that this thesis stands on the firm foundation built by countless researchers, whose work has shaped the field of statistics and paved the way for my own contributions. Equally important are the many people who have inspired, guided, and shaped me into the person I am today. While I would like to thank each of them individually, both space and words would fall short of expressing my full gratitude, so I will mention only a few here.

First and foremost, I would like to express my deepest gratitude to my supervisors, Prof. Ayanendranath Basu and Dr. Abhik Ghosh, for their invaluable guidance, support, and encouragement throughout the course of my PhD journey. Their insights, patience, and unwavering belief in my abilities have been instrumental in shaping this work, and I could not have asked for better mentors.

I am also deeply thankful to the members of my research fellow advisory committee (Prof. Bimal Roy, Dr. Partha Sarathi Mukherjee, Dr. Shyamal Krishna De, Prof. Sumitra Purakayastha, etc.) for their insightful feedback during the annual evaluation meetings. I extend my gratitude to the Dean of Studies (Prof. Debasis Sengupta and Prof. Gopal Krishna Basak) for their cooperation and assistance with the administrative processes.

I have been extremely fortunate to have had the mentorship of Vijay Krishna, CEO and founder of SysCloud Technologies Pvt. Ltd., who generously allowed me to pursue my research alongside my full-time responsibilities at SysCloud. His kindness in providing me the flexibility to work from home, coupled with his continuous encouragement toward academic excellence throughout my PhD, is something for which I am immensely grateful. No words could adequately express my thanks for his trust, mentorship, and the remarkable opportunities he has provided.

I am deeply indebted to Prof. Arnab Chakraborty, Prof. Smarajit Bose, the late Prof. Saurabh Ghosh, Prof. Rita Saha Ray, Prof. Diganta Mukherjee, and Prof. Arijit Chakrabarti for their invaluable teachings during my B.Stat. and M.Stat. years. From an 18-year-old boy in a government-aided local school who had been unsuccessful at the re-

gional mathematical olympiads due to his (then) limited mathematical knowledge, to now completing a doctoral degree in statistics, I owe much of my growth and understanding to these mentors. I also wish to thank the then Dean of Studies Prof. Pradipta Bandhopadhyay, whose words during my B.Stat. orientation instilled in me the disciplined habit of daily self-improvement, a principle that has paid immense dividends over the years.

To my friends, both near and far, your companionship has been a source of joy and balance in my life, and I am truly grateful for your constant support and encouragement. Especially, Monitirtha — without your help, this journey would have been far more difficult.

I am also thankful to Arijit da, Abir, and, Supratik for their invaluable contributions to various aspects of my research. Our collaborations have been both fruitful and intellectually stimulating, and I look forward to continuing our work together in the future. I also extend my gratitude to the editors, associate editors, and reviewers of various journals for their thoughtful feedback and constructive criticisms, which have greatly enhanced the quality of this work.

To my parents, I owe a debt of gratitude for your countless sacrifices, unwavering love, support, and belief in me. Your constant encouragement to make myself better has been the bedrock of all my achievements. To my elder brother, thank you for always being there for me during the moments of hardships and challenges. Words cannot do justice to the immense contributions you all have made to my life.

I dedicate this thesis to my late grandmother, who passed away recently at the age of ninety-two after living a long and fulfilling life. It is deeply inspiring to me that, approximately seventy-five years ago, she excelled in her higher secondary examinations, defying the prevailing cultural norms that discouraged education for women.

Finally, I thank God for granting me the serenity to accept the hardships I could not control, for granting me the courage to face and overcome the challenges I could, and for granting me the wisdom to discern between the two.

Contents

1 Background **2**

1.1 Introduction . . . 2
1.2 Notations . . . 3
1.3 Mathematical Preliminaries . . . 6
1.3.1 Singular Value Decomposition . . . 6
1.3.2 Principal Component Analysis . . . 7
1.3.3 Breakdown Point . . . 9
1.3.4 Density Power Divergence . . . 10
1.3.5 S-divergence . . . 12
1.4 Organization of the thesis . . . 13

2 Robust Singular Value Decomposition **15**

2.1 Introduction and Existing Literature . . . 15
2.2 The rSVDdpd Estimator . . . 17
2.2.1 Problem Statement . . . 17
2.2.2 Derivation . . . 18
2.2.3 Estimation Algorithm . . . 19
2.2.4 Choice of the Hyperparameters . . . 22
2.3 Mathematical Analysis . . . 23
2.3.1 Uniqueness . . . 27
2.3.2 Equivariance Properties . . . 28
2.3.3 Consistency . . . 31
2.4 Simulation Studies . . . 39
2.4.1 Simulation Setups . . . 40
2.4.2 Simulation Results . . . 41
2.4.3 Effect of matrix dimensions on rSVDdpd estimates . . . 44
2.5 Application to Background Modelling . . . 45
2.5.1 Background Modelling Challenger Dataset . . . 48
2.5.2 University of Houston Camera Tampering Dataset . . . 52
2.6 Example: Factor Analysis of Spatiotemporal Data . . . 53

3 Robust Principal Component Analysis **58**
3.1 Introduction . . . 58
3.2 Existing Works on Robust PCA . . . 58
3.3 The rPCAdpd Estimator . . . 61
3.3.1 Description . . . 61
3.3.2 Efficient Computation of the rPCAdpd Estimator . . . 62
3.3.3 Choice of the Robust Location Estimator . . . 63
3.3.4 Choice of the Hyperparameters . . . 64
3.4 Theoretical Properties . . . 64
3.4.1 Existence and Convergence . . . 64
3.4.2 Orthogonal Equivariance . . . 66
3.4.3 Consistency and Asymptotic Distribution . . . 67
3.4.4 Influence Function Analysis . . . 76
3.4.5 Breakdown Point Analysis . . . 77
3.5 Simulation Studies . . . 78
3.5.1 Simulation Settings . . . 79
3.5.2 Simulation Results . . . 80
3.6 Real Data Analysis . . . 85
3.6.1 Car Dataset . . . 85
3.6.2 Octane Data . . . 88
3.6.3 Credit Card Fraud Detection . . . 90
3.A Appendix: Proofs of some additional results . . . 93

4 Rank Estimation **95**
4.1 Introduction . . . 95
4.2 Problem Description and Existing Approaches . . . 96
4.2.1 Penalization Approaches . . . 97
4.2.2 Cross-validation Approaches . . . 99
4.3 Divergence Information Criterion for Matrix Rank Estimation . . . 102
4.4 Theoretical Studies . . . 103
4.4.1 Selection Consistency . . . 103
4.4.2 Robustness . . . 108
4.5 Simulation Studies . . . 110
4.5.1 Simulation Setups . . . 110
4.5.2 Simulation Results . . . 111
4.A Appendix: Detailed Simulation Results . . . 112

5 Breakdown Analysis of Minimum Super Divergence Estimator **121**
5.1 Introduction . . . 121
5.2 Existing Literature . . . 121
5.3 Theoretical Analysis of Breakdown Point . . . 123
5.3.1 Main Results . . . 125

5.3.2 Asymptotic Breakdown point of minimum S-divergence estimator . . 138
5.4 Examples and Illustrations . . . 143
5.4.1 Normal Location Model (Univariate) . . . 143
5.4.2 Normal Scale Model (Univariate) . . . 143
5.4.3 Normal Location Model and Scale Model (Multivariate) . . . 144
5.4.4 Gamma distribution (Scale Family) . . . 144
5.4.5 Gamma distribution (Shape Family) . . . 145
5.4.6 Exponential Distribution with Normal Contamination . . . 146
5.5 Empirical Studies . . . 148
5.5.1 Normal Model Family . . . 148
5.5.2 Gamma Distribution (Shape Family) . . . 152
5.5.3 Exponential Distribution (Scale Parameter) . . . 153
5.5.4 Binomial Model . . . 154
5.A Appendix: Proof of the useful lemmas . . . 156

6 Breakdown Analysis of Minimum Generalized Alpha-Beta Divergence Estimator 159
6.1 Introduction . . . 159
6.2 Generalized Alpha-Beta Divergence Family . . . 160
6.3 Characterization of GAB Divergence Family . . . 161
6.3.1 GAB divergence for $\alpha + \beta = 1$ . . . 162
6.3.2 Necessary and Sufficient Condition for nonnegativity of GABD for non-limiting cases . . . 164
6.3.3 Necessary and Sufficient Condition for nonnegativity of GABD for the limiting cases . . . 167
6.4 Asymptotic Breakdown Results . . . 169
6.4.1 Asymptotic Breakdown Point of MGABD functional for positive α and β . . . 170
6.4.2 Asymptotic Breakdown Point of MGABD functional for the case $\alpha > 0, \beta = 0$ . . . 178
6.5 Examples and Illustrations . . . 182
6.5.1 Normal Distributions . . . 182
6.5.2 Gamma Distribution (Scale Family) . . . 183
6.5.3 Gamma Distribution (Shape Family) . . . 184
6.6 Numerical Studies . . . 185
6.6.1 Normal Location and Scale Family . . . 186
6.6.2 Exponential Model . . . 187
6.6.3 Binomial Model . . . 188
6.6.4 Geometric Model . . . 190
6.A Appendix: Proofs of additional results . . . 193

7 Conclusion and Future Scopes 195

List of Manuscripts **198**

List of Figures

1.1 A 2-dimensional dataset with data points lying approximately on a one dimensional manifold of a circle, and a few outliers in the center which are not distinguishable as outliers in the one dimensional linear projections of the data. 3

2.1 Empirical Bias and RMSE of the first singular value estimated by rSVDdpd algorithm with $\alpha = 0.5$ of an $n \times n$ matrix comprising a matrix of rank three and an error matrix (Both the horizontal and vertical axes are in logarithmic scale). 45

2.2 The estimated bias and the RMSE of different SVD algorithms under change in the number of columns while the number of rows and the matrix rank are kept fixed; (y-axis is in the log-scale). The group of lines on the top contains performance for svd, rSVD, pcaOne and RobSVD methods, while the group of lines at the bottom consists of performances for the rSVDdpd methods for different choices of α. 46

2.3 The estimated bias and the RMSE of different SVD algorithms under change in the rank of low-rank component while the dimension of the data matrix is kept fixed; (y-axis is in the log-scale). The group of lines on the top contains performance for svd, rSVD, pcaOne and RobSVD methods, while the group of lines at the bottom consists of performances for the rSVDdpd methods for different choices of α. 46

2.4 One frame of the street video (a), ground truth (b) and estimated foreground mask from exact RPCA (c), inexact ALM (d), SRPCP (e), Variational Bayes (f), Outlier Pursuit (g), GoDec (h), GRASTA (i) and rSVDdpd (j) algorithms. 51

2.5 One frame of the rotary video (a), ground truth (b) and estimated foreground mask from exact RPCA (c), inexact ALM (d), SRPCP (e), Variational Bayes (f), Outlier Pursuit (g), GoDec (h), GRASTA (i) and rSVDdpd (j) algorithms. 52

2.6 True video frames and the estimated foreground by usual SVD, rSVDdpd, ALM and GoDec algorithms (Left to Right) for frame 5, 50, 75 in UHCTD Day 1 Camera B dataset. 53

2.7 True video frames and the estimated foreground by usual SVD, GoDec, Variational Bayes (VB), Sparse Regularized PCP and rSVDdpd algorithms (Top to Bottom) for frame 50, 100, 150, 200 in UHCTD Day 1 Camera A dataset. 54

2.8 [Top] The time series graph highlighting the daily average PM2.5 measurements in 3 stations; y-axis is in log scale. [Bottom] The histogram indicating the distribution of the PM2.5 measurements in those three stations. 55

2.9 The signal strength obtained from the rSVDdpd estimate of the low rank component $\boldsymbol{L}$ for the transformed data with extended Yeo-Johnson transformations for different values of λ_p and λ_n. 57

2.10 Histogram of the transformed PM2.5 measurements after applying extended Yeo-Johnson transformation with $\lambda_p = 0$ and $\lambda_n = 2$. 57

3.1 Diagnostic plots for the Car dataset. 86

3.2 Monitoring plot for the Cars dataset. 87

3.3 Monitoring plot for first principal eigenvalue of the Cars dataset. 88

3.4 Diagnostic plots for the Octane dataset. 89

3.5 Monitoring plot for the Octane dataset. 90

3.6 Monitoring plot for first principal eigenvalue of the Octane dataset. 91

3.7 Diagnostic plots for the Credit Card dataset for different robust PCA methods. 91

5.1 Lower bound $\widetilde{\epsilon}_{(\alpha,\lambda)}$ as in Corollary 5.2 of the asymptotic breakdown point of MSDE for different choices of α and λ. The black regions indicate either $A < 0$ or $B < 0$. The dotted green line indicates the implicit curve $(B/(1+\alpha))^{1/A} = 1/2$. 135

5.2 Behaviour of MSDE (including MDPDE) under normal location model as a function of the contamination for different choices of α and λ (denoted in the title of individual plots). 150

5.3 Behaviour of MDPD estimates under normal location-scale model as a function of the contamination proportion ϵ, with the location parameter in the left panel and the scale parameter in the right panel. 151

5.4 Behaviour of MSDE under normal location-scale model as a function of the contamination proportion ϵ, with the location parameter in the left panel and the scale parameter in the right panel, for different choices of λ (denoted in the title of individual plots). 151

5.5 Behaviour of MDPD estimates under gamma density model as a function of the contamination proportion ϵ, for two values of shape parameter $t_0 = 10$ (Left) and $t_0 = 0.001$ (Right). 152

5.6 Behaviour of MDPD estimates under exponential model as a function of the contamination proportion ϵ, for $\theta_0 = 10$ (Left) and $\theta_0 = 0.01$ (Right). . 153

5.7 Minimum S-divergence estimate of success proportion under different levels of contamination in the binomial model for different choices of α and λ (λ is denoted in the title of each individual plot). 155

6.1 Behaviour of the MLSD functional for change in contaminating proportion ϵ for location estimation in the normal family. 186
6.2 Behaviour of the MLSD functional for change in contaminating proportion ϵ for scale estimation in the normal family when the contaminating density is normal with zero mean and large variance. 187
6.3 Behaviour of the MLSD functional for change in contaminating proportion ϵ for scale estimation in the normal family when the contaminating density is normal with zero mean and a very small variance. 188
6.4 Behaviour of the MLSD functional for change in contaminating proportion ϵ for exponential densities with a very large value of contaminating inverse-scale parameter. 189
6.5 Behaviour of the MLSD functional for change in contaminating proportion ϵ for exponential densities with a very small value of contaminating inverse-scale parameter. 189
6.6 Behaviour of the MLSD functional for change in contaminating proportion ϵ for binomial model with a binomial with success probability 1. 190
6.7 Behaviour of the MLSD functional for change in contaminating proportion ϵ for the Geometric model when the contaminating density is Geometric with very low success probability. 191
6.8 Behaviour of the MLSD functional for change in contaminating proportion ϵ for the Geometric model when the contaminating density is Geometric with very high success probability. 192

List of Tables

2.1 Summary of performance measures (total squared bias, total MSE of singular values, total dissimilarity of left and right singular vectors, denoted by Left and Right respectively) and the average runtime (in milliseconds) for the existing SVD and robust SVD algorithms. 42

2.2 Summary of average time taken (in milliseconds) to get the first singular value and vectors of an $n \times 25$ matrix with uniformly distributed entries via different SVD algorithms. 44

2.3 Benchmark Results for the Street Video (Video number $Vi1j$ denotes i-th variant of the Street Video and $j = 1$ or 2 denote the training and the testing video respectively); In each column, bold symbols indicate the algorithm with best F_1 measure. 49

2.4 Benchmark Results for the Rotary Video (Video number $Vi2j$ denotes i-th variant of the Rotary Video and $j = 1$ or 2 denote the training and the testing video respectively); In each column, bold symbols indicate the algorithm with best F_1 measure. 50

2.5 Processing time (in seconds per frame) for benchmark videos in BMC dataset for different background modelling algorithms. 50

3.1 The choices of $g(\cdot)$ and $\psi(\cdot)$ functions for different elliptically symmetric family of densities. Here, $K_\nu(\cdot)$ is the modified Bessel function of the second kind. 63

3.2 Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA1). 80

3.3 Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA2a). 81

3.4 Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for simulation scenario (PCA2b). 81

3.5 Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA2c). 82

3.6 Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA2d). 82

3.7 Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA3a). 83

3.8 Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA3b). 83
3.9 Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA3c). 84
3.10 Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA3d). 84

4.1 Different Penalized Criteria for Matrix Rank Estimation. (Here $\widehat{\sigma}_\alpha$ denotes the estimated noise variance by minimizing corresponding $H_\alpha(\cdot)$). 99
4.2 Summarized results of the best performing existing rank estimation methods under different simulation scenarios (Column Prop. indicates the proportion of replications where the correct rank is estimated exactly). 112
4.3 Proportion of exact (overestimation in brackets) for different rank-estimation methods when using classical SVD for the simulation setting of 50×40-dimensional matrix with equal singular values. 113
4.4 Bias (RMSE in brackets) for different rank-estimation methods when using classical SVD for the simulation setting of 50×40-dimensional matrix with equal singular values. 114
4.5 Proportion of exact (overestimation in brackets) for different rank-estimation methods when using classical SVD for the simulation setting of 50×40-dimensional matrix with decreasing singular values. 115
4.6 Bias (RMSE in brackets) for different rank-estimation methods when using classical SVD for the simulation setting of 50×40-dimensional matrix with decreasing singular values. 116
4.7 Proportion of exact (overestimation in brackets) for different rank-estimation methods when using robust SVD for the simulation setting of 50×40-dimensional matrix with equal singular values. 117
4.8 Bias (RMSE in brackets) for different rank-estimation methods when using robust SVD for the simulation setting of 50×40-dimensional matrix with equal singular values. 118
4.9 Proportion of exact (overestimation in brackets) for different rank-estimation methods when using robust SVD for the simulation setting of 50×40-dimensional matrix with decreasing singular values. 119
4.10 Bias (RMSE in brackets) for different rank-estimation methods when using robust SVD for the simulation setting of 50×40-dimensional matrix with decreasing singular values. 120

5.1 The asymptotic breakdown of MSD functional for different situations for exponential distribution as model family and normal distribution as contamination, with mean tending to $+\infty$, as shown in Example 5.4.6 The quantity $\widetilde{\epsilon}_{(\alpha,\lambda)}$ is a solution to (5.17). 147

6.1 Different families of statistical divergence that arise as special cases of the Generalized Alpha-Beta divergence family, up to a multiplicative constant. . 161

Abbreviations

pmf	probability mass function
pdf	probability density function
iid	independent and identically distributed
inh	independent and non-homogeneous
AIC	Akaike's Information Criterion
ALM	Augmented Lagrange Multiplier
BCV	Bi-Cross Validation
BIC	Bayesian Information Criterion
BP	Breakdown point
CV	Cross Validation
DCT	Dominated Convergence Thereom
DIC	Divergence Information Criterion
DICMR	Divergence Information Criterion for Matrix Rank
DPD	Density power divergence
GABD	Generalized Alpha-Beta divergence
GCV	Gabriel-style Cross Validation
GSD	Generalized S-divergence
IC	Information Criterion
KLD	Kullback Leibler Divergence
LSD	Logarithmic S-divergence
MAE	Mean absolute error

MCT	Monotone Convergence Theorem
MDPDE	Minimum density power divergence estimator
MGABDE	Minimum generalized Alpha-Beta divergence estimator
MGSDE	Minimum generalized S-divergence estimator
MHDE	Minimum Hellinger divergence estimator
MLE	Maximum likelihood estimator
MLSDE	Minimum logarithmic S-divergence estimator
MSDE	Minimum S-divergence estimator
MSE	Mean squared error
PCA	Principal component analysis
PCP	Principal component pursuit
PD	Power divergence
RMSE	Root mean squared error
RPCA	Robust principal component analysis
RSVD	Robust singular value decomposition
rPCAdpd	Robust principal component analysis using density power divergence
rSVDdpd	Robust singular value decomposition using density power divergence
SVD	Singular value decomposition
WCV	Wold-style Cross Validation

Chapter 1

Background

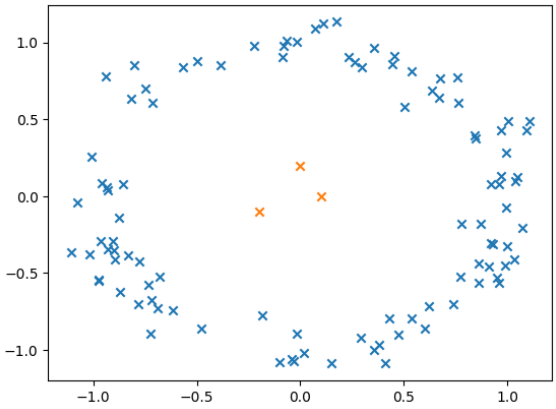
1.1 Introduction

In today's era of big data, a statistician often has to work with high-dimensional datasets. To obtain inference from these datasets, she usually aims to uncover an approximate low-dimensional structure present in these datasets, and one of the primary tools for this purpose is various matrix factorization techniques. Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) are two notable matrix factorization techniques having widespread applications across a multitude of disciplines such as machine learning, clustering techniques, network and graph analysis, signal processing, video, image and sound processing, natural language processing, financial data modelling, genetics, etc.

In these high-dimensional datasets, it is often extremely difficult to detect outlying observations. For instance, it is possible to have an outlying observation in a two-dimensional dataset that does not present itself as an outlier when one-dimensional projections of the data are considered; see Figure 1.1 for instance. It is known that the classical matrix factorization techniques are extremely sensitive to these outliers (Hawkins et al., 2001, Wright et al., 2009), even a single bad observation can lead to egregiously bad estimates. Unfortunately, the classical robust inference techniques such as M-estimation methods are not well suited for such cases, due to the large dimensionality of the data. In particular, Maronna (1976) showed that an affine equivariant M-estimator of the location and scatter of a p-dimensional data cannot have an asymptotic breakdown point larger than $1/(p+1)$, which decays rapidly for large p. Therefore, as an alternative to that, almost all robust matrix factorization techniques proposed to date, rely on some form of convex optimization problem with finite-sample theoretical guarantees. These optimization problems will be discussed in detail in Section 1.3. However, these techniques are computationally expensive and are difficult to scale for arbitrarily high dimensional datasets, which limits their applicability in practice.

In contrast, the minimum divergence estimators have been found to be extremely efficient and robust in many applications (Basu et al., 2011), and are another alternative to the classical M-estimators. In this thesis, we show that a broad subclass of these minimum divergence estimators has an asymptotic breakdown point independent of data dimension.

Figure 1.1: A 2-dimensional dataset with data points lying approximately on a one dimensional manifold of a circle, and a few outliers in the center which are not distinguishable as outliers in the one dimensional linear projections of the data.



This is a key property that makes these estimators suitable for the inference problems related to arbitrarily high-dimensional datasets. Among these minimum divergence estimators, the minimum density power divergence estimator (MDPDE) proposed by Basu et al. (1998) turns out to be both robust and efficient, striking a balance between these two competing properties with the help of a robustness tuning parameter. Also, the MDPDE is a single integral non-kernel type estimator (Jana and Basu, 2019), which avoids any nonparametric smoothing techniques leading to an elegant solution to the estimating equations in consideration. Therefore, we use the popular MDPDE to derive robust versions of SVD and PCA estimators and demonstrate efficient algorithms to compute them. Our proposed new approach leads to a highly robust and highly scalable estimator, improving upon the limitations of the existing approaches.

1.2 Notations

This section introduces several notations to be used throughout this thesis.

- The Greek and Roman letters in the normal font will indicate scalar quantities. On the other hand, symbols in bold fonts will be used to denote vectors or matrices.
- For two real numbers a and b, the symbol $a \vee b$ denotes the maximum of a and b and the symbol $a \wedge b$ is used to denote the minimum of a and b.
- The vectors $\boldsymbol{v}$ will generally be regarded as column vectors unless otherwise specified.
- For a matrix $\boldsymbol{A}$, its transpose, trace and determinant are respectively denoted by $\boldsymbol{A}^\intercal$, $\text{Trace}\,(\boldsymbol{A})$ and $|\boldsymbol{A}|$.
- If $\boldsymbol{A}$ is invertible, its inverse is denoted by $\boldsymbol{A}^{-1}$. However, if $\boldsymbol{A}$ is not invertible, then $\boldsymbol{A}^{-}$ and $\boldsymbol{A}^{+}$ denote any g-inverse and the Moore-Penrose pseudo-inverse of $\boldsymbol{A}$ respectively.

- The symbol $\text{Diag}\,(\boldsymbol{A})$ denotes the vector comprising of the diagonal elements of the matrix $\boldsymbol{A}$, and when the same symbol is used as $\text{Diag}\,(a_1, \ldots a_n)$, then it means the $n \times n$ diagonal matrix whose nonzero entries are $a_1, a_2, \ldots a_n$ in the same order.

- The symbols $\boldsymbol{A}_{i,\cdot}$ and $\boldsymbol{A}_{\cdot,j}$ denote the i-th row and the j-th column of the matrix $\boldsymbol{A}$ respectively.

- For any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, their usual matrix product is denoted by $\boldsymbol{AB}$, Kronecker product is denoted by $\boldsymbol{A} \otimes \boldsymbol{B}$ and Hadamard product is denoted by $\boldsymbol{A} \odot \boldsymbol{B}$.

- For a matrix $\boldsymbol{A}$, the notation $\boldsymbol{A}^{\odot k}$ denotes the Hadamard exponent of $\boldsymbol{A}$, i.e., $\boldsymbol{A}^{\odot k} = \boldsymbol{A} \odot \boldsymbol{A} \odot \ldots k \text{ times} \odot \boldsymbol{A}$ for any $k \geq 1$.

- For a vector $\boldsymbol{v} = (v_1, v_2, \ldots v_n)^\intercal$, its p-th order norm is defined as

$$\|\boldsymbol{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{1/p}, \; p \in (0, \infty).$$

  Clearly, $\|\boldsymbol{v}\|_2$ denotes the usual Euclidean norm. When there is no scope for confusion, we may drop the suffix 2.

- The symbol f is used as a generic symbol to denote a probability density function dominated by a measure μ on sample space Ω and its Borel subset $\mathcal{B}$. Similarly, θ is used as a generic symbol to denote a parameter lying in some parameter space Θ, where $\Theta \subseteq \mathbb{R}^d$ for some $d \geq 1$.

- The symbol $\mathbb{P}$ is used as a generic symbol to denote a probability measure. The corresponding expectation and variance operators are denoted by $\mathbb{E}(\cdot)$ and $\text{Var}(\cdot)$ respectively. The covariance operator is denoted by $\text{Cov}(\cdot, \cdot)$.

- $\text{Supp}(f)$ denotes the support of any nonnegative function f, i.e., the closure of the set $\{x : f(x) > 0\}$.

- For a probability density function f, its p-th norm is denoted by

$$\|f\|_p := \left(\int_\Omega f^p d\mu \right)^{1/p}, \; p > 0,$$

  The space of density functions such that its p-th norm is finite will be denoted by $L^p(\Omega, \mathcal{B}, \mu)$ or just L^p when there is no ambiguity in the underlying choice of Ω or $\mathcal{B}$.

- For two probability density functions f and g dominated by a common measure μ, their (A, B)-product is denoted by

$$\langle f, g \rangle_{A,B} := \int_\Omega f^A g^B d\mu,$$

  for all $A, B \in \mathbb{R}$ such that the above is well-defined.

- A sequence of nonnegative functions $\{f_n\}$ are said to be uniformly L^p-integrable if $\sup_n \|f_n\|_p < \infty$.

- For a random variable X, the symbol $X \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes that X follows a p-variate normal distribution with location $\boldsymbol{\mu}$ and dispersion parameter $\boldsymbol{\Sigma}$. When $p = 1$, we drop the suffix and use the notation $\mathcal{N}(\mu, \sigma^2)$ to indicate a univariate normal distribution with mean μ and variance σ^2.

- Given two sequences of real numbers $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we denote

  (a) $a_n = \mathcal{O}(b_n)$ if there exists a constant $C \in (0, \infty)$ independent of n such that $|a_n| \leq C|b_n|$ for sufficiently large n.

  (b) $a_n \asymp b_n$ if these two sequences are asymptotically equivalent, i.e., there exist two constants $C_1, C_2 \in (0, \infty)$ independent of n such that $C_1 a_n < b_n < C_2 a_n$ for all sufficiently large n.

- We use the symbol $\mathcal{C}^k([a, b])$ to denote the class of all real-valued functions that are k-times differentiable with a continuous k-th order derivative on $[a, b]$ for $-\infty \leq a < b \leq \infty$. Extending this notation, we use the symbol $\mathcal{C}^0([a, b])$ to denote the class of all continuous functions on $[a, b]$.

- For any set A, the symbol $\mathbf{1}_A$ denotes the indicator function on the set A, i.e.,

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A \end{cases}.$$

- For a function $f(x, y)$, its first order partial derivatives with respect to x and y are denoted by $\partial_x f$ and $\partial_y f$ respectively. The second order partial derivative with respect to x and y are denoted by $\partial_x^2 f$ and $\partial_y^2 f$ respectively. The second-order mixed derivative is denoted by $\partial_{x,y}^2 f$.

- In a parametric inference problem for any parameter $\boldsymbol{\theta}$, the superscript $\boldsymbol{\theta}^g$ will denote the true value of the parameter that is being estimated, when the true data are generated from a distribution function G with density function g; the superscript $\boldsymbol{\theta}^{(t)}$ will denote the value of the estimated parameter at t-th iteration of an iterative algorithm and $\boldsymbol{\theta}^*$ will be used to indicate its limit, provided that the sequence of iterated estimates converge.

- The symbol S_r^n denotes the r-Stiefel manifold of order n, i.e., the space of all $n \times r$ real orthogonal matrices. When $r = 1$, the suffix may be dropped. Consequently, the symbol S^{n+} denotes the set of all orthonormal vectors of length n such that its first nonzero entry is positive.

1.3 Mathematical Preliminaries

1.3.1 Singular Value Decomposition

The singular value decomposition (SVD) is a matrix factorization technique that breaks down any matrix into a product of three matrices: two orthogonal matrices whose columns consist of the left and the right singular vectors and a diagonal matrix consisting of the nonnegative singular values. SVD is commonly viewed as a low-rank approximation of a linear transformation, representing the transformation as a sequence of rotations and dilations. In the statistical context, the problem of SVD begins with a data matrix $\boldsymbol{X}$ having n rows and p columns comprising measurements across two "fixed-effect" domains of interest (e.g., time, space, age groups, genes, demographics, industry sectors, etc.). The data are arranged so that the ordinal variables corresponding to the rows and columns are organized in their natural order. If there has been no error in the measurements, then it is expected that the measurements across each row from left to right will exhibit some smoothness restrictions. Similar behaviour should also be present for each column as we move from the top to bottom. This setup is prevalent in many functional data analysis literature, see Ferraty and Vieu (2006) for further details.

In literature (Yang, 2014), this structure is modelled through a representation of the form

$$\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{E}, \tag{1.1}$$

where the matrix $\boldsymbol{L}$ is a low-rank matrix and the matrix $\boldsymbol{E}$ is an error matrix consisting of the measurement errors. The smoothness restrictions on the rows and columns here translate to the restriction on the rank of $\boldsymbol{L}$ matrix. In this model, the entries of the matrix $\boldsymbol{E}$, i.e., e_{ij}s are distributed as independent and identically distributed (i.i.d.) mean zero real-valued random variables. Applying classical SVD of $\boldsymbol{L}$ into (1.1), we can also rewrite the decomposition as

$$\boldsymbol{X}_{n\times p} = \boldsymbol{U}_{n\times r}\boldsymbol{D}_{r\times r}\boldsymbol{V}^{\intercal}_{r\times p} + \boldsymbol{E}_{n\times p}, \tag{1.2}$$

where $r = \text{rank}(\boldsymbol{L})$ which may be unknown, and $\boldsymbol{U}, \boldsymbol{V}$ are orthogonal matrices consisting of the left and right singular vectors and $\boldsymbol{D}$ is the diagonal matrix containing the singular values of $\boldsymbol{L}$ in its diagonal. The objective of the problem in SVD is to estimate the components $\boldsymbol{U}, \boldsymbol{D}$ and $\boldsymbol{V}$ given the observation matrix $\boldsymbol{X}$.

The applications of SVD are immense. Some of its mathematical applications include computing pseudo-inverses or Moore-Penrose inverses of matrices, efficiently solving systems of homogeneous linear equations, determining range space, null space, and rank of a linear transformation, and finding ordinary least square solutions (Golub and Reinsch, 1970). In statistical data analysis and modelling, SVD is used to perform different dimensionality reduction techniques such as Principal Component Analysis (PCA) (Jolliffe, 2002), correspondence analysis (Greenacre, 2017), latent semantic indexing (Anandarajan et al., 2019), etc. Many clustering techniques (Drineas et al., 2004, Cheng et al., 2019), image processing algorithms (Sadek, 2012), signal processing methods (Vaccaro, 1991,

Moonen and De Moor, 1995) rely on SVD as their fundamental component. In the field of image and sound processing, interesting applications of SVD include image watermarking schemes (Dappuri et al., 2020), signal denoising and feature enhancements (Zhao and Jia, 2017), audio watermarking (Rezaei and Khalili, 2019), sound source localization (Grondin and Glass, 2019), and sound recovery techniques (Zhang et al., 2016a), etc. SVD has also gained prominence in the field of bioinformatics, where it is used for analysing protein functional associations (Franceschini et al., 2015), clustering gene expression data (Bustamam et al., 2018), and predicting protein-coding regions (Das et al., 2017). In the realm of geographical science, Kumar et al. (2011) employed SVD-based techniques, including its robust variant, to generate accurate graphical representations of climate data, mitigating the impact of extreme weather events such as thunderstorms and heavy rainfall. Interested readers are referred to Johnson et al. (2002) and references therein for further applications of SVD and related techniques. Such a wide range of applications clearly underscores the relevance of SVD as an extremely integral component of data analysis across a plethora of disciplines.

1.3.2 Principal Component Analysis

The classical problem of finding the principal components aims to approximate the covariance structure of a high dimensional sample of many features by the covariance structure of a lower dimensional sample of "principal components", obtained as linear combinations of the original feature variables. Mathematically, starting with an independent and identically distributed (i.i.d.) sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots \boldsymbol{X}_n$, where each $\boldsymbol{X}_i \in \mathbb{R}^p$, and a scale measure $S_n(y_1, \dots y_n)$ to measure the dispersion in a univariate sample $\{y_1, \dots y_n\}$, the first eigenvector associated with the principal components is defined as the unit length vector $\boldsymbol{v}$ maximizing the function

$$\boldsymbol{v} \to S_n(\boldsymbol{v}^\intercal \boldsymbol{X}_1, \dots \boldsymbol{v}^\intercal \boldsymbol{X}_n);\ \boldsymbol{v} \in \mathbb{R}^p. \tag{1.3}$$

Similarly, assuming that the first $(k-1)$ eigenvectors $\widehat{\boldsymbol{v}}_1, \widehat{\boldsymbol{v}}_2, \dots \widehat{\boldsymbol{v}}_{k-1}$ have already been found, one can obtain the subsequent k-th eigenvector as the unit vector maximizing the same function as in (1.3), but under the additional set of restrictions $\boldsymbol{v}^\intercal \widehat{\boldsymbol{v}}_i = 0$ for all $i = 1, \dots (k-1)$. The corresponding eigenvalues are defined as the maximum values of the scale function, i.e.,

$$\widehat{\lambda}_k = S_n(\widehat{\boldsymbol{v}}_k^\intercal \boldsymbol{X}_1, \dots \widehat{\boldsymbol{v}}_k^\intercal \boldsymbol{X}_n).$$

In essence, principal component analysis (PCA) takes input n observations of dimension p, where p is presumably very large, and outputs a set of pairs $\{(\widehat{\lambda}_k, \widehat{\boldsymbol{v}}_k) : k = 1, 2, \dots r\}$ where r is a pre-specified number of components, generally much smaller compared to both n and p. For each k, the former of the pair $\widehat{\lambda}_k$ denotes the maximum variability expressed by the k-th principal component, and the latter of the pair $\widehat{\boldsymbol{v}}_k$ denotes the direction along which this maximum variability can be found in the given i.i.d. sample. The k-th principal component is then defined by the variable obtained from projecting the observations along the k-th eigenvector scaled by the k-th eigenvalue, i.e., $\{\widehat{\lambda}_k \widehat{\boldsymbol{v}}_k^\intercal \boldsymbol{X}_i : i = 1, \dots n\}$.

Since S_n is a measure of dispersion, it is independent of the change in mean. Hence, without the loss of generality, one may assume $\boldsymbol{X}_i$s to have a mean $\mathbf{0}$. Otherwise, one can subtract $\overline{\boldsymbol{X}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i$ from each observation and consider $\boldsymbol{X}_i - \overline{\boldsymbol{X}}$ as the new sample. Denoting the data matrix $\boldsymbol{X}$ by stacking these observations as rows

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1^\intercal \\ \boldsymbol{X}_2^\intercal \\ \vdots \\ \boldsymbol{X}_n^\intercal \end{bmatrix}_{n\times p},$$

and using the usual standard derivation measure as S_n yields that the direction of the first principal component is the same as the direction of the eigenvector corresponding to the largest eigenvalue of the covariance matrix $\boldsymbol{X}^\intercal\boldsymbol{X}/n$. Subsequently, the direction of the k-th principal component matches with the direction of the eigenvector corresponding to the k-th largest eigenvalue of the same covariance matrix, for all $k \geq 1$. Hence, if r is a prespecified number of principal components to extract, then the estimation of PCA is the same as estimating the eigen-decomposition of the covariance matrix

$$\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X} = \sum_{k=1}^{r}\widehat{\lambda}_k\widehat{\boldsymbol{v}}_k\widehat{\boldsymbol{v}}_k^\intercal + \boldsymbol{E}, \tag{1.4}$$

where $\boldsymbol{E}$ is the error matrix consisting of the subsequent $(k = (r+1), \ldots, (n \wedge p))$ eigenvalues and corresponding eigenvectors that are expected to be negligible. The decomposition (1.4) can also be reformulated as

$$\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{E} = \boldsymbol{U}_{n\times r}\boldsymbol{D}_{r\times r}\boldsymbol{V}_{p\times r}^\intercal + \boldsymbol{E}, \tag{1.5}$$

where $\boldsymbol{D}$ is the diagonal matrix consisting of the square root of n times the eigenvalues, and $\boldsymbol{V}$ is a $p \times r$ matrix whose columns consist of the eigenvectors $\widehat{\boldsymbol{v}}_k$s. Here, $\boldsymbol{U}$ is a $n \times r$ random matrix having i.i.d. mean zero rows, contributing the stochastic component of $\boldsymbol{X}$. As a result of the above factorization, the matrix $\boldsymbol{L}$ becomes a low-rank matrix. Whereas the classical approach to PCA was to consider the variance maximization problem as in (1.3) (Croux and Ruiz-Gazen, 1996), recent authors (Zhou et al., 2010, Candès et al., 2011) consider the problem given in (1.5) with sparse $\boldsymbol{E}$ matrix as the robust PCA problem instead.

Since usually a small number of principal components can explain most of the variation present in the random sample, its most prominent use is in dimensionality reduction. PCA provides a simple method of visualizing any high-dimensional data by plotting the first two or three principal components, aiding in exploratory data analysis (Tukey, 1977). Locantore et al. (1999) used this technique of projecting data onto the first two principal components to identify potential outliers in ophthalmology data. Jolliffe (2002) discusses an application of PCA for variable selection in the regression context. In machine learning and pattern recognition, PCA has been used abundantly for both supervised and unsupervised paradigms (Vathy-Fogarassy and Abonyi, 2013). PCA has also found its applications

across many disciplines ranging from multi-sensor data fusion (Lock et al., 2013), image and video processing (Bouwmans et al., 2018) to material and chemical sciences (Smilde et al., 2005). The readers are referred to Sanguansat (2012) and the references therein for further details on various applications of PCA.

Although both SVD and PCA lead to similar decompositions as illustrated in (1.2) and (1.5), there are some fundamental differences between their stochastic models.

1. In SVD, the data matrix $\boldsymbol{X}$ are observations across two domains of interest having multiple categories (or levels), hence it is expected that there will be some correlations present between the rows and the columns. On the other hand, PCA considers the rows of the data matrix $\boldsymbol{X}$ to be independent observations, but across multiple (more than two) variables. Hence, in PCA, only the columns are allowed to be correlated.

2. In SVD, the entries of $\boldsymbol{E}$ are i.i.d. mean zero random variables, as they represent measurement errors. In PCA, only the rows of $\boldsymbol{E}$ are independently distributed. They have zero mean since the mean vector is already subtracted from each row of the data matrix.

3. The matrix $\boldsymbol{U}$ consisting of the left singular vectors is fixed, deterministic but unknown in SVD, namely it is a parameter of the model. The only stochastic component of the SVD model is the matrix $\boldsymbol{E}$ alone. In contrast, the rows of $\boldsymbol{U}$ in PCA are random variables. Hence the PCA model has two stochastic components, the matrices $\boldsymbol{U}$ and $\boldsymbol{E}$.

A detailed discussion on the differences between the PCA and the SVD can be found in Yang (2014).

1.3.3 Breakdown Point

The breakdown point is a popular measure of robustness and global reliability for an estimator. While developing the robust Hodges-Lehman estimator of location, Hodges (1967) motivates the idea of a finite-sample breakdown point of a location estimator as the maximum proportion of incorrect or arbitrary observations in the sample that an estimator can tolerate without returning any egregiously unreasonable value i.e., the maximum proportion of arbitrary observations in the sample such that the estimate cannot be made arbitrarily large or arbitrarily small. The asymptotic breakdown point of the estimator is defined as the limit of this finite-sample breakdown point, provided that the limit exists. However, such a definition makes sense only in the case of location estimators. For instance, the estimator of the scale parameter can break down when the estimate either "explodes" to infinity or "implodes" to zero (Maronna et al., 2019). Thus, generalizing the idea of Hodges (1967), similar to the definition present in Hampel (1971), we define

the asymptotic breakdown point ϵ^* of a sequence of estimators $\{T_n : n \geq 1\}$ as

$$\epsilon^* := \sup\Bigg\{\epsilon : \epsilon \in [0, 1/2], \text{ and, } \inf_{\theta_\infty \in \partial\Theta} \liminf_{m\to\infty} \liminf_{n\to\infty} P_{G_{\epsilon,m}}(\|T_n - \theta_\infty\| > 0) = 1, \text{ for all } \{K_m\}_{m=1}^{\infty}\Bigg\}, \quad (1.6)$$

where $G_{\epsilon,m} = (1-\epsilon)G + \epsilon K_m$ is the class of ϵ-contaminated distributions introduced by Huber and Donoho (1983), $\{K_m\}_{m=1}^{\infty}$ is an infinite sequence of contaminating distributions, and $P_{G_{\epsilon,m}}$ denotes the probability measure under the contaminated distribution $G_{\epsilon,m}$. Also, in Eq. (1.6), $\partial\Theta$ denotes the boundary of the parameter space $\Theta \subset \mathbb{R}^d$ $(d \geq 1)$ in the extended real number system. Hence, for a location estimator in the univariate setup, $\Theta = (-\infty, \infty)$ and $\partial\Theta = \{-\infty, \infty\}$, and for a scale estimator $\Theta = (0, \infty)$ with $\partial\Theta = \{0, \infty\}$.

Modern literature such as Maronna et al. (2019) adopt this definition for functionals, and correspondingly define the asymptotic breakdown point of a functional T estimating a parameter θ in a parameter space $\Theta \subset \mathbb{R}^d$ $(d \geq 1)$ as

$$\epsilon^* := \sup\left\{\epsilon : \epsilon \in [0, 1/2] \text{ and, } \inf_{\theta_\infty \in \partial\Theta} \liminf_{m\to\infty} \|T(G_{\epsilon,m}) - \theta_\infty\| > 0, \text{ for all } \{K_m\}_{m=1}^{\infty}\right\}, \quad (1.7)$$

The readers are referred to Definition 3.1 of Maronna et al. (2019) and its associated discussion for further details.

1.3.4 Density Power Divergence

Minimum divergence estimators have a natural appeal in parametric inference problems. These estimators aim to obtain the best choice of the parameter that minimizes a quantification of the degree of distance between the parameterized model and the data. These usually come in two major types:

1. The first type measures the distance between the model family of distribution functions and the "true" distribution of the data. Examples of this type of distance measures include Kolmogorov-Smirnov distance, Cramér-von Mises distance, Anderson-Darling distance, etc.

2. The second type measures a statistical divergence between the model family of density functions and the "true" density function approximating the data. The statistical divergence considered here may not be a valid distance in the sense of metric spaces, in particular, it may not satisfy the triangle inequality. However, such a divergence measure $d(f, g)$ is nonnegative for all pairs of densities f and g and equals 0 if and only if $f = g$ almost surely with respect to a common dominating measure.

Among these density-based minimum divergence estimators, particular classes require special attention due to their highly robust and efficient behaviours. Cressie and Read

(1984) introduced the power divergence family of statistical divergences including several important divergences like Kullback-Leibler (KL) divergence, Pearson's chi-square, Neyman's chi-square, Hellinger distance, and chi-square type divergences in general. The family of power divergence is a subclass of the general class of ϕ-divergences (Csiszár, 1963, Morimoto, 1963). Another notable divergence measure, namely the density power divergence (DPD), has been found to be particularly robust and efficient in different applications (Ghosh and Basu, 2013, Xiong and Zhu, 2021, Pyne et al., 2024). Let, f and g be two densities dominated by a common dominating measure μ. Then, Basu et al. (1998) define the DPD between two densities f and g as a family of divergence measures parameterized by a hyperparameter α, having the form

$$d_\alpha(f,g) = \begin{cases} \int f^{1+\alpha} d\mu - \left(1+\frac{1}{\alpha}\right)\int f^\alpha g d\mu + \frac{1}{\alpha}\int g^{1+\alpha} d\mu, & \text{if } \alpha > 0, \\ \int g \ln\left(\frac{g}{f}\right) d\mu, & \text{if } \alpha = 0, \end{cases} \tag{1.8}$$

provided that both f and g are $L^{1+\alpha}$-integrable, i.e., the $(1+\alpha)$-th norm $\|f\|_{1+\alpha}$ and $\|g\|_{1+\alpha}$ exist and are finite. The form given for $\alpha = 0$ can be obtained by taking limit $\alpha \to 0+$ of the general form for $\alpha > 0$, to ensure continuity of the divergence as a function of α. The parameter α is also called the robustness controlling parameter. By substituting $\alpha = 1$, the density power divergence becomes the same as the squared L^2 distance between f and g which is known to have satisfactory robustness properties in parametric estimation problems. On the other hand, for $\alpha = 0$, the DPD reduces to the Kullback-Leibler divergence, whose minimization results in the maximum likelihood estimator (MLE), an efficient but non-robust estimator. The parameter α thus provides a bridge between these two divergences, striking a balance between robustness and efficiency in estimation.

Given an independent and identically distributed sample of observations $X_1, \ldots X_n$, the minimum density power divergence estimator (MDPDE) is then defined as

$$\widehat{\theta}_\alpha = \arg\min_{\theta\in\Theta} d_\alpha(\widehat{g}_n, f_\theta), \tag{1.9}$$

provided that such a minimum exists. Here f_θ is an element of the model family of densities $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ indexed by the parameter θ and $\widehat{g}_n$ is an empirical density estimate of g based on the sample $X_1, \ldots X_n$. Note that, the minimization in (1.9) does not depend on the last term $\int \widehat{g}_n^{1+\alpha}(x)dx/\alpha$, hence one can substitute the integral $\int f_\theta^\alpha(x)\widehat{g}_n(x)dx$ by $\int f_\theta^\alpha(x)dG_n(x)$ where G_n is the empirical distribution function of the sample observations $X_1, \ldots X_n$. This yields the MDPDPE to be a solution to the reduced objective function as

$$\widehat{\theta}_\alpha = \arg\min_{\theta\in\Theta}\left[\int f_\theta^{1+\alpha}(x)dx - \left(1+\frac{1}{\alpha}\right)\frac{1}{n}\sum_{i=1}^n f_\theta^\alpha(X_i)\right].$$

Ghosh and Basu (2013) extended this estimation technique to accommodate independent but non-homogeneous samples (including, for example, the linear regression case). In this case, the sample of observations $X_1, X_2, \ldots X_n$ are modelled by a sequence of family

of densities $\{\mathcal{F}_i\}_{i=1}^n$ with $\mathcal{F}_i := \{f_{i,\theta} : \theta \in \Theta\}$, indexed by a common parameter θ. The corresponding objective function is then given by the average of density power divergences between the empirical density of the Dirac delta distribution at X_i and the corresponding model density $f_{i,\theta}$. In this case, the MDPDE is given by

$$\widehat{\theta}_\alpha = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}\left[\int f_{i,\theta}^{1+\alpha}(x)dx - \left(1+\frac{1}{\alpha}\right) f_{i,\theta}^{\alpha}(X_i)\right].$$

1.3.5 S-divergence

The power divergence (PD) family introduced by Cressie and Read (1984) and the density power divergence (DPD) family introduced by Basu et al. (1998) have later been subsumed in a larger "S-divergence" family by Ghosh et al. (2017). The DPD family (indexed by an α) connects the KL divergence ($\alpha = 0$) with the squared L^2 distance ($\alpha = 1$). Similarly, the S-divergence family (indexed by α and λ) connects the entire Cressie-Read family of power divergences (PD) smoothly to the squared L^2 distance at the other end but also contains the DPD class of divergences as a special case. The S-divergence between two densities g and f is defined as

$$S_{(\alpha,\lambda)}(g,f) = \frac{1}{A}\|f\|_{1+\alpha}^{1+\alpha} - \frac{1+\alpha}{AB}\langle f,g\rangle_{B,A} + \frac{1}{B}\|g\|_{1+\alpha}^{1+\alpha}, \tag{1.10}$$

where $A = 1+\lambda(1-\alpha)$, $B = \alpha - \lambda(1-\alpha)$. The choice of α is usually restricted in the unit interval $[0,1]$, but λ is allowed to take any real value. However, a sufficient (but not strictly necessary) condition for the integrals appearing in (1.10) to be well-defined is to take $A > 0, B > 0$ and f and g to be both $L^{1+\alpha}$-integrable.

If either $A = 0$ or $B = 0$, the corresponding cases are defined by the continuous limits of the divergence form (1.10) under $A \to 0$ and $B \to 0$, respectively, and are given by

$$S_{(\alpha,\lambda:A=0)}(g,f) = \lim_{A\to 0} S_{(\alpha,\lambda)}(g,f) = \int f^{1+\alpha}\ln\left(\frac{f}{g}\right) - \frac{1}{(1+\alpha)}\left(\|f\|_{1+\alpha}^{1+\alpha} - \|g\|_{1+\alpha}^{1+\alpha}\right), \tag{1.11}$$

and

$$S_{(\alpha,\lambda:B=0)}(g,f) = \lim_{B\to 0} S_{(\alpha,\lambda)}(g,f) = \int g^{1+\alpha}\ln\left(\frac{g}{f}\right) - \frac{1}{(1+\alpha)}\left(\|g\|_{1+\alpha}^{1+\alpha} - \|f\|_{1+\alpha}^{1+\alpha}\right). \tag{1.12}$$

In this case, a sufficient condition for the integrals appearing above to be finite is to let f and g be $L^{1+\alpha+\delta}$-integrable for some $\delta > 0$, and the cross-integral terms $\int f^{1+\alpha}\ln(g)$ and $\int g^{1+\alpha}\ln(f)$ to be finite as well. Additionally, in (1.11), we require $\text{Supp}(f) \subseteq \text{Supp}(g)$. In the same spirit, we require $\text{Supp}(g) \subseteq \text{Supp}(f)$ for Eq. (1.12) to be well-defined. Note that, here we use the convention that $0\ln(0) := 0$, by using the continuity of the function $x\ln(x)$ near the origin.

The S-divergence class is also a subclass of the Alpha-Beta divergence (also known as AB divergence) proposed by Cichocki et al. (2011). AB divergence allows one to take A and B to any real value, instead of putting the restriction $A + B = (1+\alpha) \geq 1$ like the

S-divergence family. For two density functions f and g, the AB divergence, parametrized by two hyperparameters α and β, is defined as

$$d_{AB}^{(\alpha,\beta)}(f,g) = \begin{cases} -\frac{1}{\alpha\beta}\left(\langle f,g\rangle_{\alpha,\beta} - \frac{\alpha}{\alpha+\beta}\|f\|_{\alpha+\beta}^{\alpha+\beta} - \frac{\beta}{\alpha+\beta}\|g\|_{\alpha+\beta}^{\alpha+\beta}\right) & \text{if } \alpha,\beta,\alpha+\beta\neq 0, \\ \frac{1}{\alpha^2}\left(\int f^\alpha \ln(f^\alpha/g^\alpha) - \|f\|_\alpha^\alpha + \|g\|_\alpha^\alpha\right) & \text{if } \alpha\neq 0,\beta=0, \\ \frac{1}{\beta^2}\left(\int g^\beta \ln(g^\beta/f^\beta) - \|g\|_\beta^\beta + \|f\|_\beta^\beta\right) & \text{if } \alpha=0,\beta\neq 0, \\ \frac{1}{\alpha^2}\left(\int \ln(g^\alpha/f^\alpha) + \int f^\alpha/g^\alpha - 1\right) & \text{if } \alpha,\beta\neq 0,(\alpha+\beta)=0, \\ \frac{1}{2}\int(\ln f - \ln g)^2 & \text{if } \alpha=\beta=0. \end{cases} \tag{1.13}$$

For $\alpha = 0$, the S-divergence family reduces to the PD family with parameter λ; for $\alpha = 1$, $S_{1,\lambda}$ equals the robust squared L^2 distance irrespective of the value of λ. On the other hand, $\lambda = 0$ generates the DPD family as a function of α. In Ghosh et al. (2017), it was shown that the S-divergence family defined in (1.10)-(1.12) indeed represents a class of genuine statistical divergence measures in the sense that, whenever finitely defined, the quantity $S_{\alpha,\lambda}(g,f) \geq 0$ for all pair of densities g, f and all $\alpha \geq 0, \lambda \in \mathbb{R}$, and $S_{\alpha,\lambda}(g,f)$ is equal to zero if and only if $g = f$ almost surely under their common dominating measure. As in the case of MDPDE, the minimum S-divergence (MSD) functional at a distribution G with density g is defined as

$$T_{(\alpha,\lambda)}(G) = \arg\min_{\theta\in\Theta} S_{(\alpha,\lambda)}(g, f_\theta), \tag{1.14}$$

provided that the minimum exists. Here, $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is the model family of densities. Unfortunately, the empirical distribution function cannot be used directly in (1.14) to obtain the estimate given a random sample $X_1, \ldots X_n$ except the case for MDPDE. One of the approaches is to replace the true density g with a suitable nonparametric density estimate $\widehat{g}_n$ and subsequently perform the minimization in (1.14). Further details on this estimation process can be found in Ghosh (2015) and Ghosh et al. (2017).

1.4 Organization of the thesis

In this chapter, we described the general aim of the thesis in Section 1.1, several notations in Section 1.2 and some mathematical preliminaries in Section 1.3. The rest of the thesis is organized into several chapters as follows: In Chapter 2, we propose a novel robust singular value decomposition algorithm called "rSVDdpd" using the minimum density power divergence estimator (MDPDE). We also derive efficient algorithms to compute it, as well as establish theoretical guarantees about the convergence of the algorithm and the consistency of the estimator. In the same chapter, we also demonstrate the applicability of this method in background modelling and foreground extraction of video surveillance data. In a similar spirit, we introduce a robust principal component analysis technique "rPCAdpd" in Chapter 3. Standard theoretical properties such as consistency, equivariance and asymptotic normality of the proposed estimator are also established

in Chapter 3. In Chapter 4, we indicate methods to determine the rank of the low-rank component $\boldsymbol{L}$ of a data matrix $\boldsymbol{X}$ having a decomposition as in (1.1). To this regard, we propose a new criterion function that improves over existing criterion functions, is consistent and also robust. Extensive simulation studies have been performed and corresponding results have been provided to support the theoretical derivations.

Moving on to Chapter 5, we analyse the robustness of the minimum S-divergence estimator (MSDE) introduced in Section 1.3.5 through an analysis of its asymptotic breakdown point. We establish that under reasonable assumptions, the MSDE (including the MDPDE) has a lower bound to the asymptotic breakdown point which is independent of the dimension of the data. This result provides a theoretical justification for the superior performance of "rSVDdpd" and "rPCAdpd" methods (that use MDPDE) for high-dimensional datasets. In Chapter 6, we generalize the results from Chapter 5 further and demonstrate that the dimension-free lower bound of the asymptotic breakdown point can be obtained for a very general superfamily of MSDE. In this direction, we propose a new generalized Alpha-Beta divergence family of statistical divergences, provide its characterization and some interesting properties, and then establish the asymptotic breakdown point related results.

Chapter 7 concludes the thesis and mentions some future directions of research including potential extensions of the proposed methods and the results present in the thesis and applications to more complex datasets.

Chapter 2

Robust Singular Value Decomposition

2.1 Introduction and Existing Literature

Several authors (Hawkins et al., 2001, Liu et al., 2003) have highlighted the vulnerability of the classical method of computing Singular Value Decomposition (SVD) to outliers present in the data matrix. Specifically, the eigenvalue stability inequality (Tao, 2012) only guarantees that the eigenvalues of a matrix $\boldsymbol{A}$ perturbed with a noise matrix $\boldsymbol{E}$ may shift the eigenvalue up to the operator norm of $\boldsymbol{E}$, i.e.,

$$|\lambda_i(\boldsymbol{A}+\boldsymbol{E}) - \lambda_i(\boldsymbol{A})| \leq \|\boldsymbol{E}\|_{op},$$

where $\lambda_i(\boldsymbol{A})$ is the i-th eigenvalue of the matrix $\boldsymbol{A}$ and $\|\boldsymbol{E}\|_{op}$ is the operator norm of $\boldsymbol{E}$. In the case where the error matrix $\boldsymbol{E}$ is of dimension $n \times n$, then even if its entries are independent and identically distributed (i.i.d.) random variables with zero mean and constant variance, the operator norm can grow at the rate of $\sqrt{n}$. This poses significant challenges in applying classical SVD methods to the real-world applications described in Section 1.3.1, where the data matrix $\boldsymbol{X}$ is often contaminated with outliers. Therefore, robust alternatives to the classical SVD methods are essential for producing reliable estimates in such scenarios.

One notable example is the application of automated foreground and background extraction from a noisy video surveillance data. In this context, outliers can manifest in several ways, such as noisy frames, low frame rate of the camera, change in illumination, multimodal backgrounds, presence of small moving objects and instances of camera tampering (Mantini and Shah, 2019a). Almost all existing methods addressing this problem (Garcia-Garcia et al., 2020) in an unsupervised manner rely on robust SVD or robust PCA techniques.

In contrast to the literature on robust PCA, the literature on robust SVD is rather limited (Yang, 2014). Ammann (1993) was one of the early pioneers in developing a robust version of SVD. He treated it as a special case of the projection pursuit problem to be solved using the transposed QR algorithm. Other researchers, such as Hawkins

et al. (2001), Liu et al. (2003) and Ke and Kanade (2005), approached the computation of SVD as the least squares problem and proposed robust extensions using alternating L_1 regression algorithms with the least absolute deviation (LAD) loss function. However, a simple LAD approach is sensitive to high leverage points, which led to the exploration of weighted LAD approaches and the utilization of the Huber weight function (Jung, 2010). In a different attempt, Rey (2007) introduced a robust method called "Total" SVD, which employed Huber's weight function and "Total" least squares (Markovsky and Van Huffel, 2007). This approach accounted for errors in both the data matrix and the singular vectors, in contrast to the usual least squares where the only source of error is the response variable. Although this resulted in a more robust SVD estimate, the method faced several convergence issues as mentioned by Rey (2007) himself. Alternatively, Zhang et al. (2013) incorporated the Huber weight function in the loss function and combined it with a squared error-based penalty function for regularization, creating another robust SVD estimator. Wang (2017) used an estimator derived from an α-stable distribution with a cost function $\rho(x) = \log(x^2 + K^2)$, where the tuning parameter K provides a balance between robustness and efficiency. Nevertheless, finding the appropriate tuning parameter K for this estimator was challenging. Apart from a simple alternating L^1 regression approach, there is a lack of theoretical guarantees or properties for the resulting SVD estimates in the literature (Gabriel and Zamir, 1979); in addition, the orthogonality of the singular vectors are not assured. Yang (2014) provides some solution to this using a thresholding approach and estimating all singular values jointly, and provides an optimal guarantee in terms of the minimax error rate. However, this approach only considers sparse and smooth structures present in the data matrix, and there are no robustness guarantees in terms of common robustness measures such as influence functions or breakpoint points.

A major limitation of many existing algorithms is that they rely on several matrix inversion steps, which are numerically unstable for high-dimensional matrices. These algorithms also have a super-quadratic computational complexity (at least $\mathcal{O}(n^{2.3078})$) for inverting an $n \times n$ dense matrix (Ambainis et al., 2015). This can limit the practical usefulness of these algorithms in real-life applications. For example, in the background modelling problem, it takes about 13 seconds per frame for the exact PCP (Principal Component Pursuit) method to converge for a 176×144 low-resolution greyscale video as mentioned in Candès et al. (2011). In comparison, the robust SVD method that we will propose in this Chapter avoids these matrix inversion steps and takes only 0.183 seconds per frame for the same video. Some algorithms avoid the matrix inversion step by performing only the rank-one decomposition in the SVD and outputting only the largest singular value and the corresponding singular vectors. For obtaining subsequent singular value estimates, one has to work with the residual matrices, which may produce singular vectors that are not orthogonal to each other.

To address these issues, we introduce a novel robust SVD methodology, called **Robust SVD using Density Power Divergence** (rSVDdpd), based on the popular minimum density power divergence estimator introduced by Basu et al. (1998). The minimum den-

sity power divergence has been proven to be robust and efficient in estimation in several contexts (Ghosh and Basu, 2013, Xiong and Zhu, 2021). We also introduce an iterative algorithm using an alternating weighted regression approach as in Rey (2007), that avoids the matrix inversion steps and reduces the computational complexity of each iteration to linear time with the help of parallel computation. We also ensure the orthogonality of resulting singular vectors by using the Gram-Schmidt orthogonalization trick. Furthermore, we establish convergence guarantee of the estimation algorithm and prove the equivariance and statistical consistency of the estimates under some reasonable assumptions.

The rest of this Chapter is organized as follows. In Section 2.2, we formulate the robust SVD problem and introduce the rSVDdpd estimator. An efficient algorithm to compute this estimate using alternating weighted regression is also presented in Section 2.2.3. Then we present several theoretical properties of the rSVDdpd estimator in Section 2.3. In Section 2.4, we corroborate the theoretical findings with extensive simulation studies and compare the performance of the rSVDdpd estimator with several existing robust SVD algorithms. Following that, we apply the rSVDdpd estimator to the video surveillance background modelling problem in Section 2.4.

2.2 The rSVDdpd Estimator

2.2.1 Problem Statement

For the problem of robust singular value decomposition (RSVD), we consider an extended version of the classical SVD model given in (1.1) to incorporate the presence of outliers. We begin with a data matrix $\boldsymbol{X}$ of dimension $n \times p$ (n and p may be different), admitting an approximate low-rank representation of the form

$$\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S} + \boldsymbol{N}, \tag{2.1}$$

where $\boldsymbol{L}$ is a low-rank matrix, $\boldsymbol{S}$ is a sparse matrix with a small number of non-zero entries, and $\boldsymbol{N}$ is a dense perturbation matrix (such as a matrix with i.i.d. random entries with mean zero and constant variance) representing the measurement errors. As mentioned in Section 1.3.1, the entries of $\boldsymbol{X}$ matrix correspond to observations from two "fixed effect" domains. In contrast to (1.1), the non-zero entries of the sparse matrix $\boldsymbol{S}$ in (2.1) correspond to the outlying measurements that necessitate a robust SVD procedure. Similar to (1.2), using the classical SVD of the low rank matrix $\boldsymbol{L}$, we can rewrite the decomposition in (2.1) as

$$\boldsymbol{X} = \sum_{k=1}^{r} \lambda_k \boldsymbol{u}_k \boldsymbol{v}_k^\intercal + \boldsymbol{E}, \tag{2.2}$$

where $\{\boldsymbol{u}_k\}_{k=1}^r$ is a set of r vectors of length n and $\{\boldsymbol{v}_k\}_{k=1}^r$ is a set of r vectors of length p satisfying the orthogonality conditions

$$\boldsymbol{u}_k^\intercal \boldsymbol{u}_l = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{if } k \neq l \end{cases}, \; \boldsymbol{v}_k^\intercal \boldsymbol{v}_l = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{if } k \neq l \end{cases}, \; k, l \in \{1, 2, \ldots, r\}. \tag{2.3}$$

The $n \times p$ dimensional matrix $\boldsymbol{E} := \boldsymbol{N} + \boldsymbol{S}$ consists of the errors e_{ij}s, which are generally expected to be smaller in magnitude than the corresponding entries of the data matrix $\boldsymbol{X}$, except at a few coordinates with outlying observations due to the nonzero elements of the sparse component $\boldsymbol{S}$.

Given such a decomposition as in (2.1) or (2.2), the goal of the robust SVD problem is to estimate the unknown rank r of the low-rank component of $\boldsymbol{X}$, the left and right singular vectors $\boldsymbol{u}_k$s and $\boldsymbol{v}_k$s and the nonnegative singular values λ_ks. For notational convenience, let us denote $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r]_{n\times r}$, $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r]_{p\times r}$ and $\boldsymbol{\Lambda}$ as the $r \times r$ diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_r$. Also, the entries of the vectors $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ are denoted by u_{ki} and v_{kj} respectively, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$. In view of these notations, the orthonormality constraints in (2.3) simply mean that the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ belong to the r-Stiefel manifolds S_r^n of order n and S_r^p of order p respectively.

2.2.2 Derivation

For the time being, we assume that the rank r of the low-rank component $\boldsymbol{L}$ is known, and focus on the estimation of singular values and vectors. In many practical problems, the choice of the rank r may be pre-determined by the context. However, when it is unknown, one may use the methods indicated in Section 2.2.4 or in Chapter 4 to estimate the rank.

The description of RSVD as in (2.1) is equivalent to the LSN decomposition in Zhou et al. (2010) and is a generalization of the LS decomposition of Candès et al. (2011). In both cases, the authors assume that the only source of randomness comes from the choice of the support of the nonzero entries of the sparse $\boldsymbol{S}$ matrix, and all the theoretical results are proved for this setup. In fact, it is assumed that the entries $\mathbf{1}_{\{s_{ij}\neq 0\}} \sim \text{Ber}(\delta)$ for some $\delta \in (0, 1)$. In decomposition (2.1), if we additionally assume that the entries n_{ij}s of the matrix $\boldsymbol{N}$ are independently distributed random variables, then it follows that the entries e_{ij} of the combined matrix $\boldsymbol{E} = \boldsymbol{S} + \boldsymbol{N}$ follow a mixture distribution of the form $G_{ij} = (1 - \delta)G_{1,ij} + \delta G_{2,ij}$ for some $\delta \in [0, 1]$, denoting the proportion of contamination through non-zero entries in the sparse component. Here, $G_{1,ij}$ and $G_{2,ij}$ are the distribution functions corresponding to the dense perturbation and the sparse outlying components of e_{ij}, i.e., of n_{ij} and s_{ij} respectively. We assume that the distribution G_{ij} admits a density g_{ij} with respect to the Lebesgue measure for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

On the flip side, although e_{ij}s may follow different densities g_{ij} due to incorrect choice of the rank of $\boldsymbol{L}$ and differing means arising from the contribution of subsequent singular values $\sum_{k=(r+1)}^{\min\{n,p\}} \lambda_k u_{ki} v_{kj}$, an underlying assumption of performing robust SVD up to rank r is that all subsequent singular values $\lambda_{(r+1)}, \lambda_{(r+2)}, \ldots, \lambda_{\min\{n,p\}}$ are small and same. This can be reflected by modelling these errors to be i.i.d. observations from a symmetrically distributed scale family of densities, $\mathcal{F} = \{\sigma^{-1} f(\cdot/\sigma) : \sigma \in (0, \infty)\}$ with $f(x) = f(-x)$ for all $x \in \mathbb{R}$, where the functional form of f is known. A popular and standard choice of f could be the standard normal density function.

Hence, the problem of estimating SVD robustly as in decomposition (2.2) can be regarded as a standard robust parametric inference problem. To solve this, we use the min-

imum density power divergence estimator (MDPDE) introduced in Section 1.3.4. Many authors (Basu et al., 1998, Ghosh and Basu, 2013, Toma et al., 2020) have noted that MDPDE possesses strong robustness properties without sacrificing much efficiency in estimation. In the high dimensional context where efficiency loss in the popular robust M-estimation techniques can be rapid (Maronna, 1976), MDPDE provides a suitable alternative method of estimation with high asymptotic breakdown as shown in Chapter 5. In the case of low-rank decomposition as in (2.2), the MDPDE is given by the minimizer of the objective function

$$H_{\alpha}^{(r)}(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p} V_{ij,\alpha}^{(r)}(\boldsymbol{\theta}) \tag{2.4}$$

where

$$V_{ij,\alpha}^{(r)}(\boldsymbol{\theta}) = \sigma^{-\alpha}\left[\|f\|_{1+\alpha}^{1+\alpha} - \left(1+\frac{1}{\alpha}\right) f^{\alpha}\left(\left|\frac{X_{ij}-\sum_{k=1}^{r}\lambda_k u_{ki} v_{kj}}{\sigma}\right|\right)\right]. \tag{2.5}$$

Here, the parameter $\boldsymbol{\theta} = \left(\boldsymbol{\Lambda}, \boldsymbol{U}, \boldsymbol{V}, \sigma^2\right)$ is restricted in the parameter space $[0,\infty)^r \times S_r^n \times S_r^p \times (0,\infty)$, where S_r^n is the r-Stiefel manifold of order n. The resulting matrices $\widehat{\boldsymbol{\Lambda}}$, $\widehat{\boldsymbol{U}}$ and $\widehat{\boldsymbol{V}}$ as a solution to the MDPDE objective function given in (2.4) is then defined to be the Robust SVD using Density Power Divergence (rSVDdpd) estimator of the data matrix $\boldsymbol{X}$ up to rank r.

A standard and popular choice for the scale family of densities is to consider the normal densities with mean 0 and unknown variance σ^2. In this case, the V-function given in (2.5) reduces to

$$V_{ij,\alpha}^{(r),\phi}(\boldsymbol{\theta}) = \sigma^{-\alpha}\left[\frac{1}{\sqrt{1+\alpha}} - \left(1+\frac{1}{\alpha}\right) e^{-\alpha(X_{ij}-\sum_{k=1}^{r}\lambda_k u_{ki} v_{kj})^2/2\sigma^2}\right]. \tag{2.6}$$

2.2.3 Estimation Algorithm

It is extremely difficult to perform a direct minimization of the objective function given in (2.4), since the quantities $\boldsymbol{U}$ and $\boldsymbol{V}$ are restricted to the nonconvex Stiefel manifolds. Following the footsteps of Rey (2007), we reformulate the decomposition in (2.2) as

$$\boldsymbol{X} = \sum_{k=1}^{r} \boldsymbol{a}_k \boldsymbol{b}_k^{\intercal} + \boldsymbol{E}, \tag{2.7}$$

where $\boldsymbol{a}_k$s and $\boldsymbol{b}_k$s are still orthogonal sets of vectors for $k = 1, 2, \ldots, r$, but not necessarily normalized. Once the estimates of $\boldsymbol{a}_k$s and $\boldsymbol{b}_k$s are known, they can be normalized to obtain the $\boldsymbol{u}_k$s and $\boldsymbol{v}_k$s and the singular values are then given by $\lambda_k = \|\boldsymbol{a}_k\|\,\|\boldsymbol{b}_k\|$ for each $k = 1, \ldots, r$. Similar to Rey (2007), we consider the rank one decomposition first, i.e.,

$$X_{ij} = a_i b_j + e_{ij},\ i = 1, \ldots n; j = 1, \ldots, p. \tag{2.8}$$

For notational convenience, we drop the suffix for the rank and use a_i and b_j to indicate the entries a_{1i} and b_{1j} respectively. For a fixed index j, the set of equations in (2.8) can be interpreted as a linear regression model with X_{ij}s as the observed responses, a_is as the covariates, b_js as the regression slope parameters to be estimated and e_{ij}s as the random error components modelled by the scale family of densities $\mathcal{F}$. As we vary the column

index $j = 1, 2, \ldots, p$, we are posed with p such linear regression problems, solving each of them will jointly yield an estimate of $\boldsymbol{b} = (b_1, \ldots, b_p)$ given the values of a_is. Now, one can interchange the role of a_is and b_js and estimate $\boldsymbol{a} = (a_1, \ldots, a_n)$ given the values of $\boldsymbol{b}$. To obtain the rank one "rSVDdpd" estimate of $\boldsymbol{X}$, we alternatively solve these regression problems robustly, using the MDPDE described by Ghosh and Basu (2013).

To find the MDPDE for the alternating regression problems, a standard differentiation of $H_\alpha^{(1)}$ as in (2.4) with respect to a_i, yields the estimating equation

$$\sum_{j=1}^{p} f^\alpha \left(|e_{ij}/\sigma|\right) s\left(|e_{ij}|/\sigma\right) \frac{e_{ij}}{2|e_{ij}|} b_{kj} = 0,$$

where $s(\cdot)$ is the score function of the density f, i.e.,

$$s(x) := \frac{f'(x)}{f(x)}, \ x \in \mathbb{R}, \tag{2.9}$$

assuming f' exists. The existence of f' will be later justified through Assumption (A1). Denoting

$$\psi(x) = -f^\alpha(|x|)s(|x|)/|x|, \tag{2.10}$$

this can be rearranged as

$$a_i \left(\sum_{j=1}^{p} b_j^2 \psi(e_{ij}/\sigma) \right) = \sum_{j=1}^{p} X_{ij} b_j \psi(e_{ij}/\sigma),$$

for all $i = 1, 2, \ldots, n$. Performing these steps for both the regression problems and for the estimating equation of σ^2 yields

$$a_i = \frac{\sum_j b_j X_{ij} \psi(e_{ij}/\sigma)}{\sum_j b_j^2 \psi(e_{ij}/\sigma)}, \qquad i = 1, \ldots n, \tag{2.11}$$

$$b_j = \frac{\sum_i a_i X_{ij} \psi(e_{ij}/\sigma)}{\sum_i a_i^2 \psi(e_{ij}/\sigma)}, \qquad j = 1, \ldots p, \tag{2.12}$$

$$\sigma^2 = \frac{(np)^{-1} \sum_i \sum_j e_{ij}^2 \psi(e_{ij}/\sigma)}{(np)^{-1} \sum_i \sum_j \psi(e_{ij}/\sigma) - \alpha(1+\alpha)^{-1} \|f\|_{1+\alpha}^{1+\alpha}}. \tag{2.13}$$

Therefore, one can alternatively use the iterative equations (2.11)-(2.13) based on the current estimates of a_is, b_js and σ^2 until convergence. To ensure that the estimate of σ^2 obtained using (2.13) remains positive throughout the iterations, we restrict $\sigma^2 = \epsilon$ for some small $\epsilon > 0$ when it becomes less than ϵ. As will be shown later in Theorem 2.1, under some reasonable assumptions on f, the sequence of estimators obtained through this set of iterating equations converge to the rSVDdpd estimator, a local minimizer of $H_\alpha^{(r)}(\boldsymbol{\theta})$ as in (2.4).

Remark 2.1. *An interesting observation is that the right-hand side (RHS) of Eq. (2.11) can be interpreted as a weighted average of the estimates of a_i from all the different columns. Namely, since $X_{ij} \approx a_i b_j$, it follows that X_{ij}/b_j is an approximation of a_i, and each such approximation is averaged with weights $b_j^2 \psi(e_{ij}/\sigma^2)$. Here, the $\psi(\cdot)$ function controls the robustness of the estimates. For instance, if it is decreasing, then errors*

e_{ij} with large magnitude get down-weighted, i.e., it controls the effect of outliers. Conversely, if ψ is an increasing function, then inliers are down-weighted. Here, we use the term "inlier" to indicate the observations that are extremely close to the model density with an error term very close to 0. A relevant discussion on inliers may be found in Section 2.3.5 of Basu et al. (2011).

Remark 2.2. *If f is the standard normal density, then $\psi(x) = e^{-\alpha x^2/2}$ for all $x > 0$. For $\alpha > 0$, this is a decreasing function of x and hence as mentioned in Remark 2.1, it leads to a robust SVD estimator by reducing the effect of large errors. However, for $\alpha = 0$, $\psi(x) = 1$ and the iterative equations (2.11)-(2.13) reduces to the estimation procedure of classical SVD (Hawkins et al., 2001). Hence, the proposed algorithm produces a class of SVD estimators including both robust and non-robust estimators.*

Algorithm 1 Iterative Algorithm for rSVDdpd estimator. Here, QR($\boldsymbol{M}$) denotes the QR-decomposition of matrix $\boldsymbol{M}$.

Require: $\boldsymbol{X}_{n\times p}$, $\alpha \in (0, 1]$ and $r \geq 0$.

Initialize $\boldsymbol{\Lambda}$ as an empty array, $\boldsymbol{A}_{n\times r}$ and $\boldsymbol{B}_{p\times r}$ with random orthogonal matrices of appropriate dimensions.
for $k = 1$ to r **do**
  $\boldsymbol{a}, \boldsymbol{b}, \lambda \leftarrow \boldsymbol{A}[, k], \boldsymbol{B}[, k], 1$
  $\sigma \leftarrow 1.4826 \times \text{med}_{i,j} \, |X_{ij} - \text{med}_{i,j} X_{ij}|$
  repeat
    $\boldsymbol{E} \leftarrow (\boldsymbol{X} - \lambda \boldsymbol{a}\boldsymbol{b}^\intercal)/\sigma$
    $\boldsymbol{a} \leftarrow \text{diag}(\psi(\boldsymbol{E})\boldsymbol{b}^{\odot 2})^{-1}(\boldsymbol{X} \odot \psi(\boldsymbol{E}))\boldsymbol{b}$
    $\boldsymbol{A}[, k] \leftarrow \boldsymbol{a}$
    $\boldsymbol{Q}, \boldsymbol{R} \leftarrow \text{QR}(\boldsymbol{A}[, 1 : k])$
    $\boldsymbol{a}, \lambda \leftarrow \boldsymbol{Q}[, k], \boldsymbol{R}[k, k]$
    $\boldsymbol{E} \leftarrow (\boldsymbol{X} - \lambda \boldsymbol{a}\boldsymbol{b}^\intercal)/\sigma$
    $\boldsymbol{b} \leftarrow \text{diag}(\psi(\boldsymbol{E})^\intercal \boldsymbol{a}^{\odot 2})^{-1}(\boldsymbol{X} \odot \psi(\boldsymbol{E}))^\intercal \boldsymbol{a}$
    $\boldsymbol{B}[, k] \leftarrow \boldsymbol{b}$
    $\boldsymbol{Q}, \boldsymbol{R} \leftarrow \text{QR}(\boldsymbol{B}[, 1 : k])$
    $\boldsymbol{b}, \lambda \leftarrow \boldsymbol{Q}[, k], \boldsymbol{R}[k, k]$
    $\boldsymbol{E} \leftarrow (\boldsymbol{X} - \lambda \boldsymbol{a}\boldsymbol{b}^\intercal)$
    $c \leftarrow \left(\sum_{i,j} \psi(e_{ij}/\sigma) - np\alpha \, \|f\|_{1+\alpha}^{1+\alpha} / (1 + \alpha)\right)$
    $\sigma^2 \leftarrow \max\left\{\epsilon, \left(\sum_{i,j} e_{ij}^2 \psi(e_{ij}/\sigma)\right) / c\right\}$
  until convergence
  Append λ to the array $\boldsymbol{\Lambda}$
  $\boldsymbol{X} \leftarrow \boldsymbol{X} - \boldsymbol{a}\boldsymbol{b}^\intercal$
end for

Given that such a rank-one decomposition as in (2.8) can be obtained, one can stack the estimates of a_is to get $\widehat{\boldsymbol{a}}$, and combine estimates of b_js to get $\widehat{\boldsymbol{b}}$, the non-normalized

first set of the singular vectors. Then these can be normalized to obtain an estimate of the first singular value as $\widehat{\lambda} = \|\widehat{\boldsymbol{a}}\| \|\widehat{\boldsymbol{b}}\|$ and an estimate of the corresponding singular vectors as $\widehat{\boldsymbol{u}} = \widehat{\boldsymbol{a}}/\|\widehat{\boldsymbol{a}}\|$ and $\widehat{\boldsymbol{v}} = \widehat{\boldsymbol{b}}/\|\widehat{\boldsymbol{b}}\|$. For the subsequent singular values, one can apply the same algorithm on the residual matrix $\boldsymbol{X} - \widehat{\lambda}\widehat{\boldsymbol{u}}\widehat{\boldsymbol{v}}^{\intercal}$. Such a method to reuse rank one approximation is quite common in the matrix factorization literature (Hawkins et al., 2001, Cichocki et al., 2011). However, this method does not guarantee that the subsequent singular vectors remain orthonormal to the previous set of singular vectors (Yang, 2014). The orthogonality property usually degrades as one estimates more and more singular values. However, in between two singular value estimation steps, one can perform a Gram-Schmidt orthogonalization trick (Giraud et al., 2005) using QR decomposition to restore the orthogonal property. The exact technical details of the estimation algorithm are outlined in Algorithm 1.

2.2.4 Choice of the Hyperparameters

There are two hyperparameters to the rSVDdpd algorithm: the robustness parameter $\alpha \in [0, 1]$ and the rank r of the low-rank component $\boldsymbol{L}$ of $\boldsymbol{X}$ matrix. In this section, we discuss the choice of these parameters in a data-driven manner.

The robustness parameter α in the objective function (2.4) provides a bridge between robustness and efficiency of estimation. Through extensive simulation, we have seen that increasing α reduces bias in estimation at the cost of increasing variability, under any kind of contamination. Thus, to determine the optimal choice of α, we consider a conditional MSE criterion as in Huang et al. (2009). Some elementary calculations yield that the optimal choice of the robustness parameter is the minimizer of the criterion

$$(n+p)(\widehat{\sigma}^{(\alpha)})^2 \left(1 + \frac{\alpha^2}{1+2\alpha}\right)^{3/2} + \frac{1}{r}\sum_{k=1}^{r} \|\widehat{\boldsymbol{a}}_k^{(\alpha)} - \widehat{\boldsymbol{a}}_k^{(1)}\|_2^2 + \frac{1}{r}\sum_{k=1}^{r} \|\widehat{\boldsymbol{b}}_k^{(\alpha)} - \widehat{\boldsymbol{b}}_k^{(1)}\|_2^2 \qquad (2.14)$$

where $\widehat{\boldsymbol{a}}_k^{(\alpha)}, \widehat{\boldsymbol{b}}_k^{(\alpha)}$ are the estimates of k-th non-normalized singular vectors, and $(\widehat{\sigma}^{(\alpha)})^2$ is the estimated error variance as obtained by rSVDdpd algorithm with robustness parameter α. Therefore, one may choose an equally spaced grid of values of α in the range $[0, 1]$, and then perform a grid search over these values to minimize the criterion indicated in (2.14).

The rank r is often determined by the context in which the robust SVD is being used. For example, for video surveillance background modelling applications, considering a rank-one or rank-two decomposition is often enough to accurately capture the variations present in the background content. In general, one may use the simple Elbow method or a generalized bi-cross validation method as described by Owen and Perry (2009) to estimate the rank r in a data-dependent manner. Based on our simulations, we have found that the Elbow method (i.e., finding out when there is a large drop in two subsequent singular values followed by a significantly smaller drop) works well in our simulation settings, and hence is used for its computational simplicity. However, later in Chapter 4, we propose a new criterion to tackle the problem of rank estimation in detail.

2.3 Mathematical Analysis

Since the estimation process of the subsequent singular values and vectors follows from the same estimation technique of rank one decomposition on the residual matrix, we shall restrict our attention to the study of different properties of the rSVDdpd estimator for the rank one decomposition only. To ensure all the results developed in this section are true for the subsequent singular values and vectors as well, the assumptions on the distribution of the data matrix $\boldsymbol{X}$ must hold for the residual matrix after subtracting the effect of the previous singular values and vectors. We assume that this is indeed the case.

In the context of rank-one estimation, we redefine the parameter comprising only the first singular value and corresponding vectors, $\boldsymbol{\theta} = \left(\lambda, \{u_i\}_{i=1}^n, \{v_j\}_{j=1}^p, \sigma^2\right)$. The corresponding parameter space becomes $\boldsymbol{\Theta} = (0,\infty) \times S^{n+} \times S^p \times [\epsilon, \infty)$, where $\epsilon > 0$ is a small positive quantity and S^{n+} is the space of orthogonal vectors of length n such that its first nonzero coordinate is positive. Such a restriction on the parameter space is necessary to ensure the identifiability of the rank-one robust SVD problem, since one can switch the signs of both $\boldsymbol{u}$ and $\boldsymbol{v}$ resulting in the same decomposition. In view of (2.11)-(2.13), the sequence of estimates $\boldsymbol{\theta}^{(t)}$ is then related in the following way,

$$\lambda^{(t+1)} u_i^{(t+1)} = \frac{\sum_j v_j^{(t)} X_{ij} \psi(e_{ij}^{(t)}/\sigma^{(t)})}{\sum_j (v_j^{(t)})^2 \psi(e_{ij}^{(t)}/\sigma^{(t)})} \tag{2.15}$$

$$\lambda^{(t+1)} v_j^{(t+1)} = \frac{\sum_i u_i^{(t+1)} X_{ij} \psi(e_{ij}^{(t)}/\sigma^{(t)})}{\sum_i (u_i^{(t+1)})^2 \psi(e_{ij}^{(t)}/\sigma^{(t)})} \tag{2.16}$$

$$(\sigma^2)^{(t+1)} = \frac{(np)^{-1} \sum_{i,j} (e_{ij}^{(t)})^2 \psi(e_{ij}^{(t)}/\sigma^{(t)})}{(np)^{-1} \sum_{i,j} \psi(e_{ij}^{(t)}/\sigma^{(t)}) - \alpha \left\| f \right\|_{1+\alpha}^{1+\alpha} / (1+\alpha)} \tag{2.17}$$

for all $t = 0, 1, \ldots$. Here, we use $\boldsymbol{\theta}^{(t)}$ and its elements to denote the value of the estimated parameter at t-th iteration of the algorithm. Before proceeding with the statistical properties of the rSVDdpd estimator, we establish the convergence of the above iterative procedure under three simple assumptions.

(A1) Assume that the model density f is twice differentiable with respect to its arguments.

(A2) The model density f is symmetric and satisfies $f'(x) \leq 0$ for all $x > 0$ and

$$\frac{1}{x} > \alpha \frac{f'(x)}{f(x)} + s'(x) \frac{f(x)}{f'(x)}, \text{ for } x > 0. \tag{2.18}$$

where $s(x)$ is the score function as defined in (2.9).

(A3) There exists a constant K such that $x^2 \psi(x) < K$ for all $x \geq 0$ where ψ is as given in (2.10).

Remark 2.3. *The Assumption (A2) is not a very strict condition on the choice of f. As a result of $f'(x) \leq 0$, we obtain $\psi(x) \geq 0$, and the condition in (2.18) implies only that ψ is decreasing. Thus, the provided conditions simply amount to the requirement that the*

weights in (2.11)-(2.13) are nonnegative, and the larger errors are down-weighted so that it leads to a robust estimator. It also means when the errors are large in magnitude, their reduced effect due to small value of $\psi(\cdot)$ ensures that the estimates remain stable from one iteration step to another, leading to accelerated convergence.

Remark 2.4. *When f is standard normal density, $\psi(x) = e^{-\alpha x^2/2}$. This means that Assumption (A2) is automatically satisfied due to the nonnegativity and decreasing nature of ψ. Assumption (A3) is satisfied in this case with $K = 2/\alpha e$.*

Theorem 2.1. *For a fixed n and p and the data matrix $\boldsymbol{X}_{n\times p}$, if the Assumptions (A1)-(A3) hold, then the sequence of estimates $\boldsymbol{\theta}^{(t)}$ obtained through (2.15)-(2.17) converges to a local minimizer of $H_\alpha^{(1)}(\boldsymbol{\theta})$ given in (2.4).*

Proof. Note that, each $V_{ij,\alpha}^{(1)}(\boldsymbol{\theta})$ as in (2.5) bounded below by the finite quantity

$$\epsilon^{-\alpha}\left(\|f\|_{1+\alpha}^{1+\alpha} - \left(1+\frac{1}{\alpha}\right) f^\alpha(0)\right),$$

due to the decreasing nature of f. Hence the same lower bound also applies for $H_\alpha^{(1)}(\boldsymbol{\theta})$. Therefore, for any compact subset $K \subset \boldsymbol{\Theta}$, there exists at least one local minimum of $H_\alpha^{(1)}(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in K$.

We first show that iterative equations (2.15)-(2.17) reduces the value of the objective function $H_\alpha^{(1)}(\boldsymbol{\theta})$. We shall show this only for Eq. (2.15), as the rest can be shown similarly. Let, $e_{ij}^{(t+1/2)} = X_{ij} - \lambda^{(t+1)} u_i^{(t+1)} v_j^{(t)}$. Then,

$$e_{ij}^{(t)} = e_{ij}^{(t+1/2)} + v_j^{(t)} \frac{\sum_k v_k^{(t)} e_{ik}^{(t)} \psi(e_{ik}^{(t)}/\sigma^{(t)})}{\sum_k (v_k^{(t)})^2 \psi(e_{ik}^{(t)}/\sigma^{(t)})}.$$

Let's call the second term in the above sum as $v_j^{(t)} e_i^*$. We aim to obtain an upper bound of $|e_i^*|$. Note that, we can divide both sides of the equation by $\sigma^{(t)}$ and consider $e_{ik}^{(t)}/\sigma^{(t)}$ as the rescaled errors, hence without the loss of generality, we can assume that $\sigma^{(t)} = 1$. By an application of Cauchy-Schwartz inequality and the Assumption (A3), it follows that

$$\begin{aligned}
\left|\sum_k v_k^{(t)} e_{ik}^{(t)} \psi(e_{ik}^{(t)})\right| &\le \max\left\{1, \left|\sum_k v_k^{(t)} e_{ik}^{(t)} \psi(e_{ik}^{(t)})\right|^2\right\} \\
&\le \max\left\{1, \left(\sum_k (v_k^{(t)})^2 \psi(e_{ik}^{(t)})\right)\left(\sum_k (e_{ik}^{(t)})^2 \psi(e_{ik}^{(t)})\right)\right\} \\
&\le \max\left\{1, nK\left(\sum_k (v_k^{(t)})^2 \psi(e_{ik}^{(t)})\right)\right\},
\end{aligned}$$

where the last inequality follows from Assumption (A3). Therefore, it follows that

$$|e_i^*| \le \max\left\{\frac{1}{\sum_k (v_k^{(t)})^2 \psi(e_{ik}^{(t)})}, nK\right\} \le \max\left\{nK, 1/\psi(\max_{i,j} |e_{ij}^{(t)}|)\right\},$$

due to the decreasing nature of ψ. Note that, $\max_{i,j} |e_{ij}^{(t)}|$ is a finite quantity since it is maximum over a finite set, for fixed n, p and t. We call the bound at the right-hand side as $B_1(\max_{i,j} |e_{ij}^{(t)}|)$.

Moving on, in view of the definition of $H_\alpha^{(1)}(\boldsymbol{\theta})$, it is now enough to show that

$$\sum_{i,j}\left[f^\alpha\left(\frac{|e_{ij}^{(t+1/2)}|}{\sigma^{(t)}}\right)-f^\alpha\left(\frac{|e_{ij}^{(t)}|}{\sigma^{(t)}}\right)\right]\geq 0.$$

Again, due to the rescaling, we assume $\sigma^{(t)}=1$, so an application of Taylor's theorem yields

$$f^\alpha\left(|e_{ij}^{(t+1/2)}|\right)-f^\alpha\left(|e_{ij}^{(t)}|\right)=-\alpha\psi\left(|e_{ij}^{(t)}|\right)\left(|e_{ij}^{(t)}-v_j^{(t)}e_i^*|-|e_{ij}^{(t)}|\right)$$
$$-\frac{\alpha}{2}\psi'(c)(|e_{ij}^{(t)}-v_j^{(t)}e_i^*|-|e_{ij}^{(t)}|)^2,$$

where c is some value between $e_{ij}^{(t+1/2)}$ and $e_{ij}^{(t)}$. Because of Assumption (A2), we have $\psi'(c)<0$ and hence the second term on the right hand side is nonnegative. For the first term, consider the inequality

$$\sum_j\psi\left(|e_{ij}^{(t)}|\right)\left(|e_{ij}^{(t)}|-|e_{ij}^{(t)}-v_j^{(t)}e_i^*|\right)=\sum_j\psi\left(|e_{ij}^{(t)}|\right)\frac{(e_{ij}^{(t)})^2-(e_{ij}^{(t)}-v_j^{(t)}e_i^*)^2}{|e_{ij}^{(t)}|+|e_{ij}^{(t)}-v_j^{(t)}e_i^*|}$$
$$\geq\sum_j\psi\left(|e_{ij}^{(t)}|\right)\frac{(e_{ij}^{(t)})^2-(e_{ij}^{(t)}-v_j^{(t)}e_i^*)^2}{2\max_{i,j}|e_{ij}^{(t)}|+B_1(\max_{i,j}|e_{ij}^{(t)}|)},$$

This lower bound is nonnegative since by the structure of e_i^*, it minimizes the weighted squared error

$$\sum_j\psi\left(|e_{ij}^{(t)}|\right)(e_{ij}^{(t)}-v_j^{(t)}a)^2,$$

over choice of all possible $a\in\mathbb{R}$. Adding these quantities for all i and putting it back to the Taylor's series shows that each iteration decreases the value of the objective function $H_\alpha^{(1)}$. Now, the sequence $\{H_\alpha^{(1)}(\boldsymbol{\theta}^{(t)})\}_{t=0}^\infty$ becomes a decreasing sequence of real numbers bounded below (as shown before), and hence has a convergent subsequence by Bolzano-Weierstrass Theorem. Let $\|\boldsymbol{X}\|_F$ denote the Frobenius norm of $\boldsymbol{X}$ given by

$$\|\boldsymbol{X}\|_F=\left(\sum_{i=1}^n\sum_{j=1}^p|X_{ij}|^2\right)^{1/2}.$$

Then, the facts that $H_\alpha^{(1)}$ is a continuous function of $\boldsymbol{\theta}$ due to continuity of f, and that $\boldsymbol{\Theta}$ can effectively be restricted to a compact set $[0,\|\boldsymbol{X}\|_F]\times S^{n+}\times S^p\times[\epsilon,\|\boldsymbol{X}\|_F]$, imply that $\boldsymbol{\theta}^{(t)}$ converges to some $\boldsymbol{\theta}^*$. Finally, since $\boldsymbol{\theta}^*$ satisfies the iterating equations (2.15)-(2.17), it in turn, satisfies the estimating equations, i.e., the gradient of $H_\alpha^{(1)}$ at $\boldsymbol{\theta}^*$ is zero. This implies that $\boldsymbol{\theta}^*$ is a local minimum as we wanted. □

Let us denote the converged estimator as $\boldsymbol{\theta}^*=(\lambda^*,\{u_i^*\}_{i=1}^n,\{v_j^*\}_{j=1}^p,(\sigma^*)^2)$, whose existence is now assured by Theorem 2.1. On the other hand, let us denote the population counterpart as $\boldsymbol{\theta}^g=(\lambda^g,\{u_i^g\}_{i=1}^n,\{v_j^g\}_{j=1}^p,(\sigma^g)^2)$, which is the true value of the parameters that are ultimately being estimated. Similar to the iteration rules (2.15)-(2.17) for the rSVDdpd estimator, the true value $\boldsymbol{\theta}^g$ is expected to satisfy such fixed point criteria, in the

sense of overall population based measures rather than its empirical counterparts. With this in mind, we define these "best" fitting parameters for the particular setup of SVD under consideration.

Definition 1. *Let, $\boldsymbol{X}$ be a data matrix of order $n \times p$ such that its entries X_{ij}s are independent but not identically distributed with respective density functions g_{ij}, for all $i = 1, \ldots n$ and $j = 1, \ldots p$. Then, $\boldsymbol{\theta}^g = (\lambda^g, \{u_i^g\}_{i=1}^n, \{v_j^g\}_{j=1}^p, (\sigma^g)^2)$ is called a "best" fitting parameter for the robust SVD problem if the following conditions hold*

1. *The $\{u_i^g\}$s and $\{v_j^g\}$s constitute entries of unit vectors, i.e.,*

$$\sum_{i=1}^{n}(u_i^g)^2 = 1, \quad \text{and} \quad \sum_{j=1}^{p}(v_j^g)^2 = 1. \tag{2.19}$$

2. *For any $i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, p$,*

$$\lambda^g u_i^g = \arg\min_a \int V_f(\cdot; a, v_j^g, (\sigma^g)^2) g_{ij}. \tag{2.20}$$

3. *For any $i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, p$,*

$$\lambda^g v_j^g = \arg\min_b \int V_f(\cdot; u_i^g, b, (\sigma^g)^2) g_{ij}. \tag{2.21}$$

4. *For any $i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, p$,*

$$(\sigma^g)^2 = \arg\min_{\sigma^2} \int V_f(\cdot; \lambda^g u_i^g, v_j^g, \sigma^2) g_{ij} = \arg\min_{\sigma^2} \int V_f(\cdot; u_i^g, \lambda^g v_j^g, \sigma^2) g_{ij}. \tag{2.22}$$

where

$$V_f(x; a, b, \sigma^2) = \sigma^{-\alpha} \left[\|f\|_{1+\alpha}^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) f^\alpha \left(\left| \frac{x - ab}{\sigma} \right| \right) \right]. \tag{2.23}$$

Here, (2.20) shows that the minimizer of the quantity on the right-hand side of the equation is always $\lambda^g a_i^g$, independent of the choice of column index j. One example where this assumption holds is when the true densities g_{ij} are densities of the normal distributions with the location parameters being elements from the best rank one approximation of $\boldsymbol{X}$, i.e., the entries of the data matrix X_{ij}s are normally distributed with mean μ_{ij} and constant variance σ^2, and the matrix $\boldsymbol{\mu} = (\mu_{ij})_{i=1,j=1}^{n,p}$ is of unit rank.

For the popular case of modelling the errors via normal distribution, the V_f function as in (2.23) reduces to

$$V_\phi(x; a, b, \sigma^2) = \sigma^{-\alpha} \left(\frac{1}{\sqrt{1+\alpha}} - \left(1 + \frac{1}{\alpha}\right) \phi^\alpha \left(\frac{x - ab}{\sigma} \right) \right) \tag{2.24}$$

where ϕ is the standard normal density function given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},\ x \in \mathbb{R}.$$

2.3.1 Uniqueness

Since the aim of the rSVDdpd algorithm is to robustly estimate the singular values and the singular vectors of a given data matrix, it is required to show that the "best" fitting parameters as introduced by Definition 1 resemble the behaviour of the usual singular values and vectors. Regarding this, the following theorem claims that if the elements of the data matrix $\boldsymbol{X}$ follow a decomposition as in (2.2), then the "best" fitting parameter given by Definition 1 matches exactly with the usual singular values and vectors.

Theorem 2.2. *Let the data matrix $\boldsymbol{X}$ be such that $X_{ij} = \lambda^* u_i^* v_j^* + \epsilon_{ij}$ where ϵ_{ij}s are i.i.d. random variables with density $(\sigma^*)^{-1} f(\cdot/\sigma^*)$. Then $\boldsymbol{\theta}^g$ is the unique "best" fitting parameter if $\boldsymbol{\theta}^g = \left(\lambda^*, \{u_i^*\}_{i=1}^n, \{v_j^*\}_{j=1}^p, (\sigma^*)^2\right) \in \boldsymbol{\Theta}$, the parameter space.*

Proof. The proof follows from verifying that the postulated $\boldsymbol{\theta}^g$ indeed satisfies all the conditions present in Definition 1. Relation (2.19) is verified by the implications that u_i^* and v_j^*s belong to the respective Stiefel manifolds. To verify (2.20), note that with $v_j^g = v_j^*$ and $\sigma^g = (\sigma^*)$, the quantity in (2.20) is equal to the minimizer of

$$\int V_f(x; a, v_j^*, \sigma^*) + \frac{1}{\alpha} \int g_{ij}^{\alpha},$$

over a, since the last term is independent of the minimization over a. But this is exactly same as the density power divergence (DPD) between the model density f and true density g_{ij}. From Theorem 2.1 of Basu et al. (1998), it follows that this divergence is minimized if and only if two densities match, i.e., $a = \lambda^* u_i^*$. By similar logic and interchanging the roles of u_i and v_j, (2.21) and (2.22) can also be verified. This proves that $\boldsymbol{\theta}^* = (\lambda^*, \{u_i^*\}_{i=1}^n, \{v_j^*\}_{j=1}^p, (\sigma^*)^2)$ is a "best" fitting parameter for the given setup.

In order to prove uniqueness, suppose $\widetilde{\boldsymbol{\theta}} = (\widetilde{\lambda}, \{\tilde{u}_i\}_{i=1}^n, \{\tilde{v}_j\}_{j=1}^p, \tilde{\sigma}^2)$ be another "best" fitting parameter. Then again, the DPD with v_j and σ substituted for v_j^* and σ^* respectively, is minimized at $a = \widetilde{\lambda}\tilde{u}_i$ independently of the choice of j. However, this divergence can be made equal to its minimum value 0 if and only if $\tilde{\sigma}^2 = (\sigma^*)^2$, and

$$\lambda u_i v_j = \lambda^* u_i^* v_j^*, \;\; i = 1, \ldots, n; j = 1, \ldots, p, \tag{2.25}$$

which follows from Theorem 2.1 of Basu et al. (1998). Since, both $\widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ are "best" fitting parameters, they must satisfy (2.19). Hence, we have a chain of equality

$$(\widetilde{\lambda}\tilde{u}_i)^2 = \sum_j (\widetilde{\lambda}\tilde{u}_i\tilde{v}_j)^2 = \sum_j (\lambda^* u_i^* v_j^*)^2 = (\lambda^* u_i^*)^2.$$

Taking sum over the row index i now yields $\widetilde{\lambda}^2 = (\lambda^*)^2$. Since both $\widetilde{\lambda}, \lambda^* \geq 0$, it follows that $\widetilde{\lambda} = \lambda^*$, and consequently, $|\tilde{u}_i| = |u_i^*|$ and $|\tilde{v}_j| = |v_j^*|$.

Now suppose $\tilde{v}_j = (-v_j^*)$ for some j. Along with Eq. (2.25), it means that $\tilde{u}_i = (-u_i^*)$ for all $i = 1, 2, \ldots, n$. This leads to a contradiction since both $\tilde{\boldsymbol{u}} = (\tilde{u}_1, \ldots, \tilde{u}_n)$ and $\boldsymbol{u}^* = (u_1^*, \ldots, u_n^*)$ cannot be in S^{n+}, as one of them will have the first nonzero coordinate negative. □

The following corollaries of Theorem 2.2 now immediately establish the validity of the best fitting parameter, for the case when the entries of the data matrix $\boldsymbol{X}$ follow a normal distribution, or are deterministic (which is a special case of the family of normal distribution with variance parameter equal to 0).

Corollary 2.1. *Let the data matrix $\boldsymbol{X}$ be such that X_{ij} are i.i.d. $\mathcal{N}(\lambda^* u_i^* v_j^*, (\sigma^*)^2)$. Then $\boldsymbol{\theta}^g$ is the unique "best" fitting parameter with normal model family of densities if $\boldsymbol{\theta}^g = \left(\lambda^*, \{u_i^*\}_{i=1}^n, \{v_j^*\}_{j=1}^p, (\sigma^*)^2\right) \in \boldsymbol{\Theta}$. In this case, the V-function will be considered as given in (2.24).*

Corollary 2.2. *If the data matrix $\boldsymbol{X}$ is of rank 1 such that $X_{ij} = \lambda^* u_i^* v_j^*$ for all $i = 1, \ldots, n$, $j = 1, \ldots, p$, with $\sum_i (u_i^*)^2 = \sum_j (v_j^*)^2 = 1$, then there exists a unique "best" fitting parameter given by $\boldsymbol{\theta}^g = (\lambda^*, \{u_i^*\}_{i=1}^n, \{v_j^*\}_{j=1}^p, 0)$ if $\boldsymbol{\theta}^g \in \boldsymbol{\Theta}$.*

2.3.2 Equivariance Properties

We have seen that if the entries of $\boldsymbol{X}$ are of special structure as in (2.1) without the sparse term $\boldsymbol{S}$, i.e., only low-rank component plus the i.i.d. noise component, under correct specification of the family of densities f, the best fitting parameter equals the singular values and the corresponding vectors. However, in case of misspecification or if $\boldsymbol{X}$ includes the sparse component $\boldsymbol{S}$, it still contains independent entries and, therefore, we can show that a "best" fitting parameter satisfies equivariance properties similar to what the traditional singular values and vectors will satisfy. One such equivariance property is that whenever the data matrix is multiplied by a positive scalar quantity, the singular values are also multiplied by the same scalar quantity. However, the singular vectors remain unchanged. The following theorem presents this equivariance property for a "best" fitting parameter.

Theorem 2.3. *If a "best" fitting parameter for matrix $\boldsymbol{X}$ is $\boldsymbol{\theta}^g$, then a "best" fitting parameter for the matrix $c\boldsymbol{X}$ is $\widetilde{\boldsymbol{\theta}}^g = (c\lambda^g, \{u_i^g\}_{i=1}^n, \{v_j^g\}_{j=1}^p, (c\sigma^g)^2)$ for any positive real constant c.*

Proof. It is obvious that $\widetilde{\boldsymbol{\theta}}^g$ satisfies (2.19) as $\boldsymbol{\theta}^g$ is given to be a "best" fitting parameter. Considering the matrix $\boldsymbol{Y} = c\boldsymbol{X}$, let us denote the true density of Y_{ij} as $g_{ij}^Y(\cdot)$, as opposed to $g_{ij}(\cdot)$ denoting the true density of X_{ij}. A change of variable formula yields that $g_{ij}^Y(y) = c^{-1} g_{ij}(y/c)$. Hence, from the substitution principle of integration, it follows that

$$\int V_f\left(y; a, v_j^g, c^2(\sigma^g)^2\right) g_{ij}^Y(y)dy = c^{-\alpha} \int V_f\left(z; a/c, v_j^g, (\sigma^g)^2\right) g_{ij}(z)dz.$$

Since the right-hand side is minimized at $a/c = \lambda^g u_i^g$, the left-hand side is minimized at $a = c\lambda^g u_i^g$. This verifies (2.20). In a similar manner, (2.21) can also be established by interchanging the role of $\boldsymbol{u}$ and $\boldsymbol{v}$ above. For the parameter σ, again a substitution principle applies, and we obtain

$$\int V_f(y; c\lambda^g u_i^g, v_j^g, \sigma^2) g_{ij}^Y(y)dy = c^{-\alpha} \int V_f(z; \lambda u_i^g, v_j^g, \sigma^2/c^2) g_{ij}(z)dz.$$

Again by the hypothesis that $\boldsymbol{\theta}^g$ is a "best" fitting parameter, the latter is minimized when $\sigma/c = \sigma^g$, hence the former is minimized at $\sigma^2 = c^2(\sigma^g)^2$. This verifies (2.22). □

Another equivariance property of the traditional singular value decomposition is that under any row (or column) permutation of the data matrix, the entries of the left (or right) singular vectors permute accordingly, while the singular values remain unaffected. Such a property also holds for a "best" fitting parameter. Let π_R and π_C denote some such permutations on the row and column indices of the data matrix $\boldsymbol{X}$ respectively.

Theorem 2.4. *Let $\boldsymbol{P}, \boldsymbol{Q}$ be permutation matrices corresponding to the permutations π_R and π_C, on the row and column indices of the matrix $\boldsymbol{X}$. If a "best" fitting parameter for the matrix $\boldsymbol{X}$ is $\boldsymbol{\theta}^g$, then a "best" fitting parameter for permuted matrix $\boldsymbol{PXQ}^\intercal$ is the correspondingly permuted version of $\boldsymbol{\theta}^g$ given by $\widetilde{\boldsymbol{\theta}}^g = (\lambda^g, \{(\boldsymbol{Pu^g})_i\}_{i=1}^n, \{(\boldsymbol{Qv^g})_j\}_{j=1}^p, (\sigma^g)^2)$.*

Proof. Let, $\boldsymbol{Y} = \boldsymbol{PXQ}^\intercal$. Then, (2.19) is satisfied for the new setup as $\sum_{i=1}^n (u^g_{\pi_R(i)})^2 = \sum_{i=1}^n (u_i^g)^2 = 1$ and similarly $\sum_{j=1}^p (v^g_{\pi_C(j)})^2 = \sum_{j=1}^p (v_j^g)^2 = 1$. To see that (2.20) hold for the new setup with $\widetilde{\theta}^g$, note that for every $j = 1, 2, \ldots, p$, the minimizer of the integral in (2.20) is u_i^g, independent of the choice of j. Now, considering (2.20) for $\pi_R^{-1}(i)$ and $\pi_C^{-1}(j)$ instead of i and j, we obtain

$$\lambda^g u^g_{\pi_R^{-1}(i)} = \arg\min_a \int V_f(x; a, v^g_{\pi_C^{-1}(j)}, (\sigma^g)^2) g_{\pi_R^{-1}(i), \pi_C^{-1}(j)}(x) dx. \tag{2.26}$$

Note that, if we consider a vector $\boldsymbol{u}^g$ with its entries u_i^g, then $u^g_{\pi_R^{-1}(i)}$ is the i-th entry of $\boldsymbol{Pu}$. Similarly, $v^g_{\pi_C^{-1}(j)}$ is the j-th entry of $\boldsymbol{Qv}$. And finally, the (i,j)-th entry of the new matrix $\boldsymbol{Y}$ is same as $X_{\pi_R^{-1}(i), \pi_C^{-1}(j)}$, thus the density for the element Y_{ij} is $g^Y_{ij}(y) = g_{\pi_R^{-1}(i), \pi_C^{-1}(j)}(y)$ which can also be verified by a change of variable formula. Combining these, (2.26) can be reformulated as

$$\lambda^g (\boldsymbol{Pu}^g)_i = \arg\min_a \int V_f(y; a, (\boldsymbol{Qv}^g)_j, (\sigma^g)^2) g^Y_{ij}(y) dy.$$

This shows that (2.20) holds for new matrix $\boldsymbol{PXQ}^\intercal$ with the given best fitting parameter $\widetilde{\boldsymbol{\theta}}^g$. The relation (2.21) holds by imitating the same proof, except interchanging the role of $\boldsymbol{u}$ and $\boldsymbol{v}$. Finally, relation (2.22) for the permuted matrix follows from noting that

$$\begin{aligned} & \int V_f(x; \lambda^g u^g_{\pi_R^{-1}(i)}, v^g_{\pi_C^{-1}(j)}, \sigma^2) g_{\pi_R^{-1}(i), \pi_C^{-1}(j)}(x) dx \\ = & \int V_f\left(y; \lambda^g (\boldsymbol{Pu}^g)_i, (\boldsymbol{Qv}^g)_j, \sigma^2\right) g^Y_{ij}(y) dy. \end{aligned}$$

□

Please note that the Theorems 2.3 and 2.4 impose no restrictions or assumptions on the model density f or on the true density g underlying the entries of the data matrix $\boldsymbol{X}$. Hence, even in cases of model misspecification or the presence of outliers in the data, the estimated singular values and vectors obtained from the rSVDdpd algorithm still adhere to these desirable properties. Consequently, it's possible to appropriately scale the entries of the data matrix $\boldsymbol{X}$ before initializing Algorithm 1, and then adjust the scaling constant

to the singular values in the final step. This rescaling process significantly enhances the numerical stability of the rSVDpd algorithm. Furthermore, this can also be justified from another viewpoint: By demonstrating that these equivariance properties hold not only for the "best" fitting parameter but also for the converged rSVDdpd estimator, provided that the initial estimates exhibit a similar equivariance structure.

Theorem 2.5. *Let $\boldsymbol{\theta}^*$ be the converged rSVDdpd estimator for the matrix $\boldsymbol{X}$, starting from an initial estimate $\boldsymbol{\theta}^{(0)}$. Then, for any constant $c > 0$, the rSVDdpd estimator for the data matrix $\boldsymbol{X'} = c\boldsymbol{X}$ converges to $(c\lambda^*, \{u_i^*\}_{i=1}^n, \{v_j^*\}_{j=1}^p, (c\sigma^*)^2)$ provided the initial estimate is $(c\lambda^{(0)}, \{u_i^{(0)}\}_{i=1}^n, \{v_j^{(0)}\}_{j=1}^p, (c\sigma^{(0)})^2)$.*

Proof. Let us denote $\boldsymbol{\theta}^{(t)}$ denote the estimate at t-th iteration for data matrix $\boldsymbol{X}$ and let $\widetilde{\boldsymbol{\theta}}^{(t)}$ denote the same for the matrix $c\boldsymbol{X}$. Clearly, it is then enough to show that for all $t = 1, 2, \ldots,$

$$\widetilde{\lambda}^{(t)} = c\lambda^{(t)},\ (\widetilde{\sigma}^{(t)})^2 = c^2(\sigma^{(t)})^2,\ \widetilde{u}_i^{(t)} = u_i^{(t)},\ \widetilde{v}_j^{(t)} = v_j^{(t)}.$$

We will show this by using the principle of mathematical induction. For $t = 0$, the claim is validated by the equivariance of the initial estimate. To show the inductive step, we first consider (2.15). Note that, the error at t-th step for the new matrix satisfy

$$\widetilde{e}_{ij}^{(t)} = cX_{ij} - \widetilde{\lambda}^{(t)}\widetilde{u}_i^{(t)}\widetilde{v}_j^{(t)} = ce_{ij}^{(t)},$$

by induction hypothesis. Therefore,

$$\widetilde{\lambda}^{(t+1)}\widetilde{u}_i^{(t+1)} = \frac{\sum_j c\widetilde{v}_j^{(t)} X_{ij}\psi(\widetilde{e}_{ij}^{(t)}/\widetilde{\sigma}^{(t)})}{\sum_j (\widetilde{v}_j^{(t)})^2\psi(\widetilde{e}_{ij}^{(t)}/\widetilde{\sigma}^{(t)})} = c\frac{\sum_j v_j^{(t)} X_{ij}\psi(ce_{ij}^{(t)}/c\sigma^{(t)})}{\sum_j (v_j^{(t)})^2\psi(ce_{ij}^{(t)}/c\sigma^{(t)})} = c\lambda^{(t+1)}u_i^{(t+1)}.$$

Performing the same steps with (2.16) and (2.17) ensure that

$$\widetilde{\lambda}^{(t+1)}\widetilde{v}_j^{(t+1)} = c\lambda^{(t+1)}v_j^{(t+1)},\ \widetilde{\sigma}^{(t+1)} = c\sigma^{(t+1)}.$$

Finally, since the estimates of the singular vectors are normalized and restricted to be in the parameter space $\boldsymbol{\Theta} = [0, \infty) \times S^{n+} \times S^p \times [0, \infty)$, the inductive step follows from the normalization. □

Theorem 2.6. *Let $\boldsymbol{\theta}^*$ be the converged rSVDdpd estimator for the matrix $\boldsymbol{X}$, starting from an initial estimate $\boldsymbol{\theta}^{(0)}$. Also, let $\boldsymbol{P}$ and $\boldsymbol{Q}$ be the permutation matrices corresponding to the permutations π_R and π_C respectively. Then, starting with the new initial estimate $(\lambda^{(0)}, \{u_{\pi_R^{-1}(i)}^{(0)}\}_{i=1}^n, \{v_{\pi_C^{-1}(j)}^{(0)}\}_{j=1}^p, (\sigma^{(0)})^2)$, the rSVDdpd estimator for the data matrix $\boldsymbol{PXQ}^\intercal$ converges to the corresponding permuted version of $\boldsymbol{\theta}^*$ which is given by $(\lambda^*, \{u_{\pi_R^{-1}(i)}^*\}_{i=1}^n, \{v_{\pi_C^{-1}(j)}^*\}_{j=1}^p, (\sigma^*)^2)$.*

Proof. This proof is very similar to the proof of Theorem 2.5. We shall again denote $\boldsymbol{\theta}^{(t)}$ and $\widetilde{\boldsymbol{\theta}}^{(t)}$ as the estimates at the t-th iteration for the data matrices $\boldsymbol{X}$ and $\boldsymbol{PXQ}^\intercal$ respectively. Again, it is enough to show that for all $t = 1, 2, \ldots,$

$$\widetilde{\lambda}^{(t)} = \lambda^{(t)},\ (\widetilde{\sigma}^{(t)})^2 = (\sigma^{(t)})^2,\ \widetilde{u}_i^{(t)} = u_{\pi_R^{-1}(i)}^{(t)},\ \widetilde{v}_j^{(t)} = v_{\pi_C^{-1}(j)}^{(t)},$$

which we shall show using the principle of mathematical induction. The initial case $t = 0$ follows from the equivariance of the initial estimate. Note that,

$$\widetilde{e}_{ij}^{(t)} = (\boldsymbol{PXQ}^\intercal)_{ij} - \widetilde{\lambda}^{(t)}\widetilde{u}_i^{(t)}\widetilde{v}_j^{(t)} = (\boldsymbol{PXQ}^\intercal)_{ij} - \lambda^{(t)}u_{\pi_R^{-1}(i)}^{(t)}v_{\pi_C^{-1}(j)}^{(t)} = e_{\pi_R^{-1}(i),\pi_C^{-1}(j)}^{(t)},$$

by the induction hypothesis. Here we use the fact that the (i, j)-th entry of the permuted matrix $\boldsymbol{PXQ}^\intercal$ is same as the $(\pi_R^{-1}(i), \pi_C^{-1}(j))$-th entry of $\boldsymbol{X}$. Now,

$$\begin{aligned}
\widetilde{\lambda}^{(t+1)}\widetilde{u}_i^{(t+1)} &= \frac{\sum_j \widetilde{v}_j^{(t)} X_{\pi_R^{-1}(i),\pi_C^{-1}(j)}\psi(\widetilde{e}_{ij}^{(t)}/\widetilde{\sigma}^{(t)})}{\sum_j (\widetilde{v}_j^{(t)})^2\psi(\widetilde{e}_{ij}^{(t)}/\widetilde{\sigma}^{(t)})} \\
&= \frac{\sum_j v_{\pi_C^{-1}(j)}^{(t)} X_{\pi_R^{-1}(i),\pi_C^{-1}(j)}\psi(e_{\pi_R^{-1}(i),\pi_C^{-1}(j)}^{(t)}/\sigma^{(t)})}{\sum_j (v_{\pi_C^{-1}(j)}^{(t)})^2\psi(e_{\pi_R^{-1}(i),\pi_C^{-1}(j)}^{(t)}/\sigma^{(t)})}, \text{ by induction hypothesis} \\
&= \frac{\sum_j v_j^{(t)} X_{\pi_R^{-1}(i),j}\psi(e_{\pi_R^{-1}(i),j}^{(t)}/\sigma^{(t)})}{\sum_j (v_j^{(t)})^2\psi(e_{\pi_R^{-1}(i),j}^{(t)}/\sigma^{(t)})}, \text{ since } \pi_C \text{ is a permutation} \\
&= \lambda^{(t+1)}u_{\pi_R^{-1}(i)}^{(t+1)}.
\end{aligned}$$

This completes the inductive step for u-values. Performing the same steps now with Eq. (2.16) and (2.17) completes the proof. □

2.3.3 Consistency

The convergence result presented in Theorem 2.1 ensures that under some minimal assumptions, the rSVDdpd algorithm given by the iterations (2.15)-(2.17) converges to the rSVDdpd estimator, i.e., a local minimizer $\widehat{\boldsymbol{\theta}}^*$ of $H_\alpha^{(1)}(\boldsymbol{\theta})$ given in (2.4). However, in view of Definition 1 of a "best" fitting parameter $\boldsymbol{\theta}^g$, it is necessary to know whether such a local optimum remains close to a "best" fitting parameter in an asymptotic sense. In this asymptotic regime, we allow both the matrix dimensions n and p to grow to infinity, subject to a constant ratio in the limit, i.e., $n/p \to c$ for some $c \in (0, \infty)$.

Answering this question about the statistical consistency of the rSVDdpd estimator has two technical difficulties. Firstly, the parameter space $\boldsymbol{\Theta}$ is not necessarily convex due to the presence of the coordinates related to singular vectors. This problem can be circumvented using an inverse stereographic projection which transforms this non-convex parameter space $\boldsymbol{\Theta}$ into a convex parameter space $\boldsymbol{\Xi} \subseteq \mathbb{R}^{(n+p)}$. We call this parameter space $\boldsymbol{\Xi}$ as the natural parameter space in the given setup. The one-one transformation $\mathcal{T}$ between these two parameter spaces $\boldsymbol{\Theta}$ and $\boldsymbol{\Xi}$ are governed by the following two equations

$$\mathcal{T}(\lambda, \{u_i\}_{i=1}^n, \{v_j\}_{j=1}^p, \sigma^2) = \left(\lambda, \left\{\frac{u_i}{(1-u_n)}\right\}_{i=1}^{(n-1)}, \left\{\frac{v_j}{(1-v_p)}\right\}_{j=1}^{(p-1)}, \sigma^2\right), \tag{2.27}$$

and,

$$\begin{aligned}
&\mathcal{T}^{(-1)}\left(\lambda, \{\alpha_i\}_{i=1}^{(n-1)}, \{\beta_j\}_{j=1}^{(p-1)}, \sigma^2\right) \\
&\qquad = \left(\lambda, \left\{\frac{2\alpha_i}{U^2+1}\right\}_{i=1}^{(n-1)}, \frac{U^2-1}{U^2+1}, \left\{\frac{2\beta_j}{V^2+1}\right\}_{j=1}^{(p-1)}, \frac{V^2-1}{V^2+1}, \sigma^2\right),
\end{aligned} \tag{2.28}$$

where $U^2 = \sum_{i=1}^{(n-1)} \alpha_i^2$ and $V^2 = \sum_{j=1}^{(p-1)} \beta_j^2$. The above transformation is well defined for all points such that $u_n \neq 1$ and $v_p \neq 1$. Accordingly, we denote $\boldsymbol{\eta}$ as an element of this natural parameter space $\boldsymbol{\Xi}$, where the corresponding transformed parameter $\boldsymbol{\theta} = \mathcal{T}^{(-1)}(\boldsymbol{\eta})$ denotes an element of $\boldsymbol{\Theta}$.

The second problem is that the dimension of the singular vectors is not fixed and grows with the dimension of the matrix. Thus, as $n, p \to \infty$, the dimension of the parameter space, i.e., $(n + p + 2)$ grows linearly in n or p and increases to infinity. This means that although rSVDdpd uses the MDPDE proposed by Ghosh and Basu (2013) to solve the linear regression problems, their consistency results cannot be used directly as it assumes the dimension of the parameter space to be fixed. This varying dimension problem has been of considerable interest to many authors under the M-estimation setup (Huber, 1973, Portnoy, 1984). Most of these results assume convexity of the objective function (He and Shao, 2000), but that cannot be employed in our case. Here, the objective function is convex in any of the parameters individually when the other parameters are kept fixed, but becomes non-convex if all the parameters are taken together. To deal with this problem, we need to apply some concentration bounds on the error terms, for which we will restrict our attention to the scenario where the model density f is the standard normal density function. Note that, this choice does not put any restriction on the data matrix $\boldsymbol{X}$ or its distribution, rather it focuses on a specific class of rSVDdpd algorithm. As will be indicated later in Remark 2.6, it is possible to relax this assumption and consider f to be any symmetric subgaussian density in general.

Now that the necessary foundations are laid out, we can present the consistency theorem, which claims that under some reasonable assumptions indicated below, the minimizer $\boldsymbol{\theta}^*$ of $H_\alpha^{(1)}$ as given in Theorem 2.1, is a consistent estimator of the best fitting parameter $\boldsymbol{\theta}^g$. However, note that the description of a "best" fitting parameter indicated in Definition 1 is applicable for fixed n and p. In contrast, statistical consistency is an asymptotic property requiring the dimensions of the matrix n and p to grow towards infinity. To resolve this conflict in a unified setup, we assume that for varying n and p, a sequence of "best" fitting parameters exists. We denote the (i, j)-th entry of the data matrix $\boldsymbol{X}_{n,p}$ of order $n \times p$ by the random variable $(X_{ij})_{n,p}$, and indicate the assumptions about it as follows

(B1) There exists a sequence of the best fitting parameters $\{\boldsymbol{\theta}_{n,p}^g\}_{n,p=1}^{\infty}$ such that $\boldsymbol{\theta}_{n,p}^g = (\lambda^g, \{u_{i,n}^g\}_{i=1}^n, \{v_{j,p}^g\}_{j=1}^p, (\sigma_{n,p}^g)^2)$ and X_{ij}s are independently distributed as

$$X_{ij} = \lambda^g u_{i,n}^g v_{j,p}^g + \sigma_{n,p}^g Z_{ij},$$

for all $i = 1, 2, \ldots, n; j = 1, 2, \ldots, p$ and Z_{ij}s are i.i.d. following a common density function g.

(B2) The density function g is such that the integral $\int e^{-\alpha z^2/2} g(z) dz$ exists, is finite, and is three times differentiable and the derivatives can be taken under the integral sign. Also, the integrals $\int z^k e^{-\alpha z^2/2} g(z)$ are finite for $k = 0, 1, \ldots, 4$.

(B3) The density function g is symmetric.

(B4) For each pair of positive integers n and p, there exists an open rectangle $R_{n,p}$ inside $\boldsymbol{\Xi}_{n,p} \subset \mathbb{R}^{(n+p)}$ containing the natural parametrization $\boldsymbol{\eta}^g_{n,p}$ of $\boldsymbol{\theta}^g_{n,p}$ such that the sequence of sets $\mathcal{T}^{-1}(R_{n,p})$ does not contain a limit point with $\boldsymbol{a} = (0, \ldots, 0, 1)$ or $\boldsymbol{b} = (0, \ldots, 0, 1)$ for all sufficiently large n and p.

(B5) The converged rSVDdpd estimate for the data matrix $\boldsymbol{X}_{n,p}$, i.e, $\boldsymbol{\theta}^*_{n,p}$ (the minimizer of $H^{(1)}_{n,p}$ as indicated by Theorem 2.1) satisfy $(\boldsymbol{a}^*_n)^\intercal \neq (0, \ldots, 0, 1)$ and $(\boldsymbol{b}^*_p)^\intercal \neq (0, \ldots, 0, 1)$ for all sufficiently large n and p.

(B6) The variance $(\sigma^g_{n,p})^2$ in best fitting parameters satisfy $(\sigma^g_{n,p})^2 \asymp (np)^{-1/2}$, where $\asymp$ denotes the asymptotic equivalence as both $n, p \to \infty$.

The first Assumption (B1) is simply a description of the setup. Assumptions (B2) and (B4) are standard assumptions connected to the MDPDE (Ghosh and Basu, 2013). Assumptions (B3) and (B6) are required to provide concentration bound on the covariance terms and tail probabilities respectively. It is well known from random matrix theory that the Gaussian ensemble with each entry following a standard normal distribution has the singular values asymptotically at the order of $(\sqrt{n} + \sqrt{p})$ (Tracy and Widom, 1993, Mehta, 2004). However, since Assumption (B1) indicates that the same λ^g acts as the singular value for any n and p, the variance of the entries of the data matrix has to go down asymptotically to ensure that the singular value does not grow with increasing n or p.

Theorem 2.7. *Under Assumptions (B1)-(B6), if the model density f is the standard normal density, then there exists a sequence of rSVDdpd estimates $\boldsymbol{\theta}^*_{n,p}$ which is consistent for the sequence of "best" fitting parameters $\boldsymbol{\theta}^g_{n,p}$ given in Assumption (B1). That means, as both dimensions of the data matrix $\boldsymbol{X}_{n,p}$ (i.e., n and p) tend to infinity subject to a constant ratio in limit, i.e., $\lim_{\substack{n\to\infty \\ p\to\infty}} \frac{n}{p} = c$ for some $c \in (0, \infty)$,*

$$\left\|\boldsymbol{\theta}^*_{n,p} - \boldsymbol{\theta}^g_{n,p}\right\| \to 0,$$

in probability.

Proof. First, we observe that the stereographic transformation mentioned above can be employed and would remain valid because of Assumptions (B4) and (B5). Now, to prove the consistency, we shall take a route similar to the one taken by Ghosh and Basu (2013) as in the case of MDPDE for independent and non-homogeneous (inh) setup. Instead of showing that the rSVDdpd estimator, i.e., $\boldsymbol{\theta}^*_{n,p}$ is consistent for $\boldsymbol{\theta}^g_{n,p}$, we shall instead show that $\boldsymbol{\eta}^*_{n,p}$ is consistent for $\boldsymbol{\eta}^g_{n,p}$. Let us denote $\mathcal{H}_{n,p}(\boldsymbol{\eta})$ to indicate the H-function as in (2.4) evaluated at $\boldsymbol{\theta} = \mathcal{T}^{-1}(\boldsymbol{\eta})$, for fixed n and p with V_f substituted by V_ϕ given in (2.24). To prove that $\boldsymbol{\eta}^*_{n,p}$ is consistent for $\boldsymbol{\eta}^g_{n,p}$, we shall show that for any sufficiently small $r > 0$, $\mathcal{H}_{n,p}(\boldsymbol{\eta}_{n,p}) > \mathcal{H}_{n,p}(\boldsymbol{\eta}^g_{n,p})$ for sufficiently large n and p, for any $\boldsymbol{\eta}_{n,p}$ with $\|\boldsymbol{\eta}_{n,p} - \boldsymbol{\eta}^g_{n,p}\|_2 = r$. This means that the value of $\mathcal{H}_{n,p}$ at the surface of the ball of radius r centered at $\boldsymbol{\eta}^g_{n,p}$ would be higher than its value at $\boldsymbol{\eta}^g_{n,p}$, and hence by the smoothness of $\mathcal{H}_{n,p}$, it is ensured

that there will be a local minimum strictly inside that ball. Proceeding as in Ghosh and Basu (2013), we start with the Taylor series expansion of $\mathcal{H}_{n,p}(\boldsymbol{\eta}_{n,p})$ about $\boldsymbol{\eta}^g_{n,p}$, for any fixed n and p. For notational convenience, we suppress the subscripts n and p from $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^g$ which should be obvious from the context. We also use the symbol $\partial_{x_{i_1},\ldots,x_{i_k}}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}}$ to denote the k-th order partial derivative of $\mathcal{H}_{n,p}$ in the direction of the variables $x_{i_1},\ldots,x_{i_k}$ respectively, at the parameter value $\boldsymbol{\eta}$.

$$\begin{aligned}
\mathcal{H}_{n,p}(\boldsymbol{\eta}) - \mathcal{H}_{n,p}(\boldsymbol{\eta}^g) &= \partial_\lambda \mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}(\lambda-\lambda^g) + \sum_{i=1}^{(n-1)} \partial_{\alpha_i}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}(\alpha_i-\alpha_i^g) \\
&+ \sum_{j=1}^{(p-1)} \partial_{\beta_j}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}(\beta_j-\beta_j^g) + \partial_{\sigma^2}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}(\sigma^2-(\sigma^g)^2) \\
&+ \frac{1}{2}\sum_{k_1,k_2}\partial^2_{\eta_{k_1},\eta_{k_2}}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}(\eta_{k_1}-\eta^g_{k_1})(\eta_{k_2}-\eta^g_{k_2}) \\
&+ \frac{1}{6}\sum_{k_1,k_2,k_3}\partial^3_{\eta_{k_1},\eta_{k_2},\eta_{k_3}}|_{\boldsymbol{\eta}'}\mathcal{H}_{n,p}(\eta_{k_1}-\eta^g_{k_1})(\eta_{k_2}-\eta^g_{k_2})(\eta_{k_3}-\eta^g_{k_3}) \\
&= S_{1,1}+S_{1,2}+S_{1,3}+S_{1,4}+\frac{1}{2}S_2+\frac{1}{6}S_3,
\end{aligned}$$

where $\boldsymbol{\eta}'$ is a point on the line joining $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^g$, and the quantities $S_{1,1}, S_{1,2}, S_{1,3}, S_{1,4}, S_2$ and S_3 respectively denote the summands they are replacing. Here, η_k denotes the k-th coordinate of the vector $\boldsymbol{\eta}_{n,p}$. Also, α_i's and β_j's are the natural parametric representation of the elements of left $(u_{i,n})$ and right singular vectors $(v_{j,p})$ respectively, where the dimension subscripts (n and p) have been suppressed for notational convenience as indicated before.

Clearly, the smoothness of $\mathcal{H}_{n,p}$ along with Assumption (B1) on the normality of the errors, indicates that $\int \mathcal{H}_{n,p} g_{ij}(x)dx$ can be differentiated thrice with respect to the natural parameter $\boldsymbol{\eta}_{n,p}$, and the derivative can be taken under the integral sign. Hence, we have

$$\mathbb{E}\left[\partial_{\eta_k}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}\right] = \partial_{\eta_k}\mathbb{E}\left(\mathcal{H}_{n,p}\right)|_{\boldsymbol{\eta}^g} = 0, \tag{2.29}$$

since the population version of the objective function $\mathbb{E}\mathcal{H}_{n,p}$ is minimized at the true parameter $\boldsymbol{\eta}^g$. Thus, by a generalized version of Khinchin's Weak Law of Large numbers, it follows that as n and p both increase to infinity, each of the first-order partial derivatives goes in probability to 0. However, the problem arises as there are potentially infinitely many terms (as the parameter space increases in dimension). This jeopardizes any approach to naturally extending the proof of Theorem 3.1 of Ghosh and Basu (2013).

Before proceeding further, we note that since $\sum_{k=1}^n (u_k^g)^2 = 1$, its derivative yields $\sum_{k=1}^n u_k^g \partial_{\alpha_i} u_k|_{\boldsymbol{\eta}^g} = 0$ for any $i = 1,\ldots,(n-1)$. Similarly, $\sum_{l=1}^p v_l^g \partial_{\beta_j} v_l|_{\boldsymbol{\eta}^g} = 0$ for all $j = 1,\ldots,(p-1)$. Also, for notational convenience, we denote $w_{ij} = e^{-\alpha Z_{ij}^2/2}$.

Let us consider each of the sums $S_{1,1}, S_{1,2}, S_{1,3}$ and $S_{1,4}$ pertaining to the first order derivative separately. Since $\partial_\lambda \mathcal{H}_{n,p}$ and $\partial_{\sigma^2}\mathcal{H}_{n,p}$ both converges in probability to 0, hence for sufficiently large n and p, we have $|S_{1,1}| < r^3$ and $|S_{1,4}| < r^3$ with probability tending to 1. Now, to deal with an increasing number of summands in $S_{1,2}$ or $S_{1,3}$ we apply

Chebyshev's inequality after bounding its expectation and variance separately. By the chain rule of differentiation,

$$s_{n,p} := \sum_{i=1}^{(n-1)} \partial_{\alpha_i} \mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g} = \sum_{k=1}^{n} \partial_{u_k} \mathcal{H}_{n,p}|_{\boldsymbol{u}_k^g} \sum_{i=1}^{(n-1)} \partial_{\alpha_i} u_k|_{\boldsymbol{\eta}^g}, \tag{2.30}$$

where $\partial_{\alpha_i} u_k|_{\boldsymbol{\eta}^g}$ denotes the partial derivative of the entry of the left singular vector u_k with respect to the stereographic projection variables α_i at $\boldsymbol{\eta}^g$. As in the case of (2.29), one can verify that $\mathbb{E}\left[\partial_{u_k} \mathcal{H}_{n,p}|_{\boldsymbol{u}_k^g}\right] = 0$ for all $k = 1, 2, \ldots, n$, and, therefore, Eq. (2.30) implies that $\mathbb{E}(s_{n,p}) = 0$. Turning to its variance, it follows that

$$\begin{aligned}
\mathrm{Var}(s_{n,p}) &= \sum_{k=1}^{n} \left(\sum_{i=1}^{(n-1)} \partial_{\alpha_i} u_k|_{\boldsymbol{u}_k^g} \right)^2 \mathrm{Var}\left(\partial_{u_k} \mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}\right) \\
&= \sum_{k=1}^{n} \left(\sum_{i=1}^{(n-1)} \partial_{\alpha_i} u_k|_{\boldsymbol{u}_k^g} \right)^2 \frac{(\alpha+1)^2 (\lambda^g)^2}{(2\pi)^\alpha \sigma^{2(\alpha+1)} n^2 p^2} B_1,
\end{aligned}$$

where $B_1 = \mathbb{E}(Z_{ij}^2 w_{ij}^2)$. Here we use the fact that for $k \neq l$,

$$\mathrm{Cov}\left(\partial_{u_k} \mathcal{H}_{n,p}|_{\boldsymbol{u}_k^g}, \partial_{u_l} \mathcal{H}_{n,p}|_{\boldsymbol{u}_l^g}\right) = 0,$$

which follows by noting that the part of $\mathcal{H}_{n,p}$ dependent on u_k would consist of only the k-th row of the data matrix $\boldsymbol{X}$ and similarly $\partial_{u_l} \mathcal{H}_{n,p}$ will depend only l-th row of $\boldsymbol{X}$, which are assumed to be independently distributed in the current setup. Also, an application of Cauchy-Schwartz inequality yields that

$$S_u := \sum_{i=1}^{(n-1)} u_i^g \leq \left| \sum_{i=1}^{(n-1)} u_i^g \right| \leq \sqrt{n-1} \left(\sum_{i=1}^{(n-1)} (u_i^g)^2 \right) \leq \sqrt{n-1}.$$

Therefore, as $u_n^g \in (-1, 1)$, the sum

$$\begin{aligned}
\sum_{k=1}^{n} \left(\sum_{i=1}^{(n-1)} \partial_{\alpha_i} u_k|_{\boldsymbol{\eta}^g} \right)^2 &= (1-u_n^g)^2 S_u^2 + \sum_{k=1}^{n} \left((1-u_n^g) - u_k^g S_u \right)^2 \\
&= (1-u_n^g)^2 S_u^2 + n(1-u_n^g)^2 - S_u^2 + 2S_u^2 u_n^g - 2S_u u_n^g (1-u_n^g), \\
&\leq 4S_u^2 + 4n - S_u^2 + 2S_u^2 + 4S_u \\
&\leq 4n + 4n + n + 2\sqrt{n-1},
\end{aligned}$$

which is further bounded by $11n$ in magnitude. Therefore, for sufficiently large n and p,

$$\mathrm{Var}(s_{n,p}) = \mathcal{O}\left((\sigma^g)^{-(2\alpha+2)} / np^2 \right).$$

Since we have $\sigma^g \asymp (np)^{-1/4}$ and $\alpha \leq 1$, it follows that $\mathrm{Var}(s_{n,p}) \to 0$ as n and p tends to infinity. Therefore, we have $|\sum_{i=1}^{(n-1)} \partial_{\alpha_i} \mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}| \to 0$, with probability tending to one. Along with $|\alpha_i - \alpha_i^g| < r$, we have $|S_{1,2}| < r^3$, for sufficiently large n and p, with probability tending to 1.

Reversing the role of n and p, and considering $\sum_{j=1}^{(p-1)} \partial_{\beta_j} \mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}$ instead, one can show that $|S_{1,3}| < r^3$ for sufficiently large n and p with probability tending to 1. Thus, combining everything so far, we obtain

$$|S_1| \leq |S_{1,1}| + |S_{1,2}| + |S_{1,3}| + |S_{1,4}| < 4r^3,$$

for sufficiently large n and p, with probability tending to 1.

Now, turning our attention to the second order term S_2, we start by writing the expressions for each second-order derivative term. Let, $C_\alpha = -(\alpha+1)(2\pi)^{-\alpha/2}(\sigma^g)^{-(\alpha+2)}/np$, then

$$\mathbb{E}\left[\partial^2_\lambda \mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}\right] = C_\alpha \sum_{i=1}^{n}\sum_{j=1}^{p} (u_i^g)^2 (v_j^g)^2 \mathbb{E}\left[w_{ij}(\alpha Z_{ij}^2 - 1)\right] = C_\alpha B_2$$

where, $B_2 = \mathbb{E}\left[w_{ij}(\alpha Z_{ij}^2 - 1)\right]$. For the mixed derivative,

$$\begin{aligned}
\mathbb{E}\left[\partial^2_{\lambda,\alpha_i} \mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}\right] &= \sum_{k=1}^{n} \mathbb{E}\left[\partial^2_{\lambda,u_k} \mathcal{H}_{n,p}|_{\boldsymbol{u}_k^g}\right] \partial_{\alpha_i} u_k|_{\boldsymbol{\eta}^g} \\
&= \sum_{k=1}^{n} C_\alpha \lambda^g \sum_{j=1}^{p} v_j^g \mathbb{E}\left[w_{kj}(\sigma^g Z_{kj} + \lambda^g u_k^g v_j^g (\alpha Z_{kj}^2 - 1))\right] \partial_{\alpha_i} u_k \\
&= \quad 0
\end{aligned}$$

since, $\mathbb{E}(Z_{ij} w_{ij}) = 0$ by symmetry of g and we know $\sum_{i=1}^{n} u_k^g \partial_{\alpha_i} u_k|_{\boldsymbol{\eta}^g} = 0$. Exchanging the role of u_i^gs and v_j^gs, we obtain

$$\mathbb{E}\left[\partial^2_{\lambda,\beta_j} \mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}\right] = 0.$$

A chain rule of differentiation can be used to obtain the second order derivatives of $\mathcal{H}_{n,p}$ with respect to α_i's as

$$\begin{aligned}
\mathbb{E}\left[\partial^2_{\alpha_i,\alpha_j} \mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}\right] &= \mathbb{E}\left[\sum_{k=1}^{n} \partial_{u_k} \mathcal{H}_{n,p}|_{\boldsymbol{u}_k^g} \partial^2_{\alpha_i,\alpha_j} u_k|_{\boldsymbol{\eta}^g} + \sum_{k=1}^{n}\sum_{l=1}^{n} \partial^2_{u_k,u_l} \mathcal{H}_{n,p}|_{\boldsymbol{u}_k^g,\boldsymbol{u}_l^g} \partial_{\alpha_i} u_k|_{\boldsymbol{\eta}^g} \partial_{\alpha_j} u_l|_{\boldsymbol{\eta}^g}\right], \\
&= \sum_{k=1}^{n} \mathbb{E}\left[\partial^2_{u_k} \mathcal{H}_{n,p}|_{\boldsymbol{u}_k^g}\right] \partial_{\alpha_i} u_k|_{\boldsymbol{\eta}^g} \partial_{\alpha_j} u_k|_{\boldsymbol{\eta}^g},
\end{aligned}$$

since, $\mathbb{E}\left[\partial_{u_k} \mathcal{H}_{n,p}|_{\boldsymbol{u}_k^g}\right] = 0$ and for $k \neq l$, $\mathbb{E}\left[\partial^2_{u_k,u_l} \mathcal{H}_{n,p}|_{\boldsymbol{u}_k^g,\boldsymbol{u}_l^g}\right] = 0$. A similar calculation as above reveals that

$$\mathbb{E}(\partial^2_{u_k} \mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}) = C_\alpha (\lambda^g)^2 B_2.$$

Combining this with the fact that

$$\sum_{k=1}^{n} \partial_{\alpha_i} u_k|_{\boldsymbol{\eta}^g} \partial_{\alpha_j} u_k|_{\boldsymbol{\eta}^g} = \begin{cases} (1-u_n^g)^2 & \text{if, } i = j \\ 0 & \text{if, } i \neq j \end{cases},$$

yields

$$\mathbb{E}\left[\partial^2_{\alpha_i,\alpha_j} \mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}\right] = \begin{cases} C_\alpha (\lambda^g)^2 B_2 (1-u_n^g)^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

Exact same calculation also holds for $\mathbb{E}\left(\partial^2_{\beta_i,\beta_j}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}\right)$ with u_n^g replaced by v_p^g. Because of Assumption (B4), $(1-u_n^g)^2$ and $(1-v_p^g)^2$ can be bounded below by some $\delta > 0$ independent of n and p. Also, note that

$$\begin{aligned}\mathbb{E}\left[\partial^2_{\alpha_i,\beta_j}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}\right] &= \sum_{k=1}^{n}\sum_{l=1}^{p}\partial_{\alpha_i}u_k|_{\boldsymbol{\eta}^g}\partial_{\beta_j}v_l|_{\boldsymbol{\eta}^g}\mathbb{E}\left[\partial^2_{u_k,v_l}\mathcal{H}_{n,p}|_{\boldsymbol{u}_k^g,\boldsymbol{v}_l^g}\right]\\ &= \sum_{k=1}^{n}\sum_{l=1}^{p}\partial_{\alpha_i}u_k|_{\boldsymbol{\eta}^g}\partial_{\beta_j}v_l|_{\boldsymbol{\eta}^g}C_\alpha\lambda^g\mathbb{E}\left[\sigma^g w_{kl}Z_{kl} + \lambda^g u_k^g v_l^g(\alpha Z_{kl}^2 - 1)\right]\\ &= 0\end{aligned}$$

which follows from noting that $\mathbb{E}(z_{kl}w_{kl}^2) = 0$ by symmetry of the density function g and the fact that $\sum_k u_k^g\partial_{\alpha_i}u_k|_{\boldsymbol{u}_k^g} = \sum_l v_l^g\partial_{\beta_j}v_l|_{\boldsymbol{v}_l^g} = 0$. Furthermore, as shown in Ghosh and Basu (2013), the scale and the location estimator become asymptotically uncorrelated for the classical linear regression setup with normally distributed errors. Therefore, we have

$$\mathbb{E}\left[\partial^2_{\alpha_i,\sigma^2}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}\right] = \mathbb{E}\left[\partial^2_{\beta_j,\sigma^2}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}\right] = \mathbb{E}\left[\partial^2_{\lambda,\sigma^2}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}\right] = 0,$$

and

$$\mathbb{E}\left[\partial^2_{\sigma^2}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}\right] = (2\pi)^{-\alpha/2}(\sigma^g)^{-(\alpha+4)}\left[\frac{\alpha(\alpha+2)}{4\sqrt{1+\alpha}} - \frac{(\alpha+1)}{2}B_3\right] \asymp \sigma^{-(\alpha+4)},$$

where $B_3 = \mathbb{E}\left(w_{ij}(1-2Z_{ij}^2+\alpha(1-Z_{ij}^2)^2/2)\right)$. Therefore, if we consider the $(n+p)\times(n+p)$ matrix $\boldsymbol{\Psi}_{n,p}$ whose (k_1,k_2)-th element is given by $\mathbb{E}(\partial^2_{\eta_{k_1},\eta_{k_2}}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g})$, then $\boldsymbol{\Psi}_{n,p}$ turns out to be a diagonal matrix with nonzero entries of the order of $(\sigma^g)^{-(\alpha+2)}/np$ and $\sigma^{-(\alpha+4)}$, among which the minimum is at the order of $(\sigma^g)^{-(\alpha+2)}/np$ due to Assumption (B6). Hence, the minimum eigenvalue of $\boldsymbol{\Psi}_{n,p}$ is bounded below by $K_1(\sigma^g)^{-(\alpha+2)}/np$ for some positive finite constant K_1.

Now, we decompose S_2 by considering elements of $\boldsymbol{\Psi}_{n,p}$ as follows

$$\begin{aligned}&\sum_{k_1,k_2}\partial^2_{\eta_{k_1},\eta_{k_2}}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g}(\eta_{k_1}-\eta_{k_1}^g)(\eta_{k_2}-\eta_{k_2}^g)\\ &\qquad= \sum_{k_1,k_2}\left[\partial^2_{\eta_{k_1},\eta_{k_2}}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}^g} - (\boldsymbol{\Psi}_{n,p})_{k_1,k_2}\right](\eta_{k_1}-\eta_{k_1}^g)(\eta_{k_2}-\eta_{k_2}^g)\\ &\qquad\qquad + \sum_{k_1,k_2}(\boldsymbol{\Psi}_{n,p})_{k_1,k_2}(\eta_{k_1}-\eta_{k_1}^g)(\eta_{k_2}-\eta_{k_2}^g).\end{aligned}$$

Here, we can apply an orthogonal transformation on $(\boldsymbol{\eta}-\boldsymbol{\eta}^g)$ to express it as a linear combination of the eigenvectors of $\Psi_{n,p}$, so that the second term can be made greater than or equal to $K_1(\sigma^g)^{-(\alpha+2)}r^2/np$. Also, it is evident that the first summation has the expected value equal to 0. By a similar routine calculation as above, one can show that the variance of the first term also goes to 0. Therefore, for sufficiently large n and p, with probability tending to 1, $S_2 > (-r^3 + K_1(\sigma^g)^{-(\alpha+2)}r^2/np)$.

Finally, turning to S_3, we note that the expected values of the third-order derivatives

are bounded as shown below.

$$\begin{aligned}
\left|\mathbb{E}\left[\partial_{\lambda}^{3}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}'}\right]\right| &= M_1\left|\frac{\sigma^{-(\alpha+3)}}{np}\sum_{i,j}(u_i v_j)^3\right|,\\
\left|\mathbb{E}\left[\partial_{a_i}^{3}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}'}\right]\right| &= M_2\left|\frac{\sigma^{-(\alpha+3)}}{np}\sum_{j}(\lambda v_j)^3\right|,\\
\left|\mathbb{E}\left[\partial_{b_j}^{3}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}'}\right]\right| &= M_3\left|\frac{\sigma^{-(\alpha+3)}}{np}\sum_{i}(\lambda u_i)^3\right|,\\
\left|\mathbb{E}\left[\partial_{\sigma^2}^{3}\mathcal{H}_{n,p}|_{\boldsymbol{\eta}'}\right]\right| &= M_4\sigma^{-(\alpha+6)},
\end{aligned}$$

where M_1, M_2, M_3 and M_4 are positive finite constants. The first three among these are $\mathcal{O}(\sigma^{-(\alpha+2)}/np)$ and the last one is $\mathcal{O}(\sigma^{-(\alpha+6)})$, which follows from Cauchy-Schwartz inequality and the normalization of u_is and v_js. Combining these bounds along with the continuity of the third order derivative of $\mathcal{H}_{n,p}$ and Assumption (B6), we obtain that $|S_3| \leq M\sigma^{-(\alpha+6)}$ for sufficiently large n and p, and for some sufficiently large finite positive constant M independent of n and p. Therefore, using the bounds for the individual terms of the Taylor's series, we have

$$\mathcal{H}_{n,p}(\boldsymbol{\eta}) - \mathcal{H}_{n,p}(\boldsymbol{\eta}^g) > \left(-5r^3 + \frac{K_1}{np}(\sigma_{n,p}^g)^{-(\alpha+2)}r^2 - M(\sigma_{n,p}^g)^{-(\alpha+6)}r^3\right), \tag{2.31}$$

with probability tending to 1 for sufficiently large n and p. Now since $(\sigma^g)^4 \asymp (np)^{-1}$ due to Assumption (B6) and as $\sigma^g \to 0$ as n and p tends to infinity, it follows that

$$\lim_{n,p\to\infty}\frac{K_1(\sigma_{n,p}^g)^{-(\alpha+2)}/np}{5 + M(\sigma_{n,p}^g)^{-(\alpha+6)}} = \frac{K_1(\sigma_{n,p}^g)^{-(\alpha+6)}}{5 + M(\sigma_{n,p}^g)^{-(\alpha+6)}} = K_2 \in (0,\infty).$$

Choosing $r < K_2$ ensures that the lower bound in (2.31) remains positive, i.e., $\mathcal{H}_{n,p}(\boldsymbol{\eta}) > \mathcal{H}_{n,p}(\boldsymbol{\eta}^g)$ for any $\boldsymbol{\eta}$ satisfying $\|\boldsymbol{\eta} - \boldsymbol{\eta}^g\|_2 = r$. This is exactly what we intended to show at the beginning.

Finally, since $\boldsymbol{\eta}^*$ is consistent for $\boldsymbol{\eta}^g$, an application of the continuous mapping theorem completes the proof. □

Remark 2.5. *Theorems 2.5 and 2.6 are, respectively, the empirical counterparts of Theorems 2.3 and 2.4. In view of Theorems 2.5 and 2.6, the equivariance properties hold given that the initial values of the iterations of rSVDdpd algorithm also satisfy such equivariance properties. However, from the convergence and the consistency of the rSVDdpd estimator, it follows that for sufficiently large n and p, the converged estimator can be made arbitrarily close to the true "best" fitting parameters with probability close to 1. Since these "best" fitting parameters obey the equivariance properties as assured by Theorems 2.3 and 2.4, it follows that for sufficiently large n and p, the converged estimator will also approximately satisfy these equivariance properties, irrespective of the equivariance of initial values.*

Remark 2.6. *Instead of restricting the rSVDdpd estimator to consider only the normal family of model densities for the consistency result in Theorem 2.7, it is possible to choose f*

to be any symmetric subgaussian density. In this case, Assumption (B2) needs to be appropriately modified to ensure that the corresponding $\psi(\cdot)$ function defined in (2.10) is two-times continuously differentiable and the quantities $\int z^k \psi(z) g(z) dz$, $\int z^2 (\psi'(z))^2 g(z) dz$ and $\int z^2 \psi''(z) g(z)$, are all uniformly bounded.

Also note that Theorem 2.7 ensures the consistency of the rSVDdpd under a general setup where the errors may follow any arbitrary density function g subject to the Assumptions (B1)-(B6). Therefore, it also allows density functions of the form $g = (1-\epsilon)g_1 + \epsilon g_2$, which is a contaminated version of density g_1 contaminated by density g_2, provided that both g_1 and g_2 are symmetric density functions. Additionally, to ensure that Assumption (B2) is satisfied, a sufficient condition is that the density function g is thrice continuously differentiable and the random variable Z with density g has finite fourth-order moments. However, even if the moment condition does not hold, one can directly show Assumption (B2) for all $\alpha > 0$ in certain setups. For instance, in the case of Cauchy density,

$$\mathbb{E}(Z^4 e^{-\alpha Z^2/2}) = \int_{-\infty}^{\infty} \frac{x^4 e^{-\alpha x^2/2}}{\pi(1+x^2)} dx = \frac{2}{\sqrt{\pi}} e^{\alpha/2} \int_{\sqrt{\alpha}/\sqrt{2}}^{\infty} e^{-t^2} dt - \frac{\alpha - 1}{\sqrt{2\pi}\alpha^{3/2}},$$

which is finite for all $\alpha > 0$. Therefore, except for $\alpha = 0$ (i.e., in the case of MLE), the consistency of the rSVDdpd estimator is ensured by Theorem 2.7 even when the errors follow a Cauchy distribution with heavy tails.

2.4 Simulation Studies

In this section, we compare the performance of the rSVDdpd estimator with existing robust SVD estimators. As mentioned in Section 2.1, the amount of literature concerning the robust SVD is limited, and the publicly available implementations of these existing algorithms are scarce. Here, we consider two existing robust SVD estimators for comparison, namely the ones proposed by Hawkins et al. (2001) and Zhang et al. (2013). Implementation of the robust SVD algorithm proposed by Hawkins et al. (2001) (to be referred to here as **pcaSVD**) is available as an R package `pcaMethods` (Stacklies et al., 2007), which outputs all singular values and vectors of the input data matrix. The second algorithm by Zhang et al. (2013) obtains the first pair of singular vectors based on the minimization procedure

$$(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}) = \underset{(\boldsymbol{u},\boldsymbol{v})}{\arg\min} \left[\rho\left(\frac{\boldsymbol{X} - \boldsymbol{u}\boldsymbol{v}^\intercal}{\sigma} \right) + \mathcal{P}_\lambda(\boldsymbol{u}, \boldsymbol{v}) \right],$$

where $\rho(\cdot)$ is the Huber's loss function and $\mathcal{P}_\lambda$ is a regularization penalty term to motivate smoothing in the entries of the singular vectors. For an extensive comparison, we consider two variants of this algorithm. In one variation, we perform the minimization with only Huber's loss function without any penalty term, which we shall refer to as **RobSVD**, while in the other variation, we follow the recommended procedure of minimization with

a penalty term, which we shall call **RobRSVD**. Implementations of both of these variants are available in the R package `RobRSVD` (Zhang and Pan, 2013). To compare these algorithms on a level ground, we also add a wrapper on these two variants which outputs the subsequent singular values by sequentially applying the same method on the residual matrix. Additionally, we consider two randomized SVD methods: **rSVD** by Halko et al. (2011) and **pcaOne** by Li et al. (2023); the implementations of these algorithms are available as R packages under the same names (Erichson et al., 2019, Li, 2022). Finally, we have made available the implementation of the proposed rSVDdpd estimator in the R package `rsvddpd` (Roy, 2021) which we have used for the entirety of the simulation studies.

2.4.1 Simulation Setups

To compare the performance of different robust SVD approaches, we consider $B = 1000$ Monte Carlo replications of signal-plus-noise matrices as in (2.1). The signal, or the low rank component $\boldsymbol{L}$ of the data matrix $\boldsymbol{X}$ was generated as

$$\boldsymbol{L} = \boldsymbol{U} \begin{bmatrix} 10 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 3 \end{bmatrix} \boldsymbol{V}^{\intercal},$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices with the first three columns as the coefficients of the first three orthogonal polynomial contrasts of order 10 and 4. The final data matrix is then generated as $\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{E}$ where the entries e_{ij} of the error matrix $\boldsymbol{E}$ follow a pre-specified distribution based on one of the options as described below.

(SVD1) The errors follow the standard normal distribution (no outliers per se).

(SVD2) The errors are distributed according to a contaminated standard normal distribution namely

$$e_{ij} \sim (1 - \epsilon)\mathcal{N}(0, 1) + \epsilon\delta_{25},$$

where ϵ is the amount of contamination and δ_{25} denotes the degenerate distribution at 25. Based on the amount of contamination, we consider three subcases of this simulation setting.

- (SVD2a) $\epsilon = 0.05$, denotes only 5% contamination, corresponding to a relatively light amount of outlying observations.
- (SVD2b) $\epsilon = 0.1$, denoting medium level contamination with the presence of approximately 10% outlying values.
- (SVD2c) $\epsilon = 0.2$, denoting heavy contamination with approximately 20% outlying observations.

(SVD3) The errors are distributed according to a standard normal distribution with 2×2 block-based contamination as presented in Zhang et al. (2013).

(SVD4) The errors are distributed according to a standard Cauchy distribution. This setup helps us to study the effect of heavy-tailed errors in robust SVD.

(SVD5) The errors are distributed according to a standard log-normal distribution, which is used to study the effect of an asymmetric error distribution.

To measure the performance of a robust SVD algorithm, we track the estimates for each Monte Carlo sample and compute the mean squared error (MSE) and bias for each singular value, as well as the sum of squared biases and the sum of MSE across all singular values. For comparing the left and right singular vectors, we consider a dissimilarity score between two normalized vectors, denoted as $\text{Diss}(\boldsymbol{u}, \boldsymbol{v})$, which is equal to $1-|\langle \boldsymbol{u}, \boldsymbol{v}\rangle|$ where $\langle \cdot, \cdot \rangle$ denote the usual Euclidean inner product. This is connected to the usual metric for measuring similarly between two orthogonal matrices by considering the trace of the cross-product of their projection matrices. In particular,

$$\text{Trace}\left(P_{\boldsymbol{u}}^{\intercal} P_{\boldsymbol{v}}\right) = \text{Trace}\left((\boldsymbol{u}\boldsymbol{u}^{\intercal})^{\intercal}(\boldsymbol{v}\boldsymbol{v}^{\intercal})\right) = \text{Trace}\left((\boldsymbol{u}^{\intercal}\boldsymbol{v})(\boldsymbol{v}^{\intercal}\boldsymbol{u})\right) = |\langle \boldsymbol{u}, \boldsymbol{v}\rangle|^2,$$

where $P_{\boldsymbol{u}}$ and $P_{\boldsymbol{v}}$ are the projection matrices corresponding to the column space of $\boldsymbol{u}$ and $\boldsymbol{v}$ respectively. Further details of this equivalence are illustrated in Yang (2014). The dissimilarity score is 0 if $\boldsymbol{u} = \boldsymbol{v}$ or $\boldsymbol{u} = (-\boldsymbol{v})$, and it is equal to 1 if $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthogonal. The average dissimilarity score between the estimated singular vectors and the true singular vectors, computed over all Monte Carlo samples, is used as a performance measure in our simulation studies.

2.4.2 Simulation Results

Table 2.1 depicts the comparative results of performance measures for the usual SVD method, the three existing robust SVD algorithms, i.e, pcaSVD (Stacklies et al., 2007) and two variants of RobSVD (Zhang and Pan, 2013), two randomized SVD algorithms, i.e., rSVD (Halko et al., 2011) and pcaOne (Li et al., 2023), and the proposed rSVDdpd method (Roy, 2021) for different choices of robustness parameter α, under different simulation setups (SVD1)-(SVD5). Our findings are summarized in the following key points.

1. Classical SVD leads to a biased estimator of the singular values for the Gaussian error case, which is also supported by well-established theory (Rudelson and Vershynin, 2010).

2. The randomized SVD algorithms are only fast alternatives to the traditional SVD methods, they do not provide any robustness benefit, hence they perform poorly in the presence of contamination.

3. As the robustness parameter α increases, the proposed rSVDdpd leads to a less biased estimator. This is more prevalent in scenarios with some contamination present. On the other hand, in the absence of contamination, increasing α results in a higher variance of the estimated singular values.

Table 2.1: Summary of performance measures (total squared bias, total MSE of singular values, total dissimilarity of left and right singular vectors, denoted by Left and Right respectively) and the average runtime (in milliseconds) for the existing SVD and robust SVD algorithms.

Method	Metric	(SVD1)	(SVD2a)	(SVD2b)	(SVD2c)	(SVD3)	(SVD4)	(SVD5)
Usual SVD	Bias	7.96	519.35	1397.36	2434.02	1677.95	41825.79	93.63
	MSE	10.46	729.83	1661.82	2712.69	1686.36	2.17×10^6	146.19
	Left (Right)	0.70 (0.42)	1.93 (1.53)	2.15 (1.70)	2.21 (1.73)	2.05 (1.92)	2.09 (1.61)	2.02 (1.78)
	Time	0.035	0.034	0.034	0.033	0.034	0.034	0.032
rSVD (Halko et al., 2011)	Bias	7.95	519.36	1397.36	2434.05	1677.95	41825.79	93.62
	MSE	10.44	729.83	1661.82	2712.70	1686.37	2.17×10^6	146.19
	Left (Right)	0.70 (0.42)	1.93 (1.53)	2.15 (1.70)	2.21 (1.73)	2.05 (1.92)	2.09 (1.61)	2.02 (1.78)
	Time	0.312	0.311	0.309	0.294	0.312	0.316	0.299
pcaOne (Li et al., 2023)	Bias	7.95	519.36	1397.36	2434.05	1677.96	41825.78	93.62
	MSE	10.44	729.83	1661.82	2712.71	1686.37	2.17×10^6	146.20
	Left (Right)	0.70 (0.42)	1.93 (1.53)	2.15 (1.70)	2.21 (1.73)	2.05 (1.92)	2.09 (1.61)	2.02 (1.78)
	Time	0.075	0.075	0.074	0.072	0.073	0.076	0.071
pcaSVD (Hawkins et al., 2001)	Bias	15.07	350.4	1086.34	2110.99	1640.71	14224.14	77.5
	MSE	24.24	793.61	1706.46	2782.58	1654.5	2.16×10^6	108.51
	Left (Right)	1.20 (0.97)	1.68 (1.29)	1.96 (1.51)	2.11 (1.63)	2.00 (1.84)	1.99 (1.49)	2.03 (1.81)
	Time	3.909	3.450	3.372	2.694	3.375	3.728	3.052
RobSVD (Zhang et al., 2013)	Bias	8.81	523.86	1427.87	2488.98	1679.88	41779.81	94.9
	MSE	11.61	748.04	1722.44	2803.6	1688.71	2.17×10^6	149.13
	Left (Right)	0.82 (0.52)	1.93 (1.52)	2.15 (1.66)	2.20 (1.69)	1.94 (1.83)	2.11 (1.59)	1.97 (1.73)
	Time	87.218	60.789	42.005	40.391	44.165	42.131	46.110
RobRSVD (Zhang et al., 2013)	Bias	1.68	17.67	60.11	193.66	1090.28	540.8	29.14
	MSE	6.86	114.8	278.58	617.09	1248.28	2.91×10^4	49.05
	Left (Right)	0.73 (0.51)	1.32 (1.07)	1.67 (1.37)	1.93 (1.61)	2.16 (2.17)	1.71 (1.31)	1.97 (1.82)
	Time	300.784	202.921	200.545	199.526	396.132	300.776	297.976
rSVDdpd (ours) ($\alpha = 0.1$)	Bias	7.95	294.05	1039.98	2094.11	1355.18	469.52	69.98
	MSE	10.46	622.82	1529.37	2587.79	1634.41	1629.48	85.1
	Left (Right)	0.70 (0.42)	1.61 (1.16)	2.01 (1.49)	2.17 (1.65)	2.01 (1.84)	1.97 (1.50)	1.97 (1.75)
	Time	0.744	0.599	0.397	0.372	0.413	0.641	0.549
rSVDdpd (ours) ($\alpha = 0.3$)	Bias	7.94	18.14	130.47	662.6	114.71	265.14	67.09
	MSE	10.47	91.89	494.34	1453.4	458.47	1197.88	83.27
	Left (Right)	0.71 (0.42)	0.94 (0.62)	1.28 (0.89)	1.67 (1.23)	1.21 (0.91)	1.88 (1.40)	1.95 (1.74)
	Time	0.862	0.793	0.689	0.389	0.659	0.698	0.759
rSVDdpd (ours) ($\alpha = 0.5$)	Bias	7.92	10.88	44.15	298.01	28.13	198.81	62.43
	MSE	10.55	44.02	237.68	938.7	156.99	859.04	78.83
	Left (Right)	0.72 (0.43)	0.9 (0.58)	1.11 (0.76)	1.44 (1.04)	1.02 (0.69)	1.84 (1.37)	1.94 (1.73)
	Time	1.017	0.865	0.846	0.407	0.671	0.878	0.872
rSVDdpd (ours) ($\alpha = 0.7$)	Bias	7.89	9.43	30.04	196.51	21.59	169.08	56.64
	MSE	10.68	32.33	167.24	739.15	113.33	807.87	72.46
	Left (Right)	0.73 (0.45)	0.88 (0.57)	1.08 (0.73)	1.36 (0.96)	1.01 (0.68)	1.81 (1.36)	1.94 (1.73)
	Time	1.119	1.014	0.858	0.458	0.774	1.365	0.955
rSVDdpd (ours) ($\alpha = 1$)	Bias	7.76	8.78	23.26	141.06	18.05	140.65	49.42
	MSE	10.84	27.14	132.89	612.16	96.68	842.77	65.68
	Left (Right)	0.77 (0.49)	0.90 (0.59)	1.08 (0.73)	1.31 (0.93)	1.01 (0.67)	1.79 (1.32)	1.93 (1.73)
	Time	1.289	1.122	0.870	0.488	0.815	1.794	0.996

4. When contamination is present, as in setups (SVD2a)-(SVD2c), the classical SVD, pcaSVD and RobSVD do not yield reliable estimates. Although the regularized RobRSVD provides reasonable estimates, our proposed method rSVDdpd produces better SVD estimates with a choice of α close to 1.

5. Except for simulation setups (SVD1) and (SVD5), for all other setups, rSVDdpd turns out to be the best among the algorithms in comparison.

   (a) In simulation setup (SVD1) when there is no outlying observation present, the proposed rSVDdpd does not add any additional smoothness requirements in the entries of the singular vectors. Thus the estimator is expected to perform similar or slightly worse than the classical SVD due to increased variance in estimation.

   (b) In setup (SVD5) with an asymmetric error distribution, the proposed rSVDdpd estimator is second best and is competitive with the RobRSVD algorithm, which achieves the least bias and MSE at the cost of higher variance and dissimilarity scores. Most of the contribution to the MSE for rSVDdpd algorithm comes from the high bias, which can be attributed to the fact that the Assumption (B3) of the consistency theorem is not satisfied in this case.

Although RobRSVD closely outperforms the proposed rSVDdpd in some simulation scenarios, it does so at the cost of extremely high computational complexity. This is precisely due to the matrix inversion step to compute $(\mathcal{V}^\intercal \mathcal{W}^* \mathcal{V} + 2\Omega_{\boldsymbol{u}|\boldsymbol{v}})^{-1}$ (see Eq. (9) of Zhang et al. (2013)). Currently, the fastest matrix inversion algorithm, i.e., a variant of Coppersmith-Winograd algorithm (Alman and Williams, 2021) achieves a computational complexity of $\mathcal{O}(n^{2.3728596})$ for inverting an $n \times n$ matrix. It is also known that such variants of Strassen's algorithm or Coppersmith-Winograd algorithm (the two fastest matrix multiplication techniques known till today) cannot achieve a computational complexity lower than $\mathcal{O}(n^{2.3078})$ (Ambainis et al., 2015), which is again super-quadratic in matrix dimension n. Therefore, each iteration of the RobRSVD algorithm has a time complexity of at least $\mathcal{O}(\max\{n,p\}^{2.3078})$. On the other hand, each iteration of rSVDdpd performs only a weighted average computation as indicated in (2.15)-(2.17), which requires a computational complexity of $\mathcal{O}(\max\{n,p\}^2)$, if run sequentially. Also, since the update for each row can be performed independently, one can boost the performance of rSVDdpd using parallel computation so that every iteration can be completed even faster.

To demonstrate this, we consider $n \times p$ matrices with uniformly distributed entries for different choices of n and fixed $p = 25$, and apply different methods of computing SVD on them. Table 2.2 summarizes the time taken (in units of milliseconds) to obtain the first singular value using a computer with a 2.30GHz processor and 8 GB of RAM. As seen from Table 2.2, the computational budget of rSVDdpd is similar to pcaSVD, which is lower by several orders of magnitude than RobSVD and RobRSVD. This extremely high computational cost of RobRSVD can be avoided if the matrix $(\mathcal{V}^\intercal \mathcal{W}^* \mathcal{V} + 2\Omega_{\boldsymbol{u}|\boldsymbol{v}})$ becomes a diagonal matrix, which happens if the penalty parameter is zero and RobRSVD is

Table 2.2: Summary of average time taken (in milliseconds) to get the first singular value and vectors of an $n \times 25$ matrix with uniformly distributed entries via different SVD algorithms.

Number of	Existing methods for computing SVD				Choice of α in rSVDdpd		
rows (n)	SVD	pcaSVD	RobSVD	RobRSVD	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$
5	0.014	2.209	3.098	73.666	0.545	0.841	1.417
10	0.011	4.388	3.966	99.077	0.975	1.508	2.100
25	0.008	4.965	3.989	81.045	3.861	6.416	11.394
50	0.013	8.519	7.824	149.4	4.396	7.104	12.526
100	0.017	15.696	34.649	826.44	5.136	8.683	14.362
250	0.026	32.066	402.125	7839.942	7.756	13.252	22.379
500	0.041	35.828	2948.001	54209.15	13.622	21.413	32.404
750	0.058	58.697	10363.441	210494.564	24.67	36.563	55.422
1000	0.072	69.76	26282.893	531362.110	27.727	40.309	62.234

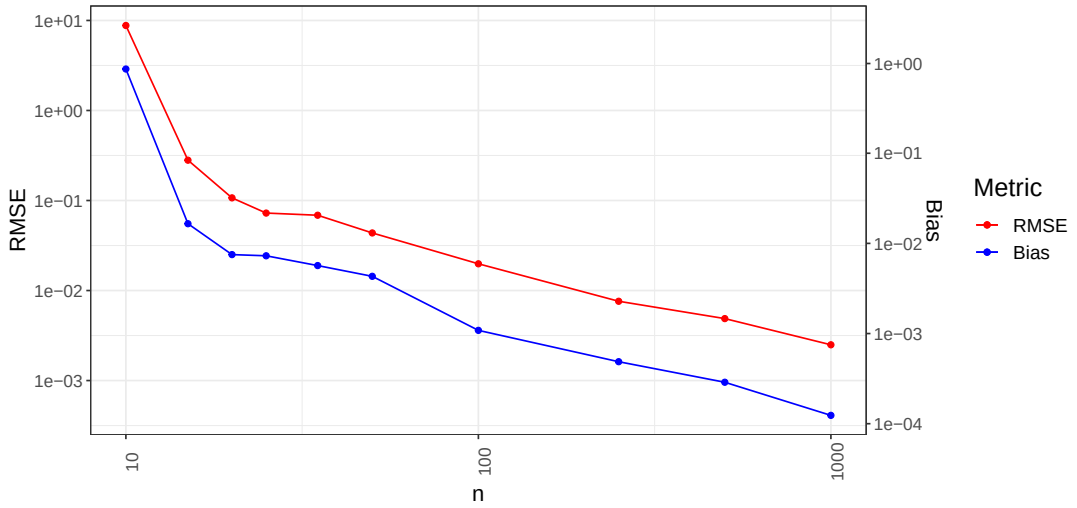
reduced to its non-regularized variant RobSVD. However, as Table 2.1 shows, the RobSVD algorithm without the regularization cannot provide a reliable estimate of singular values, even using Huber's robust loss function.

2.4.3 Effect of matrix dimensions on rSVDdpd estimates

We also perform some additional simulation studies to empirically verify the consistency of the rSVDdpd estimator as ensured by Theorem 2.7. Figure 2.1 depicts the empirical bias and the RMSE of the first singular values (obtained through the rSVDdpd algorithm) of 1000 simulations of $n \times n$ matrix $\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{E}$, where $\boldsymbol{L}$ is a matrix with rank three and $\boldsymbol{E}$ is a matrix with i.i.d. Gaussian entries with mean 0 and variance $1/n$ (as required in Assumption (B6)). As evident from Figure 2.1, both the bias and the RMSE of the estimated first singular value tend to 0 as the dimension n tends to infinity, supporting the consistency of the estimator as ensured by Theorem 2.7.

Another simulation study is undertaken to observe the effect of the dimensions and the rank of the matrix on the bias and RMSE of the estimated singular values, as obtained by the different SVD algorithms. For this simulation, we generate 1000 replications of $n \times p$ data matrix following the decomposition as in Eq. (2.2), keeping the number of rows fixed at $n = 100$ and varying the choices of the number of columns p. The low-rank matrix was generated by simulating two random orthogonal matrices of requisite dimensions, and taking the top three singular values all equal to 1 and the rest of the singular values equal to 0, thus, effectively fixing the rank at $r = 3$. The error component was generated according to the simulation setup (SVD2b) with 10% level of contamination, and the noise variance was kept equal to $1/\sqrt{np}$ as necessitated by Assumption (B6). The average bias across three singular value estimates and the average RMSE are depicted in Figure 2.2 as a function of $c = (p/n)$. As evident from Figure 2.2, the rSVDdpd algorithm produces singular value estimates that are more accurate than the ones obtained by the traditional

Figure 2.1: Empirical Bias and RMSE of the first singular value estimated by rSVDdpd algorithm with $\alpha = 0.5$ of an $n \times n$ matrix comprising a matrix of rank three and an error matrix (Both the horizontal and vertical axes are in logarithmic scale).



SVD or existing robust SVD methods. Also, the bias and the RMSE of almost all the methods increase with an increase in c. In a second experiment which is otherwise similar to the first one, we consider varying the rank of the low-rank component but keep the number of columns fixed. In this case, we fix the dimension of the data matrix at $n = 100$ and $p = 50$. Figure 2.3 shows the estimated bias and RMSE from 100 replications for different SVD algorithms as a function of $\delta = r/p$, where r is the rank of the low-rank component. Similar to the first experiment, the proposed rSVDdpd method outperforms the existing robust SVD approaches. Also, an increase in δ improves the signal-to-noise ratio of the data matrix, which results in a downward trend for the bias and RMSE for all SVD algorithms, except pcaSVD.

2.5 Application to Background Modelling

Automated surveillance from noisy videos is an extremely important problem with applications in different areas such as defence, security, research and monitoring, etc. In such video surveillance, a basic algorithmic task is to separate the background of a video from the foreground or moving objects, based on the input image frames from a surveillance video. The modelled background and foreground are then widely used in different image processing and computer vision applications such as monitoring human activities in traffic surveillance systems, movements of animals and insects to study their behaviours, vision-based hand gesture recognition, autonomous vehicle pilot systems, content-based video coding, etc. Interested readers are referred to Garcia-Garcia et al. (2020) and the references therein.

Various methods for background modelling and foreground detection for video surveil-

Figure 2.2: The estimated bias and the RMSE of different SVD algorithms under change in the number of columns while the number of rows and the matrix rank are kept fixed; (y-axis is in the log-scale). The group of lines on the top contains performance for svd, rSVD, pcaOne and RobSVD methods, while the group of lines at the bottom consists of performances for the rSVDdpd methods for different choices of α.

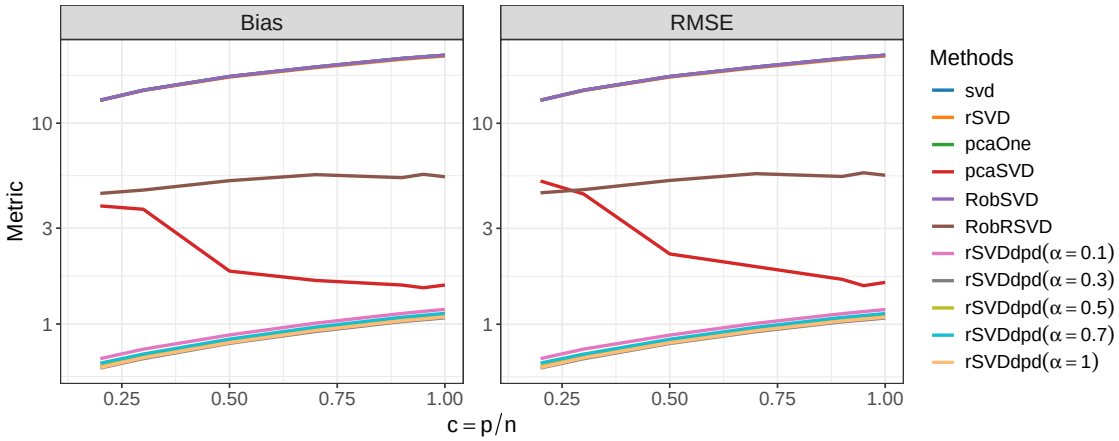
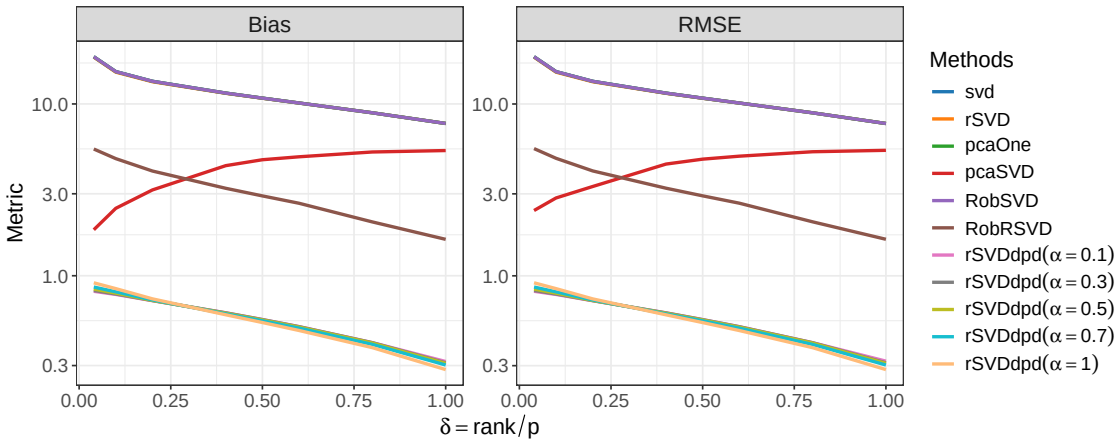


Figure 2.3: The estimated bias and the RMSE of different SVD algorithms under change in the rank of low-rank component while the dimension of the data matrix is kept fixed; (y-axis is in the log-scale). The group of lines on the top contains performance for svd, rSVD, pcaOne and RobSVD methods, while the group of lines at the bottom consists of performances for the rSVDdpd methods for different choices of α.



lance data have been developed in the past two decades. The different models can be classified into mathematical models (mean based (Lee and Hedley, 2002, McIvor, 2000) and fuzzy theory based (Subudhi et al., 2022, Zhao et al., 2012)), statistical models (Bouwmans et al., 2010, Lee, 2005), machine learning models (subspace learning (Zhang et al., 2016b, Vaswani et al., 2018), neural networks (Maddalena and Petrosino, 2012) and deep learning methods (Giraldo et al., 2021, Mandal and Vipparthi, 2022)), and signal processing methods (filter based (Messelodi et al., 2005), sparse representation based (Stagliano et al., 2015) and graph signal processing (Giraldo et al., 2022)). Among these, some of these methods are unsupervised, such as subspace learning methods; some are supervised such as deep learning methods and some are semi-supervised such as the graph-based signal processing methods. Recently, deep learning methods have emerged offering state-of-the-art performances; however, there have been a few serious drawbacks to them.

1. The deep learning methods are supervised and require a large volume of labelled data. Although in today's world, the data are usually abundant; corrupted and noisy data along with error-prone labelling can lead to drastically wrong results (Song et al., 2023).

2. Deep learning methods do not model the underlying video data generation process in a proper theoretical (or statistical) way, rather these methods solely focus on improving the accuracy of the results by minimising an appropriate loss function. Thus, their performances decrease dramatically in the presence of unseen data not present in the training set. Even if sometimes they generalize well to the unseen data, much of their characteristics remained largely unexplained (Neyshabur et al., 2017).

3. These methods also require sophisticated hardware that may not be available for all practical applications (e.g., a street surveillance camera). In this case, the camera must send the required data over the network to a computing device with the necessary hardware to perform the background modelling. This increases latency and introduces more errors and noises in the video data during transmission.

In contrast, unsupervised statistical methods are backed by rich theoretical properties, require no labelled training data apriori and can be used in real-time applications. Among these, methods based on robust principal component analysis (PCA) have been shown to deliver state-of-the-art performances (Bouwmans, 2014, Bouwmans and Zahzah, 2014) compared to the traditional approaches of background modelling. Here, we focus on comparing the proposed rSVDdpd method with some existing robust PCA algorithms for this particular application. To perform background and foreground extraction from a given video data, we construct a data matrix $\boldsymbol{X}$ by stacking the pixel values in all frames of a video as columns. This means, for a p-frame long video with each frame being h pixels long and w pixels wide, the $\boldsymbol{X}$ matrix becomes of dimension $hw \times p$. As the background content of the video changes very slowly from one frame to another, it can be represented by the low-rank component $\boldsymbol{L} = \sum_{k=1}^{r} \lambda_k \boldsymbol{u}_k \boldsymbol{v}_k^\intercal$ as in (2.2), where r is usually

predetermined (1 or 2). Consequently, the foreground part of the video can be obtained from the error $\boldsymbol{E}$.

Although the temporal correlation present in the video data indicates that the columns of the low-rank matrix $\boldsymbol{L}$ satisfy some smoothness constraints, these may not be available for the rows of the matrix. For example, when the vectorization transformation converts a single frame of the video into a column vector, the spatial correlation present in the frame is usually lost when switching from one row to another row of the same image frame. However, one can follow a procedure similar to the diagonal enumeration technique of rational numbers to recover the spatial correlation. To illustrate this point, let $((I_{ij}))_{h\times w}$ be the pixel values of the t-th image frame of the video. Then, we can obtain by t-th column of $\boldsymbol{X}$ as

$$X_{1t} = I_{11},\ X_{2t} = I_{12},\ X_{3t} = I_{21},\ X_{4t} = I_{31},\ X_{5t} = I_{22}, X_{6t} = I_{13}, \ldots.$$

This process transforms the spatial correlation present in the image pixels to a smoothness condition on the rows of $\boldsymbol{X}$, and consequently on the rows of $\boldsymbol{L}$. This enables us to use a robust SVD technique such as the proposed rSVDdpd algorithm to solve the problem as opposed to tackling it via the existing robust PCA approaches.

In this section, we use the rSVDdpd algorithm to perform this low-rank decomposition for the background modelling problem and compare its performance to several existing algorithms. Among robust PCA-based approaches, we consider the exact Principal Component Pursuit (PCP) via Augmented Lagrangian Method (ALM) (Candès et al., 2011), inexact ALM via alternating direction (Shen et al., 2014), Outlier Pursuit (OP) (Xu et al., 2012), Sparse Regularized PCP (SRPCP) (Liu et al., 2017) and Variational Bayesian method (VB) (Babacan et al., 2012). In addition, we also consider one statistical algorithm Go Decomposition (GoDec) (Zhou and Tao, 2011) and an incremental algorithm, namely the Grassmannian Robust Adaptive Subspace Tracking Algorithm (GRASTA) (He et al., 2011, 2012). The implementation of the exact PCP method is available in the R package `RPCA` (Sykulski, 2015), while the implementations of the rest of the existing algorithms are available in Background Subtraction Website[1].

2.5.1 Background Modelling Challenger Dataset

To compare these methods and measure performances, we consider the Background Models Challenger (BMC) benchmark dataset (Vacavant et al., 2013). This benchmark dataset has also been used in earlier studies to provide a detailed comparison of several background modelling algorithms (Bouwmans and Zahzah, 2014). The dataset contains 20 synthetic videos with ground truth foreground masks. The synthetic videos are generated by the SIVIC software and closely resemble different kinds of natural tampering present in real-life video surveillance data such as noises, artefacts, shadows, changes of illumination, the presence of fog and rain, etc. The whole set of videos comprises two kinds of backgrounds, one of a straight street highway (Street), while the other contains the simulated video

[1] https://sites.google.com/site/backgroundsubtraction/Home

Table 2.3: Benchmark Results for the Street Video (Video number $Vi1j$ denotes i-th variant of the Street Video and $j = 1$ or 2 denote the training and the testing video respectively); In each column, bold symbols indicate the algorithm with best F_1 measure.

Method	Measures	V111	V112	V211	V212	V311	V312	V411	V412	V511	V512
RPCA (Candès et al., 2011)	Precision	0.766	0.741	0.788	0.764	0.751	0.761	0.983	0.784	0.617	0.487
	Recall	0.768	0.767	0.747	0.744	0.7	0.698	0.655	0.675	0.568	0.49
	F1	0.767	0.754	0.767	0.754	0.725	0.728	0.787	0.726	0.592	0.489
ALM (Shen et al., 2014)	Precision	0.766	0.741	0.811	0.765	0.752	0.762	0.984	0.785	0.626	0.486
	Recall	0.768	0.767	0.729	0.745	0.7	0.698	0.655	0.675	0.568	0.49
	F1	0.767	0.754	0.768	0.755	0.725	0.729	0.787	0.726	0.595	0.488
SRPCP (Liu et al., 2017)	Precision	0.766	0.741	0.811	0.765	0.752	0.762	0.984	0.785	0.626	0.486
	Recall	0.768	0.767	0.729	0.745	0.7	0.698	0.657	0.675	0.568	0.49
	F1	0.767	0.754	0.768	0.755	0.725	0.729	**0.788**	0.726	0.595	0.488
VB (Babacan et al., 2012)	Precision	0.774	0.749	0.809	0.764	0.75	0.761	0.985	0.781	0.323	0.38
	Recall	0.767	0.766	0.729	0.745	0.701	0.698	0.654	0.649	0.48	0.435
	F1	0.77	0.757	0.767	0.754	0.725	0.728	0.786	0.709	0.386	0.405
OP Xu et al. (2012)	Precision	0.783	0.776	0.825	0.794	0.757	0.837	0.985	0.797	0.686	0.619
	Recall	0.769	0.748	0.727	0.723	0.699	0.532	0.653	0.643	0.601	0.5
	F1	0.776	0.762	0.773	**0.757**	**0.727**	0.65	0.785	0.712	0.641	**0.553**
GoDec (Zhou and Tao, 2011)	Precision	0.844	0.8	0.786	0.762	0.749	0.771	0.981	0.51	0.205	0.414
	Recall	0.686	0.676	0.748	0.745	0.701	0.595	0.654	0.443	0.364	0.214
	F1	0.757	0.733	0.767	0.754	0.724	0.672	0.785	0.474	0.263	0.282
GRASTA (He et al., 2012)	Precision	0.646	0.766	0.613	0.758	0.764	0.89	0.987	0.165	0.338	0.374
	Recall	0.764	0.644	0.727	0.712	0.613	0.282	0.652	0.189	0.395	0.367
	F1	0.700	0.699	0.665	0.734	0.68	0.428	0.785	0.176	0.364	0.371
rSVDdpd (ours)	Precision	0.783	0.776	0.78	0.769	0.753	0.842	0.987	0.8	0.785	0.626
	Recall	0.772	0.768	0.769	0.746	0.699	0.683	0.653	0.735	0.601	0.497
	F1	**0.777**	**0.772**	**0.774**	**0.757**	0.725	**0.754**	0.787	**0.766**	**0.681**	**0.553**
	$\widehat{\alpha}, \widehat{r}$	0.30, 1	0.30, 1	0.38, 1	0.39, 1	0.34, 2	0.34, 2	0.42, 2	0.44, 2	0.39, 3	0.39, 3

footage of a traffic junction (Rotary). Due to the computational limitations of the existing algorithms, we perform the low-rank decomposition by considering batches of 120 frames from the video at a time. The $F1$-scores between the estimated foreground and the provided ground truth frames for each algorithm are calculated, and they are summarized in Tables 2.3 and 2.4, respectively for the Street and the Rotary video. The estimated rank $\widehat{r}$ and robustness parameter $\widehat{\alpha}$ for the rSVDdpd algorithm for the corresponding videos are also indicated in the last row of these two tables.

As evident from Tables 2.3 and 2.4, the proposed rSVDdpd algorithm outperforms all the existing algorithms under consideration for almost all the benchmark videos. For some of the videos where rSVDdpd does not achieve the best $F1$-measure, the difference in $F1$-measure compared to the best output is too small to visually distinguish the extracted foregrounds. It turns out that the inexact ALM method and the SRPCP method have similar performances as the exact robust PCA algorithm, while the OP and VB methods perform better than these only in some specific scenarios, in particular when the background shows some continuous natural movement of the trees. Both GoDec and GRASTA algorithms do not work well under these situations. In comparison, the rSVDdpd algorithm can tackle all these scenarios by allowing the background matrix $\boldsymbol{L}$ to be of rank

Table 2.4: Benchmark Results for the Rotary Video (Video number $Vi2j$ denotes i-th variant of the Rotary Video and $j = 1$ or 2 denote the training and the testing video respectively); In each column, bold symbols indicate the algorithm with best F_1 measure.

Method	Measures	V121	V122	V221	V222	V321	V322	V421	V422	V521	V522
RPCA (Candès et al., 2011)	Precision	0.771	0.771	0.792	0.795	0.781	0.781	0.893	0.698	0.425	0.496
	Recall	0.771	0.769	0.733	0.738	0.656	0.659	0.612	0.547	0.592	0.599
	F1	0.771	0.77	0.761	0.766	0.713	0.714	0.726	0.614	0.495	0.542
ALM (Shen et al., 2014)	Precision	0.772	0.771	0.792	0.796	0.781	0.781	0.898	0.698	0.422	0.491
	Recall	0.771	0.769	0.733	0.738	0.656	0.659	0.611	0.546	0.592	0.599
	F1	0.772	0.77	0.761	0.766	0.713	0.714	0.727	0.613	0.493	0.54
SRPCP (Liu et al., 2017)	Precision	0.772	0.771	0.792	0.796	0.781	0.781	0.898	0.698	0.422	0.491
	Recall	0.771	0.769	0.733	0.738	0.656	0.659	0.611	0.546	0.592	0.599
	F1	0.772	0.77	0.761	0.766	0.713	**0.715**	0.727	0.613	0.493	0.54
VB Babacan et al. (2012)	Precision	0.772	0.772	0.793	0.797	0.776	0.773	0.904	0.683	0.447	0.483
	Recall	0.771	0.769	0.733	0.738	0.658	0.661	0.611	0.518	0.526	0.527
	F1	0.772	0.77	**0.762**	0.766	**0.712**	0.713	0.729	0.589	0.483	0.504
OP (Xu et al., 2012)	Precision	0.806	0.79	0.792	0.802	0.751	0.797	0.883	0.655	0.551	0.544
	Recall	0.728	0.733	0.713	0.7	0.657	0.468	0.611	0.517	0.467	0.488
	F1	0.765	0.76	0.75	0.749	0.701	0.59	0.722	0.577	0.506	0.514
GoDec (Zhou and Tao, 2011)	Precision	0.815	0.795	0.758	0.762	0.737	0.796	0.987	0.57	0.502	0.512
	Recall	0.624	0.688	0.734	0.739	0.661	0.539	0.57	0.587	0.476	0.5
	F1	0.707	0.738	0.746	0.75	0.697	0.642	0.723	0.578	0.488	0.506
GRASTA (He et al., 2012)	Precision	0.82	0.56	0.604	0.797	0.758	0.762	0.892	0.45	0.334	0.346
	Recall	0.684	0.692	0.69	0.7	0.628	0.301	0.61	0.283	0.603	0.611
	F1	0.746	0.619	0.644	0.748	0.687	0.432	0.724	0.347	0.43	0.442
rSVDdpd (ours)	Precision	0.772	0.784	0.798	0.805	0.735	0.685	0.901	0.697	0.742	0.634
	Recall	0.792	0.769	0.729	0.733	0.686	0.718	0.65	0.58	0.56	0.586
	F1	**0.781**	**0.776**	**0.762**	**0.767**	0.710	0.702	**0.755**	**0.633**	**0.638**	**0.609**
	$\widehat{\alpha}, \widehat{r}$	0.32, 1	0.32, 1	0.39, 1	0.39, 1	0.47, 2	0.47, 2	0.49, 2	0.49, 2	0.43, 3	0.43, 3

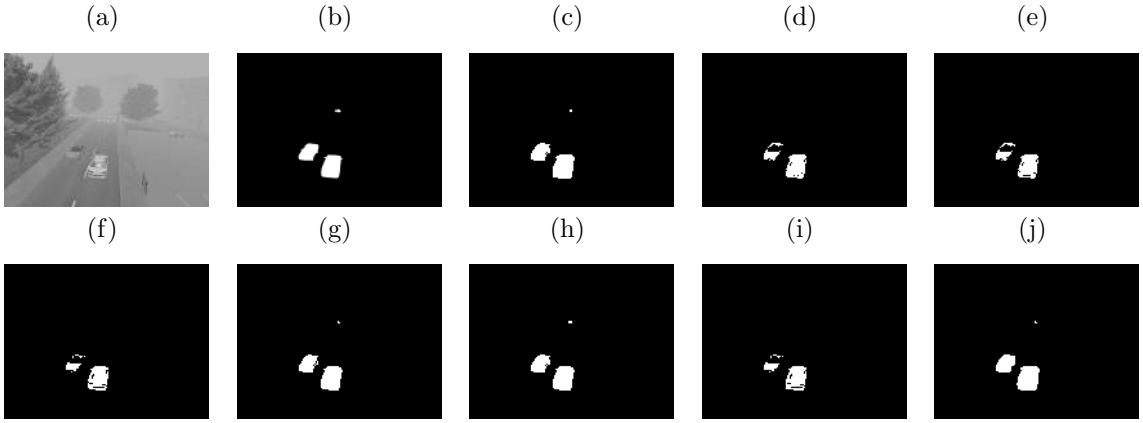
Table 2.5: Processing time (in seconds per frame) for benchmark videos in BMC dataset for different background modelling algorithms.

Method	Average Time (in seconds/frame)	Std. dev. (in seconds/frame)
GRASTA (He et al., 2012)	0.37	0.01
rSVDdpd (ours)	2.86	0.84
VB (Babacan et al., 2012)	7.06	1.39
GoDec (Zhou and Tao, 2011)	12.06	0.29
OP (Xu et al., 2012)	23.42	0.22
ALM (Shen et al., 2014)	34.85	7.40
SRPCP (Liu et al., 2017)	54.25	9.12
RPCA (Candès et al., 2011)	136.41	27.26

more than one, usually 2 or 3.

We also demonstrate a comparison in the speed of these background modelling algorithms through Table 2.5. Only GRASTA and rSVDdpd algorithms complete the foreground extraction task in a reasonable amount of time to be useful in real-time video

Figure 2.4: One frame of the street video (a), ground truth (b) and estimated foreground mask from exact RPCA (c), inexact ALM (d), SRPCP (e), Variational Bayes (f), Outlier Pursuit (g), GoDec (h), GRASTA (i) and rSVDdpd (j) algorithms.



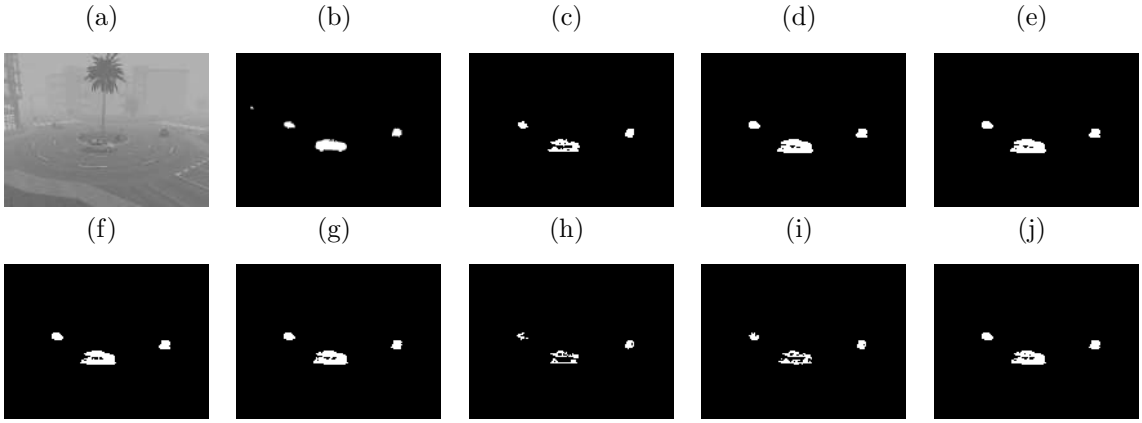
surveillance and object tracking. In contrast, the exact robust PCA method takes a ridiculously large amount of time to converge, more than 2 minutes per frame, thus having very limited practical utility. The variants of robust PCA, i.e., inexact ALM and SRPCP methods reduce the computational time significantly, but they are still not on par with the speed of rSVDdpd or GRASTA.

As seen from the above discussions, the rSVDdpd algorithm achieves superior or near-optimal performance compared to the existing robust PCA and background modelling methods by spending only a fraction of computational time. GRASTA, the only algorithm that is faster than the proposed rSVDdpd, achieves this speed at the cost of very poor performance under different kinds of noises that are usually present in a real-life video data. A fast algorithm with little performance guarantee can only be of limited use. The performance improvement of rSVDdpd over GRASTA is substantial in most cases (e.g., videos 312, 511, 512 in Table 2.3 and 322, 521, 522 in Table 2.4). The rSVDdpd holds its own even when practically all the other algorithms show a largely degraded performance, e.g., as in the case of video 512. Therefore, the proposed rSVDdpd is competitive or better than existing algorithms in consideration, either in performance accuracy or speed or both, in the context of video surveillance background modelling problems.

For streamlining the presentation, we demonstrate the estimated foreground masks as obtained from only two frames of the two videos, using different existing algorithms and the proposed rSVDdpd algorithm in Figures 2.4 and 2.5. The results obtained from all the videos in this benchmark dataset are available on the accompanying rSVDdpd webpage[2]. As shown in Figures 2.4 and 2.5, both the exact robust PCA method and the rSVDdpd method lead to visually indistinguishable results; however rSVDdpd achieves the same at the cost of a significantly reduced computational time as indicated earlier. The other alternative competing methods for video surveillance, namely GRASTA and GoDec estimate a noisy background in general. Although these noises remain imperceptible at

[2] https://subroy13.github.io/rsvddpd/

Figure 2.5: One frame of the rotary video (a), ground truth (b) and estimated foreground mask from exact RPCA (c), inexact ALM (d), SRPCP (e), Variational Bayes (f), Outlier Pursuit (g), GoDec (h), GRASTA (i) and rSVDdpd (j) algorithms.



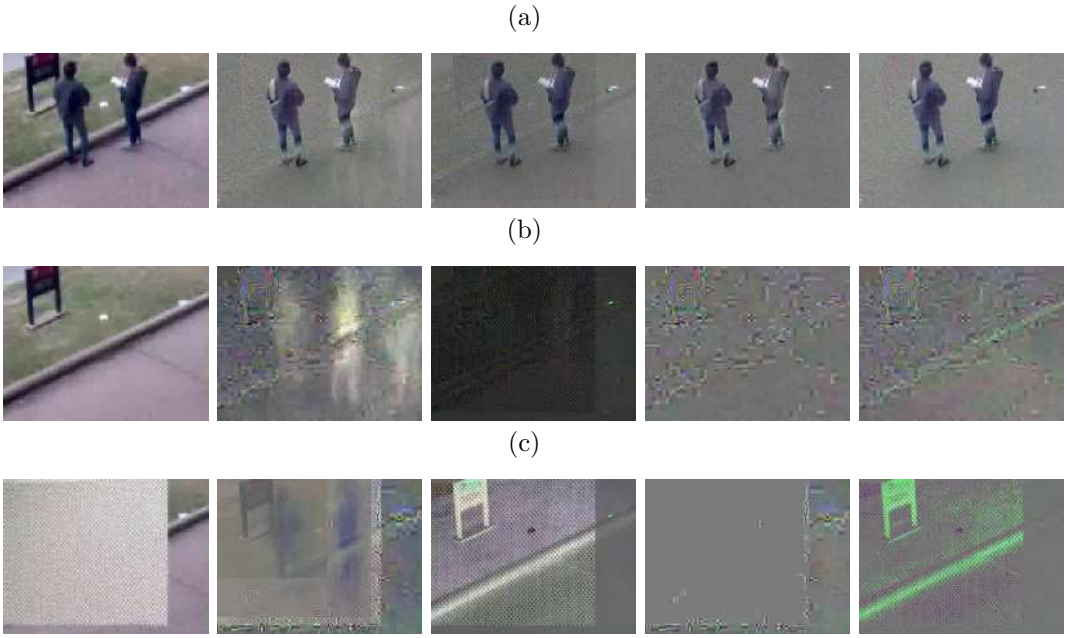
first, they result in a more noisy foreground, and such noisy pixels sometimes become visible in the thresholded foreground mask.

2.5.2 University of Houston Camera Tampering Dataset

Among different types of noises and outliers present in a video surveillance data, camera tampering poses a major problem (Sitara and Mehtre, 2019). To illustrate the applicability of rSVDdpd for background modelling under camera tampering, we measure its performance on some video extracts from the University of Houston Camera Tampering Detection Dataset (UHCTD). Mantini and Shah (2019b) compiled this comprehensive large-scale dataset with over 288 hours of surveillance video footage ranging across 6 days from two cameras in the University of Houston campus, and synthesized different types of camera tampering methods (covered, defocused, moved, etc.) in the captured video data.

Figure 2.6 depicts the true frames along with the estimated foregrounds obtained from the traditional SVD, rSVDdpd (ours), robust PCA method via ALM and GoDec algorithms. The estimated backgrounds obtained from other robust PCA methods are visually similar to the output of the ALM algorithm and hence are omitted here. In this surveillance scene from Camera B, a noisy image is synthesized to obstruct the view of the camera. The background estimated from the traditional SVD shows a clear indication of the noise even in the frames where camera tampering is not induced, and such an effect is amplified further by the presence of shadow-like artefacts in the estimated background content. Such an effect is also prevalent across the existing techniques. In comparison, the background estimated via the proposed rSVDdpd algorithm removes such artefacts and is less affected by tampering. This robustness property can also be seen in the estimated foreground content when the true frame is exactly the same as the non-tampered background. In this case, since there is no foreground content, the proposed rSVDdpd estimates the foreground as a very dark black image without any distinguishing feature as desired.

Figure 2.6: True video frames and the estimated foreground by usual SVD, rSVDdpd, ALM and GoDec algorithms (Left to Right) for frame 5, 50, 75 in UHCTD Day 1 Camera B dataset.



In Figure 2.7, we depict similar estimated foreground images for another surveillance video footage captured from another camera, during night-time. Since night-time footage contains mostly darker pixel values, the signal strength is low. These results demonstrate that rSVDdpd is capable of extracting the background and foreground content correctly even under this low signal-to-noise ratio situations.

2.6 Example: Factor Analysis of Spatiotemporal Data

In this section, we consider an example of spatio-temporal data analysis. We consider a dataset comprising of hourly PM2.5 measurements at 38 different pollution measurement stations across Delhi region, during the time period of January-June, 2020. The dataset is publicly available to download from the website[3] of Central Pollution Control Board (CPCB), India. We arrange the data into a matrix $\boldsymbol{X}$ where (i, j)-th entry X_{ij} denotes the PM2.5 measurement at i-th timepoint and j-th location (measurement station) resulting in a matrix of dimensions 4344×38. In this case, the temporal correlation present in the measurements appears as the correlation between the rows while the spatial correlation appears as the correlation between the columns of the $\boldsymbol{X}$ matrix. An important objective to analyze such spatiotemporal dataset is to extract the latent factors, which are useful to detect spatial clusters, determine regions of temporal stationarity, etc.; see Cressie and Wikle (2011) for details. The proposed robust SVD estimator rSVDdpd is naturally suitable to extract such latent factors from the data matrix $\boldsymbol{X}$.

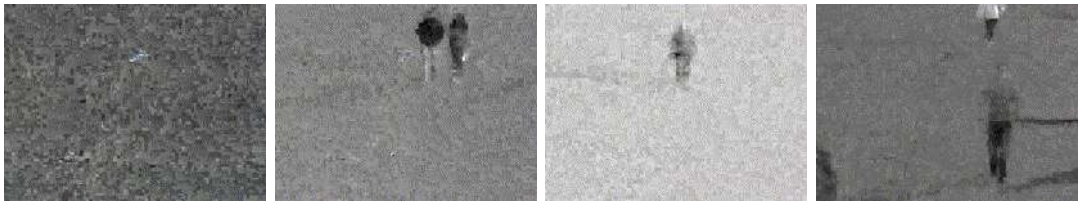
[3] https://airquality.cpcb.gov.in/AQI_India/

Figure 2.7: True video frames and the estimated foreground by usual SVD, GoDec, Variational Bayes (VB), Sparse Regularized PCP and rSVDdpd algorithms (Top to Bottom) for frame 50, 100, 150, 200 in UHCTD Day 1 Camera A dataset.

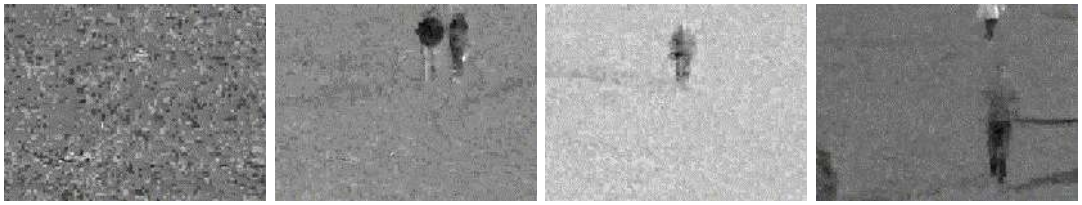
(a)

(b)

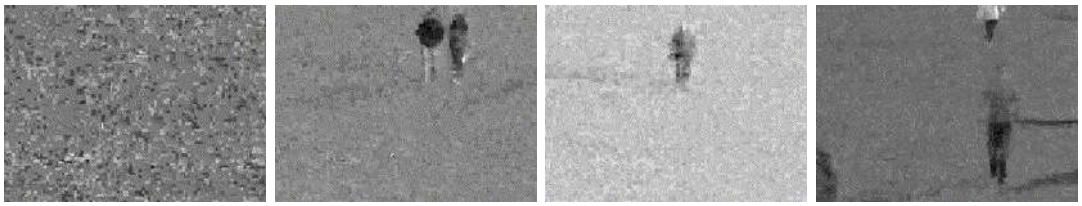
(c)



(d)



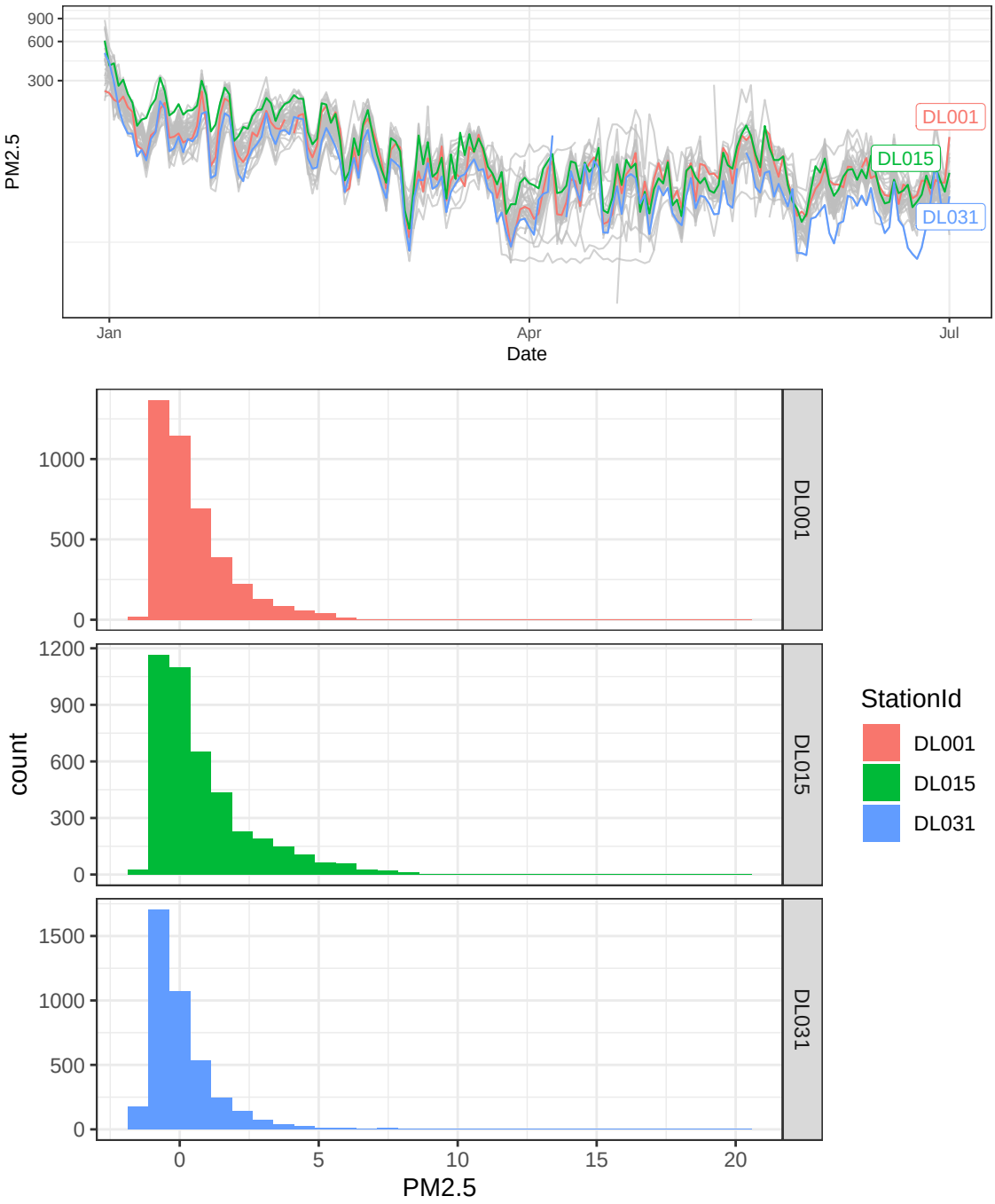
(e)



(f)

However, the PM2.5 measurements are rarely normally distributed. As seen from bottom part of Figure 2.8, the distribution of PM2.5 measurements for all stations are positively skewed, thus violating Assumption (B3). In this scenario, the performance

Figure 2.8: [Top] The time series graph highlighting the daily average PM2.5 measurements in 3 stations; y-axis is in log scale. [Bottom] The histogram indicating the distribution of the PM2.5 measurements in those three stations.



of rSVDdpd algorithm is slightly degraded, as seen in the simulation results presented in Section 2.4.2 for lognormally distributed errors. A popular technique to handle such skewed data is to apply some transformation to the data so that the transformed data approximately follows a Gaussian distribution (or a symmetric distribution). As suggested in Atkinson et al. (2020), it may be useful to apply the robustified version of extended Yeo-Johnson transformation on the PM2.5 measurements, after a proper centering and scaling of the values. The extended Yeo-Johnson transformation for a datapoint x is given by the formula

$$z_{\text{EYJ}}(x;\lambda_p,\lambda_n) = \begin{cases} \frac{(x+1)^{\lambda_p}-1}{\lambda_p \dot{x_n}^{\lambda_n-1}\dot{x_p}^{\lambda_p-1}} & \text{if } x \geq 0, \lambda_p \neq 0, \\ (\dot{x_p}/\dot{x_n}^{\lambda_n-1})\log(x+1) & \text{if } x \geq 0, \lambda_p = 0, \\ -\frac{(1-x)^{2-\lambda_n}-1}{(2-\lambda_n)\dot{x_n}^{\lambda_n-1}\dot{x_p}^{\lambda_p-1}} & \text{if } x < 0, \lambda_n \neq 2, \\ -\log(1-x)/\dot{x_n}\dot{x_p}^{\lambda_p-1} & \text{if } x < 0, \lambda_n = 2 \end{cases}$$

where $\lambda_p \in [0,1]$ and $\lambda_n \in [0,2]$ are hyperparameters to be chosen respectively for positive and negative datapoints. Here,

$$\dot{x_p} = \exp\left(\frac{1}{n}\sum_{x_{ij}\geq 0}\log(1+x_{ij})\right), \text{ and,}$$
$$\dot{x_n} = \exp\left(-\frac{1}{n}\sum_{x_{ij}<0}\log(1-x_{ij})\right).$$

In Figure 2.9, we have plotted the signal strength (i.e., the ratio $\|\widehat{\boldsymbol{L}}\|_2^2/\|\boldsymbol{Z}\|_2^2$ where $\widehat{\boldsymbol{L}}$ is the low-rank estimate of $\boldsymbol{Z}$) obtained from the rSVDdpd estimate of the transformed data matrix $\boldsymbol{Z}$ (with entries $z_{\text{EYJ}}(x_{ij};\lambda_p,\lambda_n)$) with extended Yeo-Johnson transformations for different values of λ_p and λ_n. As evident from the plot, the signal strength is considerably improved with the choice $\lambda_p = 0$ (corresponding to the logarithmic transformation) compared to the choice $\lambda_p = 1$ (corresponding to no transformation of the data). A similar effect is also prominent for the hyperparameter λ_n, which maximizes the signal strength at $\lambda_n = 2$, again corresponding to the logarithmic transformation. This suggests that an initial logarithmic transformation on the PM2.5 measurements is appropriate before applying the rSVDdpd algorithm to obtain robust latent factors present in this spatio-temporal pollutants dataset. The histogram of the transformed data is shown in Figure 2.10, which clearly shows a unimodal and less skewed distribution.

Figure 2.9: The signal strength obtained from the rSVDdpd estimate of the low rank component $\boldsymbol{L}$ for the transformed data with extended Yeo-Johnson transformations for different values of λ_p and λ_n.

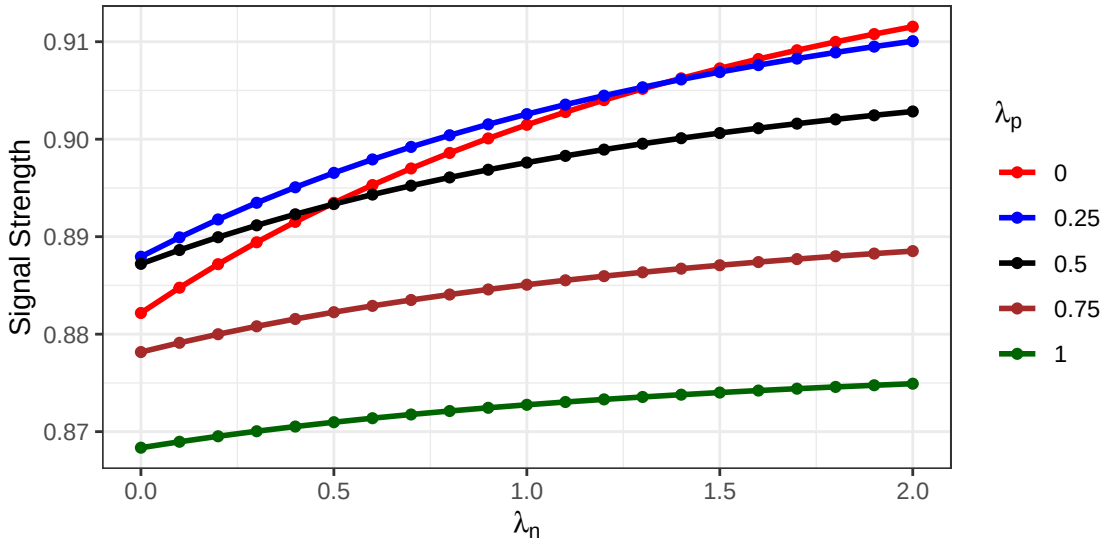
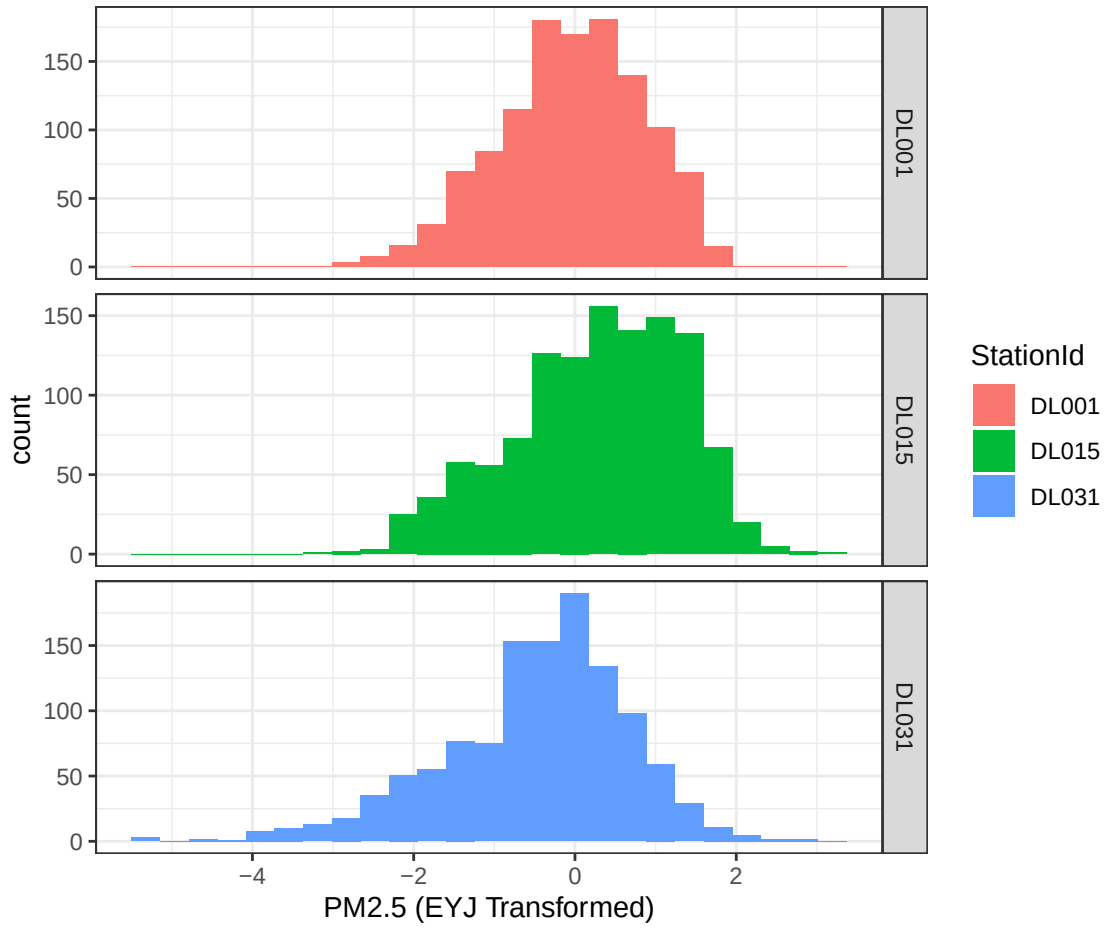


Figure 2.10: Histogram of the transformed PM2.5 measurements after applying extended Yeo-Johnson transformation with $\lambda_p = 0$ and $\lambda_n = 2$.



Chapter 3

Robust Principal Component Analysis

3.1 Introduction

The objective of the Principal Component Analysis (PCA) is to find a linear combination of the sample features to explain the maximum possible variability. In Section 1.3.2, we have discussed the mathematical definition of PCA in detail. For the traditional PCA, the scale estimator $S_n(\cdot)$ as in (1.3) is taken to be the square root of the sample variance. As a result, the solution to the PCA problem becomes the same as the eigenvalues and the eigenvectors of the sample covariance matrix of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. It is well known that the sample covariance matrix is very sensitive to outliers, hence the principal components resulting from the traditional PCA algorithm also suffer from the presence of outlying observations in the data (Hubert et al., 2005, Candès et al., 2011). In the context of high dimensional datasets related to the applications mentioned before in Section 1.3.2, it is very challenging to locate these outlying observations beforehand for proper treatment and subsequent scrutiny. Thus, any practitioner relying solely on the output of the traditional PCA algorithm to interpret multivariate data may end up with a distorted visualization of the data, false detection of outliers, and a wrong conclusion about the data. Several robustified versions of PCA have been proposed to date to provide reliable estimates of the principal components even under the presence of outlying observations (Jolliffe, 2002).

3.2 Existing Works on Robust PCA

Most of the early literature to derive a robust principal component analysis (RPCA) followed one of the two primary approaches. The first class of estimators estimated the principal components robustly from the eigenvalues and the eigenvectors of a robust covariance matrix of the sample. Notable among this class of estimators are those due to Maronna (1976) and Campbell (1980), where the authors create affine-equivariant principal component estimates from robust M-estimators of the covariance matrix. Devlin et al. (1981) proposed to use the minimum covariance determinant (MCD) estimator and the minimum

volume ellipsoid (MVE) estimator (Rousseeuw, 1985) for this purpose due to their high breakdown compared to the M-estimators.

The other approach considered robustifying PCA by using a robust scale function S_n in Eq. (1.3). This idea was first presented by Li and Chen (1985) and was further developed later by Croux and Ruiz-Gazen (1996) where they considered the median absolute deviation about the sample median as the scale function. Various theoretical properties like the influence function, asymptotic distribution and the breakdown point of this estimator have also been established in the literature (Croux and Haesbroeck, 2000, Croux and Ruiz-Gazen, 2005). These estimators and their variants primarily restricted their attention to the elliptically symmetric family of distributions, i.e., the random sample of observations $\boldsymbol{X}_i$ for $i = 1, 2 \ldots, n$ were assumed to follow a density function of the form

$$f(\boldsymbol{x}) \propto g\left((\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{3.1}$$

where $g : \mathbb{R}^+ \to \mathbb{R}$ is a known function governing the shape of the density. It turns out that under this model,

$$\mathbb{E}(\boldsymbol{X}_i) = \boldsymbol{\mu}, \quad \text{and,} \quad \mathbb{E}\left((\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^\intercal\right) = \boldsymbol{\Sigma}.$$

Although a popular notion considers that the variance of $\boldsymbol{X}_i$ is $k_g\boldsymbol{\Sigma}$ where k_g is a constant depending on the function g, but it is possible to include such k_g in the dispersion matrix $\boldsymbol{\Sigma}$ itself by modifying the function g appropriately. Even though these statistical Robust PCA (RPCA) approaches guarantee the highest possible asymptotic breakdown point of 1/2, they show low asymptotic efficiency and sometimes large bias even at considerably lower levels of contaminations than their breakdown points (Fishbone and Mili, 2023).

Recent advances in the area of RPCA view the estimation of the principal components in a different light through the guise of a factor model. Wright et al. (2009) define the RPCA problem as the problem of recovering $\boldsymbol{L}$ from the unknown decomposition of the data matrix $\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}$, where $\boldsymbol{L}$ is a low rank matrix and $\boldsymbol{S}$ is a sparse noise component. The direct solution to this problem would consider the optimization problem

$$\min_{\boldsymbol{L},\boldsymbol{S}} \operatorname{Rank}(\boldsymbol{L}) + \gamma \|\boldsymbol{S}\|_0, \tag{3.2}$$

subject to the restriction that $\|\boldsymbol{S}\|_0 \leq k$ and $\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}$, for a predetermined value of k. Here, $\|\boldsymbol{A}\|_0$ denotes the L^0-norm of the matrix $\boldsymbol{A}$, i.e., the number of the nonzero entries of $\boldsymbol{A}$ and γ is a tuning parameter to control the balance between the rank of $\boldsymbol{L}$ and the sparsity of $\boldsymbol{S}$. As noted in Candès et al. (2011), the classical PCA seeks the best low-rank component $\boldsymbol{L}$ in terms of minimizing the usual Euclidean L^2 norm, i.e., it is related to the optimization problem $\min_{\boldsymbol{L}} \|\boldsymbol{X} - \boldsymbol{L}\|_2$ subject to the restriction that $\operatorname{Rank}(\boldsymbol{L}) \leq k$. However, the problem given in (3.2) is notoriously difficult to solve, hence Wright et al. (2009) and Candès et al. (2011) considered a simpler convex optimization problem

$$\min_{\boldsymbol{L},\boldsymbol{S}} \|\boldsymbol{L}\|_* + \gamma\|\boldsymbol{S}\|_1 \tag{3.3}$$

where $\|\boldsymbol{L}\|_*$ is the nuclear norm of the matrix $\boldsymbol{L}$, i.e., the sum of its singular values and $\|\boldsymbol{S}\|_1$ is the L^1 norm of the matrix $\boldsymbol{S}$. Various algorithmic techniques like Principal Component Pursuit (PCP) method (Candès et al., 2011), Augmented Lagrange Multiplier (ALM) method (Lin et al., 2010) and Alternating Projection (AltProj) algorithm (Cai et al., 2019) have been developed to solve this optimization problem efficiently. These newer approaches radically differ from the traditional statistical approaches in several aspects:

1. These methods are non-parametric in nature and do not make distributional assumptions on the rows of $\boldsymbol{X}$.

2. These methods assume that the only source of randomness comes from the positions of the nonzero entries of the sparse matrix $\boldsymbol{S}$.

3. The convergence and the correctness guarantees of these methods are then provided based on the bounds on the entries of the component matrices $\boldsymbol{L}$ and $\boldsymbol{S}$ directly, leading to an algorithm that solves the relaxed convex optimization problem as in (3.3) exactly.

However, this exact decomposition is often far from the truth, as in any practical application, every entry of the data $\boldsymbol{X}$ is subject to measurement errors. To mitigate this, Zhou et al. (2010) added an additional component with the noise $\boldsymbol{N}$ to the decomposition, giving rise to the LSN decomposition as shown in (2.1). Although such a decomposition is considered, the analysis of the algorithm still assumed $\boldsymbol{X}$ to be deterministic and considered the Frobenius norm of the noise matrix to satisfy $\|\boldsymbol{N}\|_F \leq \delta$, a prespecified level of noise variance to maintain a high signal-to-noise ratio.

In this chapter, we demonstrate how the rSVDdpd algorithm introduced in Chapter 2 can be used to solve the robust PCA problem. However, the mathematical analysis present in Section 2.3 does not hold immediately in this case, since the PCA considers the data matrix to have i.i.d. rows distributed according to a multivariate distribution as described in Section 1.3.2. In the subsequent discussions in Section 3.4, we present the modified theoretical studies with proper treatment of this setup. As will be shown below, our proposal enjoys two key theoretical advantages over the existing approaches.

1. It possesses various theoretical properties such as equivariance, $\sqrt{n}$-consistency and asymptotic distribution of the proposed RPCA estimator akin to the widely used robust M-estimators. There exists very little literature on the theoretical behaviour of the existing PCP methods and often the asymptotic distributions of these estimators are non-Gaussian (Bickel et al., 2018).

2. We can also theoretically demonstrate the robustness of the proposed estimator by showing that its influence function is bounded. Also, a global measure of robustness, namely the asymptotic breakdown point of the proposed estimator has a lower bound independent of the data dimension p but only a function of a robustness tuning parameter α, unlike the affine-equivariant M-estimators whose breakdown decreases

at the order of $1/(p+1)$. This ensures that the proposed estimator is scalable and can be used for arbitrarily high dimensional random samples.

In addition to these, the proposed algorithm is also computationally simple compared to many existing robust PCA algorithms. It turns out that the convex optimization problem as in (3.3) that needs to be solved by many of our competitors can present significant challenges in optimization in a high dimensional setting (Jain and Kar, 2017). This also results in an increase in the computational complexity of the existing algorithms as demonstrated in Section 2.5.

3.3 The rPCAdpd Estimator

3.3.1 Description

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a p-variate sample such that each of the observations $\boldsymbol{X}_i$ follows an elliptically symmetric family of distributions with a density function of the form

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = c_g^{-1} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left[g\left(\boldsymbol{x}^\intercal \sum_{k=1}^{p} \gamma_k^{-1} \boldsymbol{v}_k \boldsymbol{v}_k^\intercal \boldsymbol{x} \right) \right], \tag{3.4}$$

where $\boldsymbol{\Sigma} = \sum_{k=1}^{p} \gamma_k \boldsymbol{v}_k \boldsymbol{v}_k^\intercal$ is the eigen-decomposition of the dispersion matrix. A derivation on the particular form of the normalizing constant can be found in Appendix 3.A.1. The parameter $\boldsymbol{\theta} = (\gamma_1, \ldots, \gamma_p, \boldsymbol{\eta})$ in (3.4) consists of the eigenvalues $\gamma_1, \ldots, \gamma_p$ of the dispersion matrix $\boldsymbol{\Sigma}_{p\times p}$ and a parameter $\boldsymbol{\eta}$ parametrizing the eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ residing in the p-Stiefel manifold of order p, i.e., S_p^p. A description of this parameter $\boldsymbol{\eta}$ is given by the stereographic projections defined through the transformations (2.27) and (2.28) in the previous Chapter 2. As described in Eq. (3.1), g is a known scalar function. For instance, the multivariate Gaussian family of distributions corresponds to $g(x) = (-x/2)$. The quantity c_g is the normalizing constant depending on this function g. Note that, since the principal components primarily deal with the variance structure of the data, the location parameter $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X}_i)$ is a nuisance parameter, hence it is assumed to be a known constant. Without the loss of generality, we take this known location parameter equal to $\mathbf{0}$, otherwise, one may treat $\boldsymbol{Y}_i = \boldsymbol{X}_i - \boldsymbol{\mu}$ as the i.i.d. sample under consideration. Since the location parameter $\boldsymbol{\mu}$ is unknown in most practical situations, one can substitute $\boldsymbol{\mu}$ by any consistent robust estimate of the location parameter (some choices will be described later in Section 3.3.3). We shall show later in Section 3.4 that the choice of this location estimator does not affect the asymptotic properties of the proposed robust estimator of $\boldsymbol{\theta}$ described below.

Based on the above formulation, we shall again use the popular minimum density power divergence estimator (MDPDE) (Basu et al., 1998) to estimate these parameters in $\boldsymbol{\theta}$. In this case, we can recover MDPDE as the solution to the optimization problem

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\min}\, c_g^{-\alpha} \prod_{k=1}^{p} \gamma_k^{-\alpha/2} \left[\frac{c_{(1+\alpha)g}}{c_g} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^{n} e^{\alpha g(\mathrm{qf}(\boldsymbol{X}_i))} \right], \tag{3.5}$$

where

$$\mathrm{qf}(\boldsymbol{X}_i) = \boldsymbol{X}_i^\intercal \sum_{k=1}^{p} \gamma_k^{-1} \boldsymbol{v}_k(\boldsymbol{\eta}) \boldsymbol{v}_k^\intercal(\boldsymbol{\eta}) \boldsymbol{X}_i, \tag{3.6}$$

and $c_{(1+\alpha)g}$ is the normalizing constant as in (3.4) when we take the generating function in the elliptically symmetric family of densities as $(1+\alpha)g(\cdot)$. We refer to the above as the **rPCAdpd estimator** of the principal components under the general elliptically symmetric family of distributions. The existence and the uniqueness of this estimator have been later established in Section 3.4.1. This estimator assumes the description of the model family through the specification of the completely known function $g(\cdot)$. In particular, when $g(x) = (-x/2)$, i.e., the model family is a p-variate Gaussian distribution, then the corresponding optimization problem in (3.5) translates to

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\min} (2\pi)^{-\alpha p/2} \prod_{k=1}^{p} \gamma_k^{-\alpha/2} \left[(1+\alpha)^{-p/2} - \left(1+\frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^{n} e^{-\frac{\alpha}{2}\mathrm{qf}(\boldsymbol{X}_i)} \right], \tag{3.7}$$

where $\mathrm{qf}(\boldsymbol{X}_i)$ is as in (3.6).

3.3.2 Efficient Computation of the rPCAdpd Estimator

Clearly, if the minimization given in (3.5) was to be performed on the entries of the dispersion matrix to obtain a robust estimate of covariance directly, it would be difficult to restrict the optimization space to the space of all positive definite matrices. Thus, the optimization is deliberately made with respect to the eigenvalues and the eigenvectors of the dispersion matrix to ensure that the estimated dispersion matrix remains positive definite and symmetric. While it is easy to optimize the objective function in (3.5) with respect to the eigenvalues, it still remains computationally expensive to solve it for the eigenvectors due to the non-convexity of the Stiefel manifold S_p^p. Although there exist some efficient optimization algorithms on the Riemannian manifold as proposed by Wen and Yin (2013), Jiang and Dai (2015), Li et al. (2020), these general-purpose optimization techniques require complicated iteration steps via Cayley transformation and curvilinear searches.

To circumvent this direct optimization, we can apply the same rSVDdpd algorithm introduced earlier in Section 2.2.3. We start by assuming that the unknown location parameter $\boldsymbol{\mu}$ is already estimated using a robust consistent estimator of the location. For our purpose, we use the L_1-median as the location estimator; however, in Section 3.3.3, we shall describe some alternative choices that may be used. Applying the rSVDdpd algorithm on the centered data matrix $\boldsymbol{X} - \mathbf{1}^\intercal\widehat{\boldsymbol{\mu}}$, one obtains the estimates of the singular values $\widehat{\lambda}_k^*$ and the left and right singular vectors $\widehat{\boldsymbol{u}}_k^*$ and $\widehat{\boldsymbol{v}}_k^*$ for $k = 1, 2, \ldots, r$, where r is the rank of the low-rank component of the data matrix $\boldsymbol{X}$. By definition of the PCA, the left singular vectors are random variables, hence we discard them. The quantities $(\widehat{\lambda}_k^{(*)})^2/n$ and $\widehat{\boldsymbol{v}}_k^*$s are outputted as the k-th eigenvector and k-th eigenvalue corresponding to the principal components of the i.i.d. sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ respectively.

Although the rSVDdpd algorithm has been extensively studied in previous Chapter 2 by considering the Gaussian density for the modelling the errors, it is possible to choose

Table 3.1: The choices of $g(\cdot)$ and $\psi(\cdot)$ functions for different elliptically symmetric family of densities. Here, $K_\nu(\cdot)$ is the modified Bessel function of the second kind.

Density family	$g(x)$	$\psi(x)$
Multivariate Normal	$-x/2$	$e^{-\alpha x^2/2}$
Multivariate t	$-\frac{\nu+p}{2}\log(1+x/\nu)$	$(1+p/\nu)(1+x^2/\nu)^{-\alpha(\nu+p)/2-1}$
Multivariate Logistic	$-x-2\log(1+e^{-x})$	$2e^{\alpha x^2}(e^{x^2}-1)(1+e^{x^2})^{-(2\alpha+1)}$
Multivariate Laplace	$\frac{\nu}{2}\log(x)+\log(K_\nu(\sqrt{2}x))$	$2\sqrt{2}\lvert x\rvert^{\alpha\nu}K_\nu^{\alpha-1}(\sqrt{2}x^2)K_{\nu-1}(\sqrt{2}x^2)$ $+\nu\lvert x\rvert^{\alpha\nu-2}K_\nu^{\alpha}(\sqrt{2}x^2)$

different ψ function in the iterative equations (2.15)-(2.17) to tackle different choices of the $g(\cdot)$ function in Eq. (3.5). Simple calculations yield the relationship between the weight function ψ and g function as

$$\psi(x) = -2e^{\alpha g(x^2)}g'(x^2),\ x \geq 0. \tag{3.8}$$

For some standard elliptically symmetric family of distributions, the choices of corresponding g and the ψ functions are indicated in Table 3.1.

3.3.3 Choice of the Robust Location Estimator

There are several choices for a robust location estimator to be used alongside the rPCAdpd algorithm. We shall discuss only a few of these estimators which are quick and simple since the primary focus of this chapter is to estimate the principal components. As we shall show later in Section 3.4, the asymptotic properties of the estimated principal components are free of the choice of this location estimator, as long as the location estimator is robust and asymptotically consistent.

Naturally, we may want to use the MDPDE (Basu et al., 1998) for a normal location model family, extended to a multivariate setup. However, estimating the location parameter in this way would force us to estimate the unknown dispersion matrix $\boldsymbol{\Sigma}$ as well, which is already taken care of using the rPCAdpd algorithm. Also, as will be discussed later in Section 3.4.2, this multivariate MDPDE does not satisfy the desirable orthogonal equivariance property, and in particular, the permutation equivariance property. So instead, we can resort to a coordinate-wise MDPDE under the normal location model family. In this case, the coordinates of the estimated location vector satisfy

$$\widehat{\mu}_j = \arg\min_{\mu}\min_{\sigma}\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha}\left[\frac{1}{\sqrt{1+\alpha}} - \left(\frac{1+\alpha}{\alpha}\right)\frac{1}{n}\sum_{i=1}^{n}\exp\left\{-\alpha\frac{(X_{ij}-\mu)^2}{2\sigma^2}\right\}\right],$$

for each $j = 1, \ldots, p$, where α is the robustness parameter lying between 0 and 1, X_{ij} is the j-th coordinate of the sample observation $\boldsymbol{X}_i$. This coordinate-wise MDPDE still retains some of its robustness properties while being permutation and scale equivariant. However, it still does not satisfy the orthogonal equivariance property for a general orthogonal matrix. Detailed properties of such a procedure are present in Chakraborty et al. (2024).

Alternative choices of a robust and consistent estimator of the location parameter would include the L_1 median (Vardi and Zhang, 2000), coordinate-wise median or any M-estimator for location (Huber, 1964a). The L_1 median possesses the desirable orthogonal equivariance property. Based on extensive simulation studies, we have found that L_1 median fits our purpose and provides a desirable balance between speed (computational advantage) and accuracy (robustness and efficiency), and hence it is chosen to be used as a robust location estimator during the rPCAdpd algorithm for all our subsequent studies.

3.3.4 Choice of the Hyperparameters

The two hyperparameters associated with the rPCAdpd estimator are the rank of the $\boldsymbol{L}$ matrix, i.e., the number of significant eigenvalues or the number of principal components to output, and the robustness parameter α in the objective function (3.5). The robustness parameter α can be estimated in a data-driven manner by minimizing a conditional MSE criterion as described before for the rSVDdpd algorithm as in Section 2.2.4.

To determine the rank of the matrix $\boldsymbol{L}$, we robustly estimate all the $(n \wedge p)$ eigenvalues and the corresponding eigenvectors using the rPCAdpd algorithm. Subsequently, we select the rank as the minimum possible value $r \leq (n \wedge p)$, such that the first r eigenvalues can account for at least $(1-\delta)$ proportion of variation present in the sample. Common choices for δ are typically 0.1 or 0.25. Mathematically, the estimate of the rank of the matrix $\boldsymbol{L}$ is given as

$$\widehat{r} = \min \left\{ 1 \leq r \leq (n \wedge p) : \sum_{k=1}^{r} \widehat{\gamma}_k^{(\alpha)} > (1-\delta) \sum_{k=1}^{(n \wedge p)} \widehat{\gamma}_k^{(\alpha)} \right\},$$

where $\widehat{\gamma}_k^{(\alpha)}$ is the k-th eigenvalue as estimated by rPCAdpd method with robustness parameter α. Similar criteria have been used to determine the number of significant principal components by multiple authors (He et al., 2012, Xu et al., 2012). Later in Chapter 3, we discuss this choice in detail.

3.4 Theoretical Properties

3.4.1 Existence and Convergence

Among various theoretical properties of the rPCAdpd estimator, we begin by showing the existence and uniqueness of the estimator. We start by writing the objective function present in (3.5) as a function of the individual terms of the parameter vector $\boldsymbol{\theta}$ as

$$Q(\gamma_1, \ldots, \gamma_p, \boldsymbol{\eta}) = \prod_{k=1}^{p} \gamma_k^{-\alpha/2} \left[\frac{c_{(1+\alpha)g}}{c_g} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^{n} \exp^{\alpha g(\mathrm{qf}(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}))} \right]. \tag{3.9}$$

where $\widehat{\boldsymbol{\mu}}$ is a robust consistent estimate of the location and $\mathrm{qf}(\cdot)$ is as given in (3.6). The following result establishes the existence of the rPCAdpd estimator by showing that the above objective function $Q(\cdot)$ remains bounded below.

Theorem 3.1. *If the generating function $g : [0, \infty) \to \mathbb{R}$ of the elliptically symmetric family of distributions is a decreasing continuous function, then for a sufficiently large number of sample observations n, there exists a minimum of the objective function $Q(\cdot)$ given in (3.9) with probability tending to 1.*

Proof. First note that the eigenvectors $\boldsymbol{v}_k$ lie in the Stiefel manifold of order p, which is a closed and bounded subset of $\mathbb{R}^p$, hence is compact.

Also, since $g(x)$ is a continuous decreasing function, $\lim_{x\to\infty} e^{g(x)} = 0$. Otherwise if $\lim_{x\to\infty} e^{g(x)} = \epsilon > 0$, it implies that the integral $\int_0^\infty e^{g(x)}$ diverges by comparison test, contradicting the existence of the elliptically symmetric probability density function.

Fixing $\boldsymbol{\mu} \in \mathbb{R}^p$, let us now observe how the objective function Q behaves for extreme values of the eigenvalues $\gamma_1, \ldots, \gamma_p$. If $\gamma_1 \to 0$, then it follows that

$$\lim_{\gamma_1\to 0} Q(\gamma_1, \ldots, \gamma_p, \boldsymbol{\eta}) = \lim_{\gamma_1\to 0} \gamma_1^{-1/2} \left[\frac{c_{(1+\alpha)g}}{c_g} - \lim_{x\to\infty} e^{g(x)} \right] \geq 0,$$

since $c_g > 0$ for any choice of g function by definition. On the other hand, if $\gamma_1 \to \infty$, then for every $i = 1, 2, \ldots, n$, the quadratic form

$$(\boldsymbol{X}_i - \boldsymbol{\mu})^\intercal \sum_{k=1}^{p} \gamma_k^{-1} \boldsymbol{v}_k(\boldsymbol{\eta}) \boldsymbol{v}_k(\boldsymbol{\eta})^\intercal (\boldsymbol{X}_i - \boldsymbol{\mu}) \to (\boldsymbol{X}_i - \boldsymbol{\mu})^\intercal \sum_{k=2}^{p} \gamma_k^{-1} \boldsymbol{v}_k(\boldsymbol{\eta}) \boldsymbol{v}_k(\boldsymbol{\eta})^\intercal (\boldsymbol{X}_i - \boldsymbol{\mu}).$$

Then by the law of large numbers, it follows that as $n \to \infty$, with probability 1,

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^{n} \exp\left\{ \alpha g \left((\boldsymbol{X}_i - \boldsymbol{\mu})^\intercal \sum_{k=2}^{p} \gamma_k^{-1} \boldsymbol{v}_k(\boldsymbol{\eta}) \boldsymbol{v}_k(\boldsymbol{\eta})^\intercal (\boldsymbol{X}_i - \boldsymbol{\mu}) \right) \right\} \\
\to\ & \mathbb{E}\left[\exp\left\{ \alpha g \left((\boldsymbol{X} - \boldsymbol{\mu})^\intercal \sum_{k=2}^{p} \gamma_k^{-1} \boldsymbol{v}_k(\boldsymbol{\eta}) \boldsymbol{v}_k(\boldsymbol{\eta})^\intercal (\boldsymbol{X} - \boldsymbol{\mu}) \right) \right\} \right] \\
\geq\ & \mathbb{E}\left[\exp\left\{ \alpha g \left((\boldsymbol{X} - \boldsymbol{\mu})^\intercal \sum_{k=1}^{p} \gamma_k^{-1} \boldsymbol{v}_k(\boldsymbol{\eta}) \boldsymbol{v}_k(\boldsymbol{\eta})^\intercal (\boldsymbol{X} - \boldsymbol{\mu}) \right) \right\} \right] \\
=\ & c_g^{-1} \int_{\mathbb{R}^p} \exp\left\{ (1+\alpha) g \left((\boldsymbol{x} - \boldsymbol{\mu})^\intercal \sum_{k=1}^{p} \gamma_k^{-1} \boldsymbol{v}_k(\boldsymbol{\eta}) \boldsymbol{v}_k(\boldsymbol{\eta})^\intercal (\boldsymbol{x} - \boldsymbol{\mu}) \right) \right\} d\boldsymbol{x} \\
=\ & \frac{c_{(1+\alpha)g}}{c_g},
\end{aligned}$$

where the inequality on the third line uses the fact that g is monotonically decreasing. Therefore, for sufficiently large n, with probability 1, $Q(\gamma_1, \ldots, \gamma_p, \boldsymbol{\eta})$ increases to 0 as γ_1 increases to ∞. Therefore, for any given $\epsilon > 0$, there exists $0 < a_1(\epsilon) < b_1(\epsilon) < \infty$ such that $Q(\gamma_1, \gamma_2, \ldots, \gamma_p) > (-\epsilon)$ for any $\gamma_1 \notin [a_1(\epsilon), b_1(\epsilon)]$. Note that, since γ_1 is chosen arbitrarily, the same conclusion also holds for all other eigenvalues, possibly with different choices of $a_k(\cdot)$ and $b_k(\cdot)$ for $k = 1, 2, \ldots, p$. Now consider the set $K = \prod_{k=1}^{p} [a_k(1), b_k(1)] \times S_p^p$, which is clearly compact. If $\boldsymbol{\theta} \notin K$, then by the above construction we must have $Q(\boldsymbol{\theta}) > (-1)$. If $\boldsymbol{\theta} \in K$, then by continuity of $Q(\cdot)$ and compactness of K, we get that $Q(\boldsymbol{\theta})$ must be bounded below. Therefore, $\inf Q(\gamma_1, \ldots, \gamma_p, \boldsymbol{\eta}) = q_0$ must exist and $q_0 < 0$.

Now let $\epsilon = -q_0/2 > 0$ and consider $K' = \prod_{k=1}^{p} [a_k(-q_0/2), b_k(-q_0/2)] \times S_p^p$. Note that, the infimum q_0 must exist within this set K'. Since K' is a compact subset of $\mathbb{R}^p$, it

follows by the Extreme Value Theorem that the infimum must be attained. This proves that the rPCAdpd estimator exists for any arbitrary value of μ, including the location estimate $\widehat{\mu}$. □

The condition that the generating function g in the elliptically symmetric distribution is monotonically decreasing and continuous is not a strict imposition. For instance, for all the popular choices of the elliptically symmetric densities given in Table 3.1, the generating function g is a decreasing continuous function. Hence, the rPCAdpd estimator exists for all these multivariate families of distributions.

Once the existence of the rPCAdpd estimator is established, the convergence of the iterative algorithm follows directly from the convergence of the rSVDdpd algorithm as presented in Theorem 2.1 in Chapter 2.

3.4.2 Orthogonal Equivariance

As mentioned in Rousseeuw (1985), orthogonal equivariance is one of the fundamental properties that an estimator of principal components should possess. Let, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be a transformed sample $\boldsymbol{Y}_i = a\boldsymbol{P}\boldsymbol{X}_i + \boldsymbol{b}$ for $i = 1, 2, \ldots, n$, where $\boldsymbol{P}_{p\times p}$ is an orthogonal matrix, $a \in (0, \infty)$ and $\boldsymbol{b}$ is a p-length vector. Then, an orthogonally equivariant estimator $T_\lambda(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ of an eigenvalue should satisfy $T_\lambda(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) = a^2 T_\lambda(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. Similarly, for an orthogonally equivariant estimate $T_{\boldsymbol{v}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ of the corresponding eigenvector, it should satisfy $T_{\boldsymbol{v}}(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) = \boldsymbol{P}T_{\boldsymbol{v}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. For any orthogonal equivariant estimate of the principal components, both of these two conditions should hold for all eigenvalues and their corresponding eigenvectors.

Given the orthogonal equivariance property of the location estimator, it follows that the resulting rPCAdpd estimator also satisfies the same equivariance property. The choice of L_1-median as a robust estimator of location satisfies this property.

Theorem 3.2. *The rPCAdpd estimators of the eigenvalues and eigenvectors are equivariant under the transformation*

$$\boldsymbol{Y}_i = a\boldsymbol{P}\boldsymbol{X}_i + \boldsymbol{b}, \;\; i = 1, 2, \ldots, n, \tag{3.10}$$

where $\boldsymbol{P}_{p\times p}$ is an orthogonal matrix, $a \in (0, \infty)$ and $\boldsymbol{b}$ is a p-length vector provided that the location estimator used in the rPCAdpd procedure also satisfy same equivariance property, i.e.,

$$\widehat{\mu}(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) = a\boldsymbol{P}\widehat{\mu}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) + \boldsymbol{b}.$$

Proof. Let $\widehat{\boldsymbol{\mu}}_Y$ and $\widehat{\boldsymbol{\mu}}_X$ be the robust estimates of the location based on the sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ respectively. Then by the orthogonal equivariance of the location estimator, we have that $\widehat{\boldsymbol{\mu}}_Y = a\boldsymbol{P}\widehat{\boldsymbol{\mu}}_X + \boldsymbol{b}$. The equivariance property for the estimated eigenvalues and eigenvectors by the rPCAdpd algorithm then follows from the

observation that the quadratic form of the transformed data can be expressed as

$$\begin{aligned}\text{qf}(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_Y) &= a(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_X)^\intercal \boldsymbol{P}^\intercal \left(\sum_{k=1}^{p} \gamma_k^{-1} \boldsymbol{v}_k \boldsymbol{v}_k^\intercal\right) (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_X)\boldsymbol{P}a \\ &= (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_X)^\intercal \left(\sum_{k=1}^{p} (\gamma_k/a^2)^{-1} \boldsymbol{P}^\intercal \boldsymbol{v}_k \boldsymbol{v}_k^\intercal \boldsymbol{P}\right) (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_X).\end{aligned}$$

Now if rPCAdpd estimates of the eigenvalues for the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are γ_k^* for $k = 1, 2, \ldots, p$ respectively, then it is the minimizer of the objective function $Q(\cdot)$ as in (3.9). Equivalently, $a^2\gamma_k^*$ for $k = 1, \ldots, p$ becomes the minimizer for the Q-function for the transformed sample observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ due to the equivalence of the quadratic form as shown above. Therefore, the rPCAdpd estimate of the same for the transformed sample would be $a^2\gamma_k^*$.

A similar conclusion can be drawn for the rPCAdpd estimates of the eigenvectors. □

Corollary 3.1. *As in the case of the rSVDdpd estimator discussed in Section 2.3.2, the rPCAdpd estimator also satisfies scale and permutation equivariance. This follows from the observation that both are special cases of the transformation mentioned in Eq. (3.10). In particular, with $\boldsymbol{P} = \boldsymbol{I}_p$, the identity matrix of order p, we get scale equivariance. Also, if $a = 1$ and $\boldsymbol{P}$ is a permutation matrix, then permutation equivariance follows.*

3.4.3 Consistency and Asymptotic Distribution

One of the integral components of the proposed rPCAdpd estimator is the MDPDE. As shown in Basu et al. (1998), the MDPDE, being an M-estimator and a minimum distance estimator, enjoys a vast set of nice asymptotic properties including consistency and asymptotic normality. In this subsection, we will investigate how these properties carry over to the special scenario of principal component estimation under elliptically symmetric models.

Throughout this entire subsection, we assume that the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ consists of i.i.d. random observations from a p-variate elliptically symmetric distribution with generating function g as in (3.1) with unknown mean $\boldsymbol{\mu}^*$ and unknown dispersion matrix $\boldsymbol{\Sigma}^*$. The covariance matrix $\boldsymbol{\Sigma}^*$ is assumed to have an eigen-decomposition $\boldsymbol{\Sigma}^* = \sum_{k=1}^{p} \gamma_k^* \boldsymbol{v}_k^* (\boldsymbol{v}_k^*)^\intercal$ where $\gamma_k^* \geq 0$ are eigenvalues and $\boldsymbol{v}_k^*$s are the corresponding eigenvectors of the covariance matrix. We wish to estimate the parameter of interest $\boldsymbol{\theta}^* = (\gamma_1^*, \ldots, \gamma_p^*, \boldsymbol{\eta}^*)$, comprising of the eigenvalue $\gamma_1^*, \ldots, \gamma_p^*$ and the natural parameter $\boldsymbol{\eta}^*$ parametrizing the eigenvectors in the Stiefel manifold S_p^p. The location parameter $\boldsymbol{\mu}^*$ is a nuisance parameter in this setup as indicated in Section 3.3.1.

Following the footsteps of Basu et al. (1998), we consider the following key quantities

$$\begin{aligned}\boldsymbol{\xi}_{\boldsymbol{\theta}} &= \int u_{\boldsymbol{\theta}}(\boldsymbol{x}) f_{\boldsymbol{\theta}}^{(1+\alpha)} dx, \\ \boldsymbol{J}_{\boldsymbol{\theta}} &= \int u_{\boldsymbol{\theta}}(\boldsymbol{x}) u_{\boldsymbol{\theta}}(\boldsymbol{x})^\intercal f_{\boldsymbol{\theta}}^{(1+\alpha)} dx, \\ \boldsymbol{K}_{\boldsymbol{\theta}} &= \int u_{\boldsymbol{\theta}}(\boldsymbol{x}) u_{\boldsymbol{\theta}}(\boldsymbol{x})^\intercal f_{\boldsymbol{\theta}}^{(1+2\alpha)} dx - \boldsymbol{\xi}_{\boldsymbol{\theta}} \boldsymbol{\xi}_{\boldsymbol{\theta}}^\intercal,\end{aligned}$$

which are essential for obtaining different asymptotic properties of the MDPDE. Here, $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ denotes the same elliptical family of densities as in (3.4) at parameter $\boldsymbol{\theta}$ and corresponding score function is denoted by $u_{\boldsymbol{\theta}}(\boldsymbol{x}) = \partial_{\boldsymbol{\theta}} \log(f_{\boldsymbol{\theta}}(\boldsymbol{x}))$. To calculate all of these quantities, we will resort to the following assumptions.

(C1) The generating function $g(\cdot)$ for the elliptically symmetric family of distributions is thrice differentiable and the third order derivative is continuous.

(C2) The true eigenvalues $\gamma_1^*, \ldots, \gamma_p^*$ are distinct.

(C3) The functions $s^2 g'(s) e^{g(s)}$, $s^4 (g'(s))^2 e^{g(s)}$, $s^4 g''(s) e^{g(s)}$ and $s^4 g'''(s) e^{g(s)}$ are uniformly bounded above by some constant M^* for any $s \geq 0$, where $g'(s), g''(s)$ and $g'''(s)$ denotes the first, second and third order derivatives of g.

Assumptions (C1) and (C3) are similar in spirit to Assumptions (R1) and (R2) of Ghosh and Basu (2013), which in turn imply Assumptions (A1)-(A5) of Basu et al. (1998). One of the standard regularity conditions for such asymptotic results is the exchangeability of the differentiation and integral signs, i.e., the integral $\int f_{\boldsymbol{\theta}}^{(1+\alpha)}(\boldsymbol{z}) d\boldsymbol{z}$ should be differentiable with respect to $\boldsymbol{\theta}$ for any $\alpha \in [0, 1]$ and the derivative can be taken under the integral sign. However, this fact follows as a consequence of Assumption (C1) for the elliptically symmetric family of distributions. Assumption (C2) makes the calculation simpler, but it is not strictly necessary to establish the asymptotic properties of the rPCAdpd estimator. However, it is also known that a random matrix with i.i.d. entries has negligible probability to have repeated eigenvalues (Tao, 2012). Kumar and Ahmed (2017) verify similar conclusions for a broader range of distribution of random matrices using simulations. Thus, Assumption (C2) holds for almost all positive definite matrices $\boldsymbol{\Sigma}^*$.

We begin with two generic lemmas describing the quantity $\boldsymbol{\xi}_{\boldsymbol{\theta}}$ and $\boldsymbol{J}_{\boldsymbol{\theta}}$ as a function of the integral of the model density function and its derivatives. These lemmas are generic; they are applicable in any MDPDE setup, not only in particular to the elliptically symmetric family of density or the RPCA problem. Hence, we present only the statement of these Lemmas here, while we defer their proofs till Appendices 3.A.2 and 3.A.3.

Lemma 3.1. *For any general setup with thrice differentiable model density function $f_{\boldsymbol{\theta}}(x)$, let $c_\alpha(\boldsymbol{\theta}) = \|f_{\boldsymbol{\theta}}\|_{1+\alpha}^{1+\alpha}$. Then, under the assumption of exchangeability of the differentiation and integral signs, we have*

$$\boldsymbol{\xi}_{\boldsymbol{\theta}} = (1+\alpha)^{-1} c_\alpha(\boldsymbol{\theta}) \partial_{\boldsymbol{\theta}} \log(c_\alpha(\boldsymbol{\theta})).$$

Lemma 3.2. *For any general setup with thrice differentiable model density function $f_{\boldsymbol{\theta}}(x)$, let $c_\alpha(\boldsymbol{\theta}) = \|f_{\boldsymbol{\theta}}\|_{1+\alpha}^{1+\alpha}$. Then, under the assumption of exchangeability of the differentiation and integral signs, we have*

$$\boldsymbol{J}_{\boldsymbol{\theta}} = \frac{c_\alpha(\boldsymbol{\theta})}{(1+\alpha)^2} \left(i^h(\boldsymbol{\theta}) + \left(\frac{\partial}{\partial \boldsymbol{\theta}} \log(c_\alpha(\boldsymbol{\theta})) \right) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \log(c_\alpha(\boldsymbol{\theta})) \right)^\intercal \right)$$

where $i^h(\boldsymbol{\theta})$ is the expected Fisher information matrix for a single observation $\boldsymbol{x}$ following the density function $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = c_\alpha^{-1}(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}^{(1+\alpha)}(\boldsymbol{x})$.

Before proceeding with the computation of these quantities $\boldsymbol{\xi}_{\boldsymbol{\theta}}, \boldsymbol{J}_{\boldsymbol{\theta}}$ and $\boldsymbol{K}_{\boldsymbol{\theta}}$ for the particular setup of the rPCAdpd estimator, we recognize that the estimation of the principal components is essentially a two-step procedure. In the first step, we use an initial consistent robust estimator $\widehat{\boldsymbol{\mu}}$ to estimate the location parameter $\boldsymbol{\mu}^*$. In the next step, the rSVDdpd procedure was used to obtain the MDPDE of $\boldsymbol{\theta}$ using the model family densities $f_{\boldsymbol{\theta}}(x)$ as in (3.4) by replacing $\boldsymbol{\mu}^*$ with its estimate $\widehat{\boldsymbol{\mu}}$ from the first step. Therefore, in the following, we compute the quantities $\boldsymbol{\xi}_{\boldsymbol{\theta}}, \boldsymbol{J}_{\boldsymbol{\theta}}$ and $\boldsymbol{K}_{\boldsymbol{\theta}}$ conditional on the value of $\widehat{\boldsymbol{\mu}}$, which will lead to the conditional asymptotic distribution of $\widehat{\boldsymbol{\theta}}$. However, as we shall show later in Theorem 3.4, this conditional distribution turns out to be free of $\widehat{\boldsymbol{\mu}}$, hence the unconditional asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ will also remain the same.

We start by using Lemma 3.1 in combination with Assumption (C2) for our specific use case. To compactly write $\boldsymbol{\xi}_{\boldsymbol{\theta}^*}$, we introduce the diagonal matrix $\boldsymbol{\Gamma}_{p\times p}$ with the diagonal entries equal to $\gamma_1^*, \ldots, \gamma_p^*$.

Corollary 3.2. *If $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is a density function belonging to an elliptically symmetric family of distributions with generating function $g(\cdot)$ as shown in (3.4), then under the Assumptions (C1)-(C3) when the location parameter μ^* is a fixed quantity,*

$$\boldsymbol{\xi}_{\boldsymbol{\theta}^*} = \frac{c_{(1+\alpha)g}}{(1+\alpha)(c_g)^{(1+\alpha)}} \prod_{k=1}^{p} (\gamma_k^*)^{-\alpha/2} \begin{bmatrix} -\frac{\alpha}{2}\text{Diag}\left(\boldsymbol{\Gamma}^{-1}\right) \\ \boldsymbol{0} \end{bmatrix}.$$

Proof. Since the normalized density $c_\alpha^{-1}(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}^{(1+\alpha)}$ also belongs to an elliptically symmetric class of densities, it follows that

$$c_\alpha(\boldsymbol{\theta}) = c_{(1+\alpha)g} \prod_{k=1}^{p} (\gamma_k)^{1/2} c_g^{-(1+\alpha)} \prod_{k=1}^{p} (\gamma_k)^{-(1+\alpha)/2} = \frac{c_{(1+\alpha)g}}{c_g^{1+\alpha}} \prod_{k=1}^{p} (\gamma_k)^{-\alpha/2}.$$

Putting this value and its derivative with respect to $\boldsymbol{\theta}$ into Lemma 3.1 completes the proof. □

Similarly, we can obtain the expression of $\boldsymbol{J}_{\boldsymbol{\theta}^*}$ for the elliptically symmetric family of densities as given by the following Corollary.

Corollary 3.3. *If $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is a density function belonging to an elliptically symmetric family of distributions with generating function $g(\cdot)$ as given in (3.4), then under the Assumptions (C1)-(C3) when the location parameter μ^* is a fixed quantity,*

$$\boldsymbol{J}_{\boldsymbol{\theta}^*} = \frac{c_{(1+\alpha)g}}{(1+\alpha)^2 c_g^{(1+\alpha)}} \prod_{k=1}^{p} (\gamma_k^*)^{-\alpha/2} \widetilde{\boldsymbol{J}}_{\boldsymbol{\theta}^*},$$

where

$$\widetilde{\boldsymbol{J}}_{\boldsymbol{\theta}^*} = \begin{bmatrix} i^h(\boldsymbol{\gamma},\boldsymbol{\gamma}) + \frac{\alpha^2}{4}\left(\text{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)\right)\left(\text{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)\right)^\intercal & i^h(\boldsymbol{\gamma},\boldsymbol{\eta}) \\ i^h(\boldsymbol{\gamma},\boldsymbol{\eta})^\intercal & i^h(\boldsymbol{\eta},\boldsymbol{\eta}) \end{bmatrix}$$

The quantities $i^h(\cdot,\cdot)$ are given by the following formulae

$$\begin{aligned} i^h(\boldsymbol{\gamma},\boldsymbol{\gamma}) &= -\frac{1}{4}\left(\text{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)\right)\left(\text{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)\right)^\intercal + \boldsymbol{\Gamma}^{-2}\boldsymbol{V}^\intercal A_4((1+\alpha)g)\boldsymbol{V}\boldsymbol{\Gamma}^{-2}, \\ i^h(\boldsymbol{\gamma},\boldsymbol{\eta}) &= -2\boldsymbol{\Gamma}^{-2}\boldsymbol{V}^\intercal(\boldsymbol{I}_p \otimes \boldsymbol{\Gamma}^{-1})A_4((1+\alpha)g)\boldsymbol{G}^\intercal, \\ i^h(\boldsymbol{\eta},\boldsymbol{\eta}) &= 4\boldsymbol{G}(\boldsymbol{I}_p \otimes \boldsymbol{\Gamma}^{-1})A_4((1+\alpha)g)(\boldsymbol{I}_p \otimes \boldsymbol{\Gamma}^{-1})^\intercal\boldsymbol{G}^\intercal, \end{aligned}$$

where

$$\boldsymbol{V}_{p^2\times p} = \begin{bmatrix} \boldsymbol{v}_1^* & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{v}_2^* & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{v}_p^* \end{bmatrix},$$
$$\boldsymbol{G}_{p(p+1)/2\times p^2} = \left[\frac{\partial \boldsymbol{v}_1}{\partial \boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\boldsymbol{\eta}^*} \quad \frac{\partial \boldsymbol{v}_2}{\partial \boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\boldsymbol{\eta}^*} \quad \dots \quad \frac{\partial \boldsymbol{v}_p}{\partial \boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\boldsymbol{\eta}^*} \right]^\intercal,$$

and $A_4(g)$ *is the* $p^2 \times p^2$ *matrix comprising of the partitions* $A_4(g; \boldsymbol{v}_i^*, \boldsymbol{v}_j^*)$ *for* $i, j = 1, 2, \dots, p$*, where*

$$A_4(g; \boldsymbol{u}, \boldsymbol{v}) = \prod_{k=1}^{p} \gamma_k^{-1/2} \int \left(g' \left(Q \left(\boldsymbol{x}^\intercal \sum_{k=1}^{p} (\gamma_k^*)^{-1} \boldsymbol{v}_k^* (\boldsymbol{v}_k^*)^\intercal \boldsymbol{x} \right) \right) \right)^2 \boldsymbol{x}\boldsymbol{x}^\intercal \boldsymbol{u}\boldsymbol{v}^\intercal \boldsymbol{x}\boldsymbol{x}^\intercal$$
$$c_g^{-1} e^{g\left(\boldsymbol{x}^\intercal \sum_{k=1}^{p} (\gamma_k^*)^{-1} \boldsymbol{v}_k^* (\boldsymbol{v}_k^*)^\intercal \boldsymbol{x}\right)} d\boldsymbol{x}.$$

Proof. We begin by defining a few notations as follows,

$$Q(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}^*)^\intercal \left(\sum_{k=1}^{p} \gamma_k^{-1} \boldsymbol{v}_k \boldsymbol{v}_k^\intercal \right) (\boldsymbol{x} - \boldsymbol{\mu}^*),$$
$$A_2(g) = \int g'(Q(\boldsymbol{x}))(\boldsymbol{x} - \boldsymbol{\mu}^*)(\boldsymbol{x} - \boldsymbol{\mu}^*)^\intercal c_1(\boldsymbol{\theta})^{-1} \exp(g(Q(\boldsymbol{x})))d\boldsymbol{x},$$

where

$$c_1(\boldsymbol{\theta}) = c_g \prod_{k=1}^{p} \gamma_k^{1/2}.$$

Both of these quantities are well defined due to the Assumptions (C1) and (C3). Now, starting with the identity

$$c_1(\boldsymbol{\theta}) = \int \exp\left(g(Q(\boldsymbol{x}))\right) d\boldsymbol{x},$$

and differentiating both sides by γ_k and $\boldsymbol{\eta}$ respectively, we obtain the identities

$$\gamma_k^{-2} \boldsymbol{v}_k^\intercal A_2(g) \boldsymbol{v}_k = -\frac{1}{2\gamma_k}, \ \sum_{k=1}^{p} \gamma_k^{-1} \boldsymbol{G}_k A_2(g) \boldsymbol{v}_k = 0, \tag{3.11}$$

both of which will be used later in the proof.

Let $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = c_{(1+\alpha)}^{-1}(\boldsymbol{\theta}) e^{(1+\alpha)g(Q(\boldsymbol{x}))}$ be a density belonging to the same elliptically symmetric family. Then, the score function $u_{\boldsymbol{\theta}}^h(\boldsymbol{x})$ corresponding to $h_{\boldsymbol{\theta}}$ turns out as

$$u_{\boldsymbol{\theta}}^h(\boldsymbol{x}) = \begin{bmatrix} \frac{1}{2}\text{Diag}\left(\boldsymbol{\Gamma}^{-1}\right) - (1+\alpha) g'(Q(\boldsymbol{x})) \boldsymbol{\Gamma}^{-2} \boldsymbol{V}^\intercal (\boldsymbol{I}_p \otimes (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu}^*)^\intercal) \boldsymbol{V} \\ 2(1+\alpha) g'(Q(\boldsymbol{x})) \boldsymbol{G} (\boldsymbol{\Gamma}^{-1} \otimes (\boldsymbol{x} - \boldsymbol{\mu}^*)(\boldsymbol{x} - \boldsymbol{\mu}^*)^\intercal) \boldsymbol{V} \mathbf{1}_p \end{bmatrix}.$$

Using the expression for $u_{\boldsymbol{\theta}}^h(\boldsymbol{x})$, we can further differentiate this with respect to the entries

of $\boldsymbol{\theta}$ and take expectation. This leads to the Fisher Information matrix $i^h(\boldsymbol{\theta})$ with entries,

$$\begin{aligned} i^h(\gamma_k, \gamma_l) &= \left(\partial_{\gamma_k} q_{(1+\alpha)g}\right)\left(\partial_{\gamma_l} q_{(1+\alpha)g}\right) + \left(\partial_{\gamma_k} q_{(1+\alpha)g}\right)\gamma_l^{-2}\boldsymbol{v}_l^\intercal A_2((1+\alpha)g)\boldsymbol{v}_l \\ &\quad + \left(\partial_{\gamma_l} q_{(1+\alpha)g}\right)\gamma_k^{-2}\boldsymbol{v}_k^\intercal A_2((1+\alpha)g)\boldsymbol{v}_k + \frac{\boldsymbol{v}_k^\intercal A_4((1+\alpha)g;\boldsymbol{v}_k,\boldsymbol{v}_l)\boldsymbol{v}_l}{\gamma_k^2\gamma_l^2} \\ &= -\left(\partial_{\gamma_k} q_{(1+\alpha)g}\right)\left(\partial_{\gamma_l} q_{(1+\alpha)g}\right) + \frac{\boldsymbol{v}_k^\intercal A_4((1+\alpha)g;\boldsymbol{v}_k,\boldsymbol{v}_l)\boldsymbol{v}_l}{\gamma_k^2\gamma_l^2}, \end{aligned}$$

for each $k, l = 1, 2, \ldots, p$; and

$$\begin{aligned} i^h(\gamma_k, \boldsymbol{\eta}) &= -2\left(\partial_{\gamma_k} q_{(1+\alpha)g}\right)\sum_{k=1}^{p}\boldsymbol{G}_k\frac{A_2((1+\alpha)g)\boldsymbol{v}_k}{\gamma_k} - \frac{2}{\gamma_k^2}\sum_{l=1}^{p}\frac{\boldsymbol{v}_k^\intercal A_4((1+\alpha)g;\boldsymbol{v}_k,\boldsymbol{v}_l)}{\gamma_l}\boldsymbol{G}_l^\intercal \\ &= -\frac{2}{\gamma_k^2}\sum_{l=1}^{p}\gamma_l^{-1}\boldsymbol{v}_k^\intercal A_4((1+\alpha)g;\boldsymbol{v}_k,\boldsymbol{v}_l)\boldsymbol{G}_l^\intercal \\ &\qquad \text{for each } k = 1, 2, \ldots, p, \\ i^h(\boldsymbol{\eta}, \boldsymbol{\eta}) &= 4\sum_{k=1}^{p}\sum_{l=1}^{p}\gamma_k^{-1}\gamma_l^{-1}\boldsymbol{G}_k A_4((1+\alpha)g;\boldsymbol{v}_k,\boldsymbol{v}_l)\boldsymbol{G}_l^\intercal, \end{aligned}$$

where we use the identities given in (3.11). In all of the above expressions, the quantity q_g denoted the logarithm of the normalizing constant, i.e., $q_g = \log(c_1(\boldsymbol{\theta}))$ and $q_{(1+\alpha)g} = \log(c_\alpha(\boldsymbol{\theta}))$. Finally, Corollary 3.3 follows from using Lemma 3.2 and the expression of $\boldsymbol{\xi}_{\boldsymbol{\theta}}$ given in Corollary 3.2. □

For the particular setup of robust estimation of the principal components for the elliptically symmetric family, Assumptions (C1)-(C3) imply all the necessary assumptions (A1)-(A5) of Basu et al. (1998). Thus, we can readily use Theorem 2.2 of the same to establish the asymptotic properties such as consistency and the asymptotic normality of the converged rPCAdpd estimator of the principal components. However, since the quantities $\boldsymbol{\xi}_{\boldsymbol{\theta}}, \boldsymbol{J}_{\boldsymbol{\theta}}$ are obtained for a fixed value of $\widehat{\boldsymbol{\mu}}$, the resulting asymptotic normal distribution is also obtained conditional on the values of $\widehat{\boldsymbol{\mu}}$. However, the conditional asymptotic distribution is independent of $\widehat{\boldsymbol{\mu}}$, hence the unconditional distribution also turns out to be the same. For further technical details, one may refer to the proof of Theorem 3.4.

Theorem 3.3. *Suppose the Assumptions (C1)-(C3) hold, $\alpha \in [0, 1]$ and the location estimator $\widehat{\boldsymbol{\mu}}$ is consistent for $\boldsymbol{\mu}^*$. Then, as the sample size $n \to \infty$, there exists a sequence of rPCAdpd estimators $\widehat{\boldsymbol{\theta}}_n = ((\widehat{\gamma}_1)_n, \ldots, (\widehat{\gamma}_p)_n, \widehat{\boldsymbol{\eta}}_n)$ as given in (3.5) such that*

1. *The estimated eigenvalue $(\widehat{\gamma}_j)_n$ is $\sqrt{n}$-consistent for γ_j^* for $j = 1, 2, \ldots, p$.*
2. *Similarly, the corresponding estimated eigenvector $(\widehat{\boldsymbol{v}}_j)_n$ is also $\sqrt{n}$-consistent for the true eigenvector $\boldsymbol{v}_j^*$ for $j = 1, 2, \ldots, p$.*

Remark 3.1. *The consistency of $(\widehat{\boldsymbol{v}}_j)_n$ for $\boldsymbol{v}_j^*$ follows from the fact that $\widehat{\boldsymbol{\eta}}_n$ is consistent for $\boldsymbol{\eta}^*$ and the continuous mapping theorem, as the parameter $\boldsymbol{\eta}$ is a smooth parametrization of the Stiefel manifold.*

Theorem 3.4. *Suppose the Assumptions (C1)-(C3) hold and the location estimator $\widehat{\boldsymbol{\mu}}$ is n^c-consistent for $\boldsymbol{\mu}^*$ for $c \geq 1/2$. Then, there exists a sequence of converged rPCAdpd estimator $\widehat{\boldsymbol{\theta}}_n = ((\widehat{\gamma}_1)_n, \ldots, (\widehat{\gamma}_p)_n, \widehat{\boldsymbol{\eta}}_n)$ as given in (3.5) for the general elliptically symmetric family such that after proper centering and scaling, it has an asymptotic normal distribution as $n \to \infty$. In particular,*

$$\sqrt{n}\boldsymbol{J}_{\boldsymbol{\theta}^*}\boldsymbol{K}_{\boldsymbol{\theta}^*}^{-1/2}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right)$$

converges in distribution to a standard normal random variable as $n \to \infty$. Here,

$$\begin{aligned}
\boldsymbol{J}_{\boldsymbol{\theta}^*} &= \frac{c_{(1+\alpha)g}}{(1+\alpha)^2 c_g^{(1+\alpha)}}\begin{bmatrix}\boldsymbol{J}_{11} & \boldsymbol{J}_{12}\\ \boldsymbol{J}_{12}^\intercal & \boldsymbol{J}_{22}\end{bmatrix},\\
\boldsymbol{J}_{11} &= \frac{(\alpha^2-1)}{4}\left(\mathrm{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)\right)\left(\mathrm{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)\right)^\intercal + \boldsymbol{\Gamma}^{-2}\boldsymbol{V}^\intercal\boldsymbol{A}_4((1+\alpha)g)\boldsymbol{V}\boldsymbol{\Gamma}^{-2},\\
\boldsymbol{J}_{12} &= -2\boldsymbol{\Gamma}^{-2}\boldsymbol{V}^\intercal\left(\boldsymbol{I}_p\otimes\boldsymbol{\Gamma}^{-1}\right)A_4((1+\alpha)g)\boldsymbol{G}^\intercal,\\
\boldsymbol{J}_{22} &= 4\boldsymbol{G}\left(\boldsymbol{I}_p\otimes\boldsymbol{\Gamma}^{-1}\right)A_4((1+\alpha)g)\left(\boldsymbol{I}_p\otimes\boldsymbol{\Gamma}^{-1}\right)^\intercal\boldsymbol{G}^\intercal,
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{K}_{\boldsymbol{\theta}^*} &= \frac{c_{(1+2\alpha)g}}{(1+2\alpha)^2 c_g^{(1+2\alpha)}}\begin{bmatrix}\boldsymbol{K}_{11} & \boldsymbol{K}_{12}\\ \boldsymbol{K}_{12}^\intercal & \boldsymbol{K}_{22}\end{bmatrix} - \frac{c_{(1+\alpha)g}^2}{(1+\alpha)^2 c_g^{(2+2\alpha)}}\begin{bmatrix}\frac{\alpha^2}{4}\mathrm{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)\mathrm{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)^\intercal & 0\\ 0 & 0\end{bmatrix},\\
\boldsymbol{K}_{11} &= \frac{(4\alpha^2-1)}{4}\left(\mathrm{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)\right)\left(\mathrm{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)\right)^\intercal + \boldsymbol{\Gamma}^{-2}\boldsymbol{V}^\intercal\boldsymbol{A}_4((1+2\alpha)g)\boldsymbol{V}\boldsymbol{\Gamma}^{-2},\\
\boldsymbol{K}_{12} &= -2\boldsymbol{\Gamma}^{-2}\boldsymbol{V}^\intercal\left(\boldsymbol{I}_p\otimes\boldsymbol{\Gamma}^{-1}\right)A_4((1+2\alpha)g)\boldsymbol{G}^\intercal,\\
\boldsymbol{K}_{22} &= 4\boldsymbol{G}\left(\boldsymbol{I}_p\otimes\boldsymbol{\Gamma}^{-1}\right)A_4((1+2\alpha)g)\left(\boldsymbol{I}_p\otimes\boldsymbol{\Gamma}^{-1}\right)^\intercal\boldsymbol{G}^\intercal.
\end{aligned}$$

Proof. The proof of this closely resembles the proof of Theorem 3.1 of Ghosh and Basu (2013). For brevity, we shall only indicate the modifications pertinent to the special scenario of principal components. Given the location estimator $\widehat{\boldsymbol{\mu}}$, using the same notation as in Ghosh and Basu (2013), we define

$$V(\boldsymbol{X},\boldsymbol{\theta}) = \prod_{k=1}^{p}\gamma_k^{-\alpha/2}\left[\frac{c_{(1+\alpha)g}}{c_g} - \left(1+\frac{1}{\alpha}\right)e^{\alpha g\left((\boldsymbol{X}-\widehat{\boldsymbol{\mu}})^\intercal\sum_{k=1}^{p}\gamma_k^{-1}\boldsymbol{v}_k(\boldsymbol{\eta})\boldsymbol{v}_k(\boldsymbol{\eta})^\intercal(\boldsymbol{X}-\widehat{\boldsymbol{\mu}})\right)}\right]$$

which are the summands in the objective function present in (3.9). Now, conditional on $\widehat{\boldsymbol{\mu}}$, by an application of the Law of Large Numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n}\partial_{\boldsymbol{\theta}}V(\boldsymbol{X}_i,\boldsymbol{\theta}^*)\Big|\widehat{\boldsymbol{\mu}} \xrightarrow{P} 0, \text{ and, } \frac{1}{n}\sum_{i=1}^{n}\partial_{\boldsymbol{\theta}}^2V(\boldsymbol{X}_i,\boldsymbol{\theta}^*)\Big|\widehat{\boldsymbol{\mu}} \xrightarrow{P} \boldsymbol{J}_{\boldsymbol{\theta}^*}$$

where $\boldsymbol{\theta}^*$ is the true value of the parameters and the symbol $\xrightarrow{P}$ denotes convergence in probability. Now, since the right-hand sides of both of these are continuous functions of $\widehat{\boldsymbol{\mu}}$ and as $\widehat{\boldsymbol{\mu}} \xrightarrow{P} \boldsymbol{\mu}^*$ (the true location parameter) due to the consistency of the location estimator, it follows that the unconditional random variables also converges in probability to the same value. As the support of the elliptically symmetric family of distributions is assumed to be the entire space $\mathbb{R}^p$, $\boldsymbol{J}_{\boldsymbol{\theta}^*}$ becomes free of the choice of location which makes

this convergence possible. Now, one can replicate the proof for consistency to show that the rPCAdpd estimator is consistent.

To prove the asymptotic normality, we need to show that the average Hessian $T_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \partial_{\boldsymbol{\theta}}^2 V(\boldsymbol{X}_i, \boldsymbol{\theta}^*)$ converges in distribution to a random variable $\boldsymbol{Z}$ following a multivariate normal distribution with mean 0 and variance $\boldsymbol{K}_{\boldsymbol{\theta}^*}$. Due to Portmanteau's theorem, it is enough to show that for any bounded continuous function h, $|\mathbb{E}(h(T_n)) - \mathbb{E}(h(\boldsymbol{Z}))| \to 0$ as $n \to \infty$. Since both the mean and the variance of $\boldsymbol{Z}$ are free of the choice of location, we have $\mathbb{E}(h(\boldsymbol{Z}) \mid \widehat{\boldsymbol{\mu}}) = \mathbb{E}(h(\boldsymbol{Z}))$. Also, an application of Lindeberg-Levy Central Limit Theorem and Portmanteau's theorem yields that for any $\widehat{\boldsymbol{\mu}}$,

$$|\mathbb{E}(h(T_n) \mid \widehat{\boldsymbol{\mu}}) - \mathbb{E}(h(\boldsymbol{Z}) \mid \widehat{\boldsymbol{\mu}})| \to 0, \text{ as } n \to \infty. \tag{3.12}$$

Since $\widehat{\boldsymbol{\mu}}$ is n^c-consistent for some $c \geq 1/2$, it follows that $|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*| < C/n^c$ for some sufficiently large but fixed $C > 0$ and for all sufficiently large n with probability tending to 1. As a result,

$$\begin{aligned}
& |\mathbb{E}(h(T_n)) - \mathbb{E}(h(\boldsymbol{Z}))| \\
= \; & |\mathbb{E}(\mathbb{E}(h(T_n) \mid \widehat{\boldsymbol{\mu}})) - \mathbb{E}(h(T_n) \mid \boldsymbol{\mu}^*) + \mathbb{E}(h(T_n) \mid \boldsymbol{\mu}^*) - \mathbb{E}(h(\boldsymbol{Z}))| \\
\leq \; & \sup_{|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*| < C/n^c} |\mathbb{E}(\mathbb{E}(h(T_n) \mid \widehat{\boldsymbol{\mu}}))) - \mathbb{E}(\mathbb{E}(h(T_n) \mid \boldsymbol{\mu}^*))| + |\mathbb{E}(h(T_n) \mid \boldsymbol{\mu}^*) - \mathbb{E}(h(\boldsymbol{Z}) \mid \boldsymbol{\mu}^*)|.
\end{aligned}$$

Since $\mathbb{E}(h(T_n) \mid \widehat{\boldsymbol{\mu}})$ is also a bounded and continuous function of $\widehat{\boldsymbol{\mu}}$, it follows that the first term converges to 0 in probability as $n \to \infty$. The convergence of the second term is assured by (3.12) for the fixed $\boldsymbol{\mu} = \boldsymbol{\mu}^*$. The rest of the proof follows as in Ghosh and Basu (2013). □

It is worthwhile to note that while the true eigenvalues $\gamma_1^*, \ldots, \gamma_p^*$ is in decreasing order, the estimated eigenvalues $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_p$ may not be. If two eigenvalues γ_i^* and γ_{i+1}^* are close to each other, it is possible that the corresponding estimates satisfy $\widehat{\gamma}_{i+1} > \widehat{\gamma}_i$. Thus, one may be interested in finding out the asymptotic distribution of the order statistics of estimated eigenvalues. However, because of the presence of a strong correlation between the estimated eigenvalues, it is difficult to obtain a tractable closed form of this distribution. It is only possible to derive some probabilistic bounds on the extreme eigenvalues using the methods described in Ross (2010), Adler and Taylor (2007).

One may also be interested in the special case when the underlying elliptically symmetric distribution is assumed to be the Gaussian distribution. Formally, if we consider that the sample observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are distributed according to a p-variate normal distribution with unknown mean $\boldsymbol{\mu}^*$ and unknown dispersion matrix $\boldsymbol{\Sigma}^* = \sum_{k=1}^{p} \gamma_k^* \boldsymbol{v}_k^* (\boldsymbol{v}_k^*)^\intercal$, then it follows that under the same set of assumptions, one can establish the following corollary.

Corollary 3.4. *Suppose the Assumptions (C1)-(C2) hold and the location estimator $\widehat{\boldsymbol{\mu}}$ is n^c-consistent for μ^* for some $c \geq 1/2$. Then, there exists a sequence of converged rPCAdpd estimator $\widehat{\boldsymbol{\theta}}_n = ((\widehat{\gamma}_1)_n, \ldots, (\widehat{\gamma}_p)_n, \widehat{\boldsymbol{\eta}}_n)$ as in Eq. (3.7) for the Gaussian model family of distributions such that as the sample size $n \to \infty$,*

1. *The estimated eigenvalue* $(\widehat{\gamma}_j)_n$ *is* $\sqrt{n}$*-consistent for* γ_j^* *and the estimated eigenvector* $(\widehat{\boldsymbol{v}}_j)_n$ *is* $\sqrt{n}$*-consistent for* $\boldsymbol{v}_j^*$ *for* $j = 1, 2, \ldots, p$.

2. *The scaled and centred estimated vector of principal component eigenvalues*

$$\sqrt{n}\left(\begin{bmatrix}(\widehat{\gamma}_1)_n\\ \vdots \\ (\widehat{\gamma}_p)_n\end{bmatrix} - \begin{bmatrix}\gamma_1^*\\ \vdots \\ \gamma_p^*\end{bmatrix}\right)$$

has an asymptotic p*-variate normal distribution with mean* $\mathbf{0}$ *and dispersion matrix*

$$\frac{(1+\alpha)^{p+4}}{(1+2\alpha)^{p/2}}\boldsymbol{M}^{-1}\left(A_1(\alpha)\mathrm{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)\mathrm{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)^{\intercal} + \frac{1}{2(1+2\alpha)^2}\boldsymbol{\Gamma}^{-2}\right)\boldsymbol{M}^{-1},$$

where

$$\boldsymbol{M} = \left(\frac{\alpha^2}{4}\mathrm{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)\mathrm{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)^{\intercal} + \frac{1}{2}\boldsymbol{\Gamma}^{-2}\right),$$

$$A_1(\alpha) = \alpha^2\left[\frac{1}{(1+2\alpha)^2} - \frac{(1+2\alpha)^{p/2}}{4(1+\alpha)^{p+2}}\right].$$

3. *The scaled and centered estimated* $\widehat{\boldsymbol{\eta}}_n$ *corresponding to the principal component eigenvectors, i.e.,* $\sqrt{n}(\widehat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}^*)$ *has an asymptotic normal distribution with mean* $\mathbf{0}$ *and dispersion matrix*

$$\frac{(1+\alpha)^{p+4}}{(1+2\alpha)^{2+p/2}}\left(\sum_{k=1}^{p}\sum_{l=1}^{p}\left(1-\frac{\gamma_k^*}{\gamma_l^*}\right)\boldsymbol{G}_k(\boldsymbol{v}_l^*)(\boldsymbol{v}_k^*)^{\intercal}\boldsymbol{G}_l^{\intercal}\right)^{-1},$$

where $\boldsymbol{G}_k = \dfrac{\partial \boldsymbol{v}_k}{\partial \boldsymbol{\eta}}|_{\boldsymbol{\eta}=\boldsymbol{\eta}^*}$*, the matrix corresponding of the gradients of the eigenvector* $\boldsymbol{v}_k$ *with respect to its natural parametrization* $\boldsymbol{\eta}$.

4. *The rPCAdpd estimate of the eigenvalues* $((\widehat{\gamma}_1)_n, \ldots, (\widehat{\gamma}_p)_n)$ *and estimate of the eigenvectors* $((\widehat{\boldsymbol{v}}_1)_n, \ldots, (\widehat{\boldsymbol{v}}_p)_n)$ *are asymptotically independent.*

Proof. The generating function for the Gaussian distribution in the elliptically symmetric family of distributions is $g(x) = (-x/2)$. It follows that $g'(x) = -1/2$ and the normalizing constant $\mathcal{C}_g := (2\pi)^{p/2}\prod_{k=1}^{p}\gamma_k^{1/2}$. It is easy to verify that the Assumption (C3) holds for this specific choice of g function. For ease of notation, we also define

$$c_\alpha = \frac{\mathcal{C}_{(1+\alpha)g}}{\mathcal{C}_g} = (2\pi)^{-\alpha p/2}(1+\alpha)^{-p/2}\prod_{k=1}^{p}(\gamma_k^*)^{-\alpha/2}.$$

Now, some standard calculation using properties of normal distribution and its quadratic forms (Petersen et al., 2008) reveals that $A_2((1+\alpha)g) = (1+\alpha)\boldsymbol{\Sigma}^*/4$, and

$$A_4((1+\alpha)g; \boldsymbol{u}, \boldsymbol{v}) = \frac{1}{4}\left[\boldsymbol{\Sigma}^*\left(\boldsymbol{u}\boldsymbol{v}^{\intercal} + \boldsymbol{v}\boldsymbol{u}^{\intercal}\right)\boldsymbol{\Sigma}^* + \mathrm{Trace}\left(\boldsymbol{u}\boldsymbol{v}^{\intercal}\boldsymbol{\Sigma}^*\right)\boldsymbol{\Sigma}^*\right].$$

In particular, for any $k, l = 1, 2, \ldots, p$,

$$\begin{aligned}A_4((1+\alpha)g; \boldsymbol{v}_k^*, \boldsymbol{v}_l^*) &= \frac{1}{4}\left[\boldsymbol{\Sigma}^*\left((\boldsymbol{v}_k^*)(\boldsymbol{v}_l^*)^{\intercal} + (\boldsymbol{v}_l^*)(\boldsymbol{v}_k^*)^{\intercal}\right)\boldsymbol{\Sigma}^* + \mathrm{Trace}\left((\boldsymbol{v}_k^*)(\boldsymbol{v}_l^*)^{\intercal}\boldsymbol{\Sigma}^*\right)\boldsymbol{\Sigma}^*\right]\\ &= \frac{1}{4}\left[\gamma_k^*\gamma_l^*\left((\boldsymbol{v}_k^*)(\boldsymbol{v}_l^*)^{\intercal} + (\boldsymbol{v}_l^*)(\boldsymbol{v}_k^*)^{\intercal}\right) + \delta_{k-l}\gamma_l^*\boldsymbol{\Sigma}^*\right],\end{aligned}$$

where we use the fact that $\boldsymbol{v}_k^*$ is an eigenvector of $\boldsymbol{\Sigma}^*$ corresponding to the eigenvalue γ_k^*. Here, δ_x is the Dirac delta function, i.e., $\delta_x = 1$ if $x = 0$, and it is equal to 0 otherwise. Thus, it turns out that $j^h(\boldsymbol{\mu}^*, \boldsymbol{\mu}^*) = \frac{c_\alpha}{(1+\alpha)}(\boldsymbol{\Sigma}^*)^{-1}$, and

$$j^h(\gamma_k^*, \gamma_l^*) = \frac{c_\alpha}{4(1+\alpha)^2 \gamma_k^* \gamma_l^*} \left(\alpha^2 + 2\mathbf{1}_{\{k=l\}}\right)$$

and $j^h(\gamma_k^*, \boldsymbol{\eta}^*) = 0$, where we use the fact that $\boldsymbol{G}_k \boldsymbol{v}_k^* = 0$. This equality follows from differentiating both sides of the identity $(\boldsymbol{v}_k^*)^\intercal(\boldsymbol{v}_k^*) = 1$ with respect to the parameter $\boldsymbol{\eta}$ at $\boldsymbol{\eta} = \boldsymbol{\eta}^*$. Similarly, differentiating the identity $(\boldsymbol{v}_k^*)^\intercal(\boldsymbol{v}_l^*) = 0$ for $k \neq l$ with respect to $\boldsymbol{\eta}$ yields that $\boldsymbol{G}_k \boldsymbol{v}_l^* + \boldsymbol{G}_l \boldsymbol{v}_k^* = 0$. Some lengthy calculations and an application of this identity allow us to obtain

$$j^h(\boldsymbol{\eta}^*, \boldsymbol{\eta}^*) = \frac{c_\alpha}{(1+\alpha)^2} \left(\sum_{k=1}^{p} \sum_{l=1}^{p} \left(1 - \frac{\gamma_k^*}{\gamma_l^*}\right) \boldsymbol{G}_k (\boldsymbol{v}_l^*)(\boldsymbol{v}_k^*)^\intercal \boldsymbol{G}_l^\intercal \right).$$

A similar calculation may be performed to determine the entries of $K_{\boldsymbol{\theta}^*}$. An application of Theorem 3.4 now completes the proof. □

The independence of the rPCAdpd estimate of eigenvalues and eigenvectors can enable one to create confidence intervals for the eigenvalues and eigenvectors separately. To create the asymptotic confidence interval for the eigenvalues, the knowledge of the corresponding estimates of eigenvalues is sufficient. In contrast, the asymptotic confidence band for eigenvectors require both the eigenvalues and the eigenvectors.

Remark 3.2. *The density power divergence introduced in Basu et al. (1998) becomes the same as the Kullback-Leibler divergence between the true density and the model density $f_{\boldsymbol{\theta}}(\cdot)$ in the limit as $\alpha \to 0$. Thus, for $\alpha \to 0$, the estimating equations for the MDPDE turn out to be equivalent to the estimating equations corresponding to the log-likelihood. Consequently, the MDPDE coincides with the maximum likelihood estimator as $\alpha \to 0$. From Corollary 3.4 it then follows that the maximum likelihood estimates (MLE) of the eigenvalues of the covariance matrix under the Gaussian distribution are asymptotically normal with mean γ_j and covariance $2\gamma_j^2/n$ and are asymptotically independent. This result has been well established in the literature; see Girshick (1939) for references. A similar result for the asymptotic distribution of the MLE of eigenvectors was derived by Anderson (1963). Results on the asymptotic independence between the MLE of the eigenvalues and eigenvectors were also derived by Tyler (1981) for a general setup with repeated eigenvalues. Hence, Corollary 3.4 can be seen as a generalization of these results.*

Remark 3.3. *In contrast to Remark 3.2, for $\alpha = 1$, the form of density power divergence becomes the same as the L^2 distance between the true density and the model density $f_{\boldsymbol{\theta}}(\boldsymbol{x})$. If we denote the minimum L^2 distance estimator of the eigenvalues by $\widetilde{\boldsymbol{\gamma}} = (\widetilde{\gamma}_1, \ldots, \widetilde{\gamma}_p)^\intercal$ and the true eigenvalues by $\boldsymbol{\gamma}^* = (\gamma_1^*, \ldots, \gamma_p^*)^\intercal$, then*

$$\sqrt{n}(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, \mathcal{V}_2),$$

as $n \to \infty$. *Here,* $\xrightarrow{d}$ *denotes the convergence in law. The asymptotic variance is given by*

$$\mathcal{V}_2 = \frac{2^{(p+8)}}{3^{(p/2)}} \boldsymbol{M}_1^{-1} \left(\left(\frac{1}{9} - \frac{3^{(p/2)}}{2^{(p+4)}} \right) \text{Diag}\left(\boldsymbol{\Gamma}^{-1}\right) \text{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)^{\intercal} + \frac{1}{18} \boldsymbol{\Gamma}^{-1} \right) \boldsymbol{M}_1^{-1},$$

where $\boldsymbol{M}_1 = \left(\text{Diag}\left(\boldsymbol{\Gamma}^{-1}\right) \text{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)^{\intercal} + 2\boldsymbol{\Gamma}^{-2}\right)$. *Since the quantity* $\frac{2^{(x+8)}}{3^{(x/2)}} \left(\frac{1}{9} - \frac{3^{(x/2)}}{2^{(x+4)}} \right)$ *increases exponentially fast as* x *increases, the variance of the minimum* L^2*-distance estimator increases exponentially with increase in the dimension* p. *This shows that by using the highly robust minimum* L^2 *distance estimator to obtain the principal components, one sacrifices considerable efficiency in estimation.*

3.4.4 Influence Function Analysis

The influence function is a local measure of the sensitivity and robustness of an estimator (Hampel et al., 2011). One of the primary qualities of a robust estimator is that it should have a bounded influence function. Such estimators with bounded influence functions are also called B-robust estimators.

In this section, we investigate the influence function of the rPCAdpd estimator for the Gaussian model family of distributions. For this particular choice, the asymptotic independence of the eigenvalues and the eigenvectors as shown in Theorem 3.4 helps in deriving the influence functions quite nicely. Let us assume that instead of the true distribution $\Phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})$, the observations $\boldsymbol{X}_i$s come from a contaminated distribution

$$G_\epsilon(\boldsymbol{x}) = (1-\epsilon)\Phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}) + \epsilon \delta_{\boldsymbol{y}}(\boldsymbol{x}),$$

where $\delta_{\boldsymbol{y}}(\cdot)$ is the degenerate distribution at $\boldsymbol{y} \in \mathbb{R}^p$. Let $\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})$ be the density function corresponding to the normal distribution function $\Phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})$ with the parameter vector $\boldsymbol{\theta}^*$. Then the influence of this contamination on the estimated principal components can be readily obtained from the influence function derived in Basu et al. (1998). Due to the asymptotic independence, the influence functions for the estimators of the eigenvalues and the eigenvectors can be separately obtained along with an application of the chain rule to incorporate the influence of the robust location estimator. It turns out that the influence functions for the estimated eigenvalues and the eigenvectors are respectively,

$$I_\alpha(\boldsymbol{y}, \widehat{\boldsymbol{\gamma}}, \Phi_{\boldsymbol{\theta}^*}) = \frac{4(1+\alpha)^2}{C_\alpha} \left[\alpha^2 \text{Diag}\left(\boldsymbol{\Gamma}^{-1}\right) \text{Diag}\left(\boldsymbol{\Gamma}^{-1}\right)^{\intercal} + 2\boldsymbol{\Gamma}^{-2}\right]^{-1} \left(\begin{bmatrix} u_{\gamma_1^*}(\boldsymbol{y}) \\ \vdots \\ u_{\gamma_p^*}(\boldsymbol{y}) \end{bmatrix} \phi^\alpha_{\boldsymbol{\theta}^*}(\boldsymbol{y}) I(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}, \Phi_{\boldsymbol{\theta}^*}) - \begin{bmatrix} \xi_{\gamma_1^*} \\ \vdots \\ \xi_{\gamma_p^*} \end{bmatrix} \right),$$

and,

$$I_\alpha(\boldsymbol{y}, \widehat{\boldsymbol{\eta}}, \Phi_{\boldsymbol{\theta}^*}) = -\frac{(1+\alpha)^2}{C_\alpha} \left[\sum_{k=1}^p \frac{G_k \boldsymbol{\Sigma}^* G_k^{\intercal}}{\gamma_k^*} \right]^{-1} \sum_{k=1}^p \frac{G_k}{\gamma_k^*} (\boldsymbol{y} - \boldsymbol{\mu}^*)(\boldsymbol{y} - \boldsymbol{\mu}^*)^{\intercal} \boldsymbol{v}_k^* \phi^\alpha_{\boldsymbol{\theta}}(\boldsymbol{y}) I(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}, \Phi_{\boldsymbol{\theta}^*}).$$

This may be obtained by standard calculations as in Ghosh and Basu (2013), but for the specific values of $\boldsymbol{\xi}_{\boldsymbol{\theta}^*}$ and $\boldsymbol{J}_{\boldsymbol{\theta}^*}$ in the rPCAdpd case. Here, $u_{\gamma_j^*}(\boldsymbol{y})$ denotes the score function with respect to the j-th eigenvalue γ_j^* evaluated at the contaminating point $\boldsymbol{y}$ and $I(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}, \Phi_{\boldsymbol{\theta}^*})$ is the influence function of the location estimator $\widehat{\boldsymbol{\mu}}$ at $\boldsymbol{y}$. We assume that the location estimator $\widehat{\boldsymbol{\mu}}$ is robust and hence has a bounded influence function, which is true for the L_1-median. To show that both the above influence functions are bounded, one may note that the exponential quantity $e^{-\alpha(\boldsymbol{y}-\boldsymbol{\mu}^*)^\intercal(\boldsymbol{\Sigma}^*)^{-1}(\boldsymbol{y}-\boldsymbol{\mu}^*)/2}$ present in the Gaussian density $\phi_{\boldsymbol{\theta}^*}(\boldsymbol{y})$ is bounded below by $e^{-\alpha\|\boldsymbol{y}-\boldsymbol{\mu}^*\|^2/2\gamma_{(p)}^*}$ and bounded above by $e^{-\alpha\|\boldsymbol{y}-\boldsymbol{\mu}^*\|^2/2\gamma_{(1)}^*}$, where $\gamma_{(1)}^*$ and $\gamma_{(p)}^*$ are the largest and the smallest eigenvalues of $\boldsymbol{\Sigma}^*$ respectively. Now the boundedness of the influence function follows from Assumption (C3), which can be easily verified for $g(x) = -x/2$ corresponding to the Gaussian distribution. Therefore, we have the following result.

Theorem 3.5. *Suppose that the Assumptions (C1)-(C3) holds and the location estimator $\widehat{\boldsymbol{\mu}}$ is B-robust. Then the rPCAdpd estimator of the eigenvalues and the eigenvectors is also B-robust.*

3.4.5 Breakdown Point Analysis

The breakdown point of an estimator is another accepted measure of the robustness of an estimator besides the influence function. In contrast to the influence function, the breakdown point is a measure of the global reliability of an estimator. It is defined as the highest level of contamination that an estimator can tolerate in the sample before producing an egregiously bad estimate (Hampel, 1971). A detailed definition of the breakdown point has been presented earlier in Section 1.3.3.

Since the rPCAdpd algorithm is composed of two steps, i.e., the location estimation and the eigenvalue and eigenvector estimation using the rSVDdpd procedure; the asymptotic breakdown of the entire procedure is the minimum of the asymptotic breakdown of these individual procedures. It is well known that the robust L_1-median (used as the location estimator in our entire study) has an asymptotic breakdown point of $1/2$. Also, under fairly general conditions, it will be shown in Chapter 5 that the robust MDPDE has an asymptotic breakdown point at least $\alpha/(1+\alpha)$, where α is the robustness parameter with $\alpha \in [0, 1]$. Hence, the resulting rPCAdpd estimator also has an asymptotic breakdown point of at least $\alpha/(1+\alpha)$, independent of the data dimension p. This means the proposed rPCAdpd estimator can scale to arbitrarily high dimensional data without losing its robustness properties, unlike the standard M-estimation techniques. This scalability aspect will also be demonstrated later through numerical simulation studies in Section 3.5. Note that as $\alpha \to 0$, the lower bound of the breakdown becomes 0 suggesting a lack of robustness, while for $\alpha = 1$, one would get the highest possible breakdown $1/2$ but sacrificing considerable efficiency as indicated in Remark 3.3.

Note that, this is in contrast to the breakdown point result obtained by Maronna (1976) for an affine equivariant M-estimator, which states that an affine equivariant M-estimator has a breakdown point at most $1/(p+1)$ where p is the dimensionality of the

data. As explained in Basu et al. (1998), the MDPDE is a special case of the M-estimator, and also we showed the orthogonal equivariance property of the rPCAdpd estimator in Section 3.4.2. This discrepancy holds because the classes of the M-estimators considered by Maronna (1976) differ from the classes of minimum divergence estimators to which MDPDE belongs. In particular, Maronna (1976) considered the estimators given as the solution to the system of equations

$$\sum_{i=1}^{n} u_1\left((\boldsymbol{X}_i-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})\right)(\boldsymbol{X}_i-\boldsymbol{\mu})=0,$$
$$\sum_{i=1}^{n} u_2\left((\boldsymbol{X}_i-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})\right)(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})^\intercal=\boldsymbol{\Sigma},$$

where $u_1(s)$ and $u_2(s)$ are suitable non-increasing functions for $s \geq 0$. On the other hand, denoting $\boldsymbol{\Sigma} = \sum_{k=1}^{p} \gamma_k \boldsymbol{v}_k \boldsymbol{v}_k^\intercal$, the estimating equations for MDPDE turn out to be

$$\sum_{i=1}^{n} \exp\left(-0.5\alpha(\boldsymbol{X}_i-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})\right)(\boldsymbol{X}_i-\boldsymbol{\mu})=0,$$
$$\sum_{i=1}^{n} \exp\left(-0.5\alpha(\boldsymbol{X}_i-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})\right)((\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})^\intercal-\boldsymbol{\Sigma})=0,$$

under the Gaussian model as in Eq. (3.7). While the estimating equation for the location estimator matches, the estimating equation for the covariance matrix differs. Therefore, the breakdown point results provided by Maronna (1976) do not apply to our proposed rPCAdpd estimator.

3.5 Simulation Studies

In this section, we compare performances of the existing robust PCA algorithms and our proposed rPCAdpd algorithm through simulation studies under varying levels of contamination present in the data matrix. Among the plethora of existing RPCA methods, we take the classical PCA (Jolliffe, 2002), spherical and elliptical PCA (LOC) (Locantore et al., 1999), ROBPCA algorithm by Hubert et al. (2005), projection pursuit based methods Proj and Grid (Croux and Ruiz-Gazen, 2005), robust PCA using robust covariance matrix estimation (RobCov) (Todorov and Filzmoser, 2010), principal component pursuit (PCP) algorithm by Candès et al. (2011) and Gmedian based robust principal component analysis (Gmed) by Cardot and Godichon-Baggioni (2017), for comparison purposes. We have considered several variants of the rPCAdpd algorithm differing only in the location estimator used. Based on the empirical performances, we have found that the L_1-median as a location estimator provides a desirable balance between robustness, efficiency and computational complexity, hence it is the only variant demonstrated in the results described below.

3.5.1 Simulation Settings

In the simulation experiments, we consider a data matrix comprised of i.i.d. rows. The rows $\boldsymbol{X}_i$s are generated as $\boldsymbol{X}_i = (1-\delta_i)\widetilde{\boldsymbol{X}}_i + \delta_i \boldsymbol{\epsilon}_i$ for $i = 1, 2, \ldots, n$. The uncontaminated sample $\widetilde{\boldsymbol{X}}_i$ is normally distributed with zero mean vector and a dispersion matrix $\boldsymbol{\Sigma}$ whose elements are given by $\boldsymbol{\Sigma}_{ij} = \min(i, j)/p$ for $i, j = 1, 2, \ldots, p$. This setup is similar to the one described in Cardot and Godichon-Baggioni (2017) and can be regarded as a discretized version of a Brownian motion within the unit $(0, 1)$ interval. The random variables δ_i which control the level of contamination are i.i.d. Bernoulli random variable with success probability $\delta \in [0, 1/2)$. The contaminating variable $\boldsymbol{\epsilon}_i$s are chosen to possess different features compared to $\widetilde{\boldsymbol{X}}_i$, and in this regard, we feel that the choice of the distribution of outliers as given in Cardot and Godichon-Baggioni (2017) is too restrictive. In comparison, Hubert et al. (2005) consider outliers that have changes in both mean and variance components separately, and hence we choose to work with them. In summary, we consider the following simulation scenarios.

(PCA1) $\delta = 0$, i.e., only pure data are present and there is no contamination.

(PCA2) Here a proportion of elements are contaminated. The contaminating variable $\boldsymbol{\epsilon}_i$s are i.i.d. p-variate normal random variables with mean $\mu(f_1)$ and variance $\boldsymbol{\Sigma}/f_2$. The mean vector $\mu(f_1)$ is a p-length vector where 10% of the entries at random are equal to f_1 while the rest of the entries are equal to 0. We consider the following subcases.

- (PCA2a) Here, $f_1 = 3, f_2 = 1$ and $\delta = 0.1$. Therefore, on average 10% of the data will be contaminated.
- (PCA2b) Here, $f_1 = 3, f_2 = 1$ and $\delta = 0.2$. Therefore, on average 20% of the data will be contaminated.
- (PCA2c) Analogous to (PCA2a) but with $f_2 = 5$.
- (PCA2d) Analogous to (PCA2b) but with $f_2 = 5$.

(PCA3) This is similar to simulation scenario (PCA2) but the contaminating variable $\boldsymbol{\epsilon}_i$s are i.i.d. p-variate t-distribution with 5 degrees of freedom with dispersion matrix $\boldsymbol{\Sigma}/f_2$ and a non-centrality parameter $\mu(f_1)$. This is used to understand the behaviour of the PCA algorithms for heavy-tailed contaminating variables.

- (PCA3a) $f_1 = 3, f_2 = 1$ and $\delta = 0.1$.
- (PCA3b) $f_1 = 3, f_2 = 1$ and $\delta = 0.2$.
- (PCA3c) $f_1 = 3, f_2 = 5$ and $\delta = 0.1$.
- (PCA3d) $f_1 = 3, f_2 = 5$ and $\delta = 0.2$.

In each of the above simulation scenarios, we consider five different situations with the number of samples $n = 50$ but with different dimensions ranging from very small to moderately large ($p = 10, 25, 50, 100, 250$). Based on $B = 1000$ repetitions of each exercise,

we obtain an estimate of bias and mean absolute error (MAE) of the estimated eigenvalues as

$$\text{Bias}_k = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}_k^{(b)} - \gamma_k,\ \text{MAE}_k = \frac{1}{B}\sum_{b=1}^{B}\left|\widehat{\gamma}_k^{(b)} - \gamma_k\right|,$$

where $\widehat{\gamma}_k^{(b)}, \gamma_k$ respectively denote the estimate and the true k-th eigenvalue for the b-th sample. Similarly, to measure the discrepancy in the estimated eigenvalues we look at the Subspace Recovery Error (SRE) given by

$$\text{SRE} = \frac{1}{B}\sum_{b=1}^{B} 2\left(r - \text{Trace}\left(\widehat{\boldsymbol{P}}_b\boldsymbol{P}\right)\right),$$

where $\widehat{\boldsymbol{P}}_b = \sum_{k=1}^{r}\widehat{\boldsymbol{v}}_k^{(b)}(\widehat{\boldsymbol{v}}_k^{(b)})^\intercal$ is the projection matrix onto the span of the estimated eigenvectors corresponding to the largest r eigenvalues from b-th replication, and $\boldsymbol{P} = \sum_{k=1}^{r}\boldsymbol{v}_k\boldsymbol{v}_k^\intercal$ be the corresponding projection matrix from the true eigenvectors corresponding to the largest r eigenvalues. In each of these simulation scenarios, we keep the choice of $r = 5$ fixed, as more than 90% of the variability can be explained by the first 5 principal components.

3.5.2 Simulation Results

The simulation results from the aforementioned algorithms are demonstrated in Tables 3.2-3.10. We denote the rPCAdpd estimator with L_1-median as the location estimator in these tables as the DPD method, with the robustness parameter α shown in parenthesis. Also, the RobCov algorithm (Todorov and Filzmoser, 2010) uses MCD-based robust covariance estimation for RPCA. Thus, it is inapplicable when variables outnumber the sample size (i.e., $n \leq p$), and those entries are marked as NA in these tables.

Table 3.2: Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA1).

Metric	p	Classical	LOC	ROBPCA	Proj	RobCov	Grid	Gmed	PCP	DPD (0.25)	DPD (0.5)	DPD (0.75)	DPD (1)
Bias	10	0.059	0.723	0.194	0.229	0.431	0.416	0.043	1.066	0.06	0.062	0.065	0.068
	25	0.019	2.175	0.227	0.362	0.336	0.807	0.079	2.45	0.017	0.013	0.01	0.008
	50	0.031	4.572	0.519	0.467	NA	1.414	0.177	4.729	0.026	0.017	0.007	0.017
	100	0.194	9.366	0.944	1.058	NA	2.827	0.314	9.387	0.201	0.216	0.233	0.254
	250	0.154	23.76	2.847	2.239	NA	6.906	0.644	23.301	0.184	0.236	0.295	0.359
MAE	10	17.919	72.334	26.756	33.798	45.663	50.391	19.122	106.477	17.921	17.936	17.981	18.054
	25	38.106	217.462	50.069	75.426	49.757	123.166	40.445	244.951	38.25	38.485	38.814	39.252
	50	73.085	457.212	110.189	137.425	NA	240.293	84.614	472.875	73.126	73.225	73.489	73.803
	100	143.086	936.571	200.658	263.141	NA	381.432	154.736	938.685	143.426	144.012	144.595	145.234
	250	395.183	2375.96	536.355	731.985	NA	1010.852	434.823	2330.092	395.893	396.863	397.827	396.641
SRE	10	1	1.347	1.172	1.677	1.429	2.498	1.126	1.188	0.99	0.983	0.99	0.995
	25	0.829	1.417	1.028	1.76	1.127	3.573	1.004	1.136	0.833	0.84	0.845	0.872
	50	0.766	1.336	0.959	1.653	NA	4.038	0.931	0.871	0.766	0.771	0.795	0.818
	100	0.836	1.459	0.985	1.587	NA	3.313	1.045	0.923	0.84	0.847	0.857	0.872
	250	0.828	1.268	0.927	1.494	NA	3.265	0.939	0.899	0.828	0.832	0.84	0.859

Table 3.2 presents metrics for various PCA algorithms in simulation setup (PCA1). For uncontaminated data, traditional PCA outperforms all robust methods across all

metrics. Gmed and ROBPCA exhibit relatively less efficiency loss. However, the proposed rPCAdpd consistently outperforms both of these under any $\alpha \in [0, 1]$ and regardless of the location estimator used. Increasing α escalates the efficiency loss moderately compared to the other methods. Although the L_1-median is quite inefficient (Huber, 2004), its strong robustness properties allow rPCAdpd to achieve very low MAE.

Table 3.3: Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA2a).

Metric	p	Classical	LOC	ROBPCA	Proj	RobCov	Grid	Gmed	PCP	DPD (0.25)	DPD (0.5)	DPD (0.75)	DPD (1)
Bias	10	0.158	0.74	0.22	0.205	0.44	0.458	0.08	1.065	0.15	0.08	0.017	0.009
	25	0.304	2.183	0.386	0.416	0.459	1.046	0.152	2.452	0.281	0.097	0.023	0.025
	50	0.874	4.585	0.807	1.074	NA	2.308	0.274	4.724	0.792	0.248	0.121	0.129
	100	1.627	9.382	1.41	1.63	NA	4.246	0.433	9.386	1.445	0.251	0.111	0.132
	250	4.567	23.777	3.114	4.573	NA	11.98	1.471	23.303	4.134	0.866	0.718	0.777
MAE	10	29.382	74.012	30.359	37.423	47.943	59.314	24.69	106.383	30.81	24.729	18.938	18.165
	25	62.13	218.321	62.944	83.99	61.246	141.267	56.95	245.214	63.996	47.073	39.385	39.137
	50	138.901	458.488	124.057	163.26	NA	305.897	111.013	472.415	145.313	95.448	82.258	82.154
	100	258.437	938.246	213.108	296.41	NA	495.19	211.008	938.639	268.129	155.77	140.017	139.704
	250	693.852	2377.669	558.396	729.947	NA	1311.666	545.383	2330.337	709.073	398.446	380.915	383.593
SRE	10	1.779	1.875	1.056	2.016	1.405	2.697	1.843	1.171	1.787	1.46	1.063	1.005
	25	2.135	2.322	1.063	2.243	1.076	3.774	2.197	1.152	2.137	1.261	0.872	0.852
	50	2.185	2.43	0.998	2.2	NA	4.395	2.263	0.924	2.172	1.145	0.847	0.871
	100	2.251	2.482	1.084	2.3	NA	3.544	2.351	0.986	2.228	1.075	0.898	0.901
	250	2.231	2.504	0.991	2.229	NA	3.599	2.317	0.912	2.196	0.936	0.869	0.882

Table 3.4: Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for simulation scenario (PCA2b).

Metric	p	Classical	LOC	ROBPCA	Proj	RobCov	Grid	Gmed	PCP	DPD (0.25)	DPD (0.5)	DPD (0.75)	DPD (1)
Bias	10	0.321	0.757	0.381	0.429	0.589	0.757	0.14	1.065	0.329	0.281	0.138	0.067
	25	0.553	2.198	0.368	0.635	1.004	1.344	0.235	2.451	0.568	0.364	0.073	0.036
	50	1.467	4.602	0.829	1.796	NA	3.221	0.583	4.617	1.48	0.97	0.323	0.182
	100	2.66	9.414	1.028	2.692	NA	6.019	1.235	9.159	2.766	2.005	0.533	0.2
	250	7.033	23.805	3.245	8.006	NA	15.969	2.746	22.799	7.089	4.447	1.446	0.299
MAE	10	41.99	75.693	43.08	52.712	60.646	82.448	30.185	106.511	45.196	42.803	29.261	22.409
	25	85.197	219.781	63.246	93.376	112.498	165.495	65.545	245.114	90.83	74.713	45.453	41.413
	50	194.589	460.223	130.581	236.929	NA	406.956	144.635	462.199	209.841	172.386	110.373	96.321
	100	364.678	941.397	221.517	400.614	NA	665.195	267.786	916.897	394.475	317.498	173.981	142.885
	250	957.207	2380.505	518.404	1066.532	NA	1696.499	658.65	2283.607	1060.277	838.361	545.85	432.927
SRE	10	1.812	2.049	1.109	2.346	1.424	2.886	1.889	1.197	1.811	1.774	1.405	1.111
	25	2.14	2.422	1.021	2.645	2.212	4.19	2.26	1.276	2.152	1.832	1.111	1.03
	50	2.219	2.472	1.02	2.828	NA	4.985	2.314	2.265	2.24	1.819	1.201	1.049
	100	2.227	2.453	1.043	2.868	NA	3.86	2.326	2.272	2.242	1.868	1.153	1.007
	250	2.249	2.549	1.066	2.976	NA	3.901	2.362	2.302	2.262	1.767	1.16	1.007

Tables 3.3 and 3.4 respectively show results for setups (PCA2a) and (PCA2b) which differs in the level of contamination. As the contamination level increases, classical PCA worsens as expected, spherical PCA (Locantore et al., 1999) yields biased estimates when the number of variables p is large, and the projection pursuit-based methods also perform poorly. The ROBPCA algorithm by Hubert et al. (2005) and the Gmedian algorithm by Cardot and Godichon-Baggioni (2017) stand out to be the most promising among the

Table 3.5: Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA2c).

Metric	p	Classical	LOC	ROBPCA	Proj	RobCov	Grid	Gmed	PCP	DPD (0.25)	DPD (0.5)	DPD (0.75)	DPD (1)
Bias	10	0.215	0.746	0.166	0.159	0.396	0.434	0.161	1.065	0.234	0.131	0.112	0.11
	25	0.328	2.182	0.249	0.305	0.604	1.168	0.286	2.458	0.335	0.127	0.112	0.105
	50	0.855	4.592	0.139	0.494	NA	2.04	0.845	4.745	0.944	0.356	0.348	0.347
	100	1.793	9.394	0.152	1.086	NA	3.896	1.713	9.403	1.861	0.858	0.8	0.775
	250	3.839	23.786	1.204	2.871	NA	10.833	3.264	23.392	4.119	1.396	1.331	1.275
MAE	10	29.127	74.586	28.237	32.563	44.906	57.255	25.461	106.444	31.154	22.089	19.424	19.468
	25	60.747	218.249	56.479	76.163	87.508	156.242	60.369	245.818	64.18	44.461	42.378	42.696
	50	129.217	459.232	99.63	149.773	NA	314.016	127.928	474.518	140.922	83.036	82.464	81.935
	100	253.278	939.405	195.978	280.408	NA	510.712	251.441	940.342	266.573	163.12	158.681	158.084
	250	633.745	2378.584	496.981	745.615	NA	1303.445	623.223	2339.228	676.812	398.135	394.581	395.265
SRE	10	1.815	2.014	1.099	2.118	1.485	2.823	1.87	1.194	1.821	1.159	1	0.997
	25	2.167	2.43	1.013	2.408	2.127	4.035	2.261	1.151	2.064	0.992	0.888	0.902
	50	2.221	2.47	1.043	2.531	NA	4.64	2.328	1.03	2.101	0.983	0.971	0.969
	100	2.242	2.472	1.028	2.531	NA	3.721	2.34	0.96	1.942	0.918	0.882	0.893
	250	2.251	2.515	0.963	2.535	NA	3.702	2.379	0.906	2.019	0.882	0.869	0.897

Table 3.6: Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA2d).

Metric	p	Classical	LOC	ROBPCA	Proj	RobCov	Grid	Gmed	PCP	DPD (0.25)	DPD (0.5)	DPD (0.75)	DPD (1)
Bias	10	0.318	0.795	0.143	0.351	0.624	0.732	0.238	1.066	0.381	0.223	0.171	0.173
	25	0.673	2.244	0.251	0.785	0.76	1.607	0.491	2.464	0.741	0.485	0.329	0.33
	50	1.558	4.655	0.258	1.328	NA	3.931	0.893	4.662	1.873	1.132	0.747	0.738
	100	3.048	9.455	0.55	2.703	NA	7.156	1.806	9.224	3.675	2.412	1.444	1.409
	250	7.077	23.848	0.789	7.754	NA	18.827	4.491	22.946	8.81	5.079	3.244	3.285
MAE	10	37.458	79.459	29.663	51.391	69.393	83.281	30.899	106.478	43.167	30.213	23.255	21.96
	25	83.302	224.442	56.003	114.552	109.815	199.736	68.858	246.374	92.75	66.028	50.217	48.364
	50	183.396	465.533	100.959	222.596	NA	473.89	139.529	466.556	216.876	143.9	100.715	96.639
	100	365.033	945.488	194.688	453.592	NA	787.951	276.815	923.643	438.872	299.393	204.371	195.749
	250	896.782	2384.753	491.336	1154.62	NA	2015.285	682.064	2297.192	1077.574	670.359	488.388	482.877
SRE	10	1.882	2.106	1.159	2.529	2.221	3.09	1.991	1.205	1.886	1.491	1.214	1.174
	25	2.136	2.387	1.094	2.75	2.318	4.552	2.209	1.261	2.132	1.539	1.164	1.125
	50	2.233	2.506	0.978	2.915	NA	5.408	2.323	2.272	2.274	1.523	1.09	1.035
	100	2.225	2.532	0.982	2.926	NA	4.051	2.33	2.274	2.25	1.514	1.066	1.002
	250	2.256	2.479	0.933	2.907	NA	4.143	2.344	2.3	2.265	1.346	0.972	0.937

existing methods. However, the Gmedian algorithm suits applications where the outlying distribution and the true distribution have the same theoretical mean but a different covariance structure. In contrast, the ROBPCA algorithm works well where there is a significant change in mean between the outlying distribution and the true distribution. The proposed rPCAdpd algorithm, suited for similar scenarios with changes in mean, surpasses ROBPCA at high robustness parameter α, and is significantly better in high dimensions. The PCP algorithm (Candès et al., 2011) has consistent results across simulation setups (PCA1), (PCA2a) and (PCA2b). This is due to the fact that the error comes only from the perturbation matrix $\boldsymbol{E}$ in Eq. (2.1), which is non-estimable by the PCP method. Table 3.5 and 3.6 summarises the results obtained for scenario (PCA2c) and (PCA2d) respectively. These results closely mirror those in scenarios (PCA2a) and (PCA2b).

Table 3.7: Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA3a).

Metric	p	Classical	LOC	ROBPCA	Proj	RobCov	Grid	Gmed	PCP	DPD (0.25)	DPD (0.5)	DPD (0.75)	DPD (1)
Bias	10	0.212	0.743	0.23	0.263	0.429	0.519	0.05	1.066	0.2	0.138	0.082	0.064
	25	0.534	2.185	0.403	0.494	0.412	1.09	0.127	2.452	0.515	0.324	0.247	0.247
	50	1.218	4.58	1.018	0.997	NA	2.249	0.296	4.717	1.159	0.715	0.493	0.455
	100	2.39	9.389	1.431	1.876	NA	4.388	0.597	9.372	2.218	1.201	0.873	0.857
	250	5.09	23.783	2.675	2.997	NA	9.718	1.631	23.309	4.503	1.83	1.424	1.451
MAE	10	32.656	74.327	30.411	42.304	48.105	66.025	24.609	106.512	34.086	29.286	23.804	21.794
	25	76.302	218.539	58.718	82.371	62.299	144.168	53.57	245.173	78.951	61.355	54.034	54.426
	50	160.631	458.022	131.667	166.613	NA	302.974	113.246	471.68	166.973	127.601	105.154	100.407
	100	303.042	938.894	218.99	303.707	NA	503.095	218.692	937.193	314.138	220.16	188.221	189.425
	250	815.964	2378.289	589.99	858.143	NA	1279.957	662.041	2330.966	817.858	559.381	515.87	516.502
SRE	10	1.822	1.882	1.152	1.945	1.537	2.698	1.822	1.155	1.818	1.636	1.292	1.158
	25	2.154	2.272	1.003	2.128	1.048	3.757	2.185	1.142	2.155	1.4	1.054	1.053
	50	2.204	2.368	1.017	2.152	NA	4.395	2.287	0.978	2.206	1.423	0.961	0.919
	100	2.251	2.475	1.02	2.279	NA	3.587	2.311	1.01	2.23	1.282	0.951	0.955
	250	2.264	2.521	1.072	2.202	NA	3.525	2.343	1.015	2.16	1.165	0.98	0.98

Table 3.8: Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA3b).

Metric	p	Classical	LOC	ROBPCA	Proj	RobCov	Grid	Gmed	PCP	DPD (0.25)	DPD (0.5)	DPD (0.75)	DPD (1)
Bias	10	0.454	0.756	0.376	0.415	0.509	0.749	0.172	1.065	0.478	0.399	0.283	0.207
	25	0.957	2.198	0.529	0.828	1.049	1.596	0.363	2.448	0.968	0.817	0.523	0.416
	50	2.053	4.603	0.829	1.571	NA	3.221	0.729	4.613	2.108	1.738	0.952	0.674
	100	4.085	9.41	1.438	2.712	NA	5.838	1.174	9.136	4.235	3.121	1.585	1.015
	250	10.553	23.793	5.036	8.705	NA	17.022	4.318	22.72	10.683	8.138	5.083	3.947
MAE	10	52.663	75.648	40.059	50.066	53.751	81.22	29.654	106.464	56.791	50.076	39.234	32.751
	25	113.553	219.799	73.592	109.254	119.102	191.308	75.088	244.825	119.018	106.796	78.32	67.627
	50	236.518	460.337	122.695	206.081	NA	411.252	144.926	461.708	249.263	218.761	141.867	114.571
	100	492.217	941.042	244.944	399.289	NA	647.709	288.144	914.552	534.963	445.369	293.533	237.476
	250	1175.092	2379.306	659.666	1087.267	NA	1791.531	725.623	2273.85	1231.25	1015.697	716.547	604.369
SRE	10	1.839	1.993	1.134	2.326	1.465	2.825	1.886	1.198	1.85	1.797	1.556	1.351
	25	2.22	2.414	1.071	2.784	2.217	4.183	2.249	1.234	2.237	2.026	1.356	1.176
	50	2.304	2.528	0.97	2.888	NA	4.909	2.358	2.287	2.312	2.149	1.43	1.196
	100	2.286	2.491	1.034	2.838	NA	3.809	2.307	2.288	2.309	1.993	1.223	0.995
	250	2.307	2.481	0.952	2.83	NA	3.847	2.34	2.308	2.317	1.997	1.348	1.172

Table 3.9: Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA3c).

Metric	p	Classical	LOC	ROBPCA	Proj	RobCov	Grid	Gmed	PCP	DPD (0.25)	DPD (0.5)	DPD (0.75)	DPD (1)
Bias	10	0.149	0.746	0.228	0.181	0.464	0.505	0.123	1.065	0.163	0.081	0.043	0.041
	25	0.357	2.188	0.258	0.402	0.543	1.057	0.381	2.457	0.382	0.174	0.148	0.143
	50	0.777	4.59	0.33	0.455	NA	2.011	0.676	4.738	0.823	0.296	0.265	0.253
	100	1.517	9.387	0.667	1.276	NA	4.291	1.144	9.402	1.495	0.442	0.392	0.372
	250	3.886	23.785	1.012	2.727	NA	9.358	3.513	23.383	3.883	1.464	1.386	1.359
MAE	10	25.443	74.609	32.294	37.131	50.803	64.494	22.821	106.453	26.557	19.682	15.649	15.393
	25	59.571	218.815	53.687	74.807	76.628	147.007	60.024	245.654	65.782	44.871	42.176	41.994
	50	133.973	459.029	108.565	159.38	NA	319.77	123.649	473.841	141.108	87.353	84.297	84.503
	100	256.543	938.692	193.825	284.691	NA	496.1	225.919	940.223	263.121	159.77	154.23	154.083
	250	676.504	2378.474	524.335	711.74	NA	1230.442	649.349	2338.346	678.719	435.267	427.714	430.439
SRE	10	1.796	2.006	1.052	2.079	1.468	2.768	1.872	1.144	1.788	1.286	0.952	0.958
	25	2.141	2.342	0.981	2.397	1.775	3.969	2.211	1.186	2.057	0.994	0.861	0.87
	50	2.179	2.418	0.954	2.389	NA	4.572	2.289	0.858	2.038	0.908	0.841	0.851
	100	2.23	2.487	1.053	2.476	NA	3.66	2.315	0.929	1.983	0.899	0.849	0.883
	250	2.254	2.555	0.999	2.479	NA	3.719	2.364	0.927	1.908	0.841	0.826	0.834

Table 3.10: Estimated Bias, Mean Absolute Error and Subspace Recovery Error (SRE) for different PCA algorithms for the simulation scenario (PCA3d).

Metric	p	Classical	LOC	ROBPCA	Proj	RobCov	Grid	Gmed	PCP	DPD (0.25)	DPD (0.5)	DPD (0.75)	DPD (1)
Bias	10	0.268	0.79	0.139	0.347	0.633	0.724	0.183	1.066	0.324	0.217	0.115	0.113
	25	0.673	2.231	0.243	0.71	0.725	1.615	0.486	2.461	0.748	0.458	0.304	0.297
	50	1.325	4.641	0.13	1.474	NA	3.602	0.709	4.649	1.669	0.934	0.522	0.481
	100	2.772	9.436	0.101	2.798	NA	6.713	1.462	9.206	3.428	2.046	1.142	1.098
	250	6.861	23.832	0.34	6.809	NA	17.565	3.587	22.922	8.555	5.192	3.156	2.813
MAE	10	35.048	79.01	31.374	49.556	67.262	81.659	27.681	106.519	39.974	32.114	20.677	19.517
	25	83.555	223.086	58.178	105.35	112.409	201.416	70.421	246.107	91.476	65.052	49.229	46.971
	50	179.148	464.061	112.288	227.737	NA	458.559	129.521	465.585	210.767	138.549	96.856	89.979
	100	370.199	943.586	216.485	440.313	NA	738.531	272.546	921.936	430.631	287.974	191.931	187.771
	250	908.806	2383.166	566.649	1118.667	NA	1918.17	688.634	2294.489	1054.348	709.767	492.689	450.457
SRE	10	1.849	2.045	1.084	2.42	2.053	3.117	1.955	1.026	1.851	1.608	1.225	1.16
	25	2.145	2.383	0.948	2.836	2.283	4.553	2.233	1.077	2.134	1.409	1.04	0.965
	50	2.235	2.495	0.983	2.951	NA	5.331	2.34	2.274	2.286	1.496	1.112	1.022
	100	2.269	2.529	0.985	2.928	NA	3.971	2.351	2.311	2.296	1.503	1.01	0.953
	250	2.28	2.506	1.039	3.009	NA	4.132	2.388	2.326	2.295	1.551	1.176	1.076

In scenarios (PCA3a)-(PCA3d), the contaminating distribution changes to t-distribution with 5 degrees of freedom which has a heavier tail. In these scenarios, ROBPCA (Hubert et al., 2005) and Gmedian (Cardot and Godichon-Baggioni, 2017) algorithms and the proposed rPCAdpd methods perform closely. In (PCA3a), the rPCAdpd method excels for large values of α. As the contamination rises to 20%, as shown in Table 3.8, all of the chosen algorithms show a significant increase in MAE. However, the proposed estimator maintains a low bias for all components even for large p relative to n, consistent with its theoretical breakdown point behaviour as pointed out in Section 3.4.5.

In essence, the proposed rPCAdpd algorithm excels at detecting and removing low-variance, different-location contaminating components, compared to the primary data distribution component. In all other cases, its performance is closely comparable to the best of the existing algorithms. Also, across all of the simulation setups considered, the proposed rPCAdpd algorithm yields significantly better estimates of principal components than the existing algorithms when the dimension of the data p is large, which is also theoretically justified by its dimension-free asymptotic breakdown point.

3.6 Real Data Analysis

In this section, we demonstrate applications of the proposed rPCAdpd estimator on three real-life datasets. The first two datasets, namely the "Car dataset" and the "Octane dataset" are popular benchmark datasets used to compare performances of different RPCA algorithms (see Hubert et al. (2005) for details). We also consider a novel "Credit Card Fraud Detection dataset" to demonstrate how the proposed estimator can serve as a preliminary preprocessing step to identify fraudulent credit card transactions before applying any sophisticated binary classification algorithms.

3.6.1 Car Dataset

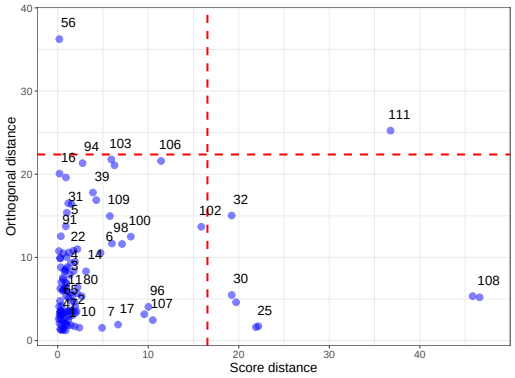
The Car dataset comprises $n = 111$ observations of cars with $p = 11$ variables, including the length, width, and height of the car. This dataset has served as a benchmark for various robust PCA methods (Hubert et al., 2005, Croux et al., 2007). We utilize it to assess the performance of our proposed rPCAdpd method on outlier detection. For visual evaluation, we adopt the orthogonal distances and the score distances as diagnostic metrics (Hubert et al., 2005).

Analysing the screeplots for both the rPCAdpd and the classical PCA for the Car dataset reveals that the first four principal components capture more than 90% of the variance. Therefore, we apply both these algorithms to extract the first four principal components only and compute the orthogonal distances and the score distances for each observation. Figure 3.1 illustrates this diagnostic analysis. Classical PCA identifies a cluster of influential points (observations $25, 30, 32, 34$, and 36), which are also detected by the rPCAdpd estimator. These points share a value of (-2) for 4 of the 11 variables: `Rear.Hd`, `Rear.Seat`, `Rear.Shld`, and `Luggage`. However, classical PCA assigns low orthogonal dis-

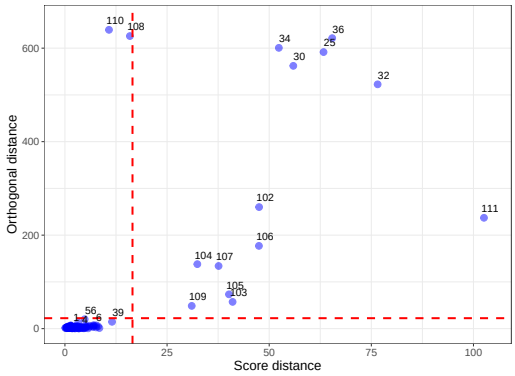
tances to these outliers, indicating their good fit, thus inflating distances for almost all the points. Conversely, rPCAdpd assigns high orthogonal distances to these outliers. Additionally, rPCAdpd identifies a different set of outliers (observations $102 - 107, 109$), unnoticed by the traditional PCA, which is consistent with the findings by Hubert et al. (2005). As demonstrated in Figure 3.1, the ROBPCA and Gmedian algorithms also spot these outliers.

Figure 3.1: Diagnostic plots for the Car dataset.

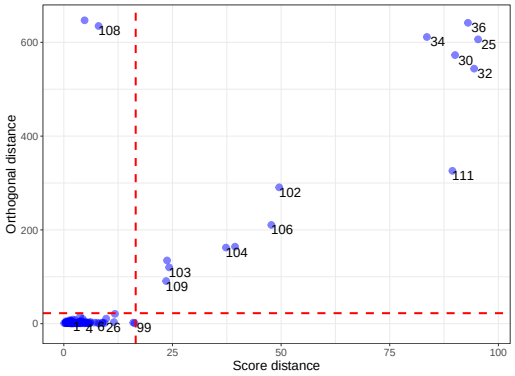
(a) Classical PCA.



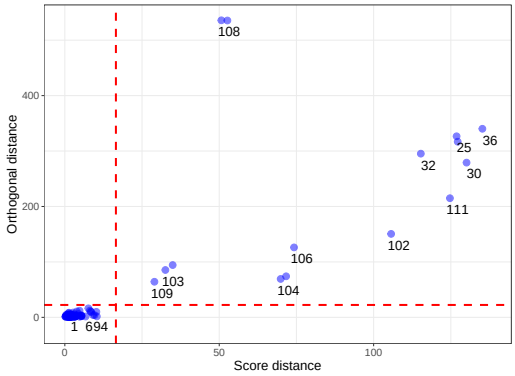
(b) rPCAdpd.



(c) ROBPCA.



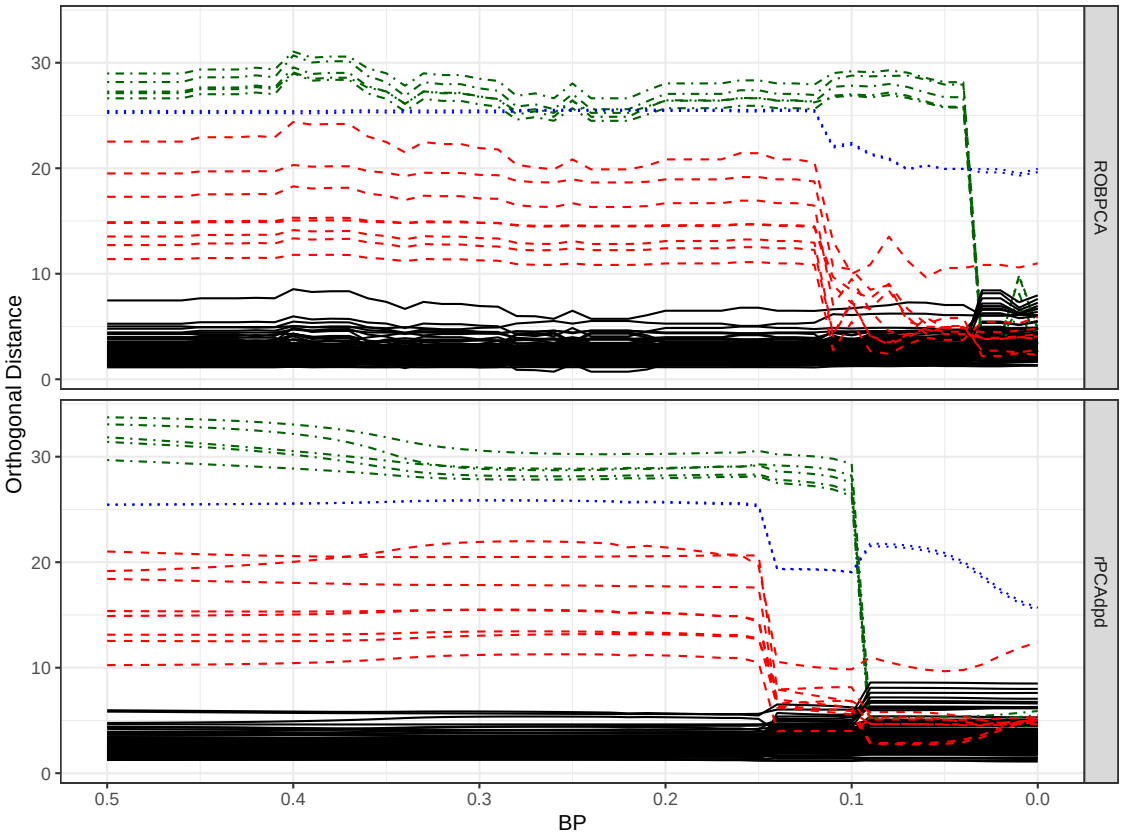
(d) Gmedian.



Another interesting way to compare the robustness of different methods is through the monitoring approach considered in Riani et al. (2014, 2020). The idea is to monitor an error metric for each observation (e.g., residuals in linear regression model) as the tuning parameters of the robust method are varied along with its asymptotic breakdown point. In the case of principal component analysis, one may track the orthogonal distances for each observation obtained from different robust PCA methods as the breakdown point changes from 0.5 (the maximum possible value) to 0 (the minimum possible value). For example, for the proposed rPCAdpd method, we know that the asymptotic breakdown point is at least $\alpha/(1+\alpha)$ when the robustness tuning parameter is α (see Section 3.4.5). Therefore, conversely, given the breakdown level at b, we can consider the tuning parameter for rPCAdpd algorithm as $\alpha = b/(1-b)$. For the ROBPCA algorithm, given the breakdown point b, the tuning parameter α denoting the proportion of observations considered for the

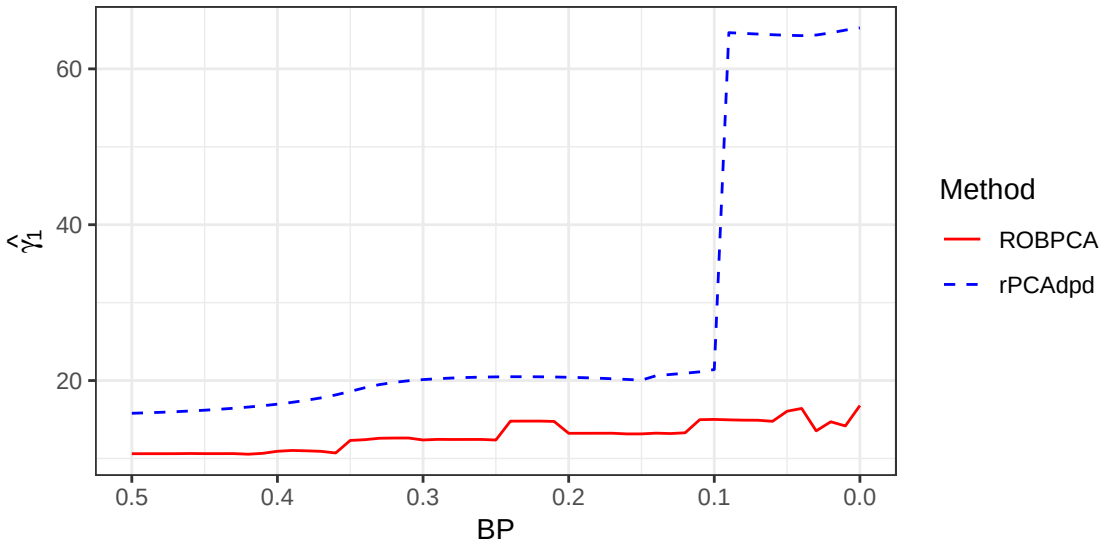
MCD type covariance estimation, satisfy $\alpha = (1-b)$. For these two algorithms, we indicate the monitoring plots in Figure 3.2. As seen from the figure, there are multiple groups of observations as per their orthogonal distances, and as the breakdown point decreases, they merge together to form a single group. However, for the ROBPCA algorithm, the two bad leverage points (cases 108 and 110) continue to demonstrate very high orthogonal distance, even when the breakdown point becomes 0.

Figure 3.2: Monitoring plot for the Cars dataset.



In order to investigate further, we consider another monitoring plot where we track the estimated first eigenvalue for each of these methods as the breakdown point varies. This plot, shown in Figure 3.3, clearly demonstrates that the ROBPCA algorithm shrinks the eigenvalue even when the tuning parameter is adjusted such that the breakdown point is equal to 0. However, the rPCAdpd estimate of the first eigenvalue increases as expected and becomes same as the maximum likelihood estimate as both breakdown b and tuning parameter α reduce to 0. Hence, the rPCAdpd estimator retains a desirable balance between robustness and efficiency and does not apply unnecessary shrinkage to the estimate. The figure also brings out that the rPCAdpd estimator starts providing a robust estimate for any $\alpha \geq 0.1/(1-0.1) \approx 0.11$.

Figure 3.3: Monitoring plot for first principal eigenvalue of the Cars dataset.



3.6.2 Octane Data

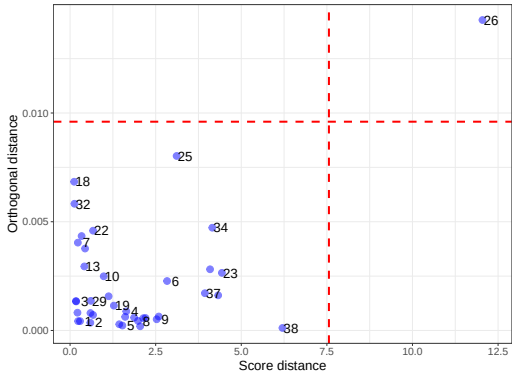
The Octane dataset, sourced from Esbensen et al. (2002), features spectroscopic data with octane numbers derived from near-infrared (NIR) absorbance spectra of 39 gasoline samples. The measurements span across 226 electromagnetic radiation wavelengths (1102 nm to 1552 nm), each of which gives rise to a feature variable. With 39 observations and 226 features, principal component analysis (PCA) is pivotal for dimension reduction and subsequent analysis. Six sample observations (i.e., serial numbers 25, 26, and $36-39$) are known to contain additional alcoholic substances, making them quite different from the others (Hubert et al., 2005).

As in the case of the Car dataset, a screeplot analysis reveals that there are only 2 significant principal components present in the Octane dataset. However, the first principal value estimated by the classical PCA (0.132) is several magnitudes higher than the first principal value estimated by rPCAdpd (0.01075), which aligns with the estimates obtained from existing robust PCA algorithms (Hubert et al., 2005). Diagnostic plots in Figure 3.4 demonstrate that traditional PCA fails to detect the outliers, except observation 26, while the rPCAdpd algorithm identifies all the alcohol-mixed gasoline sample observations accurately. The ROBPCA algorithm also detects these outliers, with a similar score and orthogonal distances.

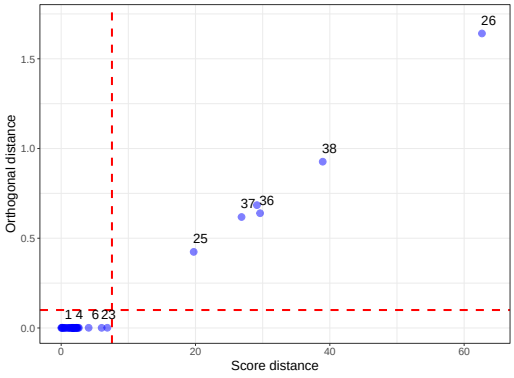
As in Section 3.6.1, we indicate the monitoring plots in Figures 3.5 and 3.6. Again in this case, there are few different groups of outliers. As the breakdown decreases to 0, all the orthogonal distances are merged together to provide a non-robust fit to the principal components. Comparing the monitoring plots, we see that even with a very small positive tuning parameter $\alpha \approx 0.025$, the rSVDdpd algorithm is able to provide a robust estimate of the principal components, while ROBPCA algorithm obtains the robust estimate much later when $\alpha \approx 0.13$. Figure 3.6 yields a similar conclusion showing a robust

Figure 3.4: Diagnostic plots for the Octane dataset.

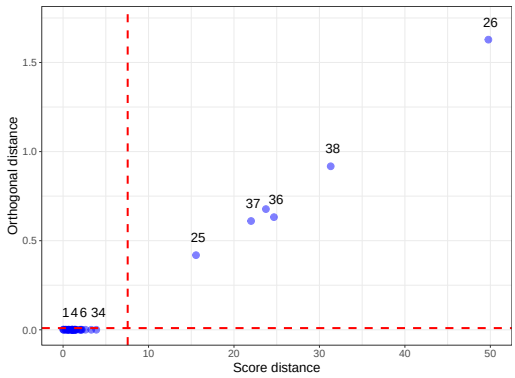
(a) Classical PCA.



(b) rPCAdpd.



(c) ROBPCA.



(d) GMedian.

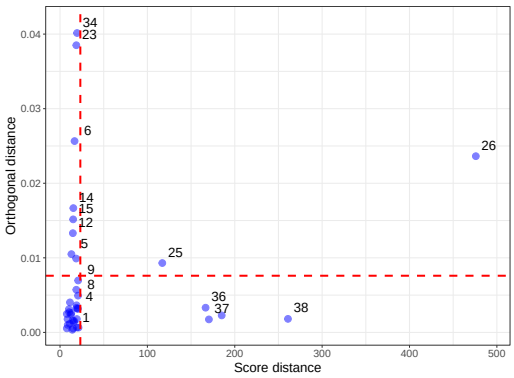
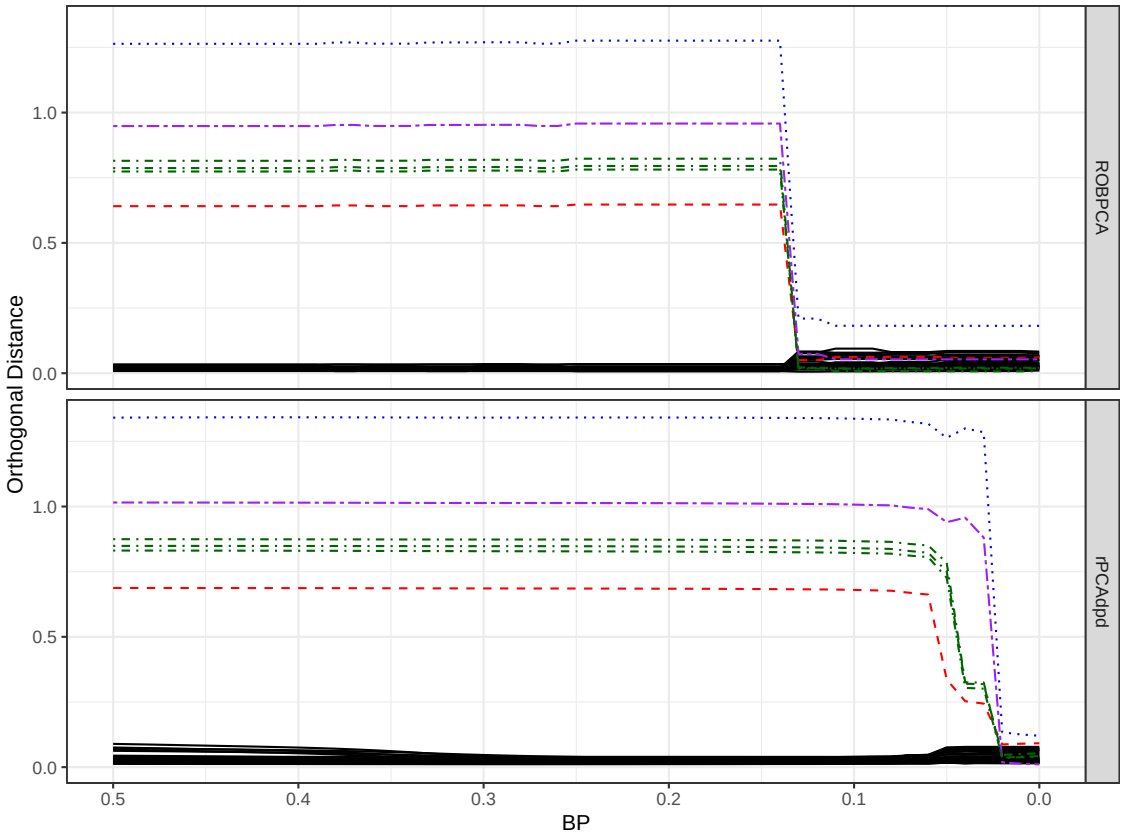


Figure 3.5: Monitoring plot for the Octane dataset.



estimate ($\widehat{\gamma}_1 = 0.015$) of the first principal eigenvalue as the existing ROBPCA method while moving to the efficient maximum likelihood estimate ($\widehat{\gamma}_1 = 0.15$) as the breakdown decreases to 0.

3.6.3 Credit Card Fraud Detection

Credit card fraud detection is a very challenging problem because of the specific nature of transaction data and the labelling process. Most of the real-life transaction data are highly imbalanced, and the number of fraudulent transactions is far too small compared to the extremely large number of valid transactions made on a day-to-day basis. There are primarily two kinds of strategies to detect such fraudulent transactions: the first one models the situation as a binary classification problem with a resampling procedure (e.g., undersampling, oversampling) to counter the class imbalance issue. The second approach assumes that the fraudulent transactions are outliers in the data and applies an outlier detection algorithm (Carcillo et al., 2018, 2019). These methods often begin with a principal component analysis (PCA) to reduce the dimension of the data and training time for a real-time application.

To this end, we anticipate that the proposed robust PCA algorithm will outperform traditional PCA in dimensionality reduction and provide reliable principal component estimates. We demonstrate this using the Credit Card Fraud Detection Dataset

Figure 3.6: Monitoring plot for first principal eigenvalue of the Octane dataset.

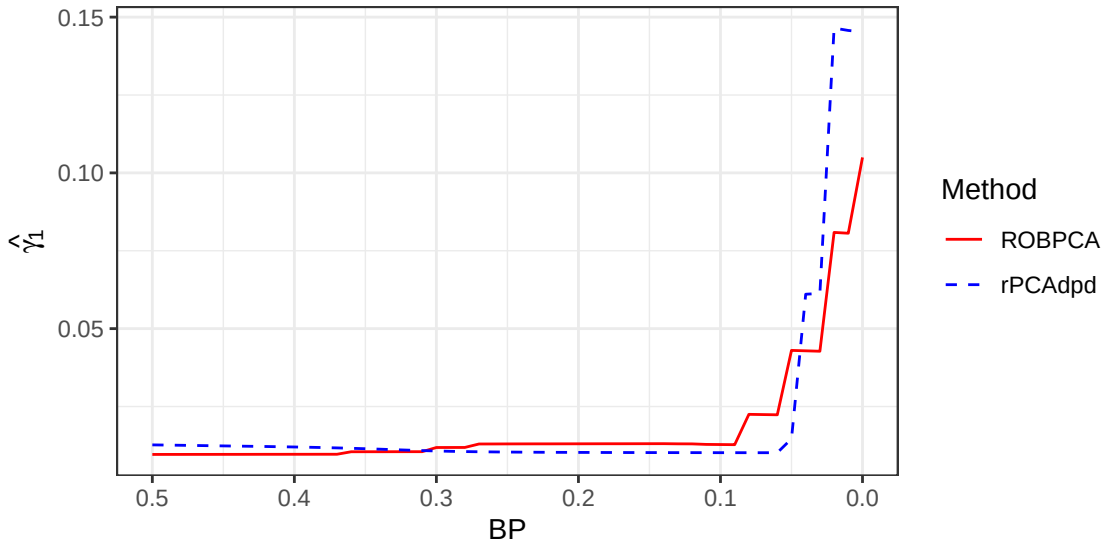
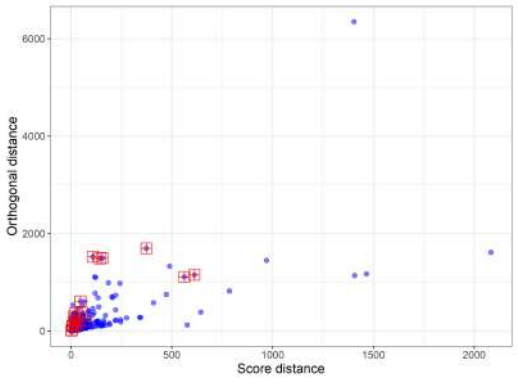
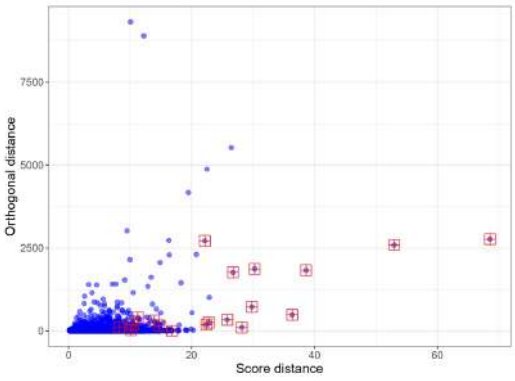


Figure 3.7: Diagnostic plots for the Credit Card dataset for different robust PCA methods.

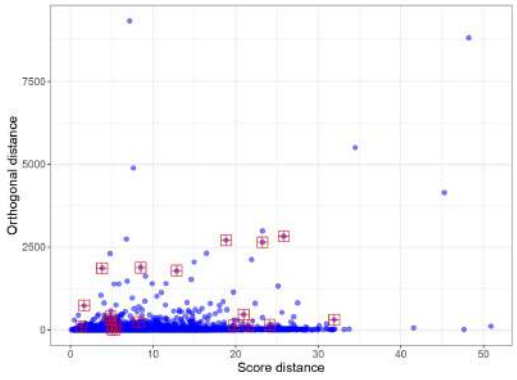
(a) Classical PCA.



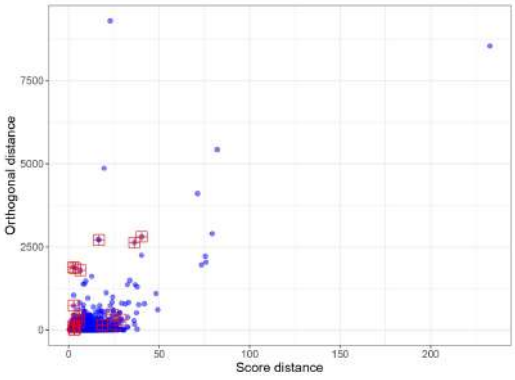
(b) rPCAdpd.



(c) ROBPCA.



(d) GMedian.



from Le Borgne and Bontempi (2004). The dataset encompasses 28 anonymized features over 284807 transactions, with only 0.1% (492) transactions being fraudulent. For demonstration, we randomly sample 5% of the dataset, which includes 19 fraudulent transactions. The first 5 principal components, explaining over 80% of variation, are retained for both the traditional and RPCA algorithms. In Figure 3.7, we portray the orthogonal distance and the score distance for each algorithm in consideration, and mark the fraudulent transactions with red squares. As evident from Figure 3.7, the classical PCA method fails to separate most of the fraudulent transactions, correctly identifying only 5 of them. In contrast, the rPCAdpd algorithm separates 13 out of 19 outliers. Existing robust PCA methods such as ROBPCA and GMedian spot 7 and 6 outliers respectively, which are marginally better than classical PCA but at the cost of many false positives (outliers without the red squares). Thus, substituting the traditional PCA with the robust rPCAdpd algorithm in the preprocessing step of this analysis can greatly enhance the projection onto a reduced-dimensional space. By doing so, valuable insights about fraudulent transactions can assist the existing outlier detection and classification algorithms on this transformed, lower-dimensional data.

3.A Appendix: Proofs of some additional results

Lemma 3.A.1. *For an elliptically symmetric family of densities*

$$f(\boldsymbol{x}) \propto \exp\left[g\left((\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right],\ \boldsymbol{x}\in\mathbb{R}^d,$$

the normalizing constant for the density is given by $c_g \det(\boldsymbol{\Sigma})^{1/2}$ *where* c_g *is a constant depending on the choice of function* g*.*

Proof. To see this, we note that the normalizing constant can be expressed as

$$\mathcal{C}_g = \int_{\mathbb{R}^p} \exp\left[g\left((\boldsymbol{x}-\boldsymbol{\mu})^\intercal \sum_{k=1}^{p} \gamma_k^{-1}\boldsymbol{v}_k\boldsymbol{v}_k^\intercal(\boldsymbol{x}-\boldsymbol{\mu})\right)\right] d\boldsymbol{x}.$$

Let $\boldsymbol{P}$ be the $p\times p$ orthogonal matrix whose rows are the vectors $\boldsymbol{v}_k^\intercal$ for $k=1,2,\ldots,p$. Then, applying a change of variable $\boldsymbol{z}=\boldsymbol{P}^\intercal(\boldsymbol{x}-\boldsymbol{\mu})$, we can rewrite the integral as

$$\mathcal{C}_g = \int_{\mathbb{R}^p} \exp\left[g\left(\sum_{k=1}^{p}\gamma_k^{-1}z_k^2\right)\right] d\boldsymbol{z},$$

where $\boldsymbol{z}=(z_1,z_2,\ldots,z_p)^\intercal$. Finally, another change of variable with $u_k = z_k/\sqrt{\gamma_k}$ for $k=1,2,\ldots,p$ yields,

$$\mathcal{C}_g = \int_{\mathbb{R}^p} \prod_{k=1}^{p}\gamma_k^{1/2} \exp\left[g\left(\sum_{k=1}^{p}u_k^2\right)\right] du_1du_2\ldots du_p = \det(\boldsymbol{\Sigma})^{1/2}c_g,$$

where the constant c_g is the integral it is replacing. Clearly, the term c_g is free of the mean $\boldsymbol{\mu}$ and the dispersion $\boldsymbol{\Sigma}$ matrix, and hence is a constant depending only on the g function. □

Lemma 3.A.2. *For any general setup with thrice differentiable model density function* $f_{\boldsymbol{\theta}}(x)$*, let* $c_\alpha(\boldsymbol{\theta}) = \|f_{\boldsymbol{\theta}}\|_{1+\alpha}^{1+\alpha}$*. Then, under the assumption of exchangeability of the differentiation and integral signs, we have*

$$\boldsymbol{\xi}_{\boldsymbol{\theta}} = (1+\alpha)^{-1}c_\alpha(\boldsymbol{\theta})\partial_{\boldsymbol{\theta}}\log(c_\alpha(\boldsymbol{\theta})).$$

Proof. Let, $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = c_\alpha^{-1}(\boldsymbol{\theta})f_{\boldsymbol{\theta}}^{(1+\alpha)}(\boldsymbol{x})$ be another density function. Note that

$$u_{\boldsymbol{\theta}}^h(\boldsymbol{x}) = \frac{\partial}{\partial\boldsymbol{\theta}}\log(h_{\boldsymbol{\theta}}(\boldsymbol{x})) = -\frac{\partial}{\partial\boldsymbol{\theta}}\log(c_\alpha(\boldsymbol{\theta})) + (1+\alpha)u_{\boldsymbol{\theta}}(\boldsymbol{x}), \tag{3.13}$$

where $u_\theta(\boldsymbol{x})$ is the score function corresponding to $f_{\boldsymbol{\theta}}(\boldsymbol{x})$. Under the standard regularity conditions, one can exchange the differentiation and the integral sign to obtain that the expectation of the score function is equal to 0. Therefore,

$$0 = \int \frac{\partial}{\partial\boldsymbol{\theta}}\log(h_{\boldsymbol{\theta}}(\boldsymbol{x}))h_{\boldsymbol{\theta}}(\boldsymbol{x})d\boldsymbol{x} = -\frac{\partial}{\partial\boldsymbol{\theta}}\log(c_\alpha(\boldsymbol{\theta})) + \frac{(1+\alpha)}{c_\alpha(\boldsymbol{\theta})}\xi_{\boldsymbol{\theta}}.$$

Interchanging the sides and solving for $\xi_{\boldsymbol{\theta}}$ yields the result. □

Lemma 3.A.3. *For any general setup with thrice differentiable model density function $f_{\boldsymbol{\theta}}(x)$, let $c_\alpha(\boldsymbol{\theta}) = \|f_{\boldsymbol{\theta}}\|_{1+\alpha}^{1+\alpha}$. Then, under the assumption of exchangeability of the differentiation and integral signs, we have*

$$\boldsymbol{J}_{\boldsymbol{\theta}} = \frac{c_\alpha(\boldsymbol{\theta})}{(1+\alpha)^2}\left(i^h(\boldsymbol{\theta}) + \left(\frac{\partial}{\partial \boldsymbol{\theta}}\log(c_\alpha(\boldsymbol{\theta}))\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}}\log(c_\alpha(\boldsymbol{\theta}))\right)^\intercal\right) \tag{3.14}$$

where $i^h(\boldsymbol{\theta})$ is the expected Fisher information matrix for a single observation $\boldsymbol{x}$ following the density function $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = c_\alpha^{-1}(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}^{(1+\alpha)}(\boldsymbol{x})$.

Proof. Starting with the equality (3.13), it follows that

$$\begin{aligned}\left(u_{\boldsymbol{\theta}}^h(\boldsymbol{x})\right)\left(u_{\boldsymbol{\theta}}^h(\boldsymbol{x})\right)^\intercal = &\left(\frac{\partial}{\partial \boldsymbol{\theta}}\log(c_\alpha(\boldsymbol{\theta}))\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}}\log(c_\alpha(\boldsymbol{\theta}))\right)^\intercal \\ &- 2(1+\alpha)u_{\boldsymbol{\theta}}(\boldsymbol{x})\left(\frac{\partial}{\partial \boldsymbol{\theta}}\log(c_\alpha(\boldsymbol{\theta}))\right)^\intercal + (1+\alpha)^2 u_{\boldsymbol{\theta}}(\boldsymbol{x})u_{\boldsymbol{\theta}}^\intercal(\boldsymbol{x}).\end{aligned}$$

Multiplying both sides with $h_{\boldsymbol{\theta}}(\boldsymbol{x})$ and integrating with respect to $\boldsymbol{x}$ yields

$$i^h(\boldsymbol{\theta}) = \left(\frac{\partial_{\boldsymbol{\theta}} c_\alpha(\boldsymbol{\theta})}{c_\alpha(\boldsymbol{\theta})}\right)\left(\frac{\partial_{\boldsymbol{\theta}} c_\alpha(\boldsymbol{\theta})}{c_\alpha(\boldsymbol{\theta})}\right)^\intercal - 2\left(\frac{\partial_{\boldsymbol{\theta}} c_\alpha(\boldsymbol{\theta})}{c_\alpha(\boldsymbol{\theta})}\right)\left(\frac{\partial_{\boldsymbol{\theta}} c_\alpha(\boldsymbol{\theta})}{c_\alpha(\boldsymbol{\theta})}\right)^\intercal + \frac{(1+\alpha)^2}{c_\alpha(\boldsymbol{\theta})}\boldsymbol{J}_{\boldsymbol{\theta}}.$$

Solving for $\boldsymbol{J}_{\boldsymbol{\theta}}$ yields Eq. (3.14). □

Chapter 4

Rank Estimation

4.1 Introduction

In previous Chapters 2 and 3, we introduced a novel estimator to robustly estimate the singular values, vectors and the principal components of a signal-plus-noise data matrix $\boldsymbol{X}$ as in (2.1), provided an apriori knowledge about the rank r of its low-dimensional component. In other existing robust SVD or PCA methods (Candès et al., 2011, Zhang et al., 2013), some tuning parameter tied to the low-rank component $\boldsymbol{L}$ must be specified, which indirectly controls the rank. Sometimes, this rank is guessed based on the domain knowledge. For instance, we usually take rank to be 1 or 2 for the background modelling problem, between 3 and 5 in image reconstruction, etc. However, for many practical purposes, this rank is not known and must be estimated in a data-dependent fashion.

Primarily there are two approaches to determine the rank of the low-rank component $\boldsymbol{L}$. The first approach relies on creating a general model selection criterion based on distributional assumptions on the entries of the error matrix $\boldsymbol{E}$. The model selection criterion is built by combining a statistical discrepancy measure of fit from a partial SVD or similar estimate and a penalty function of the model complexity. The estimate of the rank is then given by the choice of rank r that minimizes the criterion function. Classical criteria such as Akaike's Information Criterion (AIC) (Akaike, 1973) and Bayesian Information Criterion (BIC) (Schwarz, 1978) are usually inadequate for this purpose, and they need to be appropriately modified for the matrix rank estimation problem (Bai and Ng, 2002).

The second major approach estimates the rank of a matrix using a resampling technique known as cross-validation. In this case, the rows and the columns of the data matrix $\boldsymbol{X}$ are subdivided into groups; these groups are deleted in each step and a partial SVD up to a chosen rank is used to estimate the deleted entries based on the remaining elements. The estimate of the rank is then determined by the choice yielding the lowest possible prediction error. Wold (1978), Eastment and Krzanowski (1982), Gabriel (2002) and Owen and Perry (2009) are examples of such cross-validation techniques. The theoretical results on various statistical properties of the cross-validation techniques are rather limited, compared to the penalization approaches. Later in Sections 4.2.1 and 4.2.2, we describe these approaches in detail.

Beyond these two classes of estimates, there are a few other approaches present in the literature. The simplest one is the "Elbow" method, conceptualized by Thorndike (1953), which proceeds by looking at the estimated singular values (or the squared estimated eigenvalues) of the data matrix $\boldsymbol{X}$ and finding the position in the screeplot where a major slope change occurs in the graph of singular values, leading to the shape of an elbow. In contrast, Shabalin and Nobel (2013) proposed another simple rank estimation method by counting the number of singular values beyond a certain threshold as estimated from data matrix $\boldsymbol{X}$. Their studies involve deriving a distributional result of the singular values for a random signal-plus-noise matrix. Recently, Xu et al. (2021) presented a similar thresholding approach, but using the Gershgorin discs of the singular values of the data matrix, for robust PCA methods. In a different direction, Hoff (2007) proposed a matrix rank estimation method using Bayesian modelling of the data matrix $\boldsymbol{X}$. His approach considers the von-Mises distribution for the singular vectors and uses the properties of this distribution to derive the posterior distribution of the rank. Gibb's sampling is then used to generate the posterior samples. Since his approach relies heavily on the properties of the classical SVD, it cannot be readily applied to a robust SVD procedure such as rSVDdpd without any non-trivial modifications.

Despite their popularity, the penalized and the cross-validation approaches perform poorly in the presence of outlying observations, as will be demonstrated later in Section 4.5.2. Even if a robust SVD procedure is used, the existing penalized approaches result in an unsatisfactory level of accuracy due to the misspecification of the penalty function, and the cross-validation approaches usually face some computational bottlenecks.

In this chapter, we present a novel penalized criterion for estimating the rank of a data matrix robustly. The proposed penalized criterion is a modified version of the divergence information criterion (DIC) as in Kurata and Hamada (2018) which uses the robust "rSVDdpd" algorithm introduced in Chapter 2 to estimate the singular values and vectors. As will be demonstrated in Section 4.5.2, this method provides a reliable estimate of the rank compared to the existing methods, even in the presence of contaminated data entries. The rest of the Chapter is organized as follows. In Section 4.2 we formally describe the rank estimation problem and describe the mathematical foundations of the penalization and cross-validation approaches to solve the problem. In Section 4.3, we derive our proposed criterion for robust rank estimation. Subsequently, we demonstrate its selection consistency and robustness in Section 4.4. Finally, we perform extensive simulation studies to demonstrate the superior performance of our proposed criterion over the existing rank estimation methods.

4.2 Problem Description and Existing Approaches

As in the robust SVD problem, we consider an $n \times p$ data matrix decomposition as

$$\boldsymbol{X} = \boldsymbol{L} + \frac{\sigma}{\sqrt{n \vee p}}\boldsymbol{E}, \tag{4.1}$$

where $\boldsymbol{L}$ is a low-rank matrix and $\boldsymbol{E}$ is a matrix with i.i.d. entries with density function g such that $\mathbb{E}(e_{ij}) = 0$ and $\text{Var}(e_{ij}) = 1$ for all $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$. Based on this decomposition (4.1), the aim is to robustly estimate $r = \text{Rank}\,(\boldsymbol{L})$. In Eq. (4.1), the usage of $\sqrt{n \vee p}$ in the denominator for the noise matrix is crucial. There are two reasons for this: One, the error variance should asymptotically go to 0 as n or p increases to infinity for the rank estimation problem to be identifiable, otherwise, the largest singular value of the noise matrix $\boldsymbol{E}$ and its corresponding singular vectors can be added to $\boldsymbol{L}$ to constitute a new low-rank component with a different rank than that of $\boldsymbol{L}$ (see Candès et al. (2011) for details of this identifiability issue). Two, as shown in Shabalin and Nobel (2013), asymptotically any reasonable matrix decomposition technique cannot consistently estimate the singular values of $\boldsymbol{L}$ (or its functions like the rank) from the data matrix $\boldsymbol{X}$ if the noise variance is larger than an order of $1/\sqrt{np}$. Also note that, this exact scaling factor is present in Assumption (B6) which was required to prove the statistical consistency of the robust rSVDdpd estimates earlier in Chapter 2.

4.2.1 Penalization Approaches

The penalization approach to the matrix rank estimation problem considers a model selection criterion combining a statistical discrepancy measure of fit from a partial SVD estimate to the data matrix $\boldsymbol{X}$ and a penalty function of the rank used in the partial SVD estimate. As one includes higher rank components in the estimate, the error in estimation reduces by yielding a small discrepancy measure, but the penalty function compensates by being an increasing function of the rank. In mathematical terms, the penalization approach considers a model selection criterion of the form

$$C(r) = d(\widehat{\boldsymbol{\theta}}^{(r)}, \boldsymbol{X}) + g(r), \;\; r = 0, 1, 2, \ldots, (n \wedge p), \tag{4.2}$$

where $d(\widehat{\boldsymbol{\theta}}^{(r)}, \boldsymbol{X})$ is a measure of discrepancy between $\boldsymbol{X}$ and the estimate of $\boldsymbol{X}$ based on a partial SVD estimate $\widehat{\boldsymbol{\theta}}^{(r)}$ up to rank r. The quantity $g(r)$ is an appropriately chosen penalty function of the rank r. The estimate of rank is then given by

$$\widehat{r} = \underset{0 \le r \le (n \wedge p)}{\arg\min} \; C(r).$$

For majority of the existing rank estimation algorithms, $d(\cdot)$ is taken to be the squared Frobenius norm of the difference between $\boldsymbol{X}$ and $\sum_{k=1}^{r} \widehat{\lambda}_k \widehat{u}_{ki} \widehat{v}_{kj}$, where $\widehat{\lambda}_k, \widehat{\boldsymbol{u}}_k$ and $\widehat{\boldsymbol{v}}_k$ are the estimates of the k-th largest singular value and corresponding left and right singular vectors. For the typical model selection criterion like Akaike's Information Criterion (AIC) (Akaike, 1973) and Bayesian Information Criterion (BIC) (Schwarz, 1978), the penalty function $g(r)$ does not turn out to be strictly increasing function of the rank r, but an inverted parabola like $r(n + p - r)$. This choice of penalty function usually leads to overestimation. Thus, Bai and Ng (2002) proposed six modified criteria named PC1, PC2, PC3, IC1, IC2 and IC3 for the matrix rank estimation problem. All of these criteria use a linearly increasing function of r as the penalty function.

Since the classical SVD estimates are highly influenced by the presence of outlying observations, one may substitute robust rSVDdpd estimates $\widehat{\boldsymbol{\theta}}_\alpha^{(r)}$ in calculating the discrepancy measure. However, due to the squared residuals in the form of the Frobenius norm, even these modified procedures still produce poor estimates of the rank, as will be shown later in Section 4.5.2. As an alternative, Karagrigoriou and Papaioannou (2008) proposed a "Divergence Information Criterion" (DIC) based on the density power divergence (Basu et al., 1998) which is demonstrated to be robust in selecting the correct model in regression setups. For the matrix rank estimation problem, the DIC is given as

$$\mathrm{DIC}_\alpha(r) = H_\alpha^{(r)}(\widehat{\boldsymbol{\theta}}_\alpha^{(r)}) + r(\alpha+1)(2\pi)^{-\alpha/2}\left(\frac{1+\alpha}{1+2\alpha}\right)^{3/2}, \tag{4.3}$$

where

$$H_\alpha^{(r)}(\boldsymbol{\theta}) = \sigma^{-\alpha}\left[\|f\|_{1+\alpha}^{1+\alpha} - \left(1+\frac{1}{\alpha}\right)\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p} f^\alpha\left(\frac{|X_{ij}-\sum_{k=1}^{r}\lambda_k u_{ki}v_{kj}|}{\sigma}\right)\right]. \tag{4.4}$$

Here, the parameter $\boldsymbol{\theta}$ consists of the singular values λ_k and corresponding left and right singular vectors $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ for $k = 1, 2, \ldots, (n \wedge p)$, and the noise variance σ^2. Note that, (4.4) is exactly same as the objective function (2.4) of the "rSVDdpd" estimator. Hence, the quantity $\widehat{\boldsymbol{\theta}}_\alpha^{(r)}$ denotes the "rSVDdpd" estimator up to rank r for the robustness tuning parameter α.

We know that as $\alpha \to 0$, the density power divergence takes on the form of Kullback-Leibler divergence and the corresponding MDPDE becomes the maximum likelihood estimator (MLE). It is easy to see that if f is taken to be the standard normal density, then

$$H_0^{(r),\phi}(\boldsymbol{\theta}) := \lim_{\alpha\to 0} H_\alpha^{(r),\phi}(\boldsymbol{\theta}) = (2\sigma^2)^{-1}\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(X_{ij} - \sum_{k=1}^{r}\lambda_k u_{ki}v_{kj}\right)^2, \tag{4.5}$$

where the superscript ϕ indicates that $H(\cdot)$ is calculated based on the standard normal density. Eq. (4.5) establishes that the limiting form of the discrepancy measure $H(\cdot)$ relates to the usual Frobenius norm used in the classical criteria like AIC, BIC, etc. This also means, the minimizer of $H_0^{(r),\phi}(\boldsymbol{\theta})$, i.e., $\widehat{\boldsymbol{\theta}}_0^{(r),\phi}$ is the traditional nonrobust SVD estimate up to rank r. To deal with the general case beyond normal distribution for the errors, Kurata and Hamada (2018) proposes a new criterion called BHHJ-C given by

$$\text{BHHJ-C}_\alpha = H_\alpha^{(r)}(\widehat{\boldsymbol{\theta}}_\alpha^{(r)}) + \frac{1}{n}\mathrm{Trace}\left(J_\alpha^{-1}(\widehat{\boldsymbol{\theta}}_\alpha^{(r)})K_\alpha(\widehat{\boldsymbol{\theta}}_\alpha^{(r)})\right), \tag{4.6}$$

where

$$J_\alpha(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{E}_{g_{ij}}\left[\nabla_{\boldsymbol{\theta}}^2 V_{ij,\alpha}^{(r)}(\boldsymbol{\theta})\right],$$

$$K_\alpha(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{E}_{g_{ij}}\left[\left(\nabla_{\boldsymbol{\theta}} V_{ij,\alpha}^{(r)}(\boldsymbol{\theta})\right)\left(\nabla_{\boldsymbol{\theta}} V_{ij,\alpha}^{(r)}(\boldsymbol{\theta})\right)^\intercal\right],$$

where $V_{ij,\alpha}(\cdot)$ is the V-function defined in (2.5), g_{ij} is the true density of the entries X_{ij} of the matrix, and $\nabla_{\boldsymbol{\theta}}$ and $\nabla^2_{\boldsymbol{\theta}}$ respectively denote the gradient and the Hessian operator with respect to $\boldsymbol{\theta}$. Since g_{ij}s are unknown, one computes the above expectations assuming that the densities belong to the model families of densities, and then replaces $\boldsymbol{\theta}$ by its consistent estimate $\widehat{\boldsymbol{\theta}}^{(r)}_{\alpha}$ to compute the BHHJ-C. While both DIC and BHHJ-C are modelled after AIC, Kurata and Hamada (2020) proposed a simpler criterion based on BIC as follows

$$\mathrm{RCC}_{\alpha} = H^{(r)}_{\alpha}(\boldsymbol{\theta}; \boldsymbol{X}) + \frac{p}{2n}\log(n).$$

The RCC is demonstrated to achieve selection consistency and robustness. Unfortunately, this derivation only works when the number of rows n asymptotically grows to infinity but the number of columns p remains fixed. However, the analysis of SVD considers the asymptotic regime where both n and p grow to infinity maintaining a constant ratio in limit, as described in Chapter 2. Recently, Kurata (2022) showed that many different combinations of the divergence based metric and the penalty function $p\log(n)/2n$ also achieve selection consistency and robustness. On the other hand, Toma et al. (2020) generalized the DIC by replacing the density power divergence with the logarithmic density power divergence (LDPD) measure, while appropriately adjusting the penalty term as necessary. A summary of different penalized criteria is indicated in Table 4.1.

Table 4.1: Different Penalized Criteria for Matrix Rank Estimation. (Here $\widehat{\sigma}_{\alpha}$ denotes the estimated noise variance by minimizing corresponding $H_{\alpha}(\cdot)$).

Criterion	**Definition**
AIC (Akaike, 1973)	$2H^{(r),\phi}_0(\widehat{\boldsymbol{\theta}}^{(r),\phi}_0)\widehat{\sigma}^2_0 + r\widehat{\sigma}^2_0(n+p-r)/np$
BIC (Schwarz, 1978)	$2H^{(r),\phi}_0(\widehat{\boldsymbol{\theta}}^{(r),\phi}_0)\widehat{\sigma}^2_0 + r\widehat{\sigma}^2_0(n+p-r)\log(np)/np$
PC1 (Bai and Ng, 2002)	$2H^{(r),\phi}_0(\widehat{\boldsymbol{\theta}}^{(r),\phi}_0)\widehat{\sigma}^2_0 + r\widehat{\sigma}^2_0\log(\frac{n+p}{np})/np$
PC2 (Bai and Ng, 2002)	$2H^{(r),\phi}_0(\widehat{\boldsymbol{\theta}}^{(r),\phi}_0)\widehat{\sigma}^2_0 + r\widehat{\sigma}^2_0\log(n\wedge p)/np$
PC3 (Bai and Ng, 2002)	$2H^{(r),\phi}_0(\widehat{\boldsymbol{\theta}}^{(r),\phi}_0)\widehat{\sigma}^2_0 + r\widehat{\sigma}^2_0\log(n\wedge p)/(n\wedge p)$
IC1 (Bai and Ng, 2002)	$2H^{(r),\phi}_0(\widehat{\boldsymbol{\theta}}^{(r),\phi}_0) + r\log(\frac{n+p}{np})/np$
IC2 (Bai and Ng, 2002)	$2H^{(r),\phi}_0(\widehat{\boldsymbol{\theta}}^{(r),\phi}_0) + r\log(n\wedge p)/np$
IC3 (Bai and Ng, 2002)	$2H^{(r),\phi}_0(\widehat{\boldsymbol{\theta}}^{(r),\phi}_0) + r\log(n\wedge p)/(n\wedge p)$
DIC (Karagrigoriou and Papaioannou, 2008)	$H^{(r)}_{\alpha}(\widehat{\boldsymbol{\theta}}^{(r)}_{\alpha}) + r(\alpha+1)(2\pi)^{-\alpha/2}\left(\frac{1+\alpha}{1+2\alpha}\right)^{1+r/2}$
RCC (Kurata and Hamada, 2020)	$H^{(r)}_{\alpha}(\widehat{\boldsymbol{\theta}}^{(r)}_{\alpha}) + r\log(n)/2n$
DICMR (ours)	$H^{(r)}_{\alpha}(\widehat{\boldsymbol{\theta}}^{(r)}_{\alpha}) + \frac{r(n+p)}{2np}(2\pi)^{-\alpha/2}\widehat{\sigma}^{-\alpha}_{\alpha}\left(\frac{1+\alpha}{1+2\alpha}\right)^{3/2}$

4.2.2 Cross-validation Approaches

Another approach to the estimation of the rank of a matrix uses a resampling technique known as cross-validation. In this case, the rows and the columns of the data matrix $\boldsymbol{X}$ are subdivided into groups; these groups are deleted in respective turns. The remaining entries are used to compute a partial SVD estimate up to a chosen rank, and it is used to

predict the deleted entries. The choice of the rank that yields the lowest possible prediction error is declared to be the estimate of the rank. In mathematical terms, let $R_1, R_2, \ldots, R_B$ be B randomly chosen subsets of the rows having indices from $\{1, 2, \ldots, n\}$ and similarly, let $C_1, C_2, \ldots, C_B$ be such subsets of the columns having indices from $\{1, 2, \ldots, p\}$. The choice of B may depend on the dimensions of the matrix. Typically, $B = \binom{n}{n_1}\binom{p}{p_1}$ for some $n_1 \leq n$ and $p_1 \leq p$. For any choice of $b = 1, 2, \ldots, B$, the data matrix $\boldsymbol{X}$ is then partitioned into

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{R_b, C_b} & \boldsymbol{X}_{R_b, C_b^c} \\ \boldsymbol{X}_{R_b^c, C_b} & \boldsymbol{X}_{R_b^c, C_b^c} \end{bmatrix}.$$

The tuple (R_b, C_b) is often called the holdout set following the terminology in the cross-validation literature. Based on this partitioned matrix, one then finds an estimate of $\boldsymbol{X}_{R_b, C_b}$ using partial SVD up to rank r of the other entries,

$$\widehat{\boldsymbol{X}}^{(r)}_{R_b, C_b} = T_r\left(\boldsymbol{X}_{R_b, C_b^c}, \boldsymbol{X}_{R_b^c, C_b}, \boldsymbol{X}_{R_b^c, C_b^c}\right), \tag{4.7}$$

where T_r denotes the prediction function based on the partial SVD estimate. Finally, we combine the error in estimation to create a cross-validation criterion

$$\mathrm{CV}(r) := S\left(\left\{\left\|\boldsymbol{X}_{R_b, C_b} - \widehat{\boldsymbol{X}}^{(r)}_{R_b, C_b}\right\|\right\}_{b=1}^{B}\right), \; r = 1, 2, \ldots, (n \wedge p), \tag{4.8}$$

where $S(\cdot)$ is a suitable measure of the scale of univariate samples. If the rank r is the true rank, then the distribution of these errors should be homogeneous across all such random choice of partitions, hence yielding a small value of the cross-validation metric given in (4.8). Based on this idea, the estimated rank is obtained as $\widehat{r} = \underset{0 \leq r \leq (n \wedge p)}{\arg\min}\ \mathrm{CV}(r)$.

Wold (1978) considers "speckled"-type holdout sets where the sets R_b and C_b are singleton sets covering each row and column indices, resulting in $B = np$ holdout sets. The prediction function T_r based on the partial SVD estimates is calculated in an iterative fashion based on Expectation-Maximization (EM) algorithm. For a fixed choice of holdout set combination (R_b, C_b), it starts with an imputation of the entry X_{R_b, C_b} as the mean of remaining entries and performs a partial r-SVD of the entire imputed matrix. The entry X_{R_b, C_b} is then imputed by its revised estimate from this partial r-SVD and then a further partial SVD is performed on this new imputed matrix. These two steps are repeated until for two successive runs the imputed value of X_{R_b, C_b} remains stable. Finally, the scale measure for calculating the cross-validation error is simply given by the root mean squared error (RMSE).

On the other hand, Gabriel (2002) consider a "block"-type holdout set where the sets R_b and C_b are all possible fixed-size sets from all possible combinations of row and column indices. His proposed estimate T_r is given by

$$T_r\left(\boldsymbol{X}_{R_b, C_b^c}, \boldsymbol{X}_{R_b^c, C_b}, \boldsymbol{X}_{R_b^c, C_b^c}\right) = \boldsymbol{X}_{R_b, C_b^c}\left(\boldsymbol{X}_{R_b^c, C_b^c}\right)^{(r),+}\boldsymbol{X}_{R_b^c, C_b},$$

where $\left(\boldsymbol{X}_{R_b^c, C_b^c}\right)^{(r),+}$ is a generalized inverse of the partial rank r-SVD of $\boldsymbol{X}_{R_b^c, C_b^c}$, given by

$$\left(\boldsymbol{X}_{R_b^c, C_b^c}\right)^{(r),+} = \sum_{k=1}^{r} \frac{1}{\widehat{\lambda}_{k,0}} \widehat{\boldsymbol{u}}_{k,0} \widehat{\boldsymbol{v}}_{k,0}^{\intercal},$$

where $\widehat{\lambda}_{k,0}, \widehat{\boldsymbol{u}}_{k,0}, \widehat{\boldsymbol{v}}_{k,0}$ are classical SVD estimates of singular values and vectors based on the partitioned matrix $\boldsymbol{X}_{R_b^c, C_b^c}$ only. While the initial proposal by Gabriel (2002) considers only singleton holdout sets, Owen and Perry (2009) demonstrated that generalizing the holdout sets to $n_r \times n_c$-size partitions (where $n_r, n_c > 1$) leads to a more reliable estimate of the rank for large data matrices. Their recommendation was to use $n_r = [n/2]$ and $n_c = [p/2]$ to have the best bias-variance trade-off. However, this increases the number of holdout sets from np to $O(n^{n_r} p^{n_c})$, significantly increasing the time complexity of the entire cross-validation procedure.

In another direction, Eastment and Krzanowski (1982) consider the holdout for rows and columns separately. For a fixed b, they choose $R_b = C_{b+1} = \phi$, the null set and $R_{b+1} = \{i_0\}$ and $C_b = \{j_0\}$, where i_0 and j_0 are randomly chosen row and column indices. Subsequently let us denote $\widehat{\lambda}_{k,0}^{(b)}, \widehat{\boldsymbol{u}}_{k,0}^{(b)}, \widehat{\boldsymbol{v}}_{k,0}^{(b)}$ as the SVD estimates of the $n \times (p-1)$ data matrix after deleting j_0-th column and $\widehat{\lambda}_{k,0}^{(b+1)}, \widehat{\boldsymbol{u}}_{k,0}^{(b+1)}, \widehat{\boldsymbol{v}}_{k,0}^{(b+1)}$ as the same for $(n-1) \times p$ data matrix after deleting i_0-th row. They are then combined to produce an estimate of X_{i_0,j_0} as

$$\widehat{X}_{i_0,j_0}^{(r)} = \sum_{k=1}^{r} \left(\widehat{\lambda}_{k,0}^{(b)}\right)^{1/2} \left(\widehat{\lambda}_{k,0}^{(b+1)}\right)^{1/2} \widehat{\boldsymbol{u}}_{k,0}^{(b)} \left(\widehat{\boldsymbol{v}}_{k,0}^{(b+1)}\right)^{\mathsf{T}}. \tag{4.9}$$

Note that, since both of these estimates ignore the entry X_{i_0,j_0} either by removing row i_0 or column j_0, Eq. (4.9) yields a proper cross-validation estimate of X_{i_0,j_0} based on partial SVD up to rank r. However, Bro et al. (2008) performed extensive simulation studies and found that it is useful to add a scaling factor to the singular value estimates for small matrices, such as

$$\widehat{X}_{i_0,j_0}^{(r)} = \sum_{k=1}^{r} \left(\widehat{\lambda}_{k,0}^{(b)} \sqrt{p/(p-1)}\right)^{1/2} \left(\widehat{\lambda}_{k,0}^{(b+1)} \sqrt{n/(n-1)}\right)^{1/2} \widehat{\boldsymbol{u}}_{k,0}^{(b)} \left(\widehat{\boldsymbol{v}}_{k,0}^{(b+1)}\right)^{\mathsf{T}}. \tag{4.10}$$

Asymptotically, this scaling factor goes to 1, hence both variants of cross-validation algorithms of Eastment and Krzanowski (1982) yield similar performances for large matrices.

All three methods described above, produce poor estimates for data matrix with outliers, due to the usage of classical SVD and the RMSE as the scale measure to aggregate the prediction error. This is illustrated through the simulation studies in Section 4.5.2. For this reason, we also consider different variants of each cross-validation method where we replace the RMSE and the Frobenius norm of the cross-validation errors from different folds with more robust choices, i.e., Mean Absolute Error (MAE) or Median Absolute Deviation (MAD). To introduce more robustness, we also consider replacing the classical SVD with the robust "rSVDdpd" estimator in these estimators. However, it is computationally challenging to use a partial robust SVD estimate in place of the partial classical SVD estimate for each combination of holdout sets. For instance, the Gabriel style cross-validation of a 50×40-dimensional matrix requires computing 2000 SVDs. Even if "rSVDdpd" is quite fast compared to its peers as shown in Chapter 2, completing the robust SVD of a 50×40 matrix in a standard consumer-grade computer takes about a second, resulting in the rank estimation procedure take approximately 33 minutes in its entirety. This severely

limits the practical usage of such a procedure. Arciniegas-Alarcón et al. (2022) considered a workaround where a robust SVD up to rank $(n \wedge p)$ is used on the data matrix $\boldsymbol{X}$ first to estimate a full-rank proxy component $\boldsymbol{X}'$ free of outlying observations. Then, a cross-validation procedure based on the traditional SVD is applied on the proxy $\boldsymbol{X}'$, which should be able to recover its true rank.

4.3 Divergence Information Criterion for Matrix Rank Estimation

In the previous section, we described the two major approaches for rank estimation in detail. In this section, we introduce a new rank estimation method that is quite fast and computationally simple unlike the cross-validation methods, yet achieves high accuracy in robust rank estimation compared to the existing penalized approaches. This new criterion, called Divergence Information Criterion for Matrix Rank (DICMR) is a penalized criterion that uses the robust rSVDdpd estimate along with some adjustments of the usual DIC for the matrix rank estimation problem.

We begin with a derivation similar to the derivation of DIC as in Karagrigoriou and Papaioannou (2008). The objective is to find a robust unbiased estimator of $\mathbb{E}(Q_\alpha^{(r)}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_\alpha^{(r)})$, where

$$Q_\alpha^{(r)}(\boldsymbol{\theta}) := \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \int V_{ij,\alpha}^{(r)}(\boldsymbol{\theta}; x) g_{ij}(x) dx,$$

and $V_{ij,\alpha}^{(r)}(\cdot)$ are as defined in (2.5) and g_{ij}s are the true densities of the entries X_{ij} of the data matrix. Let us rewrite $\widehat{\boldsymbol{\theta}}_\alpha^{(r)} = \left(\widehat{\boldsymbol{D}}_\alpha^{(r)}, \widehat{\boldsymbol{U}}_\alpha^{(r)}, \widehat{\boldsymbol{V}}_\alpha^{(r)}\right)$, the components describing the largest r singular values, corresponding left singular vectors and right singular vectors. To estimate $\mathbb{E}(Q_\alpha^{(r)}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_\alpha^{(r)})$, we shall use the BHHJ-C presented in (4.6). Also note that, as presented before in Chapter 2, conditional on the event $\boldsymbol{V}_\alpha^{(r)} = \widehat{\boldsymbol{V}}_\alpha^{(r)}$, the estimation of the other parameters of $\widehat{\boldsymbol{\theta}}_\alpha^{(r)}$ is simply solving a linear regression problem. Applying a derivation similar to the normal regression setup as done in Ghosh and Basu (2013), we can calculate the $J_\alpha(\boldsymbol{\theta})$ and $K_\alpha(\boldsymbol{\theta})$ matrices for a general model family f, and substitute them in the BHHJ-C as in (4.6). The resulting penalty term of the BHHJ-C conditioned on $\boldsymbol{V}_\alpha^{(r)} = \widehat{\boldsymbol{V}}_\alpha^{(r)}$ turns out to be

$$\text{Trace}\left(J_\alpha^{-1}(\widehat{\boldsymbol{\theta}}_\alpha^{(r)}) K_\alpha(\widehat{\boldsymbol{\theta}}_\alpha^{(r)})\right) = \text{Trace}\left(\frac{\widehat{\sigma}^{-(2\alpha+2)} C_{2\alpha}^f}{\widehat{\sigma}^{-(\alpha+2)} C_\alpha^f} \boldsymbol{I}_r\right) = \text{Trace}\left(\widehat{\sigma}^{-\alpha} C_{2\alpha}^f / C_\alpha^f\right), \quad (4.11)$$

where $\widehat{\sigma}^2$ is the estimated noise variance and

$$C_\alpha^f = \int (f'(|x|))^2 f^{\alpha-1}(|x|) dx, \quad (4.12)$$

assuming f' exists. In the special case of the normal model family, we have $f(x) = \phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$ and hence

$$C_\alpha^\phi = (2\pi)^{-\alpha/2} (1+\alpha)^{-3/2}.$$

Note that, here we also use the orthogonality of the estimated right singular vectors $\widehat{\boldsymbol{V}}_\alpha^{(r)}$ to conclude that $(\widehat{\boldsymbol{V}}_\alpha^{(r)})^\intercal \widehat{\boldsymbol{V}}_\alpha^{(r)} = \boldsymbol{I}_r$, the identity matrix. Using this approach, we obtain

$$\begin{aligned}
\mathbb{E}\left(Q_\alpha^{(r)}(\widehat{\boldsymbol{\theta}}_\alpha^{(r)})\right) &= \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{g_{ij}}V_{ij}^{(r)}(\widehat{\boldsymbol{\theta}}_\alpha^{(r)}; X)\right] \\
&= \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left[\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{g_{ij}}V_{ij}^{(r)}(\widehat{\boldsymbol{\theta}}_\alpha^{(r)}; X)\right) \mid \boldsymbol{V}_\alpha^{(r)} = \widehat{\boldsymbol{V}}_\alpha^{(r)}\right] \\
&= \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}V_{ij}^{(r)}(\widehat{\boldsymbol{\theta}}_\alpha^{(r)}; X_{ij}) + \frac{r}{n}\widehat{\sigma}^{-\alpha}C_{2\alpha}^f/C_\alpha^f + o(1/n)\right], \quad \text{by (4.11)}, \\
&= \mathbb{E}\left[H_\alpha^{(r)}(\widehat{\boldsymbol{\theta}}_\alpha^{(r)}, \boldsymbol{X}) + \frac{r}{n}\widehat{\sigma}^{-\alpha}C_{2\alpha}^f/C_\alpha^f + o(1/n)\right].
\end{aligned}$$

Here, $o(1/n)$ refers to a term which goes to 0 faster than $1/n$ as $n \to \infty$. We can similarly take conditional expectation conditioned on the left singular vectors to obtain

$$\mathbb{E}\left(Q_\alpha^{(r)}(\widehat{\boldsymbol{\theta}}_\alpha^{(r)})\right) = \mathbb{E}\left[H_\alpha^{(r)}(\widehat{\boldsymbol{\theta}}_\alpha^{(r)}, \boldsymbol{X}) + \frac{r}{p}\widehat{\sigma}^{-\alpha}C_{2\alpha}^f/C_\alpha^f + o(1/p)\right].$$

Now since both n and p tend towards infinity, and if $r = O((n \wedge p))$, we obtain an estimate by averaging the above two unbiased estimators, resulting in a new criterion given as

$$\text{DICMR}_\alpha(r) := H_\alpha^{(r)}(\widehat{\boldsymbol{\theta}}_\alpha^{(r)}, \boldsymbol{X}) + \frac{r(n+p)}{2np}\widehat{\sigma}^{-\alpha}C_{2\alpha}^f/C_\alpha^f. \tag{4.13}$$

We shall refer to this as the "Divergence Information Criterion for Matrix Rank" (DICMR) in the subsequent discussion. In the special case of normally distributed errors, the DICMR reduces to

$$\text{DICMR}_\alpha^\phi(r) := H_\alpha^{(r),\phi}(\widehat{\boldsymbol{\theta}}_\alpha^{(r)}, \boldsymbol{X}) + \frac{r(n+p)}{2np}(2\pi)^{-\alpha/2}\widehat{\sigma}^{-\alpha}\left(\frac{1+\alpha}{1+2\alpha}\right)^{3/2}. \tag{4.14}$$

As we will demonstrate next in Section 4.4, this criterion achieves selection consistency for all $\alpha \in (0, 1]$ under the same asymptotic regime of SVD, and also maintains robustness. These will also be corroborated through numerical simulations in Section 4.5.

4.4 Theoretical Studies

4.4.1 Selection Consistency

The selection consistency of a model selection criterion refers to the property that if two competing models are present; one of which is nested under the other, the criterion selects the smaller model if it is adequate (i.e., contains the true distribution) and chooses the larger model if the smaller model is not adequate, asymptotically with probability tending to 1. In the case of the rank estimation problem, the selection consistency for the DICMR implies that if the low rank component $\boldsymbol{L}$ of the data matrix $\boldsymbol{X}$ as in (4.1) is of rank r, then the corresponding criterion should satisfy

$$\mathbb{P}\left(\text{DICMR}_\alpha(r) \leq \min\{\text{DICMR}_\alpha(r-1), \text{DICMR}_\alpha(r+1)\}\right) \to 1,$$

for any $r \geq 1$ as both n and p tend to infinity, subject to the restriction that $\lim_{n,p\to\infty}(n/p) = c$ for some $c \in (0, \infty)$.

Since the rSVDdpd estimator introduced earlier in Chapter 2 proceeds by sequentially estimating the singular values and vectors one by one, we restrict our attention to two competing models of rank zero and rank one for its simplicity. The same results will continue to hold for other choices of rank $r \geq 1$, provided that the assumptions are modified appropriately. We will begin by considering the situation when the low rank component $\boldsymbol{L}$ is of rank zero. This means the data matrix $\boldsymbol{X}$ consists of only pure noise such that they are independently (but may not be identically) distributed with zero mean and variance $\sigma_0^2 = 1/\sqrt{np}$. Then, the following theorem establishes that under some suitable assumptions, the DICMR will choose the correct rank equal to zero between a choice of rank zero and rank one.

Theorem 4.1. *Let us assume that the Assumptions (B1)-(B6) of Chapter 2 hold with the density g replaced by a bounded model density f. Also assume that in Assumption (B1), $\lambda^g = 0$, i.e., the low rank component of the data matrix $\boldsymbol{X}$ is the $\mathbf{0}$ matrix. Then, for any $\alpha > 0$ and any $\epsilon > 0$, we have*

$$\lim_{\substack{n\to\infty,p\to\infty \\ (n/p)\to c\in(0,\infty)}} \mathbb{P}\left(\mathrm{DICMR}_\alpha(0) < \mathrm{DICMR}_\alpha(1)\right) \geq \frac{1}{2} + \max\left\{0, \frac{1}{2} - \frac{1}{2(1-\epsilon)t_\alpha^2}\right\}$$

where

$$t_\alpha = \frac{(1+c)}{2\sqrt{c}} \frac{C_{2\alpha}^f}{C_\alpha^f \|f\|_{1+\alpha}^{1+\alpha}} \frac{B_\alpha^f}{\sqrt{B_{2\alpha}^f - (A_\alpha^f)^2}}, \tag{4.15}$$

and

$$A_\alpha^f = \int \left[f(|x|) + |x|\, f'(|x|)\right] f^\alpha(|x|)dx, \tag{4.16}$$

$$B_\alpha^f = \int f^{1+\alpha}(|x|)dx - 2\int |x|\, f'(|x|) f^\alpha(|x|)dx + \int x^2 \left(f'(|x|)\right)^2 f^{\alpha-1}(|x|)dx. \tag{4.17}$$

Proof. Let us denote the penalty function of the DICMR as in (4.13) as $r\mathcal{P}_\alpha$. Now, consider the difference in the DICMR between rank zero and rank one,

$$\begin{aligned}
& \mathrm{DICMR}_\alpha(0) - \mathrm{DICMR}_\alpha(1) \\
= \;& H_\alpha^{(0)}(\widehat{\boldsymbol{\theta}}_\alpha^{(0)}; \boldsymbol{X}) - H_\alpha^{(1)}(\widehat{\boldsymbol{\theta}}_\alpha^{(1)}; \boldsymbol{X}) - \mathcal{P}_\alpha \\
= \;& \left(\widehat{\sigma}_0^{-\alpha} - \widehat{\sigma}_1^{-\alpha}\right) \|f\|_{1+\alpha}^{1+\alpha} \\
& + \left(1 + \frac{1}{\alpha}\right) \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left[\widehat{\sigma}_1^{-\alpha} f^\alpha\left(\frac{|X_{ij} - \widehat{\lambda}_1 \widehat{u}_{1i} \widehat{v}_{1j}|}{\widehat{\sigma}_1}\right) - \widehat{\sigma}_0^{-\alpha} f^\alpha\left(\frac{|X_{ij}|}{\widehat{\sigma}_0}\right)\right] - \mathcal{P}_\alpha
\end{aligned} \tag{4.18}$$

where $\widehat{\sigma}_0^2$ and $\widehat{\sigma}_1^2$ are the estimated noise variance for rank zero and rank one robust SVD estimates. Similarly, let σ_0^2 and σ_1^2 be the corresponding population variance based on rank zero and rank one partial SVD approximation. Now, due to the Law of Large Numbers, it follows that as n and p both tend to infinity, we have

$$\frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \widehat{\sigma}_0^{-\alpha} f^\alpha\left(\frac{|X_{ij}|}{\widehat{\sigma}_0}\right) = \widehat{\sigma}_0^{-\alpha} \|f\|_{1+\alpha}^{1+\alpha} + \mathcal{O}(1/\sqrt{np}).$$

Because of Assumptions (B1)-(B6) and an application of Theorem 2.7 about the consistency of the "rSVDdpd" estimator, it follows that we have $\left\|\widehat{\sigma}_0^2 - \sigma_0^2\right\|_2 \xrightarrow{P} 0$. Similarly, it also follows from a generalized version of Khinchin's law of large numbers that as both n and p tend to infinity, we have

$$\begin{aligned}
&\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\widehat{\sigma}_1^{-\alpha} f^{\alpha}\left(\frac{|X_{ij}-\widehat{\lambda}_1\widehat{u}_{1i}\widehat{v}_{1j}|}{\widehat{\sigma}_1}\right)\\
=\;& \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{E}\left[\int \widehat{\sigma}_1^{-\alpha}\sigma_0^{-1} f^{\alpha}\left(\frac{|x-\widehat{\lambda}_1\widehat{u}_{1i}\widehat{v}_{1j}|}{\widehat{\sigma}_1}\right) f\left(\frac{|x|}{\sigma_0}\right) dx\right] + \mathcal{O}(1/\sqrt{np})\\
=\;& \mathbb{E}\left[\int \widehat{\sigma}_1^{-\alpha}\sigma_0^{-1} f^{\alpha}\left(\frac{|x-\widehat{\lambda}_1\widehat{u}_{11}\widehat{v}_{11}|}{\widehat{\sigma}_1}\right) f\left(\frac{|x|}{\sigma_0}\right) dx\right] + \mathcal{O}(1/\sqrt{np})\\
&\qquad \text{since, the rows and the columns of } \boldsymbol{X} \text{ matrix are exchangeable}\\
=\;& \mathbb{E}\left[\widehat{\sigma}_1^{-\alpha}\int f^{\alpha}\left(\left|r_\sigma z - \frac{\widehat{\lambda}_1\widehat{u}_{11}\widehat{v}_{11}}{\widehat{\sigma}_1}\right|\right) f(|z|)dz\right] + \mathcal{O}(1/\sqrt{np}),
\end{aligned}$$

where $r_\sigma := \dfrac{\sigma_0}{\widehat{\sigma}_1}$. Now note that the entries X_{ij} of the data matrix can be represented as $X_{ij} = 0.u_{1i}.v_{1j} + e_{ij}$, where u_{1i}, v_{1j} are any randomly chosen but fixed vector in respective Stiefel manifold and e_{ij}s are independent and normally distributed errors with the same distribution as X_{ij}. Due to this decomposition, from the result on the consistency of the "rSVDdpd" estimator as given in Theorem 2.7, it follows that $\widehat{\lambda}_1 \xrightarrow{P} 0$, as both n and p tend to infinity. More importantly, it also shows that the error estimate $\widehat{\sigma}_1$ is also a consistent estimator of the original noise variance σ_0, as the first singular value and vectors asymptotically have no contribution. As a result, $r_\sigma \xrightarrow{P} 1$ as both n and p tend to infinity. Now, since the model density f is bounded by $f(0)$ because of its decreasing nature, by an application of the Dominated Convergence Theorem (DCT), it follows that

$$\begin{aligned}
&\int f^{\alpha}\left(\left|r_\sigma z - \frac{\widehat{\lambda}_1\widehat{u}_{11}\widehat{v}_{11}}{\widehat{\sigma}_1}\right|\right) f(|z|)dz\\
=\;& \int f^{\alpha}\left(\left|z - \frac{\widehat{\lambda}_1\widehat{u}_{11}\widehat{v}_{11}}{\widehat{\sigma}_1}\right|\right) f(|z|)dz + \mathcal{O}(1/\sqrt{np})\\
=\;& \int f^{\alpha}\left(\left|z - \frac{\widehat{\lambda}_1\widehat{u}_{11}\widehat{v}_{11}}{2\widehat{\sigma}_1}\right|\right) f\left(\left|z + \frac{\widehat{\lambda}_1\widehat{u}_{11}\widehat{v}_{11}}{2\widehat{\sigma}_1}\right|\right) dz + \mathcal{O}(1/\sqrt{np})\\
\leq\;& \int f^{\alpha}(|z|)\, f(|z|)\, dz + \mathcal{O}(1/\sqrt{np})\\
&\qquad \text{since, } f \text{ is decreasing in the magnitude of its arguments}\\
=\;& \|f\|_{1+\alpha}^{1+\alpha} + \mathcal{O}(1/\sqrt{np}).
\end{aligned}$$

Putting everything back in Eq. (4.18) now yields that

$$\begin{aligned}
&\sigma_0^{\alpha}\left(\mathrm{DICMR}_\alpha(0) - \mathrm{DICMR}_\alpha(1)\right)\\
=\;& \|f\|_{1+\alpha}^{1+\alpha}\left[-\frac{1}{\alpha} - r_\sigma^{\alpha} + \left(1+\frac{1}{\alpha}\right) r_\sigma^{\alpha}\right] - \frac{(n+p)}{2np} r_\sigma^{\alpha}\frac{C_{2\alpha}^{f}}{C_{\alpha}^{f}} + \mathcal{O}\left((np)^{-(1+\alpha)/2}\right).
\end{aligned}$$

This can further be rearranged as

$$(r_\sigma^\alpha - 1)\left[\frac{\|f\|_{1+\alpha}^{1+\alpha}}{\alpha} - \frac{(n+p)}{2np}\frac{C_{2\alpha}^f}{C_\alpha^f}\right] - \frac{(n+p)}{2np}\frac{C_{2\alpha}^f}{C_\alpha^f} + \mathcal{O}\left((np)^{-(1+\alpha)/2}\right).$$

Since $r_\sigma \xrightarrow{P} 1$, it remains bounded away from 0 for sufficiently large n and p. Therefore, the first term in the above sum converges in probability to 0. The second term also converges to 0, but from the negative side, at the rate of $(np)^{-1/2}$, as $\lim_{n,p\to\infty}(n/p) = c \in (0,\infty)$. Finally, the remainder term converges to 0 at a faster rate $(np)^{-(1+\alpha)/2}$ for all $\alpha > 0$. Therefore, the entire difference becomes negative for all sufficiently large n and p if and only if the rate at which the first term goes to 0 is smaller than the rate of the second term going to 0.

To obtain the rate of convergence, we apply the linear regression analysis present in Ghosh and Basu (2013) to obtain the asymptotic distribution of r_σ^α. By direct calculation, one may obtain that

$$J_\alpha\left(\widehat{\sigma}_1^2\right) = \frac{1}{4}\sigma_0^{-(\alpha+4)}B_\alpha^f,$$
$$K_\alpha\left(\widehat{\sigma}_1^2\right) = \frac{1}{4}\sigma_0^{-(2\alpha+4)}B_{2\alpha}^f - \frac{1}{4}\sigma_0^{-2(2+\alpha)}(A_\alpha^f)^2,$$

where we use the fact that since the low-rank component $\boldsymbol{L} = \boldsymbol{0}$, hence $\mathbb{E}(\sigma_1^2) \to \sigma_0^2$ as $n, p \to \infty$. The quantities A_α^f and B_α^f are as given in (4.16)-(4.17). As a result, conditional of the estimates of the singular values and vectors, the estimated noise variance $\widehat{\sigma}_1^2$ has an asymptotic normal distribution as

$$\sqrt{np}\left(\widehat{\sigma}_1^2 - \sigma_0^2\right) \xrightarrow{d} \mathcal{N}\left(0, 4\sigma_0^4\frac{B_{2\alpha}^f - (A_\alpha^f)^2}{(B_\alpha^f)^2}\right).$$

An application of delta method now yields the asymptotic distribution of r_σ^α as

$$\sqrt{np}(r_\sigma^\alpha - 1) \xrightarrow{d} \mathcal{N}\left(0, \alpha^2\frac{B_{2\alpha}^f - (A_\alpha^f)^2}{(B_\alpha^f)^2}\right),$$

as both $n, p \to \infty$. Now fix any $t > 0$. By an application of Berry-Esseen inequality, we obtain that with probability at most $1/2t^2$,

$$(r_\sigma^\alpha - 1) > \frac{\alpha t}{\sqrt{np}}\frac{\sqrt{B_{2\alpha}^f - (A_\alpha^f)^2}}{B_\alpha^f}.$$

Also, if $(r_\sigma^\alpha - 1) \leq 0$, we have $\mathrm{DICMR}_\alpha(0) < \mathrm{DICMR}_\alpha(1)$, trivially. Therefore, for any $\alpha > 0$ and for sufficiently large n and p, we have

$$\mathrm{DICMR}_\alpha(0) - \mathrm{DICMR}_\alpha(1) < 0,$$

with probability at least $1/2 + \max\left\{0, 1/2 - 1/2t^2\right\}$ for any t such that

$$\frac{\|f\|_{1+\alpha}^{1+\alpha}}{\alpha}\frac{\alpha t}{\sqrt{np}}\frac{\sqrt{B_{2\alpha}^f - (A_\alpha^f)^2}}{B_\alpha^f} < \frac{(n+p)}{2np}\frac{C_{2\alpha}^f}{C_\alpha^f} \implies t < \frac{(1+c)}{2\sqrt{c}}\frac{C_{2\alpha}^f}{C_\alpha^f\|f\|_{1+\alpha}^{1+\alpha}}\frac{B_\alpha^f}{\sqrt{B_{2\alpha}^f - (A_\alpha^f)^2}}.$$

Choosing $t = (1-\epsilon)t_\alpha$ as in (4.15) implies that the DICMR is minimized for rank zero with the required probability. This completes the proof. □

Since the standard normal density $\phi(x) = (2\pi)^{-1/2}e^{-x^2/2} \leq (2\pi)^{-1/2}$ for any $x \in \mathbb{R}$, it is uniformly bounded. It is also easy to see that Assumptions (B2) and (B3) both hold trivially in this case. As a result, under Assumptions (B1) and (B5)-(B6), Theorem 4.1 follows for the special case of normally distributed errors. In this case, the quantity t_α reduces to

$$t_\alpha = \frac{(1+c)}{2\sqrt{c}} \frac{(2+\alpha^2)}{(1+\alpha)^{7/2}} \left[\frac{(2+4\alpha^2)}{(1+2\alpha)^{5/2}} - \frac{\alpha^2}{(1+\alpha)^3} \right]^{-1/2}.$$

When $n = p$, the quantity t_α is very small and hence Theorem 4.1 tells that DICMR is not reliable in picking the correct model. However, when $c = \lim_{n,p}(n/p)$ is away from 1, i.e., rectangular data matrices are considered, the probability of selecting correct model by DICMR is much higher.

Now let us consider the other case where $\boldsymbol{L}$ is a matrix of rank one, i.e., the entries $X_{ij} = \lambda_1 u_{1i} v_{1j} + \sigma_1 e_{ij}$ for all $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$ where e_{ij}s are independent and identically distributed according to the model density f and the variance $\sigma_1^2 = 1/\sqrt{np}$. In this case, a selection consistent criterion would select the larger model of rank one since the smaller model with rank zero is not adequate. The following theorem establishes this fact for the DICMR.

Theorem 4.2. *Let us assume that the Assumptions (B1)-(B6) hold with the density g replaced by a bounded model density f. Also assume that the estimated noise variance $\widehat{\sigma}_0^2$ under rank zero assumption satisfies $\sigma_1/\widehat{\sigma}_0 \xrightarrow{P} 0$ asymptotically as both n and p tend to infinity, where σ_1^2 is the true noise variance after the first rank component is discarded. Then, for any $\alpha > 0$, we have*

$$\lim_{\substack{n\to\infty, p\to\infty \\ (n/p)\to c\in(0,\infty)}} P\left(\mathrm{DICMR}_\alpha(0) > \mathrm{DICMR}_\alpha(1)\right) = 1.$$

Before proceeding with the proof, we note that the assumption on the ratio of the noise variance to be asymptotically negligible stems from the assumption that the low rank component $\boldsymbol{L}$ of the data matrix $\boldsymbol{X}$ is of rank one. However, as explained earlier in Section 4.2, it is impossible to determine this rank unless the singular values are significantly larger than a threshold, depending on n, p and the noise variance σ_1^2. Thus, we require the estimated noise variance $\widehat{\sigma}_0^2$ under rank zero assumption to be significantly larger than the true noise variance σ_1^2, since the former will combine effects from both the noise variance and the first singular value. Also, the use of robust rSVDdpd estimate $\widehat{\sigma}_0^2$ here avoids the non-identifiability issue pointed out by Candès et al. (2011). For instance, the matrix $\boldsymbol{L} = \boldsymbol{e}_1\boldsymbol{e}_1^\intercal$, where $\boldsymbol{e}_1$ is the first vector of the usual set of Euclidean basis vectors, is both rank 1 and sparse, and hence, can be regarded as either a part of the low-rank component or the noise component. Using the robust rSVDdpd estimate $\widehat{\sigma}_0^2$ would ignore the effect of these entries as outlying observations and classify the above as part of the noise component.

Proof. We begin with a decomposition of the difference between DICMR evaluated at rank zero and rank one as shown in (4.18). Since the "rSVDdpd" estimate is asymptotically

consistent under the Assumptions (B1)-(B6) as established by Theorem 2.7, it follows that the estimate of singular value $\widehat{\lambda}_1 \xrightarrow{P} \lambda_1$ as both n and p tend to infinity. Similar conclusions hold for the components of the singular vectors as well. Therefore, by a generalized version of Khinchin's law of large numbers, we have

$$\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\widehat{\sigma}_1^{-\alpha} f^{\alpha}\left(\frac{|X_{ij}-\widehat{\lambda}_1\widehat{u}_{1i}\widehat{v}_{1j}|}{\widehat{\sigma}_1}\right) = \sigma_1^{-\alpha}\left\|f\right\|_{1+\alpha}^{1+\alpha} + \mathcal{O}(1/\sqrt{np}).$$

Similarly, we also have

$$\begin{aligned}
&\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\widehat{\sigma}_0^{-\alpha} f^{\alpha}\left(\frac{|X_{ij}|}{\widehat{\sigma}_0}\right)\\
=\;& \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\int \widehat{\sigma}_0^{-\alpha}\sigma_1^{-1} f^{\alpha}\left(\frac{|x|}{\widehat{\sigma}_0}\right) f\left(\frac{|x-\lambda_1 u_{1i}v_{1j}|}{\sigma_1}\right) dx + \mathcal{O}(1/\sqrt{np})\\
=\;& \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\int \widehat{\sigma}_0^{-\alpha} f^{\alpha}\left(\left|\widetilde{r}_{\sigma} z + \frac{\lambda_1 u_{1i}v_{1j}}{\widehat{\sigma}_0}\right|\right) f\left(|z|\right) dz + \mathcal{O}(1/\sqrt{np})
\end{aligned}$$

where $\widetilde{r}_{\sigma} = \sigma_1/\widehat{\sigma}_0$. Since, f is decreasing in the magnitude of its argument, symmetric about 0 and it is a density function, it follows that

$$\int f^{\alpha}\left(\left|\widetilde{r}_{\sigma} z + \frac{\lambda_1 u_{1i}v_{1j}}{\widehat{\sigma}_0}\right|\right) f\left(|z|\right) dz \leq f^{\alpha}(0)\int f(|z|)dz = f^{\alpha}(0).$$

Now, putting all these back in (4.18), we obtain that

$$\begin{aligned}
&\sigma_1^{\alpha}\left(\mathrm{DICMR}_{\alpha}(0) - \mathrm{DICMR}_{\alpha}(1)\right)\\
\geq\;& \left[\widetilde{r}_{\sigma}^{\alpha}\left\|f\right\|_{1+\alpha}^{1+\alpha} + \frac{\left\|f\right\|_{1+\alpha}^{1+\alpha}}{\alpha} - \left(1+\frac{1}{\alpha}\right)\widetilde{r}_{\sigma}^{\alpha} f^{\alpha}(0)\right] - \frac{(n+p)C_{2\alpha}^{f}}{2npC_{\alpha}^{f}} + O\left((np)^{-(1+\alpha)/2}\right).
\end{aligned}$$

Now, because of the assumption that $\widetilde{r}_{\sigma} \xrightarrow{P} 0$, it follows that the first term of the above lower bound goes to $\left\|f\right\|_{1+\alpha}^{1+\alpha}/\alpha$. The fact that the second term (the penalty function) and the third term (the remainder part) go to 0 as both n and p tend to infinity is obvious. Therefore, for sufficiently large n and p, with probability tending to 1, we must have

$$\sigma_1^{\alpha}\left(\mathrm{DICMR}_{\alpha}(0) - \mathrm{DICMR}_{\alpha}(1)\right) \geq \left\|f\right\|_{1+\alpha}^{1+\alpha}/\alpha > 0,$$

which completes the proof. □

4.4.2 Robustness

A popular measure of robustness for estimators is the influence function that measures the impact on an estimate due to the change in the distribution of the sample observations. However, in the case of a model selection criterion, there is no general consensus about such robustness measures. Let, G_{ij} be the true distributions of the entries X_{ij} of the data matrix, and let K_{ij} be a contaminating distribution for $i = 1, \ldots, n; j = 1, \ldots, p$. Define the ϵ-contaminated distributions as

$$G_{\epsilon,ij} = (1-\epsilon)G_{ij} + \epsilon K_{ij};\;\; i = 1, 2, \ldots, n; j = 1, 2, \ldots, p.$$

Under this setup, Kurata (2022) proposes to measure the robustness of a model selection criterion $T(\cdot)$ as the boundedness of the absolute difference $|T(\{G_{\epsilon,ij}\}_{i,j=1}^{n,p}) - T(\{G_{ij}\}_{i,j=1}^{n,p})|$ for any ϵ in a neighbourhood of 0. Rewriting the model selection criterion $T(\cdot)$ as a function $\tilde{T}(\epsilon)$ of $\epsilon \in [0,1]$, Kurata (2022) demonstrated that such robustness can be established if the first order derivative of $\tilde{T}$ is finite at $\epsilon = 0$, i.e.,

$$|\tilde{T}'(0)| < \infty.$$

A model selection criterion satisfying this property is called a first-order B-robust criterion.

Theorem 4.3. *For the normal model family of densities, the DICMR as given in* (4.14) *as a model selection criterion is first-order B-robust for any* $\alpha > 0$.

Proof. Let us assume that the corresponding densities are given by g_{ij}, k_{ij} and $g_{\epsilon,ij}$. Then, we can rewrite the DICMR measure defined in (4.14) as a functional as

$$T(G) = (2\pi)^{-\alpha/2}\sigma^{-\alpha}(G)\left[\frac{1}{\sqrt{1+\alpha}} - \frac{\alpha+1}{\alpha np}\sum_{i=1}^{n}\sum_{j=1}^{p}\int \exp\left(-\frac{\alpha(x-\sum_{k=1}^{r}\lambda_k(G)u_{ki}(G)v_{kj}(G))^2}{2\sigma^2(G)}\right) g_{ij}(x)dx + \frac{(n+p)}{2np}\left(\frac{1+\alpha}{1+2\alpha}\right)^{3/2}\right].$$

In the above, we write the terms $\sigma(G), \lambda_k(G)$, etc. to indicate that these parameters are estimated based on the collection of distributions $G = \left(\{G_{ij}\}_{i,j=1}^{n,p}\right)$. Correspondingly, we write $\tilde{T}(\epsilon)$ as a weighted sum of two terms $\tilde{T}_1(\epsilon)$ and $\tilde{T}_2(\epsilon)$ with a suitable choice of weights, where

$$\tilde{T}_1(\epsilon) = \sigma^{-\alpha}(G_\epsilon),$$
$$\tilde{T}_2(\epsilon) = \sigma^{-\alpha}(G_\epsilon)\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\int e^{-\frac{\alpha(x-\sum_{k=1}^{r}\lambda_k(G_\epsilon)u_{ki}(G_\epsilon)v_{kj}(G_\epsilon))^2}{2\sigma^2(G_\epsilon)}}((1-\epsilon)g_{ij}(x)+\epsilon k_{ij}(x))dx.$$

Here, g_{ij} and k_{ij} denote the density functions of G_{ij} and K_{ij} respectively. Starting with the first term, it follows that

$$\tilde{T}_1'(0) = -\alpha(\sigma(G))^{-(\alpha+1)}\partial_\epsilon\sigma(G_\epsilon)\mid_{\epsilon=0},$$

where $\partial_\epsilon\sigma(G_\epsilon)\mid_{\epsilon=0}$ indicates the von-Mises derivative of the functional $\sigma(G_\epsilon)$ with respect to ϵ at 0. One can notice that this is directly linked to the influence curve of the estimate of error variance σ.

Proceeding with the second term, it now follows that

$$\tilde{T}_2'(0) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left[-\alpha\sigma^{-(\alpha+1)}(G)\partial_\epsilon\sigma(G_\epsilon)\mid_{\epsilon=0}\int e^{-\frac{\alpha(x-\mu(G))^2}{2\sigma^2(G)}}g_{ij}(x)dx + \sigma^{-\alpha}(G)\int e^{-\frac{\alpha(x-\mu(G))^2}{2\sigma^2(G)}}(k_{ij}(x)-g_{ij}(x))dx - \alpha\sigma^{-\alpha}(G)\partial_\epsilon\mu(G_\epsilon)\mid_{\epsilon=0}\int\frac{(x-\mu(G))}{\sigma^2(G)}e^{-\frac{\alpha(x-\mu(G))^2}{2\sigma^2(G)}}g_{ij}(x)dx\right], \quad (4.19)$$

where $\mu(G) = \sum_{k=1}^{r} \lambda_k(G) u_{ki}(G) v_{kj}(G)$. For the first summand of (4.19), we have $\int e^{-\frac{\alpha(x-\mu(G))^2}{2\sigma^2(G)}} g_{ij}(x)dx \leq \int g_{ij}(x)dx = 1$, since g_{ij} is a density function. Similarly, for the second summand of (4.19), we have

$$\left| \int e^{-\frac{\alpha(x-\mu(G))^2}{2\sigma^2(G)}} (k_{ij}(x) - g_{ij}(x))dx \right| \leq \int e^{-\frac{\alpha(x-\mu(G))^2}{2\sigma^2(G)}} (k_{ij}(x) + g_{ij}(x))dx \leq 2,$$

by triangle inequality and noting that k_{ij} is also a density function. For the third summand in (4.19), note that we have $|xe^{-ax^2}| \leq (2ae)^{-1/2}$ for any $x \in \mathbb{R}$. Therefore,

$$\left| \int \frac{(x-\mu(G))}{\sigma^2(G)} e^{-\frac{\alpha(x-\mu(G))^2}{2\sigma^2(G)}} g_{ij}(x)dx \right| \leq \sigma^{-1}(G)(\alpha e)^{-1/2}.$$

Putting these back together into (4.19) yields,

$$|\tilde{T}_2'(0)| \leq \alpha(\sigma(G))^{-(\alpha+1)} |\partial_\epsilon \sigma(G_\epsilon) \mid_{\epsilon=0} | + 2\sigma^{-\alpha}(G) + \sqrt{\frac{\alpha}{e}} (\sigma(G))^{-(\alpha+1)} |\partial_\epsilon \mu(G_\epsilon) \mid_{\epsilon=0} |.$$

Now the influence curve for the mean and variance parameters are known to be bounded for $\alpha > 0$, as illustrated in Section 3.4.4 (also see Basu et al. (1998)). As a result, the quantities $|\partial_\epsilon \mu(G_\epsilon) \mid_{\epsilon=0} |$ and $|\partial_\epsilon \sigma(G_\epsilon) \mid_{\epsilon=0} |$ are bounded for $\alpha > 0$. Consequently, the above derivation shows that both $|\tilde{T}_1'(0)|$ and $|\tilde{T}_2'(0)|$ are bounded, hence the DICMR is first-order B-robust. □

4.5 Simulation Studies

4.5.1 Simulation Setups

We shall consider a simulation setup that is similar to the one mentioned by Owen and Perry (2009). In each simulation setup, the data matrix is generated following the LSN decomposition as in (2.1). To generate the low-rank component $\boldsymbol{L}$, we start by simulating an $n \times p$-size matrix $\boldsymbol{A}$ with each entry a_{ij} generated as i.i.d. standard normal random variable. Then, we construct $\boldsymbol{L}$ as

$$\boldsymbol{L} = \boldsymbol{U_A} \boldsymbol{D} \boldsymbol{V_A}^\intercal,$$

where $\boldsymbol{U_A}, \boldsymbol{V_A}$ are the left and right singular vectors for the random matrix $\boldsymbol{A}$. The singular values of $\boldsymbol{L}$, i.e., the diagonal elements of $\boldsymbol{D}$ matrix are chosen in one of the two following ways.

1. **Equal Singular Values:** First r singular values are taken to be equal to 1 and rest of the singular values are kept 0.

2. **Decreasing Singular Values:** First r singular values are chosen as $2, 1+\frac{r-2}{r-1}, \ldots, 1+\frac{1}{r-1}, 1$ and rest of the diagonal elements are kept equal to 0.

The entries n_{ij} of the noise component $\boldsymbol{N}$ are independently generated according to a normal random variable with zero mean and variance σ^2. The σ^2 is chosen as σ_e^2/np so

that the noise-to-signal ratio is σ_e^2, i.e., $\mathbb{E}(\|\boldsymbol{N}\|_F^2) = \sigma_e^2 \|\boldsymbol{L}\|_F^2$. Finally, the entries s_{ij} of the sparse component $\boldsymbol{S}$ are generated as

$$s_{ij} = 5\delta_{ij}(2\eta_{ij} - 1) \max_{i,j} |L_{ij}|, \ \ i = 1, 2, \ldots, n; j = 1, 2, \ldots, p.$$

where δ_{ij} and η_{ij} are independent Bernoulli random variables with success probabilities δ and $1/2$ respectively.

In the experiments, we consider $n = 50, p = 40$ and $r = 10$. We take different variants of the above setup with multiple combinations of noise-to-signal ratio σ_e^2 as $0.05, 0.5$ and 1, and varying levels of contamination proportion δ ranging from 0 (no contamination) to $0.05, 0.1$ and 0.2 (high level of contamination). These simulation setups are indicated as Sij with i-th level of contamination proportion and j-th level of noise-to-signal ratio. For example, (S12) indicate the simulation with noise-to-signal ratio 0.5 and 5% proportion of contamination.

4.5.2 Simulation Results

In Table 4.2, we summarize the results from all the simulation exercises that we have performed. In particular, we demonstrate the best performing cross-validation method and the best performing penalized approach for each simulation scenario considered above. For brevity of presentation, detailed results on the performances of each method are provided in Tables 4.3-4.10 of Appendix 4.A.

As evident from Table 4.2, the proposed DICMR is often the best performing penalized criterion, except in a few cases when there is no contamination. The Wold style cross-validation (WCV) method along with the robust rSVDdpd algorithm is usually the best for robust rank estimation problems which also take the longest amount of time. For a quicker solution, DICMR provides a reasonably good estimate that is often closely competitive to the WCV method, and is sometimes better than other cross-validation methods such as Gabriel-style cross-validation (GCV) or Bi-cross validation (BCV) methods.

We also note that when the errors are generated from a contaminated distribution, all of the existing methods (both penalized and CV approaches) that use traditional SVD fail to provide a reasonable estimate of the rank of the matrix, even when the contamination proportion is low. Even when the cross-validation methods are modified to use MAE or MAD as a scale measure for aggregating cross-validation errors, they do not produce a robust estimate of rank. This can be understood by the following phenomenon: If there is a single outlier in a sample of n observations, then the usual k-fold cross-validation produces $(k-1)$ subsamples which contain the outlier. Hence, each of those $(k-1)$ subsamples out of total k subsamples will produce an egregiously bad estimate of the parameter of interest, resulting in $100(1-1/k)\%$ outlying predictions. Because of this extremely high proportion of contamination, the final estimate will also be nonrobust, despite using any robust metric to obtain the final cross-validation measure. Readers are referred to Tables 4.3-4.6 for further details.

Table 4.2: Summarized results of the best performing existing rank estimation methods under different simulation scenarios (Column Prop. indicates the proportion of replications where the correct rank is estimated exactly).

δ	σ_e^2 (SNR)	Singular Values	Best CV Method Name	Prop.	RMSE	Best Penalty Method Name	Prop.	RMSE
0	0.05	Equal	Classical SVD + WCV, BCV	1.00	0.00	Classical SVD + PC3, IC3	1.00	0.00
	0.5	Equal	Classical SVD + WCV, BCV	1.00	0.00	rSVDdpd + DICMR	0.53	1.33
	1	Equal	rSVDdpd + WCV	0.85	0.39	rSVDdpd + DICMR	0.39	1.28
0	0.05	Decreasing	Classical SVD + WCV	1.00	0.00	Classical SVD + PC3, IC3	1.00	0.00
	0.5	Decreasing	Classical SVD + ECV	0.93	0.32	rSVDdpd + DICMR	0.40	1.37
	1	Decreasing	rSVDdpd + WCV	0.32	1.56	rSVDdpd + DICMR	0.20	1.51
0.05	0.05	Equal	rSVDdpd + WCV	1.00	0.00	rSVDdpd + IC3	0.92	0.28
	0.5	Equal	rSVDdpd + WCV	1.00	0.00	rSVDdpd + DICMR	0.57	1.04
	1	Equal	rSVDdpd + ECV	0.70	1.99	rSVDdpd + DICMR	0.31	1.24
0.05	0.05	Decreasing	rSVDdpd + WCV, ECV	1.00	0.00	rSVDdpd + DICMR	0.90	0.32
	0.5	Decreasing	rSVDdpd + ECV	0.89	0.41	rSVDdpd + DICMR	0.52	0.87
	1	Decreasing	rSVDdpd + WCV	0.24	1.88	rSVDdpd + DICMR	0.22	1.87
0.1	0.05	Equal	rSVDdpd + WCV, ECV	1.00	0.00	rSVDdpd + DICMR	0.84	0.40
	0.5	Equal	rSVDdpd + WCV	1.00	0.00	rSVDdpd + DICMR	0.63	1.00
	1	Equal	rSVDdpd + WCV	0.56	1.17	rSVDdpd + DICMR	0.40	1.22
0.1	0.05	Decreasing	rSVDdpd + WCV, ECV	1.00	0.00	rSVDdpd + DICMR	0.94	0.24
	0.5	Decreasing	rSVDdpd + ECV	0.79	1.11	rSVDdpd + DICMR	0.52	0.84
	1	Decreasing	rSVDdpd + WCV	0.22	1.82	rSVDdpd + DICMR	0.13	2.14
0.2	0.05	Equal	rSVDdpd + WCV, ECV	1.00	0.00	rSVDdpd + DICMR	0.88	0.35
	0.5	Equal	rSVDdpd + WCV	0.99	0.26	rSVDdpd + DICMR	0.78	0.53
	1	Equal	rSVDdpd + WCV	0.39	1.30	rSVDdpd + DICMR	0.09	2.95
0.2	0.05	Decreasing	rSVDdpd + ECV	1.00	0.0	rSVDdpd + DICMR	0.96	0.20
	0.5	Decreasing	rSVDdpd + WCV	0.33	1.17	rSVDdpd + DICMR	0.52	0.93
	1	Decreasing	rSVDdpd + WCV	0.26	1.57	rSVDdpd + DICMR	0.03	3.06

When the robust rSVDdpd procedure is used instead of the traditional SVD, most of the rank estimation methods show improvements, except for the classical information criteria like AIC and BIC. Interestingly, even the simple elbow method using a robust estimate of singular values performs so well that it becomes competitive with some of the cross-validation methods. While we incorporate the rSVDdpd for most of the rank estimation procedures for comparison purposes, there is no obvious way to incorporate it into the Bayesian procedure of Hoff (2007), so we refrain from considering that method in this scenario. The corresponding results are illustrated in Tables 4.7-4.10.

4.A Appendix: Detailed Simulation Results

In this section, we present the detailed tables containing the simulation results for robust rank estimation using existing model selection criteria and the proposed DICMR.

Table 4.3: Proportion of exact (overestimation in brackets) for different rank-estimation methods when using classical SVD for the simulation setting of 50×40-dimensional matrix with equal singular values.

Method	S01	S02	S03	S11	S12	S13	S21	S22	S23	S31	S32	S33
Elbow	0.62	0.62	0.36	0.02	0.03	0.02	0.05	0.03	0.04	0.01	0.02	0.01
	(0.00)	(0.00)	(0.05)	(0.15)	(0.10)	(0.15)	(0.11)	(0.17)	(0.14)	(0.12)	(0.09)	(0.11)
AIC	0.09	0.00	0.00	0.10	0.04	0.04	0.06	0.05	0.05	0.13	0.12	0.15
	(0.00)	(0.00)	(0.00)	(0.11)	(0.06)	(0.03)	(0.07)	(0.07)	(0.07)	(0.49)	(0.50)	(0.47)
BIC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PC1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PC2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PC3	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
IC1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
IC2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
IC3	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
WCV (MSE)	1.00	1.00	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.10)	(0.32)	(0.31)	(0.28)	(0.11)	(0.09)	(0.10)	(0.04)	(0.05)	(0.06)
WCV (MAE)	1.00	1.00	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.10)	(0.32)	(0.29)	(0.28)	(0.11)	(0.08)	(0.09)	(0.04)	(0.05)	(0.06)
WCV (MAD)	1.00	0.99	0.66	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.00
	(0.00)	(0.01)	(0.21)	(0.38)	(0.34)	(0.31)	(0.20)	(0.20)	(0.20)	(0.08)	(0.10)	(0.08)
ECV (MSE)	1.00	1.00	0.83	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.02)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
ECV (MAE)	1.00	0.99	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
ECV (MAD)	1.00	0.93	0.35	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
GCV (MSE)	0.84	0.62	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.16)	(0.38)	(0.64)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GCV (MAE)	0.85	0.64	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.15)	(0.36)	(0.66)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GCV (MAD)	0.56	0.47	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.44)	(0.53)	(0.68)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BCV (MSE)	0.89	0.53	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.11)	(0.47)	(0.83)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BCV (MAE)	0.89	0.56	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.11)	(0.44)	(0.83)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BCV (MAD)	0.68	0.43	0.18	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.32)	(0.57)	(0.80)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Bayes	1.00	0.28	0.06	0.05	0.03	0.03	0.01	0.01	0.01	0.02	0.01	0.02
	(0.00)	(0.42)	(0.69)	(0.48)	(0.46)	(0.43)	(0.17)	(0.22)	(0.15)	(0.12)	(0.16)	(0.07)

Table 4.4: Bias (RMSE in brackets) for different rank-estimation methods when using classical SVD for the simulation setting of 50×40-dimensional matrix with equal singular values.

Method	S01	S02	S03	S11	S12	S13	S21	S22	S23	S31	S32	S33
Elbow	−0.38	−0.38	−1.07	−4.74	−5.38	−4.92	−5.24	−4.87	−5.12	−5.28	−5.70	−5.56
	(0.62)	(0.62)	(2.21)	(6.28)	(6.65)	(6.57)	(6.94)	(7.14)	(7.24)	(6.70)	(7.04)	(7.01)
AIC	−8.03	−9.00	−9.00	−2.90	−4.08	−4.79	−3.48	−3.53	−3.55	0.58	0.54	0.51
	(8.49)	(9.00)	(9.00)	(4.24)	(4.95)	(5.44)	(4.25)	(4.30)	(4.37)	(2.83)	(2.81)	(2.79)
BIC	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
PC1	−9.00	−9.00	−9.00	−8.37	−8.63	−8.79	−8.74	−8.75	−8.77	−8.10	−8.08	−8.11
	(9.00)	(9.00)	(9.00)	(8.46)	(8.69)	(8.82)	(8.77)	(8.78)	(8.80)	(8.19)	(8.17)	(8.20)
PC2	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
PC3	0.00	4.90	8.61	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
	(0.00)	(4.95)	(8.64)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)
IC1	−9.00	−9.00	−9.00	−8.37	−8.63	−8.79	−8.74	−8.75	−8.77	−8.10	−8.08	−8.11
	(9.00)	(9.00)	(9.00)	(8.46)	(8.69)	(8.82)	(8.77)	(8.78)	(8.80)	(8.19)	(8.17)	(8.20)
IC2	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
IC3	0.00	4.90	8.61	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
	(0.00)	(4.95)	(8.64)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)
WCV (MSE)	0.00	0.00	0.05	−3.05	−3.18	−3.72	−6.94	−7.29	−7.10	−8.25	−8.06	−7.87
	(0.00)	(0.00)	(0.57)	(9.21)	(9.25)	(9.25)	(9.07)	(9.07)	(9.09)	(9.03)	(9.04)	(9.05)
WCV (MAE)	0.00	0.00	0.05	−3.01	−3.53	−3.74	−6.94	−7.48	−7.29	−8.25	−8.06	−7.87
	(0.00)	(0.00)	(0.57)	(9.24)	(9.26)	(9.23)	(9.07)	(9.06)	(9.07)	(9.03)	(9.04)	(9.05)
WCV (MAD)	0.00	0.01	0.10	−2.21	−2.76	−3.32	−5.13	−5.19	−5.28	−7.51	−7.11	−7.50
	(0.00)	(0.10)	(0.76)	(8.85)	(8.98)	(8.97)	(9.03)	(9.08)	(9.03)	(9.04)	(8.93)	(9.01)
ECV (MSE)	0.00	0.00	−0.14	−8.80	−8.90	−8.94	−8.97	−8.91	−8.97	−8.93	−9.00	−9.00
	(0.00)	(0.00)	(0.66)	(8.88)	(8.93)	(8.96)	(8.97)	(8.95)	(8.97)	(8.96)	(9.00)	(9.00)
ECV (MAE)	0.00	−0.01	−6.37	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(0.00)	(0.10)	(7.54)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
ECV (MAD)	0.00	−0.17	−3.91	−8.74	−9.00	−8.95	−9.00	−8.83	−9.00	−9.00	−9.00	−9.00
	(0.00)	(0.96)	(5.58)	(8.88)	(9.00)	(8.96)	(9.00)	(8.99)	(9.00)	(9.00)	(9.00)	(9.00)
GCV (MSE)	0.28	0.61	1.37	−8.79	−8.77	−8.74	−8.77	−8.76	−8.77	−8.71	−8.76	−8.79
	(0.82)	(1.15)	(1.99)	(8.82)	(8.80)	(8.77)	(8.81)	(8.79)	(8.80)	(8.74)	(8.79)	(8.81)
GCV (MAE)	0.24	0.57	1.34	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(0.72)	(1.14)	(1.94)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
GCV (MAD)	0.92	1.35	1.81	−8.97	−8.95	−8.94	−8.96	−8.97	−8.98	−9.00	−9.00	−9.00
	(1.69)	(2.30)	(2.64)	(8.97)	(8.95)	(8.94)	(8.96)	(8.97)	(8.98)	(9.00)	(9.00)	(9.00)
BCV (MSE)	0.14	0.74	1.25	−8.94	−8.92	−8.92	−8.90	−8.91	−8.91	−8.93	−8.91	−8.93
	(0.47)	(1.28)	(1.82)	(8.94)	(8.93)	(8.93)	(8.91)	(8.92)	(8.92)	(8.94)	(8.92)	(8.93)
BCV (MAE)	0.14	0.71	1.26	−8.95	−8.96	−8.94	−8.91	−8.93	−8.92	−8.94	−8.93	−8.93
	(0.47)	(1.27)	(1.84)	(8.95)	(8.96)	(8.94)	(8.92)	(8.94)	(8.93)	(8.94)	(8.93)	(8.93)
BCV (MAD)	0.53	1.04	1.64	−7.99	−7.64	−7.75	−8.00	−7.85	−7.75	−8.04	−8.15	−8.12
	(1.11)	(1.63)	(2.30)	(8.17)	(7.97)	(8.02)	(8.11)	(7.99)	(7.95)	(8.19)	(8.23)	(8.23)
Bayes	0.00	0.36	3.49	4.37	2.75	1.56	−4.58	−3.65	−5.28	−5.55	−5.38	−6.89
	(0.00)	(1.87)	(6.85)	(12.95)	(11.48)	(10.10)	(8.07)	(8.05)	(8.39)	(9.32)	(8.82)	(8.65)

Table 4.5: Proportion of exact (overestimation in brackets) for different rank-estimation methods when using classical SVD for the simulation setting of 50×40-dimensional matrix with decreasing singular values.

Method	S01	S02	S03	S11	S12	S13	S21	S22	S23	S31	S32	S33
Elbow	1.00	0.65	0.13	0.02	0.03	0.01	0.05	0.05	0.04	0.03	0.03	0.02
	(0.00)	(0.01)	(0.02)	(0.16)	(0.14)	(0.17)	(0.15)	(0.18)	(0.17)	(0.14)	(0.13)	(0.13)
AIC	0.00	0.00	0.00	0.03	0.04	0.15	0.03	0.02	0.01	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.95)	(0.89)	(0.72)	(0.96)	(0.97)	(0.98)	(1.00)	(1.00)	(1.00)
BIC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PC1	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.03	0.03	0.06	0.06	0.07
	(0.00)	(0.00)	(0.00)	(0.05)	(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.05)	(0.05)	(0.05)
PC2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PC3	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
IC1	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.03	0.03	0.06	0.06	0.07
	(0.00)	(0.00)	(0.00)	(0.05)	(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.05)	(0.05)	(0.05)
IC2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
IC3	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
WCV (MSE)	1.00	0.80	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.03)	(0.35)	(0.32)	(0.28)	(0.09)	(0.12)	(0.11)	(0.04)	(0.06)	(0.06)
WCV (MAE)	1.00	0.80	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.03)	(0.35)	(0.34)	(0.29)	(0.10)	(0.09)	(0.09)	(0.04)	(0.06)	(0.06)
WCV (MAD)	1.00	0.75	0.30	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00
	(0.00)	(0.00)	(0.08)	(0.42)	(0.36)	(0.34)	(0.23)	(0.21)	(0.19)	(0.08)	(0.09)	(0.09)
ECV (MSE)	1.00	0.93	0.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.04)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
ECV (MAE)	1.00	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
ECV (MAD)	1.00	0.40	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GCV (MSE)	0.86	0.58	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.14)	(0.42)	(0.67)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GCV (MAE)	0.85	0.56	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.15)	(0.44)	(0.67)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GCV (MAD)	0.49	0.36	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.51)	(0.64)	(0.61)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BCV (MSE)	0.89	0.41	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.11)	(0.59)	(0.53)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BCV (MAE)	0.89	0.42	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.11)	(0.58)	(0.52)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BCV (MAD)	0.66	0.32	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.34)	(0.68)	(0.52)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Bayes	0.98	0.21	0.12	0.03	0.03	0.02	0.04	0.00	0.02	0.00	0.01	0.00
	(0.02)	(0.28)	(0.29)	(0.55)	(0.44)	(0.47)	(0.11)	(0.16)	(0.18)	(0.12)	(0.21)	(0.11)

Table 4.6: Bias (RMSE in brackets) for different rank-estimation methods when using classical SVD for the simulation setting of 50×40-dimensional matrix with decreasing singular values.

Method	S01	S02	S03	S11	S12	S13	S21	S22	S23	S31	S32	S33
Elbow	0.00	−1.39	−4.87	−4.82	−5.06	−5.05	−4.74	−4.61	−4.52	−4.95	−4.85	−5.13
	(0.00)	(2.93)	(5.89)	(6.43)	(6.49)	(6.89)	(6.69)	(6.98)	(6.92)	(6.61)	(6.59)	(6.74)
AIC	−2.48	−9.00	−9.00	4.44	3.30	2.55	5.07	5.12	5.01	8.99	8.98	9.04
	(2.53)	(9.00)	(9.00)	(5.29)	(4.42)	(3.96)	(5.71)	(5.74)	(5.63)	(9.08)	(9.08)	(9.13)
BIC	−9.00	−9.00	−9.00	−8.90	−8.94	−8.99	−9.00	−9.00	−9.00	−8.99	−8.99	−8.99
	(9.00)	(9.00)	(9.00)	(8.92)	(8.95)	(8.99)	(9.00)	(9.00)	(9.00)	(8.99)	(8.99)	(8.99)
PC1	−6.60	−9.00	−9.00	−4.76	−5.52	−6.07	−5.27	−5.41	−5.42	−3.11	−3.09	−3.05
	(6.62)	(9.00)	(9.00)	(5.28)	(5.93)	(6.38)	(5.60)	(5.76)	(5.76)	(3.61)	(3.62)	(3.59)
PC2	−9.00	−9.00	−9.00	−8.92	−8.95	−8.99	−9.00	−9.00	−9.00	−8.99	−8.99	−8.99
	(9.00)	(9.00)	(9.00)	(8.93)	(8.95)	(8.99)	(9.00)	(9.00)	(9.00)	(8.99)	(8.99)	(8.99)
PC3	0.00	9.46	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
	(0.00)	(9.48)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)
IC1	−6.60	−9.00	−9.00	−4.76	−5.52	−6.07	−5.27	−5.41	−5.42	−3.11	−3.09	−3.05
	(6.62)	(9.00)	(9.00)	(5.28)	(5.93)	(6.38)	(5.60)	(5.76)	(5.76)	(3.61)	(3.62)	(3.59)
IC2	−9.00	−9.00	−9.00	−8.92	−8.95	−8.99	−9.00	−9.00	−9.00	−8.99	−8.99	−8.99
	(9.00)	(9.00)	(9.00)	(8.93)	(8.95)	(8.99)	(9.00)	(9.00)	(9.00)	(8.99)	(8.99)	(8.99)
IC3	0.00	9.46	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
	(0.00)	(9.48)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)	(10.00)
WCV (MSE)	0.00	−0.20	−1.16	−2.48	−3.02	−3.76	−7.31	−6.74	−6.91	−8.24	−7.87	−7.87
	(0.00)	(0.45)	(1.71)	(9.24)	(9.23)	(9.21)	(9.05)	(9.09)	(9.10)	(9.02)	(9.05)	(9.05)
WCV (MAE)	0.00	−0.20	−1.15	−2.48	−2.65	−3.57	−7.12	−7.29	−7.29	−8.25	−7.87	−7.87
	(0.00)	(0.45)	(1.69)	(9.24)	(9.24)	(9.22)	(9.07)	(9.07)	(9.07)	(9.03)	(9.05)	(9.05)
WCV (MAD)	0.00	−0.29	−0.94	−1.28	−2.40	−2.85	−4.66	−5.08	−5.45	−7.55	−7.30	−7.40
	(0.00)	(0.61)	(1.51)	(9.02)	(8.96)	(8.88)	(9.08)	(9.03)	(9.04)	(8.99)	(8.92)	(8.96)
ECV (MSE)	0.00	−0.08	−1.76	−8.78	−8.94	−8.83	−8.97	−8.97	−8.97	−9.00	−9.00	−9.00
	(0.00)	(0.32)	(2.75)	(8.89)	(8.96)	(8.92)	(8.97)	(8.97)	(8.97)	(9.00)	(9.00)	(9.00)
ECV (MAE)	0.00	−2.18	−8.74	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(0.00)	(3.58)	(8.82)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
ECV (MAD)	0.00	−1.61	−6.49	−8.87	−9.00	−8.93	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(0.00)	(2.53)	(7.22)	(8.96)	(9.00)	(8.96)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
GCV (MSE)	0.23	0.68	1.24	−8.76	−8.74	−8.72	−8.75	−8.72	−8.75	−8.76	−8.78	−8.76
	(0.71)	(1.24)	(1.98)	(8.79)	(8.78)	(8.76)	(8.79)	(8.76)	(8.79)	(8.79)	(8.80)	(8.79)
GCV (MAE)	0.24	0.69	1.27	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(0.72)	(1.23)	(1.98)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
GCV (MAD)	1.10	1.35	1.42	−8.94	−8.95	−8.92	−8.98	−8.96	−8.98	−9.00	−9.00	−8.99
	(1.94)	(2.03)	(2.36)	(8.94)	(8.95)	(8.93)	(8.98)	(8.96)	(8.98)	(9.00)	(9.00)	(8.99)
BCV (MSE)	0.14	1.03	0.64	−8.94	−8.92	−8.92	−8.85	−8.91	−8.90	−8.91	−8.95	−8.95
	(0.47)	(1.53)	(1.56)	(8.94)	(8.93)	(8.93)	(8.87)	(8.92)	(8.91)	(8.92)	(8.95)	(8.95)
BCV (MAE)	0.14	0.96	0.62	−8.94	−8.95	−8.95	−8.86	−8.93	−8.93	−8.94	−8.94	−8.95
	(0.47)	(1.46)	(1.56)	(8.94)	(8.95)	(8.95)	(8.88)	(8.94)	(8.94)	(8.94)	(8.94)	(8.95)
BCV (MAD)	0.55	1.25	0.85	−7.89	−7.57	−7.76	−8.03	−7.76	−7.68	−8.01	−8.15	−8.21
	(1.08)	(1.79)	(2.18)	(8.05)	(7.83)	(8.00)	(8.15)	(7.93)	(7.88)	(8.15)	(8.24)	(8.29)
Bayes	0.02	−0.43	−0.71	4.70	2.86	2.71	−5.09	−4.94	−4.59	−5.96	−4.84	−5.95
	(0.14)	(2.05)	(4.48)	(12.63)	(11.80)	(10.30)	(8.27)	(7.92)	(8.56)	(8.59)	(8.52)	(8.55)

Table 4.7: Proportion of exact (overestimation in brackets) for different rank-estimation methods when using robust SVD for the simulation setting of 50×40-dimensional matrix with equal singular values.

Method	S01	S02	S03	S11	S12	S13	S21	S22	S23	S31	S32	S33
Elbow (rsvd)	0.41	0.54	0.30	0.44	0.61	0.26	0.42	0.55	0.12	0.69	0.50	0.04
	(0.00)	(0.00)	(0.04)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.05)	(0.00)	(0.00)	(0.00)
AIC (rsvd)	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BIC (rsvd)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PC1 (rsvd)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PC2 (rsvd)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PC3 (rsvd)	0.98	0.00	0.00	0.92	0.00	0.00	0.73	0.00	0.00	0.13	0.00	0.00
	(0.02)	(1.00)	(1.00)	(0.08)	(1.00)	(1.00)	(0.27)	(1.00)	(1.00)	(0.87)	(1.00)	(1.00)
IC1 (rsvd)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
IC2 (rsvd)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
IC3 (rsvd)	0.98	0.00	0.00	0.92	0.00	0.00	0.73	0.00	0.00	0.13	0.00	0.00
	(0.02)	(1.00)	(1.00)	(0.08)	(1.00)	(1.00)	(0.27)	(1.00)	(1.00)	(0.87)	(1.00)	(1.00)
DICMR	0.89	0.53	0.39	0.88	0.57	0.31	0.84	0.63	0.40	0.88	0.78	0.09
	(0.11)	(0.47)	(0.45)	(0.12)	(0.43)	(0.40)	(0.16)	(0.37)	(0.17)	(0.12)	(0.11)	(0.03)
WCV (MSE)	1.00	1.00	0.85	1.00	1.00	0.69	1.00	1.00	0.56	1.00	0.93	0.38
	(0.00)	(0.00)	(0.09)	(0.00)	(0.00)	(0.19)	(0.00)	(0.00)	(0.24)	(0.00)	(0.00)	(0.31)
WCV (MAE)	1.00	1.00	0.85	1.00	1.00	0.68	1.00	1.00	0.57	1.00	0.93	0.39
	(0.00)	(0.00)	(0.09)	(0.00)	(0.00)	(0.20)	(0.00)	(0.00)	(0.23)	(0.00)	(0.00)	(0.31)
WCV (MAD)	1.00	1.00	0.78	1.00	1.00	0.59	1.00	1.00	0.51	0.99	0.89	0.34
	(0.00)	(0.00)	(0.15)	(0.00)	(0.00)	(0.25)	(0.00)	(0.00)	(0.34)	(0.00)	(0.01)	(0.39)
ECV (MSE)	0.98	1.00	0.75	1.00	1.00	0.70	1.00	1.00	0.60	1.00	0.96	0.23
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.03)	(0.00)	(0.01)	(0.09)
ECV (MAE)	0.98	0.97	0.13	1.00	0.94	0.11	1.00	0.82	0.09	1.00	0.56	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
ECV (MAD)	0.99	0.88	0.25	1.00	0.87	0.22	1.00	0.74	0.12	0.99	0.65	0.06
	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.02)	(0.00)	(0.01)	(0.06)
GCV (MSE)	0.76	0.63	0.39	0.70	0.44	0.17	0.57	0.33	0.16	0.25	0.00	0.00
	(0.24)	(0.37)	(0.61)	(0.30)	(0.56)	(0.83)	(0.43)	(0.67)	(0.84)	(0.75)	(1.00)	(1.00)
GCV (MAE)	0.78	0.61	0.31	0.72	0.43	0.16	0.55	0.29	0.14	0.27	0.00	0.00
	(0.22)	(0.39)	(0.69)	(0.28)	(0.57)	(0.84)	(0.45)	(0.71)	(0.86)	(0.73)	(1.00)	(1.00)
GCV (MAD)	0.55	0.39	0.24	0.42	0.37	0.18	0.35	0.24	0.09	0.30	0.03	0.00
	(0.45)	(0.61)	(0.76)	(0.58)	(0.63)	(0.82)	(0.65)	(0.76)	(0.91)	(0.70)	(0.97)	(1.00)
BCV (MSE)	0.88	0.47	0.22	0.72	0.38	0.12	0.68	0.24	0.12	0.50	0.03	0.01
	(0.12)	(0.53)	(0.77)	(0.28)	(0.62)	(0.88)	(0.32)	(0.76)	(0.88)	(0.50)	(0.97)	(0.99)
BCV (MAE)	0.89	0.48	0.23	0.73	0.41	0.13	0.67	0.22	0.13	0.46	0.03	0.01
	(0.11)	(0.52)	(0.77)	(0.27)	(0.59)	(0.87)	(0.33)	(0.78)	(0.87)	(0.54)	(0.97)	(0.98)
BCV (MAD)	0.55	0.37	0.23	0.51	0.38	0.18	0.47	0.28	0.15	0.33	0.04	0.05
	(0.45)	(0.63)	(0.75)	(0.49)	(0.62)	(0.79)	(0.53)	(0.72)	(0.84)	(0.67)	(0.96)	(0.90)

Table 4.8: Bias (RMSE in brackets) for different rank-estimation methods when using robust SVD for the simulation setting of 50×40-dimensional matrix with equal singular values.

Method	S01	S02	S03	S11	S12	S13	S21	S22	S23	S31	S32	S33
Elbow (rsvd)	−0.59	−0.54	−2.11	−0.56	−0.55	−3.64	−0.66	−0.70	−4.16	−0.55	−2.10	−5.90
	(0.77)	(1.12)	(3.73)	(0.75)	(1.41)	(5.16)	(1.17)	(1.70)	(5.64)	(1.65)	(3.73)	(6.62)
AIC (rsvd)	−7.40	−9.00	−9.00	−8.76	−9.00	−9.00	−8.68	−9.00	−9.00	−8.74	−9.00	−8.99
	(7.97)	(9.00)	(9.00)	(8.78)	(9.00)	(9.00)	(8.70)	(9.00)	(9.00)	(8.75)	(9.00)	(8.99)
BIC (rsvd)	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
PC1 (rsvd)	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
PC2 (rsvd)	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
PC3 (rsvd)	0.02	4.84	8.57	0.08	5.20	8.57	0.31	5.66	8.59	1.21	6.29	8.63
	(0.14)	(4.89)	(8.60)	(0.28)	(5.25)	(8.59)	(0.64)	(5.70)	(8.62)	(1.38)	(6.32)	(8.66)
IC1 (rsvd)	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
IC2 (rsvd)	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
IC3 (rsvd)	0.02	4.84	8.57	0.08	5.20	8.57	0.31	5.66	8.59	1.21	6.29	8.63
	(0.14)	(4.89)	(8.60)	(0.28)	(5.25)	(8.59)	(0.64)	(5.70)	(8.62)	(1.38)	(6.32)	(8.66)
DICMR	0.11	0.81	0.49	0.12	0.61	0.12	0.16	0.53	−0.44	0.12	0.02	−2.41
	(0.33)	(1.33)	(1.28)	(0.35)	(1.04)	(1.24)	(0.40)	(1.00)	(1.22)	(0.35)	(0.53)	(2.95)
WCV (MSE)	0.00	0.00	0.03	0.00	0.00	0.07	0.00	0.00	0.08	0.00	−0.07	0.02
	(0.00)	(0.00)	(0.39)	(0.00)	(0.00)	(0.56)	(0.00)	(0.00)	(0.75)	(0.00)	(0.26)	(1.29)
WCV (MAE)	0.00	0.00	0.03	0.00	0.00	0.08	0.00	0.00	0.07	0.00	−0.07	0.04
	(0.00)	(0.00)	(0.39)	(0.00)	(0.00)	(0.57)	(0.00)	(0.00)	(0.74)	(0.00)	(0.26)	(1.30)
WCV (MAD)	0.00	0.00	0.09	0.00	0.00	0.14	0.00	0.00	0.36	−0.01	−0.10	0.31
	(0.00)	(0.00)	(0.56)	(0.00)	(0.00)	(0.81)	(0.00)	(0.00)	(1.17)	(0.10)	(0.37)	(1.66)
ECV (MSE)	−0.18	0.00	−0.95	0.00	0.00	−0.69	0.00	0.00	−1.03	0.00	−0.02	−1.86
	(1.27)	(0.00)	(2.64)	(0.00)	(0.00)	(1.99)	(0.00)	(0.00)	(2.40)	(0.00)	(0.20)	(3.23)
ECV (MAE)	−0.18	−0.11	−7.76	0.00	−0.23	−7.56	0.00	−0.92	−7.89	0.00	−1.58	−7.93
	(1.27)	(0.91)	(8.35)	(0.00)	(1.30)	(8.21)	(0.00)	(2.73)	(8.40)	(0.00)	(3.17)	(8.27)
ECV (MAD)	−0.09	−0.65	−5.05	0.00	−0.49	−4.95	0.00	−0.91	−5.31	−0.01	−0.82	−4.22
	(0.90)	(2.30)	(6.53)	(0.00)	(1.86)	(6.34)	(0.00)	(2.41)	(6.50)	(0.10)	(2.12)	(5.52)
GCV (MSE)	0.35	0.62	1.13	0.49	1.49	2.48	0.81	2.25	3.77	3.46	6.46	6.85
	(0.82)	(1.21)	(1.66)	(1.01)	(2.56)	(3.39)	(1.47)	(3.27)	(4.61)	(4.73)	(6.89)	(7.19)
GCV (MAE)	0.36	0.63	1.37	0.40	1.52	2.64	0.79	2.34	3.96	3.07	6.39	7.22
	(0.89)	(1.20)	(1.89)	(0.84)	(2.63)	(3.55)	(1.37)	(3.30)	(4.76)	(4.30)	(6.88)	(7.53)
GCV (MAD)	0.84	1.38	1.84	1.41	1.69	3.00	1.73	3.17	4.51	2.83	6.12	7.56
	(1.48)	(2.09)	(2.55)	(2.23)	(2.64)	(3.95)	(2.74)	(4.23)	(5.36)	(3.93)	(6.72)	(7.91)
BCV (MSE)	0.17	0.83	1.25	0.42	1.26	1.82	0.65	1.68	2.40	1.30	4.38	4.83
	(0.54)	(1.29)	(1.87)	(0.93)	(1.83)	(2.31)	(1.49)	(2.18)	(2.88)	(2.19)	(4.95)	(5.38)
BCV (MAE)	0.16	0.80	1.31	0.34	1.12	1.78	0.58	1.64	2.25	1.34	4.31	4.72
	(0.53)	(1.26)	(1.66)	(0.71)	(1.65)	(2.25)	(1.18)	(2.10)	(2.73)	(2.25)	(4.93)	(5.30)
BCV (MAD)	0.75	1.13	1.64	0.90	1.50	1.91	1.22	2.07	2.76	2.23	4.31	4.54
	(1.26)	(1.68)	(2.38)	(1.52)	(2.22)	(2.73)	(2.07)	(2.80)	(3.62)	(3.37)	(5.10)	(5.35)

Table 4.9: Proportion of exact (overestimation in brackets) for different rank-estimation methods when using robust SVD for the simulation setting of 50×40-dimensional matrix with decreasing singular values.

Method	S01	S02	S03	S11	S12	S13	S21	S22	S23	S31	S32	S33
Elbow (rsvd)	1.00	0.55	0.09	0.97	0.36	0.03	0.93	0.32	0.01	0.86	0.11	0.04
	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.02)	(0.00)	(0.01)	(0.02)
AIC (rsvd)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BIC (rsvd)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PC1 (rsvd)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PC2 (rsvd)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PC3 (rsvd)	1.00	0.00	0.00	0.74	0.00	0.00	0.27	0.00	0.00	0.00	0.00	0.00
	(0.00)	(1.00)	(1.00)	(0.26)	(1.00)	(1.00)	(0.73)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
IC1 (rsvd)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
IC2 (rsvd)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
IC3 (rsvd)	1.00	0.00	0.00	0.74	0.00	0.00	0.27	0.00	0.00	0.00	0.00	0.00
	(0.00)	(1.00)	(1.00)	(0.26)	(1.00)	(1.00)	(0.73)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
DICMR	0.85	0.44	0.20	0.90	0.52	0.22	0.94	0.52	0.13	0.96	0.52	0.03
	(0.15)	(0.50)	(0.27)	(0.10)	(0.42)	(0.27)	(0.06)	(0.28)	(0.12)	(0.04)	(0.11)	(0.01)
WCV (MSE)	1.00	0.78	0.24	1.00	0.75	0.24	1.00	0.55	0.22	0.93	0.33	0.23
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)	(0.00)	(0.00)	(0.09)	(0.00)	(0.00)	(0.19)
WCV (MAE)	1.00	0.78	0.24	1.00	0.75	0.24	1.00	0.54	0.22	0.94	0.33	0.23
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)	(0.00)	(0.00)	(0.09)	(0.00)	(0.00)	(0.19)
WCV (MAD)	1.00	0.65	0.20	1.00	0.65	0.31	0.99	0.48	0.14	0.90	0.27	0.26
	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)	(0.10)	(0.00)	(0.00)	(0.09)	(0.00)	(0.02)	(0.18)
ECV (MSE)	1.00	0.89	0.24	1.00	0.89	0.16	1.00	0.79	0.15	1.00	0.55	0.14
	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)	(0.09)	(0.00)	(0.06)	(0.27)
ECV (MAE)	1.00	0.36	0.00	1.00	0.24	0.01	1.00	0.24	0.00	1.00	0.12	0.01
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
ECV (MAD)	1.00	0.41	0.07	1.00	0.34	0.09	0.99	0.38	0.05	0.99	0.21	0.05
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.03)	(0.05)	(0.00)	(0.07)	(0.15)
GCV (MSE)	0.82	0.56	0.27	0.73	0.45	0.17	0.79	0.29	0.16	0.46	0.03	0.00
	(0.18)	(0.44)	(0.60)	(0.27)	(0.55)	(0.75)	(0.21)	(0.71)	(0.80)	(0.54)	(0.97)	(1.00)
GCV (MAE)	0.84	0.53	0.27	0.79	0.39	0.16	0.72	0.25	0.15	0.42	0.01	0.00
	(0.16)	(0.47)	(0.63)	(0.21)	(0.61)	(0.78)	(0.28)	(0.75)	(0.81)	(0.58)	(0.99)	(1.00)
GCV (MAD)	0.44	0.39	0.20	0.49	0.26	0.09	0.53	0.15	0.09	0.25	0.04	0.01
	(0.56)	(0.61)	(0.69)	(0.51)	(0.74)	(0.83)	(0.47)	(0.83)	(0.86)	(0.75)	(0.96)	(0.99)
BCV (MSE)	0.92	0.38	0.29	0.81	0.34	0.22	0.80	0.24	0.19	0.58	0.03	0.05
	(0.08)	(0.62)	(0.57)	(0.19)	(0.66)	(0.64)	(0.20)	(0.76)	(0.63)	(0.42)	(0.97)	(0.94)
BCV (MAE)	0.93	0.41	0.28	0.78	0.35	0.22	0.79	0.23	0.18	0.60	0.03	0.05
	(0.07)	(0.59)	(0.57)	(0.22)	(0.65)	(0.63)	(0.21)	(0.77)	(0.65)	(0.40)	(0.97)	(0.94)
BCV (MAD)	0.58	0.29	0.27	0.58	0.39	0.16	0.56	0.22	0.12	0.39	0.04	0.02
	(0.42)	(0.71)	(0.52)	(0.42)	(0.61)	(0.61)	(0.44)	(0.78)	(0.63)	(0.61)	(0.96)	(0.94)

Table 4.10: Bias (RMSE in brackets) for different rank-estimation methods when using robust SVD for the simulation setting of 50×40-dimensional matrix with decreasing singular values.

Method	S01	S02	S03	S11	S12	S13	S21	S22	S23	S31	S32	S33
Elbow (rsvd)	0.00	−2.12	−5.26	−0.03	−3.04	−5.69	−0.07	−3.89	−5.99	−0.27	−5.81	−6.43
	(0.00)	(3.79)	(6.19)	(0.17)	(4.54)	(6.45)	(0.26)	(5.24)	(6.62)	(1.12)	(6.69)	(7.05)
AIC (rsvd)	−2.47	−9.00	−9.00	−3.00	−9.00	−9.00	−3.45	−8.97	−9.00	−4.19	−8.49	−8.87
	(2.52)	(9.00)	(9.00)	(3.03)	(9.00)	(9.00)	(3.49)	(8.97)	(9.00)	(4.23)	(8.51)	(8.88)
BIC (rsvd)	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
PC1 (rsvd)	−6.65	−9.00	−9.00	−6.96	−9.00	−9.00	−7.12	−9.00	−9.00	−7.56	−9.00	−9.00
	(6.67)	(9.00)	(9.00)	(6.98)	(9.00)	(9.00)	(7.14)	(9.00)	(9.00)	(7.58)	(9.00)	(9.00)
PC2 (rsvd)	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
PC3 (rsvd)	0.00	9.39	10.00	0.27	9.49	10.00	1.05	9.55	10.00	2.82	9.59	10.00
	(0.00)	(9.41)	(10.00)	(0.54)	(9.51)	(10.00)	(1.33)	(9.57)	(10.00)	(2.92)	(9.60)	(10.00)
IC1 (rsvd)	−6.65	−9.00	−9.00	−6.96	−9.00	−9.00	−7.12	−9.00	−9.00	−7.56	−9.00	−9.00
	(6.67)	(9.00)	(9.00)	(6.98)	(9.00)	(9.00)	(7.14)	(9.00)	(9.00)	(7.58)	(9.00)	(9.00)
IC2 (rsvd)	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00	−9.00
	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)	(9.00)
IC3 (rsvd)	0.00	9.39	10.00	0.27	9.49	10.00	1.05	9.55	10.00	2.82	9.59	10.00
	(0.00)	(9.41)	(10.00)	(0.54)	(9.51)	(10.00)	(1.33)	(9.57)	(10.00)	(2.92)	(9.60)	(10.00)
DICMR	0.19	0.75	−0.49	0.10	0.45	−0.68	0.06	0.15	−1.47	0.04	−0.37	−2.73
	(0.54)	(1.37)	(1.51)	(0.32)	(0.87)	(1.87)	(0.24)	(0.84)	(2.14)	(0.20)	(0.93)	(3.06)
WCV (MSE)	0.00	−0.23	−1.50	0.00	−0.27	−1.31	0.00	−0.53	−1.23	−0.07	−0.90	−0.74
	(0.00)	(0.50)	(1.93)	(0.00)	(0.56)	(1.86)	(0.00)	(0.83)	(1.83)	(0.26)	(1.17)	(1.57)
WCV (MAE)	0.00	−0.23	−1.49	0.00	−0.27	−1.32	0.00	−0.54	−1.22	−0.06	−0.90	−0.74
	(0.00)	(0.50)	(1.92)	(0.00)	(0.56)	(1.87)	(0.00)	(0.84)	(1.82)	(0.24)	(1.17)	(1.57)
WCV (MAD)	0.00	−0.46	−1.59	0.00	−0.40	−1.06	−0.01	−0.69	−1.41	−0.11	−0.97	−0.80
	(0.00)	(0.85)	(2.19)	(0.00)	(0.71)	(1.88)	(0.10)	(1.03)	(2.01)	(0.36)	(1.32)	(1.67)
ECV (MSE)	0.00	−0.14	−2.47	0.00	−0.13	−2.84	0.00	−0.28	−2.38	0.00	−0.38	−1.05
	(0.00)	(0.45)	(3.52)	(0.00)	(0.41)	(3.81)	(0.00)	(1.11)	(3.78)	(0.00)	(0.98)	(2.89)
ECV (MAE)	0.00	−2.86	−8.81	0.00	−3.55	−8.44	0.00	−2.88	−8.40	0.00	−3.54	−7.49
	(0.00)	(4.39)	(8.85)	(0.00)	(4.88)	(8.63)	(0.00)	(3.99)	(8.56)	(0.00)	(4.62)	(7.92)
ECV (MAD)	0.00	−1.96	−6.39	0.00	−2.19	−5.81	−0.01	−1.64	−5.11	−0.01	−1.86	−3.22
	(0.00)	(3.19)	(7.15)	(0.00)	(3.27)	(6.66)	(0.10)	(2.80)	(6.25)	(0.10)	(3.13)	(5.09)
GCV (MSE)	0.30	0.71	1.00	0.43	1.30	2.18	0.44	2.56	3.21	2.16	6.25	7.12
	(0.82)	(1.25)	(1.74)	(0.93)	(2.16)	(3.14)	(1.10)	(3.66)	(4.30)	(3.53)	(6.81)	(7.45)
GCV (MAE)	0.24	0.75	1.19	0.35	1.57	2.68	0.68	2.87	3.76	2.16	6.66	7.30
	(0.71)	(1.28)	(2.02)	(0.88)	(2.46)	(3.73)	(1.46)	(3.97)	(4.86)	(3.36)	(7.16)	(7.62)
GCV (MAD)	1.20	1.49	1.88	1.16	2.53	3.15	1.57	3.41	4.33	3.29	6.59	7.26
	(2.04)	(2.29)	(2.85)	(1.90)	(3.71)	(4.17)	(2.85)	(4.47)	(5.37)	(4.42)	(7.22)	(7.61)
BCV (MSE)	0.11	0.96	0.70	0.21	1.36	1.04	0.28	1.87	1.48	1.00	4.21	4.30
	(0.44)	(1.37)	(1.51)	(0.50)	(1.96)	(2.05)	(0.68)	(2.55)	(2.72)	(1.90)	(4.70)	(4.91)
BCV (MAE)	0.10	0.88	0.68	0.26	1.35	1.03	0.31	1.92	1.54	0.84	4.18	4.22
	(0.42)	(1.29)	(1.52)	(0.58)	(1.97)	(2.09)	(0.74)	(2.60)	(2.75)	(1.64)	(4.70)	(4.85)
BCV (MAD)	0.75	1.30	0.86	0.82	1.62	1.14	1.06	2.38	1.38	1.47	4.59	4.61
	(1.32)	(1.76)	(2.17)	(1.46)	(2.40)	(2.51)	(1.95)	(3.28)	(3.05)	(2.20)	(5.32)	(5.37)

Chapter 5

Breakdown Analysis of Minimum Super Divergence Estimator

5.1 Introduction

The influence function and the breakdown point are two of the most popular metrics to measure the robustness of an estimator. Unlike the influence function which is a local measure of robustness, the breakdown point of an estimator (Hampel, 1971) can determine its global reliability. It refers to the minimum proportion of observations in the sample which can be replaced to arbitrarily modify the value of an estimator. Previously in Section 1.3.3 of Chapter 1, we described the notion of breakdown point in detail.

In previous chapters, we discussed how the minimum density power divergence estimator (MDPDE) can be used to perform matrix factorization techniques (such as SVD and PCA) robustly, which is a primary ingredient of many high-dimensional statistical inference problems. In this chapter, we aim to quantify the robustness of this estimator and provide a justification for its applicability and strong robustness properties in the high-dimensional context. To illustrate this point, we shall show that a broad class of minimum distance estimators (including the MDPDE) has an asymptotic breakdown point no less than a dimension-independent lower bound. In comparison, the classical M-estimators that are affine-equivariant have a diminishing asymptotic breakdown point as the data dimension increases (Maronna, 1976).

5.2 Existing Literature

Among various traditional robust estimators of location, the sample median and Hodges-Lehman estimator (Hodges, 1967) have asymptotic breakdown points as $1/2$ and $(1-1/\sqrt{2})$ respectively. To gain more efficiency compared to these classical estimators, Huber (1964b) introduced M-estimators which were further developed later on by Maronna (1976) for multivariate location and scatter estimation. However, Rousseeuw (1985) showed that all affine-equivariant M-estimators have an asymptotic breakdown point at most $1/(p+1)$ where p is the dimension of the data. Therefore, for high-dimensional data, the robustness

of an M-estimator can decay rapidly. To counter this, Rousseeuw (1985) introduced minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimators, and Rousseeuw and Yohai (1984) introduced the general class of "S-estimators", which were shown to be less efficient than M-estimators but with a relatively high asymptotic breakdown point. Davies (1987) extended the S-estimators to the setup of multivariate location and scatter and derived the asymptotic properties of the same for the exponential family of distributions. Till now, numerous studies have investigated both finite-sample and asymptotic breakdown points of different classes of M-estimators and S-estimators; see Smucler (2019), Fishbone and Mili (2021), Park et al. (2022) and the references therein for further details.

Compared to the vast literature on the asymptotic breakdown points of various classical robust estimators, the results on the breakdown points of the minimum divergence estimators are rather limited. They are often restricted to some special classes of divergence measures and restricted setups of location or scale families of model densities. Park and Basu (2004) demonstrated that the asymptotic breakdown point of a minimum divergence estimator within the ϕ-divergence family is $1/2$ under suitable conditions. However, Park and Basu (2004) only consider cases of breakdown when the absolute value of the estimator goes to infinity, which does not encompass all kinds of breakdown, e.g., the breakdown of the scale parameter with the scale going to 0. Ghosh and Basu (2013) established similar results for the minimum density power divergence estimator (MDPDE) of the location parameter for independent but non-homogeneous sample observations from a location-scale family of distributions with a fixed scale parameter. In particular, the authors restricted their attention towards the model family of densities of the form

$$\mathcal{F} = \left\{ \frac{1}{\sigma} f\left(\frac{y - l(\mu)}{\sigma} \right) : \theta = (\mu, \sigma) \in \mathbb{R} \times (0, \infty) \right\}, \tag{5.1}$$

where $l : \mathbb{R} \to \mathbb{R}$ is an arbitrary but known one-one function. Under such models, they derived the breakdown point of the MDPDE of the location parameter μ to be equal to $1/2$ at the model density, while the scale parameter σ was assumed to be fixed at σ_0; for example, σ_0 can be substituted with any suitable robust scale estimator. They also assumed that the true data generating density g to reside in the model family $\mathcal{F}$.

Under a different set of assumptions, the minimum Hellinger distance estimator (MHDE) has been shown to attain an asymptotic breakdown point of at least $1/4$ by (Tamura and Boos, 1986) for multivariate location and scatter estimation. Toma (2008), later, improved this bound to $1/2$. On the other hand, Simpson (1987) showed that the asymptotic breakdown point for MHDE can be as large as $1/2$ under multinomial setup models with discrete countable support. These specific results are remarkable as they show the highly robust properties of the MHDE, which is a minimum divergence estimator, by achieving an asymptotic breakdown point free of the dimension of the data, unlike the shrinking bounds offered for the M-estimator. For the minimum generalized negative exponential disparity estimator (MGNEDE), a different family of minimum divergence estimator, Bhandari et al. (2006) proved that the asymptotic breakdown point of MGNEDE is $1/2$ when the

true distribution belongs to the model family, under assumptions similar to those of Park and Basu (2004).

However, to the best of our knowledge, there does not exist any literature on the asymptotic breakdown point of the minimum super-divergence estimator (MSDE) as defined in Section 1.3.5, except for the few special cases of density power divergence or power divergence family as discussed above. In this section, we aim to investigate the breakdown properties of this estimator, and demonstrate that under some reasonable assumptions, the resulting breakdown point has a dimension-free breakdown point. This makes MSDE a very attractive choice among robust estimators for multivariate problems, such as matrix factorization problems. As a consequence, our results also generalized many of these scattered theoretical results mentioned above, under a single unified framework.

The rest of this chapter is arranged as follows. In Section 5.3, we present our theoretical results. We present a few key assumptions under which we establish our general result, and then indicate some different sufficient conditions that are often easier to verify. In Section 5.4, we show multiple examples to demonstrate the applicability of our results across various setups. Section 5.5 corroborates these theoretical findings by simulation experiments, and also provides additional empirical illustrations beyond the scope of the results present in this chapter.

5.3 Theoretical Analysis of Breakdown Point

In this section, we first investigate the asymptotic breakdown properties of the MSD functional given in (1.14). Subsequently, we shall show, in Theorem 5.2, that the same breakdown analysis also applies when data density g is substituted by a suitable density estimate $\widehat{g}_n$ based on the sample observations, provided that the density estimate $\widehat{g}_n$ satisfies a few suitable assumptions. The results pertaining to the MDPDE and the minimum power divergence functional will then follow from these general results with $\lambda = 0$ and $\alpha = 0$ respectively.

In view of the sufficient conditions given in Section 1.3.5 to ensure the finiteness of the S-divergence, we restrict our attention to the situations where both $A, B \geq 0$ and the densities f and g in consideration are $L^{1+\alpha+\delta}$-integrable for some $\delta > 0$. Since the form of S-divergence is valid as long as at least one of $A = 1 + \lambda(1-\alpha)$ and $B = \alpha - \lambda(1-\alpha)$ are nonzero, technically, therefore, one can allow the tuning parameter α to go beyond 1. However, as discussed in Basu et al. (1998), choices of $\alpha > 1$ lead to unacceptably low efficiencies, and are generally avoided in practical implementations. Hence, we restrict the choice of α in the unit interval $[0, 1]$. This, together with the nonnegativity of A and B implies that either $\alpha = 1$, or $\alpha \in [0, 1)$ and $-1/(1-\alpha) \leq \lambda \leq \alpha/(1-\alpha)$. These choices are not too restrictive, they include the special classes of density power divergences at $\lambda = 0$, power divergences at $\alpha = 0$, and, S-Hellinger distances (SHD) at $\lambda = -1/2$. For the particular cases when $A = 0$ or $B = 0$, we additionally restrict our attention to the pair of densities satisfying the support condition and the condition on the finiteness of the

cross-integrals as required (see Section 1.3.5 for details). Note that, both A and B cannot be equal to 0 simultaneously as $(A+B)=(1+\alpha)\geq 1$.

Let G be the true distribution with density g, K_m denote a sequence of contaminating distributions with densities k_m and let $G_{\epsilon,m}=(1-\epsilon)G+\epsilon K_m$ denote the ϵ-contaminated distributions with densities $g_{\epsilon,m}$. To investigate the asymptotic breakdown point of the MSD functional $T_{(\alpha,\lambda)}(G)$ as defined in (1.14), we shall look into how $T_{(\alpha,\lambda)}(G_{\epsilon,m})$ changes as a function of ϵ as m tends to infinity. In the following, we present the key assumptions underlying our results. These are similar in spirit but cover more general model setups, compared to the assumptions considered by Ghosh and Basu (2013) to obtain the asymptotic breakdown point of the MDPDE for the location parameter.

(BP1) The sequence of contaminating densities k_m becomes asymptotically singular to the true density g, i.e., $\int \min\{g,k_m\}d\mu\to 0$ as $m\to\infty$.

(BP2) The sequence of contaminating densities k_m is such that for any compact subset $S\subset\Theta$ with $S\cap\partial\Theta=\phi$, we have $\int\min\{f_\theta,k_m\}d\mu\to 0$ as $m\to\infty$ uniformly on $\theta\in S$.

(BP3) The density f_θ belonging to the model family is such that for any sequence of parameters $\theta_m\to\theta_\infty$ where $\theta_\infty\in\partial\Theta$, the integral $\int\min\{g,f_{\theta_m}\}d\mu\to 0$ as $m\to\infty$.

(BP4) If $A>0$, the model family and the family of contaminating densities are uniformly $L^{1+\alpha}$-integrable, i.e.,

$$\limsup_{m\to\infty}\int k_m^{1+\alpha}<\infty,\ \text{and},\ \sup_{\theta\in\Theta}\int f_\theta^{1+\alpha}<\infty,$$

For $A=0$, these families of densities are uniformly $L^{1+\alpha+\delta}$-integrable for some $\delta>0$, and additionally, the integrals

$$\sup_{\theta\in\Theta}\left|\int f_\theta^{1+\alpha}\ln(g)\right|,\ \text{and},\ \sup_{\theta\in\Theta}\sup_m\left|\int f_\theta^{1+\alpha}\ln(k_m)\right|,$$

exist and are finite.

The Assumptions (BP1) and (BP2) ensure that the sequence of contaminating densities is asymptotically singular to the true density g and the model family of densities for the parameters lying in the interior of the parameter space. The Assumption (BP3) ensures that the true density g is asymptotically singular to the model family of densities when the parameter tends towards the boundary of the parameter space. These three assumptions are standard in the breakdown point literature, see Park and Basu (2004), Ghosh and Basu (2013). Finally, Assumption (BP4) ensures finiteness of the S-divergence itself. This assumption holds in many different situations as follows:

1. In any setup with $\alpha=0$, i.e., when restricting the S-divergence to only the power divergence family.

2. For a location estimation problem where both the model family and the contaminating densities belong to a location model family.

3. For a scale estimation problem where the model family and the contaminating densities belong to a scale model family, and the contaminating scale remains bounded away from zero, including the case when the contaminating scale parameter "explodes" to infinity.

5.3.1 Main Results

We begin by indicating the general result regarding the asymptotic breakdown point of MSD functional.

Theorem 5.1. *Let $B > 0$ and the Assumptions (BP1)-(BP4) hold. Additionally, assume that there exists $\widetilde{\epsilon}_{(\alpha,\lambda)} \in (0, 1/2]$ such that for all $\epsilon \in [0, \widetilde{\epsilon}_{(\alpha,\lambda)})$, the S-divergence between ϵk_m and f_{θ_m} satisfy the inequality*

$$\liminf_{m\to\infty} S_{(\alpha,\lambda)}(\epsilon k_m, f_{\theta_m}) > \limsup_{m\to\infty} \frac{\epsilon^{(1+\alpha)}}{B} \|k_m\|_{1+\alpha}^{1+\alpha} + q_{(\alpha,\lambda)}(1-\epsilon) \|g\|_{1+\alpha}^{1+\alpha}, \tag{5.2}$$

for any sequence of parameters $\{\theta_m\}_{m=1}^{\infty}$ satisfying $\theta_m \to \theta_\infty \in \partial\Theta$ as $m \to \infty$, where

$$q_{(\alpha,\lambda)}(\epsilon) = \begin{cases} \frac{1}{A} - \frac{1+\alpha}{AB}\epsilon^A & \text{if } A > 0, \\ \ln(1/\epsilon) - (1+\alpha)^{-1} & \text{if } A = 0, \end{cases} \quad \text{for any } \epsilon \in (0, 1]. \tag{5.3}$$

Then if the true density g belongs to the interior of the model family of densities $\mathcal{F}$, i.e., $g = f_{\theta^g}$ for some $\theta^g \in \Theta \setminus \partial\Theta$, then the MSD functional $T_{(\alpha,\lambda)}(G)$ has an asymptotic breakdown point at least $\min\{1/2, \widetilde{\epsilon}_{(\alpha,\lambda)}\}$.

There are a few key points to note here. Firstly, the function ϵk_m is not a density per se, but there is no mathematical difficulty in constructing the quantity $S_{(\alpha,\lambda)}(\epsilon k_m, f_{\theta_m})$ in the usual way in (5.2). Secondly, since the true density g is assumed to belong to the model family $\mathcal{F}$, the minimum S-divergence estimator exists and is unique (Ghosh et al., 2017), hence discussing its asymptotic breakdown point is a meaningful endeavour.

Remark 5.1. *Note that, Theorem 5.1 only refers to the boundary $\partial\Theta$ through a sequence of estimates $\theta_m \to \theta_\infty \in \partial\Theta$. It does not use any special property of the boundary $\partial\Theta$. Therefore, if one wishes to consider only a specific type of breakdown, it can be done by considering a restriction to any non-empty subset of $\partial\Theta$, including singletons containing only a specific point.*

For example, a scale parameter $\theta \in (0, \infty)$ can exhibit both "imploding" and "exploding" breakdowns as $\partial\Theta = \{0, \infty\}$. However, if one wishes to consider only the "exploding" type of breakdown (i.e., $\sigma_m \to \infty$ where σ_m is the scale parameter for contaminating density k_m for $m = 1, 2, \dots$), one may consider a restriction on the parameter space as $\Theta = [\delta_0, \infty)$ for some small $\delta_0 > 0$. In this case, the boundary is $\partial\Theta = \{\delta_0, \infty\}$. However, for the point $\delta_0 \in \partial\Theta$, the singularity assumptions (BP1)-(BP3) are not usually satisfied. In this case, to consider only "explosion"-type breakdown, one can apply Theorem 5.1 only for the proper subset $\{+\infty\}$ of the boundary $\partial\Theta = \{\delta_0, \infty\}$.

Now, we proceed with the proof of Theorem 5.1.

Proof. (of Theorem 5.1) We shall follow an approach similar to that of Park and Basu (2004). Let, ϵ ($< \epsilon^*$) be a fixed level of contamination where the breakdown occurs. This means, there exists a sequence of densities $\{k_m\}$ for contamination such that the minimum S-divergence (MSD) functional θ_m minimizing $S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_\theta)$ satisfy $\theta_m \to \theta_\infty$ as $m \to \infty$ where θ_∞ is a boundary point of Θ.

Step 1:

In the first step of the proof, we shall find the limit of the S-divergence between the sequence of contaminated densities $g_{\epsilon,m}$ and the model densities f_{θ_m}. Note that

$$S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta_m}) = \int_{A_m} \mathcal{S}_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta_m}) + \int_{A_m^c} \mathcal{S}_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta_m}), \tag{5.4}$$

where

$$\mathcal{S}_{(\alpha,\lambda)}(g, f) = \begin{cases} \frac{1}{A} f^{1+\alpha} - \frac{(1+\alpha)}{AB} f^B g^A + \frac{1}{B} g^{1+\alpha} & \text{if } A > 0 \\ f^{1+\alpha} \ln(f/g) - \frac{1}{1+\alpha} f^{1+\alpha} + \frac{1}{1+\alpha} g^{1+\alpha} & \text{if } A = 0 \end{cases}$$

and $A_m = \{x : g(x) > \max\{k_m(x), f_{\theta_m}(x)\}\}$. We will show that the first integral in (5.4) is asymptotically equivalent to $(1-\epsilon)^{1+\alpha} \|g\|_{1+\alpha}^{1+\alpha}/B$ as $m \to \infty$ and the second integral in (5.4) is asymptotically same as $S_{(\alpha,\lambda)}(\epsilon k_m, f_{\theta_m})$ as $m \to \infty$. We show this separately for $A > 0$ and $A = 0$ cases.

Step 1, First term, $A > 0$:

It follows from Assumption (BP2)-(BP3) that $\int_{A_m} k_m = \int k_m \mathbf{1}_{A_m} \to 0$ and $\int f_{\theta_m} \mathbf{1}_{A_m} \to 0$ as $m \to \infty$. Therefore, we have $g_{\epsilon,m}\mathbf{1}_{A_m} \to (1-\epsilon)g$ pointwise, and $f_{\theta_m}\mathbf{1}_{A_m} \to 0$ pointwise. Also, $\sup_m \|g_{\epsilon,m}\mathbf{1}_{A_m}\|_{1+\alpha} \le \|g^{1+\alpha}\| < \infty$. Hence, by Lemma 5.A.1 with $g_m = g_{\epsilon,m}\mathbf{1}_{A_m}$, $f_m = f_{\theta_m}\mathbf{1}_{A_m}$ and $\widetilde{g}_m = (1-\epsilon)g\mathbf{1}_{A_m}$ it follows that as $m \to \infty$, we get

$$\begin{aligned} \|g_{\epsilon,m}\mathbf{1}_{A_m}\|_{1+\alpha}^{1+\alpha} &\asymp (1-\epsilon)^{1+\alpha} \|g\mathbf{1}_{A_m}\|_{1+\alpha}^{1+\alpha}, \\ \langle f_{\theta_m}, g_{\epsilon,m}\mathbf{1}_{A_m}\rangle_{B,A} &\asymp (1-\epsilon)^{A} \langle f_{\theta_m}, g\mathbf{1}_{A_m}\rangle_{B,A}, \end{aligned}$$

where the symbol $\asymp$ denote the asymptotic equivalence as $m \to \infty$. Also, applying Lemma 5.A.1 with $g_m = f_{\theta_m}\mathbf{1}_{A_m}$, $\widetilde{g}_m \equiv 0$ and $f_m = g$, we further obtain

$$\|f_{\theta_m}\mathbf{1}_{A_m}\|_{1+\alpha}^{1+\alpha} \to 0, \text{ and } \left|\langle f_{\theta_m}, g\mathbf{1}_{A_m}\rangle_{B,A}\right| \to 0, \text{ as } m \to \infty.$$

Combining all these above and noting that $A_m \to \{x : g(x) > 0\}$, we obtain for $A > 0$, as $m \to \infty$,

$$\int_{A_m} \mathcal{S}_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta_m}) \asymp \frac{(1-\epsilon)^{1+\alpha}}{B} \|g\|_{1+\alpha}^{1+\alpha}.$$

Step 1, Second term, $A > 0$:

Now, focusing on the behaviour of the divergence on the set A_m^c and by using Assumptions (BP2) and (BP3) we obtain that $\int_{A_m^c} g = \int g\mathbf{1}_{A_m^c} \to 0$ as $m \to \infty$. It implies that $(g_{\epsilon,m} - \epsilon k_m)\mathbf{1}_{A_m^c} \to 0$ pointwise. Additionally, we note that

$$\sup_m \int g_{\epsilon,m}^{1+\alpha}\mathbf{1}_{A_m^c} \le \sup_m \int \max\{f_{\theta_m}^{1+\alpha}, k_m^{1+\alpha}\} \le \sup_m \int f_{\theta_m}^{1+\alpha} + \sup_m \int k_m^{1+\alpha} < \infty,$$

where we use the fact that $g \le \max\{f_{\theta_m}, k_m\}$ on A_m^c and $\max\{a,b\} \le (a+b)$ for $a, b \ge 0$, along with Assumption (BP4). For $A > 0$, an application of Lemma 5.A.1 with $g_m = g_{\epsilon,m}\mathbf{1}_{A_m^c}$, $f_m = f_{\theta_m}\mathbf{1}_{A_m^c}$ and $\widetilde{g}_m = \epsilon k_m \mathbf{1}_{A_m^c}$ yields that as $m \to \infty$, we get

$$\left| \int g_{\epsilon,m}^{1+\alpha}\mathbf{1}_{A_m^c} - \epsilon^{1+\alpha} \int k_m^{1+\alpha}\mathbf{1}_{A_m^c} \right| \to 0, \text{ and } \left| \int f_{\theta_m}^B g_{\epsilon,m}^A \mathbf{1}_{A_m^c} - \epsilon^A \int f_{\theta_m}^B k_m^A \mathbf{1}_{A_m^c} \right| \to 0.$$

Hence, for $A > 0$, we must have the following convergence

$$\left| \int_{A_m^c} \mathcal{S}_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta_m}) - \int \mathcal{S}_{(\alpha,\lambda)}(\epsilon k_m, f_{\theta_m}) \right| \to 0.$$

Combining the above two subcases, it follows that as $m \to \infty$, for all $B > 0$ and $A > 0$, we have

$$S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta_m}) \asymp \frac{(1-\epsilon)^{1+\alpha}}{B} \|g\|_{1+\alpha}^{1+\alpha} - S_{(\alpha,\lambda)}(\epsilon k_m, f_{\theta_m}).$$

Now, we turn to the case when $A = 0$.

Step 1, First term, $A = 0$:

Starting with the integral over A_m, we note that

$$\int_{A_m} \mathcal{S}_{(\alpha,\lambda;A=0)}(g_{\epsilon,m}, f_{\theta_m}) = \int_{A_m} f_{\theta_m}^{1+\alpha} \ln(f_{\theta_m}/g_{\epsilon,m}) - \frac{1}{1+\alpha}\int_{A_m} f_{\theta_m}^{1+\alpha} + \frac{1}{1+\alpha}\int_{A_m} g_{\epsilon,m}^{1+\alpha}.$$

By the same argument for the case with $A > 0$ as before, it follows that $\int_{A_m} f_{\theta_m}^{1+\alpha} \to 0$ and $\int_{A_m} g_{\epsilon,m}^{1+\alpha} \to (1-\epsilon)^{1+\alpha} \|g\|_{1+\alpha}^{1+\alpha}$ as $m \to \infty$. To deal with the cross-integral term, we additionally have by the triangle inequality

$$\left| \int f_{\theta_m}^{1+\alpha}\mathbf{1}_{A_m} \ln(f_{\theta_m}/g_{\epsilon,m}) \right| \le \left| \int f_{\theta_m}^{1+\alpha}\mathbf{1}_{A_m} \ln(f_{\theta_m}) \right| + \left| \int f_{\theta_m}^{1+\alpha}\mathbf{1}_{A_m} \ln(g_{\epsilon,m}) \right|.$$

We will show that both of these terms go to 0 as $m \to \infty$. For the first term, we take $g_m = f_m = f_{\theta_m}\mathbf{1}_{A_m}$ and $\widetilde{g}_m = \widetilde{f}_m \equiv 0$, and apply Lemma 5.A.2. For the second term, we take $g_m = f_{\theta_m}\mathbf{1}_{A_m}$, $f_m = g_{\epsilon,m}\mathbf{1}_{A_m}$, $\widetilde{g}_m \equiv 0$ and $\widetilde{f}_m = (1-\epsilon)g$ and again apply Lemma 5.A.2. To show the integrability of the cross-entropy terms in this case, we notice the chain of inequalities

$$\begin{aligned}
\int f_{\theta_m}^{1+\alpha}\mathbf{1}_{A_m} \ln(g_{\epsilon,m}) &\le \int f_{\theta_m}^{1+\alpha}\mathbf{1}_{A_m} \ln(g), \text{ (since, } g > k_m \text{ on } A_m) \\
&= \int_{A_m \cap \{x: g(x) \le 1\}} f_{\theta_m}^{1+\alpha} \ln(g) + \int_{A_m \cap \{x: g(x) > 1\}} f_{\theta_m}^{1+\alpha} \ln(g) \\
&\le \int_{A_m \cap \{x: g(x) > 1\}} f_{\theta_m}^{1+\alpha} \ln(g), \text{ (since the first term is nonpositive)} \\
&\le \int_{A_m} g^{1+\alpha} \ln(g), \text{ (since } g > f_{\theta_m} \text{ on } A_m),
\end{aligned}$$

which is finite as shown in the proof of Lemma 5.A.2. On the other hand, we have

$$\begin{aligned}
& \int f_{\theta_m}^{1+\alpha}\mathbf{1}_{A_m} \ln(g_{\epsilon,m}) \\
= & \int_{A_m \cap \{x: f_{\theta_m}(x) \ge k_m(x)\}} f_{\theta_m}^{1+\alpha} \ln(g_{\epsilon,m}) + \int_{A_m \cap \{x: f_{\theta_m}(x) < k_m(x)\}} f_{\theta_m}^{1+\alpha}\mathbf{1}_{A_m} \ln(g_{\epsilon,m}) \\
\ge & \int_{A_m \cap \{x: f_{\theta_m}(x) \ge k_m(x)\}} f_{\theta_m}^{1+\alpha} \ln((1-\epsilon) f_{\theta_m}) + \int_{A_m \cap \{x: f_{\theta_m}(x) < k_m(x)\}} f_{\theta_m}^{1+\alpha} \ln(f_{\theta_m}) \\
= & \int_{A_m} f_{\theta_m}^{1+\alpha} \ln(f_{\theta_m}) + \ln(1-\epsilon) \int_{A_m \cap \{x: f_{\theta_m}(x) \ge k_m(x)\}} f_{\theta_m}^{1+\alpha},
\end{aligned}$$

which is again finite due to the uniform $L^{1+\alpha+\delta}$-integrability of f_{θ_m} and noting that $\theta_m \in \Theta$. This shows that the cross-entropy term satisfy

$$\sup_m \left| \int f_{\theta_m}^{1+\alpha} \mathbf{1}_{A_m} \ln(g_{\epsilon,m}) \right| < \infty.$$

For $\widetilde{g}_m \equiv 0$ and $\widetilde{f}_m = (1-\epsilon)g$, the boundedness of the cross-entropy term follows trivially from $L^{1+\alpha}$-integrability of g. Therefore, for $A = 0$, we obtain

$$\left| \int_{A_m} \mathcal{S}_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta_m}) - \frac{(1-\epsilon)^{1+\alpha}}{1+\alpha} \|g\|_{1+\alpha}^{1+\alpha} \right| \to 0,$$

as $m \to \infty$.

Step 1, Second term, $A = 0$:

Now we turn to the integral over A_m^c for $A = 0$ case. To show this, consider the chain of inequalities

$$\begin{aligned}
&\left| \int_{A_m^c} \mathcal{S}_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta_m}) - \int \mathcal{S}_{(\alpha,\lambda)}(\epsilon k_m, f_{\theta_m}) \right| \\
&= \left| \int_{A_m^c} f_{\theta_m}^{1+\alpha} \ln(\epsilon k_m / g_{\epsilon,m}) + \frac{1}{1+\alpha} \int_{A_m^c} g_{\epsilon,m}^{1+\alpha} - \frac{1}{1+\alpha} \int_{A_m^c} \epsilon^{1+\alpha} k_m^{1+\alpha} \right| \\
&\leq \left| \int f_{\theta_m}^{1+\alpha} \mathbf{1}_{A_m^c} \ln(g_{\epsilon,m}) - \int f_{\theta_m}^{1+\alpha} \mathbf{1}_{A_m^c} \ln(\epsilon k_m) \right| + \frac{1}{1+\alpha} \left| \int g_{\epsilon,m}^{1+\alpha} \mathbf{1}_{A_m^c} - \epsilon^{1+\alpha} \int k_m^{1+\alpha} \mathbf{1}_{A_m^c} \right|.
\end{aligned}$$

That the second term goes to 0 follows from an application of Lemma 5.A.1 as indicated earlier. For the first term, we use Lemma 5.A.2 instead. For this to be valid, we again need to show the boundedness of the cross-entropy terms, i.e.,

$$\sup_m \left| \int f_{\theta_m}^{1+\alpha} \mathbf{1}_{A_m^c} \ln(\epsilon k_m) \right| < \infty, \text{ and } \sup_m \left| \int f_{\theta_m}^{1+\alpha} \mathbf{1}_{A_m^c} \ln(g_{\epsilon,m}) \right| < \infty.$$

For the first cross-entropy term, we have

$$\left| \int f_{\theta_m}^{1+\alpha} \mathbf{1}_{A_m^c} \ln(\epsilon k_m) \right| \leq \left| \int f_{\theta_m}^{1+\alpha} \mathbf{1}_{A_m^c} \ln(k_m) \right| + \ln(1/\epsilon) \left| \int f_{\theta_m}^{1+\alpha} \mathbf{1}_{A_m^c} \right|,$$

which is finite due to Assumption (BP4) and the uniform $L^{1+\alpha+\delta}$-integrability of f_{θ_m} for any $\epsilon > 0$. For the second cross-entropy term, we again bound the integral from two sides as shown below.

$$\begin{aligned}
&\int f_{\theta_m}^{1+\alpha} \mathbf{1}_{A_m^c} \ln(g_{\epsilon,m}) \\
= &\int_{A_m^c \cap \{x : f_{\theta_m}(x) \leq k_m(x)\}} f_{\theta_m}^{1+\alpha} \ln(g_{\epsilon,m}) + \int_{A_m^c \cap \{x : f_{\theta_m}(x) > k_m(x)\}} f_{\theta_m}^{1+\alpha} \ln(g_{\epsilon,m}) \\
\leq &\int_{A_m^c \cap \{x : f_{\theta_m}(x) \leq k_m(x)\}} f_{\theta_m}^{1+\alpha} \ln(k_m) + \int_{A_m^c \cap \{x : f_{\theta_m}(x) > k_m(x)\}} f_{\theta_m}^{1+\alpha} \ln((1-\epsilon) f_{\theta_m}),
\end{aligned}$$

both the integrals are finite due to Assumption (BP4) for all $\epsilon \in [0, 1)$. For the lower bound, consider

$$\begin{aligned}
\int f_{\theta_m}^{1+\alpha} \mathbf{1}_{A_m^c} \ln(g_{\epsilon,m}) &\geq \int_{A_m^c} f_{\theta_m}^{1+\alpha} \ln(\epsilon k_m) \\
&= \int_{A_m^c} f_{\theta_m}^{1+\alpha} \ln(k_m) + \ln(\epsilon) \int_{A_m^c} f_{\theta_m}^{1+\alpha},
\end{aligned}$$

which is finite for all $\epsilon \in (0,1)$. For $\epsilon = 0$, the finiteness of $\int f_{\theta_m}^{1+\alpha} \mathbf{1}_{A_m^c} \ln(g)$ is immediate as $\theta_m \in \Theta$.

Therefore, for $A = 0$, we obtain that

$$S_{(\alpha,\lambda);A=0}(g_{\epsilon,m}, f_{\theta_m}) \asymp \frac{(1-\epsilon)^{1+\alpha}}{1+\alpha} \|g\|_{1+\alpha}^{1+\alpha} - S_{(\alpha,\lambda);A=0}(\epsilon k_m, f_{\theta_m}),$$

as $m \to \infty$.

This completes Step 1 of the proof.

Step 2:

Now, since the true density g belongs to the interior of the model family of densities, there exists a $\theta^g \in \Theta \setminus \partial\Theta$ such that $g = f_{\theta^g}$. In this case, our objective is to find the asymptotic behaviour of $S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta^g})$ and demonstrate that

$$\begin{aligned} S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta^g}) &\asymp \frac{\epsilon^{1+\alpha}}{B} \|k_m\|_{1+\alpha}^{1+\alpha} + r_{(\alpha,\lambda)}(1-\epsilon) \|f_{\theta^g}\|_{1+\alpha}^{1+\alpha} \\ &= \frac{\epsilon^{1+\alpha}}{B} \|k_m\|_{1+\alpha}^{1+\alpha} + r_{(\alpha,\lambda)}(1-\epsilon) \|g\|_{1+\alpha}^{1+\alpha}, \quad (5.5) \end{aligned}$$

where

$$r_{(\alpha,\lambda)}(\epsilon) = q_{(\alpha,\lambda)}(\epsilon) + \epsilon^{1+\alpha}/B. \quad (5.6)$$

In this case, we again consider splitting the integral as in (5.4) but over the sets $B_m = \{x : k_m(x) > \max\{g(x), f_{\theta^g}(x)\}\} = \{x : k_m(x) > g(x)\}$ and B_m^c. As before, due to Assumptions (BP2) and (BP3), we get that the set B_m is asymptotically negligible set under g (or f_{θ^g}) and the set B_m^c is asymptotically negligible set under k_m.

Step 2, $A > 0$:

Since $f_{\theta^g} = g$, the S-divergence for the case $A > 0$ reduces to

$$S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta^g}) = \frac{1}{A} \|g\|_{1+\alpha}^{1+\alpha} - \frac{1+\alpha}{AB} \langle g, g_{\epsilon,m} \rangle_{B,A} + \frac{1}{B} \|g_{\epsilon,m}\|_{1+\alpha}^{1+\alpha}.$$

Similar to the argument before for Step 1, it is easy to see that as $m \to \infty$,

$$\|g_{\epsilon,m}\|_{1+\alpha}^{1+\alpha} \asymp (1-\epsilon)^{1+\alpha} \|g\|_{1+\alpha}^{1+\alpha} - \epsilon^{1+\alpha} \|k_m\|_{1+\alpha}^{1+\alpha}.$$

For the cross-integral term, we split the integral into two sets, on B_m and on B_m^c. Due to Assumptions (BP2) and (BP3), we have that $g\mathbf{1}_{B_m} \to 0$ and $k_m \mathbf{1}_{B_m^c} \to 0$ pointwise as $m \to \infty$. Therefore, $g_{\epsilon,m}\mathbf{1}_{B_m} \to 0$ pointwise as it is dominated by $k_m \mathbf{1}_{B_m}$. On the other hand, $(g_{\epsilon,m}\mathbf{1}_{B_m^c} - (1-\epsilon)g\mathbf{1}_{B_m^c}) \to 0$ as $m \to \infty$. Now, by an application of Lemma 5.A.1 with $f_m = g$, $g_m = g_{\epsilon,m}\mathbf{1}_{B_m}$ and $\tilde{g}_m \equiv 0$, it follows that

$$\left| \int g^B g_{\epsilon,m}^A \mathbf{1}_{B_m} \right| \to 0.$$

On the other hand, again applying Lemma 5.A.1 with $f_m = g$, $g_m = g_{\epsilon,m}\mathbf{1}_{B_m^c}$ and $\tilde{g}_m = (1-\epsilon)g\mathbf{1}_{B_m^c}$ yields that

$$\left| \int g^B g_{\epsilon,m}^A \mathbf{1}_{B_m^c} - \int_{B_m^c} g^{1+\alpha} \right| \to 0.$$

Note that, the uniform integrability of the respective quantities follows directly from the $L^{1+\alpha}$-integrability of g and k_m as mentioned in Assumption (BP4). Therefore, we have

$$\int g^B g_{\epsilon,m}^A \asymp (1-\epsilon)^A \|g\|_{1+\alpha}^{1+\alpha}.$$

Combining all these, the claim given in (5.5) readily follows for the case $A > 0$.

Step 2, $A = 0$:

For the $A = 0$ case with $f_{\theta^g} = g$, the S-divergence between $g_{\epsilon,m}$ and f_{θ^g} reduces to

$$S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta^g}) = \int g^{1+\alpha} \ln(g/g_{\epsilon,m}) - \frac{1}{1+\alpha}\|g\|_{1+\alpha}^{1+\alpha} + \frac{1}{1+\alpha}\|g_{\epsilon,m}\|_{1+\alpha}^{1+\alpha}. \tag{5.7}$$

As before, the last term is asymptotically equivalent to $(\epsilon^{1+\alpha}\|k_m\|_{1+\alpha}^{1+\alpha} + (1-\epsilon)^{1+\alpha}\|g\|_{1+\alpha}^{1+\alpha})/(1+\alpha)$ as $m \to \infty$. For the first term, we split the integral as

$$\int g^{1+\alpha} \ln(g/g_{\epsilon,m}) = \int g^{1+\alpha} \ln(g) - \int g^{1+\alpha} \ln(g_{\epsilon,m}) \mathbf{1}_{B_m} - \int g^{1+\alpha} \ln(g_{\epsilon,m}) \mathbf{1}_{B_m^c}.$$

As before, due to Assumptions (BP2) and (BP3), we have that $g\mathbf{1}_{B_m} \to 0$, $k_m \mathbf{1}_{B_m^c} \to 0$, $g_{\epsilon,m}\mathbf{1}_{B_m} \to 0$ and $(g_{\epsilon,m}\mathbf{1}_{B_m^c} - (1-\epsilon)g\mathbf{1}_{B_m^c}) \to 0$ pointwise as $m \to \infty$. In order to apply Lemma 5.A.2, we just need to show that the corresponding cross-integrals are uniformly integrable.

It is easy to see that the first integral $\int g^{1+\alpha}\ln(g)$ is finite due to the $L^{1+\alpha+\delta}$-integrability of g, for details see the proof of Lemma 5.A.2.

For the second integral, we have

$$\int g^{1+\alpha} \ln(g) \mathbf{1}_{B_m} \le \int g^{1+\alpha} \ln(g_{\epsilon,m}) \mathbf{1}_{B_m} \le \int g^{1+\alpha} \ln(k_m) \mathbf{1}_{B_m}.$$

The lower-bound is finite due to the $L^{1+\alpha+\delta}$-integrability of g, and the upper bound is also uniformly bounded due to Assumption (BP4). Reverse inequality holds for the third term $\int g^{1+\alpha}\ln(g_{\epsilon,m})\mathbf{1}_{B_m^c}$. Therefore, an application of Lemma 5.A.2 yields that

$$\int g^{1+\alpha} \ln(g/g_{\epsilon,m}) \asymp -\int g^{1+\alpha} \ln(1-\epsilon) = -\ln(1-\epsilon)\|g\|_{1+\alpha}^{1+\alpha}.$$

Combining this with (5.7) yields that

$$S_{(\alpha,\lambda);A=0}(g_{\epsilon,m}, f_{\theta^g}) \asymp \left[-\ln(1-\epsilon) - \frac{1}{1+\alpha} + \frac{(1-\epsilon)^{1+\alpha}}{1+\alpha}\right]\|g\|_{1+\alpha}^{1+\alpha} + \frac{\epsilon^{1+\alpha}}{1+\alpha}\|k_m\|_{1+\alpha}^{1+\alpha},$$

as $m \to \infty$, which we intended to show.

Step 3:

Now, we will have a contradiction to our assumption that ϵ is a breakdown point if this constant θ^g lying away from the boundary $\partial\Theta$ produces a value of S-divergence smaller than the asymptotic limit of $S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta_m})$. Then the sequence $\{\theta_m\}$ would not minimize $S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_\theta)$ over $\theta \in \Theta$ for every m.

Therefore, asymptotically there is no breakdown at ϵ if, for sufficiently large m, we have

$$S_{(\alpha,\lambda)}(\epsilon k_m, f_{\theta_m}) + \frac{(1-\epsilon)^{1+\alpha}}{B}\|g\|_{1+\alpha}^{1+\alpha} > \frac{\epsilon^{1+\alpha}}{B}\|k_m\|_{1+\alpha}^{1+\alpha} + r_{(\alpha,\lambda)}(1-\epsilon)\|g\|_{1+\alpha}^{1+\alpha}$$
$$\ge S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_{\theta^g}),$$

which can be rearranged as

$$S_{(\alpha,\lambda)}(\epsilon k_m, f_{\theta_m}) > \frac{\epsilon^{1+\alpha}}{B} \|k_m\|_{1+\alpha}^{1+\alpha} + q_{(\alpha,\lambda)}(1-\epsilon) \|g\|_{1+\alpha}^{1+\alpha}.$$

Clearly, the condition (5.2) ensures that the above inequality is true for all $\epsilon < \widetilde{\epsilon}_{(\alpha,\lambda)}$. This completes the proof.

□

The result in Theorem 5.1 depends heavily on the condition (5.2) to provide the choice of $\widetilde{\epsilon}_{(\alpha,\lambda)}$, which acts as a lower bound of the breakdown point ϵ^*. This condition compares the extremity of the contaminating density with respect to the model family of the densities when the parameter tends to the boundary of the parameter space. They are similar to Assumption (BP3) of Ghosh and Basu (2013) and Ghosh et al. (2017) in spirit, but with a more complicated lower bound. However, this condition is difficult to verify in practice in different parametric setups. Therefore, in the following discussion, we shall show three other sufficient conditions which imply the condition (5.2) and help in deriving the exact value of $\widetilde{\epsilon}_{(\alpha,\lambda)}$ under many general parametric setups. With the knowledge of the model family of densities and the contaminating densities, one can easily check these alternative conditions as we will demonstrate later through some specific examples in Section 5.4.

We start with a lemma that provides a lower bound to the S-divergence itself.

Lemma 5.1. *Let, f and g be two densities such that $\|f\|_{1+\alpha} \geq \|g\|_{1+\alpha}$. Assume $A \geq 0, B \geq 0$ and $\alpha \in [0,1]$ and define*

$$\epsilon^*_{(\alpha,\lambda)} := \begin{cases} [B/(1+\alpha)]^{1/A} & \text{if } A > 0 \\ e^{-1/(1+\alpha)} & \text{if } A = 0 \end{cases}. \tag{5.8}$$

Then, the corresponding S-divergences satisfy $S_{(\alpha,\lambda)}(\epsilon g, f) \geq S_{(\alpha,\lambda)}(\epsilon g, g)$ for any choice of $\epsilon \leq \epsilon^*_{(\alpha,\lambda)}$, provided that these S-divergences on both the sides are well-defined.

Proof. We will illustrate the proof for the cases $A = 0$, $B = 0$ and positive A and B separately.

Case $A = 0$:

In respect to Eq. (1.11), we need to show that

$$\int f^{1+\alpha} \ln(f/\epsilon g) - \frac{1}{1+\alpha} \|f\|_{1+\alpha}^{1+\alpha} \geq \int g^{1+\alpha} \ln(1/\epsilon) - \frac{1}{1+\alpha} \|g\|_{1+\alpha}^{1+\alpha}.$$

Note that, it can be rearranged as

$$\int f^{1+\alpha} \ln(f/g) + \left(\ln(1/\epsilon) - \frac{1}{1+\alpha}\right) (\|f\|_{1+\alpha}^{1+\alpha} - \|g\|_{1+\alpha}^{1+\alpha}) \geq 0.$$

The first term of the left-hand side is bounded below by a scalar multiple of the KL divergence between two densities proportional to $f^{1+\alpha}$ and $g^{1+\alpha}$ respectively, and hence nonnegative. Since, $\epsilon \leq \lim_{A\to 0+}(B/(1+\alpha))^{1/A} = e^{-1/(1+\alpha)}$ and $\|f\|_{1+\alpha}^{1+\alpha} \geq \|g\|_{1+\alpha}^{1+\alpha}$, the second term is also nonnegative.

Case $B = 0$:

When $B = 0$, we have $\epsilon^*_{(\alpha,\lambda)} = 0$. By continuity, since $\lim_{x\to 0} x^{1+\alpha}\log(x) = 0$ for all $\alpha \in (0, 1]$, it follows that $S_{(\alpha,\lambda;B=0)}(0, f) = \|f\|^{1+\alpha}_{1+\alpha}/(1+\alpha)$. Since $\|f\|^{1+\alpha}_{1+\alpha} \geq \|g\|^{1+\alpha}_{1+\alpha}$, it readily follows that

$$S_{(\alpha,\lambda)}(0, f) \geq S_{(\alpha,\lambda)}(0, g).$$

Case $A > 0$ and $B > 0$:

Using the form of S-divergence as in (1.10) for positive A and B, and applying Hölder's inequality,

$$\begin{aligned} S_{(\alpha,\lambda)}(\epsilon g, f) &= \frac{1}{A}\|f\|^{1+\alpha}_{1+\alpha} - \frac{\epsilon^A(1+\alpha)}{AB}\langle f, g\rangle_{B,A} + \frac{\epsilon^{1+\alpha}}{B}\|g\|^{1+\alpha}_{1+\alpha} \\ &\geq \|f\|^{1+\alpha}_{1+\alpha} - \frac{\epsilon^A(1+\alpha)}{AB}\|f\|^B_{1+\alpha}\|g\|^A_{1+\alpha} + \frac{\epsilon^{1+\alpha}}{B}\|g\|^{1+\alpha}_{1+\alpha} \end{aligned}$$

since $A + B = (1+\alpha)$. As $\epsilon \leq [B/(1+\alpha)]^{1/A}$, hence $0 < \epsilon^A(1+\alpha)/B < 1$. Using $\|f\|_{1+\alpha} \geq \|g\|_{1+\alpha}$, we obtain the following two inequalities;

$$\|f\|^{1+\alpha}_{1+\alpha}\left[\frac{1}{A} - \epsilon^A\frac{(1+\alpha)}{AB}\right] \geq \|g\|^{1+\alpha}_{1+\alpha}\left[\frac{1}{A} - \epsilon^A\frac{(1+\alpha)}{AB}\right], \tag{5.9}$$

and

$$\epsilon^A\frac{(1+\alpha)}{AB}\|f\|^{1+\alpha}_{1+\alpha} \geq \epsilon^A\frac{(1+\alpha)}{AB}\|f\|^B_{1+\alpha}\|g\|^A_{1+\alpha}. \tag{5.10}$$

Summing up both the inequalities in (5.9) and (5.10), we obtain

$$\frac{1}{A}\|f\|^{1+\alpha}_{1+\alpha} - \epsilon^A\frac{(1+\alpha)}{AB}\|f\|^B_{1+\alpha}\|g\|^A_{1+\alpha} \geq \frac{1}{A}\|g\|^{1+\alpha}_{1+\alpha} - \epsilon^A\frac{(1+\alpha)}{AB}\|g\|^{1+\alpha}_{1+\alpha}. \tag{5.11}$$

Adding $\epsilon^{\alpha+1}\|g\|^{1+\alpha}_{1+\alpha}/B$ to both sides of the above inequality (5.11) yields the result. □

A corollary of the above lemma yields a similar bound in the case of the DPD which immediately follows by taking $A = 1$ and $B = \alpha$.

Corollary 5.1. *Let, f and g be two densities, and let $d_\alpha(g, f)$ denote the DPD between these two densities for $\alpha \geq 0$. If $\epsilon \leq \alpha/(1+\alpha)$ and $\|f\|_{1+\alpha} \geq \|g\|_{1+\alpha}$, then $d_\alpha(\epsilon g, f) \geq d_\alpha(\epsilon g, g)$, provided the DPDs on both sides are well-defined.*

Now, we turn to describe our first alternative condition which is sufficient (but not necessary) for the generic condition (5.2) provided in Theorem 5.1.

Corollary 5.2. *Assume A and B are both nonnegative and the Assumptions (BP1)-(BP4) are true. Also, assume that the contaminating densities $\{k_m\}$ satisfy the inequality*

$$\|k_m\|_{1+\alpha} \leq \|f_{\theta_m}\|_{1+\alpha} \tag{5.12}$$

for sufficiently large m for any sequence $\{\theta_m\}$ with a limit point on the boundary $\partial\Theta$ of the parameter space. Then if the true density g belongs to the interior of the model family of densities $\mathcal{F}$, then the MSD functional $T_{(\alpha,\lambda)}(G)$ has an asymptotic breakdown point at least $\widetilde{\epsilon}_{(\alpha,\lambda)} = \min\{\epsilon^*_{(\alpha,\lambda)}, 1 - \epsilon^*_{(\alpha,\lambda)}\}$, where $\epsilon^*_{(\alpha,\lambda)}$ is as given in (5.8).

Proof. We shall show that the condition given in (5.12) implies the generic condition(5.2) so that an application of Theorem 5.1 then completes the proof.

If $B = 0$, then $A > 0$ and so $\widetilde{\epsilon}_{(\alpha,\lambda)} = 0$. Since any estimator has an asymptotic breakdown point at least 0, the result holds trivially. So, we shall focus on the case with $B > 0$. Under condition (5.12), an application of Lemma 5.1 yields that for all $\epsilon < \epsilon^*_{(\alpha,\lambda)}$ the inequality

$$S_{(\alpha,\lambda)}(\epsilon k_m, f_{\theta_m}) \geq r_{(\alpha,\lambda)}(\epsilon) \|k_m\|^{1+\alpha}_{1+\alpha},$$

holds for all sufficiently large m, where $r_{(\alpha,\lambda)}(\epsilon)$ is as defined in (5.6). Therefore, the condition (5.2) follows if we can show that the quantity $r_{(\alpha,\lambda)}(\epsilon) \|k_m\|^{1+\alpha}_{1+\alpha}$ is greater than or equal to the right-hand side of inequality (5.2), for all $\epsilon < \widetilde{\epsilon}_{(\alpha,\lambda)}$, where $\widetilde{\epsilon}_{(\alpha,\lambda)} = \min\{\epsilon^*_{(\alpha,\lambda)}, 1 - \epsilon^*_{(\alpha,\lambda)}\}$. We shall show this for the $A > 0$ and $A = 0$ cases separately.

To this end, for $A > 0$, we have

$$\begin{aligned}
& r_{(\alpha,\lambda)}(\epsilon) \|k_m\|^{1+\alpha}_{1+\alpha} > \frac{\epsilon^{1+\alpha}}{B} \|k_m\|^{1+\alpha}_{1+\alpha} + \left[\frac{1}{A} - \frac{(1+\alpha)}{AB}(1-\epsilon)^A\right] \|g\|^{1+\alpha}_{1+\alpha} \\
\iff \quad & \|k_m\|^{1+\alpha}_{1+\alpha} - \|k_m\|^{1+\alpha}_{1+\alpha} \frac{(1+\alpha)}{B}\epsilon^A > \|g\|^{1+\alpha}_{1+\alpha} - \frac{(1+\alpha)}{B}(1-\epsilon)^A \|g\|^{1+\alpha}_{1+\alpha} \\
\iff \quad & (\epsilon^A b_m - (1-\epsilon)^A) < \frac{B}{(1+\alpha)}(b_m - 1) \\
\iff \quad & b_m \left(\epsilon^A - \frac{B}{(1+\alpha)}\right) + \left(\frac{B}{(1+\alpha)} - (1-\epsilon)^A\right) < 0,
\end{aligned}$$

where $b_m = \|k_m\|^{1+\alpha}_{1+\alpha} / \|g\|^{1+\alpha}_{1+\alpha}$. This inequality holds true if $\epsilon < (B/(1+\alpha))^{1/A}$ and $\epsilon < 1 - (B/(1+\alpha))^{1/A}$. Since $\epsilon < \widetilde{\epsilon}_{(\alpha,\lambda)}$, the inequality holds.

For $A = 0$, the proof follows similarly. Since $\epsilon^*_{(\alpha,\lambda)} = e^{-1/(1+\alpha)}$, choosing $\epsilon < \min\{\epsilon^*_{(\alpha,\lambda)}, 1 - \epsilon^*_{(\alpha,\lambda)}\}$ ensures that

$$-\left[\ln(\epsilon) + \frac{1}{1+\alpha}\right] \|k_m\|^{1+\alpha}_{1+\alpha} \geq 0 \geq \left[\ln(1-\epsilon) - \frac{1}{1+\alpha}\right] \|g\|^{1+\alpha}_{1+\alpha}.$$

Now add $\epsilon^{1+\alpha} \|k_m\|^{1+\alpha}_{1+\alpha}/(1+\alpha)$ to both sides of the above inequality yields condition (5.2). □

The analogous result for MDPDE follows from substituting $A = 1$ and $B = \alpha$ in Corollary 5.2.

Corollary 5.3. *Under the same conditions as in Corollary 5.2, the MDPDE functional minimizing $d_\alpha(g, f_\theta)$ has an asymptotic breakdown point at least $\alpha/(1+\alpha)$ for $\alpha \in [0, 1]$, provided that the true density g belongs to the interior of the model family of densities.*

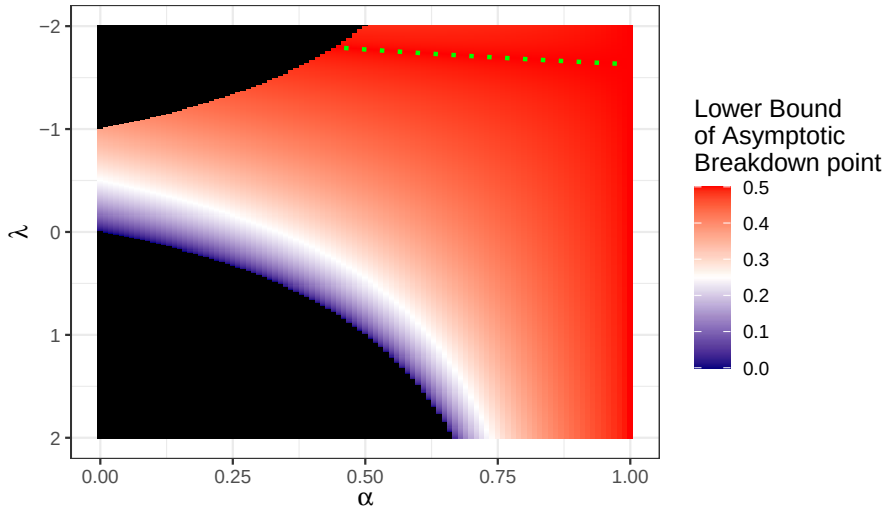
Remark 5.2. *When $A = 0$, Corollary 5.2 assures that the asymptotic breakdown point of the MSD functional is at least $\min\{e^{-1/(1+\alpha)}, 1 - e^{-1/(1+\alpha)}\}$ for all $\alpha \in [0, 1]$. Note that, this bound is increasing in α from 0 to $\alpha^* = 1/\ln(2) - 1 \approx 0.443$, and then decreasing in the interval $[\alpha^*, 1]$. This shows that while increasing the tuning parameter α results in more robust estimators for the case of MDPDE with $A = 1$, this may not be true in general for the MSD functional.*

Remark 5.3. *When $B = 0$, Corollary 5.2 ensures that the asymptotic breakdown point of the MSD functional is at least 0, which contains no meaningful information. However, this bound cannot be improved further. To see this, consider the example of estimating the location parameter in a normal model family of densities $\mathcal{N}(\theta, 1)$ where $\theta \in \mathbb{R}$. Clearly, here $\|f_\theta\|_{1+\alpha} = \|g\|_{1+\alpha}$ for any $\theta \in \mathbb{R}$. In this case, the minimization of the MSD functional is equivalent to the minimization of the Kullback-Leibler (KL) divergence between the densities proportional to $f_\theta^{1+\alpha}$ and $g^{1+\alpha}$. This minimization results in the mean of the true density g for any $\alpha \in [0, 1]$, because of the specific choice of normal densities. It is known that the mean functional is nonrobust with an asymptotic breakdown point of 0.*

It is remarkable that unlike the shrinking breakdown offered by the usual M-estimators for multivariate location and scatter (Maronna, 1976), the MDPDE has an asymptotic breakdown point with a lower bound independent of the dimension of the data, as observed from Corollary 5.3. Thus, the MDPDE is more suited for robust inference of the parameters for arbitrarily high-dimensional or multivariate data. In Figure 5.1, we show the quantity $\widetilde{\epsilon}_{(\alpha,\lambda)}$ as a heatmap for different combinations of α and λ. The choices of α and λ for which either $A < 0$ or $B < 0$ are indicated by the black regions; for these situations, as the assumptions do not hold, our analysis does not guarantee any useful lower bound of the asymptotic breakdown point of the corresponding MSDE. The green dotted line indicates the pairs (α, λ) for which $(B/(1+\alpha))^{1/A} = 1/2$, which represents the highest possible asymptotic breakdown point 1/2 when $\alpha < 1$. It follows that, as α increases, the asymptotic breakdown point also increases considerably for all λ approximately larger than (-1.63). However, when $\lambda < (-1.63)$, as α increases from 0 to 1, the asymptotic breakdown point increases till the green line and then decreases. On the other hand, increasing λ has an opposite effect of reducing the breakdown point. This figure is in agreement with Figure 5 of Ghosh et al. (2017) where the authors indicate the region where both A and B are positive, corresponding MSDE has an asymptotic breakdown point of 1/2 for location estimators. Figure 5.1 also depicts that the MSDE is highly robust for a wide range of choices of α and λ; the majority of the choices of α and λ lead to an estimator with an asymptotic breakdown point at least 1/4, for which the corresponding contour is denoted by the solid white line.

When $\alpha = 1$ in the MSD functional, MSDE coincides with the robust minimum L^2 distance estimator irrespective of the value of λ. It has an asymptotic breakdown point of 1/2 which can also be seen from Figure 5.1. Another one of the important divergences among the S-divergence family of divergences is the Hellinger distance ($A = B = 1/2$). Minimum Hellinger Distance Estimator (MHDE) which aims to minimize the Hellinger distance between the empirical density and the model density function has been found to be extremely robust, and is often a popular choice for multivariate location and scatter estimate. Tamura and Boos (1986) and Toma (2008) showed that under some suitable assumptions, the asymptotic breakdown point of MHDE is at least 1/4, independent of the data dimension. This is a special case of Corollary 5.2 where we obtain the same bound under a different set of assumptions.

Figure 5.1: Lower bound $\widetilde{\epsilon}_{(\alpha,\lambda)}$ as in Corollary 5.2 of the asymptotic breakdown point of MSDE for different choices of α and λ. The black regions indicate either $A < 0$ or $B < 0$. The dotted green line indicates the implicit curve $(B/(1+\alpha))^{1/A} = 1/2$.



When the parameter of interest is the location parameter and the support of the model family of densities is the whole $\mathbb{R}^d$, for some positive integer d, Assumption (BP4) and the condition (5.12) are satisfied automatically if the contaminating density also belongs to the same family. Hence, for the location estimation problem, Corollary 5.2 implies that the asymptotic breakdown point of the MSDE functional is at least $\min\{(B/(1+\alpha))^{1/A}, 1-(B/(1+\alpha))^{1/A}\}$ for $A > 0$, and $\min\{e^{-1/(1+\alpha)}, 1-e^{-1/(1+\alpha)}\}$ for $A = 0$, regardless of the dimension d.

However, when $A > 0$, a better lower bound can be obtained by directly applying Theorem 5.1. Note that since the model family is a location family of densities, the quantity $\|f_\theta\|_{1+\alpha}$ is independent of the choice of θ, the location parameter. Also, assume that the contaminating densities k_m belong to the same location family. Then, for $A, B > 0$, applying Hölder's inequality to the term $\int f_{\theta_m}^B(x) k_m^A(x) dx$ appearing in the S-divergence $S_{(\alpha,\lambda)}(\epsilon k_m, f_{\theta_m})$, a sufficient condition for the inequality in (5.2) turns out to be $\epsilon^A - (1-\epsilon)^A < 0$. This yields, $\widetilde{\epsilon}_{(\alpha,\lambda)} = 1/2$. Therefore, the asymptotic breakdown point for MSDE when both $A > 0$ and $B > 0$ is $1/2$ for the location parameter when both the model family and the contaminating distributions are from the same location family of distributions. The same result is also obtained independently by Ghosh et al. (2017) via investigating this special case only. The breakdown results obtained for the MDPDE for location parameter in Basu et al. (1998) and Ghosh and Basu (2013), and for MHDE in Toma (2007) are all special cases of this situation.

The condition (5.12) considers a dynamic bound to the sequence of contaminating densities via the model family of densities. On the other hand, because of Assumption (BP4), we know that there must exists some constant $C > 0$ such that $\sup_m \|k_m\|_{1+\alpha} < C$. In view of this, we can have another sufficient condition when the previous sufficient condition (5.12) is not satisfied.

Corollary 5.4. *Assume that $A \geq 0, B > 0$, the Assumptions (BP1)-(BP4) hold true and the model family of densities contain the true density g in its interior. Additionally, the contaminating densities $\{k_m\}$ satisfy*

$$\limsup_{m\to\infty} \|k_m\|_{1+\alpha}^{1+\alpha} \leq C \tag{5.13}$$

for some deterministic constant $C \geq 0$. Then, the MSD functional $T_{(\alpha,\lambda)}(G)$ has an asymptotic breakdown point at least $\min\{\epsilon'_{(\alpha,\lambda)}, 1/2\}$ where $\epsilon'_{(\alpha,\lambda)} \in (0,1)$ is the unique solution to the equation

$$\frac{C}{B}x^{1+\alpha} + q_{(\alpha,\lambda)}(1-x)\,\|g\|_{1+\alpha}^{1+\alpha} = 0.$$

Proof. We shall show that the specific choice of $\epsilon'_{(\alpha,\lambda)}$ along with the condition (5.13) implies the generic condition (5.2). Note that, the quantity on the left-hand side of the inequality in (5.2) is nonnegative, as they are valid S-divergences. Therefore, it is sufficient to show that the right-hand side of the inequality is negative.

Consider the case $A > 0$ first. Because of the condition (5.13), it is enough to ensure that

$$h(\epsilon) = \frac{C}{B}\epsilon^{1+\alpha} + \left(\frac{1}{A} - \frac{1+\alpha}{AB}(1-\epsilon)^A\right)\|g\|_{1+\alpha}^{1+\alpha} < 0,$$

for all $\epsilon < \epsilon'_{(\alpha,\lambda)}$. Note that, $h(0) = -\|g\|_{1+\alpha}^{1+\alpha}/B < 0$ and $h(1) = C/B + \|g\|_{1+\alpha}^{1+\alpha}/A > 0$ as $C \geq 0$. Also, $h(\cdot)$ is continuous and strictly increasing in the interval $(0,1)$. Therefore, by Intermediate Value Theorem, there must exist a unique root $\epsilon'_{(\alpha,\lambda)} \in (0,1)$ such that $h(\epsilon) < 0$ for all $\epsilon < \epsilon'_{(\alpha,\lambda)}$, as we intended.

For the case $A = 0$, again it is enough to ensure that

$$h_0(\epsilon) = \frac{C}{B}\epsilon^{1+\alpha} - \left(\frac{1}{1+\alpha} + \ln(1-\epsilon)\right)\|g\|_{1+\alpha}^{1+\alpha} < 0.$$

Similar to the argument presented before, we have $h_0(0) = -\|g\|_{1+\alpha}^{1+\alpha}/(1+\alpha) < 0$ and $\lim_{\epsilon\to(1-)} h_0(\epsilon) = \infty > 0$. By using the continuity and strictly increasing nature of $h_0(\epsilon)$ and an application of the Intermediate Value Theorem, the same conclusion follows, completing the proof. □

Since the conditions (5.12) and (5.13) cover all uniformly integrable contaminating densities, such conditions may not provide a reasonably sharp lower bound to the asymptotic breakdown points. Also, Corollary 5.4, which is dependent on the condition (5.13), requires the knowledge of $\|g\|_{1+\alpha}$ which may be unknown. However, in view of the asymptotic singularities between the model family, the true density and the contaminating densities as asserted by Assumptions (BP2) and (BP3), one might hope for some orthogonality to be present between the model family densities f_{θ_m} and the contaminating densities k_m, i.e., $\int f_{\theta_m}^B k_m^A$ to tend to 0 at a faster rate than $\|f_{\theta_m}\|_{1+\alpha}^{1+\alpha}$. This is formulated in the following corollary, which provides yet another sufficient condition to obtain an explicit lower bound of the asymptotic breakdown point of the MSD functional.

Corollary 5.5. *Assume that $A, B \geq 0$, the assumptions (BP1)-(BP4) are true and the model family of densities contain the true density g in its interior. Additionally, the contaminating densities $\{k_m\}$ satisfy*

$$L = \liminf_{m\to\infty, \theta_m \to \theta_\infty} L_m > 0 \tag{5.14}$$

where

$$L_m := \begin{cases} \dfrac{\|f_{\theta_m}\|_{1+\alpha}^{1+\alpha}}{\langle f_{\theta_m}, k_m\rangle_{B,A}} & \text{if } A > 0, \\ \dfrac{\int f_{\theta_m}^{1+\alpha} \ln(f_{\theta_m}/k_m)}{\|f_{\theta_m}\|_{1+\alpha}^{1+\alpha}} & \text{if } A = 0, \end{cases} \tag{5.15}$$

for any sequence of $\{\theta_m\}$ with a limit point in the boundary of parameter space $\partial\Theta$. Then, the MSD functional $T_{(\alpha,\lambda)}(G)$ has an asymptotic breakdown point at least

$$\epsilon^*_{(\alpha,\lambda)} = \begin{cases} \min\left\{\left(\frac{BL}{1+\alpha}\right)^{1/A}, 1 - \left(\frac{B}{1+\alpha}\right)^{1/A}, \frac{1}{2}\right\} & \text{if } A > 0, \\ \min\left\{e^{-1/(1+\alpha)+L}, 1 - e^{-1/(1+\alpha)}, \frac{1}{2}\right\} & \text{if } A = 0. \end{cases}$$

Proof. The scenario with $B = 0$ follows trivially since every estimator has an asymptotic breakdown point of at least 0. Therefore, we restrict our attention to the case where $B > 0$.

For $A = 0$, the condition (5.2) can be rewritten as

$$\int f_{\theta_m}^{1+\alpha} \ln(f_{\theta_m}/\epsilon k_m) - \frac{1}{1+\alpha}\|f_{\theta_m}\|_{1+\alpha}^{1+\alpha} > -\left[\frac{1}{1+\alpha} + \ln(1-\epsilon)\right]\|g\|_{1+\alpha}^{1+\alpha}.$$

Now if $\epsilon < 1 - e^{-1/(1+\alpha)}$, we have $\ln(1-\epsilon) + 1/(1+\alpha) > 0$ implying that the right-hand side of the above inequality is some negative number. Therefore, it is enough to ensure that the left-hand side of the above is a positive quantity.

This means, for sufficiently large m, we require to show

$$\int f_{\theta_m}^{1+\alpha} \ln(f_{\theta_m}/k_m) - \|f_{\theta_m}\|_{1+\alpha}^{1+\alpha}\left[\ln(\epsilon) + \frac{1}{1+\alpha}\right] > 0.$$

If we pick ϵ to be lesser than $e^{-1/(1+\alpha)+L}$, then the above inequality follows from the definition of L in (5.15). Now, an application of Theorem 5.1 completes the proof for the $A = 0$ case.

When both $A > 0$ and $B > 0$, it is enough to show that the condition (5.2) of Theorem 5.1 holds in this case. Note that, it can be rewritten as

$$\|f_{\theta_m}\|_{1+\alpha}^{1+\alpha} - \frac{1+\alpha}{B}\epsilon^A \langle f_{\theta_m}, k_m\rangle_{B,A} > \left[1 - \frac{1+\alpha}{B}(1-\epsilon)^A\right]\|g\|_{1+\alpha}^{1+\alpha}, \tag{5.16}$$

for all sufficiently large m and for all $\epsilon < \widetilde{\epsilon}_{(\alpha,\lambda)}$, where $\widetilde{\epsilon}_{(\alpha,\lambda)} \leq 1/2$ is a predetermined quantity. Since $\|g\|_{1+\alpha}$ is the integral corresponding to the true density g which is unknown, it can be arbitrarily large or arbitrarily small. For any $\epsilon < 1 - (B/(1+\alpha))^{1/A}$, we have the right-hand side of (5.16) as some negative real number. Hence, the only way

for the inequality (5.16) to hold true is to ensure that the left-hand side is nonnegative for sufficiently large m. This is possible if we choose $\epsilon < (BL/(1+\alpha))^{1/A}$. Therefore, choosing ϵ less than the minimum of $(BL/(1+\alpha))^{1/A}$ and $1-(B/(1+\alpha))^{1/A}$ ensures that the inequality (5.16) holds true. Finally, we apply Theorem 5.1. □

When $A > 0, B > 0$, the condition (5.12) is a stronger variant of (5.14), i.e., when (5.12) holds, the quantity L in (5.14) is at least 1 by a direct application of Hölder's inequality. An application of this Corollary 5.5 is demonstrated in Section 5.4.6 through an example of estimating the rate parameter for exponential distribution contaminated by a normal family of densities.

5.3.2 Asymptotic Breakdown point of minimum S-divergence estimator

So far, we have investigated the breakdown point of the MSD functional $T_{(\alpha,\lambda)}(G)$ assuming the knowledge of the true distribution G. However, in practice, the true distribution G and corresponding density g is not known, and hence $T_{(\alpha,\lambda)}(G)$ must be replaced with its empirical version $T_{(\alpha,\lambda)}(G_n)$ that minimizes $S_{(\alpha,\lambda)}(\widehat{g}_n, f_\theta)$ for a suitable density estimate $\widehat{g}_n$ based on sample observations $X_1, \ldots, X_n$. For this, we shall require a few more assumptions as follows:

(BPF1) With probability tending to 1, the density estimate $\widehat{g}_n(x)$ converges to $g(x)$ in L^p norm, i.e., $\int |\widehat{g}_n - g|^p \to 0$ as $n \to \infty$, where $p = (1+\alpha)$ when $A > 0$ and $p = (1+\alpha+\delta)$ for some $\delta > 0$ when $A = 0$.

(BPF2) The minimizer $T_{(\alpha,\lambda)}(G)$ is a well-separated minimizer of the S-divergence $S_{(\alpha,\lambda)}(g, f_\theta)$, i.e., for every $\delta > 0$,

$$S_{(\alpha,\lambda)}(g, f_{T_{(\alpha,\lambda)}(G)}) < \inf_{\|\theta - T_{(\alpha,\lambda)}(G)\| > \delta} S_{(\alpha,\lambda)}(g, f_\theta).$$

(BPF3) Either $A > 0$, or for all $\eta > 0$ there exists a compact set $K_\eta \subset \Theta$ such that for all sufficiently large n and m and any $\epsilon \in (0, 1/2)$,

$$\sup_{\theta \notin K_\eta} \left| \int f_\theta^{1+\alpha} \ln\left(\frac{(1-\epsilon)\widehat{g}_n}{\epsilon k_m} + 1\right)\right| < \eta, \ \sup_{\theta \notin K_\eta} \left| \int f_\theta^{1+\alpha} \ln\left(\frac{(1-\epsilon) g}{\epsilon k_m} + 1\right)\right| < \eta,$$

and,

$$\sup_{\theta \notin K_\eta} \left| \int f_\theta^{1+\alpha} \ln\left(\widehat{g}_n\right)\right| < \eta, \ \sup_{\theta \notin K_\eta} \left| \int f_\theta^{1+\alpha} \ln\left(g\right)\right| < \eta.$$

Under these assumptions, we establish the following theorem that ensures that the asymptotic breakdown point of the sequence of estimators $T_{(\alpha,\lambda)}(G_n)$ is greater than or equal to the asymptotic breakdown point of the population functional $T_{(\alpha,\lambda)}(G)$.

Theorem 5.2. *Assume that $A \geq 0$ and $B > 0$, Assumptions (BPF1)-(BPF3) hold along with Assumption (BP4). Let $\epsilon_1^* \in [0, 1/2)$ be the asymptotic breakdown point of the MSD functional $T_{(\alpha,\lambda)}(G)$ as defined in (1.7) and $\epsilon_2^* \in [0, 1/2)$ be the asymptotic breakdown point of the corresponding sequence of MSDE $T_{(\alpha,\lambda)}(G_n)$ as defined in (1.6). Then, $\epsilon_2^* \geq \epsilon_1^*$.*

Before we discuss the proof of this theorem, let us spend some time understanding the nature of the Assumptions (BPF1)-(BPF3). The first assumption (BPF1) ensures that the choice of the density estimate is reasonable and sufficiently close to the true density. One choice of $\widehat{g}_n$ is a kernel density estimate based on the sample observations $X_1, X_2, \ldots, X_n$ given by

$$\widehat{g}_n(x) = \frac{1}{nb_n} \sum_{j=1}^{n} w\left(\frac{x - X_j}{b_n}\right),$$

where b_n is a bandwidth sequence and $w(\cdot)$ is a properly chosen weight function called kernel function (Bickel and Rosenblatt, 1973). The basic premise under which asymptotic properties of g_n are derived requires $b_n \to 0$, $nb_n \to \infty$ as $n \to \infty$. Under some reasonable assumptions on the density function g, if $b_n = o(n^{-2/9})$, $n^{-1/4}(\log n)^{1/2}(\log\log n)^{1/4} = o(b_n)$ as $n \to \infty$, Bickel and Rosenblatt (1973) showed that

$$b_n^{-1/2}\left[nb_n \int [\widehat{g}_n(x) - g(x)]^2 a(x)dx - \int g(x)a(x)dx \int w^2(z)dz\right],$$

is asymptotically normally distributed with mean 0 and a constant variance σ_g^2 depending on the choice of w, g and a. If we choose $a(x) = 1$, then it follows that

$$\begin{aligned}
& b_n^{-1/2}\left[nb_n \int [\widehat{g}_n(x) - g(x)]^2 dx - \int w^2(z)dz\right] \xrightarrow{d} \mathcal{N}(0, \sigma_g^2), \\
\Rightarrow \quad & nb_n \int [\widehat{g}_n(x) - g(x)]^2 dx - \int w^2(z)dz \xrightarrow{d} \mathcal{N}(0, b_n\sigma_g^2), \\
\Rightarrow \quad & nb_n \int [\widehat{g}_n(x) - g(x)]^2 dx \xrightarrow{d} \mathcal{N}\left(\int w^2(z)dz, b_n\sigma_g^2\right), \\
\Rightarrow \quad & \int [\widehat{g}_n(x) - g(x)]^2 dx \xrightarrow{d} \mathcal{N}\left(\frac{1}{nb_n}\int w^2(z)dz, \frac{1}{n^2 b_n}\sigma_g^2\right), \\
\Rightarrow \quad & \int [\widehat{g}_n(x) - g(x)]^2 \xrightarrow{P} 0,
\end{aligned}$$

where $\xrightarrow{d}$ denotes the convergence in distribution, $\xrightarrow{P}$ denotes convergence in probability as $n \to \infty$. This establishes L^2-type convergence of $\widehat{g}_n$ to g, which, in turn, implies Assumption (BPF1) for all nonnegative A and B except when $A = 0$ and $B = 2$.

The second condition (BPF2) about the well-separated minimizer is quite common in the literature of M-estimation; for further details, see Vaart (1998). The third assumption (BPF3) holds trivially for $A > 0$. For the $A = 0$ case, it is a stronger variant of the asymptotic singularity condition present in Assumptions (BP2)-(BP3). To see this, note that by choosing sufficiently large K_η, we can ensure that $\theta \notin K_\eta$ implies θ is close to the boundary $\partial\Theta$, and as a result, when $f_\theta > 0$, g is asymptotically negligible. Hence, one would expect

$$g_{\epsilon,m}\mathbf{1}_{\{f_\theta>0\}} = (1-\epsilon)g\mathbf{1}_{\{f_\theta>0\}} + \epsilon k_m \mathbf{1}_{\{f_\theta>0\}} \approx \epsilon k_m \mathbf{1}_{\{f_\theta>0\}},$$

resulting in

$$\begin{aligned}
\int f_\theta^{1+\alpha} \ln\left(\frac{(1-\epsilon)g}{\epsilon k_m} + 1\right) = \int f_\theta^{1+\alpha}\mathbf{1}_{\{f_\theta>0\}} \ln\left(\frac{g_{\epsilon,m}}{\epsilon k_m} + 1\right) \\
\approx \int f_\theta^{1+\alpha}\mathbf{1}_{\{f_\theta>0\}} \ln(1) = 0.
\end{aligned}$$

Assumption (BPF3) entails that this approximation holds uniformly.

Proof. (of Theorem 5.2)

We will prove it by contradiction. Assume that $\epsilon_2^* < \epsilon_1^*$. Then, pick some $\epsilon \in (\epsilon_2^*, \epsilon_1^*)$. Clearly, for this particular choice of ϵ, the sequence of estimators $T_{(\alpha,\lambda)}(G_n)$ breaks down at ϵ-level of contamination. Hence, by the definition of the breakdown point as in (1.6), we have a $\theta_\infty \in \partial\Theta$ and a contaminating sequence of distributions $\{K_m\}_{m=1}^\infty$ with densities $\{k_m\}_{m=1}^\infty$ such that $\|T_{(\alpha,\lambda)}(G_n, \epsilon, m) - \theta_\infty\| \to 0$ as $n \to \infty$ for any fixed m with nonzero probability. Here, we use the notation $G_{n,\epsilon,m} = (1-\epsilon)G_n + \epsilon K_m$. Since $\epsilon < \epsilon_1^*$, we will have a contradiction if we show that $\|T_{(\alpha,\lambda)}(G_{\epsilon,m}) - \theta_\infty\| \to 0$ as $m \to \infty$. We will show this in three steps.

Step 1:

Consider the difference between the S-divergences evaluated at the densities $g_{\epsilon,m} = (1-\epsilon)g + \epsilon k_m$ and $g_{n,\epsilon,m} = (1-\epsilon)\widehat{g}_n + \epsilon k_m$, i.e.,

$$\left|S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_\theta) - S_{(\alpha,\lambda)}(g_{n,\epsilon,m}, f_\theta)\right|.$$

In this step, we shall show that for any fixed m, the above difference goes to 0 as $n \to \infty$ uniformly over $\theta \in \Theta$. We show this separately for $A > 0$ and $A = 0$ cases.

Step 1 with $A > 0$:

First, let us consider the case with both $A > 0$ and $B > 0$. Fix any m and $\epsilon \in [0, 1/2)$. Then,

$$\begin{aligned}
&\left|S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_\theta) - S_{(\alpha,\lambda)}(g_{n,\epsilon,m}, f_\theta)\right| \\
\leq &\frac{1}{B}\left|\int (g_{n,\epsilon,m}^{1+\alpha} - g_{\epsilon,m}^{1+\alpha})\right| + \frac{1+\alpha}{AB}\left|\int f_\theta^B (g_{n,\epsilon,m}^A - g_{\epsilon,m}^A)\right| \\
\leq &\frac{1}{B}\left|\int (g_{n,\epsilon,m}^{1+\alpha} - g_{\epsilon,m}^{1+\alpha})\right| + \frac{1+\alpha}{AB}\|f_\theta\|_{1+\alpha}^B \left(\int \delta_{n,\epsilon,m}^{1+\alpha}\right)^{\frac{A}{1+\alpha}},
\end{aligned}$$

where $\delta_{n,\epsilon,m} = |g_{n,\epsilon,m}^A - g_{\epsilon,m}^A|^{1/A}$. Now, since $\int |\widehat{g}_n - g|^{1+\alpha} \to 0$ as $n \to \infty$, we have the pointwise convergence $\widehat{g}_{n_j}(x) \to g(x)$ almost surely for all x as $n_j \to \infty$, for some subsequence $\{n_j\}_{j=1}^\infty$; see Theorem 3.12 of Rudin (1987). This implies, for any fixed m and any fixed $\epsilon \in [0, 1/2)$, we have pointwise convergence $g_{n_j,\epsilon,m} \to g_{\epsilon,m}$ as $n_j \to \infty$ almost surely for all x.

Now, note that

$$\begin{aligned}
\sup_n \int g_{n,\epsilon,m}^{1+\alpha} = \sup_n \int ((1-\epsilon)\widehat{g}_n + \epsilon k_m)^{1+\alpha} &\leq \sup_n \int \widehat{g}_n^{1+\alpha} + \int k_m^{1+\alpha} \\
&\leq \sup_n \int (|\widehat{g}_n - g| + g)^{1+\alpha} + \int k_m^{1+\alpha} \\
&\leq 2^{1+\alpha}\left(\sup_n \int |\widehat{g}_n - g|^{1+\alpha} + \int g^{1+\alpha}\right) + \int k_m^{1+\alpha} < \infty,
\end{aligned}$$

where we use the $L^{1+\alpha}$ convergence of $\widehat{g}_n$ to g along with Assumption (BP4). Therefore, by applying Lemma 5.A.1 with $g_n = g_{n,\epsilon,m}, \widetilde{g}_n = g_{\epsilon,m}$ for fixed ϵ and m, we note that the first term of the bound goes to 0 for the subsequence $\{n_j\}$ as $n_j \to \infty$. For the second term,

due to Assumption (BP4), we have $\|f_\theta\|_{1+\alpha} < M$ for some sufficiently large $M > 0$ and $\int \delta_{n_j,\epsilon,m}^{1+\alpha}$ converges to 0 as $n_j \to \infty$ (see the proof of Lemma 5.A.1 for details). Therefore, by taking supremum over $\theta \in \Theta$, we obtain that

$$\sup_{\theta\in\Theta} |S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_\theta) - S_{(\alpha,\lambda)}(g_{n_j,\epsilon,m}, f_\theta)| \to 0,$$

as $n_j \to \infty$ for any $A, B > 0$ and for any fixed m and $\epsilon \in [0, 1/2)$.

Step 1 with $A = 0$:

For $A = 0$, we have

$$\begin{aligned}
& \sup_{\theta\in\Theta} \left|S_{(\alpha,\lambda)}(g_{\epsilon,m}, f_\theta) - S_{(\alpha,\lambda)}(g_{n,\epsilon,m}, f_\theta)\right| \\
\leq\ & \sup_{\theta\in\Theta} \left|\int f_\theta^{1+\alpha} \ln\left(\frac{g_{n,\epsilon,m}}{g_{\epsilon,m}}\right)\right| + \frac{1}{1+\alpha}\left|\int (g_{n,\epsilon,m}^{1+\alpha} - g_{\epsilon,m}^{1+\alpha})\right| \\
\leq\ & \sup_{\theta\in\Theta} \left|\int f_\theta^{1+\alpha} \ln(g_{n,\epsilon,m}) - \int f_\theta^{1+\alpha} \ln(g_{\epsilon,m})\right| + \frac{1}{1+\alpha}\left|\int (g_{n,\epsilon,m}^{1+\alpha} - g_{\epsilon,m}^{1+\alpha})\right|.
\end{aligned}$$

Since $g_{n_j,\epsilon,m}$ converges to $g_{\epsilon,m}$ pointwise as $n_j \to \infty$ for any fixed m and $\epsilon \in (0, 1/2)$, and both $g_{n,\epsilon,m}$ and $g_{\epsilon,m}$ are $L^{1+\alpha+\delta}$-integrable (as shown in the previous case), by an application of Lemma 5.A.1, it follows that the second term goes to 0 as $n_j \to \infty$.

For the first term, note that for a fixed $\theta \in \Theta$, we can apply Lemma 5.A.2 along with $L^{1+\alpha+\delta}$-integrability of g_n and g. In particular, choosing $g_n = \tilde{g}_n = f_\theta$, and $f_j = g_{n_j,\epsilon,m}, \tilde{f}_n = g_{\epsilon,m}$ in Lemma 5.A.2 yields that

$$\left|\int f_\theta^{1+\alpha} \ln(g_{n_j,\epsilon,m}) - \int f_\theta^{1+\alpha} \ln(g_{\epsilon,m})\right| \to 0,$$

as $n_j \to \infty$, for any fixed $\theta \in \Theta$, fixed m and fixed $\epsilon \in [0, 1/2)$.

Now, pick any $\eta > 0$. By Assumption (BPF3), there exists a compact set K_η such that for sufficiently large m and n, we have for any $\epsilon \in (0, 1/2)$,

$$\sup_{\theta\notin K_\eta} \left|\int f_\theta^{1+\alpha} \ln\left(\frac{(1-\epsilon)\widehat{g}_n}{\epsilon k_m} + 1\right)\right| < \eta, \text{ and, } \sup_{\theta\notin K_\eta} \left|\int f_\theta^{1+\alpha} \ln\left(\frac{(1-\epsilon)g}{\epsilon k_m} + 1\right)\right| < \eta.$$

Therefore,

$$\begin{aligned}
& \sup_{\theta\notin K_\eta} \left|\int f_\theta^{1+\alpha} \ln(g_{n,\epsilon,m}) - \int f_\theta^{1+\alpha} \ln(g_{\epsilon,m})\right| \\
=\ & \sup_{\theta\notin K_\eta} \left|\int f_\theta^{1+\alpha} \ln(g_{n,\epsilon,m}) - \int f_\theta^{1+\alpha} \ln(\epsilon k_m) - \int f_\theta^{1+\alpha} \ln(g_{\epsilon,m}) + \int f_\theta^{1+\alpha} \ln(\epsilon k_m)\right| \\
\leq\ & \sup_{\theta\notin K_\eta} \left|\int f_\theta^{1+\alpha} \ln(g_{n,\epsilon,m}) - \int f_\theta^{1+\alpha} \ln(\epsilon k_m)\right| + \sup_{\theta\notin K_\eta} \left|\int f_\theta^{1+\alpha} \ln(g_{\epsilon,m}) - \int f_\theta^{1+\alpha} \ln(\epsilon k_m)\right| \\
=\ & \sup_{\theta\notin K_\eta} \left|\int f_\theta^{1+\alpha} \ln\left(\frac{(1-\epsilon)\widehat{g}_n + \epsilon k_m}{\epsilon k_m}\right)\right| + \sup_{\theta\notin K_\eta} \left|\int f_\theta^{1+\alpha} \ln\left(\frac{(1-\epsilon)g + \epsilon k_m}{\epsilon k_m}\right)\right| \\
\leq\ & 2\eta,
\end{aligned}$$

for all $\epsilon \in (0, 1/2)$. For $\epsilon = 0$, we have

$$\sup_{\theta\notin K_\eta} \left|\int f_\theta^{1+\alpha} \ln(g_{n,\epsilon,m}) - \int f_\theta^{1+\alpha} \ln(g_{\epsilon,m})\right| = \sup_{\theta\notin K_\eta} \left|\int f_\theta^{1+\alpha} \ln(\widehat{g}_n) - \int f_\theta^{1+\alpha} \ln(g)\right| < 2\eta$$

On the other hand, by the compactness of K_η and the pointwise convergence established before, it follows that

$$\sup_{\theta\in K_\eta}\left|\int f_\theta^{1+\alpha}\ln(g_{n_j,\epsilon,m})-\int f_\theta^{1+\alpha}\ln(g_{\epsilon,m})\right|<\eta<2\eta,$$

for all sufficiently large subsequence indices n_j. As $\eta>0$ was arbitrary, we get that for all sufficiently large m and fixed $\epsilon\in[0,1/2)$,

$$\sup_{\theta\in\Theta}\left|\int f_\theta^{1+\alpha}\ln(g_{n_j,\epsilon,m})-\int f_\theta^{1+\alpha}\ln(g_{\epsilon,m})\right|\to 0,$$

as $n_j\to\infty$ for any fixed m and fixed $\epsilon\in[0,1/2)$. This completes Step 1 for the $A=0$ case.

Step 2:

Since $T_{(\alpha,\lambda)}(G_{\epsilon,m})$ minimizes the function $\theta\to S_{(\alpha,\lambda)}(g_{\epsilon,m},f_\theta)$, we have

$$\begin{aligned}
& S_{(\alpha,\lambda)}(g_{\epsilon,m},f_{T_{(\alpha,\lambda)}(G_{n,\epsilon,m})})-S_{(\alpha,\lambda)}(g_{\epsilon,m},f_{T_{(\alpha,\lambda)}(G_{\epsilon,m})})\\
\leq\ & S_{(\alpha,\lambda)}(g_{\epsilon,m},f_{T_{(\alpha,\lambda)}(G_{n,\epsilon,m})})-S_{(\alpha,\lambda)}(g_{\epsilon,m},f_{T_{(\alpha,\lambda)}(G_{\epsilon,m})})\\
& \quad +S_{(\alpha,\lambda)}(g_{n,\epsilon,m},f_{T_{(\alpha,\lambda)}(G_{\epsilon,m})})-S_{(\alpha,\lambda)}(g_{n,\epsilon,m},f_{T_{(\alpha,\lambda)}(G_{n,\epsilon,m})})\\
\leq\ & 2\sup_{\theta\in\Theta}\left|S_{(\alpha,\lambda)}(g_{n,\epsilon,m},\theta)-S_{(\alpha,\lambda)}(g_{\epsilon,m},\theta)\right|,
\end{aligned}$$

where the first inequality follows from the definition of $T_{(\alpha,\lambda)}(G_{n,\epsilon,m})$. Now, the first step assures that this quantity converges to 0 for the specific subsequence $\{n_j\}_{j=1}^\infty$ as $n_j\to\infty$, for any fixed m and any fixed $\epsilon\in[0,1/2)$. Combining it with Assumption (BPF2) of a well-separated minimizer, it then follows that $\left\|T_{(\alpha,\lambda)}(G_{n_j,\epsilon,m})-T_{(\alpha,\lambda)}(G_{\epsilon,m})\right\|\to 0$ as $n_j\to\infty$, for any fixed choice of m and any $\epsilon\in[0,1/2)$, with probability tending to 1.

Step 3:

Now, we pick any $\delta>0$. By the hypothesis (of the contradiction) at the start of the proof, there exists M_1 such that for any $m>M_1$ and $n>n_{1m}$, we have

$$\left\|T_{(\alpha,\lambda)}(G_{n,\epsilon,m})-\theta_\infty\right\|<\delta/2,$$

with some nonzero probability p_0. Now by Step 2 above, we also have that for every m, there exists n_{2m} from the subsequence indices $\{n_j\}_{j=1}^\infty$ such that $n_{2m}>n_{1m}$. Then,

$$\left\|T_{(\alpha,\lambda)}(G_{\epsilon,m})-T_{(\alpha,\lambda)}(G_{n_{2m},\epsilon,m})\right\|<\delta/2,$$

with probability at least $(1-p_0/2)$. Now, due to Bonferroni's inequality, both of these inequalities hold for any $m>M_1$ with probability at least $p_0/2$ which is positive. Hence, it is possible to pick a sample ω of size n_{2m} for which both inequalities are true. Now, we have by the triangle inequality

$$\begin{aligned}
\|T_{(\alpha,\lambda)}(G_{\epsilon,m})-\theta_\infty\|\leq\ & \|T_{(\alpha,\lambda)}(G_{n_{2m},\epsilon,m})(\omega)-\theta_\infty\|\\
& +\|T_{(\alpha,\lambda)}(G_{\epsilon,m})-T_{(\alpha,\lambda)}(G_{n_{2m},\epsilon,m})(\omega)\|\leq\delta,
\end{aligned}$$

for any $m>M_1$, and we use $T_{(\alpha,\lambda)}(G_{n_{2m},\epsilon,m})(\omega)$ to denote the value of the MSDE computed for the sample ω. Since $\delta>0$ is arbitrary, this completes the proof. □

5.4 Examples and Illustrations

In this section, we discuss about several examples where the results present in this chapter are applicable and provide a lower bound to the asymptotic breakdown point of the MSD functional.

5.4.1 Normal Location Model (Univariate)

This example concerns the popular parametric setup regarding the estimation of location in a Gaussian family of distributions with a known variance. Let us assume, without loss of generality, that the known variance is 1. Therefore, we have the true distribution $\mathcal{N}(\theta^g, 1)$ and the model family of distributions is $\mathcal{N}(\theta, 1)$ for $\theta \in \mathbb{R}$. Let us also assume that the contaminating densities $\{k_m\}$ belong to the same model family with location parameters θ_{k_m} such that $|\theta_{k_m}| \to \infty$ as $m \to \infty$. Assumptions (BP1)-(BP3) clearly hold for this choice of contaminating densities. Also note that, both $\|f_{\theta_m}\|_{1+\alpha}^{1+\alpha}$ and $\|k_m\|_{1+\alpha}^{1+\alpha}$ are equal to $(2\pi)^{-\alpha/2}(1+\alpha)^{-1/2}$ for all m, which implies Assumption (BP4). The condition (5.12) also holds trivially in this case. Therefore, an application of Corollary 5.2 for the location estimation case reveals that the MSD functional has an asymptotic breakdown point of $1/2$ when $A > 0$ and $B > 0$, whereas it is at least $\min\{e^{-1/(1+\alpha)}, 1 - e^{-1/(1+\alpha)}\}$ for $A = 0$. When, $B = 0$, the lower bound $\min\{(B/(1+\alpha))^{1/A}, 1-(B/(1+\alpha))^{1/A}\}$ yields 0, which is also tight as indicated in Remark 5.3. Hence, for the MDPDE, the asymptotic breakdown point equals $1/2$ for all $\alpha \in (0, 1]$ and 0 for $\alpha = 0$.

5.4.2 Normal Scale Model (Univariate)

Another popular parametric setup is estimating the scale parameter for a known location parameter in the Gaussian family of distributions. Without the loss of generality, we assume that the known location parameter is 0; otherwise, one can work by subtracting the known location from the random variate under consideration. Let, $\mathcal{N}(0, (\sigma^g)^2)$ be the true distribution, while the model family of distributions is $\mathcal{N}(0, \sigma^2)$ with $\sigma \in (0, \infty)$ and the corresponding densities are denoted by f_σ. Therefore, breakdown occurs if the estimate of the variance parameter σ either "implodes" to 0 or "explodes" to ∞. Let us also assume that the contaminating density k_m belongs to the same Gaussian family of distributions namely $\mathcal{N}(\eta, \sigma_{k_m}^2)$ such that $\sigma_{k_m}^2$ either tends to 0 or tends to ∞ as $m \to \infty$.

It follows that $\|f_\sigma\|_{1+\alpha}^{1+\alpha} = (2\pi)^{-\alpha/2}\sigma^{-\alpha/2}(1+\alpha)^{-1/2}$. Therefore, if $\alpha = 0$, Assumption (BP4) is satisfied with $\|k_m\|_{1+\alpha} = \|f_{\theta_m}\|_{1+\alpha} = 1$. In this case, the condition (5.12) is satisfied irrespective of whether σ_{k_m} "explodes" or "implodes". So, applying Corollary 5.2, we obtain an asymptotic breakdown point of the minimum power divergence functional as at least $(-\lambda)^{1/(1+\lambda)}$ for all $\lambda \in (-1, 0]$ and $e^{-1} \approx 0.368$ for $\lambda = -1$.

If $\alpha > 0$, then $\|f_\sigma\|_{1+\alpha}$ does not remain bounded when σ goes to 0. To avoid the problem near zero, we must restrict our attention to only "explosion"-type breakdown in this case, which is indeed the more important case of scale breakdown as pointed out by Dasiou and Moyssiadis (2001) and Huber and Ronchetti (2011). As mentioned before

in the discussion following Theorem 5.1, we restrict the parameter space to $[\delta_0, \infty)$ for some small $\delta_0 > 0$ and apply the corollaries for only the subset $\{\infty\}$ of the boundary of the parameter space. So, if $\sigma^2_{k_m}$ "explodes" to ∞, then $\|k_m\|_{1+\alpha}$ tends to 0 as $m \to \infty$ and the condition (5.13) is satisfied with $C = 0$. Hence, Corollary 5.4 can be applied, resulting in an asymptotic breakdown point of the MSD functional being at least $\min\{1/2, 1 - (B/(1+\alpha))^{1/A}\}$ for $A > 0$, and $\min\{1/2, 1 - e^{-1/(1+\alpha)}\}$ for $A = 0$.

5.4.3 Normal Location Model and Scale Model (Multivariate)

For the estimation of the location parameter under the multivariate normal setup, the contaminating densities k_m are assumed to be multivariate normal with mean $\boldsymbol{\mu}_{k_m}$ and known dispersion matrix with $\|\boldsymbol{\mu}_{k_m}\| \to \infty$ as $m \to \infty$, where $\|\cdot\|$ denotes the Euclidean L^2 norm. The same conclusion as in Example 5.4.1 follows for this case; the MSD functional achieves the highest possible asymptotic breakdown of 1/2 whenever both A and B are positive and $\alpha > 0$, irrespective of the data dimension p.

For the scatter matrix estimation under the multivariate normal setup, we assume that the contaminating densities k_m are also p-variate normal densities with mean $\boldsymbol{\eta}$ and variance $\boldsymbol{\Sigma}_{k_m}$ such that $\det(\boldsymbol{\Sigma}_{k_m}) \to 0$ or $\det(\boldsymbol{\Sigma}_{k_m}) \to \infty$. It follows that

$$\begin{aligned}
\|k_m\|_{1+\alpha}^{1+\alpha} &= \int (2\pi)^{-(1+\alpha)p/2} \det(\boldsymbol{\Sigma}_{k_m})^{-(1+\alpha)/2} \exp\left[-\frac{1+\alpha}{2}(\boldsymbol{x}-\boldsymbol{\eta})^\intercal \boldsymbol{\Sigma}_{k_m}^{-1}(\boldsymbol{x}-\boldsymbol{\eta})\right] d\boldsymbol{x} \\
&= (2\pi)^{-(1+\alpha)p/2} \det(\boldsymbol{\Sigma}_{k_m})^{-(1+\alpha)/2} \int \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\eta})^\intercal \left(\frac{\boldsymbol{\Sigma}_{k_m}}{(1+\alpha)}\right)^{-1}(\boldsymbol{x}-\boldsymbol{\eta})\right] d\boldsymbol{x} \\
&= (2\pi)^{-\alpha p/2} \det(\boldsymbol{\Sigma}_{k_m})^{-\alpha/2} (1+\alpha)^{-p/2}.
\end{aligned}$$

If $\alpha = 0$, we have $\|k_m\|_{1+\alpha} = \|f_{\theta_m}\|_{1+\alpha} = 1$ for any $p \geq 1$. Similar to Example 5.4.2, for both the "exploding" and "imploding" type of contaminations, the minimum power divergence functional achieves an asymptotic breakdown point of at least $(-\lambda)^{1/(1+\lambda)}$ for all $\lambda \in (-1, 0]$, and e^{-1} for $\lambda = -1$.

If $\alpha > 0$, we again need a restriction to the parameter space as

$$\Theta = \{\boldsymbol{\Sigma} : \boldsymbol{\Sigma} \text{ is positive definite and } \det(\boldsymbol{\Sigma}) \in [\delta_0, \infty)\},$$

for some small but fixed $\delta_0 > 0$, to ensure that Assumption (BP4) is satisfied. Note that, if $\det(\boldsymbol{\Sigma}_{k_m}) \to \infty$, we get $\|k_m\| \to 0$ and hence the condition (5.13) is satisfied. Therefore, an application of Corollary 5.4 on the appropriate subset of $\partial\Theta$ then ensures the asymptotic breakdown point of the MSD functional to be at least $\min\{1/2, 1 - (B/(1+\alpha))^{1/A}\}$ for $A > 0$, and $\min\{1/2, 1 - e^{-1/(1+\alpha)}\}$ for $A = 0$.

5.4.4 Gamma distribution (Scale Family)

In this example, we consider the gamma family of distribution as the model family of distributions, with a fixed shape parameter t and unknown inverse-scale parameter θ. The true distribution is $\text{Gamma}(t, \theta^g)$, and the densities k_m are also assumed to belong to the same family with inverse scale parameter θ_{k_m}. Clearly, contamination happens if

$\theta_{k_m} \to 0$ or $\theta_{k_m} \to \infty$. It turns out that if f_θ is the gamma density function with shape t and inverse-scale θ, then

$$\|f_\theta\|_{1+\alpha}^{1+\alpha} = \theta^\alpha (1+\alpha)^{\alpha-(1+\alpha)t} \Gamma((1+\alpha)t - \alpha)\Gamma(t)^{-(1+\alpha)}.$$

Therefore, if $\alpha = 0$, both the Assumption (BP4) and the condition (5.12) are satisfied. As a result, Corollary 5.2 provided a lower bound to the asymptotic breakdown point of the minimum power divergence functional as $(-\lambda)^{1/(1+\lambda)}$ for all $\lambda \in (-1, 0]$, and e^{-1} for $\lambda = -1$.

When $\alpha > 0$, we need to restrict our attention to the subset $(0, M]$ for some large but fixed $M > 0$ as the parameter space in order to satisfy Assumption (BP4). Considering the subset $\{0\}$ of the boundary $\partial\Theta$, it follows that when $\theta_{k_m} \to 0$, condition (5.13) is satisfied with $C = 0$. Hence, the asymptotic breakdown point of the MSD functional becomes at least $\min\{1/2, 1 - (B/(1+\alpha))^{1/A}\}$ for $A > 0$, and $\min\{1/2, 1 - e^{-1/(1+\alpha)}\}$ for $A = 0$. As a special case, for the case of MDPDE, the asymptotic breakdown point is at least $\alpha/(1+\alpha)$ for all $\alpha \in (0, 1]$.

5.4.5 Gamma distribution (Shape Family)

This example also considers the model family of gamma distributions as in Example 5.4.4, but with the parameter of interest being the shape parameter instead of the rate parameter. Without loss of generality, we fix the rate parameter at 1 and consider the model density $f_t(x) = \Gamma(t)^{-1} x^{t-1} e^{-x}$, for all $x > 0$. In this case,

$$\|f_t\|_{1+\alpha}^{1+\alpha} = (1+\alpha)^{\alpha-(1+\alpha)t} \Gamma(t(1+\alpha) - \alpha)\Gamma(t)^{-(1+\alpha)}.$$

The contamination density k_m also belongs to the same family with shape parameter t_{k_m} with either $t_{k_m} \to \infty$ or $t_{k_m} \to 0$.

Let us first consider the case when $t_{k_m} \to \infty$. By using Stirling's approximation on $\|k_m\|_{1+\alpha}^{1+\alpha}$, we get

$$\begin{aligned}\|k_m\|_{1+\alpha}^{1+\alpha} &\asymp (1+\alpha)^{\alpha-(1+\alpha)t_{k_m}} (2\pi)^{-\alpha/2} \frac{((1+\alpha)(t_{k_m}-1))^{(1+\alpha)t_{k_m}-\alpha-1/2} e^{(1-t_{k_m})(1+\alpha)}}{(t_{k_m}-1)^{(t_{k_m}-1/2)(1+\alpha)} e^{(1-t_{k_m})(1+\alpha)}} \\ &= (2\pi)^{-\alpha/2} (1+\alpha)^{-1/2} (t_{k_m}-1)^{-\alpha/2},\end{aligned}$$

for sufficiently large t_{k_m}. Hence, for any $\alpha > 0$, it now follows as $\|k_m\|_{1+\alpha} \to 0$. This implies that the condition (5.13) is satisfied in this case and the Corollary 5.4 is applicable. Hence, we have a lower bound of the asymptotic breakdown point of MSD functional as $\min\{1/2, 1 - (B/(1+\alpha))^A\}$ for $A > 0$, and $\min\{1/2, 1 - e^{-1/(1+\alpha)}\}$ for $A = 0$. For the case of MDPDE, the lower bound again turns out to be $1/2$ for all $\alpha \in (0, 1]$.

On the other hand, if $t_{k_m} \to 0$, then we use the approximation $\Gamma(t_{k_m}) \sim t_{k_m}^{-1}$. It then follows that

$$\|k_m\|_{1+\alpha}^{1+\alpha} \asymp (1+\alpha)^\alpha \Gamma(-\alpha) t_{k_m}^{(1+\alpha)} \to 0, \quad \text{for } \alpha \in (0, 1).$$

Therefore, the condition (5.13) holds for any $\alpha \in (0, 1)$, and we obtain the same lower bound as in the first case. For $\alpha = 1$, $\|k_m\|_{1+\alpha}^{1+\alpha}$ is not finite as required by Assumption (BP4), hence it is out-of-scope of the results presented here.

Finally, when $\alpha = 0$, we have $\|f_t\|_{1+\alpha} = 1$ implying that the condition (5.12) is satisfied. Using the corresponding Corollary 5.2, we obtain that the asymptotic breakdown point of the minimum power divergence functional is at least $(-\lambda)^{1/(1+\lambda)}$ for any $\lambda \in (-1, 0]$, and e^{-1} for $\lambda = 0$ as before.

5.4.6 Exponential Distribution with Normal Contamination

In this example, we explore the effect of contamination on the asymptotic breakdown point of MSDE for different density families for the model family and the contaminating density family. In particular, we choose the corresponding densities as

$$f_\theta(x) = \theta e^{-\theta x}\mathbf{1}_{\{x>0\}}, \text{ and, } k_m(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_m)^2}{2}\right),$$

for $x \in \mathbb{R}$. Without the loss of generality, the true density is taken to be the standard exponential distribution, with rate parameter equal to 1. The contaminating density k_m is a Gaussian density such that the absolute value of the mean parameter $|\mu_m| \to \infty$ as $m \to \infty$. Note that, we take the variance parameter σ of the contaminating density as 1, but the same argument can be modified appropriately for any fixed $\sigma > 0$.

To begin with, we note that

$$\|f_\theta\|_{1+\alpha}^{1+\alpha} = \frac{\theta^\alpha}{(1+\alpha)}, \text{ and, } \|k_m\|_{1+\alpha}^{1+\alpha} = (2\pi)^{-\alpha/2}(1+\alpha)^{-1/2}$$

If $\mu_m \to -\infty$ and a breakdown occurs, the mean of the estimated exponential density decreases towards 0, hence the rate parameter $\theta_m \to \infty$. Therefore, we would have $\|f_{\theta_m}\|_{1+\alpha} \to \infty$ in this case, which violates uniform $L^{1+\alpha}$-integrability of $\{f_\theta : \theta \in \Theta\}$ and hence Assumption (BP4) is not satisfied. Hence, we restrict our attention to the subset of the parameter space $(0, M]$ for some predetermined large constant M as in the previous examples.

Conversely, if $\mu_m \to \infty$, then $\theta_m \to 0$, then the condition (5.12) is not satisfied. We can make use of either the condition (5.13) or the condition (5.14) in this case. Let, without the loss of generality, assume $\theta^g = 1$ and consider the case when $B > 0$. Using condition (5.13), we obtain an asymptotic breakdown point of at least $\min\{1/2, \widetilde{\epsilon}_{(\alpha,\lambda)}\}$ where $\widetilde{\epsilon}_{(\alpha,\lambda)}$ is a solution of

$$\epsilon^{1+\alpha}\frac{\sqrt{1+\alpha}}{(2\pi)^{\alpha/2}B} = \begin{cases} 1/A - (1+\alpha)(1-\epsilon)^A/AB & \text{if } A > 0, \\ (1+\alpha)^{-1} + \ln(1-\epsilon) & \text{if } A = 0. \end{cases} \tag{5.17}$$

In this case, the lower bound is not explicitly obtained, and it has to be calculated as a solution to this implicit equation. However, it is possible to find an exact lower bound using condition (5.14) instead.

In order to apply condition (5.14), we consider the quantity

$$\begin{aligned}
&\int f_{\theta_m}^B(x)k_m^A(x)dx \\
&= \theta_m^B(2\pi)^{-A/2}\int_0^\infty e^{-\theta_m Bx - A(x-\mu_m)^2/2}dx \\
&= \theta_m^B(2\pi)^{-A/2}e^{-A\mu_m^2/2}\int_0^\infty e^{-Ax^2/2 + x(A\mu_m - \theta_m B)}dx \\
&= \theta_m^B(2\pi)^{-A/2}e^{-A\mu_m^2/2}\frac{\sqrt{\pi}}{\sqrt{2A}}\exp\left[(A\mu_m - \theta_m B)^2\,\frac{1}{2A}\right]\left(1+\widetilde{\Phi}\left(\frac{A\mu_m - B\theta_m}{\sqrt{2A}}\right)\right) \\
&= \theta_m^B(2\pi)^{-A/2}\frac{\sqrt{\pi}}{\sqrt{2A}}\exp\left[\frac{\theta_m^2B^2}{2A} - \mu_m\theta_m B\right]\left(1+\widetilde{\Phi}\left(\frac{A\mu_m - B\theta_m}{\sqrt{2A}}\right)\right),
\end{aligned}$$

for $A, B > 0$. Here, $\widetilde{\Phi}(\cdot)$ is the error function given by

$$\widetilde{\Phi}(x) := \frac{\sqrt{2}}{\sqrt{\pi}}\int_0^x e^{-t^2}dt.$$

Note that, since $\mu_m \to \infty$ and $\theta_m \to 0$, it follows that $(A\mu_m - B\theta_m) \to \infty$ for $A, B > 0$. Applying Dominated Convergence Theorem (DCT) on the function $e^{-t^2}\mathbf{1}_{[0,(A\mu_m - B\theta_m)]}$ then yields that $\widetilde{\Phi}\left(\frac{A\mu_m - B\theta_m}{\sqrt{2A}}\right) \to \widetilde{\Phi}(\infty) = 1$. Hence, we have an asymptotic equivalence

$$\int f_{\theta_m}^B(x)k_m^A(x)dx \asymp \theta_m^B(2\pi)^{-A/2}\frac{\sqrt{2\pi}}{\sqrt{A}}\exp\left[\frac{\theta_m^2B^2}{2A} - \mu_m\theta_m B\right],$$

for sufficiently large m. Therefore, we have an asymptotic approximation of L as in (5.15) given by

$$L \asymp \liminf_{m\to\infty}\frac{\theta_m^{(A-1)}(2\pi)^{(A-1)/2}\sqrt{A}}{(1+\alpha)}\exp\left[B\theta_m\mu_m - \frac{\theta_m^2B^2}{2A}\right].$$

Now, depending on the limit inferior of the sequence $\mu_m\theta_m$ and the choice of A, the quantity L will change, and so will the asymptotic breakdown point in this case. The resulting cases are summarized in Table 5.1.

Table 5.1: The asymptotic breakdown of MSD functional for different situations for exponential distribution as model family and normal distribution as contamination, with mean tending to $+\infty$, as shown in Example 5.4.6 The quantity $\widetilde{\epsilon}_{(\alpha,\lambda)}$ is a solution to (5.17).

A	θ_m **and** μ_m	**Bound on** L	**Lower bound to Asymptotic BP**
$A = 0$	$\liminf \theta_m\mu_m \le 2/(1+\alpha)$	∞	$1/2$
$A = 0$	$\liminf \theta_m\mu_m > 2/(1+\alpha)$	0	0
$0 < A < 1$	no restriction	∞	$\min\{1/2, 1-(B/(1+\alpha))^{1/A}\}$
$A = 1$	$\liminf \theta_m\mu_m = \infty$	∞	$1/2$
$A = 1$	$\liminf \theta_m\mu_m = c \in [0,\infty)$	$e^{\alpha c}/(1+\alpha)$	$\min\left\{1/2, \max\left\{\widetilde{\epsilon}_{(\alpha,\lambda;A=0)}, \frac{\alpha e^{\alpha c}}{(1+\alpha)^2}\right\}\right\}$
$A > 1$	$\liminf \theta_m\mu_m = c \in [0,\infty)$	0	$\min\{1/2, \widetilde{\epsilon}_{(\alpha,\lambda)}\}$
$A > 1$	$\liminf (B\theta_m\mu_m + (A-1)\log(\theta_m)) = c \in [0,\infty)$	$\frac{(2\pi)^{(A-1)/2}\sqrt{A}e^c}{(1+\alpha)}$	$\min\left\{1/2, \max\left\{\widetilde{\epsilon}_{(\alpha,\lambda)}, \frac{(\sqrt{A}B)^{1/A}e^{c/A}}{(2\pi)^{-(A-1)/2A}(1+\alpha)^{2/A}}\right\}\right\}$
$A > 1$	$\liminf (B\theta_m\mu_m + (A-1)\log(\theta_m)) = \infty$	∞	$\min\{1/2, 1-(B/(1+\alpha))^{1/A}\}$

For $A = 0$, we consider the integral

$$\begin{aligned}\int f_{\theta_m}^{1+\alpha} \ln(f_{\theta_m}/k_m) &= \int_0^\infty \theta_m^{1+\alpha} e^{-(1+\alpha)\theta_m x} \left(\ln(\theta_m) - \theta_m x - \ln(2\pi)/2 - \frac{1}{2}(x-\mu_m)^2\right) \\ &= \theta_m^{1+\alpha}\left(\ln(\theta_m/\sqrt{2\pi}) - \mu_m^2/2\right)\int_0^\infty e^{-(1+\alpha)\theta_m x} \\ &\quad + \theta_m^{1+\alpha}(\mu_m - \theta_m)\int_0^\infty x e^{-(1+\alpha)\theta_m x} - \frac{1}{2}\theta_m^{1+\alpha}\int_0^\infty x^2 e^{-(1+\alpha)\theta_m x} \\ &= \theta_m^\alpha \frac{\left(\ln(\theta_m/\sqrt{2\pi}) - \mu_m^2/2\right)}{1+\alpha} + \theta_m^{\alpha-1}\frac{(\mu_m - \theta_m)}{(1+\alpha)^2} - \frac{\theta_m^{\alpha-2}}{(1+\alpha)^3}.\end{aligned}$$

Therefore, when $A = 0$, the quantity L can be calculated exactly as

$$\begin{aligned}L &= \liminf_{m\to\infty}\left(\ln(\theta_m/\sqrt{2\pi}) - \mu_m^2/2\right) + \frac{\theta_m^{-1}(\mu_m - \theta_m)}{(1+\alpha)} - \frac{\theta_m^{-2}}{(1+\alpha)^2} \\ &= \liminf_{m\to\infty} \ln\left(\frac{\theta_m}{\sqrt{2\pi}} e^{-\mu_m^2/2}\right) + \left(\frac{\theta_m^{-1}\mu_m}{(1+\alpha)} - \frac{\theta_m^{-2}}{(1+\alpha)^2}\right) - \frac{1}{1+\alpha} \\ &= \liminf_{m\to\infty} \frac{1}{1+\alpha}\left[\theta_m^{-1}\mu_m - \ln\left(\sqrt{2\pi}\theta_m^{-2} e^{(\mu_m^2/2 - \theta_m^{-2}/(1+\alpha))}\right) - 1\right]\end{aligned}$$

The resulting value of L now depends on the limit inferior of $(\mu_m^2/2 - \theta_m^{-2}/(1+\alpha))$ which becomes the dominating factor in the exponential. The asymptotic breakdown point, in this case, is also illustrated in Table 5.1.

5.5 Empirical Studies

In Section 5.4, we directly use the theory developed in the paper to obtain the asymptotic breakdown points of the MSD and MDPD functionals. In contrast, in this section, we empirically verify them by observing the behaviour of the estimate as a function of the contaminating proportion ϵ. Here, we consider the S-divergence between the ϵ-contaminated density and the model family of distribution for different setups, and obtain the MSD functional by minimizing the S-divergence directly. We repeat this for varying choices of ϵ and observe the changes in the estimated parameter.

5.5.1 Normal Model Family

In this simulation exercise, we consider normal distribution $\mathcal{N}(\mu, \sigma^2)$ as the model family. Let us assume, without loss of generality, that the true distribution is the standard normal distribution with density ϕ. We have the contaminating distribution as $\mathcal{N}(\mu_0, \sigma_0^2)$ where μ_0 and σ_0^2 are to be chosen according to the type of contamination. Assume that the sample observations come from the contaminated mixture model with $(1-\epsilon)$ proportion of the data coming from the standard normal distribution and ϵ proportion of the data coming from the $\mathcal{N}(\mu_0, \sigma_0^2)$ distribution. Now for any choice of μ_0, σ_0^2, one can obtain the minimizer μ and σ^2 of the S-divergence between the model density $\phi((x-\mu)/\sigma)$ and $(1-\epsilon)\phi(x) + \epsilon\phi((x-\mu_0)/\sigma_0)$ as a function of the contaminating proportion ϵ. For any positive A, B, the MSD functional for estimation of μ or σ^2 can be obtained by minimizing

the objective function

$$\frac{1}{A}\int \sigma^{-(1+\alpha)}\phi^{(1+\alpha)}\left(\frac{x-\mu}{\sigma}\right)dx - \frac{1+\alpha}{AB}\int \sigma^{-B}\phi^{B}\left(\frac{x-\mu}{\sigma}\right)\left[(1-\epsilon)\phi\left(x\right)+\epsilon\sigma_0^{-1}\phi\left(\frac{x-\mu_0}{\sigma_0}\right)\right]^{A}dx. \quad (5.18)$$

However, it is difficult to obtain a closed form of the S-divergence for general values of α and λ because the second term becomes intractable. Only when $\lambda = 0$ (i.e., $A = 1$), the S-divergence reduces to the density power divergence (DPD) which has a neat solution. This can be obtained by noting that the cross-integral between two normal densities are

$$\begin{aligned}
&\int \sigma_1^{-1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right)\times\sigma_2^{-\alpha}\phi^{\alpha}\left(\frac{x-\mu_2}{\sigma_2}\right)dx \\
=\;& \sigma_1^{-1}\sigma_2^{-\alpha}\int_{-\infty}^{\infty}\frac{1}{(2\pi)^{(1+\alpha)/2}}\exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{\alpha(x-\mu_2)^2}{2\sigma_2^2}\right]dx \\
=\;& \sigma_1^{-1}\sigma_2^{-\alpha}(2\pi)^{-\alpha/2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left\{\left(\frac{1}{\sigma_1^2}+\frac{\alpha}{\sigma_2^2}\right)\left(x-\frac{(\frac{\mu_1}{\sigma_1^2}+\frac{\alpha\mu_2}{\sigma_2^2})}{(\frac{1}{\sigma_1^2}+\frac{\alpha}{\sigma_2^2})}\right)^2\right.\right. \\
&\qquad\left.\left.+\frac{\mu_1^2}{\sigma_1^2}+\frac{\alpha\mu_2^2}{\sigma_2^2}-\frac{(\frac{\mu_1}{\sigma_1^2}+\frac{\alpha\mu_2}{\sigma_2^2})^2}{(\frac{1}{\sigma_1^2}+\frac{\alpha}{\sigma_2^2})}\right\}\right]dx \\
=\;& \sigma_1^{-1}\sigma_2^{-\alpha}(2\pi)^{-\alpha/2}\frac{\sigma_1\sigma_2}{\sqrt{\alpha\sigma_1^2+\sigma_2^2}}\exp\left[-\frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2}+\frac{\alpha\mu_2^2}{\sigma_2^2}-\frac{(\mu_1\sigma_2^2+\alpha\mu_2\sigma_1^2)^2}{\sigma_1^2\sigma_2^2(\alpha\sigma_1^2+\sigma_2^2)}\right)\right] \\
=\;& (2\pi)^{-\alpha/2}\frac{\sigma_2^{-(\alpha-1)}}{\sqrt{\alpha\sigma_1^2+\sigma_2^2}}\exp\left[-\frac{-\alpha(\mu_1-\mu_2)^2}{2(\alpha\sigma_1^2+\sigma_2^2)}\right].
\end{aligned}$$

This means the objective function for MDPDE becomes proportional to

$$\left[\frac{\sigma^{-\alpha}}{\sqrt{1+\alpha}}-\left(1+\frac{1}{\alpha}\right)\sigma^{(1-\alpha)}\left\{\frac{(1-\epsilon)}{\sqrt{\sigma^2+\alpha}}e^{-\frac{\alpha\mu^2}{2(\sigma^2+\alpha)}}+\frac{\epsilon}{\sqrt{\sigma^2+\alpha\sigma_0^2}}e^{-\frac{\alpha(\mu-\mu_0)^2}{2(\sigma^2+\alpha\sigma_0^2)}}\right\}\right]. \quad (5.19)$$

Now for any choice of μ_0, σ_0^2, one can obtain the minimizer μ and σ^2 as a function of the contaminating proportion ϵ.

For $\lambda \neq 0$, although the S-divergence is intractable, we can express the second integral in (5.18) as

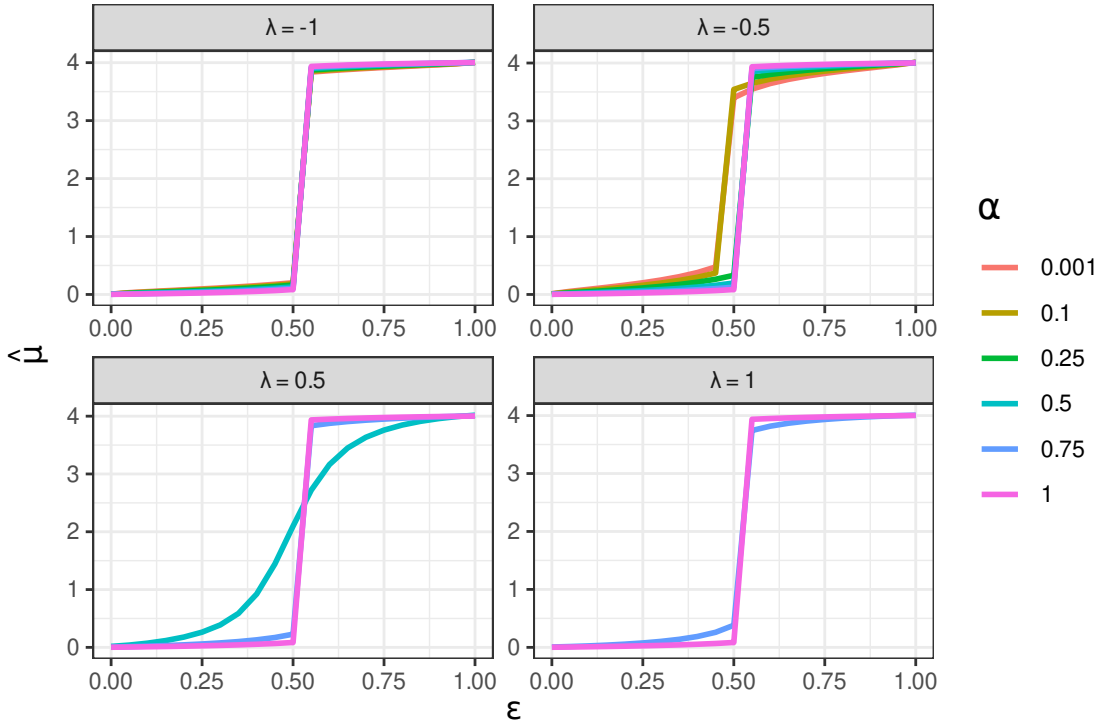
$$\sigma^{-(B-1)}\mathbb{E}_X\left[\phi^{B-1}\left(\frac{X-\mu}{\sigma}\right)\left[(1-\epsilon)\phi\left(X\right)+\epsilon\sigma_0^{-1}\phi\left(\frac{X-\mu_0}{\sigma_0}\right)\right]^{A}\right],$$

where $\mathbb{E}_X$ denote the expectation operator with respect to a random variable X following a normal distribution with mean μ and variance σ^2. Therefore, to calculate this integral, we perform a Monte Carlo integration technique and then proceed with the minimization.

First, we wish to see the effect of contamination for MSDE of a location parameter in the normal model family. For this, we pick $\sigma^2 = \sigma_0^2 = 1$, and $\mu_0 = 5$. This ensures that the mean of the contaminating density lies beyond the 3σ-limit of the true density, which approximates the asymptotic singular conditions present in Assumptions (BP2)-(BP3). The resulting estimates for the MSDE (including MDPDE) are depicted in Figure 5.2

for some choices of α and λ such that A and B are positive. It is clearly seen that the breakdown point of the location estimator is close 0.5 for all values of $\alpha > 0$, as indicated by the theoretical analysis.

Figure 5.2: Behaviour of MSDE (including MDPDE) under normal location model as a function of the contamination for different choices of α and λ (denoted in the title of individual plots).



Next, we shall consider simultaneous estimation of both location and scale parameters. Note that, the "implosion"-type breakdown of the scale parameter is out-of-scope of our results since Assumption (BP4) is not satisfied in this case when $\alpha > 0$. However, it is still possible to perform the numerical simulation exercise to gain insights into the behaviour of the MSDE in this case.

In this experiment, we fix the contaminated location as $\mu_0 = 5$ but take $\sigma_0 = 0.01$, emulating the setup where the variance of the contaminating distribution "implodes" to zero. The resulting estimates are shown in Figure 5.3. Due to the very small σ_0, the contaminating distribution is extremely spiky and almost singular to the true distribution, thus, even for a very small $\alpha > 0$, the resulting MDPDE for location parameter exhibits a high breakdown point near 0.5. However, for $\alpha = 0$, the resulting location estimate is the maximum likelihood estimate, which is a linear function of the contamination proportion ϵ, hence the MLE breaks down at any positive ϵ.

A very interesting situation ensues for the estimates of the scale parameter. For $\alpha = 0$, the resulting MLE for the precision parameter becomes linear in ϵ, which translates to the circular arc as shown in Figure 5.3 (right panel) for the variance (or scale) parameter. For $\alpha > 0$, the scale estimator remains close to the true value of 1 up to $\epsilon \in (0.2, 0.4)$

Figure 5.3: Behaviour of MDPD estimates under normal location-scale model as a function of the contamination proportion ϵ, with the location parameter in the left panel and the scale parameter in the right panel.

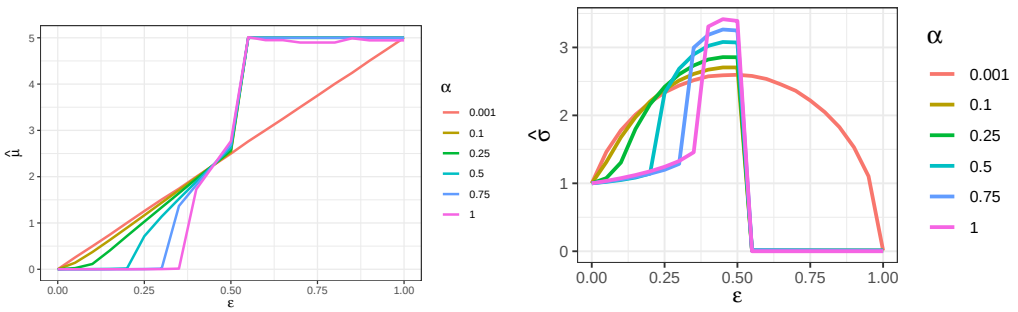
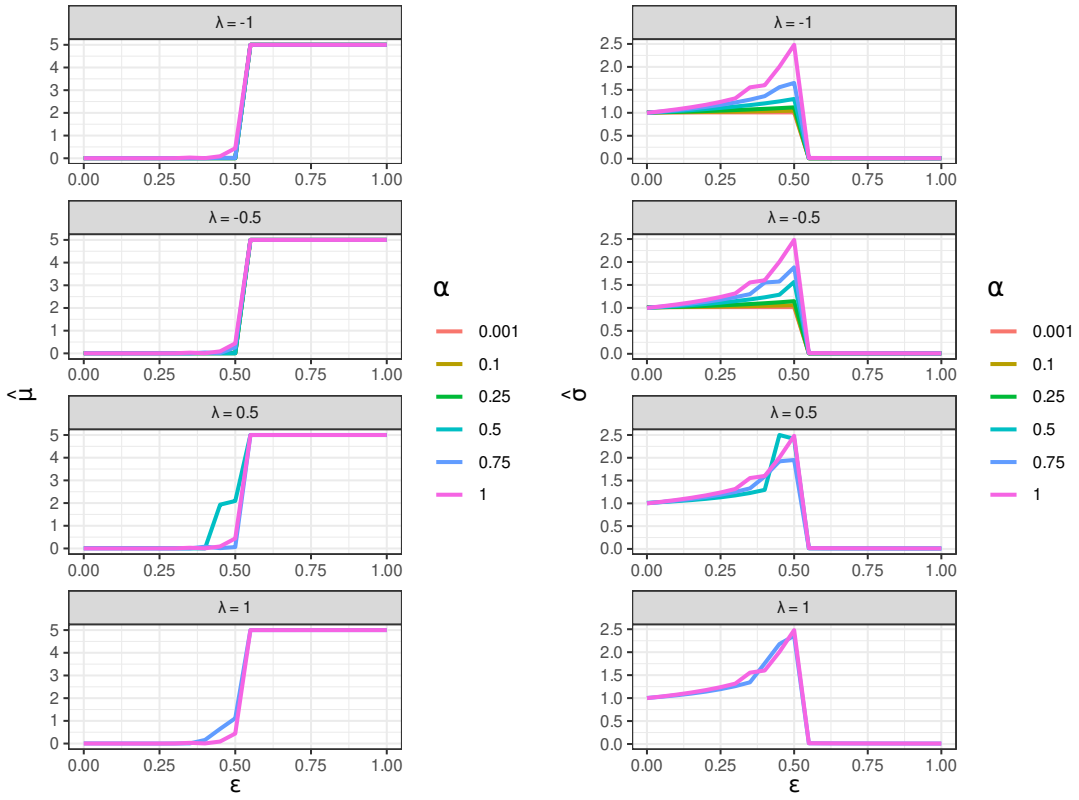


Figure 5.4: Behaviour of MSDE under normal location-scale model as a function of the contamination proportion ϵ, with the location parameter in the left panel and the scale parameter in the right panel, for different choices of λ (denoted in the title of individual plots).



depending on the value of α, after that, it immediately increases to a very high value and around $\epsilon = 0.5$, the estimator drops to the neighbourhood of 0 as the majority of the data then comes from $N(5, 0.01^2)$ distribution. In Figure 5.4, we show the MSDE for different values of λ. A similar phenomenon of increase in the estimate of the variance parameter is also observed across all values of λ, for contamination proportion between 0.3 and 0.5.

5.5.2 Gamma Distribution (Shape Family)

In Example 5.4.5, we demonstrated that the MDPDE of the shape parameter has an asymptotic breakdown of $\alpha/(1+\alpha)$, using the theoretical results derived in this chapter. In this section, we perform a simulation study to empirically verify this fact.

We consider the Gamma family of distributions Gamma$(t, 1)$ (with the rate parameter equal to 1), where the true value of the unknown parameter t is 1. The contaminating density is Gamma$(t_0, 1)$, where t_0 is a very large value or a very small value close to 0. To obtain the MDPDE estimate of the shape parameter under ϵ-level contamination, we minimize the objective function

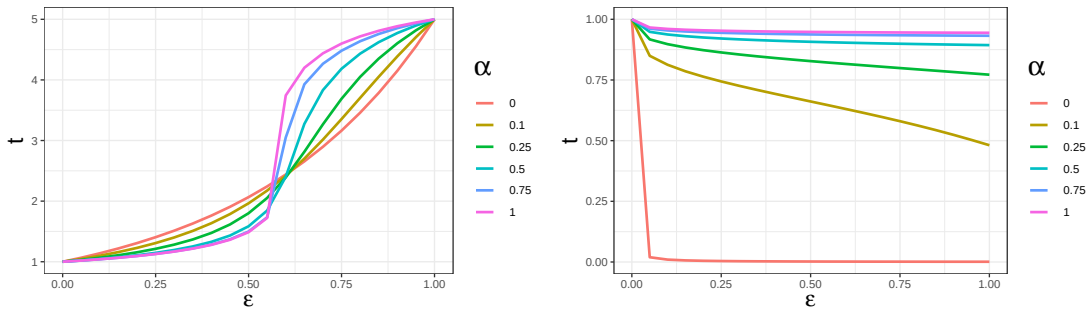
$$H(t,t) - \left(1 + \frac{1}{\alpha}\right)\left(\epsilon H(t,1) + (1-\epsilon)H(t,t_0)\right),$$

for any $\alpha > 0$, where

$$H(t_1,t_2) = (1+\alpha)^{-\alpha(t_1-1)-t_2}\frac{\Gamma(\alpha(t_1-1)+t_2)}{\Gamma(t_1)^{\alpha}\Gamma(t_2)},$$

with respect to the unknown parameter t. Then, we visualize how the estimate is changing as the contaminating proportion ϵ increases.

Figure 5.5: Behaviour of MDPD estimates under gamma density model as a function of the contamination proportion ϵ, for two values of shape parameter $t_0 = 10$ (Left) and $t_0 = 0.001$ (Right).



In the first experiment, we choose $t_0 = 5$ and in the second experiment, we choose $t_0 = 10^{-3}$, to demonstrate two contrasting types of contamination possible. The resulting estimates are depicted in Figure 5.5. As seen from both the figures, the resulting MDPDE remains very close to the true value of 1 in both cases for moderately high values of $\alpha > 0$. In comparison, for $\alpha = 0$, the resulting estimator (i.e., the MLE) breaks down at $\epsilon = 0$, as expected by its non-robust behaviour.

5.5.3 Exponential Distribution (Scale Parameter)

Let us assume, without loss of generality, that the true distribution is the standard exponential distribution with the rate parameter 1. Under ϵ-level contamination ($\epsilon \in [0,1]$), we assume that the sample observations follow a mixture of the standard exponential distribution with mixing proportion $(1-\epsilon)$ and the contaminating exponential distribution with rate parameter θ_0. The model family of distributions under consideration is again a family of exponential distributions with unknown rate parameter θ with $\theta > 0$.

In this case, the MDPDE aims to minimize the objective function

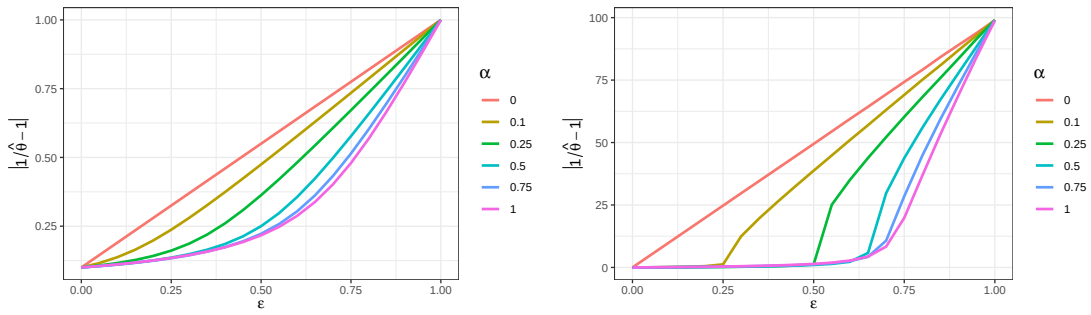
$$\int \theta^{(1+\alpha)} e^{-(1+\alpha)\theta x} dx - \left(1+\frac{1}{\alpha}\right)\left[\int \theta^{\alpha} e^{-\alpha\theta x}\left\{(1-\epsilon)e^{-x} + \epsilon\theta_0 e^{-\theta_0 x}\right\} dx\right].$$

It simplifies to

$$\theta^{\alpha}\left[\frac{1}{1+\alpha} - \left(1+\frac{1}{\alpha}\right)\left(\frac{1-\epsilon}{(\alpha\theta+1)} + \frac{\epsilon\theta_0}{(\alpha\theta+\theta_0)}\right)\right].$$

Given a choice of θ_0, we can obtain the minimizer of the MDPD objective function with respect to θ and visualize how the MDPDE behaves as a function of the contamination proportion ϵ. As indicated before, the rate parameter of the contaminating distribution can either implode to zero or explode to infinity, in view of Assumption (BP2). So, in the first experiment, we choose $\theta_0 = 10$ and in the second experiment, we consider $\theta_0 = 0.01$. The corresponding absolute biases of the MDPDE of the inverse rate parameter as a function of ϵ for varying tuning parameter α have been illustrated in Figure 5.6. For both cases, $\theta_0 = 10$ and $\theta_0 = 0.01$, the resulting MDPDE of the inverse rate parameter corresponding to $\alpha = 0$ (i.e., the MLE) turns out to be a linear function of the contamination proportion ϵ, and the absolute bias also remains linear in ϵ. When $\theta_0 = 10$, for higher values of α, the bias increases as a convex function of ϵ and the curvatures of those functions increase in α, resulting in an estimator with a higher breakdown. On the other hand, when $\theta_0 = 0.01$, the robust behaviour of the MDPDE can be effectively seen from Figure 5.6. As the contamination proportion ϵ increases, the resulting MDPDE moves away from the true inverse rate parameter 1 to the contaminated value $1/\theta_0 = 100$, but at a very slow rate for $\alpha \geq 0.25$.

Figure 5.6: Behaviour of MDPD estimates under exponential model as a function of the contamination proportion ϵ, for $\theta_0 = 10$ (Left) and $\theta_0 = 0.01$ (Right).



We also analyse the behaviour of the resulting estimate under different levels of contamination for the gamma distribution setup. The results are very similar to the results

obtained for the exponential case with the exception that the choice of the shape parameter t governs the curvatures of the curves shown in Figure 5.6; increasing t results in more robust behaviours of MDPDE with $\alpha > 0$.

5.5.4 Binomial Model

Here we consider an example from Basu et al. (2011), for which we empirically demonstrate the breakdown point of the minimum S-divergence estimator. Let the model family of distributions be a binomial distribution with size parameter 12 and unknown success probability $\theta \in [0, 1]$. We assume that the true value of θ is 0.5 which is the target quantity to estimate. Clearly, this discrete setup does not naturally come under the paradigm of minimum S-divergence estimation which assumes the existence of the densities for the model distribution. Additionally, the Assumptions (BP2) and (BP3) do not hold for any choice of the contaminating distribution which has bounded support in $[0, 12]$. Despite this, we can still obtain an estimator analogous to the MSDE by minimizing a discretized version of S-divergence, where we replace the densities with corresponding probability mass functions. To satisfy the asymptotic singularity as required by Assumptions (BP2) and (BP3) as much as possible, we may consider the contaminating distribution as the Dirac delta distribution at 12. This also yields the maximum effect of contamination possible in this discrete setup. Therefore, we denote

$$g_\epsilon(x) = (1-\epsilon)2^{-12}\binom{12}{x} + \epsilon\delta_{12}(x), \;\; x \in \{0, 1, \ldots, 12\}, \tag{5.20}$$

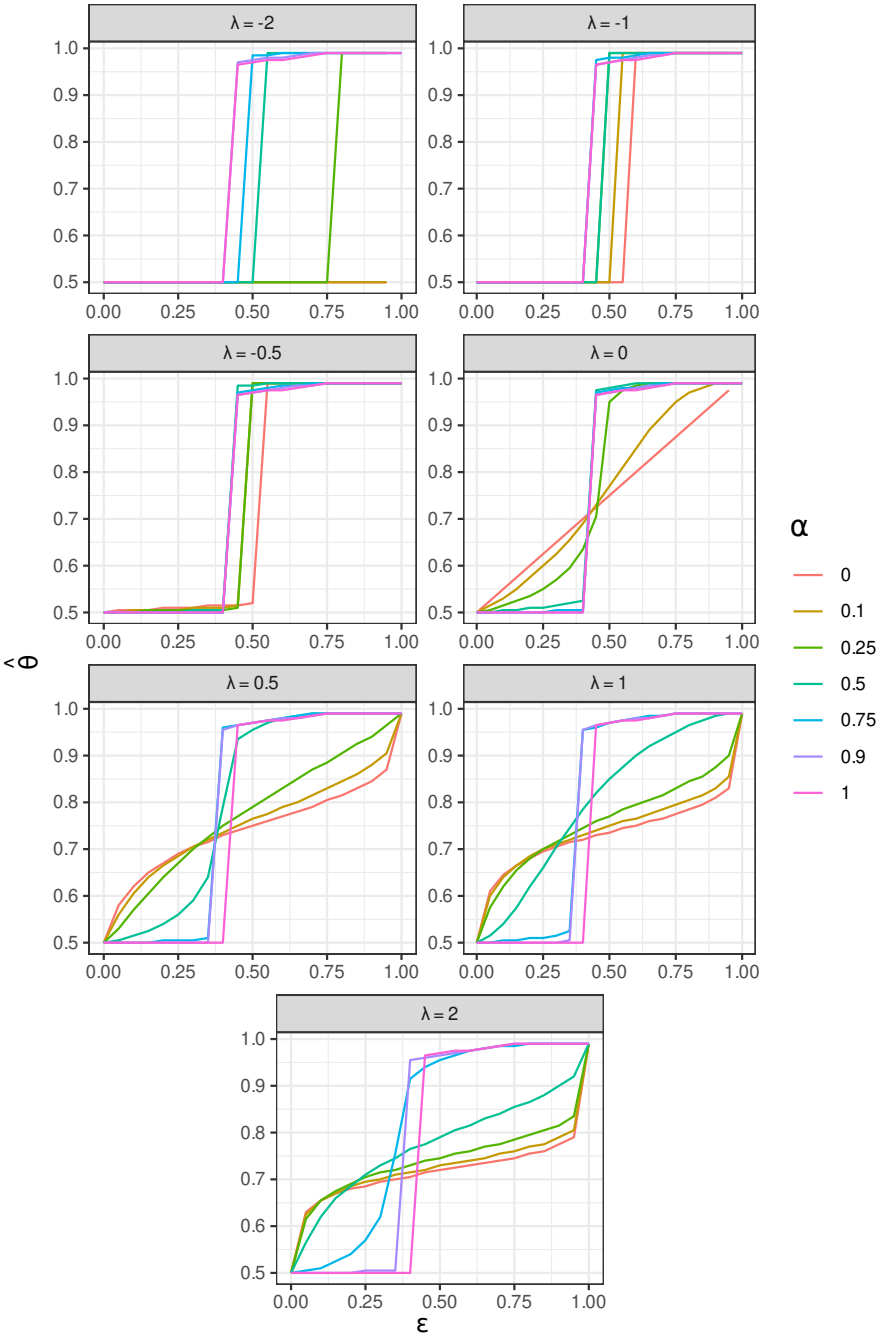
as the contaminated probability mass function (PMF) at level $\epsilon \in [0, 1]$ from which the samples are assumed to be generated. Here, $\delta_{12}(x)$ is identically equal to 0 for all $x \in \{0, 1, \ldots, 11\}$ but at $x = 12$, it is equal to 1. Now, the definition of the S-divergence can be readily extended to measure the statistical discrepancy between two probability mass functions in a discrete setup, and hence the MSD functional for this particular setup can be written as

$$\widehat{\theta}_{(\alpha,\lambda)}(\epsilon) = \underset{\theta\in[0,1]}{\arg\min} \left(\frac{1}{A}\sum_{x=0}^{12} f_\theta^{1+\alpha}(x) - \frac{1+\alpha}{AB}\sum_{x=0}^{12} f_\theta^B(x) g_\epsilon^A(x) + \frac{1}{B}\sum_{x=0}^{12} g_\epsilon^{1+\alpha}(x) \right),$$

where $g_\epsilon(x)$ is as defined in (5.20) and $f_\theta(x)$ is the probability mass function of the binomial distribution with size parameter 12 and success probability θ evaluated at x.

In Figure 5.7, we plot the values of $\widehat{\theta}_{(\alpha,\lambda)}(\epsilon)$ for different choices of α and λ as a function of ϵ. For $\alpha = 0$ and $\lambda = 0$, when the MSD functional corresponds to the minimum Kullback-Leibler divergence functional (or the maximum likelihood functional), the estimate becomes a linear function of the contamination proportion ϵ. For $\lambda = 0$, as α increases to 1, the estimate becomes more robust. As shown in Figure 5.7, for values of α closer to 1, the estimator remains near 0.5 when $\epsilon < 0.5$ and jumps rapidly to 1 when $\epsilon > 0.5$. For $\lambda < 0$, a wider range of α ensures this robust behaviour. However, when $\lambda > 0$, only comparatively higher values of α (usually $\alpha > 0.5$) demonstrate this robust behaviour.

Figure 5.7: Minimum S-divergence estimate of success proportion under different levels of contamination in the binomial model for different choices of α and λ (λ is denoted in the title of each individual plot).



5.A Appendix: Proof of the useful lemmas

Lemma 5.A.1. *Fix an $\alpha \geq 0$. Let, $\{f_m\}_{m=1}^{\infty}, \{g_m\}_{m=1}^{\infty}$ and $\{\tilde{g}_m\}_{m=1}^{\infty}$ be three sequences of nonnegative functions such that they are uniformly $L^{1+\alpha}$-integrable, i.e., $\sup_m \int f_m^{1+\alpha} < \infty$, $\sup_m \int g_m^{1+\alpha} < \infty$, and $\sup_m \int \tilde{g}_m^{1+\alpha} < \infty$. Also, let $(g_m - \tilde{g}_m) \to 0$ pointwise as $m \to \infty$. Then,*

$$\left| \int g_m^{1+\alpha} - \int \tilde{g}_m^{1+\alpha} \right| \to 0, \text{ and, } \left| \int f_m^B g_m^A - \int f_m^B \tilde{g}_m^A \right| \to 0,$$

as $m \to \infty$, where $A > 0, B > 0$ and $(A + B) = (1 + \alpha)$.

Proof. We begin with the first convergence result. Note that,

$$\begin{aligned}
\left| \int g_m^{1+\alpha} - \int \tilde{g}_m^{1+\alpha} \right| &\leq \int |g_m^{1+\alpha} - \tilde{g}_m^{1+\alpha}| \\
&\leq (1+\alpha) \int \max\{g_m^\alpha, \tilde{g}_m^\alpha\} |g_m - \tilde{g}_m|, \\
&\qquad (\text{since, } |x^{1+\alpha} - y^{1+\alpha}| < (1+\alpha)\max\{x,y\}^\alpha |x-y|) \\
&= (1+\alpha) \int g_m^\alpha \mathbf{1}_{\{g_m \geq \tilde{g}_m\}} |g_m - \tilde{g}_m| + (1+\alpha) \int \tilde{g}_m^\alpha \mathbf{1}_{\{g_m \geq \tilde{g}_m\}} |g_m - \tilde{g}_m|.
\end{aligned}$$

The proof now follows if we can show that both of the above terms go to 0 as $m \to \infty$. We shall show the convergence for the first term only; the convergence for the second term follows analogously. Now pick any $\epsilon > 0$. Since, g_m is uniformly $L^{1+\alpha}$-integrable, there exists a sufficiently large M such that

$$\int g_m^{1+\alpha} \mathbf{1}_{\{g_m \geq M\}} < \frac{\epsilon}{4(1+\alpha)},$$

for any m. Therefore, it follows that

$$\begin{aligned}
& (1+\alpha) \int g_m^\alpha \mathbf{1}_{\{g_m \geq \tilde{g}_m\}} |g_m - \tilde{g}_m| \\
= \; & (1+\alpha) \int g_m^\alpha \mathbf{1}_{\{g_m \geq \tilde{g}_m, g_m > M\}} |g_m - \tilde{g}_m| + (1+\alpha) \int g_m^\alpha \mathbf{1}_{\{g_m \geq \tilde{g}_m, g_m \leq M\}} |g_m - \tilde{g}_m| \\
\leq \; & 2(1+\alpha) \int g_m^{1+\alpha} \mathbf{1}_{\{g_m > M\}} + (1+\alpha) M^\alpha \int |g_m - \tilde{g}_m| \mathbf{1}_{\{g_m < M, \tilde{g}_m < M\}} \\
\leq \; & \frac{\epsilon}{2} + (1+\alpha) M^\alpha \int |g_m - \tilde{g}_m| \mathbf{1}_{\{g_m < M, \tilde{g}_m < M\}}.
\end{aligned}$$

Now an application of the Dominated Convergence Theorem (DCT) shows that the second term can be made arbitrarily small, in particular, smaller than $\epsilon/2(1+\alpha)M^\alpha$ by choosing sufficiently large m. Combining both of these, we obtain that

$$(1+\alpha) \int g_m^\alpha \mathbf{1}_{\{g_m \geq \tilde{g}_m\}} |g_m - \tilde{g}_m| < \epsilon,$$

as we wanted to show.

For the second convergence result, note that when $A, B > 0$, we can apply Hölder's inequality as follows

$$\begin{aligned}\left|\int f_m^B g_m^A - \int f_m^B \tilde{g}_m^A\right| &\leq \int f_m^B |g_m^A - \tilde{g}_m^A| \\ &\leq \left(\int f_m^{1+\alpha}\right)^{B/(1+\alpha)} \left(\int \delta_m^{1+\alpha}\right)^{A/(1+\alpha)} \\ &\leq \left(\sup_m \int f_m^{1+\alpha}\right)^{B/(1+\alpha)} \left(\int \delta_m^{1+\alpha}\right)^{A/(1+\alpha)}\end{aligned}$$

where $\delta_m = |g_m^A - \tilde{g}_m^A|^{1/A}$. Therefore, it is enough to show that $\int \delta_m^{1+\alpha} \to 0$ as $m \to \infty$. Similar to the proof of the first convergence result, we can split the integral $\int \delta_m^{1+\alpha}$ into two parts: one where $g_m > \tilde{g}_m$ and another where $g_m \leq \tilde{g}_m$. For the first part, note that

$$\begin{aligned}\int \delta_m^{1+\alpha} \mathbf{1}_{\{g_m > \tilde{g}_m\}} &= \int \left((g_m - \tilde{g}_m + \tilde{g}_m)^A - \tilde{g}_m^A\right)^{(1+\alpha)/A} \mathbf{1}_{\{g_m > \tilde{g}_m\}} \\ &\leq \int \left((g_m - \tilde{g}_m + \tilde{g}_m)^{1+\alpha} - \tilde{g}_m^{1+\alpha}\right)^{(1+\alpha)/(1+\alpha)} \mathbf{1}_{\{g_m > \tilde{g}_m\}},\end{aligned}$$

where we use the fact that for any $y, z \geq 0$, the quantity $[(y+z)^x - y^x]^{1/x}$ is increasing for $x > 0$. Now it follows from the first convergence result that the above tends to 0 as $m \to \infty$. The second part of the integral, i.e., the part where $g_m \leq \tilde{g}_m$ can be dealt with similarly. □

Lemma 5.A.2. *Fix an $\alpha \geq 0$. Let, $\{f_m\}_{m=1}^\infty, \{g_m\}_{m=1}^\infty$ be two sequences of nonnegative functions such that they are uniformly $L^{1+\alpha+\delta}$-integrable, i.e., $\sup_m \int f_m^{1+\alpha+\delta} < \infty$ and $\sup_m \int g_m^{1+\alpha+\delta} < \infty$, for some $\delta > 0$. Also, let $(f_m - \tilde{f}_m) \to 0$ and $(g_m - \tilde{g}_m) \to 0$ pointwise where $\tilde{f}_m$ and $\tilde{g}_m$ are also uniformly $L^{1+\alpha+\delta}$-integrable. Furthermore, assume for each m, Supp$(g_m) \subseteq$ Supp(f_m) and Supp$(\tilde{g}_m) \subseteq$ Supp$(\tilde{f}_m)$, where Supp$(\cdot)$ denotes the support of a function. Then,*

$$\left|\int g_m^{1+\alpha} \ln(f_m) - \int \tilde{g}_m^{1+\alpha} \ln(\tilde{f}_m)\right| \to 0$$

as $m \to \infty$ provided that

$$\sup_m \left|\int g_m^{1+\alpha} \ln(f_m)\right| < \infty, \ \sup_m \left|\int \tilde{g}_m^{1+\alpha} \ln(\tilde{f}_m)\right| < \infty.$$

Also, if $g_m = f_m$ or $\tilde{g}_m = \tilde{f}_m$, the corresponding conditions follow trivially from $L^{1+\alpha+\delta}$-finiteness assumption. In case, $\tilde{f}_m \equiv 0$ and $\tilde{g}_m \equiv 0$, we use the convention $0 \ln(0) := 0$ for calculating the integral $\int \tilde{g}_m^{1+\alpha} \ln(\tilde{f}_m)$ with $\alpha \geq 0$.

Proof. We are given that

$$\left|\sup_m \int g_m^{1+\alpha} \ln(f_m)\right| < \infty.$$

In case $g_m = f_m$, we have

$$\begin{aligned}\left|\int g_m^{1+\alpha} \ln(g_m)\right| &\leq \int |g_m^{1+\alpha} \ln(g_m)| \\ &= -\int g_m^{1+\alpha} \ln(g_m) \mathbf{1}_{\{g_m \leq 1\}} + \int g_m^{1+\alpha} \ln(g_m) \mathbf{1}_{\{g_m > 1\}} \\ &\leq \int g_m^{\alpha} + C \int g_m^{1+\alpha+\delta},\end{aligned}$$

where C is positive constant independent of m, and we use the fact that $x\ln(x) \geq (-1)$. By the uniform $L^{1+\alpha+\delta}$-integrability of g_m, we obtain that $\sup_m \int |g_m^{1+\alpha}\ln(g_m)| < \infty$. Similarly, it follows that $\left|\sup_m \int \tilde{g}_m^{1+\alpha}\ln(\tilde{g}_m)\right|$ also exists and is finite.

In the case when $g_m \neq f_m$ or $\tilde{g}_m \neq \tilde{f}_m$, the finiteness of the integrals $\int g_m^{1+\alpha}\ln(f_m)$ and $\int \tilde{g}_m^{1+\alpha}\ln(\tilde{f}_m)$ are guaranteed by the conditions of the Lemma 5.A.2.

Now let us denote $h_m = g_m^{1+\alpha}\ln(f_m)$ and $\tilde{h}_m = \tilde{g}_m^{1+\alpha}\ln(\tilde{f}_m)$. Then, given an $\epsilon > 0$, because of the finiteness of the above integrals, we can choose a sufficiently large M such that

$$\begin{aligned}
\left|\int h_m - \int \tilde{h}_m\right| &\leq \int |h_m - \tilde{h}_m| \\
&= \int |h_m - \tilde{h}_m|\mathbf{1}_{\{|h_m|>|\tilde{h}_m|\}} + \int |h_m - \tilde{h}_m|\mathbf{1}_{\{|h_m|\leq|\tilde{h}_m|\}} \\
&\leq 2\int |h_m|\mathbf{1}_{\{|h_m|>M\}} + 2\int |\tilde{h}_m|\mathbf{1}_{\{|\tilde{h}_m|>M\}} + \int |h_m - \tilde{h}_m|\mathbf{1}_{\{|h_m|\leq M,|\tilde{h}_m|\leq M\}} \\
&\leq \frac{\epsilon}{2} + \int |h_m - \tilde{h}_m|\mathbf{1}_{\{|h_m|\leq M,|\tilde{h}_m|\leq M\}}.
\end{aligned}$$

Now, due to the pointwise convergences $(g_m - \tilde{g}_m) \to 0$ and $(f_m - \tilde{f}_m) \to 0$ as $m \to \infty$ and the support conditions for the well-definedness, we have the pointwise convergence $(h_m - \tilde{h}_m)\mathbf{1}_{\{|h_m|\leq M,|\tilde{h}_m|\leq M\}} \to 0$ as $m \to \infty$ a.e. by continuous mapping theorem. Here, we use the convention $0\ln(0) := 0$ whenever appropriate. Now, an application of the Dominated Convergence Theorem (DCT) implies that the second term can be made arbitrarily small, in particular, smaller than $\epsilon/2$ by choosing sufficiently large m. This completes the proof. □

Chapter 6

Breakdown Analysis of Minimum Generalized Alpha-Beta Divergence Estimator

6.1 Introduction

In Chapter 5, we illustrated how the minimum S-divergence estimator achieves a high asymptotic breakdown point across many setups, independent of the dimension of the data or the parameter. It turns out that such a phenomenon holds for even a broader spectrum of minimum divergence estimators, namely the minimum Generalized Alpha-Beta divergence estimators.

As noted earlier, Ghosh et al. (2017) introduced the S-divergence family by connecting the family of Cressie-Read power divergence smoothly to the squared L^2 divergence using two hyperparameters α and λ. Parallelly, Maji et al. (2014) proposed the logarithmic density power divergence family which was later extended to the logarithmic S-divergence family by Maji (2019); members of this family have been noted to have stronger robustness properties compared to the S-divergence family (Fujisawa and Eguchi, 2008). In another direction, Jones et al. (2001) developed a (ϕ, γ)-divergence family, which includes both the density power divergence (DPD) and the logarithmic density power divergence (LDPD) families. For an application to a robust nonnegative matrix factorization problem, Cichocki et al. (2011) introduced a family of Alpha-Beta divergence (also known as AB divergence) which is a superclass of the S-divergence family but extending beyond the choices of the hyperparameters $\alpha \geq 0$ and $\lambda \in \mathbb{R}$, to allow α to take even negative values. The authors also hint at a similar generalization using logarithms which produces a superclass of the logarithmic S-divergence family.

These divergence measures and their relationships provide guidance to develop new divergences with interesting properties such as robustness, efficiency, convexity, etc., which are useful in different applications; see Cichocki and Amari (2010). As a generalization to all of the divergences mentioned above, we introduce a generalized Alpha-Beta (GAB)

divergence family in Section 6.2 of this chapter. This family was partially hinted by Maji (2019), but no proper characterization of the divergence measure was provided. Even whether this generalized class results in a well-defined statistical divergence measure has so far been unknown, until it was partially solved for the case of the functional density power divergence (FDPD) family by Ray et al. (2023). Therefore, in Section 6.3 we provide the necessary and sufficient conditions for the proposed GAB measure to be a valid divergence measure. Then, in Section 6.4, we show that under some reasonable assumptions as in (BP1)-(BP4), the minimum generalized Alpha-Beta divergence (MGABD) functional has a dimension-free lower bound to its asymptotic breakdown point. Section 6.5 provides some numerical illustrations and example applications of these results across different setups.

6.2 Generalized Alpha-Beta Divergence Family

Let F and G be two nonnegative measures dominated by a common measure μ and let f and g be their corresponding nonnegative Radon-Nikodym derivatives. Note that, F and G may not be probability measures and hence f and g may not be probability density functions. In this general cases, we shall refer to them as non-normalized density functions instead. Let us also consider a differentiable function $\psi : (0, \infty) \to \mathbb{R}$ such that its first order derivative is continuous, i.e., $\psi \in C^1((0, \infty))$, satisfying the conditions described later in Section 6.3. Then, the generalized Alpha-Beta (GAB) divergence is defined as

$$d_{GAB}^{(\alpha,\beta),\psi}(f, g) = \frac{1}{\beta(\alpha+\beta)}\psi\left(\|f\|_{\alpha+\beta}^{\alpha+\beta}\right) + \frac{1}{\alpha(\alpha+\beta)}\psi\left(\|g\|_{\alpha+\beta}^{\alpha+\beta}\right) - \frac{1}{\alpha\beta}\psi\left(\langle f, g\rangle_{\alpha,\beta}\right), \quad (6.1)$$

when $\alpha, \beta, (\alpha+\beta)$ are all nonzero, provided that the integrals $\int f^{\alpha+\beta}$, $\int g^{\alpha+\beta}$ and $\int f^{\alpha} g^{\beta}$ exist and are finite. With the choice of $\psi(x) = x$, this leads to the Alpha-Beta (AB) divergence proposed by Cichocki et al. (2011). Also note that, the S-divergence as in (1.10) is a subclass of the GAB divergence as shown below:

$$S_{(\alpha,\lambda)}(g, f) = d_{GAB}^{(1+\lambda(1-\alpha),\alpha-\lambda(1-\alpha)),\psi(x)=x}(f, g), \ \alpha \geq 0, \lambda \in \mathbb{R}.$$

Similarly, the choice $\psi(x) = \ln(x)$ leads to the logarithmic Alpha-Beta divergence family. With additional restriction that $\alpha + \beta = (1 + a)$ for some $a \geq 0$, it further reduces to the logarithmic S-divergence family. On the other hand, choosing $\psi(x) = x^{\phi}/\phi$, $\alpha = \gamma$ and $\beta = 1$ leads to the (ϕ, γ)-divergence family considered by Jones et al. (2001). A detailed summarization of these relationships is provided in Table 6.1.

Since $\psi \in C^1((0, \infty))$, one can define the form of GAB divergence for $\alpha = 0, \beta = 0$ or $\alpha + \beta = 0$ cases using the corresponding limits,

$$d_{GAB}^{(\alpha,\beta),\psi}(f, g) = \begin{cases} \frac{1}{\alpha^2}\left(\psi'(\|f\|_{\alpha}^{\alpha})\int f^{\alpha}\ln(f^{\alpha}/g^{\alpha}) - \psi(\|f\|_{\alpha}^{\alpha}) + \psi(\|g\|_{\alpha}^{\alpha})\right) & \text{if } \alpha \neq 0, \beta = 0, \\ \frac{1}{\beta^2}\left(\psi'(\|g\|_{\beta}^{\beta})\int g^{\beta}\ln(g^{\beta}/f^{\beta}) - \psi(\|g\|_{\beta}^{\beta}) + \psi(\|f\|_{\beta}^{\beta})\right) & \text{if } \alpha = 0, \beta \neq 0, \\ \frac{1}{\alpha^2}\left(\psi'(1)\int \ln(g^{\alpha}/f^{\alpha}) + \psi\left(\int f^{\alpha}/g^{\alpha}\right) - \psi(1)\right) & \text{if } \alpha = -\beta \neq 0, \\ \frac{\psi'(1)}{2}\int(\ln f - \ln g)^2 & \text{if } \alpha = \beta = 0. \end{cases} \quad (6.2)$$

Table 6.1: Different families of statistical divergence that arise as special cases of the Generalized Alpha-Beta divergence family, up to a multiplicative constant.

Divergence Family	**Form of the divergence**	$\psi(x)$	**Hyperparameters**
Alpha-Beta Family	$\frac{\int f^{\alpha+\beta}}{\beta(\alpha+\beta)} - \frac{\int f^{\alpha} g^{\beta}}{\alpha\beta} + \frac{\int g^{\alpha+\beta}}{\alpha(\alpha+\beta)}$	$\psi(x) = x$	$\alpha, \beta \in \mathbb{R} \setminus \{0\}, (\alpha+\beta) \neq 0$
Density Power Divergence	$\int f^{1+a} - \left(1+\frac{1}{a}\right) \int f^{a} g + \frac{1}{a} \int g^{1+a}$	$\psi(x) = x$	$\alpha = a, \beta = 1$
Power divergence	$\frac{1}{\lambda(\lambda-1)} \left[\int f^{\lambda} g^{1-\lambda} - \lambda \int f + (\lambda-1) \int g\right]$	$\psi(x) = x$	$\alpha = \lambda, \beta = (1-\lambda)$
Kullback Leibler divergence	$\int f \ln(f/g)$	$\psi(x) = x$	$\alpha = 1, \beta = 0$
S-divergence	$\frac{\int f^{1+a}}{A} - \frac{1+a}{AB} \int f^{B} g^{A} + \frac{\int g^{1+a}}{B}$	$\psi(x) = x$	$\alpha = B, \beta = A = (1+a-\alpha)$
AC divergence	$\ln \left[\|f\|_{\alpha+\beta}^{\beta} \|g\|_{\alpha+\beta}^{\alpha}\right] - \ln \left[\langle f, g \rangle_{\alpha,\beta}^{1/\alpha\beta}\right]$	$\psi(x) = \ln(x)$	$\alpha, \beta \in \mathbb{R} \setminus \{0\}, (\alpha+\beta) \neq 0$
Logarithmic S-divergence	$\ln \left[\|f\|_{1+a}^{(1+a)/A} \|g\|_{1+a}^{(1+a)/B}\right] - \ln \left[\langle f, g \rangle_{B,A}^{(1+a)/AB}\right]$	$\psi(x) = \ln(x)$	$\alpha = B, \beta = A = (1+a-\alpha)$
γ-divergence	$\frac{1}{\gamma(\gamma-1)} \ln \left[\frac{\int f^{\gamma} g^{1-\gamma}}{(\int f)^{\gamma} (\int g)^{1-\gamma}}\right]$	$\psi(x) = \ln(x)$	$\alpha = \gamma, \beta = (1-\gamma)$
(ϕ, γ)-divergence	$\phi^{-1}\left(\int f^{1+\gamma}\right)^{\phi} - \frac{1+\gamma}{\gamma} \phi^{-1} \left(\int f^{\gamma} g\right)^{\phi} + \frac{1}{\gamma} \phi^{-1} \left(\int g^{1+\gamma}\right)^{\phi}$	$\psi(x) = \phi^{-1} x^{\phi}$	$\alpha = \gamma, \beta = 1$

It is easy to see that when $\psi(x) = x$, $\psi'(x) = 1$ and the above limiting cases match exactly with the corresponding cases for the AB divergence family as provided in Cichocki et al. (2011).

The specific form of the GAB divergence allows us to derive some nice properties of it using elementary calculations, as described below.

1. **Duality Property:** The GAB divergence satisfy a duality property given as

$$d_{GAB}^{(\alpha,\beta),\psi}(f, g) = d_{GAB}^{(\beta,\alpha),\psi}(g, f). \tag{6.3}$$

2. **Scaling Property:** For any $c > 0$, let $\psi_c(x) := \psi(c^{\alpha+\beta} x)$. Then,

$$d_{GAB}^{(\alpha,\beta),\psi}(cf, cg) = d_{GAB}^{(\alpha,\beta),\psi_c}(f, g). \tag{6.4}$$

3. **Zooming Property:** For any $c \neq 0$, we have

$$d_{GAB}^{(\alpha,\beta),\psi}(f, g) = c^2 d_{GAB}^{(c\alpha,c\beta),\psi}(f^{1/c}, g^{1/c}). \tag{6.5}$$

   In particular, this allows us to scale either of the hyperparameter α or β to be equal to 1 (provided they are nonzero), which reduces the GAB divergence to a generalized one-parameter divergence family defined on the class of appropriately zoomed densities. Mathematically, we have

$$d_{GAB}^{(\alpha,\beta),\psi}(f, g) = \frac{1}{\beta^2} d_{GAB}^{(\alpha/\beta,1),\psi}(f^{\beta}, g^{\beta}) = \frac{1}{\alpha^2} d_{GAB}^{(1,\beta/\alpha),\psi}(f^{\alpha}, g^{\alpha}). \tag{6.6}$$

6.3 Characterization of GAB Divergence Family

Before we move forward, it is important to establish the class of generating functions ψ such that the corresponding GAB divergence measure is a well-defined statistical measure of discrepancy. In this section, we aim to characterize this class by obtaining some necessary and sufficient conditions for the class of ψ-functions. A statistical divergence measure should satisfy the following two key criteria.

1. **Nonnegativity:** For any pair of non-normalized densities f and g, $d_{GAB}^{(\alpha,\beta),\psi}(f,g) \geq 0$, provided that the corresponding integrals are finite.

2. **Strict positivity:** For any two non-normalized densities f and g, $d_{GAB}^{(\alpha,\beta),\psi}(f,g) = 0$ if and only if f/g is a constant almost surely with respect to the common dominating measure μ.

If both $\alpha, \beta > 0$, then it is easy to verify by an application of Hölder's inequality and Jensen's inequality that if ψ is monotonically increasing and convex, the form in (6.1) turns out to be nonnegative for all pair of non-normalized densities f and g (also see Lemma 6.A.3). However, this characterization does not encompass the entire class of generating functions ψ for which the resulting GAB divergence is a valid divergence. In particular, the family of logarithmic super-divergences (LSD) or the logarithmic Alpha-Beta divergence which are known to be valid divergence measures (Ghosh et al., 2017), its corresponding generating function $\psi(x) = \ln(x)$ is not convex.

To completely characterize this class of valid generating functions, we follow the footsteps of Ray et al. (2023). In fact, our results are more general than that of the functional density power divergence (FDPD) class proposed by Ray et al. (2023), as it covers a significantly larger class of divergence measures with general choices of α and β. The results for the FDPD case then follows from our general results with a simple replacement of $\beta = 1$. However, because the forms of GAB divergence as in Eq. (6.2) are only valid when the generating function ψ is differentiable with a continuous first-order derivative, we shall restrict our attention to only this class of functions throughout this Chapter. Although it is possible to have a statistically valid divergence with not-so-smooth ψ functions (as in the case of FDPD; see Ray et al. (2023)), the interesting properties of the resulting minimum divergence estimators ensue only for continuously differentiable generating functions e.g., $\psi(x) = x, \psi(x) = \ln(x), \psi(x) = x^\gamma/\gamma$, etc.

6.3.1 GAB divergence for $\alpha + \beta = 1$

We consider the special case $\alpha + \beta = 1$ separately in this section. For this special case, the GAB divergence reduces to

$$d_{GAB}^{(\alpha,\beta),\psi}(f,g) = \begin{cases} \frac{1}{\alpha(1-\alpha)}\left(\psi(1) - \psi(\langle f,g\rangle_{\alpha,1-\alpha})\right) & \text{if } \alpha \notin \{0,1\}, \\ \psi'(1)\int f\ln(f/g) & \text{if } \alpha = 1, \beta = 0, \\ \psi'(1)\int g\ln(g/f) & \text{if } \alpha = 0, \beta = 1, \end{cases}$$

when we restrict the choices of f and g to be density functions. Based on the forms for the limiting cases, it clearly follows that it is a valid divergence if and only if $\psi'(1) > 0$, i.e., ψ is strictly increasing at 1. We shall show that this is indeed a necessary and sufficient condition for this special case.

Theorem 6.1. *For $\alpha + \beta = 1$, the generalized Alpha-Beta divergence is nonnegative if and only if the corresponding generating function ψ satisfies the following:*

1. *When* $\alpha \in (0,1)$, $\psi(x) < \psi(1)$ *for all* $x < 1$, *i.e.,* ψ *is increasing on the left-hand side of* 1.

2. *When* $\alpha < 0$ *or* $\alpha > 1$, $\psi(x) > \psi(1)$ *for any* $x > 1$, *i.e.,* ψ *is increasing on the right-hand side of* 1.

3. *When* $\alpha \in \{0,1\}$, ψ *is increasing at* 1.

Proof. The cases with $\alpha = 1, \beta = 0$ and $\beta = 1, \alpha = 0$ is immediate. Therefore, we focus on the case with $\alpha \notin \{0,1\}$. Now, because the Cressie-Reed power divergence is a valid divergence, it follows that for any $\alpha \notin \{0,1\}$ and any pair of densities f and g,

$$\frac{1}{\alpha(1-\alpha)} > \frac{1}{\alpha(1-\alpha)} \langle f, g \rangle_{\alpha,1-\alpha} .$$

We will consider two separate cases now.

Case 1 with $\alpha \in (0,1)$:

Assume that ψ is increasing on the left-hand side of 1. By the form of the power divergence, it follows that for any pair of densities f and g, we have $\langle f, g \rangle_{\alpha,1-\alpha} < 1$. Using the increasing behaviour of ψ, it follows that $\psi(\langle f, g \rangle_{\alpha,1-\alpha}) < \psi(1)$. The nonnegativity of the GAB divergence now follows by rearranging this inequality.

Conversely, assume that the GAB divergence is nonnegative. Since $\langle f, g \rangle_{\alpha,1-\alpha} \in (0,1]$, it is enough to show that for any $x \in (0,1)$, $\psi(x) < \psi(1)$. Consider the densities $f(x) = (\theta+1)^{-1}\mathbf{1}_{[0,\theta+1]}(x)$ and $g(x) = (\theta+1)^{-1}\mathbf{1}_{[1,\theta+2]}(x)$. The nonnegativity of the GAB divergence between the densities f and g in this case reduces to $\psi(\theta/(\theta+1)) < \psi(1)$. Letting $\theta = x/(1-x)$ now yields $\psi(x) < \psi(1)$ as we intended.

Case 2 with $\alpha < 0$ or $\alpha > 1$:

Assume that ψ is increasing on the right-hand side of 1. As in the previous case, it follows that for any pair of densities f and g, we have $\langle f, g \rangle_{\alpha,1-\alpha} > 1$, by nonnegativity of the power divergence. Since ψ is increasing at 1, we have $\psi(\langle f, g \rangle_{\alpha,1-\alpha}) > \psi(1)$. The nonnegativity of the GAB divergence now follows.

Conversely assume that the GAB divergence is nonnegative. As before, it is enough to show that for any $x > 1$, we have $\psi(x) > \psi(1)$. Let f and g be the density functions for Gaussian distributions with mean 0 and θ respectively with common variance equal to 1. It is easy to see that $\langle f, g \rangle_{\alpha,1-\alpha} = \exp(-\theta^2\alpha(1-\alpha)/2)$. As $\alpha \notin [0,1]$, we must have $\alpha(1-\alpha) < 0$. If we pick $\theta = \sqrt{-2x/(\alpha(1-\alpha))}$, then the nonnegativity of the GAB divergence between f and g yields the inequality $\psi(1) \leq \psi(x)$ as we wanted.

□

In general, when the class of nonnegative functions f and g are not restricted to be probability densities, but are any non-normalized density functions, the integrals $\int f$ and $\int g$ can be arbitrary. In such cases, the necessary and the sufficient condition presented in Theorem 6.1 reduces to the condition that ψ function must be monotonically increasing on $(0,\infty)$, instead on increasing at a single point $x = 1$.

6.3.2 Necessary and Sufficient Condition for nonnegativity of GABD for non-limiting cases

Now that we have established a characterization for the special case $\alpha + \beta = 1$ of the GAB divergence, we restrict our attention to the case when $\alpha + \beta \neq 1$. In this section, we additionally consider the case when α, β and $(\alpha + \beta)$ are all allowed to be nonzero.

Theorem 6.2. *Define $\Psi(x) := \psi(e^x)$ for all $x \in [0, \infty)$. If Ψ is strictly increasing and convex, then the GAB divergence generated by ψ is nonnegative for all choices of $\alpha, \beta, (\alpha + \beta) \neq 0$ such that $\alpha + \beta \neq 1$.*

Proof. Based on the signs of $\alpha\beta, \alpha(\alpha+\beta)$ and $\beta(\alpha+\beta)$, we deal with different cases separately using different versions of Hölder's inequality. Let, f and g be two non-normalized density functions.

Case 1: $\alpha\beta > 0, \alpha(\alpha + \beta) > 0, \beta(\alpha + \beta) > 0$**:**

By applying Hölder's inequality on $f^{\alpha+\beta}$ and $g^{\alpha+\beta}$, we have

$$\int f^\alpha g^\beta \leq \left(\int f^{\alpha+\beta}\right)^{\alpha/(\alpha+\beta)} \left(\int g^{\alpha+\beta}\right)^{\beta/(\alpha+\beta)}. \tag{6.7}$$

Since Ψ is convex, by applying Lemma 6.A.1, we have

$$\frac{\alpha}{\alpha+\beta}\psi\left(\|f\|_{\alpha+\beta}^{\alpha+\beta}\right) + \frac{\beta}{\alpha+\beta}\psi\left(\|g\|_{\alpha+\beta}^{\alpha+\beta}\right) \geq \psi\left(\|f\|_{\alpha+\beta}^{\alpha}\|g\|_{\alpha+\beta}^{\beta}\right) \geq \psi\left(\langle f, g\rangle_{\alpha,\beta}\right),$$

where the last inequality follows from the inequality (6.7) and the strictly increasing nature of ψ. Dividing both sides by $\alpha\beta$ now shows that the GAB divergence form as in (6.1) is nonnegative. Note that, the equality holds if and only if $f^{\alpha+\beta} = cg^{\alpha+\beta}$ for some constant $c > 0$.

Case 2: $\alpha\beta < 0, \alpha(\alpha + \beta) > 0, \beta(\alpha + \beta) < 0$**:**

Note that, $(\alpha+\beta)/\alpha > 0$ and $-\beta/\alpha > 0$ and they add up to 1. Therefore, by applying Hölder's inequality on $f^\alpha g^\beta$ and $g^{\alpha+\beta}$, we get

$$\left(\int f^\alpha g^\beta\right)^{(\alpha+\beta)/\alpha} \left(\int g^{\alpha+\beta}\right)^{-\beta/\alpha} \geq \int f^{\alpha+\beta}.$$

Using the increasingness and geometric-convexity of ψ, we obtain the chain of inequalities

$$\frac{\alpha+\beta}{\alpha}\psi\left(\langle f, g\rangle_{\alpha,\beta}\right) - \frac{\beta}{\alpha}\psi\left(\|g\|_{\alpha+\beta}^{\alpha+\beta}\right) \geq \psi\left(\left(\int f^\alpha g^\beta\right)^{(\alpha+\beta)/\alpha}\left(\int g^{\alpha+\beta}\right)^{-\beta/\alpha}\right) \geq \psi\left(\int f^{\alpha+\beta}\right).$$

Dividing both sides by $\beta(\alpha + \beta)$ yields,

$$\frac{1}{\alpha\beta}\psi\left(\langle f, g\rangle_{\alpha,\beta}\right) - \frac{1}{\alpha(\alpha+\beta)}\psi\left(\|g\|_{\alpha+\beta}^{\alpha+\beta}\right) \leq \frac{1}{\beta(\alpha+\beta)}\psi\left(\|f\|_{\alpha+\beta}^{\alpha+\beta}\right).$$

A rearrangement of the above shows that the GAB divergence between f and g is nonnegative. As in the previous case, it is easy to see that the equality holds if and only if $f = cg$ almost surely for some constant c.

Case 3: $\alpha\beta < 0, \alpha(\alpha + \beta) < 0, \beta(\alpha + \beta) > 0$**:** This follows from the previous case and the duality property of the GAB divergence as presented in (6.3). □

When $\alpha = 1$ and $\beta > 0$, then Theorem 6.2 reduces to the sufficient condition for the ψ function of the functional density power divergence (FDPD) class, as in the Proposition 4.1 of Ray et al. (2023). Thus, Theorem 6.2 is a generalization of this result to the family of generalized Alpha-Beta divergences. Similar to the case of FDPD, it turns out that the condition of the geometric convexity and monotonicity of the ψ function is also a necessary condition for the GAB divergence to be a valid statistical divergence.

Theorem 6.3. *If the GAB divergence with a generating function $\psi \in C^1((0,\infty))$ is nonnegative, then $\Psi(x) := \psi(e^x)$ must be strictly increasing and convex, for any choice of $\alpha \neq 0, \beta \neq 0, (\alpha+\beta) \neq 0$ and $(\alpha+\beta) \neq 1$.*

Proof. To establish the necessary condition, we proceed by carefully constructing a pair of nonnegative measures such that the nonnegativity of the GAB divergence between them yields the necessary increasing and convexity properties. As in the proof of Theorem 6.2, we split the proof by considering three scenarios separately depending on the signs of α and β.

Case 1, $\alpha\beta > 0, \alpha(\alpha+\beta) > 0$ and $\beta(\alpha+\beta) > 0$:

Increasing Property: To show that Ψ is strictly increasing, it is enough to show that ψ is strictly increasing. Take any $y > z > 0$. Consider the non-normalized densities $f(x) = \mathbf{1}_{[0,y]}(x)$ and $g(x) = \mathbf{1}_{[y-z,2y-z]}(x)$. Note that, $\|f\|_{\alpha+\beta}^{\alpha+\beta} = \|g\|_{\alpha+\beta}^{\alpha+\beta} = y$ and $\langle f, g\rangle_{\alpha,\beta} = y-(y-z) = z$. Since there does not exist a constant c such that $f = cg$, the corresponding GAB divergence between f and g must be strictly positive. Therefore, we have

$$\frac{1}{\beta(\alpha+\beta)}\psi(y) + \frac{1}{\alpha(\alpha+\beta)}\psi(y) - \frac{1}{\alpha\beta}\psi(z) > 0,$$

which can be rearranged as $\psi(y) > \psi(z)$. Since $y > z$ is arbitrary, this establishes the strict increasing nature of ψ.

Convexity Property: To show the convexity, consider again any $y > z > 0$. Let us consider the pair of non-normalized densities f and g given by

$$f(x) = C\theta^{-(\gamma+1)}x^{\gamma}\mathbf{1}_{(0,\theta)}(x),\ g(x) = C\eta^{-(\gamma+1)}x^{\gamma}\mathbf{1}_{(0,\eta)}(x), \tag{6.8}$$

where $C = (\gamma(\alpha+\beta)+1)^{1/(\alpha+\beta)}$ and $\gamma > (-1)$. Then it is easy to verify that

$$\begin{aligned}
\|f\|_{\alpha+\beta}^{\alpha+\beta} &= \theta^{-(\alpha+\beta)+1},\\
\|g\|_{\alpha+\beta}^{\alpha+\beta} &= \eta^{-(\alpha+\beta)+1},\\
\langle f, g\rangle_{\alpha,\beta} &= \begin{cases} \theta^{(\gamma+1)\beta+(1-\alpha-\beta)}\eta^{-(\gamma+1)\beta} & \text{if } \theta \leq \eta,\\ \theta^{-(\gamma+1)\alpha}\eta^{(\gamma+1)\alpha+(1-\alpha-\beta)} & \text{if } \theta > \eta, \end{cases}
\end{aligned}$$

where θ, η, γ are parameters to be suitably chosen. We take $\theta = e^{-y/(\alpha+\beta-1)}, \eta = e^{-z/(\alpha+\beta-1)}$.

Now if $\alpha+\beta > 1$, then $y > z$ implies that $\theta < \eta$. The nonnegativity of the GAB divergence in this case reduces to

$$\frac{1}{\beta(\alpha+\beta)}\Psi(y) + \frac{1}{\alpha(\alpha+\beta)}\Psi(z) > \frac{1}{\alpha\beta}\Psi\left(\left(1 - \frac{\beta(\gamma+1)}{\alpha+\beta-1}\right)y + \frac{\beta(\gamma+1)}{(\alpha+\beta-1)}z\right).$$

Picking $\gamma = -1/(\alpha+\beta)$ and multiplying both sides by $\alpha\beta$ we get that Ψ is $\alpha/(\alpha+\beta)$-convex on $\mathbb{R}$. Now an application of Lemma 1.2 of the Appendix A of Ray et al. (2023) along with the continuity of the ψ function completes the proof that Ψ is convex.

On the other hand, $\alpha+\beta<1$, we note that as $\beta/\alpha>0$ and $(1+\beta/\alpha)>1$. Letting $\alpha'=1$ and $\beta'=\beta/\alpha$ as new hyperparameters instead of α and β, and an application of the zooming property given in (6.6), now lead to the former case with $(\alpha+\beta)>1$.

Case 2, $\alpha\beta<0, \alpha(\alpha+\beta)>0$ and $\beta(\alpha+\beta)<0$:

Increasing Property: As in the previous case, to show the increasing nature of Ψ, it is enough to show that for any $y>z>0$, $\psi(y)>\psi(z)$. To see this, we consider the pair of standard normal densities

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/2\sigma^2}, \text{ and, } g(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\theta)^2/2\sigma^2}.$$

Elementary calculations yield that

$$\|f\|_{\alpha+\beta}^{\alpha+\beta}=\|g\|_{\alpha+\beta}^{\alpha+\beta}=C,\ \langle f,g\rangle_{\alpha,\beta}=C\exp\left(-\frac{\alpha\beta\theta^2}{2(\alpha+\beta)\sigma^2}\right),$$

where $C=(2\pi)^{-(\alpha+\beta-1)/2}\sigma^{-(\alpha+\beta-1)}(\alpha+\beta)^{-1/2}$. If $(\alpha+\beta)\in(0,1)$, then we choose

$$\sigma=(2\pi)^{-1/2}(\alpha+\beta)^{-1/2(\alpha+\beta-1)}z^{-1/(\alpha+\beta-1)}, \text{ and, } \theta=\sqrt{-2(\alpha+\beta)/\alpha\beta}(\ln(y/z))^{1/2}\sigma.$$

With this specific choice, the nonnegativity of the GAB divergence leads to the inequality

$$\frac{1}{\beta(\alpha+\beta)}\psi(z)-\frac{1}{\alpha\beta}\psi(y)+\frac{1}{\alpha(\alpha+\beta)}\psi(z)>0.$$

Multiplying both sides by $\alpha\beta$ and simple rearrangement yields $\psi(y)>\psi(z)$, as we wanted. This proves the increasingness of Ψ.

If $\alpha+\beta<0$ or $\alpha+\beta>1$, then note that we have $\beta/\alpha<0$ and $(\alpha+\beta)/\alpha=(1+\beta/\alpha)\in(0,1)$. Therefore, as before, an application of the zooming property (6.6) reduces it to the former subcase with $(\alpha+\beta)\in(0,1)$ with new hyperparameters $\alpha'=\beta/\alpha$ and $\beta'=(\alpha+\beta)/\alpha$.

Convexity Property: To show that Ψ is convex, let us consider two real numbers $y>z>0$. We consider the pair of non-normalized densities given in (6.8) again. Since $\alpha\beta<0$, at least one of α and β must be positive. Without the loss of generality, assume that $\alpha>0$. Then we take $\eta=e^{-y/(\alpha+\beta-1)}$ and $\theta=\eta e^{(y-z)/(\gamma+1)\alpha}$. For this specific choice of the parameters, the nonnegativity of the GAB divergence yields

$$-\frac{\beta}{\alpha}\Psi(y)+\left(\frac{\alpha+\beta}{\alpha}\right)\Psi(z)\geq\Psi\left(\left(1-\frac{\alpha+\beta-1}{(\gamma+1)\alpha}\right)y+\frac{\alpha+\beta-1}{(\gamma+1)\alpha}z\right).$$

If $\alpha+\beta>1$, then taking $\gamma=-1/(\alpha+\beta)$ shows that Ψ is $(-\beta/\alpha)$-convex on $\mathbb{R}$. We can now apply Lemma 1.2 of the Appendix A of Ray et al. (2023) along with the continuity of the ψ function to complete the proof that Ψ is convex for $(\alpha+\beta)>1$ case.

When $\alpha+\beta<1$, we must have $(\alpha+\beta)\in(0,1)$ since $\alpha(\alpha+\beta)>0$. Hence, there exists a sufficiently large but fixed $M>0$ such that $M(\alpha+\beta)>1$. Now an application of

the zooming property (6.5), reduces it to the former case with $\alpha' + \beta' > 1$ where the new hyperparameters are $\alpha' = M\alpha$ and $\beta' = M\beta$.

Case 3, $\alpha\beta < 0, \alpha(\alpha+\beta) < 0$ **and** $\beta(\alpha+\beta) > 0$**:** This case follows from the previous case and is a direct application of the duality property of GAB divergence as shown in (6.3). □

6.3.3 Necessary and Sufficient Condition for nonnegativity of GABD for the limiting cases

Before we proceed further to establish similar necessary and sufficient conditions for the nonnegativity of the GAB divergence for either $\alpha = 0$ or $\beta = 0$ cases, we rewrite the form of GAB divergence to show its connection to the popular Kullback Leibler (KL) divergence. Note that,

$$\begin{aligned}\int f^\alpha \ln(f^\alpha/g^\alpha) &= \|f\|_\alpha^\alpha \int f^{[\alpha]} \ln\left(f^{[\alpha]}\|f\|_\alpha^\alpha / g^{[\alpha]}\|g\|_\alpha^\alpha\right) \\ &= \|f\|_\alpha^\alpha \left[d_{KL}(f^{[\alpha]}, g^{[\alpha]}) + \ln(\|f\|_\alpha^\alpha) - \ln(\|g\|_\alpha^\alpha)\right],\end{aligned}$$

where $d_{KL}(f^{[\alpha]}, g^{[\alpha]})$ is the KL divergence between the α-zoomed versions of f and g, i.e., the densities proportional to f^α and g^α. These densities will exist when both f and g are L^α-integrable, which we assume to be the case. Therefore, we can rewrite the GAB divergence for the $\alpha \neq 0, \beta = 0$ case as

$$\begin{aligned}d_{GAB}^{(\alpha,0),\psi} = \frac{1}{\alpha^2}\Big[\psi'(\|f\|_\alpha^\alpha)\|f\|_\alpha^\alpha \left(d_{KL}(f^{[\alpha]}, g^{[\alpha]}) + \ln(\|f\|_\alpha^\alpha) - \ln(\|g\|_\alpha^\alpha)\right) \\ -\psi(\|f\|_\alpha^\alpha) + \psi(\|g\|_\alpha^\alpha)\Big]. \quad (6.9)\end{aligned}$$

Similarly, for $\alpha = 0, \beta \neq 0$ case, we get

$$\begin{aligned}d_{GAB}^{(0,\beta),\psi} = \frac{1}{\beta^2}\Big[\psi'(\|g\|_\beta^\beta)\|g\|_\beta^\beta \left(d_{KL}(g^{[\beta]}, f^{[\beta]}) + \ln(\|g\|_\beta^\beta) - \ln(\|f\|_\beta^\beta)\right) \\ -\psi(\|g\|_\beta^\beta) + \psi(\|f\|_\beta^\beta)\Big]. \quad (6.10)\end{aligned}$$

This connection demonstrates that the GAB divergence for the cases with either $\alpha = 0, \beta \neq 0$ or $\alpha \neq 0, \beta = 0$ can be visualized as a localized affine transformation of the KL divergence between the zoomed versions of the corresponding densities. When $\alpha = 0$ and $\beta \neq 0$, if $\beta > 0$ holds, then the corresponding tails of these zoomed densities are down-weighted. This results in comparatively higher robustness than the usual minimum KL-divergence estimators, i.e., the maximum likelihood estimator (MLE).

In the following theorem, we present a sufficient condition for the GAB divergence to be valid for these two cases, which follows easily from Eq. (6.10)-(6.9).

Theorem 6.4. *Define* $\Psi(x) := \psi(e^x)$ *for all* $x \in \mathbb{R}$*. If* Ψ *is strictly increasing and is convex for all* $x \in \mathbb{R}$*, then the GAB divergence generated by* ψ *is nonnegative when either* $\alpha \neq 0, \beta = 0$ *or* $\alpha = 0, \beta \neq 0$*.*

Proof. We shall show the proof only for the case when $\alpha \neq 0, \beta = 0$, as the case for $\alpha = 0, \beta \neq 0$ would then follow from the duality property shown in (6.3).

Let f and g be two non-normalized densities that are L^α-integrable. Denoting $x = \ln(\|f\|_\alpha^\alpha)$ and $y = \ln(\|g\|_\alpha^\alpha)$, in the GAB divergence given in (6.9), we obtain

$$\alpha^2 d_{GAB}^{(\alpha,0),\psi} = \Psi'(x) d_{KL}(f^{[\alpha]}, g^{[\alpha]}) + \big(\Psi(y) - \Psi(x) - \Psi'(x)(y-x)\big).$$

Since, $\Psi'(x) > 0$ due to the increasing nature of Ψ and the KL-divergence is nonnegative, it is enough to show that $\Psi(y) - \Psi(x) - \Psi'(x)(y-x)$ is nonnegative. It now follows from the fact that Ψ is convex and $\Psi'(x)$ is a subgradient of Ψ at x. In particular, there can be three cases. If $y = x$, then the right-hand side is 0, hence nonnegative. If $y > x$, then by exploiting the convexity and increasing property of Ψ, we get $\Psi'(x) \leq \Psi'(y)$ and $\Psi(x) \leq \Psi(y)$. So, by Lagrange's Mean Value theorem, there exists $z \in (x, y)$ such that

$$\Psi'(x) \leq \Psi'(z) = \frac{\Psi(y) - \Psi(x)}{y - x} \leq \Psi'(y).$$

For the other case when $y < x$, reversing the roles of x and y completes the proof. □

As the cases $\alpha = 1, \beta = 0$ or $\alpha = 0, \beta = 1$ are already discussed in Section 6.3.1, while developing the necessary conditions below, we disregard these two specific cases.

Theorem 6.5. *Suppose either $\alpha \notin \{0, 1\}, \beta = 0$ or $\alpha = 0, \beta \notin \{0, 1\}$ holds. If the GAB divergence with generating function $\psi \in C^1((0, \infty))$ is nonnegative, then $\Psi(x) := \psi(e^x)$ must be strictly increasing and convex.*

Proof. As before, we shall show the proof only for the case when $\alpha \neq 0, \beta = 0$. The case for $\alpha = 0, \beta \neq 0$ would then follow from the duality property of the GAB divergence as in (6.3).

Increasing Property: Since $\psi \in C^1((0, \infty))$, to show that Ψ is strictly increasing, it is enough to show that for any $y > 0$, $\psi'(y) > 0$. Consider the pair of Gaussian densities f and g as,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}, \; g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\theta)^2/2\sigma^2}, \; x \in \mathbb{R}.$$

It follows that $\|f\|_\alpha^\alpha = \|g\|_\alpha^\alpha = (2\pi)^{-(\alpha-1)/2} \sigma^{-(\alpha-1)} \alpha^{-1/2}$. Since $\alpha \neq 1$, we may take

$$\sigma = (2\pi)^{-1/2} (\alpha)^{-1/2(\alpha-1)} y^{-1/(\alpha-1)},$$

which is well-defined and nonnegative. The nonnegativity of the GAB divergence as in (6.9) now reduces to

$$\psi'(y) y d_{KL}(f^{[\alpha]}, g^{[\alpha]}) - \psi(y) + \psi(y) > 0,$$

and since the KL-divergence is nonnegative, it implies that $\psi'(y) > 0$, as we wanted to show.

Convexity Property: Moving on to the convexity condition, let us pick any $y > z \geq 0$. We consider again the family of non-normalized densities f and g given in (6.8). If $\alpha > 1$,

then we pick $\theta = e^{-y/(\alpha-1)}$ and $\eta = e^{-z/(\alpha-1)}$. As $\theta < \eta$, elementary calculations show that

$$\int f^\alpha \ln(f/g) = (\gamma+1)(\ln(\theta) - \ln(\eta))\theta^{-(\alpha-1)} = (\gamma+1)e^y(z-y)/(\alpha-1).$$

Therefore, the nonnegativity of the GAB divergence translates to the inequality

$$\frac{(\gamma+1)\alpha}{(\alpha-1)}\psi'(e^y)e^y(z-y) - \psi(e^y) + \psi(e^z) \geq 0.$$

Choosing $\gamma = -1/\alpha$ yields $\Psi'(y)(y-z)-(\Psi(y)-\Psi(z)) \geq 0$ for any $y > z$, which establishes the convexity of Ψ function.

When $\alpha \in (0,1)$, we choose $M > 0$ sufficiently large such that $M\alpha > 1$. With the new hyperparameter $\alpha' = M\alpha$ and an application of the zooming property given in (6.5) reduces it to the former case with $\alpha > 1$. This takes care of all the cases with positive α.

If $\alpha < 0$, we again appeal to the zooming property given in (6.5) with $c = (-1)$. As the zooming property does not change the ψ function, the same argument as provided above can be applied. □

6.4 Asymptotic Breakdown Results

Now that we have characterized the generalized Alpha-Beta divergence, we shall move on to establishing its robustness properties, in particular, the asymptotic breakdown point of the corresponding minimum divergence estimator. Given a family of model densities $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ and a true density g with distribution function G, we define the minimum generalized Alpha-Beta divergence (MGABD) functional as

$$T^\psi_{(\alpha,\beta)}(G) := \arg\min_{\theta\in\Theta} d^{(\alpha,\beta),\psi}_{GAB}(f_\theta, g),$$

provided that the minimum exists.

As in Chapter 5, we shall consider ϵ-contaminated densities $g_{\epsilon,m} = (1-\epsilon)g + \epsilon k_m$ with distribution functions $G_{\epsilon,m}$ for a sequence of contaminating densities k_m with corresponding distribution functions K_m for $m = 1, 2, \ldots$. To investigate the asymptotic breakdown point of the minimum GABD functional $T^\psi_{(\alpha,\beta)}(G)$, we shall look into how $T^\psi_{(\alpha,\beta)}(G_{\epsilon,m})$ changes as a function of ϵ as m tends to infinity. The main theorem of this chapter follows an approach similar to that of Chapter 5 of this thesis and Park and Basu (2004). The key assumptions in this chapter remain the same as before, namely (BP1)-(BP3), with the exception that Assumption (BP4) must be modified as the following to accommodate both hyperparameters α and β.

(BP5) If $\beta > 0$, the model family and the family of contaminating densities are uniformly $L^{\alpha+\beta}$-integrable, i.e.,

$$\limsup_{m\to\infty} \int k_m^{\alpha+\beta} < \infty, \text{ and, } \sup_{\theta\in\Theta} \int f_\theta^{\alpha+\beta} < \infty,$$

For $\beta = 0$, these families of densities are uniformly $L^{\alpha+\beta+\delta}$-integrable for some $\delta > 0$, and additionally, the integrals

$$\sup_{\theta\in\Theta}\left|\int f_\theta^\alpha \ln(g)\right|, \text{ and, } \sup_{\theta\in\Theta}\sup_{m}\left|\int f_\theta^\alpha \ln(k_m)\right|,$$

exist and are finite.

Furthermore, though the hyperparameters α, β can take any real values, we consider only the case when both of them are nonnegative. Also, the true density g is expected to be $L^{\alpha+\beta}$ integrable for $\alpha, \beta > 0$ case and $L^{\alpha+\beta+\delta}$-integrable for some $\delta > 0$ whenever either $\alpha = 0$ or $\beta = 0$, as in the case of Assumption (BP5) for f_θ and k_m.

6.4.1 Asymptotic Breakdown Point of MGABD functional for positive α and β

We first proceed with exploring the behaviour of the asymptotic breakdown point of the MGABD functional for positive α and β. The following theorem establishes the general result as in Theorem 5.1, but with a sufficient condition which is modified appropriately to accommodate the general nature of ψ function.

Theorem 6.6. *Let us assume that (BP1)-(BP3) and (BP5) hold. Also, assume that for $\alpha > 0$ and $\beta > 0$, there exists $\tilde{\epsilon} \in [0, 1/2]$ such that for all $\epsilon < \tilde{\epsilon}^{\psi}_{(\alpha,\beta)}$ there exists $\theta^g \in \Theta$ (which may depend on ϵ) such that for sufficiently large m, we have*

$$\psi\left((1-\epsilon)^\beta \langle f_{\theta^g}, g\rangle_{\alpha,\beta}\right) - \psi\left(\epsilon^\beta \langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}\right) \geq \frac{\alpha}{\alpha+\beta}\left(\psi\left(\|f_{\theta^g}\|_{\alpha+\beta}^{\alpha+\beta}\right) - \psi\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta}\right)\right), \tag{6.11}$$

whenever $\theta_m \to \theta_\infty$ for some $\theta_\infty \in \partial\Theta$ and for all contaminating densities $\{k_m\}$. Then, the asymptotic breakdown point of the minimum generalized Alpha-Beta divergence (MGABD) functional with a continuous ψ-function is at least $\tilde{\epsilon}^{\psi}_{(\alpha,\beta)}$.

Proof. Let $\epsilon < \tilde{\epsilon}^{\psi}_{(\alpha,\beta)}$ be a level of contamination where the breakdown occurs. This means, there exists a sequence of contaminating densities $\{k_m\}$ such that for the corresponding ϵ-contaminated densities $g_{\epsilon,m} = (1-\epsilon)g + \epsilon k_m$, the MGABD functional $\theta_m = T^{\psi}_{(\alpha,\beta)}(G_{\epsilon,m})$ satisfies $\theta_m \to \theta_\infty$ for some $\theta_\infty \in \partial\Theta$.

The proof now follows exactly the same steps as the proof of Theorem 5.1. In Step 1, we find an asymptotic lower bound of the GABD between f_{θ_m} and $g_{\epsilon,m}$. Next, for a fixed θ^g as provided by condition (6.11), we find an asymptotic upper bound of the GABD between f_{θ^g} and $g_{\epsilon,m}$. In the final step, we compare these bounds by using the minimality of GABD at θ_m and aim to recover a bound on the ϵ from it.

Step 1:

Consider the set $A_m = \{x : g(x) > \max\{k_m(x), f_{\theta_m}(x)\}\}$. Due to Assumptions (BP1)-(BP3) and (BP5), similar to the proof of Theorem 5.1, we obtain that

$$\int_{A_m} f_{\theta_m}^\alpha g_{\epsilon,m}^\beta \asymp 0,$$
$$\int_{A_m^c} f_{\theta_m}^\alpha g_{\epsilon,m}^\beta \asymp \epsilon^\beta \langle f_{\theta_m}, k_m\rangle_{\alpha,\beta},$$

as $m \to \infty$, where $\asymp$ denotes asymptotic equivalence of the two sides as $m \to \infty$.

Due to the continuity of ψ-function, it follows that

$$\begin{aligned}
&d_{GAB}^{(\alpha,\beta),\psi}(f_{\theta_m}, g_{\epsilon,m}) \\
=&\frac{1}{\beta(\alpha+\beta)}\psi\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta}\right) - \frac{1}{\alpha\beta}\psi\left(\langle f_{\theta_m}, g_{\epsilon,m}\rangle_{\alpha,\beta}\right) + \frac{1}{\alpha(\alpha+\beta)}\psi\left(\|g_{\epsilon,m}\|_{\alpha+\beta}^{\alpha+\beta}\right) \\
\asymp&\frac{\psi\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta}\right)}{\beta(\alpha+\beta)} - \frac{\psi\left(\epsilon^{\beta}\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}\right)}{\alpha\beta} + \frac{\psi\left(\|g_{\epsilon,m}\|_{\alpha+\beta}^{\alpha+\beta}\right)}{\alpha(\alpha+\beta)}.
\end{aligned} \tag{6.12}$$

We denote the right-hand side of Eq. (6.12) by $\delta_m(\epsilon)$.

Step 2:

Since $\beta > 0$, we have the inequality

$$\langle f_{\theta^g}, g_{\epsilon,m}\rangle_{\alpha,\beta} = \int f_{\theta^g}^{\alpha}((1-\epsilon)g + \epsilon k_m)^{\beta} \geq (1-\epsilon)^{\beta}\int f_{\theta^g}^{\alpha} g^{\beta} = (1-\epsilon)^{\beta}\langle f_{\theta^g}, g\rangle_{\alpha,\beta}.$$

Therefore, we have

$$\begin{aligned}
&d_{GAB}^{(\alpha,\beta),\psi}(f_{\theta^g}, g_{\epsilon,m}) \\
=&\frac{1}{\beta(\alpha+\beta)}\psi\left(\|f_{\theta^g}\|_{\alpha+\beta}^{\alpha+\beta}\right) - \frac{1}{\alpha\beta}\psi\left(\langle f_{\theta^g}, g_{\epsilon,m}\rangle_{\alpha,\beta}\right) + \frac{1}{\alpha(\alpha+\beta)}\psi\left(\|g_{\epsilon,m}\|_{\alpha+\beta}^{\alpha+\beta}\right) \\
\leq&\frac{1}{\beta(\alpha+\beta)}\psi\left(\|f_{\theta^g}\|_{\alpha+\beta}^{\alpha+\beta}\right) - \frac{1}{\alpha\beta}\psi\left((1-\epsilon)^{\beta}\langle f_{\theta^g}, g\rangle_{\alpha,\beta}\right) + \frac{1}{\alpha(\alpha+\beta)}\psi\left(\|g_{\epsilon,m}\|_{\alpha+\beta}^{\alpha+\beta}\right).
\end{aligned} \tag{6.13}$$

Here, we make use of the strictly increasing behaviour of the ψ function. We denote the right-hand side of Eq. (6.13) by $\Delta_m(\epsilon)$.

Step 3:

Now, we will have a contradiction to the assumption of breakdown at ϵ, if, for all sufficiently large m, we have

$$\Delta_m(\epsilon) < \delta_m(\epsilon). \tag{6.14}$$

This comparison leads to the inequality

$$\begin{aligned}
&\frac{1}{\beta(\alpha+\beta)}\psi\left(\|f_{\theta^g}\|_{\alpha+\beta}^{\alpha+\beta}\right) - \frac{1}{\alpha\beta}\psi\left((1-\epsilon)^{\beta}\langle f_{\theta^g}, g\rangle_{\alpha,\beta}\right) \\
&\qquad < \frac{1}{\beta(\alpha+\beta)}\psi\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta}\right) - \frac{1}{\alpha\beta}\psi\left(\epsilon^{\beta}\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}\right) \\
\iff &\psi\left((1-\epsilon)^{\beta}\langle f_{\theta^g}, g\rangle_{\alpha,\beta}\right) - \psi\left(\epsilon^{\beta}\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}\right) \\
&\qquad > \frac{\alpha}{\alpha+\beta}\left(\psi\left(\|f_{\theta^g}\|_{\alpha+\beta}^{\alpha+\beta}\right) - \psi\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta}\right)\right).
\end{aligned}$$

However, according to the condition (6.11), the above inequality is satisfied for the specific choice of ϵ (since $\epsilon < \tilde{\epsilon}_{(\alpha,\beta)}^{\psi}$) and the choice of θ^g depending on ϵ. Therefore, we have a contradiction, which means the MGABD functional does not asymptotically break down at ϵ. Since $\epsilon < \tilde{\epsilon}_{(\alpha,\beta)}^{\psi}$ is arbitrary, the result follows. □

In Theorem 5.1, we have found asymptotic breakdown point of the minimum super divergence (MSD) functional under a similar condition given in Eq. (5.2), for the specific

case when the true density g belongs to the model family of densities $\mathcal{F} := \{f_\theta : \theta \in \Theta\}$. This is simply a special case of Theorem 6.6 with the choice of $\psi(x) = x$, the identity function. To see this, note that for this choice of ψ and as $g = f_{\theta^g}$, the condition (6.11) reduces to

$$(1-\epsilon)^\beta \int g^{\alpha+\beta} - \epsilon^\beta \int f_{\theta_m}^\alpha k_m^\beta > \frac{\alpha}{\alpha+\beta}\int g^{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}\int f_{\theta_m}^{\alpha+\beta}.$$

Letting $A = \beta$ and $B = \alpha = (1 + a - A)$, the S-divergence between ϵk_m and f_{θ_m} satisfy

$$\begin{aligned}
& S_{(a,\lambda)}(\epsilon k_m, f_{\theta_m}) \\
= \;& \frac{1}{A}\int f_{\theta_m}^{1+a} - \frac{1+a}{AB}\epsilon^A \int f_{\theta_m}^B k_m^A + \frac{1}{B}\int k_m^{1+a} \\
> \;& \frac{1}{A}\int f_{\theta_m}^{1+a} + \frac{1}{B}\int k_m^{1+a} - \frac{1+a}{AB}\left[\left(\frac{B}{A+B} - (1-\epsilon)^A\right)\int g^{A+B} - \frac{B}{A+B}\int f_{\theta_m}^{A+B}\right] \\
= \;& \frac{1}{B}\int k_m^{1+a} + \left[\frac{1}{A} - \frac{1+a}{AB}(1-\epsilon)^A \int g^{1+a}\right],
\end{aligned}$$

where λ is such that $A = \beta = 1 + \lambda(1-a)$. Taking limit infimum to both sides of this inequality with respect to m translates to the condition (5.2).

The condition (6.11) is usually difficult to verify directly in practice. However, there are some alternative stronger conditions including specific classes of the ψ-functions and specific parametric setups, which are sufficient to imply that the condition (6.11) holds. In the following discussion, we explore these alternate conditions.

Firstly, we note that due to Assumption (BP5), there must exist a large constant $C > 0$ such that $\|k_m\|_{\alpha+\beta} \leq C$ for all sufficiently large m. This is the weakest condition that can be imposed to obtain a lower bound on the asymptotic breakdown point of the MGABD functional. However, this does not produce the lower bound explicitly, but as an implicit solution to a nonlinear equation, depending on the choice of ψ-function.

Corollary 6.1. *Suppose that $\alpha, \beta > 0, (\alpha+\beta) \neq 1$ and the Assumptions (BP1)-(BP3) and (BP5) hold. Let $C = \limsup_{m\to\infty} \|k_m\|_{\alpha+\beta}$. Then, the asymptotic breakdown point of the MGABD functional for any valid continuous ψ-function is at least $\min\{1/2, \epsilon^*_{(\alpha,\beta)}\}$, where $\epsilon^*_{(\alpha,\beta)}$ is a solution of*

$$\psi\left(\langle f_{\theta^g}, (1-\epsilon)g\rangle_{\alpha,\beta}\right) - \frac{\alpha}{\alpha+\beta}\psi\left(\|f_{\theta^g}\|_{\alpha+\beta}^{\alpha+\beta}\right) - \frac{\beta}{\alpha+\beta}\psi\left((C\epsilon)^{\alpha+\beta}\right) = 0, \tag{6.15}$$

in the interval $[0,1]$ provided it exists. If the solution does not exist, we take $\epsilon^*_{(\alpha,\beta)}$ as 0.

Proof. First note that, since the resulting GAB divergence is well-defined and nonnegative, an application of Theorem 6.3 establishes that ψ is increasing and geometrically-convex. Therefore,

$$\begin{aligned}
\psi\left(\langle f_{\theta_m}, \epsilon k_m\rangle_{\alpha,\beta}\right) &\leq \psi\left(\|f_{\theta_m}\|_{\alpha+\beta}^\alpha \|\epsilon k_m\|_{\alpha+\beta}^\beta\right), \text{by Hölder's inequality} \\
&\leq \frac{\alpha}{\alpha+\beta}\psi\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta}\right) + \frac{\beta}{\alpha+\beta}\psi\left(\|\epsilon k_m\|_{\alpha+\beta}^{\alpha+\beta}\right), \text{ by geometric convexity.}
\end{aligned}$$

Continuing, we get

$$\psi\left(\langle f_{\theta_m}, \epsilon k_m\rangle_{\alpha,\beta}\right) - \frac{\alpha}{\alpha+\beta}\psi\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta}\right) \leq \frac{\beta}{\alpha+\beta}\psi\left(\|\epsilon k_m\|_{\alpha+\beta}^{\alpha+\beta}\right) \leq \frac{\beta}{\alpha+\beta}\psi((C\epsilon)^{\alpha+\beta}),$$

where the last inequality follows from the increasing nature of ψ function.
In view of the condition (6.11), it is therefore enough to show that for all $\epsilon < \epsilon^*_{(\alpha,\beta)}$, the following inequality holds

$$\psi\left(\langle f_{\theta^g}, (1-\epsilon)g\rangle_{\alpha,\beta}\right) - \frac{\alpha}{\alpha+\beta}\psi\left(\|f_{\theta^g}\|_{\alpha+\beta}^{\alpha+\beta}\right) - \frac{\beta}{\alpha+\beta}\psi\left((C\epsilon)^{\alpha+\beta}\right) > 0.$$

Clearly, when $\epsilon = \epsilon^*_{(\alpha,\beta)}$, the left-hand side of the above inequality is 0. Since ψ is strictly increasing and $\langle f_{\theta^g}, (1-\epsilon)g\rangle_{\alpha,\beta} = (1-\epsilon)^\beta \langle f_{\theta_g}, g\rangle_{\alpha,\beta}$ is decreasing function of ϵ, the first term in left-hand side is strictly decreasing in ϵ in $[0,1]$. Since $C \geq 0$, $\psi\left((C\epsilon)^{\alpha+\beta}\right)$ is increasing in ϵ. Hence, the left-hand side is strictly decreasing in ϵ. Hence, for all $\epsilon < \epsilon^*_{(\alpha,\beta)}$ the left-hand side of the above inequality must be positive. □

In general, Eq. (6.15) may not have a solution in $[0,1]$. However, if ψ is continuous, the range of ψ is a subset of nonnegative real numbers, and, the true density g belongs to the model family $\mathcal{F}$, i.e., $f_{\theta^g} = g$, then the equation must have a solution. To see this, note that when $\epsilon = 0$, the left-hand side of the equation is $\frac{\beta}{\alpha+\beta}\psi(\|g\|_{\alpha+\beta}^{\alpha+\beta}) > 0$. On the other hand, when $\epsilon = 1$, the left-hand side of the equation is $-\frac{\alpha}{\alpha+\beta}\psi(\|g\|_{\alpha+\beta}^{\alpha+\beta}) - \frac{\beta}{\alpha+\beta}\psi(C^{\alpha+\beta}) < 0$. Now that Eq. (6.15) has a solution follows from an application of the intermediate value theorem. This also means that when $\psi(x) = \ln(x)$, i.e., in the case of the minimum logarithmic S-divergence (MLSD) functional, Corollary 6.1 only provides the trivial bound to the asymptotic breakdown point.

When the inference problem under consideration deals with estimating a location parameter in a location model family, we can explicitly obtain a lower bound to the asymptotic breakdown point, unlike Corollary 6.1.

Corollary 6.2. *If the model family $\{f_\theta\}$ is a location family, θ is the location parameter to be estimated and the true density g and the sequence of contaminating densities $\{k_m\}$ all belong to the same location family, then under Assumptions (BP1)-(BP3), the asymptotic breakdown point of MGABD functional for any valid continuous ψ function is equal to $1/2$, for any $\alpha, \beta > 0$.*

Proof. We start by noting that due to the consideration of the location family, $\|f_\theta\|_{\alpha+\beta}^{\alpha+\beta} = c$ is independent of $\theta \in \Theta$. Also, since $g \in \{f_\theta : \theta \in \Theta\}$, the density at the best fitting parameter satisfy $f_{\theta^g} = g$. Clearly, $\|k_m\|_{\alpha+\beta}^{\alpha+\beta}$ also equals to c in this case. As a result of these, Assumption (BP5) is automatically satisfied.

Hence, in condition (6.11), the RHS of the inequality becomes equal to 0. Also, the strictly increasing nature of ψ along with its continuity implies that ψ is one-one, and its inverse function is also strictly increasing. Hence, the inequality in (6.11) reduces to

$$(1-\epsilon)^\beta \int f_{\theta_g}^\alpha g^\beta \geq \epsilon^\beta \int f_{\theta_m}^\alpha k_m^\beta.$$

Now, it follows by Hölder's inequality that

$$\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta} \leq \|f_{\theta_m}\|_{\alpha+\beta}^{\alpha} \|k_m\|_{\alpha+\beta}^{\beta} = c = \|f_{\theta^g}\|_{\alpha+\beta}^{\alpha+\beta} = \langle f_{\theta^g}, g\rangle_{\alpha,\beta}\,.$$

Therefore, to ensure the inequality in condition (6.11), it is enough to ensure $(1-\epsilon)^{\beta} \geq \epsilon^{\beta}$, i.e, $\epsilon < 1/2$. Thus, choosing $\tilde{\epsilon}^{\psi}_{(\alpha,\beta)} = 1/2$ satisfies condition (6.11), and as a result, the MGABD functional in this case has an asymptotic breakdown point equal to 1/2. □

Note that, the above Corollary 6.2 generalizes the result of the MSD functional for the location estimation case to general ψ-functions and also extending the range of hyperparameters to $\alpha + \beta \in (0, 1)$.

While the above Corollary 6.2 restricts the parametric setup under consideration, one can alternatively restrict the choice of the divergences by restricting the generating function ψ to be of a particular form, namely the ones proposed by Jones et al. (2001). The corresponding result is indicated below.

Corollary 6.3. *Suppose that $\alpha, \beta > 0$ and the Assumptions (BP1)-(BP3) and (BP5) hold. Additionally, assume that the sequence of contaminating densities is such that $\|k_m\|_{\alpha+\beta} \leq \|f_{\theta_m}\|_{\alpha+\beta}$ for sufficiently large m. If $\psi(x) = \gamma^{-1}x^{\gamma}$ for some $\gamma > 0$, then the asymptotic breakdown point of the corresponding MGABD functional is at least*

$$\min\left\{\left(\frac{\alpha}{\alpha+\beta}\right)^{1/\beta\gamma}, 1 - \left(\frac{\alpha}{\alpha+\beta}\right)^{1/\beta\gamma} \frac{\|f_{\theta^g}\|_{\alpha+\beta}^{\alpha+\beta}}{\langle f_{\theta^g}, g\rangle_{\alpha,\beta}}, \frac{1}{2}\right\}$$

Proof. Let

$$\begin{aligned}
\tilde{\epsilon}^{\psi}_{(\alpha,\beta)} &= \min\left\{\liminf_{m\to\infty}\left[\frac{\psi^{-1}\left(\frac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta})\right)}{\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}}\right]^{1/\beta}, 1 - \left[\frac{\psi^{-1}\left(\frac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_g}\|_{\alpha+\beta}^{\alpha+\beta})\right)}{\langle f_{\theta_g}, g\rangle_{\alpha,\beta}}\right]^{1/\beta}\right\} \\
&= \liminf_{m\to\infty} \min\left\{\left[\frac{\psi^{-1}\left(\frac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta})\right)}{\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}}\right]^{1/\beta}, 1 - \left[\frac{\psi^{-1}\left(\frac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_g}\|_{\alpha+\beta}^{\alpha+\beta})\right)}{\langle f_{\theta_g}, g\rangle_{\alpha,\beta}}\right]^{1/\beta}\right\}
\end{aligned}$$

Note that the above equality holds due to Lemma 6.A.2. For every $\epsilon < \tilde{\epsilon}$, for sufficiently large enough m, we must have

$$\epsilon < \min\left\{\left[\frac{\psi^{-1}\left(\frac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta})\right)}{\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}}\right]^{1/\beta}, 1 - \left[\frac{\psi^{-1}\left(\frac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_g}\|_{\alpha+\beta}^{\alpha+\beta})\right)}{\langle f_{\theta_g}, g\rangle_{\alpha,\beta}}\right]^{1/\beta}\right\}.$$

This implies that

$$\begin{aligned}
&\epsilon < \left[\frac{\psi^{-1}\left(\frac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta})\right)}{\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}}\right]^{1/\beta}, \\
\implies &\psi(\epsilon^{\beta}\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}) < \frac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta}), \\
\implies &-\psi(\epsilon^{\beta}\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}) \geq -\frac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta}). \qquad (6.16)
\end{aligned}$$

Similarly, we also get

$$\epsilon < 1 - \left[\frac{\psi^{-1}\left(\frac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_g}\|_{\alpha+\beta}^{\alpha+\beta})\right)}{\langle f_{\theta_g}, g\rangle_{\alpha,\beta}}\right]^{1/\beta},$$
$$\implies \psi((1-\epsilon)^{\beta}\langle f_{\theta_g}, g\rangle_{\alpha,\beta}) > \frac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_g}\|_{\alpha+\beta}^{\alpha+\beta}). \tag{6.17}$$

Combining the inequalities in (6.16) and (6.17), we get the condition (6.11).

For $\psi(x) = \frac{1}{\gamma}x^{\gamma}$, we obtain that

$$\begin{aligned}
&\left[\frac{\psi^{-1}\left(\frac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta})\right)}{\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}}\right]^{1/\beta} \\
&= \left(\frac{\alpha}{\alpha+\beta}\right)^{1/\beta\gamma}\left[\frac{\|f_{\theta_m}\|_{\alpha+\beta}^{(\alpha+\beta)}}{\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}}\right]^{1/\beta} \\
&\geq \left(\frac{\alpha}{\alpha+\beta}\right)^{1/\beta\gamma}\left[\frac{\|f_{\theta_m}\|_{\alpha+\beta}^{(\alpha+\beta)}}{\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha}\|k_m\|_{\alpha+\beta}^{\beta}}\right]^{1/\beta}, \text{by Hölder's inequality} \\
&\geq \left(\frac{\alpha}{\alpha+\beta}\right)^{1/\beta\gamma},
\end{aligned}$$

where the last line follows from noting that $\|f_{\theta_m}\|_{\alpha+\beta} \geq \|k_m\|_{\alpha+\beta}$ for sufficiently large m. Therefore,

$$\widetilde{\epsilon}_{(\alpha,\beta)}^{\psi} \geq \min\left\{\left(\frac{\alpha}{\alpha+\beta}\right)^{1/\beta\gamma}, 1 - \left(\frac{\alpha}{\alpha+\beta}\right)^{1/\beta\gamma}\frac{\|f_{\theta^g}\|_{\alpha+\beta}^{\alpha+\beta}}{\langle f_{\theta^g}, g\rangle_{\alpha,\beta}}\right\}.$$

Now an application of Theorem 6.6 completes the proof. □

It is possible to extend the results of Corollary 6.3 to any ψ function satisfying the following requirement: There exists some $\gamma > 0$ such that

$$\frac{\alpha}{\alpha+\beta}\psi(x) \geq \psi\left(\left(\frac{\alpha}{\alpha+\beta}\right)^{1/\gamma}x\right),$$

for all $x \in [0, \infty)$. For such ψ-functions, it is easy to verify that the inequality

$$\left[\frac{\psi^{-1}\left(\frac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta})\right)}{\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}}\right]^{1/\beta} \geq \left(\frac{\alpha}{\alpha+\beta}\right)^{1/\beta\gamma}\left[\frac{\|f_{\theta_m}\|_{\alpha+\beta}^{(\alpha+\beta)}}{\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta}}\right]^{1/\beta},$$

holds due to the strictly increasing nature of the ψ function, from which exact same steps of the proof of Corollary 6.3 can be retraced.

However, Corollary 6.3 is not applicable for the MLSD functional, since one needs to consider the limiting case $\gamma \to 0+$ for this case. So, in the following corollary, we present a similar result for the particular case of MLSD, which is an important special case of the MGABD family.

Corollary 6.4. *Suppose that $\alpha, \beta > 0$ and the Assumptions (BP1)-(BP3) and (BP5) hold. Additionally, assume that the sequence of contaminating densities satisfy*

$$\limsup_{m\to\infty} \|k_m\|_{\alpha+\beta} \leq \|g\|_{\alpha+\beta}\,.$$

Then, the asymptotic breakdown point of the MGABD functional with $\psi(x) = \ln(x)$ is at least

$$\frac{\langle f_{\theta^g}, g\rangle_{\alpha,\beta}^{1/\beta}}{\langle f_{\theta^g}, g\rangle_{\alpha,\beta}^{1/\beta} + \|f_{\theta^g}\|_{\alpha+\beta}^{\alpha/\beta}\,\|g\|_{\alpha+\beta}}. \tag{6.18}$$

Proof. Let

$$\tilde{\epsilon}^{\psi}_{(\alpha,\beta)} = \liminf_{m\to\infty}\left[\frac{\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_g}, g\right\rangle_{\alpha,\beta}\right)^{1/\beta}}{\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_g}, g\right\rangle_{\alpha,\beta}\right)^{1/\beta} + \left(\left\|f_{\theta_g}\right\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_m}, k_m\right\rangle_{\alpha,\beta}\right)^{1/\beta}}\right].$$

Then, for every $\epsilon < \tilde{\epsilon}^{\psi}_{(\alpha,\beta)}$, and for sufficiently large m we must have

$$\begin{aligned}
&\epsilon < \frac{\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_g}, g\right\rangle_{\alpha,\beta}\right)^{1/\beta}}{\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_g}, g\right\rangle_{\alpha,\beta}\right)^{1/\beta} + \left(\left\|f_{\theta_g}\right\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_m}, k_m\right\rangle_{\alpha,\beta}\right)^{1/\beta}},\\
&\implies \left(\frac{1}{\epsilon} - 1\right)^{\beta} > \frac{\left\|f_{\theta_g}\right\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_m}, k_m\right\rangle_{\alpha,\beta}}{\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_g}, g\right\rangle_{\alpha,\beta}},\\
&\implies (1-\epsilon)^{\beta}\,\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_g}, g\right\rangle_{\alpha,\beta} > \epsilon^{\beta}\left\|f_{\theta_g}\right\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_m}, k_m\right\rangle_{\alpha,\beta},\\
&\implies \psi((1-\epsilon)^{\beta}\left\langle f_{\theta_g}, g\right\rangle_{\alpha,\beta}) + \tfrac{\alpha}{\alpha+\beta}\psi(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta}) > \psi(\epsilon^{\beta}\left\langle f_{\theta_m}, k_m\right\rangle_{\alpha,\beta}) + \tfrac{\alpha}{\alpha+\beta}\psi(\left\|f_{\theta_g}\right\|_{\alpha+\beta}^{\alpha+\beta}),\\
&\implies \psi((1-\epsilon)^{\beta}\left\langle f_{\theta_g}, g\right\rangle_{\alpha,\beta}) - \psi(\epsilon^{\beta}\left\langle f_{\theta_m}, k_m\right\rangle_{\alpha,\beta}) > \tfrac{\alpha}{\alpha+\beta}\left(\psi(\left\|f_{\theta_g}\right\|_{\alpha+\beta}^{\alpha+\beta}) - \psi(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta})\right).
\end{aligned}$$

This implies that condition (6.11) is satisfied for the specific sequence of densities $\{k_m\}$ under consideration. Now, by Hölder's inequality, $\langle f_{\theta_m}, k_m\rangle_{\alpha,\beta} \leq \|f_{\theta_m}\|_{\alpha+\beta}^{\alpha}\,\|k_m\|_{\alpha+\beta}^{\beta}$. It follows that

$$\begin{aligned}
&\left[\frac{\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_g}, g\right\rangle_{\alpha,\beta}\right)^{1/\beta}}{\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_g}, g\right\rangle_{\alpha,\beta}\right)^{1/\beta} + \left(\left\|f_{\theta_g}\right\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_m}, k_m\right\rangle_{\alpha,\beta}\right)^{1/\beta}}\right]\\
&\geq \left[\frac{\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_g}, g\right\rangle_{\alpha,\beta}\right)^{1/\beta}}{\left(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha}\left\langle f_{\theta_g}, g\right\rangle_{\alpha,\beta}\right)^{1/\beta} + \left(\left\|f_{\theta_g}\right\|_{\alpha+\beta}^{\alpha}\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha}\,\|k_m\|_{\alpha+\beta}^{\beta}\right)^{1/\beta}}\right]\\
&= \left[\frac{\langle f_{\theta^g}, g\rangle_{\alpha,\beta}^{1/\beta}}{\langle f_{\theta^g}, g\rangle_{\alpha,\beta}^{1/\beta} + \|f_{\theta^g}\|_{\alpha+\beta}^{\alpha/\beta}\,\|k_m\|_{\alpha+\beta}}\right]\\
&\geq \frac{\langle f_{\theta^g}, g\rangle_{\alpha,\beta}^{1/\beta}}{\langle f_{\theta^g}, g\rangle_{\alpha,\beta}^{1/\beta} + \|f_{\theta^g}\|_{\alpha+\beta}^{\alpha/\beta}\,\|g\|_{\alpha+\beta}},
\end{aligned}$$

where the last line follows from the fact that for sufficiently large m, $\|k_m\|_{\alpha+\beta} \leq \|g\|_{\alpha+\beta}$.

Now an application of Theorem 6.6 yields that the above quantity is a lower bound of the asymptotic breakdown point of the MGABD functional when $\psi(x) = \ln(x)$. □

When the true density belongs to the model family of densities, i.e., there exists θ^g lying in the interior of Θ such that $f_{\theta^g} = g$, then the asymptotic breakdown point of the MLSD functional under the same conditions as in Corollary 6.4 is 1/2. It follows clearly by noting that $\langle f_{\theta^g}, g\rangle_{\alpha,\beta} = \|f_{\theta^g}\|_{\alpha+\beta}^{\alpha+\beta} = \|g\|_{\alpha+\beta}^{\alpha+\beta}$ in (6.18).

So far, we considered the situation when $\|k_m\|_{\alpha+\beta}$ remains bounded by the $L^{\alpha+\beta}$-norm of either the true density (i.e., $\|g\|_{\alpha+\beta}$) or the model family of densities (i.e., $\|f_\theta\|_{\alpha+\beta}$) for sufficiently large m. When such a condition is not satisfied and $\|k_m\|_{\alpha+\beta}$ remains large, the model family is expected to have some asymptotic orthogonality restrictions to ensure Assumption (BP2) is satisfied. However, even when $\|k_m\|_{\alpha+\beta}$ becomes large, it is possible to use a model family for the estimation in such a way so that $\|f_{\theta_m}\|_{\alpha+\beta}$ becomes asymptotically smaller than $\|f_{\theta^g}\|_{\alpha+\beta}$ or $\|g\|_{\alpha+\beta}$. For example, in the scale estimation problem of normal distribution setup (shown later in Section 6.5.1) one might consider a location model family instead of the usual scale family to estimate the scale parameter. Since the location model family does not affect $\|f_{\theta_m}\|_{\alpha+\beta}$, it remains bounded as $m \to \infty$. On the other hand, $\|k_m\|_{\alpha+\beta}$ increases to ∞ as the contaminating scale "implodes" to 0. To ensure such nonsensical cases are removed, we restrict our attention to the situations when both $\|k_m\|_{\alpha+\beta}$ and $\|f_{\theta_m}\|_{\alpha+\beta}$ grow more than $\|f_{\theta^g}\|_{\alpha+\beta}$ for any sequence of MGABD estimators θ_m with $\theta_m \to \theta_\infty \in \partial\Theta$. In such situations, the following corollary may be useful.

Corollary 6.5. *Let $\alpha, \beta > 0$ and the Assumptions (BP1)-(BP3) and (BP5) hold. If the contaminating densities $\{k_m\}$ is such that $\|k_m\|_{\alpha+\beta} \geq \|f_{\theta^g}\|_{\alpha+\beta}$ and $\|f_{\theta_m}\|_{\alpha+\beta} \geq \|f_{\theta^g}\|_{\alpha+\beta}$ for sufficiently large m, and*

$$\limsup_m \langle f_{\theta_m}, k_m\rangle_{\alpha,\beta} \leq C,$$

for some constant $C > 0$ for any possible $\theta_m \to \theta_\infty \in \partial\Theta$, then the asymptotic breakdown point of the MGABD functional is at least

$$\min\left\{\frac{\langle f_{\theta^g}, g\rangle_{\alpha,\beta}^{1/\beta}}{C^{1/\beta} + \langle f_{\theta^g}, g\rangle_{\alpha,\beta}^{1/\beta}}, 1/2\right\}.$$

Proof. We have, $\|k_m\|_{\alpha+\beta} \geq \|f_{\theta^g}\|_{\alpha+\beta}$ and $\|f_{\theta_m}\|_{\alpha+\beta} \geq \|f_{\theta^g}\|_{\alpha+\beta}$ for sufficiently large m because of the provided assumptions. Also since $\limsup_m \langle f_{\theta_m}, k_m\rangle_{\alpha,\beta} \leq C$, to make sure that the condition (6.11) holds, it is sufficient to ensure

$$\psi\left((1-\epsilon)^\beta \langle f_{\theta^g}, g\rangle_{\alpha,\beta}\right) - \psi(\epsilon^\beta C) > \tfrac{\alpha}{\alpha+\beta}\left(\psi(\|f_{\theta^g}\|_{\alpha+\beta}^{\alpha+\beta}) - \psi(\|f_{\theta_m}\|_{\alpha+\beta}^{\alpha+\beta})\right).$$

As ψ is strictly increasing, the right-hand side of the above inequality is nonpositive for sufficiently large m. Therefore, it is enough to have

$$\begin{aligned}
&\psi\left((1-\epsilon)^\beta \langle f_{\theta^g}, g\rangle_{\alpha,\beta}\right) > \psi(\epsilon^\beta C),\\
\iff &\left(\frac{1-\epsilon}{\epsilon}\right)^\beta > \frac{C}{\langle f_{\theta^g}, g\rangle_{\alpha,\beta}}, \text{ because of the strictly increasing nature of } \psi,\\
\iff &\epsilon < \frac{\langle f_{\theta^g}, g\rangle_{\alpha,\beta}^{1/\beta}}{C^{1/\beta} + \langle f_{\theta^g}, g\rangle_{\alpha,\beta}^{1/\beta}}.
\end{aligned}$$

An application of Theorem 6.6 now completes the proof. □

6.4.2 Asymptotic Breakdown Point of MGABD functional for the case $\alpha > 0, \beta = 0$

In the previous Section 6.4.1, we have analysed the behaviour of the asymptotic breakdown point of the MGABD functional when both the hyperparameters α and β are positive. A crucial ingredient of this analysis is Hölder's inequality, which does not hold for the case when either $\alpha = 0$ or $\beta = 0$. However, when $\alpha = 0$, usually the asymptotic breakdown point only has the trivial lower bound, as seen from Theorem 5.5 for the special case of $\psi(x) = x$. Therefore, we restrict our attention to only the case $\alpha > 0, \beta = 0$ in this section.

Theorem 6.7. *Let us assume that the Assumptions (BP1)-(BP3) and (BP5) are satisfied. Suppose, there exists $\tilde{\epsilon}^{\psi}_{(\alpha,0)} \in [0, 1/2]$ such that for all $\epsilon < \tilde{\epsilon}^{\psi}_{(\alpha,0)}$ there exists $\theta^g \in \Theta$ (may depend on ϵ) such that for sufficiently large m, we have*

$$\psi'\left(\|f_{\theta_m}\|_\alpha^\alpha\right)\int f_{\theta_m}^\alpha \log\left(\frac{f_{\theta_m}}{\epsilon k_m}\right) - \psi'\left(\|f_{\theta^g}\|_\alpha^\alpha\right)\int f_{\theta^g}^\alpha \log\left(\frac{f_{\theta^g}}{(1-\epsilon)g}\right) > \frac{1}{\alpha}\left[\psi\left(\|f_{\theta_m}\|_\alpha^\alpha\right) - \psi\left(\|f_{\theta^g}\|_\alpha^\alpha\right)\right] \quad (6.19)$$

whenever $\theta_m \to \theta_\infty$ for some $\theta_\infty \in \partial\Theta$ and for any sequence of contaminating densities $\{k_m\}_{m=1}^\infty$. Then, the asymptotic breakdown point of the MGABD functional with a valid continuous generating function ψ for the $\alpha > 0, \beta = 0$ case is at least $\min\{\tilde{\epsilon}^{\psi}_{(\alpha,0)}, 1/2\}$.

Proof. The proof is very similar to the proof of Theorem 6.6, with two small differences. First, the Hölder's inequality is not applicable due to which some convergence results need to be modified appropriately. Secondly, one has to work with the limiting form of the GAB divergence for the $\beta = 0$ case. We only indicate these differences in the proof below.

Step 1:

By choosing the set A_m as before, due to Assumptions (BP1)-(BP3), we obtain that

$$\int f_{\theta_m}^\alpha \ln(f_{\theta_m}/g_{\epsilon,m}) \asymp \int_{A_m^c} f_{\theta_m}^\alpha \ln(f_{\theta_m}/\epsilon k_m),$$

as $m \to \infty$. An application of continuity of the ψ function then implies that

$$\alpha^2 d_{GAB}^{(\alpha,0),\psi}(f_{\theta_m}, g_{\epsilon,m}) \asymp \alpha\psi\left(\|f_{\theta_m}\|_\alpha^\alpha\right)\int f_{\theta_m}^\alpha \ln(f_{\theta_m}/\epsilon k_m) - \psi\left(\|f_{\theta_m}\|_\alpha^\alpha\right) + \psi\left(\|g_{\epsilon,m}\|_\alpha^\alpha\right), \quad (6.20)$$

as $m \to \infty$.

Step 2:

Since $g_{\epsilon,m} = (1-\epsilon)g + \epsilon k_m$, we clearly have the inequality

$$\int f_{\theta^g}^\alpha \ln(g_{\epsilon,m}/f_{\theta^g}) \geq \int f_{\theta^g}^\alpha \ln((1-\epsilon)g/f_{\theta^g}),$$

for any $m \geq 1$ and any $\epsilon \in (0, 1/2]$. Since for $\alpha = 0$, the necessary conditions for GABD to be a valid divergence measure is that ψ is a strictly increasing function, it follows that

$\psi'(\|f_{\theta g}\|_\alpha^\alpha) > 0$. Therefore, we have the inequality

$$\alpha^2 d_{GAB}^{(\alpha,0),\psi}(f_{\theta g}, g_{\epsilon,m}) \leq \alpha\psi'\left(\|f_{\theta g}\|_\alpha^\alpha\right) \int f_{\theta g}^\alpha \ln(f_{\theta g}/(1-\epsilon)g) - \psi\left(\|f_{\theta g}\|_\alpha^\alpha\right) + \psi\left(\|g_{\epsilon,m}\|_\alpha^\alpha\right). \tag{6.21}$$

Step 3:

In step 3, we compare the right-hand sides of the asymptotic equivalence relation (6.20) and the inequality (6.21) to yield the condition (6.19), completing the proof. □

Now that we have established the general results, we can obtain corollaries similar to the ones described in Section 6.4. We begin with a corollary similar to that of Corollary 6.1 for the $\beta = 0$ case.

Corollary 6.6. *Suppose that $\beta = 0$ and the Assumptions (BP1)-(BP3) and (BP5) hold true. Let $C = \limsup_{m\to\infty} \|k_m\|_\alpha$. Then, the asymptotic breakdown point of the MGABD functional is at least $\min\{1/2, \epsilon^*_{(\alpha,\beta)}\}$, where $\epsilon^*_{(\alpha,\beta)}$ is a solution of*

$$\psi'\left(\|f_{\theta g}\|_\alpha^\alpha\right) \int f_{\theta g}^\alpha \log\left(\frac{(1-\epsilon)g}{f_{\theta g}}\right) - \frac{1}{\alpha}\left[\psi\left((C\epsilon)^\alpha\right) - \psi\left(\|f_{\theta g}\|_\alpha^\alpha\right)\right] = 0, \tag{6.22}$$

in the interval $[0,1]$ provided it exists. If the solution does not exist, we take $\epsilon^*_{(\alpha,\beta)}$ as 0.

Proof. We first note that, the GAB divergence between two non-normalized density functions f_{θ_m} and ϵk_m is nonnegative for valid choices of ψ-function. In mathematical terms, this means

$$\alpha^2 d_{GAB}^{(\alpha,0)}(f_{\theta_m}, \epsilon k_m) = \alpha\psi'\left(\|f_{\theta_m}\|_\alpha^\alpha\right) \int f_{\theta_m}^\alpha \ln(f_{\theta_m}/\epsilon k_m) - \psi\left(\|f_{\theta_m}\|_\alpha^\alpha\right) + \psi\left(\|\epsilon k_m\|_\alpha^\alpha\right) \geq 0.$$

Since ψ must be increasing, it follows that

$$\alpha\psi'\left(\|f_{\theta_m}\|_\alpha^\alpha\right) \int f_{\theta_m}^\alpha \ln(f_{\theta_m}/\epsilon k_m) - \psi\left(\|f_{\theta_m}\|_\alpha^\alpha\right) + \psi\left((C\epsilon)^\alpha\right) \geq 0.$$

In view of condition (6.19), it is thus enough to show that for any $\epsilon < \epsilon^*_{(\alpha,\beta)}$, we have the inequality

$$\begin{aligned} &-\frac{1}{\alpha}\psi\left((C\epsilon)^\alpha\right) > \psi\left(\|f_{\theta g}\|_\alpha^\alpha\right) \int f_{\theta g} \ln\left(\frac{f_{\theta g}}{(1-\epsilon)g}\right) - \frac{1}{\alpha}\psi\left(\|f_{\theta g}\|_\alpha^\alpha\right), \\ \iff &\psi'\left(\|f_{\theta g}\|_\alpha^\alpha\right) \int f_{\theta g}^\alpha \log\left(\frac{(1-\epsilon)g}{f_{\theta g}}\right) - \frac{1}{\alpha}\left[\psi\left((C\epsilon)^\alpha\right) - \psi\left(\|f_{\theta g}\|_\alpha^\alpha\right)\right] > 0. \end{aligned}$$

Clearly, when $\epsilon = \epsilon^*_{(\alpha,\beta)}$, the left-hand side of the above inequality is 0. The first term in left-hand side is strictly decreasing in ϵ in $[0,1]$. Since $C \geq 0$, $\psi\left((C\epsilon)^{1+\alpha}\right)$ is increasing in ϵ. Hence, the left-hand side is strictly decreasing in ϵ. Hence, for all $\epsilon < \epsilon^*_{(\alpha,\beta)}$ the left-hand side of the above inequality must be positive.

Now an application of Theorem 6.7 establishes the corollary. □

As before, when $\psi \in C^1((0,\infty))$ and g belongs to the model family so that $g = f_{\theta g}$, the condition (6.22) must have a solution. Note that, when $\epsilon = 0$, the left-hand side of

condition (6.22) is $\alpha^{-1}\psi(\|g\|_\alpha^\alpha) > 0$. When $\epsilon = 1$, the left-hand side becomes negative. The existence of the solution is then guaranteed by the intermediate value theorem.

Similar to the Corollary 6.2, one can show that for the location estimation problem, the asymptotic breakdown point of the MGABD functional for the $\alpha > 0, \beta = 0$ case also achieves the highest possible breakdown of $1/2$.

Corollary 6.7. *If $\mathcal{F} := \{f_\theta : \theta \in \Theta\}$ is a location model family of densities where θ is the location parameter and the true density g and the sequence of contaminating densities $\{k_m\}$ belong to the same location model family, then under the Assumptions (BP1)-(BP3) and (BP5), the asymptotic breakdown point of the MGABD functional is $1/2$ for any valid ψ-function and hyperparameters $\alpha > 0, \beta = 0$.*

Proof. Since we consider the location family, the quantities $\|f_\theta\|_\alpha^\alpha = c$, free of θ. Also, as the true density g belong to the same location family, we have $f_{\theta^g} = g$ and $\|g\|_\alpha^\alpha = c$. Hence, the right-hand side of the inequality (6.19) reduces to 0. In fact, the condition (6.19) reduces to the inequality

$$\psi'(c)\left[\int f_{\theta_m}^\alpha \ln\left(\frac{f_{\theta_m}}{k_m}\right) - c\ln(\epsilon) + c\ln(1-\epsilon)\right] > 0. \tag{6.23}$$

By the necessary conditions established in Theorem 6.5, we know that ψ must be strictly increasing and hence $\psi'(c) > 0$. Also note that, since both f_{θ_m} and k_m belong to the same location family, their α-zoomed densities are simply given by $f_{\theta_m}^{[\alpha]} = c^{-1} f_{\theta_m}^\alpha$ and $k_m^{[\alpha]} = c^{-1} k_m^\alpha$. Therefore, the condition (6.23) further reduces to

$$d_{KL}(f_{\theta_m}^{[\alpha]}, k_m^{[\alpha]}) > \ln\left(\frac{\epsilon}{1-\epsilon}\right),$$

where $d_{KL}(\cdot,\cdot)$ denotes the Kullback-Leibler (KL) divergence. Clearly, the left-hand side is nonnegative, and hence it is enough to choose ϵ so that the right-hand side is negative. This means, we need $\epsilon/(1-\epsilon) < 1$, i.e., $\epsilon < 1/2$. In other words, for the location parameter estimation setup, the condition (6.19) remains true for any $\epsilon < 1/2$.

Finally, an application of Theorem 6.7 yields the corollary. □

The following corollary establishes an explicit form of the lower bound of the asymptotic breakdown point of MGABD functional, under conditions similar to that of Corollary 6.3, but for the $\alpha > 1, \beta = 0$ case.

Corollary 6.8. *Let $\alpha > 1, \beta = 0$ and the Assumptions (BP1)-(BP3) and (BP5) hold. Also, assume that the contaminating densities $\{k_m\}$ satisfy $\|k_m\|_\alpha \le \|f_{\theta_m}\|_\alpha$ for all sufficiently large m. Then, the corresponding MGABD functional has an asymptotic breakdown point of at least*

$$\min\left\{\exp\left(-\frac{1}{\alpha}\sup_y \frac{\Psi(y)}{\Psi'(y)}\right), 1 - \exp\left[\frac{\int f_{\theta^g}^\alpha \ln(f_{\theta^g}/g)}{\|f_{\theta^g}\|_\alpha^\alpha} - \frac{\Psi\left(\|f_{\theta^g}\|_\alpha^\alpha\right)}{\alpha\Psi'\left(\|f_{\theta^g}\|_\alpha^\alpha\right)}\right], \frac{1}{2}\right\},$$

where $\Psi(x) = \psi(e^x)$.

Proof. Using Hölder's inequality and that fact that $\ln(1+x) < x$, it follows that

$$\begin{aligned}\int f_{\theta_m}^{\alpha} \log\left(\frac{\epsilon k_m}{f_{\theta_m}}\right) &\leq \|f_{\theta_m}\|_{\alpha}^{\alpha} \ln(\epsilon) + \int f_{\theta_m}^{\alpha-1} k_m - \|f_{\theta_m}\|_{\alpha}^{\alpha} \\ &\leq \|f_{\theta_m}\|_{\alpha}^{\alpha} (\ln(\epsilon) - 1) + \|f_{\theta_m}\|_{\alpha}^{\alpha-1} \|k_m\|_{\alpha} \\ &\leq \|f_{\theta_m}\|_{\alpha}^{\alpha} \ln(\epsilon),\end{aligned}$$

for all sufficiently large m. Here, we also use the fact that $\|k_m\|_{\alpha} \leq \|f_{\theta_m}\|_{\alpha}$ at the last step of the inequality. In view of the condition (6.19), it is now enough to show that

$$-\psi'\left(\|f_{\theta_m}\|_{\alpha}^{\alpha}\right) \|f_{\theta_m}\|_{\alpha}^{\alpha} \ln(\epsilon) \geq \frac{1}{\alpha} \psi\left(\|f_{\theta_m}\|_{\alpha}^{\alpha}\right), \tag{6.24}$$

and,

$$\psi'\left(\|f_{\theta_g}\|_{\alpha}^{\alpha}\right) \int f_{\theta_g}^{1+\alpha} \log\left(\frac{f_{\theta_g}}{(1-\epsilon)g}\right) < \frac{1}{\alpha} \psi\left(\|f_{\theta_g}\|_{\alpha}^{\alpha}\right). \tag{6.25}$$

The first inequality given in (6.24) translates to

$$\epsilon < \exp\left(-\frac{\psi(x)}{\alpha x \psi'(x)}\right) = \exp\left(-\frac{\Psi(\ln(x))}{\Psi'(\ln(x))}\right),$$

where $x = \|f_{\theta_m}\|_{\alpha}^{\alpha}$. The division is permissible here since $\psi(x)$ is strictly increasing for all $x > 0$, resulting in $\psi'(x) > 0$. This now means, we can take any ϵ such that

$$\epsilon < \exp\left(-\frac{1}{\alpha} \sup_y \frac{\Psi(y)}{\Psi'(y)}\right),$$

and the inequality in (6.24) will follow.

On the other hand, the second inequality (6.25) translates to

$$\begin{aligned}&\psi'\left(\|f_{\theta_g}\|_{\alpha}^{\alpha}\right) \int f_{\theta_g}^{\alpha} \ln(1-\epsilon) > \psi'\left(\|f_{\theta_g}\|_{\alpha}^{\alpha}\right) \int f_{\theta_g}^{\alpha} \ln(f_{\theta_g}/g) - \frac{1}{\alpha} \psi\left(\|f_{\theta_g}\|_{\alpha}^{\alpha}\right) \\ \iff &\ln(1-\epsilon) > \frac{\int f_{\theta_g}^{\alpha} \ln(f_{\theta_g}/g)}{\|f_{\theta_g}\|_{\alpha}^{\alpha}} - \frac{\Psi\left(\ln(\|f_{\theta_g}\|_{\alpha}^{\alpha})\right)}{\alpha \Psi'\left(\ln(\|f_{\theta_g}\|_{\alpha}^{\alpha})\right)} \\ \iff &\epsilon < 1 - \exp\left[\frac{\int f_{\theta_g}^{\alpha} \ln(f_{\theta_g}/g)}{z} - \frac{\Psi\left(\ln(z)\right)}{\alpha \Psi'\left(\ln(z)\right)}\right],\end{aligned}$$

where $z = \|f_{\theta_g}\|_{\alpha}^{\alpha}$.

Since the choice of ϵ given in the statement of the result satisfies both the inequalities (6.24) and (6.25), an application of Theorem 6.7 now completes the proof. □

For the special case $\psi(x) = \gamma^{-1} x^{\gamma}$ with $\gamma > 0$, the lower bound of the breakdown point asserted by Corollary 6.8 results in

$$\min\left\{\exp\left(-\frac{1}{\alpha\gamma}\right), 1 - \exp\left(\frac{\int f_{\theta_g}^{\alpha} \ln(f_{\theta_g}/g)}{\|f_{\theta_g}\|_{\alpha}^{\alpha}} - \frac{1}{\alpha\gamma}\right)\right\}.$$

Furthermore, if the true density g belongs to the model family of densities $\mathcal{F}$, i.e., we have $g = f_{\theta_g}$, then the above lower bound vastly simplifies to $\min\{e^{-1/\alpha\gamma}, 1 - e^{-1/\alpha\gamma}\}$.

However, for the special case of LSD with $\psi(x) = \ln(x)$, the quantity $\sup_x \Psi(x)/\Psi'(x) = \sup_x x = \infty$, hence the resulting lower bound turns out to be 0. Unfortunately, this lower

bound cannot be improved further. To see this, note that the GAB divergence between f and g in this special case is equivalent to the KL-divergence between the corresponding α-zoomed densities $f^{[\alpha]}$ and $g^{[\alpha]}$. This follows easily from the reduction demonstrated in (6.9). Therefore, minimization of GAB divergence in this case is equivalent to the minimum KL-divergence estimator, i.e., the maximum likelihood estimator (MLE). MLE is known to be nonrobust with asymptotic breakdown point equal to 0.

6.5 Examples and Illustrations

In this section, we demonstrate how the results described in Section 6.4 are applicable to different parametric setups.

6.5.1 Normal Distributions

In this example, we analyse the asymptotic breakdown point of the MGABD functional for estimating location and scale parameter in a model family of d-variate normal densities (for $d \geq 1$). The contaminating densities k_m are also assumed to belong to the same normal family of densities. The true density g is taken to be the standard normal density on $\mathbb{R}^d$ with mean vector $\mathbf{0}$ and dispersion matrix $\boldsymbol{I}_d$.

First, we consider the problem of location estimation. In this, we consider the model family as the normal densities with mean vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \equiv \mathbb{R}^d$ and dispersion matrix $\boldsymbol{I}_d$, and the contaminating densities are assumed to be normal densities with mean vector $\boldsymbol{\mu}_m$ and same dispersion matrix, where $\|\boldsymbol{\mu}_m\| \to \infty$ as $m \to \infty$. Corollaries 6.2 and 6.7 are applicable in this case, which ensures that the MGABD functional achieves the highest possible breakdown point of $1/2$ for any $\alpha > 0$ and $\beta \geq 0$. Remarkably, this breakdown is free of the data dimension d.

Next, we consider the problem of scale estimation. In this, we take the model family as the normal densities with mean $\mathbf{0}$ but unknown dispersion matrix $\boldsymbol{\Sigma} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ denotes the set of all $d \times d$ positive definite matrices. The contaminating densities are taken to be normal densities with mean vectors $\boldsymbol{\mu}_m$ and dispersion matrices $\boldsymbol{\tau}_m$ for each $m = 1, 2, \ldots$, such that either $\det(\boldsymbol{\tau}_m) \to 0$ or $\det(\boldsymbol{\tau}_m) \to \infty$ as $m \to \infty$. It is easy to see that for $\alpha + \beta > 0$,

$$\|k_m\|_{\alpha+\beta}^{\alpha+\beta} = (2\pi)^{-(\alpha+\beta-1)d/2} \left(\det(\boldsymbol{\tau}_m)\right)^{-(\alpha+\beta-1)/2} (\alpha+\beta)^{-d/2}.$$

Case 1 (Explosion type breakdown):

Let us consider "explosion"-type breakdown first, i.e., $\det(\boldsymbol{\tau}_m) \to \infty$ as $m \to \infty$. If $\alpha > 0, \beta \geq 0$ and $(\alpha + \beta) > 1$, then we can apply Corollaries 6.1 and 6.6 with $C = 0$. Let us denote

$$z_{\alpha,\beta} = \begin{cases} \left[\|g\|_{\alpha+\beta}^{-(\alpha+\beta)} \psi^{-1}\left(\frac{\alpha}{\alpha+\beta}\psi(\|g\|_{\alpha+\beta}^{\alpha+\beta}) + \frac{\beta}{\alpha+\beta}\psi(0) \right) \right]^{1/\beta} & \beta > 0, \\ \exp\left[\frac{\psi(0) - \psi(\|g\|_\alpha^\alpha)}{\alpha \|g\|_\alpha^\alpha \psi'(\|g\|_\alpha^\alpha)} \right] & \beta = 0. \end{cases} \tag{6.26}$$

Then, it follows that the asymptotic breakdown point $\epsilon^*_{(\alpha,\beta)}$ of the MGABD functional $T^{\psi}_{(\alpha,\beta)}(G)$ satisfies

$$\epsilon^*_{(\alpha,\beta)} \geq \min\{1/2, 1 - z_{\alpha,\beta}\}, \ (\alpha+\beta) > 1.$$

One may apply Corollaries 6.3 and 6.8 also for specific forms of ψ-function to obtain explicit lower bounds in this case.

If $\alpha+\beta = 1$, then the resulting MGABD functional is equivalent to the minimum power divergence functional as illustrated in Section 6.3.1. Hence, in this case, the asymptotic breakdown point of $T^{\psi}_{(\alpha,\beta)}(G)$ is at least $\min\{1/2, \alpha^{1/\beta}\}$ if $\beta > 0$ and e^{-1} if $\beta = 0$.

Finally, when $\alpha + \beta < 1$, then $\|k_m\|^{\alpha+\beta}_{\alpha+\beta} \to \infty$ as $m \to \infty$, thus, violating the Assumption (BP5). Hence, none of our results are applicable in this particular scenario.

Case 2 (Implosion type breakdown):

In the "implosion"-type breakdown, we consider the case when $\det(\boldsymbol{\tau}_m) \to 0$ as $m \to \infty$. Contrary to the Case 1, we can only apply our results for the case when $\alpha + \beta \leq 1$ in this scenario. When $\alpha + \beta < 1$, the lower bound to the asymptotic breakdown point of the MGABD functional is the same as before, namely $\min\{1/2, 1 - z_{\alpha,\beta}\}$ where $z_{\alpha,\beta}$ is as given in Eq. (6.26). On the other hand, if $\alpha + \beta = 1$, then equivalence of the GAB divergence and the power divergence yields the lower bound to the asymptotic breakdown point as $\min\{1/2, \alpha^{1/\beta}\}$ if $\beta > 0$ and e^{-1} if $\beta = 0$.

In all of these case, we again highlight the fact that the lower bound obtained depends on the hyperparameters α, β, the ψ-function and the norm of the true density, i.e., $\|g\|_{\alpha+\beta}$, but remains free of the dimension of the parameter space d.

6.5.2 Gamma Distribution (Scale Family)

In this example, we present the breakdown point analysis for the MGABD functional to estimate the inverse-scale parameter θ of a $\Gamma(t, \theta)$ family of densities. We assume that the shape parameter $t > 0$ is fixed and known. Additionally, let us assume that the true density g and the contaminating densities k_m also belong to the same family, i.e., g is a gamma density with shape parameter t and inverse-scale parameter θ^g, and, k_m is a sequence of gamma densities with the same shape parameter but different inverse-scale parameters τ_m where $\tau_m \to \infty$ or $\tau_m \to 0$ as $m \to \infty$.

Elementary calculations show that

$$\|k_m\|^{\alpha+\beta}_{\alpha+\beta} = \frac{\tau_m^{(\alpha+\beta-1)}\Gamma((\alpha+\beta)t - (\alpha+\beta-1))}{(\alpha+\beta)^{(\alpha+\beta)t-(\alpha+\beta-1)}\Gamma(t)^{(\alpha+\beta)}}, \ \alpha, \beta, (\alpha+\beta) > 0.$$

Therefore, depending on the limit of τ_m and whether $\alpha + \beta$ is greater than or less than 1, the lower bound to the asymptotic breakdown point changes.

When either $\alpha + \beta > 1$ and $\tau_m \to 0$, or $\alpha + \beta < 1$ and $\tau_m \to \infty$, then $\|k_m\|^{\alpha+\beta}_{\alpha+\beta} \to 0$ as $m \to \infty$. Therefore, using Corollaries 6.1 and 6.6 with $C = 0$, we obtain that the asymptotic breakdown point of the MGABD functional is at least $\min\{1/2, 1 - z_{\alpha,\beta}\}$, where $z_{\alpha,\beta}$ is as provided in (6.26).

When either $\alpha+\beta<1$ and $\tau_m \to 0$, or $\alpha+\beta>1$ and $\tau_m \to \infty$, then $\|k_m\|_{\alpha+\beta}^{\alpha+\beta} \to \infty$ as $m \to \infty$. In this case, Assumption (BP5) does not hold and hence, the results provided in this chapter do not apply.

If $\alpha+\beta=1$, then irrespective of $\tau_m \to \infty$ or $\tau_m \to 0$, we can obtain the asymptotic breakdown point of the MGABD functional. It turns out to be at least $\min\{1/2, \alpha^{1/\beta}\}$ for $\beta>0$ and e^{-1} for $\beta=0$.

Note that, even if the setup is quite different from the earlier example of the normal scale family shown in Section 6.5.1, the resulting behaviour of the asymptotic breakdown points remain the same since the convergence behaviour of $\|k_m\|_{\alpha+\beta}$ is the same. These exact bounds hold true across many different scale parameter estimation setups, beyond the few examples illustrated in this section.

6.5.3 Gamma Distribution (Shape Family)

This example is closely related to the former example presented in Section 6.5.2. Here, instead of estimating the scale parameter, now we are interested in the estimation of the shape parameter. Without the loss of generality, we assume that the scale parameter is known and is equal to 1. The model family is, therefore, given by f_t, the density function of a $\Gamma(t,1)$ random variable where $t \in (0,\infty)$. Let the true shape parameter be t^g. The contaminating densities are given by k_m which are again gamma densities with shape parameter τ_m where $\tau_m \to 0$ or $\tau_m \to \infty$. Some elementary calculation as before yields that

$$\|k_m\|_{\alpha+\beta}^{\alpha+\beta} = \frac{\Gamma((\alpha+\beta)\tau_m - (\alpha+\beta-1))}{(\alpha+\beta)^{(\alpha+\beta)\tau_m-(\alpha+\beta-1)}\Gamma(\tau_m)^{(\alpha+\beta)}}, \ \alpha,\beta,(\alpha+\beta)>0.$$

When $\tau_m \to \infty$ as $m \to \infty$, we can apply Stirling's approximation $\Gamma(n+1) \asymp \sqrt{2\pi}n^{n+1/2}e^{-n}$ to obtain

$$\begin{aligned}
\|k_m\|_{\alpha+\beta}^{\alpha+\beta} &= \frac{\Gamma((\alpha+\beta)\tau_m - (\alpha+\beta-1))}{(\alpha+\beta)^{(\alpha+\beta)\tau_m-(\alpha+\beta-1)}\Gamma(\tau_m)^{(\alpha+\beta)}} \\
&\asymp \frac{\sqrt{2\pi}((\alpha+\beta)\tau_m - (\alpha+\beta))^{(\alpha+\beta)\tau_m-(\alpha+\beta)+1/2}e^{-(\alpha+\beta)\tau_m+(\alpha+\beta)}}{(\alpha+\beta)^{(\alpha+\beta)\tau_m-(\alpha+\beta-1)}\sqrt{2\pi}^{(\alpha+\beta)}(\tau_m-1)^{(\alpha+\beta)(\tau_m-1/2)}e^{-(\tau_m-1)(\alpha+\beta)}} \\
&\asymp (2\pi)^{-(\alpha+\beta-1)/2}(\alpha+\beta)^{-1/2}(\tau_m-1)^{-(\alpha+\beta-1)/2},
\end{aligned}$$

as $m \to \infty$. Hence, if $(\alpha+\beta)>1$, we have $\|k_m\|_{\alpha+\beta}^{\alpha+\beta} \to 0$, the asymptotic breakdown point of $T^{\psi}_{(\alpha,\beta)}(G)$ has a lower bound of $\min\{1/2, 1-z_{\alpha,\beta}\}$ where $z_{\alpha,\beta}$ is as given in (6.26). On the other hand, if $(\alpha+\beta)<1$, then $\|k_m\|_{\alpha+\beta}$ diverges to infinity, which makes our results inapplicable for this case.

When $\tau_m \to 0$, we can use the approximation $\Gamma(n) \asymp n^{-1}$ for small n. Therefore, we get

$$\begin{aligned}
\|k_m\|_{\alpha+\beta}^{\alpha+\beta} &= \frac{\Gamma((\alpha+\beta)\tau_m - (\alpha+\beta-1))}{(\alpha+\beta)^{(\alpha+\beta)\tau_m-(\alpha+\beta-1)}\Gamma(\tau_m)^{(\alpha+\beta)}} \\
&\asymp \Gamma(-(\alpha+\beta-1))(\alpha+\beta)^{(\alpha+\beta-1)}\tau_m^{-(\alpha+\beta)}.
\end{aligned}$$

Since $\alpha, \beta > 0$, we have $(\alpha+\beta) > 0$ and hence $\|k_m\|_{\alpha+\beta}^{\alpha+\beta} \to 0$ for any $(\alpha+\beta) \notin \{0, 1, 2, \dots\}$. Thus, we again have a lower bound to the asymptotic breakdown point as before, i.e., $\min\{1/2, 1 - z_{\alpha,\beta}\}$ where $z_{\alpha,\beta}$ is as given in (6.26). When $(\alpha + \beta)$ is a nonnegative integer, then the integral in the gamma function $\Gamma(-(\alpha + \beta - 1))$ is not finite. Thus, corresponding GAB divergences do not fall into the scope of our results, as they violate Assumption (BP5).

Finally, when $\alpha + \beta = 1$, the asymptotic breakdown point of the MGABD functional turns out to be the same as given in Section 6.5.2, for the specific case of $\alpha + \beta = 1$.

6.6 Numerical Studies

In this section, we shall empirically verify some of the results presented in this chapter by calculating the GAB divergence between the contaminated density and the model density to obtain the minimum GAB divergence estimator, and subsequently, observing how it changes for different levels of contamination. While in Section 5.4, we illustrated these numerical verifications for the S-divergence family (i.e., GABD with $\psi(x) = x$), here we restrict our attention to the logarithmic S-divergence (LSD) family (i.e., GABD with $\psi(x) = \ln(x)$) for the purpose of illustration. Also, we indicate the LSD by the hyperparameters α and λ which relate to the hyperparameters α_d and β_d of the GAB divergence by the relation

$$\alpha_d = \alpha - \lambda(1 - \alpha), \text{ and, } \beta_d = 1 + \lambda(1 - \alpha).$$

For a specific choice of contaminating density k and a model family $\{f_\theta : \theta \in \Theta\}$ containing the true density g, we calculate the LSD between $g_\epsilon = (1 - \epsilon)g + \epsilon k$ and f_θ for a range of values of $\epsilon \in [0, 1]$ and $\theta \in \Theta$. Since we consider a single contaminating density k instead of a sequence of contaminating densities, it is only possible to choose k so that the Assumptions (BP1)-(BP3) are satisfied only approximately, but not exactly. Also, when $\alpha_d, \beta_d \geq 0$, the cross-integral term $\int f_\theta^{\alpha_d} g^{\beta_d}$ is usually intractable except when $\beta_d = 1$. To circumvent this problem, we consider the decomposition of the integral

$$\begin{aligned}
\int f_\theta^{\alpha_d} g_\epsilon^{\beta_d} &= \int f_\theta^{\alpha_d} ((1-\epsilon)g + \epsilon k)^{\beta_d} \\
&= \int f_\theta^{\alpha_d} ((1-\epsilon)g + \epsilon k)^{\beta_d} \mathbf{1}_{\{x: g(x) > k(x)\}} + \int f_\theta^{\alpha_d} ((1-\epsilon)g + \epsilon k)^{\beta_d} \mathbf{1}_{\{x: g(x) \leq k(x)\}} \\
&\approx (1-\epsilon)^{\beta_d} \int f_\theta^{\alpha_d} g^{\beta_d} + \epsilon^{\beta_d} \int f_\theta^{\alpha_d} k^{\beta_d},
\end{aligned}$$

where the final approximation holds due to the Assumptions (BP1)-(BP3). It follows that when $g(x) > k(x)$, then the term $\langle f_\theta, (1-\epsilon)g \rangle_{\alpha_d,\beta_d}$ is the dominating factor in $\langle f_\theta, g_\epsilon \rangle_{\alpha_d,\beta_d}$ for any finite θ away from the boundary $\partial\Theta$. Therefore, in the following illustrations, we compute the objective function connected to the LSD functional as

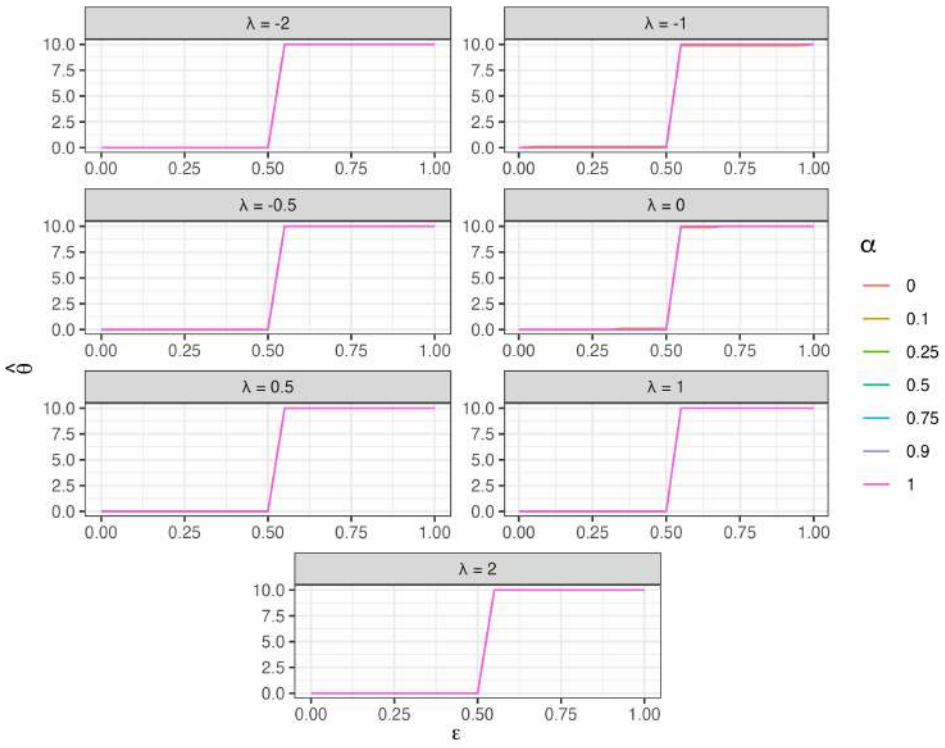
$$q^{\psi}_{(\alpha,\lambda)}(f_\theta, g_\epsilon) \approx \log \left[\frac{\|f_\theta\|_{1+\alpha}^{(1+\alpha)/\beta_d}}{\left((1-\epsilon)^{\beta_d} \langle f_\theta, g \rangle_{\alpha_d,\beta_d} + \epsilon^{\beta_d} \langle f_\theta, k \rangle_{\alpha_d,\beta_d} \right)^{(1+\alpha)/\alpha_d\beta_d}} \right], \tag{6.27}$$

and minimize it with respect to $\theta \in \Theta$. We repeat this procedure for different choices of ϵ in a grid on the unit interval $[0, 1]$, and then plot the estimates as a function of ϵ. This allows us to inspect the particular value of ϵ where the estimate θ starts to behave erratically.

6.6.1 Normal Location and Scale Family

We begin by considering the location and the scale estimation problem in normal family. The model density is taken to be a univariate normal density with mean μ and variance σ^2, which contains the true density g with mean μ_g and variance σ_g^2. The contaminating density k is taken to be $\mathcal{N}(\mu_k, \sigma_k^2)$ where μ_k and σ_k^2 are appropriately chosen to indicate different types of contamination.

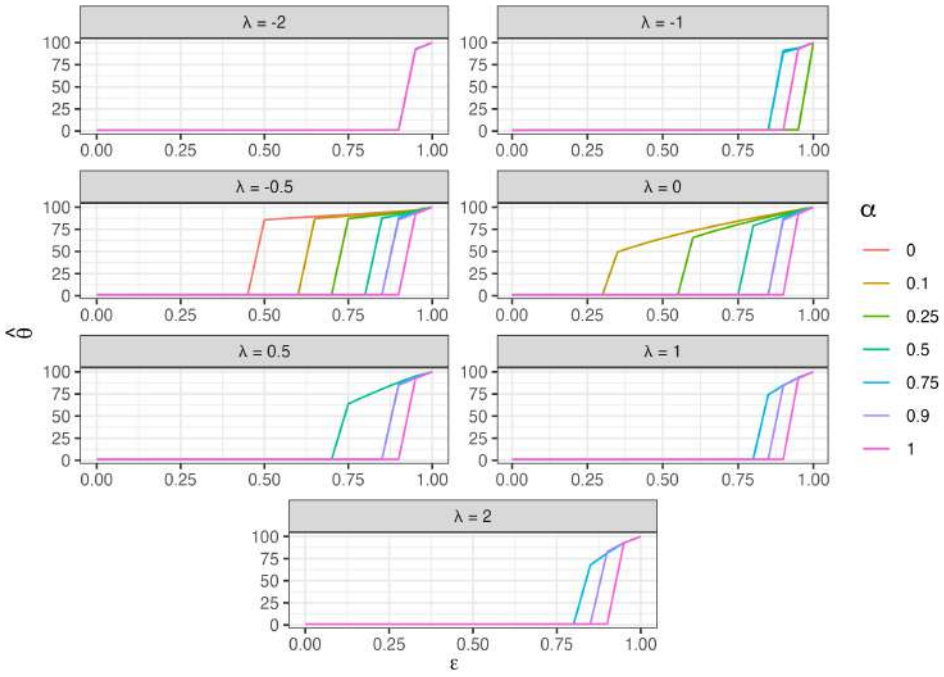
Figure 6.1: Behaviour of the MLSD functional for change in contaminating proportion ϵ for location estimation in the normal family.



In Figure 6.1, we demonstrate the change in the MLSD functional for the location problem. Here, the variance components $\sigma = \sigma_g = \tau_k$ are kept same as 1, while $\mu_g = 0$ and $\mu_k = 10$, a comparatively large value compared to the true mean. According to the earlier results, namely Theorems 6.2 and 6.7, we expect the asymptotic breakdown should be at $\epsilon = 0.5$, which is quite clear from the figure.

Figure 6.2 depicts the change in the MLSD functional for the scale model for the fixed mean 0 and large contaminating variance. Here, we choose $\mu = \mu_k = \mu_g = 0$, $\sigma_k = 10$ and $\sigma_g = 1$. From Figure 6.2, it is evident that the breakdown point is higher for large values of α indicating increased robustness. Additionally, aside from the cases $\alpha = 0, \lambda = -0.5$

Figure 6.2: Behaviour of the MLSD functional for change in contaminating proportion ϵ for scale estimation in the normal family when the contaminating density is normal with zero mean and large variance.



and $\alpha = 0.1, \lambda = 0$, the breakdown point is seen to be 0.5 for all choices of pairs (α, λ) demonstrated in our simulation exercise.

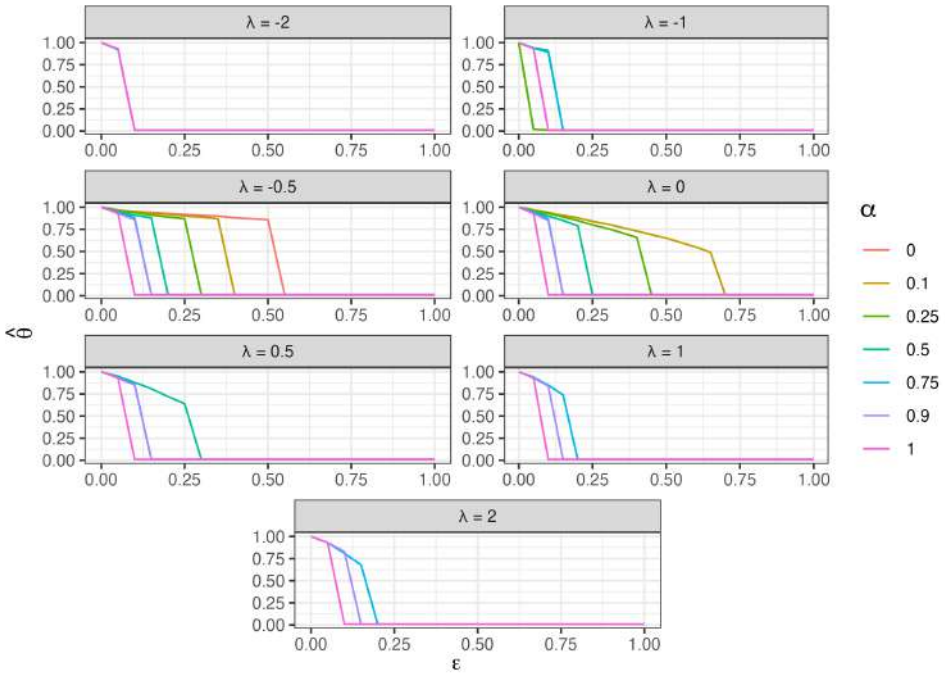
In contrast to the scenario mentioned above, we also consider the changes in the MLSD functional for the scale model when the contaminating scale σ_k is close to zero. The corresponding behaviour is shown in Figure 6.3. The breakdown point is observed to be higher for smaller values of α. This is in contrast with the usual understanding that the high values of α results in increased robustness as is the case for MDPDE and related estimators (Basu et al., 1998). Unfortunately, this particular case does not fall within the current scope of our theory as $\|k_m\|_{\alpha_d+\beta_d} = \|k_m\|_{1+\alpha} \to \infty$, so we do not have a concrete justification for such a phenomenon. However, both Fujisawa and Eguchi (2008) and Kuchibhotla et al. (2019) mentioned this phenomenon about a spurious minima for MLSD near $\sigma = 0$. The breakdown points, in this case, are not very high, indicating that the minimum logarithmic S-divergence estimator (MLSDE) for the normal scale model is unlikely to be able to withstand a large amount of contamination by another normal density with the same mean and a very low variance.

6.6.2 Exponential Model

In this example, we consider the model family of exponential distributions, with unknown inverse-scale parameter. We consider the notation $\text{Exp}(\gamma)$ to denote the form of exponential density given by

$$f_\gamma(x) = \gamma e^{-\gamma x}, x > 0,$$

Figure 6.3: Behaviour of the MLSD functional for change in contaminating proportion ϵ for scale estimation in the normal family when the contaminating density is normal with zero mean and a very small variance.



where $\gamma > 0$ is the inverse-scale parameter. The true density is then given by $\text{Exp}(\gamma_g)$ and the contaminating density by $\text{Exp}(\gamma_k)$. Let us also denote the scale parameter as $\theta = 1/\gamma$, which is also the mean of the $\text{Exp}(\gamma)$ distribution.

We conduct two types of experiments, depending on whether we consider "explosion"-type or "implosion"-type breakdown point. In the first experiment, we take $\gamma_g = 1$ and $\gamma_k = 100$ (i.e., $\theta_k = 0.01$) to be a large value, to understand the effect of "explosion"-type breakdown. In Figure 6.4, it is seen that the breakdown point (i.e., the point where the spike occurs in the graph) of the scale estimator increases with an increase in the value of α, as is expected. The resultant breakdown points are higher than 1/2 apart from a few cases. Additionally, we note that the estimators corresponding to the hyperparameters $\lambda = -2$ and $\lambda = -1$ performed exceedingly well in this case.

In the second experiment, we consider taking $\gamma_k = 1/125$, a small value close to 0. Correspondingly, the mean of the contaminating density is 125. It is easy to see that $\|k_m\|_{\alpha_d+\beta_d} \to \infty$ as $m \to \infty$ in this case, our theoretical results do not provide any guarantee on the asymptotic breakdown point of the MLSDE. However, from Figure 6.5, we again see a strange behaviour where an increase in the robustness parameter α results in a lower breakdown point.

6.6.3 Binomial Model

In this exercise, we shall be looking at the binomial model family setup as in Section 5.5.4. Although the definition of GAB divergence and the Assumptions (BP1)-(BP3) deals with

Figure 6.4: Behaviour of the MLSD functional for change in contaminating proportion ϵ for exponential densities with a very large value of contaminating inverse-scale parameter.

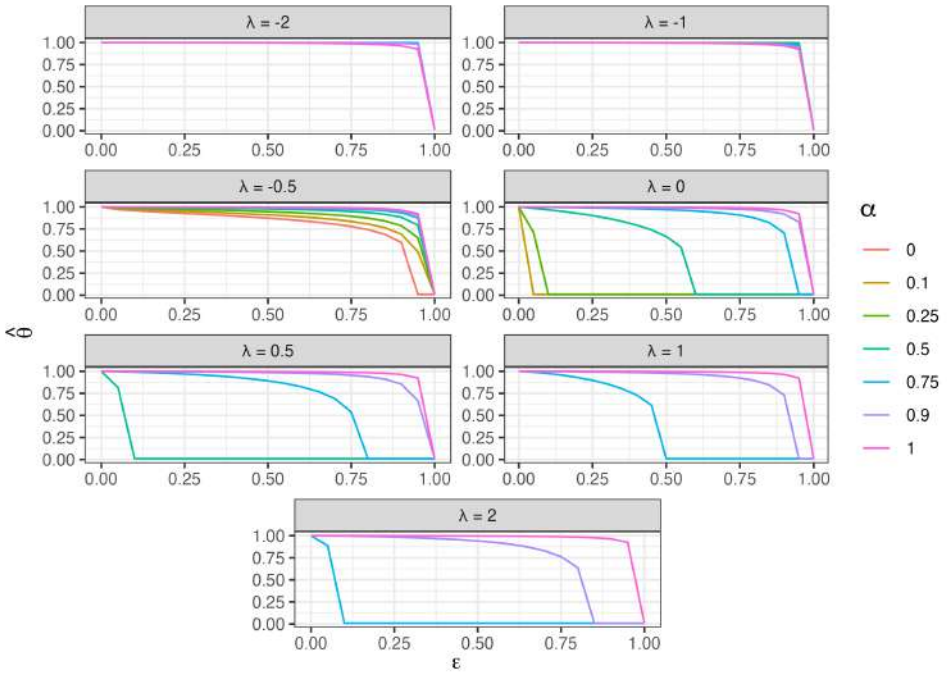
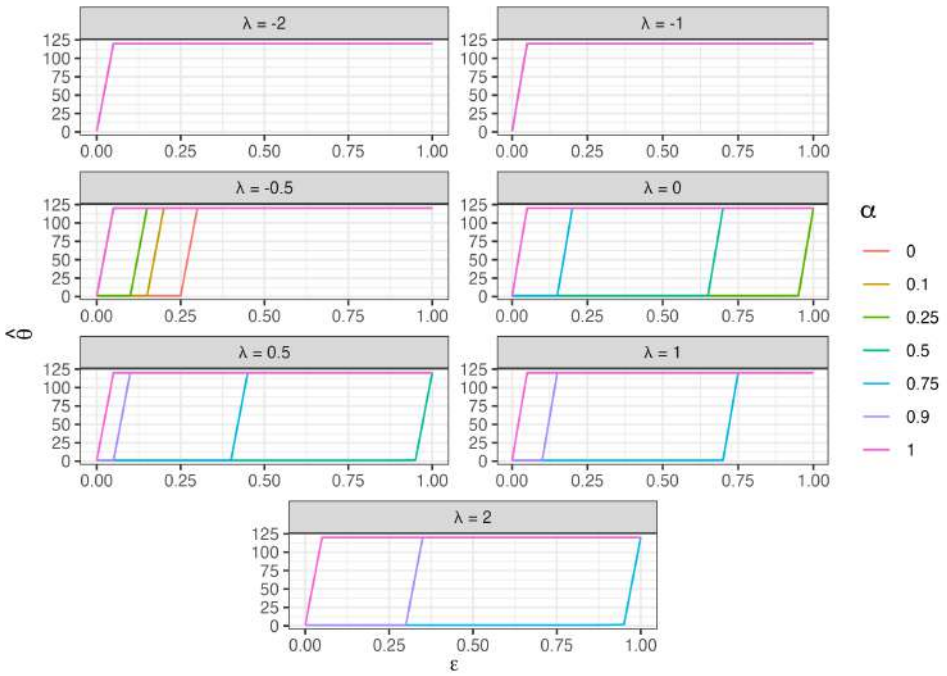


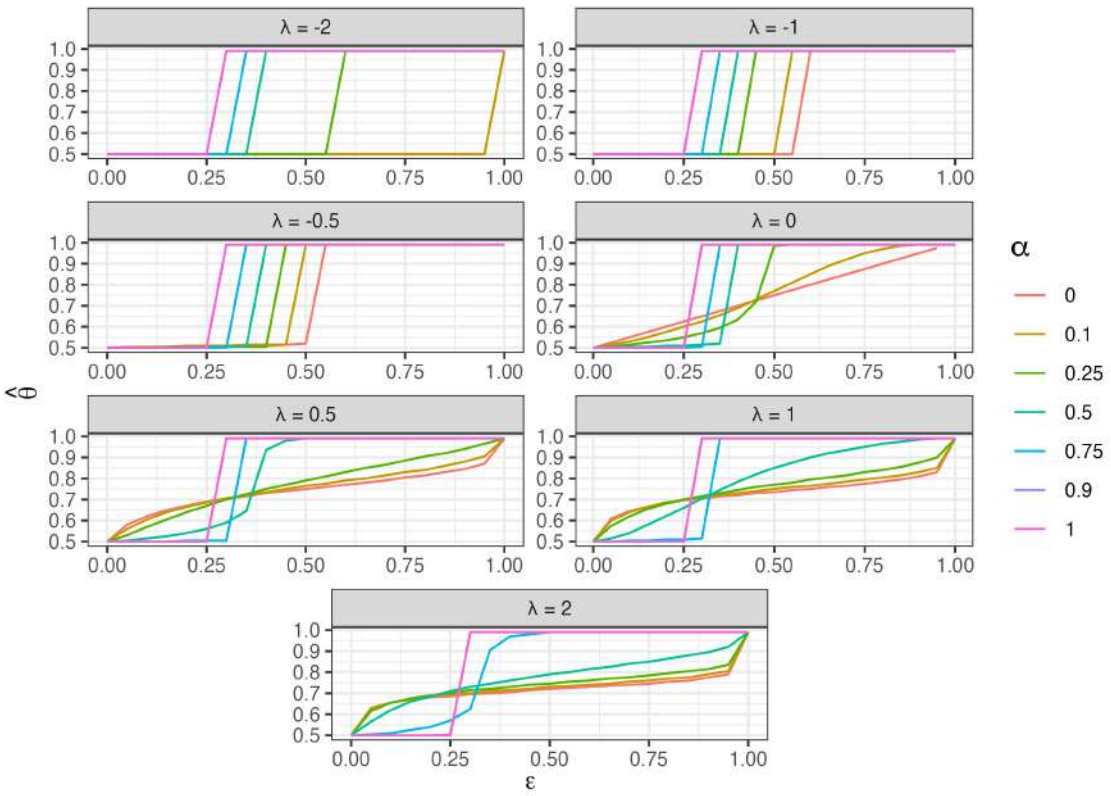
Figure 6.5: Behaviour of the MLSD functional for change in contaminating proportion ϵ for exponential densities with a very small value of contaminating inverse-scale parameter.



density functions, there is no difficulty in constructing the divergence with probability mass functions for discrete distributions. Here, the model family is the family of binomial distributions with size 12 and unknown success probability $\theta \in (0, 1)$, the true distribution has success probability $\theta_g = 1/2$ and the contaminating distribution has success probability equal to $\theta_k = 1$. As the support of the binomial distribution is compact, the singularity assumptions (BP1)-(BP3) do not hold in this scenario. However, by the specific choice of the contaminating density, we approximately satisfy these singularity assumptions.

In Figure 6.6, we illustrate the change in estimated probability θ as a function of the contamination proportion $\epsilon \in [0, 1]$. As shown in the figure, when $\lambda < 0$, the asymptotic breakdown point for the MLSD estimator seems to decrease with increasing α. On the other hand, when $\lambda > 0$, the situation reverses; the asymptotic breakdown point of the MLSD estimator becomes an increasing function of α.

Figure 6.6: Behaviour of the MLSD functional for change in contaminating proportion ϵ for binomial model with a binomial with success probability 1.

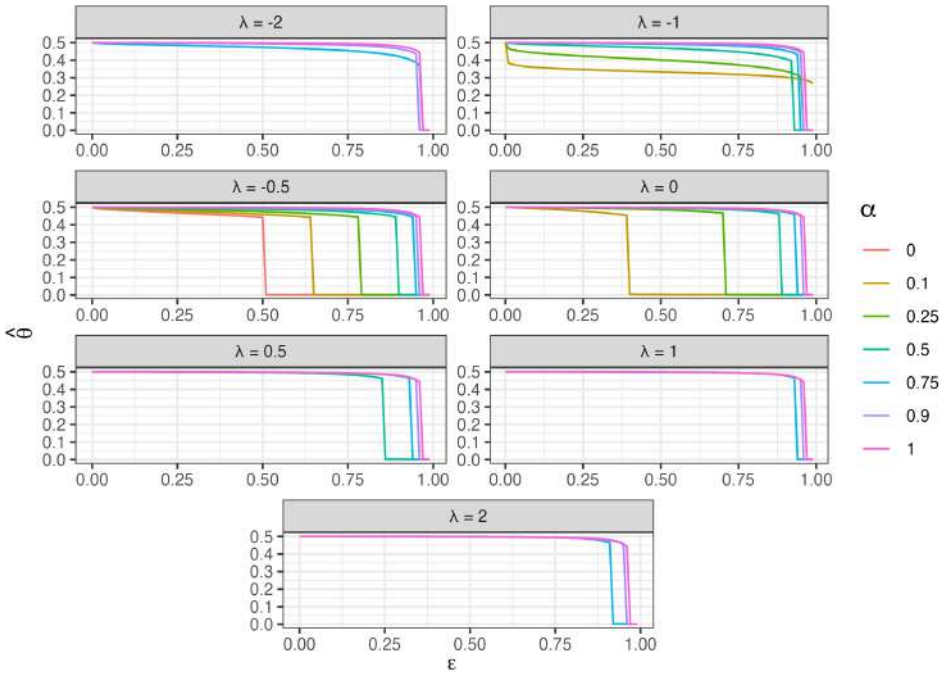


6.6.4 Geometric Model

As in the previous example involving the Binomial setup, we study the behaviour of the breakdown point for another discrete distribution setup, namely the Geometric distribution. In this case, we pick the model family as Geometric distributions with unknown success probability θ, true distribution with $\theta_g = 1/2$, and the contaminating distribution with success probability θ_k. We consider two cases here: $\theta_k = 0$ and $\theta_k = 1$ to illustrate

"implosion"-type and "explosion"-type breakdowns respectively

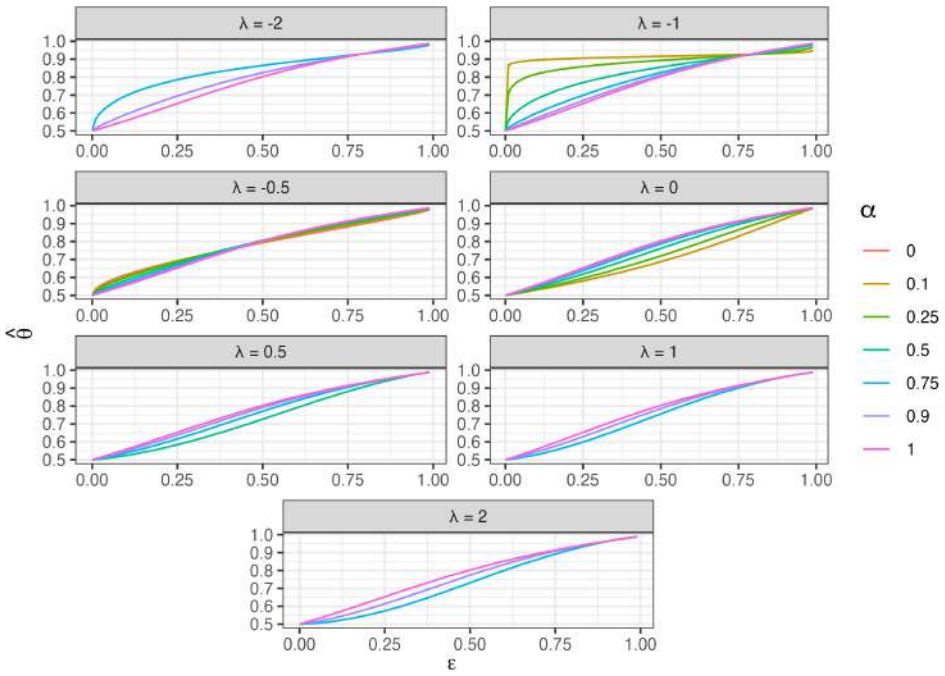
Figure 6.7: Behaviour of the MLSD functional for change in contaminating proportion ϵ for the Geometric model when the contaminating density is Geometric with very low success probability.



In the first case with $\theta_k = 0$, we depict the estimated probability as a function of ϵ in Figure 6.7. In this case, the breakdown point is indicated by the proportion of contamination ϵ where the value of the estimator experiences a sharp drop. Also, the breakdown point may appear to be beyond $1/2$ in this case because of numerical approximations, finite sample size, and the fact that Assumptions (BP1)-(BP3) are only approximately satisfied. Furthermore, with increasing α, it is observed that the breakdown point increases, as per our expectations.

However, the same cannot be said for the case when $\theta_k = 1$ since Assumption (BP5) is not satisfied in this case. As evident from Figure 6.8, the value of the estimator steadily increase to 1 in all cases except the case where $\lambda = -1$ and α is small. These are the only cases in which the breakdown point (which is pretty low) can be clearly ascertained from the plot. This indicates that the MLSD does not produce sufficiently robust estimators for the geometric family when the contaminating sequence is geometric with a very high success probability. Unfortunately, this case falls beyond the scope of the theory that we have provided in this Chapter.

Figure 6.8: Behaviour of the MLSD functional for change in contaminating proportion ϵ for the Geometric model when the contaminating density is Geometric with very high success probability.



6.A Appendix: Proofs of additional results

Lemma 6.A.1. *Let $f : \mathbb{R} \to \mathbb{R}$ be a real-valued function and $F(x) := f(e^x)$ for any $x \in \mathbb{R}$. Let $\lambda \in [0, 1]$ be a nonnegative real number. Then, F is λ-convex if and only if for any $x, y \geq 0$,*

$$\lambda f(x) + (1 - \lambda) f(y) \geq f(x^\lambda y^{1-\lambda})$$

Proof. We start with assuming F is convex. Then,

$$\begin{aligned}
\lambda f(x) + (1 - \lambda) f(y) &= \lambda F(\log x) + (1 - \lambda) F(\log y) \\
&\geq F\left(\lambda \ln(x) + (1 - \lambda) \ln(y)\right), \text{by convexity of } F \\
&= f(x^\lambda y^{1-\lambda}), \text{ by definition of } F
\end{aligned}$$

Conversely, for any $x, y \in \mathbb{R}$ and $\lambda \in [0, 1]$, we have

$$\begin{aligned}
\lambda F(x) + (1 - \lambda) F(y) &= \lambda f(e^x) + (1 - \lambda) f(e^y) \\
&\geq f((e^x)^\lambda (e^y)^{1-\lambda}) \\
&= f(e^{\lambda x + (1-\lambda) y}) \\
&= F(\lambda x + (1 - \lambda) y).
\end{aligned}$$

This proves that F is λ-convex on $\mathbb{R}$. □

Lemma 6.A.2. *Consider two non-negative sequences $\{x_m\}$ and $\{y_m\}$. Then the following holds*

$$\liminf_{m\to\infty} \min\{x_m, y_m\} = \min\left\{\liminf_{m\to\infty} x_m, \liminf_{m\to\infty} y_m\right\}$$

Proof. Note that when $\liminf x_m = \infty$ and $\liminf y_m < \infty$, we can say that there exists sufficiently large $M \in \mathbb{N}$ such that for all $m \geq M$, $y_m < x_m$ and hence, the lemma would follow immediately in this case. We, therefore, focus on the case where $\liminf x_m < \infty$ and $\liminf y_m < \infty$. We know that

$$\min\{x_m, y_m\} \leq x_m \text{ and } \min\{x_m, y_m\} \leq y_m$$

Taking $\liminf$, we observe that

$$\liminf_{m\to\infty} \min\{x_m, y_m\} \leq \liminf_{m\to\infty} x_m \text{ and } \liminf_{m\to\infty} \min\{x_m, y_m\} \leq \liminf_{m\to\infty} y_m.$$

This leads us to the fact that

$$\liminf_{m\to\infty} \min\{x_m, y_m\} \leq \min\{\liminf_{m\to\infty} x_m, \liminf_{m\to\infty} y_m\}.$$

Now, let $A = \liminf x_m$ and $B = \liminf y_m$. So, for every $\epsilon > 0$, there exists $N_1, N_2 \in \mathbb{N}$ such that $x_m > A - \epsilon$ for all $m \geq N_1$ and $y_m > B - \epsilon$ for all $m \geq N_2$. This means that for any $m \geq \max\{N_1, N_2\}$, we must have $\min\{x_m, y_m\} > \min\{A, B\} - \epsilon$. This implies the other direction

$$\liminf_{m\to\infty} \min\{x_m, y_m\} \geq \min\{A, B\}.$$

□

Lemma 6.A.3. *If both $A, B > 0$ and the function $\psi(x)$ is monotonically increasing and $\psi(x)$ is a convex function, then GAB divergence $d_{GAB}^{(\alpha,\beta),\psi}(f,g)$ is nonnegative for all f and g, and it is equal to 0 if and only if f and g are almost surely equal under their common dominated measure μ.*

Proof. Applying Hölder's inequality on the cross term, we have

$$\int f^{\alpha} g^{\beta} \leq \|f\|_{\alpha+\beta}^{\alpha} \|g\|_{\alpha+\beta}^{\beta}.$$

Applying weighted AM-GM inequality on this now yields

$$\|f\|_{\alpha+\beta}^{\alpha} \|g\|_{\alpha+\beta}^{\beta} \leq \frac{\alpha \|f\|_{\alpha+\beta}^{\alpha+\beta} + \beta \|g\|_{\alpha+\beta}^{\alpha+\beta}}{\alpha+\beta}.$$

Using this along with the monotonicity of ψ, we can write

$$\begin{aligned}
& d_{GAB}^{(\alpha,\beta),\psi}(f,g) \\
&= \frac{1}{\beta(\alpha+\beta)} \psi\left(\int f^{\alpha+\beta}\right) - \frac{1}{\alpha\beta} \psi\left(\int f^{\alpha} g^{\beta}\right) + \frac{1}{\alpha(\alpha+\beta)} \psi\left(\int g^{\alpha+\beta}\right) \\
&\geq \frac{1}{\beta(\alpha+\beta)} \psi\left(\|f\|_{\alpha+\beta}^{\alpha+\beta}\right) - \frac{1}{\alpha\beta} \psi\left(\|f\|_{\alpha+\beta}^{\alpha} \|g\|_{\alpha+\beta}^{\beta}\right) + \frac{1}{\alpha(\alpha+\beta)} \psi\left(\|g\|_{\alpha+\beta}^{\alpha+\beta}\right) \\
&\geq \frac{1}{\beta(\alpha+\beta)} \psi\left(\|f\|_{\alpha+\beta}^{\alpha+\beta}\right) - \frac{1}{\alpha\beta} \psi\left(\frac{\alpha \|f\|_{\alpha+\beta}^{\alpha+\beta} + \beta \|g\|_{\alpha+\beta}^{\alpha+\beta}}{\alpha+\beta}\right) + \frac{1}{\alpha(\alpha+\beta)} \psi\left(\|g\|_{\alpha+\beta}^{\alpha+\beta}\right) \\
&\geq \frac{1}{\beta(\alpha+\beta)} \psi\left(\|f\|_{\alpha+\beta}^{\alpha+\beta}\right) - \frac{1}{\alpha\beta} \frac{\alpha\psi(\|f\|_{\alpha+\beta}^{\alpha+\beta}) + \beta\psi(\|g\|_{\alpha+\beta}^{\alpha+\beta})}{\alpha+\beta} + \frac{1}{\alpha(\alpha+\beta)} \psi\left(\|g\|_{\alpha+\beta}^{\alpha+\beta}\right) \\
&\qquad \text{by convexity of } \psi, \\
&= 0
\end{aligned}$$

In the above derivation, we used two inequalities, namely the AM-GM inequality and the Hölder's inequality. The AM-GM inequality yields equality if and only if $\|f\|_{\alpha+\beta} = \|g\|_{\alpha+\beta}$ and the Hölder's inequality yields equality if and only if f and g are almost surely equal. □

Chapter 7

Conclusion and Future Scopes

This thesis addresses the challenge of robust matrix factorization in high dimensions – an important method in modern data analysis, from image and video processing to bioinformatics. By using minimum distance estimators, particularly the minimum density power divergence estimator, it aims to offer a robust and scalable approach to high-dimensional matrix factorization problems. We focus on two of the most popular and fundamental techniques: Singular Value Decomposition (SVD) and Principal Component Analysis (PCA). The core contribution of this work lies in demonstrating the effectiveness of these estimators, especially their robustness in performing SVD and PCA, driven by their dimension-free breakdown properties. These properties ensure reliable performance even in the presence of significant contamination in the data, across arbitrarily high-dimensional spaces.

We presented two novel algorithms – rSVDdpd and rPCAdpd – both based on the minimum density power divergence estimator (MDPDE). Both of these algorithms were evaluated theoretically and through various practical applications. The theoretical results provided guarantees that these estimators are equivariant and consistent, and our results also explored various other asymptotic properties. We prove the consistency of the proposed estimators under two different asymptotic regimes: first, when the data dimension p is fixed but the number of observations n grows to infinity; and second, when both dimensions n and p of the data matrix grow to infinity at a comparable rate. As a result, our proposed methods are applicable to a broader range of datasets (e.g., genomics data, spatio-temporal data) where the number of variables exceeds the number of observations (i.e., $n > p$), compared to most existing robust PCA algorithms which only work in the $n \leq p$ regime.

The practical usefulness of the proposed methods is demonstrated through real-world scenarios, such as background modeling in video surveillance. In this case, resilience to outliers and heavy-tailed noise is crucial for maintaining performance, especially in the presence of camera tampering. We also show the applicability of rSVDdpd beyond background modeling by using it to estimate latent factors in a spatio-temporal dataset of pollutants. Extensive simulation studies and comparisons on several benchmark datasets further establish the superiority of the proposed techniques in terms of both accuracy

and speed. Our methods maintain (and often improve upon) the performance of existing algorithms, while using only a fraction of the computational resources. The accompanying R package, `rsvddpd`, along with open-source code, makes these tools readily accessible, paving the way for adoption in both academic research and industry.

The thesis also provides some additional contributions as follows:

1. Proposing a novel penalized criterion to estimate the rank of the low-dimensional component of a data matrix robustly. Theoretical properties such as selection consistency and influence functions have been derived for this method. This provides a statistician with the complete toolset to apply the proposed rSVDdpd and rPCAdpd methods for a real-life dataset where the rank of the low-rank component $\boldsymbol{L}$ is typically unknown.

2. Introducing a broader class of statistical divergence measures, namely the Generalized Alpha-Beta Divergence (GABD) family. This class contains several important statistical divergence families, including DPD, LSD, Bridge divergence, etc. We obtained several general results on the asymptotic breakdown point of the minimum GABD functional, which encompasses many existing results on the asymptotic breakdown point of different minimum distance estimators present in the literature. We also support our theoretical results with examples and numerical illustrations.

Despite many positive outcomes, there remain some limitations in this study, opening up newer opportunities for future research.

1. **Sparse Matrix Factorization:** The proposed rSVDdpd and rPCAdpd algorithms struggle with the matrix decomposition when the observed data matrix itself is sparse with some outlying entries. In this case, the algorithm detects the low-rank component $\boldsymbol{L}$ as the $\mathbf{0}$ matrix, and assigns the entire data matrix to the sparse component $\boldsymbol{S}$. This may lead to limitations in some applications like latent semantic analysis, recommendation systems, etc. In such cases, it is necessary to develop some modifications of the proposed algorithms to handle such sparse data matrices. A viable approach to tackle this problem is to use generalized linear models (GLM) instead of the usual linear regression model when performing the alternating iterative steps in the rSVDdpd algorithm. In this case, we may consider a model given as $X_{ij} \sim F(Z_{ij})$, for $1 \leq i \leq n; 1 \leq j \leq p$ and the matrix $\boldsymbol{\eta}(\boldsymbol{Z})$ with (i,j)-th entry $\eta(Z_{ij})$ satisfies the LSN decomposition $\boldsymbol{\eta}(\boldsymbol{Z}) = \boldsymbol{L} + \boldsymbol{S} + \boldsymbol{N}$, where F is a suitable distribution parametrized by Z_{ij}s and $\eta(\cdot)$ is a suitable link function. While this extension is natural, extensive research is needed to study the convergence and the statistical properties of the resulting algorithm. We plan to tackle this problem in a future article.

2. **Breakdown Analysis:** The asymptotic breakdown point analysis presented in Chapters 5 and 6 are applicable only when the hyperparameters α and β are nonnegative. The breakdown behaviour for negative values of the hyperparameters remained

unexplored. Also, our analysis considered only the class of contaminating densities to have uniformly bounded L^p norms for some suitable choice of p. Although this boundedness is quite common, it excludes certain scenarios such as the "implosion" of the variance parameter in a normal scale model setup. Our numerical experiments in Section 6.6 highlight quite a drastic difference in the breakdown behaviour in such cases. Theoretical explorations of these scenarios and their implications on the breakdown behaviour of the minimum GABD functional represent interesting avenues for future research.

3. **Applications to new domains:** The robustness of the proposed estimators makes them attractive for a wide range of applications beyond the video surveillance background modelling example and the spatiotemporal data analysis example discussed in this thesis. Extending the model for graph-based clustering, signal processing, image and video watermarking, etc., are some of the potential applications. Additionally, latent factor estimation in domains like finance, biosciences, genomics, etc. where the signal-to-noise ratios are typically low, would be an interesting future research direction to study the real-life usefulness of the our methods.

4. **Generalization to higher-dimensional objects:** A natural extension of this work involves generalizing the proposed algorithms to higher-dimensional structures, such as tensor decomposition. This high-dimensional tensor-SVD has applications in bioinformatics, brain imaging, modelling convolutional neural networks, etc. This decomposition (called Tucker decomposition) is typically performed by flattening the tensor into a matrix and then applying singular value decomposition on the flattened matrices, and alternatively picking different dimensions to flatten the tensor. A direct robust generalization of this Higher-order Singular Value Decomposition algorithm (De Lathauwer et al., 2000) would consider replacing the traditional non-robust SVD at each step with a robust rSVDdpd algorithm. Further studies are needed to evaluate its theoretical and empirical performances. Interested readers are referred to Zhang and Xia (2018), Zhang and Han (2019) for more details on this research direction.

List of Manuscripts

In the following, I disclose the list of manuscripts that have been prepared or in preparation with the materials of this thesis.

Published Manuscripts:

1. Subhrajyoty Roy, Abhik Ghosh, and Ayanendranath Basu. Robust singular value decomposition with application to video surveillance background modelling. *Statistics & Computing*, 34(178), 2024. https://doi.org/10.1007/s11222-024-10493-7.

2. Subhrajyoty Roy, Ayanendranath Basu, and Abhik Ghosh. Robust Principal Component Analysis Using Density Power Divergence. *Journal of Machine Learning Research*, 25(324):1-40, 2024. http://jmlr.org/papers/v25/23-1096.html.

3. Subhrajyoty Roy, Abir Sarkar, Abhik Ghosh, and Ayanendranath Basu. Asymptotic Breakdown Point Analysis for a General Class of Minimum Divergence Estimators. *Bernoulli*, 2025. (accepted for publication).

Under preparation:

1. Subhrajyoty Roy, Supratik Basu, Abhik Ghosh, and Ayanendranath Basu. Characterization of Generalized Alpha-Beta Divergence and its Associated Entropy (2024+).

2. Subhrajyoty Roy, Abhik Ghosh, and Ayanendranath Basu. Divergence Information Criterion for Robust Estimation of Matrix Rank (2024+).

Bibliography

Robert J Adler and Jonathan E Taylor. Gaussian inequalities. *Random Fields and Geometry*, pages 49–64, 2007.

Hirotogu Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1973.

Josh Alman and Virginia Vassilevska Williams. *A Refined Laser Method and Faster Matrix Multiplication*, pages 522–539. 2021. doi: 10.1137/1.9781611976465.32.

Andris Ambainis, Yuval Filmus, and François Le Gall. Fast matrix multiplication: Limitations of the coppersmith-winograd method. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 585–593, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335362. doi: 10.1145/2746539.2746554.

Larry P. Ammann. Robust singular value decompositions: A new approach to projection pursuit. *Journal of the American Statistical Association*, 88(422):505–514, 1993. doi: 10.1080/01621459.1993.10476301.

Murugan Anandarajan, Chelsey Hill, and Thomas Nolan. *Semantic Space Representation and Latent Semantic Analysis*, pages 77–91. Springer International Publishing, Cham, 2019. ISBN 978-3-319-95663-3. doi: 10.1007/978-3-319-95663-3_6.

Theodore Wilbur Anderson. Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148, 1963.

Sergio Arciniegas-Alarcón, Marisol García-Peña, and Wojtek J. Krzanowski. Cross-validation for lower rank matrices containing outliers. *Applied System Innovation*, 5 (4), 2022. ISSN 2571-5577. doi: 10.3390/asi5040069.

Anthony C Atkinson, Marco Riani, and Aldo Corbellini. The analysis of transformations for profit and loss data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 69(2):251–275, 2020.

S. Derin Babacan, Martin Luessi, Rafael Molina, and Aggelos K. Katsaggelos. Sparse Bayesian Methods for Low-Rank Matrix Estimation. *IEEE Transactions on Signal Processing*, 60(8):3964–3977, 2012. doi: 10.1109/TSP.2012.2197748.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002. doi: https://doi.org/10.1111/1468-0262.00273.

Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, and M. C. Jones. Robust and Efficient Estimation by Minimising a Density Power Divergence. *Biometrika*, 85(3):549–559, 1998. ISSN 00063444.

Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park. *Statistical Inference: The Minimum Distance Approach*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2011. ISBN 9781420099669.

Subir K. Bhandari, Ayanendranath Basu, and Sahadeb Sarkar. Robust Inference in Parametric Models using the Family of Generalized Negative Exponential Disparities. *Australian & New Zealand Journal of Statistics*, 48(1):95–114, 2006. doi: https://doi.org/10.1111/j.1467-842X.2006.00428.x.

Peter J Bickel and Murray Rosenblatt. On some global measures of the deviations of density function estimates. *The Annals of Statistics*, pages 1071–1095, 1973.

Peter J. Bickel, Gil Kur, and Boaz Nadler. Projection pursuit in high dimensions. *Proceedings of the National Academy of Sciences*, 115(37):9151–9156, 2018. doi: 10.1073/pnas.1801177115.

Thierry Bouwmans. Traditional and Recent Approaches in Background Modeling for Foreground Detection: An overview. *Computer Science Review*, 11-12:31–66, 2014. ISSN 1574-0137. doi: https://doi.org/10.1016/j.cosrev.2014.04.001.

Thierry Bouwmans and El Hadi Zahzah. Robust PCA via Principal Component Pursuit: A Review for a Comparative Evaluation in Video Surveillance. *Computer Vision and Image Understanding*, 122:22–34, 2014. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2013.11.009.

Thierry Bouwmans, Fida El Baf, and Bertrand Vachon. *Statistical background modeling for foreground detection: A survey*, pages 181–199. World Scientific, 2010. doi: 10.1142/9789814273398_0008.

Thierry Bouwmans, Sajid Javed, Hongyang Zhang, Zhouchen Lin, and Ricardo Otazo. On the Applications of Robust PCA in Image and Video Processing. *Proceedings of the IEEE*, 106(8):1427–1457, 2018. doi: 10.1109/JPROC.2018.2853589.

R. Bro, K. Kjeldahl, A. K. Smilde, and H. A. L. Kiers. Cross-validation of component models: A critical look at current methods. *Analytical and Bioanalytical Chemistry*, 390(5):1241–1251, Mar 2008. ISSN 1618-2650. doi: 10.1007/s00216-007-1790-1.

A. Bustamam, S. Formalidin, and T. Siswantining. Clustering and analyzing microarray data of lymphoma using singular value decomposition (svd) and hybrid clustering. *AIP Conference Proceedings*, 2023(1):020220, 2018. doi: 10.1063/1.5064217.

HanQin Cai, Jian-Feng Cai, and Ke Wei. Accelerated Alternating Projections for Robust Principal Component Analysis. *Journal of Machine Learning Research*, 20(20):1–33, 2019.

N. A. Campbell. Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3): 231–237, 1980. ISSN 00359254, 14679876.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Fabrizio Carcillo, Yann-A"el Le Borgne, Olivier Caelen, and Gianluca Bontempi. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization. *International journal of data science and analytics (Print)*, 5(4):285–300, 2018.

Fabrizio Carcillo, Yann-A"el Le Borgne, Olivier Caelen, Yacine Kessaci, Frédéric Oblé, and Gianluca Bontempi. Combining unsupervised and supervised learning in credit card fraud detection. *Information sciences*, 2019.

Hervé Cardot and Antoine Godichon-Baggioni. Fast estimation of the median covariation matrix with application to online robust principal components analysis. *Test*, 26(3): 461–480, 2017.

Soumya Chakraborty, Ayanendranath Basu, and Abhik Ghosh. A componentwise estimation procedure for multivariate location and scatter: Robustness, efficiency and scalability. *arXiv preprint arXiv:2410.21166*, 2024.

Hongchuan Cheng, Yimin Zhang, Wenjia Lu, and Zhou Yang. A bearing fault diagnosis method based on vmd-svd and fuzzy clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, 33(12):1950018, 2019. doi: 10.1142/S0218001419500186.

Andrzej Cichocki and Shun-ichi Amari. Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities. *Entropy*, 12(6):1532–1568, 2010. ISSN 1099-4300. doi: 10.3390/e12061532.

Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, 2011. ISSN 1099-4300. doi: 10.3390/e13010134.

N. Cressie and C.K. Wikle. *Statistics for Spatio-Temporal Data*. CourseSmart Series. Wiley, 2011. ISBN 9780471692744.

Noel Cressie and Timothy R. C. Read. Multinomial Goodness-of-Fit Tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):440–464, 1984. ISSN 00359246.

C. Croux and A. Ruiz-Gazen. A Fast Algorithm for Robust Principal Components Based on Projection Pursuit. In Albert Prat, editor, *COMPSTAT*, pages 211–216, Heidelberg, 1996. Physica-Verlag HD. ISBN 978-3-642-46992-3.

Christophe Croux and Gentiane Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3):603–618, 09 2000. ISSN 0006-3444. doi: 10.1093/biomet/87.3.603.

Christophe Croux and Anne Ruiz-Gazen. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of multivariate analysis*, 95 (1):206–226, 2005.

Christophe Croux, Peter Filzmoser, and Maria Rosario Oliveira. Algorithms for projection–pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2):218–225, 2007.

Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108, 1963.

Bhasker Dappuri, M. Purnachandra Rao, and Madhu Babu Sikha. Non-blind rgb watermarking approach using svd in translation invariant wavelet space with enhanced grey-wolf optimizer. *Multimedia Tools Appl.*, 79(41–42):31103–31124, nov 2020. ISSN 1380-7501. doi: 10.1007/s11042-020-09433-0.

Lopamudra Das, J. K. Das, and Sarita Nanda. Advanced protein coding region prediction applying robust svd algorithm. In *2017 2nd International Conference on Man and Machine Interfacing (MAMI)*, pages 1–6, 2017. doi: 10.1109/MAMI.2017.8307887.

Despina Dasiou and Chronis Moyssiadis. The 50% breakdown point in simultaneous M-estimation of location and scale. *Statistical Papers*, 42(2):243–252, Apr 2001. ISSN 1613-9798. doi: 10.1007/s003620100053.

P. L. Davies. Asymptotic Behaviour of S-Estimates of Multivariate Location Parameters and Dispersion Matrices. *The Annals of Statistics*, 15(3):1269–1292, 1987. ISSN 00905364.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000. doi: 10.1137/S0895479896305696.

S. J. Devlin, R. Gnandesikan, and J. R. Kettenring. Robust Estimation of Dispersion Matrices and Principal Components. *Journal of the American Statistical Association*, 76(374):354–362, 1981. ISSN 01621459.

Petros Drineas, Alan Frieze, Ravi Kannan, Santosh Vempala, and V Vinay. Clustering large graphs via the singular value decomposition. *Machine learning*, 56(1-3):9–33, 2004.

H. T. Eastment and W. J. Krzanowski. Cross-Validatory Choice of the Number of Components from a Principal Component Analysis. *Technometrics*, 24(1):73–77, 1982. ISSN 0040-1706. doi: 10.2307/1267581. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

N. Benjamin Erichson, Sergey Voronin, Steven L. Brunton, and J. Nathan Kutz. Randomized Matrix Decompositions Using R. *Journal of Statistical Software*, 89(11):1–48, 2019. doi: 10.18637/jss.v089.i11.

K.H. Esbensen, D. Guyot, F. Westad, and L.P. Houmoller. *Multivariate Data Analysis: In Practice : an Introduction to Multivariate Data Analysis and Experimental Design*. CAMO, 2002. ISBN 9788299333030.

Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387303693.

Justin Fishbone and Lamine Mili. New highly efficient high-breakdown estimator of multivariate scatter and location for elliptical distributions. *Canadian Journal of Statistics*, n/a(n/a), 2023. doi: 10.1002/cjs.11770.

Justin A. Fishbone and Lamine Mili. New Highly Efficient High-Breakdown Estimator of Multivariate Scatter and Location for Elliptical Distributions, 2021.

Andrea Franceschini, Jianyi Lin, Christian von Mering, and Lars Juhl Jensen. SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics*, 32(7):1085–1087, 11 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv696.

Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2008.02.004.

K. Ruben Gabriel. Le biplot - outil d'exploration de données multidimensionnelles. *Journal de la Société française de statistique*, 143(3-4):5–55, 2002. ISSN 1625-7421.

K. Ruben Gabriel and S. Zamir. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21(4):489–498, 1979. ISSN 00401706.

Belmar Garcia-Garcia, Thierry Bouwmans, and Alberto Jorge Rosales Silva. Background Subtraction in Real Applications: Challenges, Current Models and Future Directions. *Computer Science Review*, 35:100204, 2020. ISSN 1574-0137. doi: https://doi.org/10.1016/j.cosrev.2019.100204.

Abhik Ghosh. Asymptotic Properties of Minimum S-Divergence Estimator for Discrete Models. *Sankhya A*, 77(2):380–407, Aug 2015. ISSN 0976-8378. doi: 10.1007/s13171-014-0063-2.

Abhik Ghosh and Ayanendranath Basu. Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. *Electronic Journal of Statistics*, 7(none):2420 – 2456, 2013. doi: 10.1214/13-EJS847.

Abhik Ghosh, Ian R. Harris, Avijit Maji, Ayanendranath Basu, and Leandro Pardo. A generalized divergence for statistical inference. *Bernoulli*, 23(4A):2746–2783, 2017.

Jhony H. Giraldo, Sajid Javed, Naoufel Werghi, and Thierry Bouwmans. Graph CNN for Moving Object Detection in Complex Environments from Unseen Videos. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 225–233, 2021. doi: 10.1109/ICCVW54120.2021.00030.

Jhony H. Giraldo, Sajid Javed, and Thierry Bouwmans. Graph Moving Object Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5): 2485–2503, 2022. doi: 10.1109/TPAMI.2020.3042093.

L. Giraud, J. Langou, and M. Rozloznik. The loss of orthogonality in the gram-schmidt orthogonalization process. *Computers & Mathematics with Applications*, 50(7):1069–1075, 2005. ISSN 0898-1221. doi: https://doi.org/10.1016/j.camwa.2005.08.009. Numerical Methods and Computational Mechanics.

MA Girshick. On the sampling theory of roots of determinantal equations. *The Annals of Mathematical Statistics*, 10(3):203–224, 1939.

G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420, Apr 1970. ISSN 0945-3245. doi: 10.1007/BF02163027.

Michael Greenacre. *Correspondence analysis in practice*. CRC press, 2017.

François Grondin and James Glass. Svd-phat: A fast sound source localization method. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4140–4144, 2019. doi: 10.1109/ICASSP.2019.8683253.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53(2):217–288, 2011. doi: 10.1137/090771806.

F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics. Wiley, 2011. ISBN 9781118150689.

Frank R. Hampel. A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971. ISSN 00034851.

Douglas M. Hawkins, Li Liu, and Stanley Young. Robust singular value decomposition. Technical Report 122, National Institute of Statistical Sciences (NISS), 12 2001.

Jun He, Laura Balzano, and John Lui. Online Robust Subspace Tracking from Partial Information. *arXiv preprint arXiv:1109.3827*, 2011.

Jun He, Laura Balzano, and Arthur Szlam. Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1568–1575. IEEE, 2012. doi: 10.1109/CVPR.2012.6247848.

Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135, 2000. ISSN 0047-259X. doi: https://doi.org/10.1006/jmva.1999.1873.

Joseph L. Hodges. Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 163–186, 1967.

Peter D. Hoff. Model Averaging and Dimension Selection for the Singular Value Decomposition. *Journal of the American Statistical Association*, 102(478):674–685, 2007. ISSN 0162-1459.

Jianhua Z. Huang, Haipeng Shen, and Andreas Buja. The Analysis of Two-Way Functional Data Using Two-Way Regularized Singular Value Decompositions. *Journal of the American Statistical Association*, 104(488):1609–1620, 2009. doi: 10.1198/jasa.2009.tm08024.

Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964a. doi: 10.1214/aoms/1177703732.

Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964b. doi: 10.1214/aoms/1177703732.

Peter J. Huber. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799 – 821, 1973. doi: 10.1214/aos/1176342503.

Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

Peter J. Huber and David L. Donoho. The notion of breakdown point. *A Festschrift for Erich L. Lehmann*, 1983.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, 2011. ISBN 9781118210338.

Mia Hubert, Peter J Rousseeuw, and Karlien Vanden Branden. ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, 47(1):64–79, 2005. ISSN 00401706.

Prateek Jain and Purushottam Kar. Non-convex Optimization for Machine Learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017. doi: 10.1561/2200000058.

Soham Jana and Ayanendranath Basu. A characterization of all single-integral, non-kernel divergence estimators. *IEEE Transactions on Information Theory*, 65(12):7976–7984, 2019. doi: 10.1109/TIT.2019.2937527.

Bo Jiang and Yu-Hong Dai. A framework of constraint preserving update schemes for optimization on Stiefel manifold. *Mathematical Programming*, 153(2):535–575, 2015.

Richard Arnold Johnson, Dean W Wichern, et al. Applied multivariate statistical analysis. 2002.

Ian. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer New York, NY, 2002. ISBN 9780387224404. doi: 10.1007/b98835.

M. C. Jones, Nils Lid Hjort, Ian R. Harris, and Ayanendranath Basu. A comparison of related density-based minimum divergence estimators. *Biometrika*, 88(3):865–873, 2001. ISSN 00063444.

Kang-Mo Jung. Robust singular value decomposition balsed on weighted least absolute deviation regression. *Communications for Statistical Applications and Methods*, 17(6): 803–810, 11 2010.

Alex Karagrigoriou and Takis Papaioannou. On measures of information and divergence and model selection criteria. *Statistical Models and Methods for Biomedical and Technical Systems*, pages 503–518, 2008.

Qifa Ke and Takeo Kanade. Robust l1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *Proceedings of (CVPR) Computer Vision and Pattern Recognition*, volume 1, pages 739 – 746, June 2005. ISBN 0-7695-2372-2. doi: 10.1109/CVPR.2005.309.

Arun Kumar Kuchibhotla, Somabha Mukherjee, and Ayanendranath Basu. Statistical inference based on bridge divergences. *Annals of the Institute of Statistical Mathematics*, 71:627–656, 2019.

Nishith Kumar, Mohammed Nasser, and Subaran Chandra Sarker. A new singular value decomposition based robust graphical clustering technique and its application in climatic data. *Journal of Geography and Geology*, 3(1):227, 2011.

Sachin Kumar and Zafar Ahmed. New Spectral Statistics for Ensembles of 2x2 Real Symmetric Random Matrices. *Acta Polytechnica*, 57(6):418, Dec 2017. doi: 10.14311/ap.2017.57.0418.

Sumito Kurata. On robustness of model selection criteria based on divergence measures: Generalizations of bhhj divergence-based method and comparison. *Communications in Statistics - Theory and Methods*, 0(0):1–18, 2022. doi: 10.1080/03610926.2022.2155788.

Sumito Kurata and Etsuo Hamada. A robust generalization and asymptotic properties of the model selection criterion family. *Communications in Statistics - Theory and Methods*, 47(3):532–547, 2018. doi: 10.1080/03610926.2017.1307405.

Sumito Kurata and Etsuo Hamada. On the consistency and the robustness in model selection criteria. *Communications in Statistics - Theory and Methods*, 49(21):5175–5195, 2020. doi: 10.1080/03610926.2019.1615093.

Yann-A Le Borgne and Gianluca Bontempi. Machine Learning for Credit Card Fraud Detection-Practical Handbook. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

B Lee and M Hedley. Background estimation for video surveillance. In *Proc. of the Image & Vision Computing New Zealand*, IVCNZ, pages 315–320, Auckland, NZ, 2002.

Dar-Shyang Lee. Effective Gaussian mixture learning for video background subtraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):827–832, 2005. doi: 10.1109/TPAMI.2005.102.

Guoying Li and Zhonglian Chen. Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo. *Journal of the American Statistical Association*, 80(391):759–766, 1985. ISSN 01621459.

Jun Li, Li Fuxin, and Sinisa Todorovic. Efficient Riemannian optimization on the Stiefel manifold via the Cayley transform. *arXiv preprint arXiv:2002.01113*, 2020.

Zilong Li. *pcaone: Randomized Singular Value Decomposition Algorithms with 'RcppEigen'*, 2022. R package version 1.0.0.

Zilong Li, Jonas Meisner, and Anders Albrechtsen. Fast and Accurate Out-of-Core PCA Framework for Large Scale Biobank Data. *Genome Research*, 33(9):1599–1608, 2023.

Zhouchen Lin, Minming Chen, and Yi Ma. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *arXiv preprint arXiv:1009.5055*, 2010.

Jing Liu, Pamela C Cosman, and Bhaskar D Rao. Sparsity Regularized Principal Component Pursuit. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4431–4435. IEEE, 2017.

Li Liu, Douglas M. Hawkins, Sujoy Ghosh, and S. Stanley Young. Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences*, 100(23):13167–13172, 2003. doi: 10.1073/pnas.1733249100.

N Locantore, JS Marron, DG Simpson, N Tripoli, JT Zhang, KL Cohen, Graciela Boente, Ricardo Fraiman, Babette Brumback, Christophe Croux, et al. Robust principal component analysis for functional data. *Test*, 8(1):1–73, 1999.

Eric F. Lock, Katherine A. Hoadley, J. S. Marron, and Andrew B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523 – 542, 2013. doi: 10.1214/12-AOAS597.

Lucia Maddalena and Alfredo Petrosino. The SOBS algorithm: What are the limits? In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 21–26, 2012. doi: 10.1109/CVPRW.2012.6238922.

Avijit Maji. *Some Contributions to Robust Inference based on Statistical Distances*. Doctoral Thesis, Indian Statistical Institute, Kolkata, India, March 2019.

Avijit Maji, Abhik Ghosh, and Ayanendranath Basu. The logarithmic super divergence and its use in statistical inference. *arXiv preprint arXiv:1407.3961*, 2014.

Murari Mandal and Santosh Kumar Vipparthi. An Empirical Review of Deep Learning Frameworks for Change Detection: Model Design, Experimental Frameworks, Challenges and Research Needs. *Trans. Intell. Transport. Sys.*, 23(7):6101–6122, jul 2022. ISSN 1524-9050. doi: 10.1109/TITS.2021.3077883.

Pranav Mantini and Shishir K Shah. Camera Tampering Detection using Generative Reference Model and Deep Learned Features. In *VISIGRAPP (5: VISAPP)*, pages 85–95, 2019a.

Pranav Mantini and Shishir K. Shah. UHCTD: A Comprehensive Dataset for Camera Tampering Detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2019b. doi: 10.1109/AVSS.2019.8909856.

Ivan Markovsky and Sabine Van Huffel. Overview of total least-squares methods. *Signal Processing*, 87(10):2283–2302, 2007. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2007.04.004. Special Section: Total Least Squares and Errors-in-Variables Modeling.

R.A. Maronna, R.D. Martin, V.J. Yohai, and M. Salibián-Barrera. *Robust Statistics: Theory and Methods (with R)*. Wiley Series in Probability and Statistics. Wiley, 2019. ISBN 9781119214687.

Ricardo Antonio Maronna. Robust M-Estimators of Multivariate Location and Scatter. *The Annals of Statistics*, 4(1):51 – 67, 1976. doi: 10.1214/aos/1176343347.

Alan M McIvor. Background Subtraction Techniques. *Proc. of Image and Vision Computing*, 4:3099–3104, 2000.

M.L. Mehta. *Random Matrices*. ISSN. Elsevier Science, 2004. ISBN 9780080474113.

Stefano Messelodi, Carla Maria Modena, N. Segata, and Michele Zanin. A Kalman Filter Based Background Updating Algorithm Robust to Sharp Illumination Changes. In *ICIAP*, 2005.

Marc Moonen and Bart De Moor. *SVD and Signal Processing, III: Algorithms, Architectures and Applications*. Elsevier, 1995.

Tetsuzo Morimoto. Markov Processes and the H-Theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, March 1963. doi: 10.1143/JPSJ.18.328.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Art B. Owen and Patrick O. Perry. Bi-cross-validation of the SVD and the nonnegative matrix factorization. *The Annals of Applied Statistics*, 3(2):564–594, June 2009. ISSN 1932-6157, 1941-7330. doi: 10.1214/08-AOAS227. Publisher: Institute of Mathematical Statistics.

Chanseok Park and Ayanendranath Basu. Minimum disparity estimation: Asymptotic normality and breakdown point results. *Bulletin of Informatics and Cybernetics*, pages 19–33, 2004. doi: 10.5109/12576.

Chanseok Park, Haewon Kim, and Min Wang. Investigation of Finite-sample Properties of Robust Location and Scale Estimators. *Communications in Statistics - Simulation and Computation*, 51(5):2619–2645, 2022. doi: 10.1080/03610918.2019.1699114.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The Matrix Cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Stephen Portnoy. Asymptotic Behavior of M-Estimators of p Regression Parameters when p^2/n is Large. I. Consistency. *The Annals of Statistics*, 12(4):1298 – 1309, 1984. doi: 10.1214/aos/1176346793.

Arijit Pyne, Subhrajyoty Roy, Abhik Ghosh, and Ayanendranath Basu. Robust and efficient estimation in ordinal response models using the density power divergence, 2024.

Souvik Ray, Subrata Pal, Sumit Kumar Kar, and Ayanendranath Basu. Characterizing the functional density power divergence class. *IEEE Transactions on Information Theory*, 69(2):1141–1146, 2023. doi: 10.1109/TIT.2022.3210436.

William Rey. Total singular value decomposition. robust svd, regression and location-scale. 2007.

Azadeh Rezaei and Mehdi Khalili. A robust blind audio watermarking scheme based on dct-dwt-svd. In Shahram Montaser Kouhsari, editor, *Fundamental Research in Electrical Engineering*, pages 101–113, Singapore, 2019. Springer Singapore. ISBN 978-981-10-8672-4.

Marco Riani, Andrea Cerioli, Anthony C. Atkinson, and Domenico Perrotta. Monitoring robust regression. *Electronic Journal of Statistics*, 8(1):646 – 677, 2014. doi: 10.1214/14-EJS897.

Marco Riani, Anthony C Atkinson, Aldo Corbellini, and Domenico Perrotta. Robust Regression with Density Power Divergence: Theory, Comparisons, and Data Analysis. *Entropy*, 22(4):399, 2020.

Andrew M Ross. Computing Bounds on the Expected Maximum of Correlated Normal Variables. *Methodology and Computing in Applied Probability*, 12:111–138, 2010.

Peter J. Rousseeuw. Multivariate Estimation with High Breakdown Point. *Mathematical Statistics and Applications*, 8(283-297):37, 1985.

Peter J. Rousseeuw and Victor J. Yohai. Robust Regression by Means of S-Estimators. In Jürgen Franke, Wolfgang Härdle, and Douglas Martin, editors, *Robust and Nonlinear Time Series Analysis*, pages 256–272, New York, NY, 1984. Springer US. ISBN 978-1-4615-7821-5.

Subhrajyoty Roy. *rsvddpd: Robust Singular Value Decomposition using Density Power Divergence*, 2021. R package version 1.0.0.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.

Walter Rudin. *Real and complex analysis*. McGraw-Hill, Inc., 1987.

Rowayda A Sadek. Svd based image processing applications: state of the art, contributions and research challenges. *International Journal of Advanced Computer Science and Applications*, 3(7):26–34, 2012. doi: 10.48550/arXiv.1211.7102.

P. Sanguansat. *Principal Component Analysis: Engineering Applications*. IntechOpen, 2012. ISBN 9789535101826.

Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2): 461 – 464, 1978. doi: 10.1214/aos/1176344136.

Andrey A. Shabalin and Andrew B. Nobel. Reconstruction of a low-rank matrix in the presence of gaussian noise. *Journal of Multivariate Analysis*, 118:67–76, 2013. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2013.03.005.

Y. Shen, Z. Wen, and Y. Zhang. Augmented Lagrangian Alternating Direction Method for Matrix Separation based on Low-rank Factorization. *Optimization Methods and Software*, 29(2):239–263, 2014. doi: 10.1080/10556788.2012.700713.

Douglas G. Simpson. Minimum Hellinger Distance Estimation for the Analysis of Count Data. *Journal of the American Statistical Association*, 82(399):802–807, 1987. ISSN 01621459.

K. Sitara and B. M. Mehtre. Automated camera sabotage detection for enhancing video surveillance systems. *Multimedia Tools and Applications*, 78(5):5819–5841, Mar 2019. ISSN 1573-7721. doi: 10.1007/s11042-018-6165-4.

A.K. Smilde, R. Bro, and P. Geladi. *Multi-way Analysis: Applications in the Chemical Sciences*. Wiley InterScience online books. Wiley, 2005. ISBN 9780470012109.

Ezequiel Smucler. Asymptotics for Redescending M-estimators in Linear Models with Increasing Dimension. *Statistica Sinica*, 29(2):1065–1081, 2019. ISSN 10170405, 19968507.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023. doi: 10.1109/TNNLS.2022.3152527.

Wolfram Stacklies, Henning Redestig, Matthias Scholz, Dirk Walther, and Joachim Selbig. pcaMethods – a Bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23:1164–1167, 2007.

Alessandra Stagliano, Nicoletta Noceti, Alessandro Verri, and Francesca Odone. Online Space-Variant Background Modeling With Sparse Coding. *IEEE Transactions on Image Processing*, 24(8):2415–2428, 2015. doi: 10.1109/TIP.2015.2421435.

Badri Narayan Subudhi, Manoj Kumar Panda, T. Veerakumar, Vinit Jakhetiya, and S. Esakkirajan. Kernel-Induced Possibilistic Fuzzy Associate Background Subtraction for Video Scene. *IEEE Transactions on Computational Social Systems*, pages 1–12, 2022. doi: 10.1109/TCSS.2021.3137306.

Maciek Sykulski. *rpca: RobustPCA: Decompose a Matrix into Low-Rank and Sparse Components*, 2015. R package version 0.2.3.

Roy N. Tamura and Dennis D. Boos. Minimum Hellinger Distance Estimation for Multivariate Location and Covariance. *Journal of the American Statistical Association*, 81 (393):223–229, 1986. ISSN 01621459.

Terence Tao. *Topics in Random Matrix Theory*, volume 132. American Mathematical Society, 2012.

Robert L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, Dec 1953. ISSN 1860-0980. doi: 10.1007/BF02289263.

Valentin Todorov and Peter Filzmoser. An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32:1–47, 2010.

Aida Toma. Minimum Hellinger distance estimators for some multivariate models: influence functions and breakdown point results. *Comptes Rendus Mathematique*, 345(6): 353–358, 2007. ISSN 1631-073X. doi: https://doi.org/10.1016/j.crma.2007.07.024.

Aida Toma. Minimum hellinger distance estimators for multivariate distributions from the johnson system. *Journal of Statistical Planning and Inference*, 138(3):803–816, 2008. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2007.05.033.

Aida Toma, Alex Karagrigoriou, and Paschalini Trentou. Robust model selection criteria based on pseudodistances. *Entropy*, 22(3), 2020. ISSN 1099-4300. doi: 10.3390/e22030304.

Craig A. Tracy and Harold Widom. Introduction to random matrices. In G. F. Helminck, editor, *Geometric and Quantum Aspects of Integrable Systems*, pages 103–130, Berlin, Heidelberg, 1993. Springer Berlin Heidelberg. ISBN 978-3-540-48090-7.

J.W. Tukey. *Exploratory Data Analysis*. Number v. 2 in Addison-Wesley series in behavioral science. Addison-Wesley Publishing Company, 1977. ISBN 9780201076165.

David E Tyler. Asymptotic inference for eigenvectors. *The Annals of Statistics*, 9(4): 725–736, 1981.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

Antoine Vacavant, Thierry Chateau, Alexis Wilhelm, and Laurent Lequièvre. A Benchmark Dataset for Outdoor Foreground/Background Extraction. In Jong-Il Park and Junmo Kim, editors, *Computer Vision - ACCV 2012 Workshops*, pages 291–300, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37410-4.

Richard J. Vaccaro. *SVD and Signal Processing II: Algorithms, Analysis and Applications*. Elsevier Science Inc., USA, 1991. ISBN 0444888969.

Yehuda Vardi and Cun-Hui Zhang. The multivariate l1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2000.

Namrata Vaswani, Thierry Bouwmans, Sajid Javed, and Praneeth Narayanamurthy. Robust Subspace Learning: Robust PCA, Robust Subspace Tracking, and Robust Subspace Recovery. *IEEE Signal Processing Magazine*, 35(4):32–55, July 2018. doi: 10.1109/MSP.2018.2826566.

Ágnes Vathy-Fogarassy and János Abonyi. *Graph-based clustering and data visualization algorithms*. Springer, 2013.

Deshen Wang. Adjustable robust singular value decomposition: Design, analysis and application to finance. *Data*, 2(3), 2017. ISSN 2306-5729. doi: 10.3390/data2030029.

Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1):397–434, 2013.

Svante Wold. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics*, 20(4):397–405, November 1978. ISSN

0040-1706. doi: 10.1080/00401706.1978.10489693. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1978.10489693.

John Wright, Arvind Ganesh, Shankar R Rao, Yigang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *NIPS*, volume 58, pages 289–298, 2009.

Lanyu Xiong and Fukang Zhu. Minimum Density Power Divergence Estimator for Negative Binomial Integer-Valued GARCH Models. *Communications in Mathematics and Statistics*, pages 1–29, April 2021. doi: 10.1007/s40304-020-00221-8.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via Outlier Pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, 2012. doi: 10.1109/TIT.2011.2173156.

Zhengqin Xu, Rui He, Shoulie Xie, and Shiqian Wu. Adaptive rank estimate in robust principal component analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6577–6586, 2021.

Dan Yang. *Singular Value Decomposition for High Dimensional Data*. PhD thesis, University of Pennsylvania, 2014.

Anru Zhang and Rungang Han. Optimal sparse singular value decomposition for high-dimensional high-order data. *Journal of the American Statistical Association*, 2019.

Anru Zhang and Dong Xia. Tensor svd: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018.

Dashan Zhang, Jie Guo, Xiujun Lei, and Chang'an Zhu. Note: Sound recovery from video using svd-based information extraction. *Review of Scientific Instruments*, 87(8):086111, 2016a. doi: 10.1063/1.4961979.

Hongyi Zhang, Sashank J Reddi, and Suvrit Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. *Advances in Neural Information Processing Systems*, 29:4592–4600, 2016b.

Lingsong Zhang and Chao Pan. *RobRSVD: Robust Regularized Singular Value Decomposition*, 2013. R package version 1.0.

Lingsong Zhang, Haipeng Shen, and Jianhua Z. Huang. Robust regularized singular value decomposition with application to mortality data. *The Annals of Applied Statistics*, 7 (3):1540 – 1561, 2013. doi: 10.1214/13-AOAS649.

Ming Zhao and Xiaodong Jia. A novel strategy for signal denoising using reweighted svd and its applications to weak fault feature enhancement of rotating machinery. *Mechanical Systems and Signal Processing*, 94:129–147, 2017. ISSN 0888-3270. doi: https://doi.org/10.1016/j.ymssp.2017.02.036.

Zhenjie Zhao, Thierry Bouwmans, Xuebo Zhang, and Yongchun Fang. A Fuzzy Background Modeling Approach for Motion Detection in Dynamic Backgrounds. In *International Conference on Multimedia and Signal Processing*, pages 177–185. Springer, 2012.

Tianyi Zhou and Dacheng Tao. GoDec: Randomized Low-Rank & Sparse Matrix Decomposition in Noisy Case. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 33–40, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Zihan Zhou, Xiaodong Li, John Wright, Emmanuel Candès, and Yi Ma. Stable Principal Component Pursuit. In *2010 IEEE International Symposium on Information Theory*, pages 1518–1522, 2010. doi: 10.1109/ISIT.2010.5513535.