

RICE UNIVERSITY

**Robust Parametric Functional Component  
Estimation Using a Divergence Family**

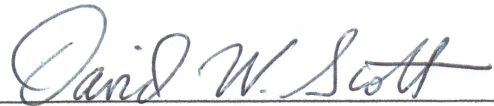
by

**Justin Lee Silver**

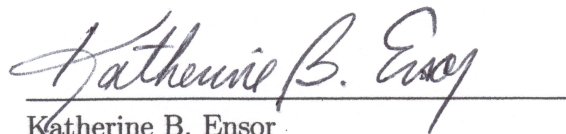
A THESIS SUBMITTED  
IN PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE

**Doctor of Philosophy**

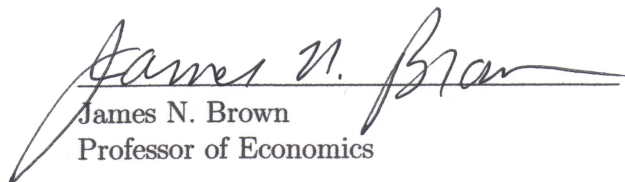
APPROVED, THESIS COMMITTEE:



David W. Scott, Chair  
Noah Harding Professor of Statistics



Katherine B. Ensor  
Professor of Statistics



James N. Brown  
Professor of Economics

HOUSTON, TEXAS  
APRIL, 2013

# Abstract

Robust Parametric Functional Component Estimation Using a Divergence Family

by

Justin Silver

The classical parametric estimation approach, maximum likelihood, while providing maximally efficient estimators at the correct model, lacks robustness. As a modification of maximum likelihood, Huber (1964) introduced M-estimators, which are very general but often ad hoc. Basu et al. (1998) developed a family of density-based divergences, many of which exhibit robustness. It turns out that maximum likelihood is a special case of this general class of divergence functions, which are indexed by a parameter  $\alpha$ . Basu noted that only values of  $\alpha$  in the  $[0, 1]$  range were of interest – with  $\alpha = 0$  giving the maximum likelihood solution and  $\alpha = 1$  the  $L_2E$  solution (Scott, 2001). As  $\alpha$  increases, there is a clear tradeoff between increasing robustness and decreasing efficiency. This thesis develops a family of robust location and scale estimators by applying Basu’s  $\alpha$ -divergence function to a multivariate partial density component model (Scott, 2004). The usefulness of  $\alpha$  values greater than 1 will be explored, and the new estimator will be applied to simulated cases and applications in parametric density estimation and regression.

# Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. David W. Scott, for his invaluable guidance, patience and encouragement. This research would not have been possible without him. I would also like to thank Dr. Katherine B. Ensor for serving on my thesis committee, for her continued support and for providing me with teaching opportunities, which further enriched my graduate experience. I also wish to thank Dr. James N. Brown for serving on my thesis committee and for mentoring me in my undergraduate studies. He showed me the power of a teacher who can truly connect with his students.

Many thanks to the faculty, staff and my fellow students of the Department of Statistics at Rice for providing me with a strong and close-knit community of support. In particular, I would like to express my appreciation for our department coordinators, Margaret, Phyllis and Carolyn, for their tireless work that often goes unrecognized.

Last but not least, I would like to thank my family and friends for always supporting me and encouraging me to keep working hard. I could not have possibly completed this thesis without all of you being by my side.

*Dedicated to Papa, Nanny and Lou. I know you would all be proud to see this.*

# Contents

Abstract . . . . .	ii
Acknowledgments . . . . .	iii
List of Figures . . . . .	viii
List of Tables . . . . .	xi
<b>1 Introduction and Background</b>	<b>1</b>
1.1 Motivation for Current Robust Methodology . . . . .	2
1.1.1 Parametric Estimation . . . . .	2
1.1.2 Breakdown Point . . . . .	10
1.1.3 Regression . . . . .	11
1.2 $\alpha$ -Divergence Function . . . . .	18
1.3 Multivariate Partial Density Component (MPDC) . . . . .	19
<b>2 MPDC-<math>\alpha</math> Divergence Estimator</b>	<b>20</b>
2.1 Derivation of MPDC- $\alpha$ Divergence Estimator . . . . .	20
2.2 Parallel to M-Estimators . . . . .	22
2.3 Properties of the MPDC- $\alpha$ Divergence Estimator . . . . .	23
2.3.1 Consistency and Asymptotic Normality of $\hat{w}_\alpha$ . . . . .	23

2.3.2	Consistency and Asymptotic Normality of $\hat{\mu}_\alpha$ and $\hat{\Sigma}_\alpha$ . . . . .	25
2.3.3	Consistency and Asymptotic Normality of $\hat{\theta}_\alpha$ . . . . .	25
2.3.4	Invariance Under Linear Transformation . . . . .	26
<b>3</b>	<b>Parametric Density Estimation</b> . . . . .	<b>29</b>
3.1	Parameter and Criterion Definition . . . . .	29
3.2	Parameter Transformations . . . . .	30
3.2.1	$w$ : <i>logit</i> transformation . . . . .	30
3.2.2	$\Sigma^{-1}$ : Cholesky Decomposition and <i>exp</i> transformation . . . . .	31
3.2.3	Computationally Efficient Criterion . . . . .	32
3.3	Simulated Cases . . . . .	32
3.3.1	Simulated Cases for $p = 2$ . . . . .	32
3.3.2	Simulated Cases for $p = 3$ . . . . .	49
3.3.3	Simulated Cases for $p = 4$ . . . . .	53
3.3.4	Simulated Cases for $p = 5$ . . . . .	57
3.3.5	Simulated Cases for $6 \leq p \leq 10$ . . . . .	61
3.3.6	Selecting $\alpha$ . . . . .	61
3.4	Application: Baseball . . . . .	69
3.4.1	Robust Parametric Estimation for PITCHf/x Variables . . . . .	69
<b>4</b>	<b>Asymptotics</b> . . . . .	<b>80</b>
4.1	Motivation: $p = 1$ . . . . .	80
4.1.1	Asymptotic Distribution of $\hat{\mu}_\alpha$ , Known $\sigma$ . . . . .	81
4.1.2	Asymptotic Distribution of $\hat{\sigma}_\alpha$ , Unknown $\sigma$ . . . . .	82

4.2	General $p$ . . . . .	83
4.2.1	Asymptotic Distribution of $\hat{\mu}_\alpha$ , Known $\Sigma$ . . . . .	83
4.2.2	Asymptotic Distribution of $\hat{\Sigma}_\alpha$ , $\Sigma = \sigma^2 I_p$ , $\sigma$ Unknown . . . . .	84
4.2.3	Asymptotic Distribution of $(\hat{\sigma}_{1,\alpha}, \dots, \hat{\sigma}_{p,\alpha})'$ , Unknown $\Sigma$ . . . . .	87
4.2.4	Asymptotic Distribution of $\hat{\rho}_\alpha$ , Unknown $\rho$ . . . . .	87
4.3	Simulation to Verify Asymptotic Results . . . . .	88
4.3.1	Verification of Asymptotic Distribution of $\hat{w}_\alpha$ . . . . .	91
4.3.2	Verification of Asymptotic Distribution of $\hat{\mu}_\alpha$ . . . . .	91
4.3.3	Verification of Asymptotic Distribution of $\hat{\sigma}_{j,\alpha}$ . . . . .	92
4.3.4	Verification of Asymptotic Distribution of $\hat{\rho}_{12,\alpha}$ . . . . .	92
<b>5</b>	<b>Regression</b> . . . . .	<b>93</b>
5.1	Criterion Definition . . . . .	94
5.2	Parameter Transformations . . . . .	95
5.2.1	$w$ : <i>logit</i> transformation . . . . .	95
5.2.2	$\sigma_\epsilon$ : <i>exp</i> transformation . . . . .	96
5.3	Simulated Cases . . . . .	96
5.4	Mixed Quadratic Example . . . . .	102
<b>6</b>	<b>Conclusions and Future Work</b> . . . . .	<b>107</b>
	<b>References</b> . . . . .	<b>109</b>
	<b>Appendix</b> . . . . .	<b>111</b>
	Optimizer: <i>nlm</i> . . . . .	111

Parametric Estimation Trace Plots for $6 \leq p \leq 10$ . . . . .	111
Parametric Estimation RMSE Plots for $6 \leq p \leq 10$ . . . . .	122
R Code . . . . .	127

# List of Figures

1.1	MLE Example . . . . .	4
1.2	$L_2E$ Example . . . . .	5
1.3	MVE/MCD Example . . . . .	8
1.4	Motivating Regression Example 1 - Single Outlier . . . . .	13
1.5	Motivating Regression Example 2 - Outlying Cluster . . . . .	15
1.6	Motivating Regression Example 3 - $L_2E$ Breakdown . . . . .	17
3.1	Pure Sample ( $p = 2$ ): Contour Plot . . . . .	34
3.2	Pure Sample ( $p = 2$ ): Trace Plots . . . . .	35
3.3	Overlap with Zero Correlation ( $p = 2$ ): Contour Plot . . . . .	37
3.4	Overlap with Zero Correlation ( $p = 2$ ): Trace Plots . . . . .	38
3.5	Overlap with Correlation ( $p = 2$ ): Contour Plot . . . . .	40
3.6	Overlap with Correlation ( $p = 2$ ): Trace Plots . . . . .	42
3.7	Separation with Zero Correlation ( $p = 2$ ): Contour Plot . . . . .	44
3.8	Separation with Zero Correlation ( $p = 2$ ): Trace Plots . . . . .	45
3.9	Separation with Correlation ( $p = 2$ ): Contour Plot . . . . .	47
3.10	Separation with Correlation ( $p = 2$ ): Trace Plots . . . . .	48



3.11	Overlap with Correlation ( $p = 3$ ): Trace Plots . . . . .	50
3.12	Separation with Correlation ( $p = 3$ ): Trace Plots . . . . .	52
3.13	Overlap with Correlation ( $p = 4$ ): Trace Plots . . . . .	54
3.14	Separation with Correlation ( $p = 4$ ): Trace Plots . . . . .	56
3.15	Overlap with Correlation ( $p = 5$ ): Trace Plots . . . . .	58
3.16	Separation with Correlation ( $p = 5$ ): Trace Plots . . . . .	60
3.17	RMSE Plot for $\alpha$ Selection ( $p = 2$ ) . . . . .	63
3.18	RMSE Plot for $\alpha$ Selection ( $p = 3$ ) . . . . .	64
3.19	RMSE Plot for $\alpha$ Selection ( $p = 4$ ) . . . . .	65
3.20	RMSE Plot for $\alpha$ Selection ( $p = 5$ ) . . . . .	66
3.21	Horizontal and Vertical Movement: Contour Plot . . . . .	72
3.22	Horizontal and Vertical Movement: Trace Plots . . . . .	73
3.23	5 PITCHf/x Variables: Trace Plots . . . . .	76
3.24	Pairwise Scatterplot Matrix for 5 PITCHF/x Variables . . . . .	78
4.1	Sampling Distributions for Components of $\hat{\theta}_\alpha$ . . . . .	90
5.1	Regression Example 1 - Outlying Cluster: Trace Plots . . . . .	97
5.2	Regression Example 1 - Outlying Cluster: Solutions . . . . .	98
5.3	Regression Example 2 - 50% Overlapping Contamination: Trace Plots	100
5.4	Regression Example 2 - 50% Overlapping Contamination: Solutions .	101
5.5	Mixed Quadratic Example: Sample Plot . . . . .	103
5.6	Mixed Quadratic Example: Random Starts . . . . .	105

A.1	Overlap with Correlation ( $p = 6$ ): Trace Plots . . . . .	112
A.2	Separation with Correlation ( $p = 6$ ): Trace Plots . . . . .	113
A.3	Overlap with Correlation ( $p = 7$ ): Trace Plots . . . . .	114
A.4	Separation with Correlation ( $p = 7$ ): Trace Plots . . . . .	115
A.5	Overlap with Correlation ( $p = 8$ ): Trace Plots . . . . .	116
A.6	Separation with Correlation ( $p = 8$ ): Trace Plots . . . . .	117
A.7	Overlap with Correlation ( $p = 9$ ): Trace Plots . . . . .	118
A.8	Separation with Correlation ( $p = 9$ ): Trace Plots . . . . .	119
A.9	Overlap with Correlation ( $p = 10$ ): Trace Plots . . . . .	120
A.10	Separation with Correlation ( $p = 10$ ): Trace Plots . . . . .	121
A.11	RMSE Plot for $\alpha$ Selection ( $p = 6$ ) . . . . .	122
A.12	RMSE Plot for $\alpha$ Selection ( $p = 7$ ) . . . . .	123
A.13	RMSE Plot for $\alpha$ Selection ( $p = 8$ ) . . . . .	124
A.14	RMSE Plot for $\alpha$ Selection ( $p = 9$ ) . . . . .	125
A.15	RMSE Plot for $\alpha$ Selection ( $p = 10$ ) . . . . .	126

# List of Tables

1.1	Comparison of Existing Robust Methods . . . . .	9
3.1	Guidelines for Selection of $\alpha$ for Various Dimension ( $p$ ) Values . . . . .	68
3.2	PITCHf/x: Horizontal and Vertical Movement . . . . .	70
3.3	PITCHf/x: Horizontal & Vertical Movement, Speed, Break, Spin . . . . .	75
3.4	Comparison of Fastballs and Breaking Balls . . . . .	79

# Chapter 1

## Introduction and Background

The detection and management of outliers remains a fundamental problem in statistics; robust methods yield parameter estimates that are unaffected by outlying points or clusters. Obtaining a robust estimator of the covariance matrix is challenging, particularly as the number of covariates increases, since most existing robust approaches such as the minimum volume ellipse (MVE) are combinatorial solutions requiring extensive computing power. Novel robust methods that can remain computationally efficient even as dimension increases, while providing consistent solutions, are needed. This thesis provides one such approach.

The classical parametric estimation approach, Maximum Likelihood Estimation, while providing maximally efficient estimators at the correct model, lacks robustness. It turns out that the MLE is a special case ( $\alpha = 0$ ) of a general class of divergence functions indexed by a parameter  $\alpha$ . Basu et al. (1998) explored these “density-based divergences” that, unlike nonparametric density estimation methods based on minimum distance, estimate parameters by minimizing a data-based estimate of a function

which measures the divergence between the assumed model density,  $f$ , and the true density,  $g$ . Basu notes that only values of  $\alpha$  in the  $[0, 1]$  range are of interest – with  $\alpha = 0$  giving the MLE solution and  $\alpha = 1$  the  $L_2E$  solution (Scott, 2001). When  $\alpha = 1$ , the  $\alpha$ -divergence criterion is both a minimum divergence and a minimum distance criterion. As  $\alpha$  increases, there is a clear tradeoff between decreasing efficiency and increasing robustness.

This thesis develops the MPDC- $\alpha$  divergence estimator, which combines Basu’s  $\alpha$ -divergence function with a multivariate partial density component (MPDC) model (Scott, 2004). The usefulness of  $\alpha$  values greater than 1 will be explored, and the MPDC- $\alpha$  estimator will be applied to simulated cases and applications in parametric estimation and regression.

## 1.1 Motivation for Current Robust Methodology

We begin by examining the usefulness of robust statistics, presenting a set of motivating examples and exploring how existing methods perform in those situations. We will also explore the amount of contamination that particular estimators can tolerate before yielding aberrant solutions. In particular, we consider the “breakdown point” of these estimators in supporting the benefit of using the MPDC- $\alpha$  estimator.

### 1.1.1 Parametric Estimation

To motivate our method in the context of parametric estimation, we begin with an example. We wish to generate a bivariate random sample of size  $n = 100$  with the

primary data centered at  $\boldsymbol{\mu}_1 \equiv (0, 0)'$  with covariance  $\boldsymbol{\Sigma}_1 \equiv I_2$  and 10% contamination centered at  $\boldsymbol{\mu}_2 \equiv (6, 0)'$  with covariance  $\boldsymbol{\Sigma}_2 \equiv I_2$ . Thus, we simulate from the mixture distribution:

$$\frac{9}{10}N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_2 \right) + \frac{1}{10}N \left( \begin{pmatrix} 6 \\ 0 \end{pmatrix}, I_2 \right).$$

Suppose we seek to estimate  $\boldsymbol{\mu}_1$  and  $\boldsymbol{\Sigma}_1$ . Maximum likelihood estimation fails to properly account for outlying points or clusters, as seen in Figure 1.1. The green line represents the 2-sigma ellipse of the MLE fit, and it contains many of the blue points, which are outliers. If we remove the outliers and recompute the MLE, we see that our estimate (dashed red 2-sigma ellipse) closely matches the true center and shape (dashed black 2-sigma ellipse). It is clear that we are in need of a robust data-based estimate of  $\boldsymbol{\mu}_1$  and  $\boldsymbol{\Sigma}_1$ , the mean vector and covariance matrix of the targeted density component, which we denote henceforth as  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ , respectively.

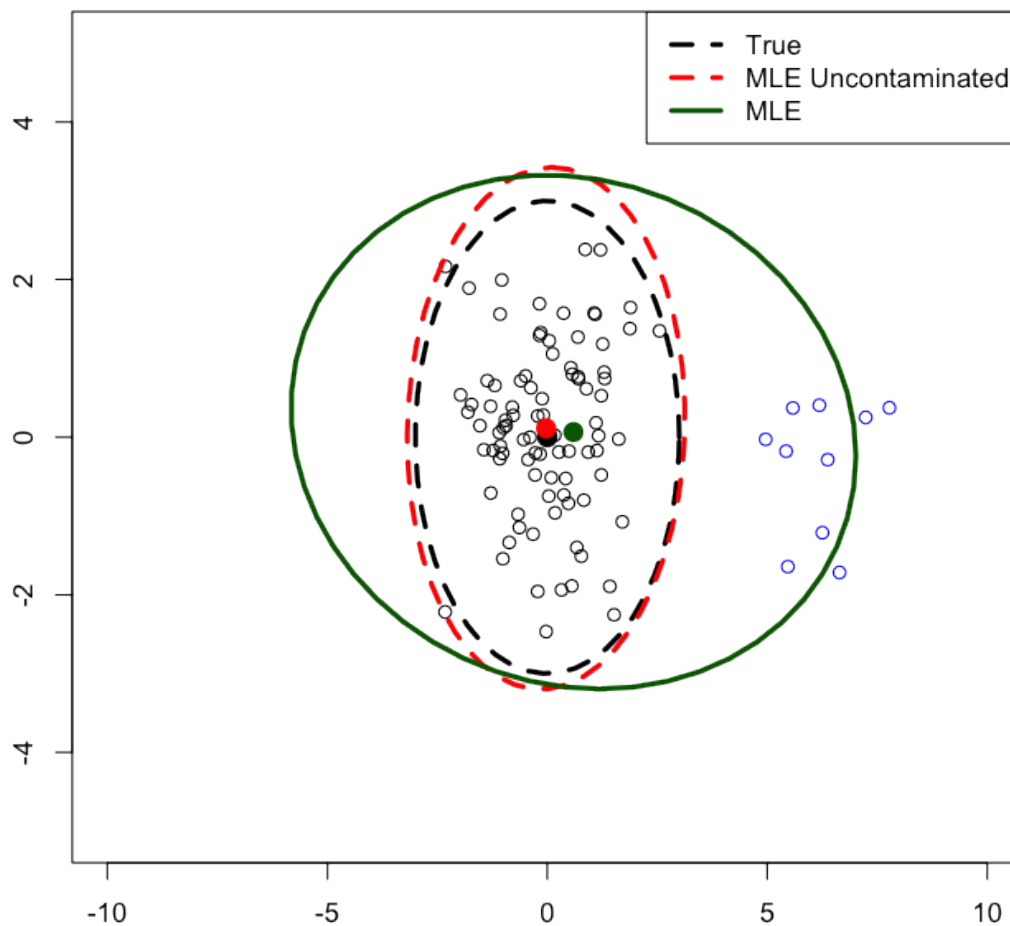


Figure 1.1: MLE estimate of mean (green point) and covariance (green 2-sigma ellipse) of targeted density component for sample of size  $n = 100$  from Normal mixture distribution with parameters:  $w = 0.9; \mu_1 = (0, 0)'; \mu_2 = (6, 0)'; \Sigma_1 = \Sigma_2 = I_2$ . The 2-sigma ellipse for MLE estimate including only uncontaminated data is in red, and the true density's 2-sigma ellipse is in black.

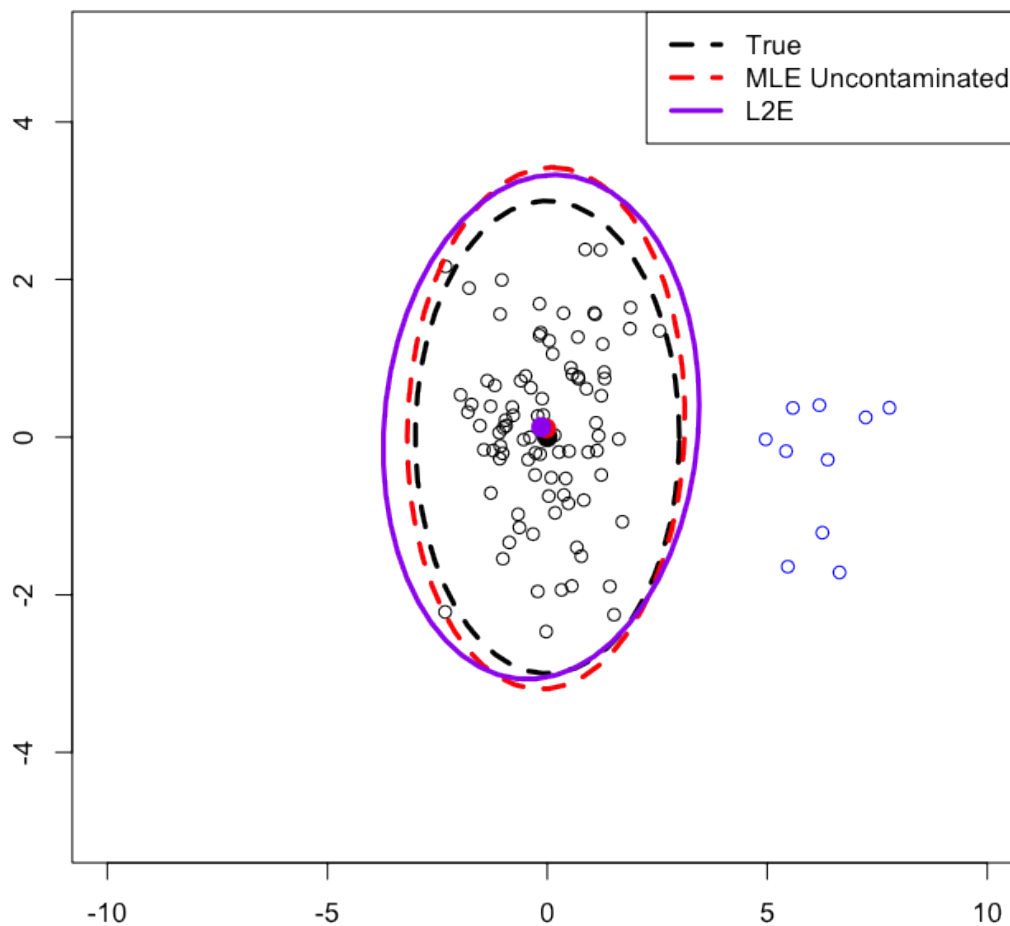


Figure 1.2:  $L_2E$  estimate of mean (purple point) and covariance (2-sigma ellipse, in purple) of targeted density component for sample of size  $n = 100$  from Normal mixture distribution with parameters:  $w = 0.9; \mu_1 = (0, 0)'; \mu_2 = (6, 0)'; \Sigma_1 = \Sigma_2 = I_2$ . The 2-sigma ellipse for MLE estimate including only uncontaminated data is in red, and the true density's 2-sigma ellipse is in black.



The  $L_2E$  solution using all the data, shown in Figure 1.2 with the purple line representing its 2-sigma ellipse, is more robust to the contamination and thus provides an estimate closer to the true solution. However, the variance of the first component for the  $L_2E$  solution is somewhat inflated compared to those of the uncontaminated MLE and true solution, as we can see in Figure 1.2 with the purple ellipse ( $L_2E$ ) having a wider minor axis than the red (MLE on just the uncontaminated data) and black (true value) ellipses. Thus, while the  $L_2E$  robustly estimates the center of the uncontaminated data, there are methods that can more robustly estimate the covariance of the uncontaminated sample. Two such methods are the minimum volume ellipsoid (MVE) and the minimum covariance determinant (MCD), which we now discuss.

### **Minimum Volume Ellipsoid (MVE)**

Rousseeuw's minimum volume ellipsoid (MVE) is the smallest ellipsoid to cover  $h$  of the  $n$  observations, where  $\frac{n}{2} \leq h < n$  (Rousseeuw, 1984). Outliers are identified by points on the boundary of this minimum volume ellipsoid. MVE provides a robust estimate of location and scatter and can be computed using a resampling algorithm. Estimating the MVE requires considerable computing power, presenting a problem which is NP-incomplete.

### **Minimum Covariance Determinant (MCD) and FAST-MCD**

A similar approach, also developed in part by Rousseeuw, is the minimum covariance determinant. Rousseeuw's minimum covariance determinant (MCD) method

provides robust estimates of location and scatter (Rousseeuw, 1984). The MCD seeks the set of  $h$  out of  $n$  points whose covariance matrix has the lowest determinant. The location estimate is then given by the mean of those  $h$  points, and the scatter estimate is the covariance matrix of those  $h$  points. This method also requires non-trivial computation, as it requires the exploration of all possible subsets of size  $h$  out of  $n$ . To remedy this, Rousseeuw and Van Driessen developed the FAST-MCD algorithm, which randomly draws many  $p + 1$  observations from the data and then constructs subsets of size  $h$  via C-steps (Rousseeuw, 1999). Although this provides a significant improvement to the computational efficiency of the MCD method, there is no improvement compared to the MCD in terms of breakdown point, which is the minimum amount of contamination needed for an estimator to “blow up.”

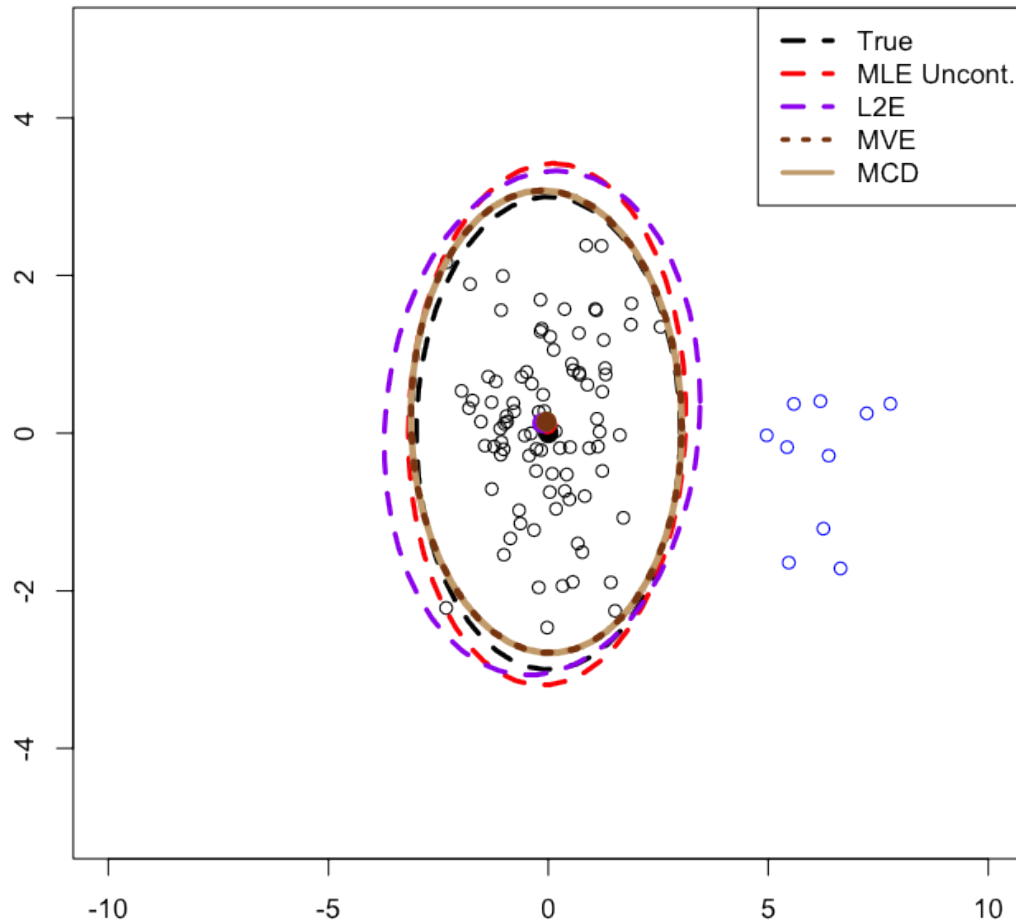


Figure 1.3: MVE and MCD estimates of mean (brown and beige points, respectively) and covariance (2-sigma ellipses, in brown and beige, respectively) of targeted density component for sample of size  $n = 100$  from Normal mixture distribution with parameters:  $w = 0.9$ ;  $\mu_1 = (0, 0)'$ ;  $\mu_2 = (6, 0)'$ ;  $\Sigma_1 = \Sigma_2 = I_2$ . Also shown are 2-sigma ellipses for  $L_2E$  estimate (purple), MLE estimate including only uncontaminated data (red), and true density (black).

Method	$\hat{\boldsymbol{\mu}}$	$\hat{\boldsymbol{\Sigma}}$
MLE	$\begin{pmatrix} 0.594 \\ 0.065 \end{pmatrix}$	$\begin{pmatrix} 4.581 & -0.217 \\ -0.217 & 1.181 \end{pmatrix}$
MLE (uncontaminated)	$\begin{pmatrix} -0.028 \\ 0.113 \end{pmatrix}$	$\begin{pmatrix} 1.108 & 0.047 \\ 0.047 & 1.219 \end{pmatrix}$
$L_2E$	$\begin{pmatrix} -0.140 \\ 0.129 \end{pmatrix}$	$\begin{pmatrix} 1.434 & 0.110 \\ 0.110 & 1.138 \end{pmatrix}$
MVE	$\begin{pmatrix} -0.044 \\ 0.145 \end{pmatrix}$	$\begin{pmatrix} 1.054 & -0.027 \\ -0.027 & 0.957 \end{pmatrix}$
MCD	$\begin{pmatrix} -0.044 \\ 0.145 \end{pmatrix}$	$\begin{pmatrix} 1.054 & -0.027 \\ -0.027 & 0.957 \end{pmatrix}$

Table 1.1: Estimates of mean vector and covariance matrix of targeted density component using various methods for a sample of size  $n = 100$  simulated from a Normal mixture distribution with parameters:  $w = 0.9$ ;  $\boldsymbol{\mu}_1 = (0, 0)'$ ;  $\boldsymbol{\mu}_2 = (6, 0)'$ ;  $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = I_2$ .

Both methods provide robust estimates of location and scatter, as seen in Figure 1.3.

We compare the estimates from all methods in Table 1.1. The MLE is affected considerably by the outlying cluster, having an inflated first component of its mean and variance.  $L_2E$  yields an improved estimate of the mean, but its covariance matrix estimate still differs somewhat from the true covariance, particularly in the first variance element. MVE and MCD yield the same estimates in this example, providing the best robust estimates of both location and scatter.

## M-Estimators

Another robust estimation procedure is the use of M-estimators (Huber, 1964), which are defined by solving the equation:

$$\sum_i \psi(X_i, t) = 0$$

for some real function  $\psi$ . We describe these further as well as their parallel to our method in Section 2.2.

### 1.1.2 Breakdown Point

The breakdown point is the minimum proportion of observations in the data that need to be replaced to push an estimate an arbitrary distance away. This measure directly corresponds to an estimator's robustness. For example, the sample mean  $\bar{x}$  has a breakdown point of  $1/n$  since the presence of one outlying point can make  $\bar{x}$  arbitrarily larger or smaller than it would be without that outlier. Thus, even for large  $n$ , the breakdown point of the sample mean is 0%, confirming that  $\bar{x}$  is not a robust estimator. The median, on the other hand, has a breakdown point of 50%, tolerating up to 50% outliers in the data. Thus, the median is a much more robust estimator than the sample mean.

The breakdown point of an M-estimator is  $\frac{1}{p+1}$ , where  $p$  is the number of parameters (Maronna, 1976). Thus, as the dimension of the data increases, the breakdown point decreases.

The MVE and MCD have a breakdown point of  $(n - h + 1)/n$ . Thus, as  $h$  approaches  $n$ , the breakdown point approaches  $1/n$ , which is the breakdown point of

the sample mean. As  $h$  approaches  $n/2$ , the breakdown point approaches  $1/2$ , which is the breakdown point of the median, for a large sample size  $n$ .

Detecting and managing outliers for the  $p = 1$  and  $p = 2$  cases is simplified by the ability to visually inspect the data. While the Mahalanobis distance provides a means to identify outliers in higher dimensions, it can be computationally expensive as it requires computing the inverse of the covariance matrix, potentially multiple times. As already noted, the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) methods provide robust estimates of location and scatter. However, because the problems are combinatorial, they require extensive computing time, particularly as the dimension of the problem increases. The FAST-MCD provides an improvement in computing time, even for a large sample size  $n$ , but the algorithm's efficiency is still highly dependent on  $p$ , and it will only find solutions that cover at least of 50% of the data.

Our new MPDC- $\alpha$  divergence estimator essentially dispels the notion of a breakdown point, allowing us to locate solutions comprising less than 50% of the data.

### 1.1.3 Regression

We will also seek to apply our method in the context of regression. Robust regression allows us to account for contamination in the data as well as to capture a mixture of regression models. The upside to using our method for this problem is that it is always one-dimensional since we apply our algorithm to the estimated residuals,  $\epsilon_i$ . To illustrate the role of robustness in regression, we will explore several simulated

examples, leading up to situations that the MPDC- $\alpha$  divergence estimator is capable of handling.

**Example 1.1:** We simulate a sample of size  $n = 100$ . Let  $x_1, x_2, \dots, x_{99} \sim iid U(-5, 5)$  and  $y_i = 2x_i - 5 + e_i$  for  $1 \leq i \leq 99$ , where  $e_1, e_2, \dots, e_{99} \sim iid N(0, 1)$ . Then, let  $x_{100} \sim U(-5, -2)$  and  $y_{100} \sim U(10, 20)$ .

The simulated data for Example 1.1 can be seen in Figure 1.4. We see that the least-squares regression line (LS, in green) is slightly affected by the single outlier (blue point), while the  $L_2E$  regression line (purple) matches up with the least-squares line that is computed with the outliers having been removed (Uncontaminated LS, in black). When we introduce additional outliers, least-squares breaks down even further.

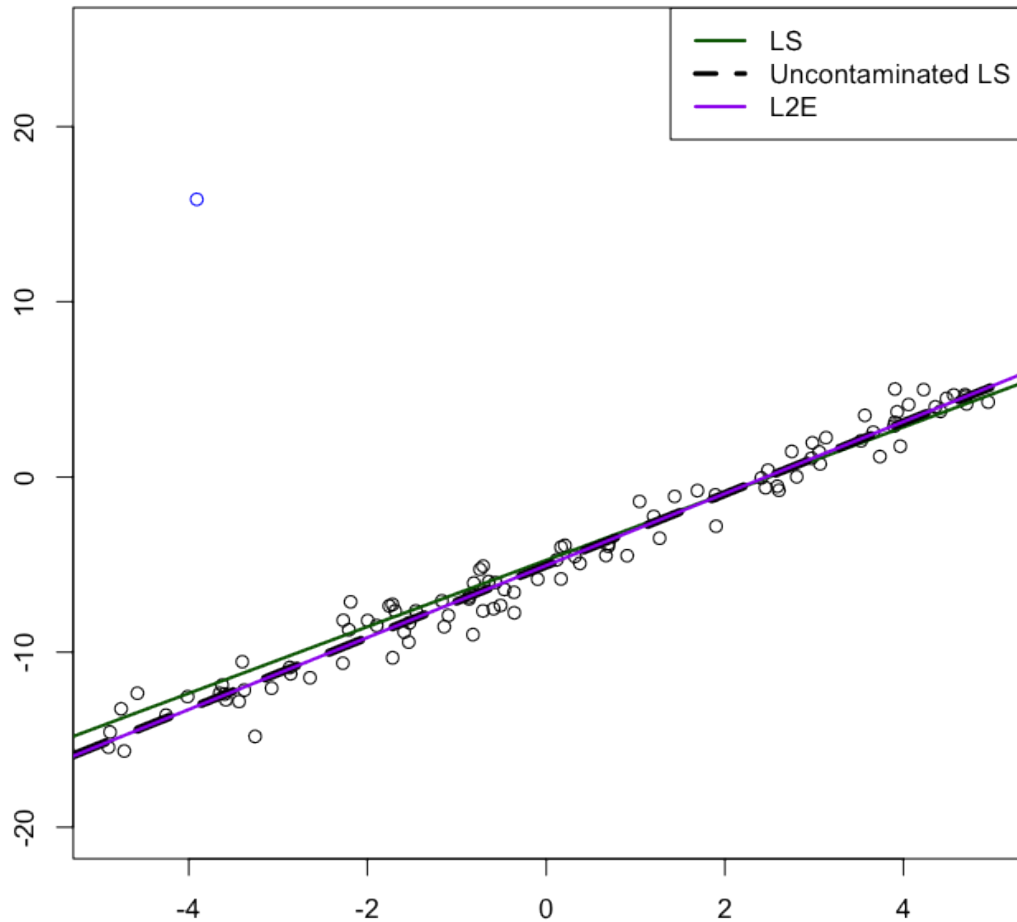


Figure 1.4: Example 1.1 - Single Outlier (1% contamination, in blue). Least-squares (LS) estimate is shown (in green) along with  $L_2E$  estimate (in purple) and least-squares estimate for only the uncontaminated data (in black).

**Example 1.2:** We simulate a sample of size  $n = 100$ . Let  $x_1, x_2, \dots, x_{80} \sim iid U(-5, 5)$  and  $y_i = 2x_i - 5 + e_i$  for  $1 \leq i \leq 80$ , where  $e_1, e_2, \dots, e_{80} \sim iid N(0, 1)$ .



Then, let  $x_{81}, x_{82}, \dots, x_{100} \sim iid U(-5, -2)$  and  $y_{81}, y_{82}, \dots, y_{100} \sim iid U(10, 20)$ .

The simulated data for Example 1.2 can be seen in Figure 1.5. The LS line is affected considerably by the outlying cluster, while the  $L_2E$  remains robust. However, we will see that  $L_2E$  is a nonconvex criterion, so multiple solutions are possible, and our resulting solution will depend on the initial values we use for our regression coefficients in the optimization routine. In this case, we initialize the  $L_2$  optimization routine with the least-squares coefficients as starting values. The next example will illustrate how  $L_2E$  can yield multiple solutions, some more robust than others.

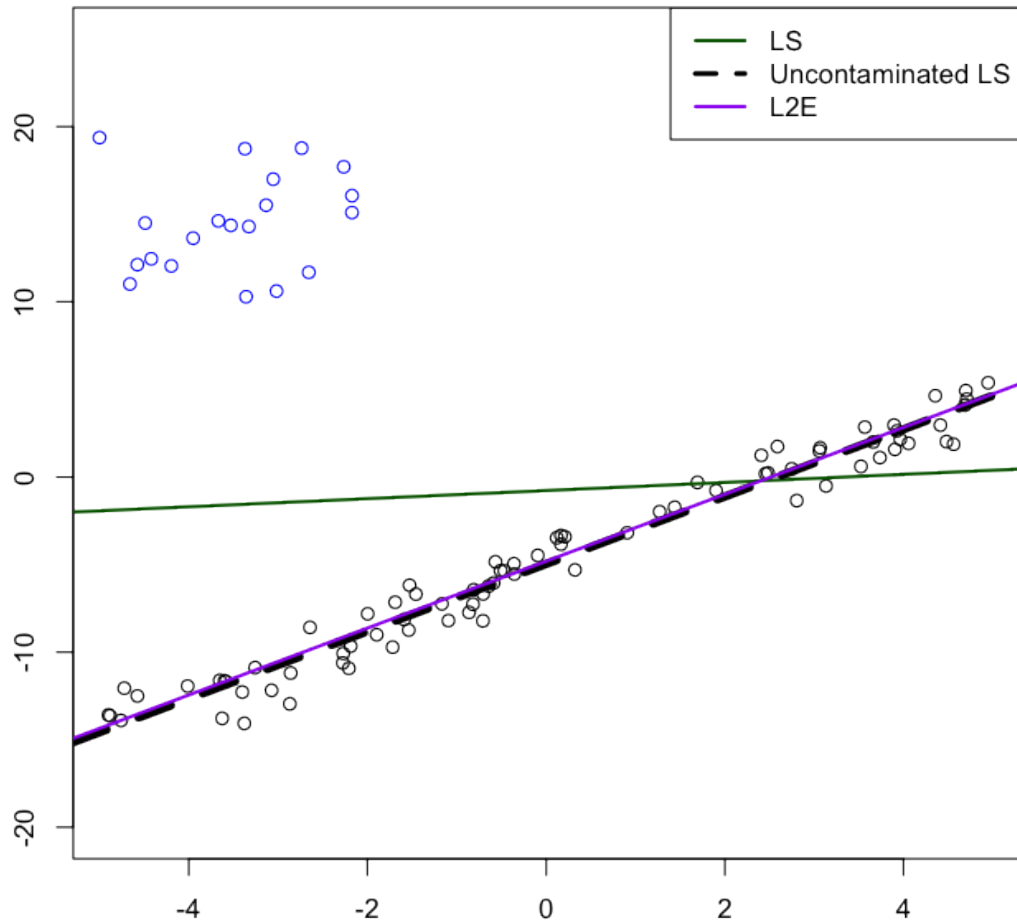


Figure 1.5: Example 1.2 - Outlying Cluster (20% contamination, in blue). Least-squares (LS) estimate is shown (in green) along with  $L_2E$  estimate (in purple) and least-squares estimate for only the uncontaminated data (in black).

**Example 1.3:** We simulate a sample of size  $n = 100$ . Let  $x_1, x_2, \dots, x_{80} \sim iid U(-5, 5)$  and  $y_i = 2x_i - 5 + e_i$  for  $1 \leq i \leq 80$ , where  $e_1, e_2, \dots, e_{80} \sim iid N(0, 1)$ .

Then, let  $x_{81}, x_{82}, \dots, x_{100} \sim iid U(-5, -2)$  and  $y_{81}, y_{82}, \dots, y_{100} \sim iid U(20, 30)$ . Note that this is almost the same setup as in Example 2, except that the  $y$  components of the 20 contaminated points have been drawn from a  $U(20, 30)$  distribution.

The simulated data for Example 1.3 can be seen in Figure 1.6. The LS line is still affected greatly by the outliers, and now the  $L_2E$  line is no longer robust to the outliers for the particular choice of starting values. Our approach, MPDC- $\alpha$  divergence regression (shown in red), is able to remain robust to the contamination and closely approximate the uncontaminated least-squares line for  $\alpha = 1.5$ .

We will develop the framework for this method and explore its practical applications.

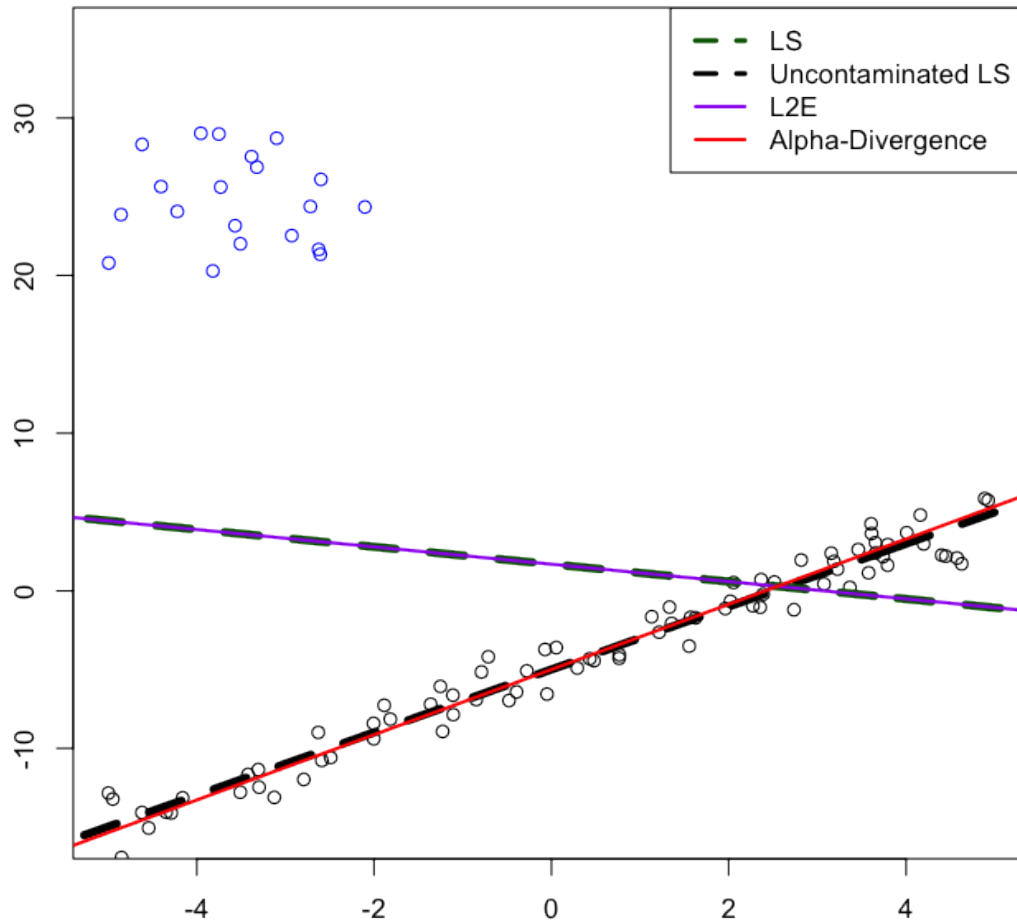


Figure 1.6: Example 1.3 - Farther Outlying Cluster (20% contamination, in blue). Least-squares (LS) estimate is shown (in green) along with  $L_2E$  estimate (in purple) and least-squares estimate for only the uncontaminated data (in black). Estimate using the MPDC- $\alpha$  criterion for  $\alpha = 1.5$  shown in red.

## 1.2 $\alpha$ -Divergence Function

The foundation for the MPDC- $\alpha$  divergence method is Basu's  $\alpha$ -divergence function. This method provides a parametric density-based divergence estimation procedure, and the parameter  $\alpha$  allows us to control the level of robustness. We first define the general setting for the  $\alpha$ -divergence function, and we will then derive its criterion for the particular case of the multivariate partial density component (MPDC) model.

Let  $\{\mathcal{F}_t\}$  be a parametric family of models indexed by unknown parameter  $t \in \Omega \subset R^S$ .  $\{\mathcal{F}_t\}$  contains a set of densities  $\{f_t\}$  with respect to the Lebesgue measure, and  $\mathcal{G}$  is the class of all distributions  $G$  having densities  $g$  with respect to the Lebesgue measure. The density power divergence (Basu, 1998) between  $g$  and  $f$  is defined as:

$$d_\alpha(g, f) = \int \left\{ f(z)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) g(z)f(z)^\alpha + \frac{1}{\alpha} g(z)^{1+\alpha} \right\} dz \quad (\alpha > 0). \quad (1.1)$$

The integrand in (1.1) is undefined when  $\alpha = 0$ , so the divergence  $d_0(g, f)$  is defined as:

$$d_0(g, f) = \lim_{\alpha \rightarrow 0} d_\alpha(g, f) = \int g(z) \log \{g(z)/f(z)\} dz. \quad (1.2)$$

Now  $d_0(g, f)$  is the Kullback-Leibler divergence function; in other words, the minimizer of the data-based version of (1.2) is the MLE. Also note:

$$\begin{aligned} d_1(g, f) &= \int \{f(z)^2 - 2g(z)f(z) + g(z)^2\} dz \\ &= \int \{f(z) - g(z)\}^2 dz. \end{aligned} \quad (1.3)$$

Minimizing an estimate of (1.3), which is the integrated squared error ( $L_2$  distance), yields the  $L_2E$  solution.

Our estimation procedure will be to select parameter values that minimize an estimate of the divergence  $d_\alpha(g, f)$ . In order to do this, an appropriate parametric model,  $f$ , must first be chosen.

### 1.3 Multivariate Partial Density Component (MPDC)

As nonparametric methods for outlier detection are notoriously error-prone (Scott, 2004), we seek a reasonable parametric approach. The most commonly assumed model is the multivariate normal, as data are often a transformation away from approximate normality. However, while it may be reasonable to impose an assumption of normality on the uncontaminated part of the data, there is less valid support for assuming the outliers or cluster(s) of outliers are normally distributed (Scott, 2004). Thus, we employ a procedure that only estimates the primary parameters of interest: an incomplete mixture model known as the multivariate partial density component (MPDC). The simplest such model is given by:

$$f(\mathbf{x}|\boldsymbol{\theta}) = w\phi(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (1.4)$$

where  $\boldsymbol{\theta} = (w, \boldsymbol{\mu}, \boldsymbol{\Sigma})$  and  $\phi$  is the Normal pdf. It is important to note that the model  $f(\mathbf{x}|\boldsymbol{\theta})$  need not be a density function, and such is the case for (1.4). For justification of this claim, see Scott (2004). We will apply the  $\alpha$ -divergence function to this MPDC model in order to obtain a robust estimate of  $\boldsymbol{\theta}$ , the MPDC- $\alpha$  divergence estimator  $\hat{\boldsymbol{\theta}}_\alpha$ .

# Chapter 2

## MPDC- $\alpha$ Divergence Estimator

This chapter explores the application of the  $\alpha$ -divergence function to a multivariate partial density component model. In particular, the  $\alpha$ -divergence function will be derived for the particular case of  $f(\mathbf{x}|\boldsymbol{\theta}) = w\phi(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . By minimizing the criterion which is an estimate of this divergence function, we obtain the MPDC- $\alpha$  divergence estimator. We will also investigate three key mathematical properties of  $\hat{\boldsymbol{\theta}}_\alpha$ , which will allow us to understand the range of cases to which we can suitably apply this robust estimator.

### 2.1 Derivation of MPDC- $\alpha$ Divergence Estimator

We begin by deriving the MPDC- $\alpha$  divergence estimator via plugging the density of our MPDC model into Basu's  $\alpha$ -divergence function. We seek to estimate the true parameter value,  $\boldsymbol{\theta} = (w, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ , from the parametric model

$$f(\mathbf{x}|\boldsymbol{\theta}) = w\phi(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

$\hat{\boldsymbol{\theta}}_\alpha$  will be the value of  $\boldsymbol{\theta}$  that minimizes our estimate of the  $\alpha$ -divergence function,  $\widehat{d}_\alpha(g, f)$ :

$$\begin{aligned} \hat{\boldsymbol{\theta}}_\alpha &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \widehat{d}_\alpha(g, f) \\ &= \begin{cases} \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \operatorname{Est} \left[ \int \left\{ f(z)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) g(z) f(z)^\alpha + \frac{1}{\alpha} g(z)^{1+\alpha} \right\} dz \right] & \alpha > 0 \\ \hat{\boldsymbol{\theta}}_{MLE} & \alpha = 0 \end{cases} \\ &= \begin{cases} \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[ \int f^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i)^\alpha \right] & \alpha > 0 \\ \hat{\boldsymbol{\theta}}_{MLE} & \alpha = 0 \end{cases} . \end{aligned}$$

Since the MLE solution when  $\alpha = 0$  is well-known, we restrict our attention to the  $\alpha > 0$  case:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_\alpha &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[ \int f^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i)^\alpha \right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[ \int [w(\phi(\mathbf{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}))]^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n [(\phi(\mathbf{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}))]^\alpha \right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[ w^{1+\alpha} \int \left[ \frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \right]^{1+\alpha} d\mathbf{x} \right. \\ &\quad \left. - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n [(\phi(\mathbf{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}))]^\alpha \right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[ \frac{w^{1+\alpha}}{(2\pi)^{p\alpha/2} (1+\alpha)^{p/2} |\boldsymbol{\Sigma}|^{\alpha/2}} \int \frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \left[\frac{\boldsymbol{\Sigma}}{1+\alpha}\right]^{-1}(\mathbf{x}-\boldsymbol{\mu})}}{(2\pi)^{p/2} \left|\frac{\boldsymbol{\Sigma}}{1+\alpha}\right|^{1/2}} d\mathbf{x} \right. \\ &\quad \left. - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n [(\phi(\mathbf{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}))]^\alpha \right] \\ &= \underset{w, \boldsymbol{\mu}, \mathbf{U}}{\operatorname{argmin}} \left[ \frac{w^{1+\alpha}}{[(2\pi)^\alpha (1+\alpha)]^{p/2} |\mathbf{U}|^\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n [(\phi(\mathbf{x}_i | \boldsymbol{\mu}, \mathbf{U}))]^\alpha \right] \end{aligned}$$



where  $\mathbf{U}$  is the Cholesky decomposition of  $\Sigma^{-1}$ , i.e.  $\Sigma^{-1} = \mathbf{U}'\mathbf{U}$ . Therefore,  $\det(\Sigma)^{-1/2} = \det(\mathbf{U})$ . We will explore the purpose and consequences of this Cholesky transformation in Chapter 3.

Thus,

$$\hat{\theta}_\alpha = \begin{cases} \underset{w, \boldsymbol{\mu}, \mathbf{U}}{\operatorname{argmin}} \left[ \frac{w^{1+\alpha}}{[(2\pi)^\alpha(1+\alpha)]^{p/2}} |\mathbf{U}|^\alpha - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n [(\phi(\mathbf{x}_i | \boldsymbol{\mu}, \mathbf{U}))^\alpha] \right] & \alpha > 0 \\ \hat{\theta}_{MLE} & \alpha = 0 \end{cases} .$$

## 2.2 Parallel to M-Estimators

The properties of the MPDC- $\alpha$  divergence estimator are clear upon realizing that all minimum divergence estimators are a particular case of M-estimators, i.e. they solve

$$\sum_i \psi(X_i, t) = 0$$

for some function  $\psi$ . For the MPDC- $\alpha$  divergence estimator

$$\psi(x, t) = u_t(x) f_t^\alpha(x) - \int u_t(z) f_t^{1+\alpha}(z) dz,$$

where  $u_t(z) = \partial \log f_t(z) / \partial t$  is the maximum likelihood score function.

Immediately following from this parallel are the consistency and asymptotic normality of the MPDC- $\alpha$  divergence estimator.

## 2.3 Properties of MPDC- $\alpha$ Divergence Estimator

Basu (1998) verifies the consistency and asymptotic normality of the  $\alpha$ -divergence estimator. We will show that the consistency and asymptotic normality of  $\hat{\theta}_\alpha$  follow directly from those results. We will also establish that  $\hat{\theta}_\alpha$  is invariant under linear transformations.

### 2.3.1 Consistency and Asymptotic Normality of $\hat{w}_\alpha$

**Lemma 2.1.** *Let  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  be a random sample from the mixture*

$$wN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + (1 - w)F^*,$$

where contamination distribution  $F^*$  is far away from the primary component. Let  $\hat{w}_\alpha$  be the weight component of the MPDC- $\alpha$  divergence estimator,  $\hat{\theta}_\alpha \equiv (\hat{w}_\alpha, \hat{\boldsymbol{\mu}}_\alpha, \hat{\boldsymbol{\Sigma}}_\alpha)$ .

Then as  $n \rightarrow \infty$ :

$$\sqrt{n}(\hat{w}_\alpha - w) \rightarrow^d N\left(0, \frac{1}{w} \left[ \frac{(1 + \alpha)^p}{(1 + 2\alpha)^{p/2}} - 1 \right]\right). \quad (2.1)$$

**Proof.**

$$\text{We define the model: } f_\theta = w\phi(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{we^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}.$$

Then we obtain an estimate of the divergence function by plugging the model into

(1.1):

$$\widehat{d}_\alpha = w^{1+\alpha} \int \frac{e^{-\frac{1+\alpha}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}}{(2\pi)^{p(1+\alpha)/2}|\boldsymbol{\Sigma}|^{\frac{1+\alpha}{2}}} - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n \left[ \frac{e^{-\frac{\alpha}{2}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})}}{(2\pi)^{p\alpha/2}|\boldsymbol{\Sigma}|^{\frac{\alpha}{2}}} \right].$$

$$\text{Letting } \gamma_1 \equiv \int \frac{e^{-\frac{1+\alpha}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}}{(2\pi)^{p(1+\alpha)/2}|\boldsymbol{\Sigma}|^{\frac{1+\alpha}{2}}} = (2\pi)^{p\alpha/2}(1 + \alpha)^{p/2}|\boldsymbol{\Sigma}|^{\alpha/2}$$

$$\text{and } \gamma_2 \equiv \frac{e^{-\frac{\alpha}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})}}{(2\pi)^{p\alpha/2} |\boldsymbol{\Sigma}|^{\frac{\alpha}{2}}},$$

$$\text{we get } E[\gamma_2] = \int w \frac{e^{-\frac{1+\alpha}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})}}{(2\pi)^{p(1+\alpha)/2} |\boldsymbol{\Sigma}|^{\frac{1+\alpha}{2}}} = w\gamma_1 \equiv \gamma_3.$$

$$\begin{aligned} \text{Thus, } \widehat{d}_\alpha &= w^{1+\alpha}\gamma_1 - \left(1 + \frac{1}{\alpha}\right) \frac{w^{1+\alpha}}{wn} \sum_{i=1}^n [(\gamma_2 - \gamma_3) + \gamma_3] \\ &= w^{1+\alpha}\gamma_1 - \left(1 + \frac{1}{\alpha}\right) w^\alpha \gamma_3 - \frac{(1 + \frac{1}{\alpha})w^{1+\alpha}}{\sqrt{wn}} Z \sqrt{\Sigma(\theta)}, \end{aligned}$$

$$\text{where } \Sigma(\theta) \equiv \text{Var}(\gamma_2).$$

To find  $\hat{w}$  we differentiate with respect to  $w$ :

$$\begin{aligned} \frac{\partial \widehat{d}_\alpha}{\partial w} &= (1 + \alpha)w^\alpha \gamma_1 - (1 + \alpha)w^{\alpha-1} \gamma_3 - \frac{1 + \alpha}{\sqrt{wn}} w^{\alpha-1} Z \sqrt{\Sigma(\theta)} = 0 \\ \iff \hat{w}\gamma_1 &= \gamma_3 + \frac{1}{\sqrt{wn}} \sqrt{\Sigma(\theta)} Z. \end{aligned}$$

$$\begin{aligned} \text{Therefore, } \sqrt{n} \left( \hat{w} - \frac{\gamma_3}{\gamma_1} \right) &\rightarrow^d N \left( 0, \frac{\Sigma(\theta)}{w\gamma_1^2} \right) \quad \text{as } n \rightarrow \infty \\ &\rightarrow^d N \left( 0, w^{-1} [(2\pi)^{p\alpha} (1 + \alpha)^p |\boldsymbol{\Sigma}|^\alpha \Sigma(\theta)] \right) \\ &\rightarrow^d N \left( 0, \frac{1}{w} \left[ \frac{(1 + \alpha)^p}{(1 + 2\alpha)^{p/2}} - 1 \right] \right). \end{aligned}$$

Thus,

$$\sqrt{n}(\hat{w}_\alpha - w) \rightarrow^d N \left( 0, \frac{1}{w} \left[ \frac{(1 + \alpha)^p}{(1 + 2\alpha)^{p/2}} - 1 \right] \right).$$

□

Lemma 2.1 says that  $\hat{w}_\alpha$  is asymptotically unbiased for  $w$ . That result is only guaranteed to hold when there is considerable separation between the targeted density component and the remaining data. When there is overlap between the target and contamination,  $\hat{w}_\alpha$  will be biased upward.

### 2.3.2 Consistency and Asymptotic Normality of $\hat{\mu}_\alpha$ and $\hat{\Sigma}_\alpha$

**Lemma 2.2.** *Let  $\mathbf{Q}_1, \mathbf{Q}_2, \dots, \mathbf{Q}_n$  be an iid sequence of random vectors from a  $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  distribution. Let  $\hat{\boldsymbol{\tau}}_\alpha(\mathbf{Q}) \equiv (\hat{\boldsymbol{\mu}}_\alpha, \hat{\boldsymbol{\Sigma}}_\alpha)$  be Basu's  $\alpha$ -divergence estimator for  $\boldsymbol{\tau} \equiv (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Note that  $\hat{\boldsymbol{\tau}}_\alpha$  and  $\boldsymbol{\tau}$  are each comprised of a  $p$ -dimensional mean vector and the vectorization of the  $\frac{p(p+1)}{2}$  elements above and including the diagonal of the covariance matrix. Then, under certain regularity conditions, there exists  $\hat{\boldsymbol{\tau}}_\alpha$  such that, as  $n \rightarrow \infty$ ,  $\sqrt{n}(\hat{\boldsymbol{\tau}}_\alpha - \boldsymbol{\tau})$  is asymptotically multivariate normal with vector mean zero and covariance matrix  $J^{-1}KJ^{-1}$ , where  $J = J(\boldsymbol{\tau})$  and  $K = K(\boldsymbol{\tau})$  are given by*

$$J = \int u_\tau(z)u_\tau^T(z)f_\tau^{1+\alpha}(z)dz + \int [i_\tau(z) - \alpha u_\tau(z)u_\tau^T(z)] [g(z) - f_\tau(z)] f_\tau^\alpha(z)dz, \quad (2.2)$$

$$K = \int u_\tau(z)u_\tau^T(z)f_\tau^{2\alpha}(z)g(z)dz - \xi\xi^T \quad (2.3)$$

with  $\xi = \int u_\tau(z)f_\tau^\alpha(z)g(z)dz$ .

**Proof.** See Basu (1998), p. 553. □

### 2.3.3 Consistency and Asymptotic Normality of $\hat{\theta}_\alpha$

**Lemma 2.3.** *Let  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  be a random sample from the mixture*

$$wN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + (1 - w)F^*,$$

where  $F^*$  is the contamination distribution, and let  $\hat{\boldsymbol{\tau}}_\alpha \equiv (1, \hat{\boldsymbol{\mu}}_\alpha, \hat{\boldsymbol{\Sigma}}_\alpha)$  be Basu's  $\alpha$ -divergence estimator for  $\boldsymbol{\theta} \equiv (w, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Let  $\hat{\boldsymbol{\theta}}_\alpha \equiv (\hat{w}_\alpha, \hat{\boldsymbol{\mu}}_\alpha, \hat{\boldsymbol{\Sigma}}_\alpha)$  be the MPDC- $\alpha$  divergence estimator for  $\boldsymbol{\theta} \equiv (w, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Then, as  $n \rightarrow \infty$ ,  $\sqrt{n}(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta})$  is asymptot-

ically multivariate normal with vector mean zero and covariance matrix  $\frac{1}{w}J^{-1}KJ^{-1}$ , where  $J = J(\boldsymbol{\tau})$  and  $K = K(\boldsymbol{\tau})$  are as defined in Lemma 2.2.

**Proof.** By Lemma 2.1, as  $n \rightarrow \infty$ :

$$\sqrt{n}(\hat{w}_\alpha - w) \rightarrow^d N\left(0, \frac{1}{w} \left[ \frac{(1+\alpha)^p}{(1+2\alpha)^{p/2}} - 1 \right]\right).$$

Thus,  $\hat{w}_\alpha$  is consistent for  $w$ , asymptotically normal, and asymptotically independent of  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ . By Lemma 2.2,  $\hat{\boldsymbol{\mu}}_\alpha$  and the elements of  $\hat{\boldsymbol{\Sigma}}_\alpha$  are asymptotically normal and consistent as the effective sample size  $wn \rightarrow \infty$  for  $\boldsymbol{\mu}$  and the elements of  $\boldsymbol{\Sigma}$ , respectively. Therefore, as  $wn \rightarrow \infty$ ,  $\sqrt{wn}(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta})$  is asymptotically multivariate normal with vector mean zero and covariance matrix  $J^{-1}KJ^{-1} \iff$  as  $n \rightarrow \infty$ ,  $\sqrt{n}(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta})$  is asymptotically multivariate normal with vector mean zero and covariance matrix  $\frac{1}{w}J^{-1}KJ^{-1}$ , where  $J = J(\boldsymbol{\tau})$  and  $K = K(\boldsymbol{\tau})$  are as defined in Lemma 2.2. □

### 2.3.4 Invariance Under Linear Transformation

**Lemma 2.4.** *The MPDC- $\alpha$  divergence estimator,  $\hat{\boldsymbol{\theta}}_\alpha$ , is invariant under linear transformation.*

**Proof.** Let  $\mathbf{X}$  be a random vector from the mixture

$$wN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + (1-w)F^*,$$

where  $F^*$  is the contamination distribution, and let  $\mathbf{Y}$  be the random vector obtained by multiplying  $\mathbf{X}$  by the eigenvectors of  $\boldsymbol{\Sigma}$ , which are the columns of matrix  $\boldsymbol{\Gamma}$ . Thus,

$\Sigma = \Gamma\Lambda\Gamma'$ , where  $\Lambda = \text{diag}(\text{eigenvalues}(\Sigma))$ . Since  $\mathbf{Y} = \Gamma\mathbf{X}$ , we can directly define the distribution of  $\mathbf{Y}$ :

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{Y}} &= \Gamma\boldsymbol{\mu}_{\mathbf{X}} = \Gamma\boldsymbol{\mu} \\ \text{cov}(\mathbf{Y}) &= \text{cov}(\Gamma\mathbf{X}) \\ &= \Gamma' \text{cov}(\mathbf{X})\Gamma \\ &= \Gamma'\Sigma\Gamma \\ &= \Gamma'(\Gamma\Lambda\Gamma')\Gamma \\ &= (\Gamma'\Gamma)\Lambda(\Gamma'\Gamma) \\ &= \Lambda \text{ since } \Gamma\Gamma' = I_p.\end{aligned}$$

Thus, since normality is preserved under linear transformations:

$$\mathbf{Y} \sim wN(\Gamma\boldsymbol{\mu}, \Lambda) + (1 - w)F^{**},$$

where  $F^{**}$  is the transformed contamination distribution. We wish to show that the MPDC- $\alpha$  divergence criterion for  $\mathbf{Y}$  is equivalent to that for  $\mathbf{X}$ . By the MLE invariance principle, we know that the result holds for the  $\alpha = 0$  case. Thus, we restrict our attention to the case when  $\alpha > 0$ . The MPDC- $\alpha$  divergence criterion for  $\mathbf{X}$  is given by:

$$\frac{w^{1+\alpha}}{[(2\pi)^\alpha(1+\alpha)]^{p/2}} |\mathbf{U}|^\alpha - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n [(\phi(\mathbf{x}_i|\boldsymbol{\mu}, \Sigma))]^\alpha$$

where  $\Sigma^{-1} = \mathbf{U}'\mathbf{U}$ .

The MPDC- $\alpha$  divergence criterion for  $\mathbf{Y}$  is given by:

$$\begin{aligned}
& \frac{w^{1+\alpha}}{[(2\pi)^\alpha(1+\alpha)]^{p/2}} |\mathbf{\Lambda}|^{-\alpha/2} - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n [(\phi(\mathbf{y}_i|\mathbf{\Gamma}\boldsymbol{\mu}, \mathbf{\Lambda}))^\alpha] \\
&= \frac{w^{1+\alpha}}{[(2\pi)^\alpha(1+\alpha)]^{p/2}} (|\mathbf{\Gamma}'\mathbf{\Gamma}||\mathbf{\Lambda}|)^{-\alpha/2} - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n [(\phi(\mathbf{y}_i|\mathbf{\Gamma}\boldsymbol{\mu}, \mathbf{\Lambda}))^\alpha] \\
&= \frac{w^{1+\alpha}}{[(2\pi)^\alpha(1+\alpha)]^{p/2}} |\mathbf{\Gamma}'\mathbf{\Gamma}\mathbf{\Lambda}|^{-\alpha/2} - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n [(\phi(\mathbf{y}_i|\mathbf{\Gamma}\boldsymbol{\mu}, \mathbf{\Lambda}))^\alpha] \\
&= \frac{w^{1+\alpha}}{[(2\pi)^\alpha(1+\alpha)]^{p/2}} |\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}'|^{-\alpha/2} - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n [(\phi(\mathbf{y}_i|\mathbf{\Gamma}\boldsymbol{\mu}, \mathbf{\Lambda}))^\alpha] \\
&= \frac{w^{1+\alpha}}{[(2\pi)^\alpha(1+\alpha)]^{p/2}} |\boldsymbol{\Sigma}|^{-\alpha/2} - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n [(\phi(\mathbf{y}_i|\mathbf{\Gamma}\boldsymbol{\mu}, \mathbf{\Lambda}))^\alpha] \\
&= \frac{w^{1+\alpha}}{[(2\pi)^\alpha(1+\alpha)]^{p/2}} |\mathbf{U}|^\alpha - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n [(\phi(\mathbf{y}_i|\mathbf{\Gamma}\boldsymbol{\mu}, \mathbf{\Lambda}))^\alpha].
\end{aligned}$$

$$\begin{aligned}
\phi(\mathbf{y}_i|\mathbf{\Gamma}\boldsymbol{\mu}, \mathbf{\Lambda}) &= \frac{e^{-\frac{\alpha}{2}(\mathbf{y}_i-\mathbf{\Gamma}\boldsymbol{\mu})'\mathbf{\Lambda}^{-1}(\mathbf{y}_i-\mathbf{\Gamma}\boldsymbol{\mu})}}{(2\pi)^{p\alpha/2}|\mathbf{\Lambda}|^{p/2}} \\
&= \frac{e^{-\frac{\alpha}{2}(\mathbf{x}_i-\boldsymbol{\mu})'\mathbf{\Gamma}'\mathbf{\Lambda}^{-1}\mathbf{\Gamma}(\mathbf{x}_i-\boldsymbol{\mu})}}{(2\pi)^{p\alpha/2}|\boldsymbol{\Sigma}|^{p/2}} \\
&= \frac{e^{-\frac{\alpha}{2}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})}}{(2\pi)^{p\alpha/2}|\boldsymbol{\Sigma}|^{p/2}} \\
&= \phi(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}).
\end{aligned}$$

Thus, the MPDC- $\alpha$  divergence criterion is invariant under linear transformation.  $\square$

Without loss of generality, for simulation purposes we will henceforth only consider  $\boldsymbol{\Sigma}$  of the form

$$\boldsymbol{\Sigma}_{jj} = 1 \quad \forall j \in [1, p],$$

since, by Lemma 2.4, any arbitrary matrix  $\boldsymbol{\Sigma}$  can be transformed to have unit variances, yielding the correlation matrix.

# Chapter 3

## Parametric Density Estimation

Our parameter vector of interest,  $\boldsymbol{\theta}$ , consists of three components for a total of  $\frac{p^2+3p+2}{2}$  parameter values: (1) the weight parameter,  $w$ ; (2) the mean vector,  $\boldsymbol{\mu}$  (which has  $p$  parameters); (3) the covariance matrix  $\boldsymbol{\Sigma}$  (for which we estimate  $\frac{p(p+1)}{2}$  parameters). Our parametric estimation study will consider simulated cases as well as an application in baseball. For the simulated cases, we will draw samples from the full two-component Normal mixture distribution and use the MPDC- $\alpha$  divergence criterion to estimate  $w$ ,  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  with  $\hat{w}_\alpha$ ,  $\hat{\boldsymbol{\mu}}_\alpha$  and  $\hat{\boldsymbol{\Sigma}}_\alpha$ , respectively.

### 3.1 Parameter and Criterion Definition

We construct the setting for our simulated cases, which will draw from a two-component Normal mixture. It should be noted that the applicability of the MPDC- $\alpha$  divergence estimator is not limited to these cases. We can utilize  $\hat{\boldsymbol{\theta}}_\alpha$  for pure samples as well as samples with outliers, and the contamination can be particular points, a single



cluster of points, or multiple clusters of points. The framework for our study will be defined as follows.

Given an *iid* sample  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  from the mixture

$$wN(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1 - w)N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

and model

$$f(\mathbf{x}|\boldsymbol{\theta}) = w\phi(\mathbf{x}|\boldsymbol{\mu} \equiv \boldsymbol{\mu}_1, \boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}_1),$$

we estimate the parameter vector  $\boldsymbol{\theta} = (w, \boldsymbol{\mu}, \boldsymbol{\Sigma})$  with  $\hat{\boldsymbol{\theta}}_\alpha$  by solving the following optimization problem for  $\alpha > 0$ :

$$\min_{\boldsymbol{\theta}} \left[ \frac{w^{1+\alpha}}{[(2\pi)^\alpha(1+\alpha)]^{p/2}|\boldsymbol{\Sigma}|^{\alpha/2}} - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n [(\phi(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}))^\alpha] \right]. \quad (3.1)$$

Before proceeding with the optimization, we apply some helpful transformations to our parameters and rewrite the criterion (3.1) in a more computationally tractable form.

## 3.2 Parameter Transformations

We wish to estimate  $\boldsymbol{\theta} \equiv (w, \boldsymbol{\mu}, \boldsymbol{\Sigma})$  via an unconstrained optimization algorithm. Since  $w$  and  $\boldsymbol{\Sigma}$  are constrained parameters, we apply transformations to them for the purpose of the optimization.

### 3.2.1 $w$ : *logit* transformation

The weight parameter,  $w$ , falls in the range of  $(0,1)$ . However, since we are solving this as an unconstrained optimization problem, we wish to optimize over some function

$\tau(w)$  such that the range of  $\tau(w)$  is  $(-\infty, \infty)$ . Thus, we define  $\tau(w)$  to be the *logit* transformation,

$$\tau(w) = \log\left(\frac{w}{1-w}\right), \quad (3.2)$$

so that when we reverse the transformation, we get:

$$w = \frac{1}{1 + e^{-\tau}}. \quad (3.3)$$

This yields the desired range of  $(0,1)$  for values of  $w$ .

### 3.2.2 $\Sigma^{-1}$ : Cholesky Decomposition and *exp* transformation

Further computational efficiency can be gained by considering the Cholesky decomposition of the precision matrix,  $\Sigma^{-1}$ . We decompose  $\Sigma^{-1}$  into the product of an upper triangular matrix,  $\mathbf{U}$  and its transpose, i.e.:

$$\Sigma^{-1} = \mathbf{U}'\mathbf{U}. \quad (3.4)$$

Thus, when we optimize over  $\mathbf{U}$ , we need only estimate  $\frac{p(p+1)}{2}$  values as opposed to  $p^2$ . However, we opt to restrict the diagonal values of  $\mathbf{U}$  to be positive. Thus, we exponentiate the diagonal values of  $\mathbf{U}$ :

$$\text{diag}(\boldsymbol{\eta}(\mathbf{U})) = \exp(\text{diag}(\mathbf{U})), \quad (3.5)$$

where  $\boldsymbol{\eta}$  is the transformed matrix over which we will perform our optimization.

### 3.2.3 Computationally Efficient Criterion

Let  $d \equiv |\mathbf{U}| = \prod_{i=1}^p U_{ii}$ . To improve computational efficiency, we can redefine the criterion (3.1) as:

$$\frac{w^{1+\alpha}}{[(2\pi)^\alpha(1+\alpha)]^{p/2}} d^\alpha - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n \left[ \frac{d^\alpha}{(2\pi)^{p\alpha/2}} e^{-\frac{\alpha}{2} \mathbf{1}'_p [\mathbf{U}(\mathbf{x}'_i - \boldsymbol{\mu})]^2} \right], \quad (3.6)$$

where  $[\ ]^2$  in the exponent is a component-wise squaring.

## 3.3 Simulated Cases

We will simulate MPDC samples with varying characteristics. Our cases span from  $p = 2$  to  $p = 10$  dimensions. Results from the simulations for  $p = 6$  through  $p = 10$  can be found in the Appendix. The uncontaminated and contaminated clusters are well-separated or overlapping, and we also consider the effect of correlation in the uncontaminated data on  $\hat{\boldsymbol{\theta}}_\alpha$ . Without loss of generality, we fix the fraction of contamination in our samples at 25%. Also, for the optimization, we initialize the parameter vector at the true values.

### 3.3.1 Simulated Cases for $p = 2$

#### Pure Sample

We first verify that the MPDC- $\alpha$  estimator yields the correct results for an uncontaminated sample. The simulated sample of size  $n = 1000$  is from the bivariate standard

Normal distribution

$$N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_2 \right).$$

Figure 3.1 shows a contour plot of the  $N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_2 \right)$  distribution overlain on the sample points.

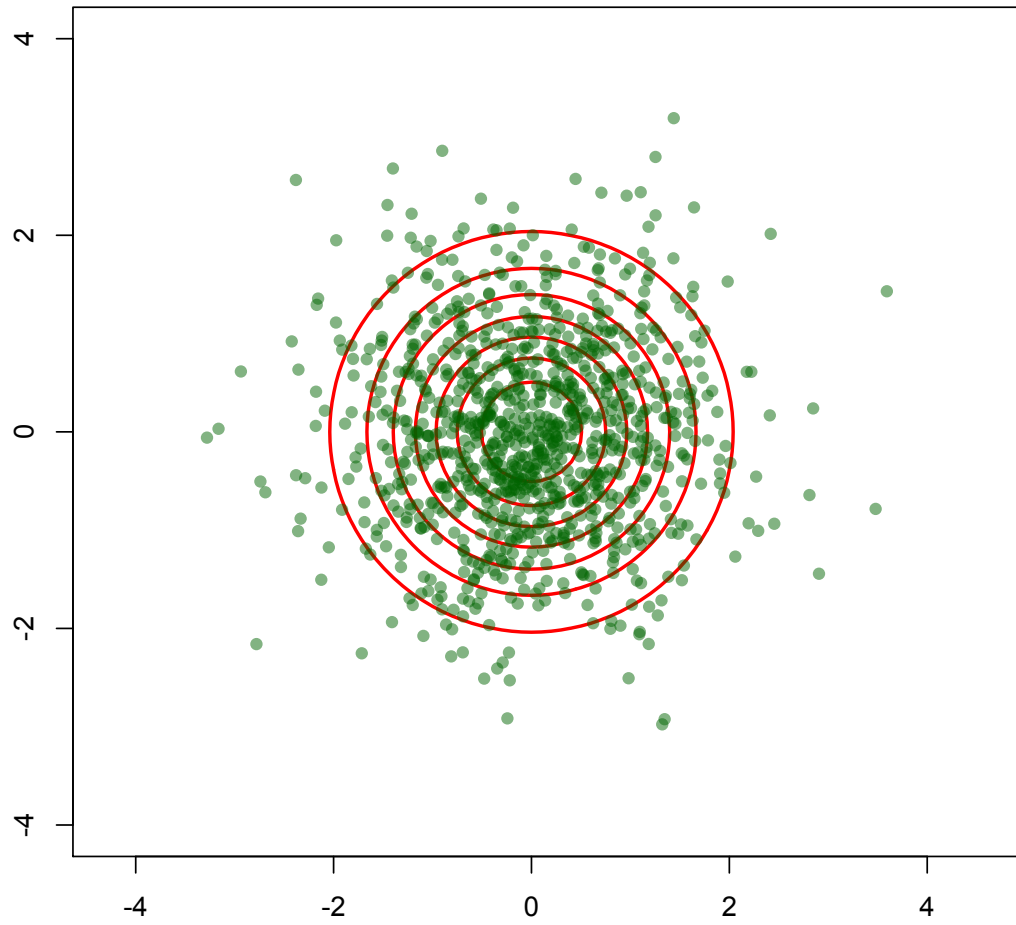


Figure 3.1: Pure sample of size  $n = 1000$  simulated from  $N((0, 0)', I_2)$  distribution with contour lines of true density (red).

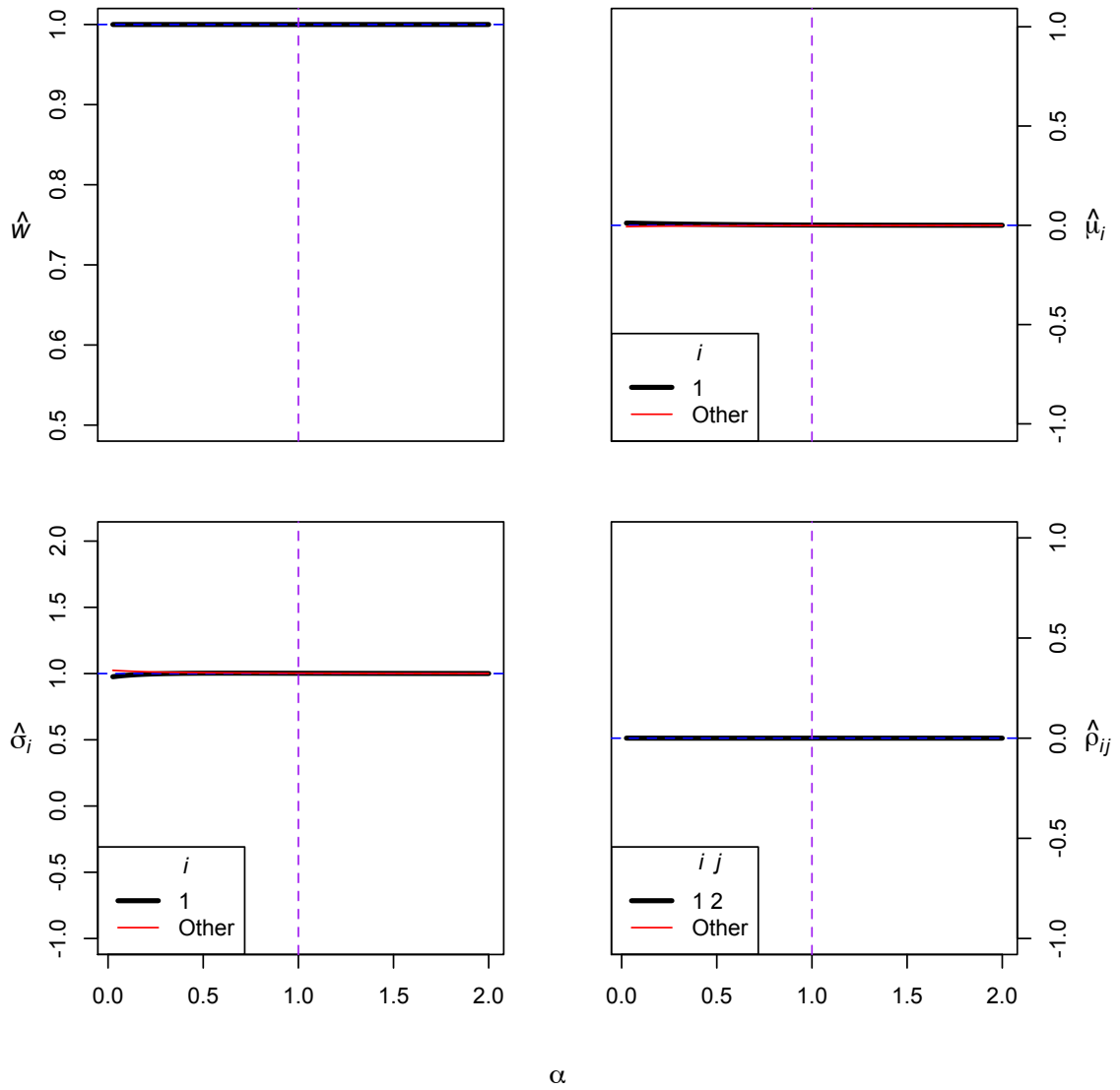


Figure 3.2: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameter estimates (in this case,  $\hat{\mu}_{2,\alpha}$  and  $\hat{\sigma}_{2,\alpha}$ ) that we track to assure algorithm stability. Estimates based on pure sample of size  $n = 1000$  simulated from  $N((0,0)', I_2)$  distribution.

Figure 3.2 shows trace plots of the parameter estimates  $\hat{w}_\alpha$ ,  $\hat{\mu}_{\alpha,i}$ ,  $\hat{\sigma}_{\alpha,i}$ , and  $\hat{\rho}_{\alpha,ij}$  for  $\alpha$  in the range  $[0, 2]$ . We can see that the MPDC- $\alpha$  estimator yields consistent values for a pure sample regardless of the value of  $\alpha$ .

### Overlapping Clusters with Zero Correlation

We begin our exploration of contaminated samples with a two-dimensional example, generating a sample of size  $n = 1000$  from the mixture distribution

$$\frac{3}{4} N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_2 \right) + \frac{1}{4} N \left( \begin{pmatrix} 3 \\ 0 \end{pmatrix}, I_2 \right).$$

Figure 3.3 shows a contour plot of the  $N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_2 \right)$  distribution overlain on the sample points (uncontaminated data in green and contamination in blue). We can see that there is considerable overlap between the main (uncontaminated) data and the contamination.

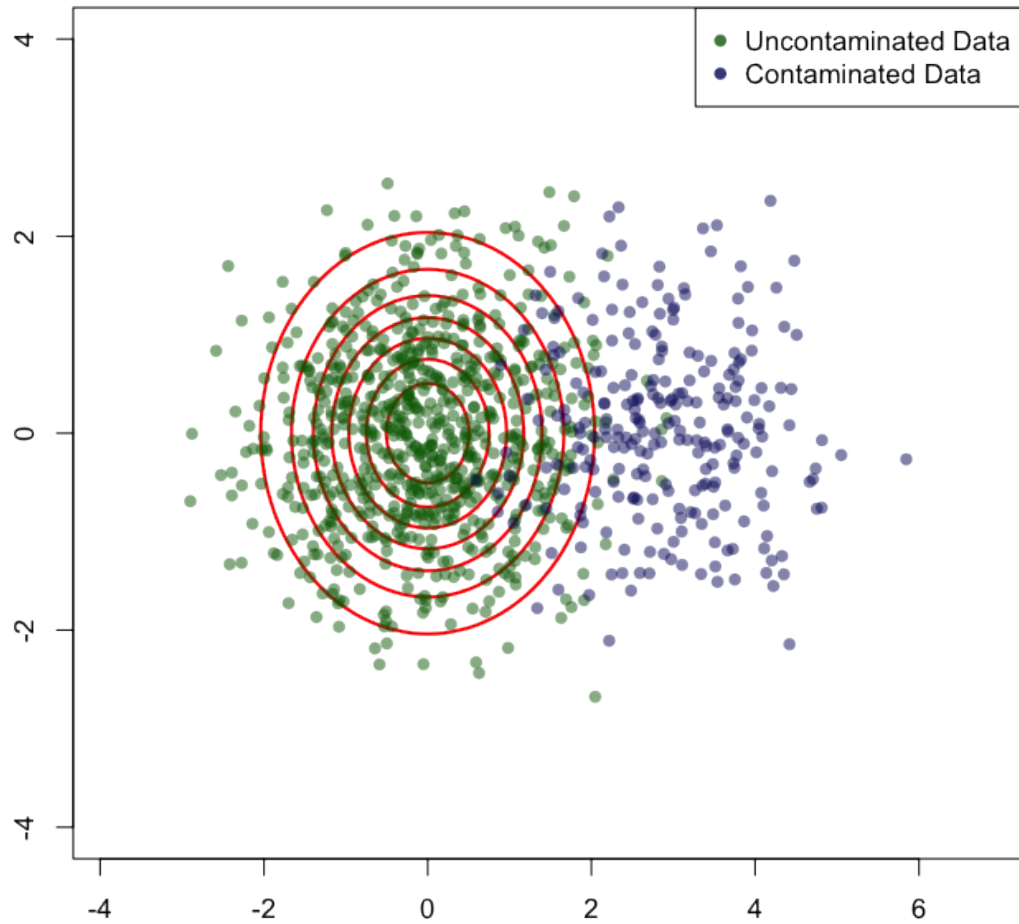


Figure 3.3: Overlapping clusters with zero correlation in the uncontaminated data. Contour lines (red) of  $N((0,0)', I_2)$  density overlain on sample of size  $n = 1000$  simulated from the Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0,0)'$ ;  $\mu_2 = (3,0)'$ ;  $\Sigma_1 = \Sigma_2 = I_2$ .



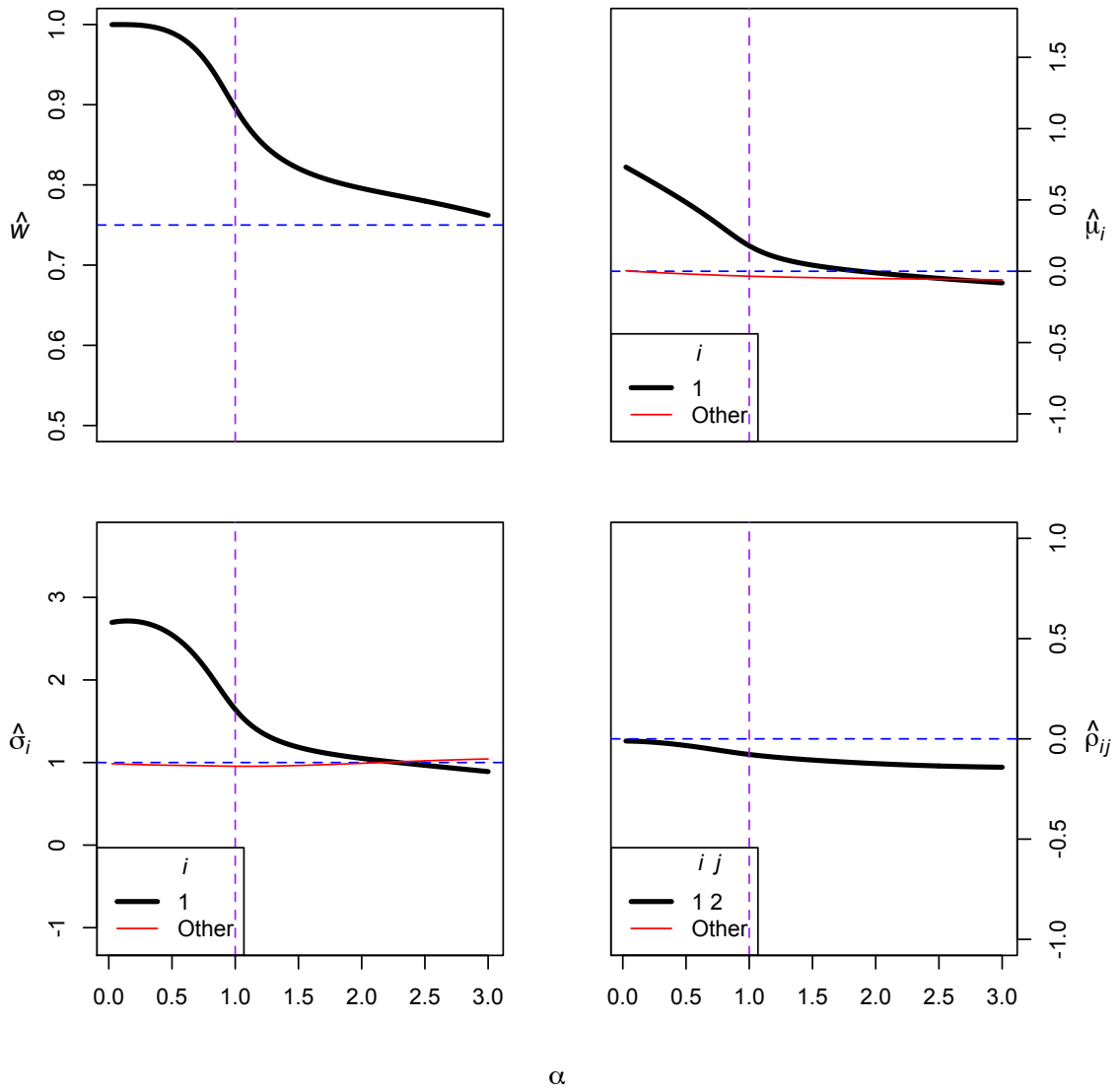


Figure 3.4: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 3 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0)'$ ;  $\mu_2 = (3, 0)'$ ;  $\Sigma_1 = \Sigma_2 = I_2$ .

Figure 3.4 shows trace plots of the parameter estimates  $\hat{w}_\alpha$ ,  $\hat{\mu}_{\alpha,i}$ ,  $\hat{\sigma}_{\alpha,i}$ , and  $\hat{\rho}_{\alpha,ij}$  for  $\alpha$  in the range  $[0, 2]$ . We see that the weight parameter estimate,  $\hat{w}_\alpha$ , starts at 1, which is the MLE estimate for  $w$ , and beyond an  $\alpha$  value of about 0.75, the estimate declines steadily to about 0.85 for  $\alpha = 2$ . While the true value of  $w$  is 0.75, because of the overlap between the main data and contaminated data in this example, we should expect our estimate for  $w$  to be somewhat inflated. Our other estimates are clearly affected by the contamination for  $\alpha$  values between 0 and 1 (denoted by the purple dashed vertical line, giving the  $L_2E$  solution as a reference point). These estimates for  $\mu$ ,  $\sigma$  and  $\rho$  begin to reach their true values (denoted by the blue dashed horizontal lines) around  $\alpha = 1.5$  and continue to improve until  $\alpha = 2$ , where they begin to level off. There is not sufficient benefit to be gained from looking at  $\alpha$  values beyond 2 in this case, particularly because of the decrease in efficiency (to be explored in Chapter 4).

### Overlapping Clusters with $\rho_{12} = 0.75$

We then investigate what happens when we introduce a non-zero correlation in the cluster centered at  $\boldsymbol{\mu}_1$ . Thus, we will consider the case where  $\rho_{12} = 0.75$ , i.e. we generate a sample of size  $n = 1000$  from the mixture model

$$\frac{3}{4} N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix} \right) + \frac{1}{4} N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_2 \right).$$

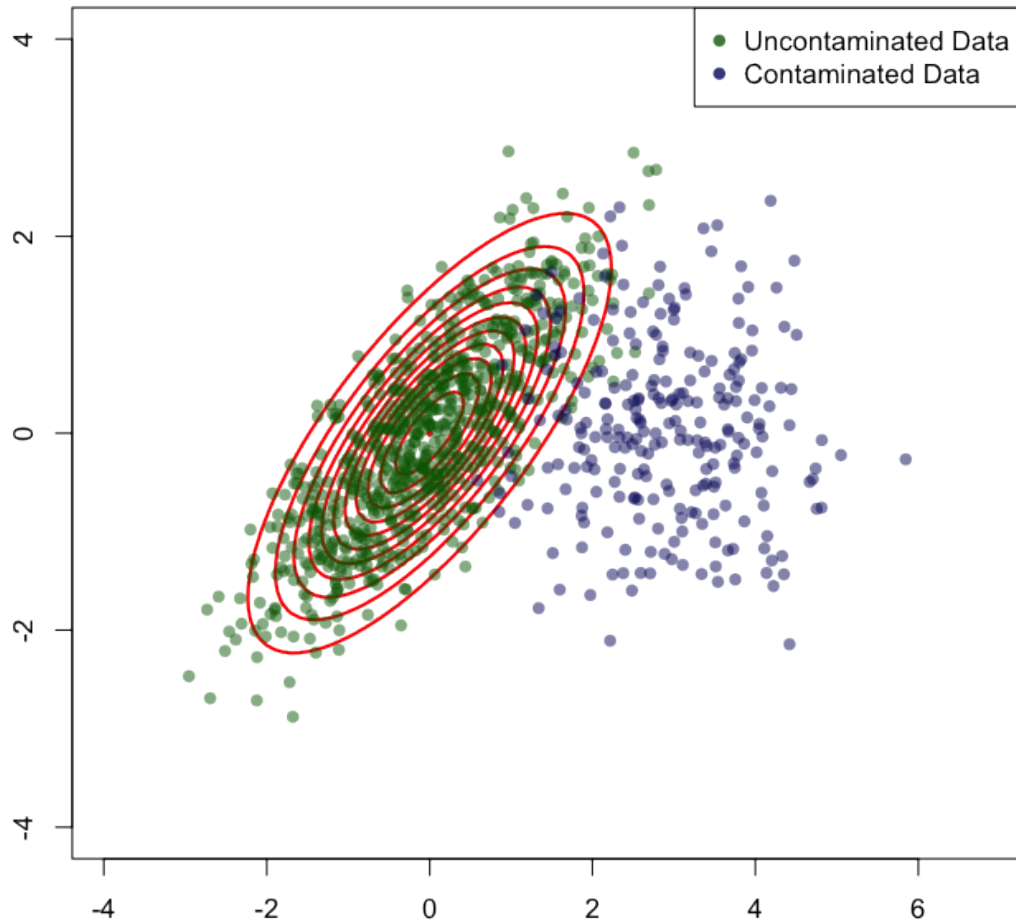


Figure 3.5: Overlapping clusters with  $\rho_{12} = 0.75$  in the uncontaminated data. Contour lines (red) of  $N((0, 0)', (\begin{smallmatrix} 1.00 & 0.75 \\ 0.75 & 1.00 \end{smallmatrix}))$  density overlain on sample of size  $n = 1000$  simulated from the Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0)'$ ;  $\mu_2 = (3, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_2$ .

Figure 3.5 shows a contour plot of the  $N((\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}), (\begin{smallmatrix} 1.00 & 0.75 \\ 0.75 & 1.00 \end{smallmatrix}))$  distribution overlain on

the sample points. Due to the correlation in the main data, we see less overlap between the uncontaminated cluster and the contamination than we saw in the previous case with zero correlation in the cluster centered at the origin.

Trace plots of the parameter estimates can be seen in Figure 3.6. In this case,  $\hat{w}_\alpha$  starts at the MLE solution of 1 and begins to decline rapidly around an  $\alpha$  value of 0.3 until  $\alpha = 0.5$ , at which point it continues to decline less rapidly until reaching a value of less than 0.8 at  $\alpha = 2$ . The first component of the mean vector,  $\hat{\mu}_{\alpha,1}$  is inflated for  $\alpha$  values between 0 and 0.5, at which point it has reached the true value of  $\mu_1 = 0$  and then stabilizes.  $\hat{\sigma}_{\alpha,1}$ , the standard deviation of the first variable, is inflated for  $\alpha$  values between 0 and 0.5, and then it steadily declines until reaching its true value around  $\alpha = 1.7$ . The correlation estimate,  $\hat{\rho}_{\alpha,12}$  actually increases for the  $\alpha$  range of (0, 0.5) and then steadily declines until reaching the true value  $\rho_{12} = 0.75$  around  $\alpha = 1.7$ . Thus, we can see that  $\alpha$  values beyond 1 provide additional robustness compared to the  $L_2E$  solution when there is overlap between the main data and the contamination. The correlation  $\rho_{12}$  can also be captured by the MPDC- $\alpha$  divergence estimator for slightly lower  $\alpha$  values than are needed when there is zero correlation. Overall, in the  $p = 2$  case, when there is overlap between the main data and contamination, we turn to  $\alpha$  values between 1.5 and 2 to reach a sufficiently robust estimate. Another scenario we will consider is when the uncontaminated and contaminated data are well-separated.

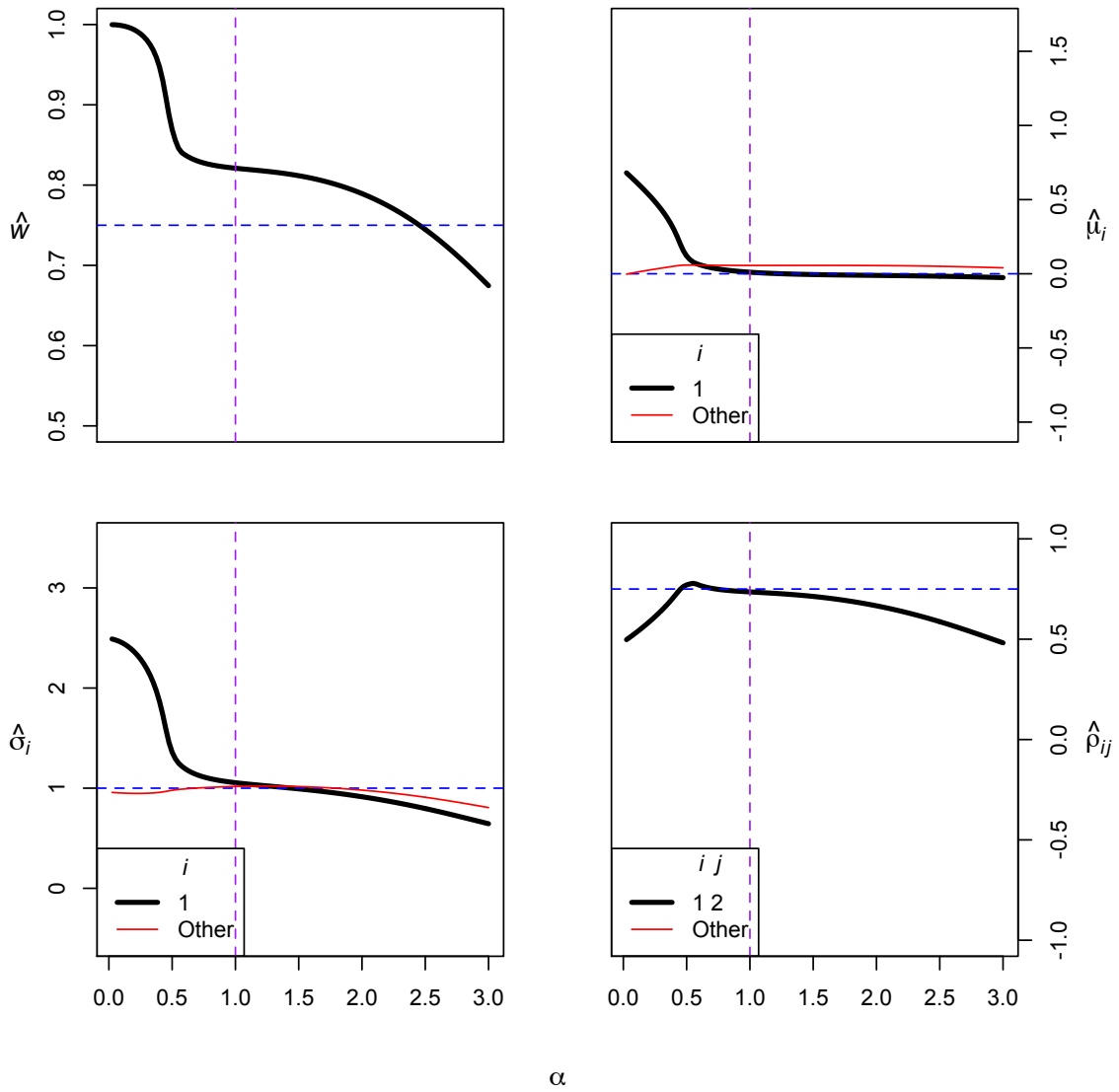


Figure 3.6: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 3 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0)'$ ;  $\mu_2 = (3, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_2$ .

### Well-Separated Clusters with Zero Correlation

We generate a sample of size  $n = 1000$  from the mixture model

$$\frac{3}{4} N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_2 \right) + \frac{1}{4} N \left( \begin{pmatrix} 7 \\ 0 \end{pmatrix}, I_2 \right).$$

Figure 3.7 shows a contour plot of the  $N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_2 \right)$  distribution overlain on the sample points. The uncontaminated and contaminated data are clearly separated, not showing the overlap we saw in the first example.

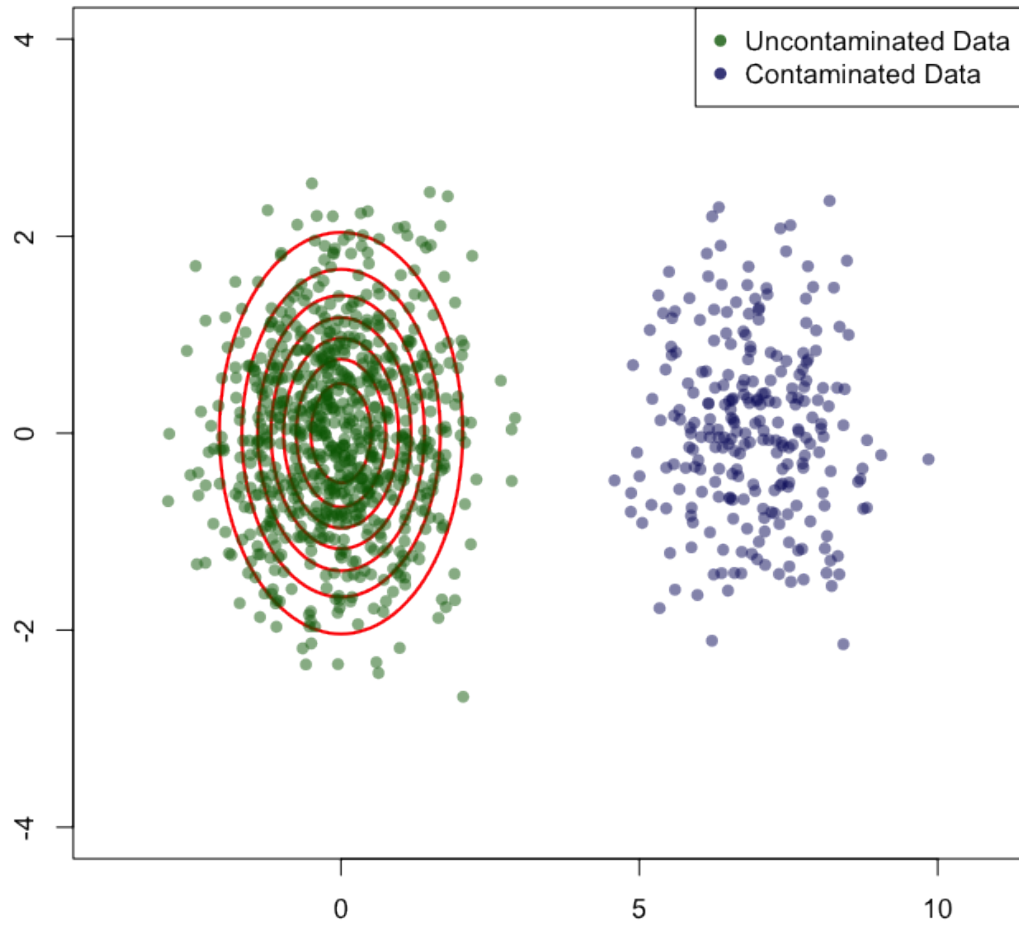


Figure 3.7: Well-separated clusters with zero correlation in the uncontaminated data. Contour lines (red) of  $N((0,0)', I_2)$  density overlain on sample of size  $n = 1000$  simulated from the Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0,0)'$ ;  $\mu_2 = (7,0)'$ ;  $\Sigma_1 = \Sigma_2 = I_2$ .

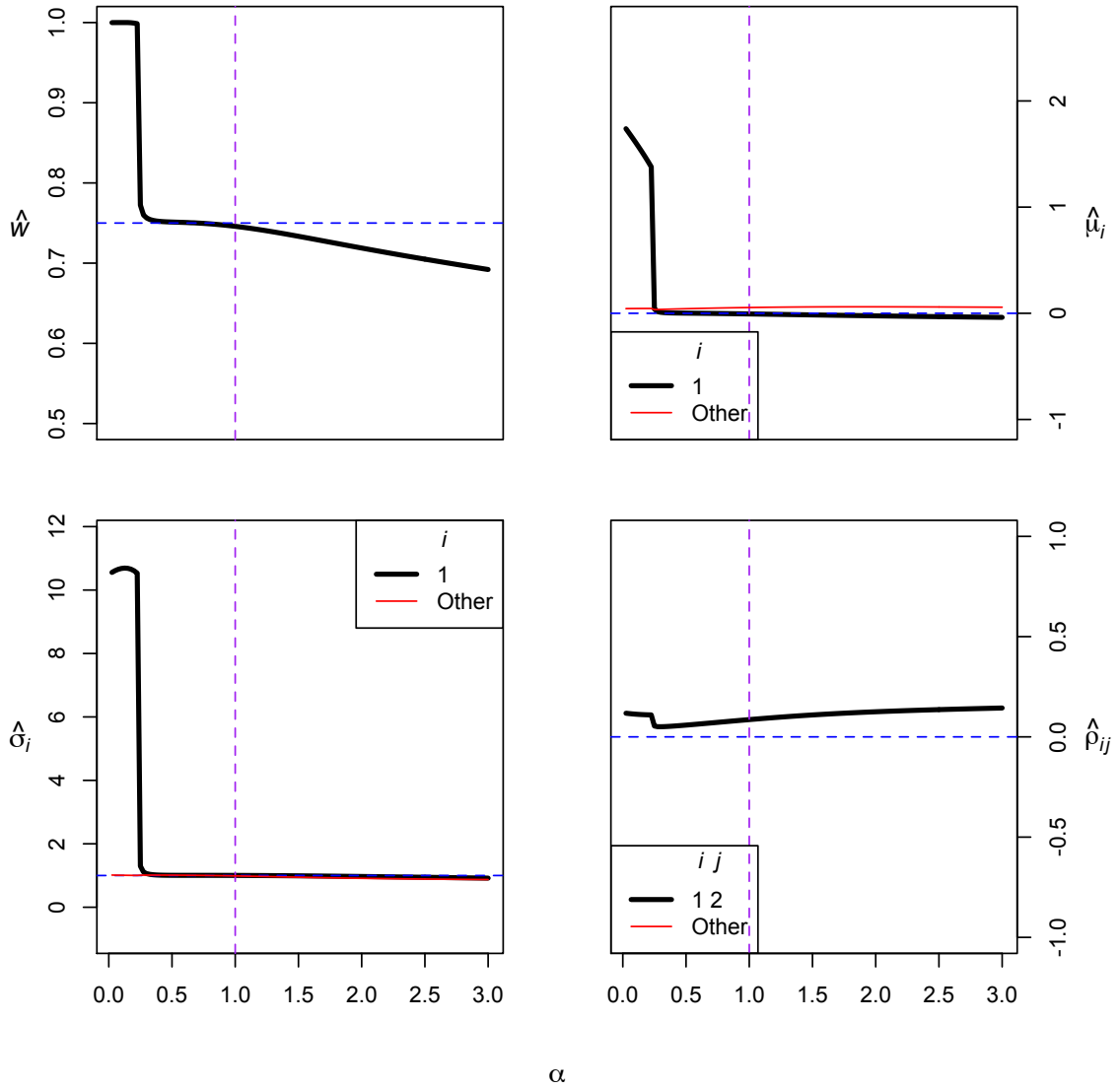


Figure 3.8: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 3 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0)'$ ;  $\mu_2 = (7, 0)'$ ;  $\Sigma_1 = \Sigma_2 = I_2$ .



As we can see in Figure 3.8, because the clusters are well-separated, we do not require as high of an  $\alpha$  value in order for  $\hat{\boldsymbol{\theta}}_\alpha$  to converge to  $\boldsymbol{\theta}$ . There is an abrupt drop in each trace plot around  $\alpha = 0.4$ , and each of the parameter estimates converges to its respective true value around  $\alpha = 0.5$ . We also note that the estimates given for  $\alpha$  values between 0 and 0.4 include the contaminated data and thus are significantly inflated, with  $\hat{\mu}_{\alpha,1}$  close to 2 (true value  $\mu_1 = 0$ ) and  $\hat{\sigma}_{\alpha,1}$  exceeding 10 (true value  $\sigma_1 = 1$ ). The correlation estimate,  $\hat{\rho}_{\alpha,12}$ , does not stray too far from the true value of  $\rho_{12} = 0$ . This begs the question of whether introducing a non-zero correlation in the main data would affect the range of  $\alpha$  values yielding consistent solutions for this example.

### Well-Separated Clusters with $\rho_{12} = 0.75$

Once again, we investigate what happens when we introduce a non-zero correlation in the cluster centered at  $\mu_1$ . Thus, we will consider the case where  $\rho_{12} = 0.75$ , i.e. we generate a sample of size  $n = 1000$  from the mixture model

$$\frac{3}{4} N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix} \right) + \frac{1}{4} N \left( \begin{pmatrix} 7 \\ 0 \end{pmatrix}, I_2 \right).$$

Figure 3.9 shows a contour plot of the  $N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.00 & 0.75 \\ 0.75 & 1.00 \end{pmatrix} \right)$  distribution overlain on the sample points.

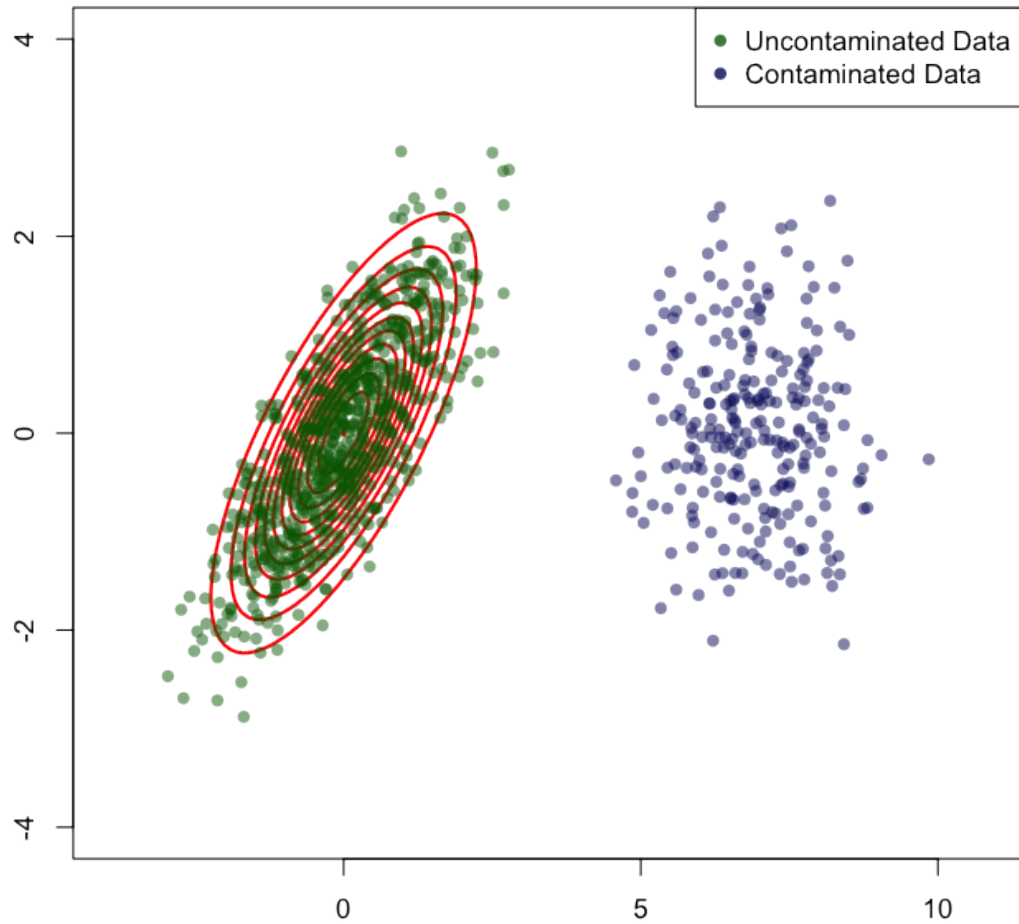


Figure 3.9: Well-separated clusters with  $\rho_{12} = 0.75$  in the uncontaminated data. Contour lines (red) of  $N((0, 0)', (\frac{1.00}{0.75} \ 0.75))$  density overlain on sample of size  $n = 1000$  simulated from the Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0)'$ ;  $\mu_2 = (7, 0)'$ ;  $\Sigma_1 = \Sigma_2 = I_2$ .

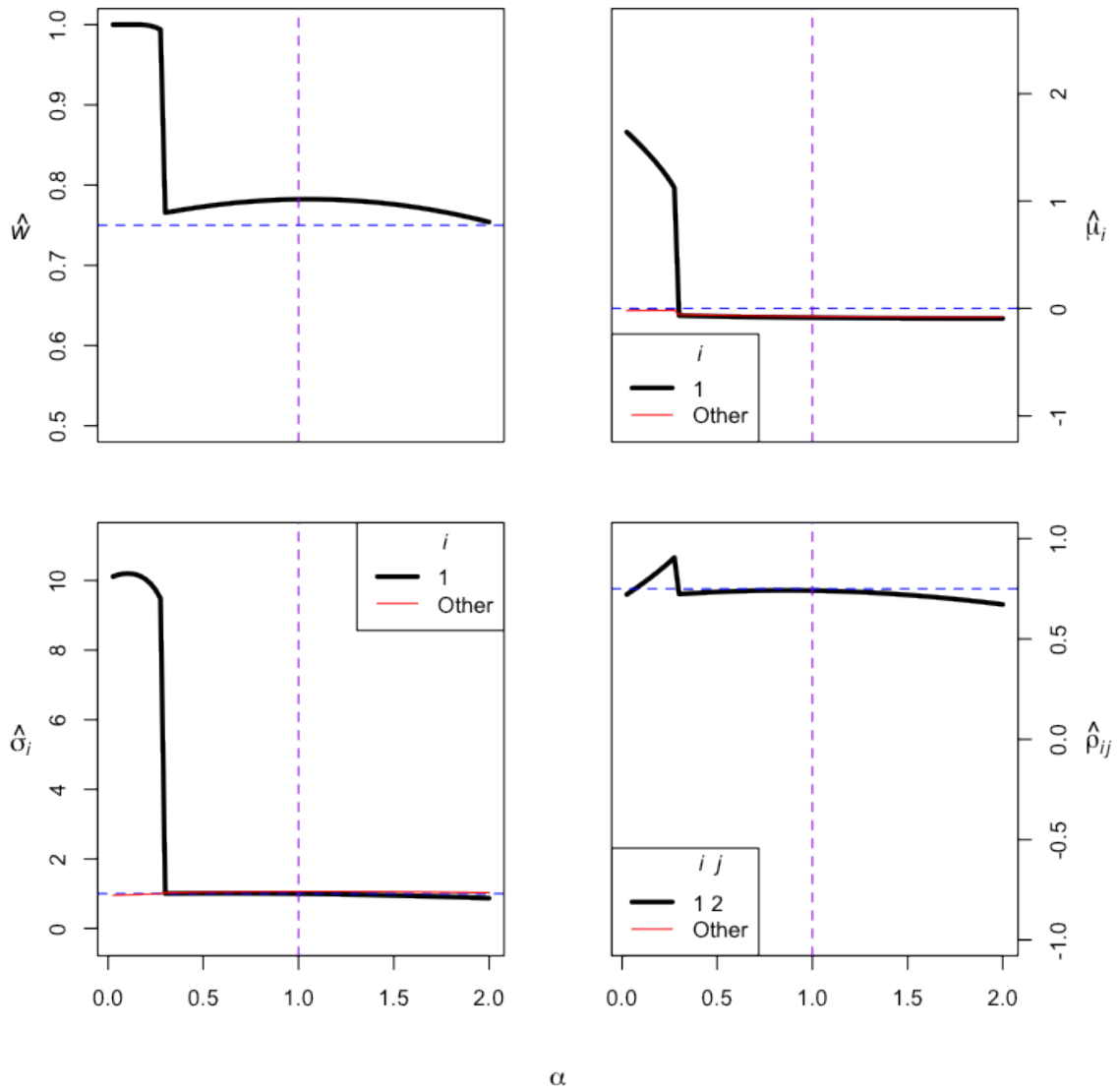


Figure 3.10: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0)'$ ;  $\mu_2 = (7, 0)'$ ;  $\Sigma_1 = \Sigma_2 = I_2$ .

In Figure 3.10 we can see that  $\hat{\boldsymbol{\theta}}_\alpha$  is not significantly affected by the correlation,  $\rho_{12}$ , as the estimates have trace plots very similar to those in the previous example. Thus, in the  $p = 2$  case, when the main data and contamination are well-separated, we do not require  $\alpha$  values much greater than 0.5 to reach a sufficiently robust estimate. We now investigate the behavior of the MPDC- $\alpha$  divergence estimator in cases of  $p$  greater than 2.

### 3.3.2 Simulated Cases for $p = 3$

We will explore analogous cases for higher dimensional problems. When  $p \geq 3$ , we cannot identify outliers by visual inspection, so we rely on alternative methods for detection and management of contamination. In our case, we would like to identify a range of  $\alpha$  values for which  $\hat{\boldsymbol{\theta}}_\alpha$  is a consistent estimate of  $\boldsymbol{\theta}$ .

#### Overlapping Clusters

We consider a 3-dimensional case, generating a sample of size  $n = 1000$  from the mixture model

$$\frac{3}{4} N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.75 & 0 \\ 0.75 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) + \frac{1}{4} N \left( \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}, I_3 \right).$$

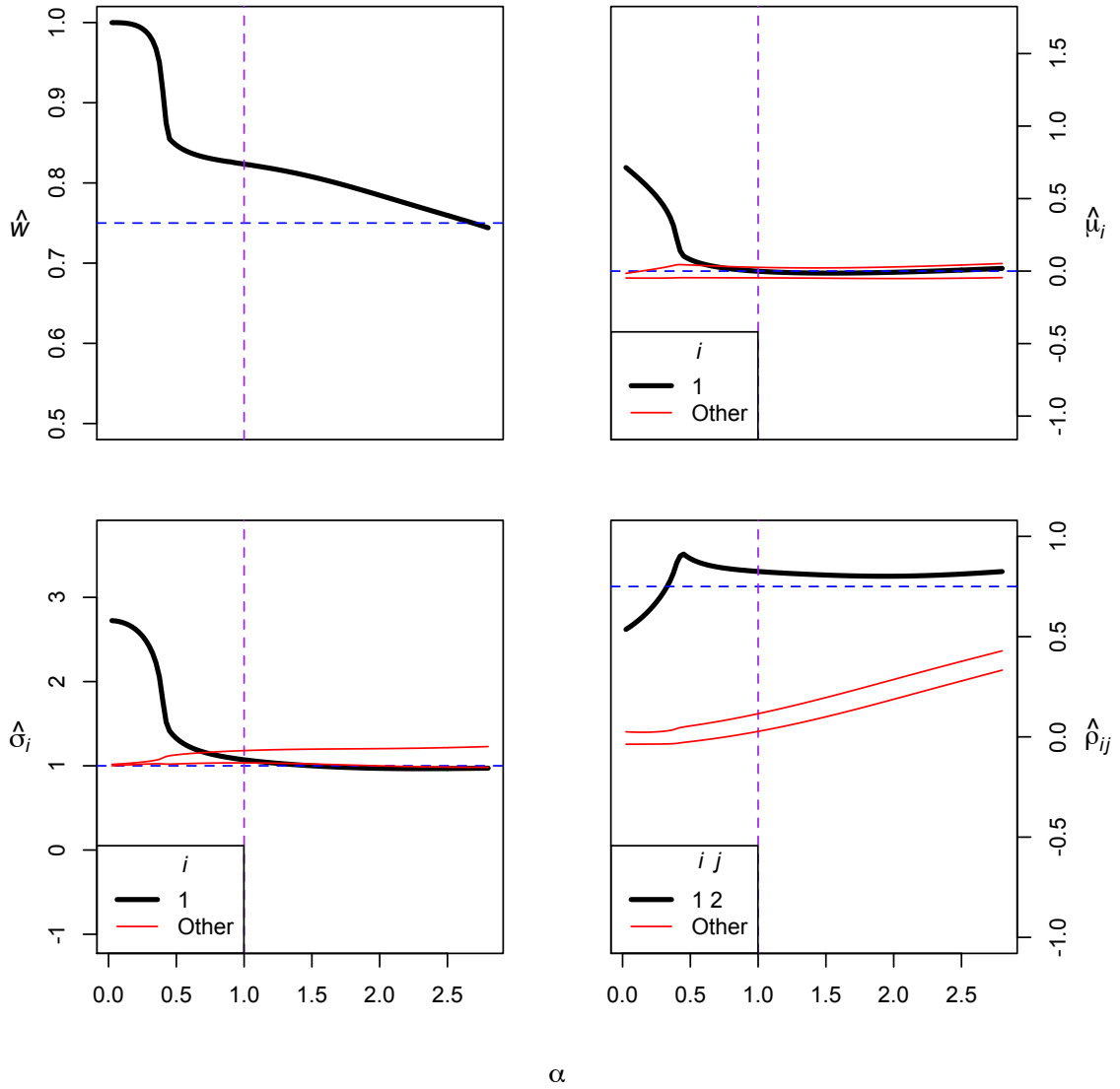


Figure 3.11: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2.5 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0)'$ ;  $\mu_2 = (3, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_3$ .

As we can see in Figure 3.11, our estimates begin to approach their true values around an  $\alpha$  value of 0.5, converging around  $\alpha = 0.7$ . Like the  $p = 2$  case, the overlap between the main data and contamination yields an estimate of the weight parameter,  $\hat{w}_\alpha$ , that is slightly higher than the true value of  $w$ . We see that as  $\alpha$  increases beyond 1, estimates for the other parameters that are not our primary focus (the red lines) begin to stray from their true values. Thus, we see that there is no additional robustness benefit from  $\alpha$  values beyond 1 in this case. As we did for the 2-dimensional case, we would like to examine whether increasing the separation between the uncontaminated and contaminated data will change the range of  $\alpha$  values yielding consistent estimates.

### Well-Separated Clusters

To explore the effect of this increased separation, we generate a sample of size  $n = 1000$  from the mixture model

$$\frac{3}{4} N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.75 & 0 \\ 0.75 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) + \frac{1}{4} N \left( \begin{pmatrix} 7 \\ 0 \\ 0 \end{pmatrix}, I_3 \right).$$

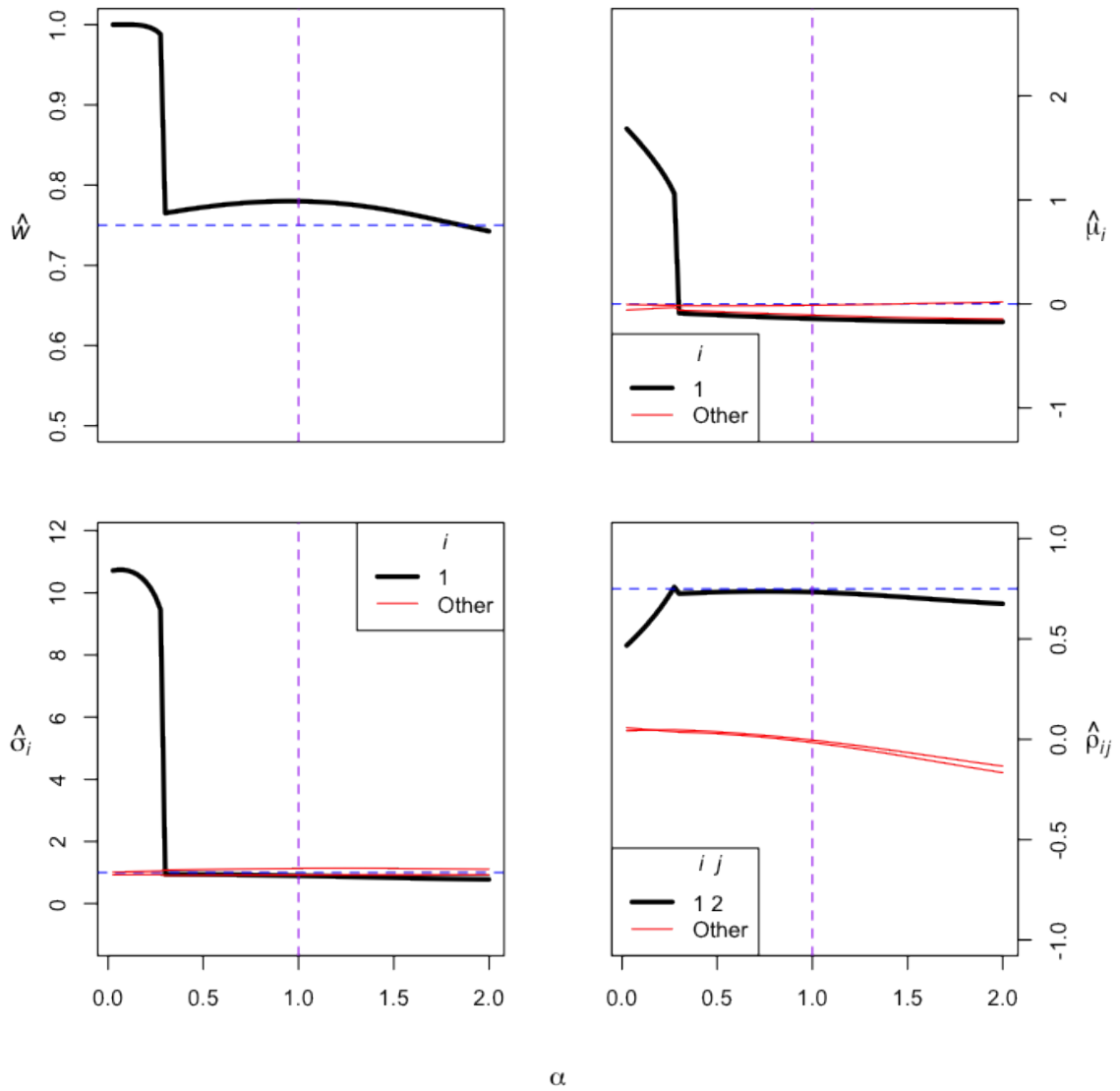


Figure 3.12: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0)'$ ;  $\mu_2 = (7, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_3$ .

Figure 3.12 shows that the estimates reach their true values around  $\alpha = 0.3$ . The difference between this setting and the previous one is that in this situation, rather than smoothly transitioning to a consistent solution, the estimates quickly switch (around  $\alpha = 0.3$ ) from incorporating all of the data to just considering the uncontaminated cluster. Thus, just as in the 2-dimensional case, when the clusters are well-separated we do not require  $\alpha$  values as high as are needed for cases with overlapping clusters to achieve consistent solutions. Overall, the  $\alpha$  values needed to provide robust estimates are lower for  $p = 3$  than they are for analogous cases with  $p = 2$ . We will see if this trend continues as we push the dimension,  $p$ , even higher.

### 3.3.3 Simulated Cases for $p = 4$

#### Overlapping Clusters

Continuing to increase the dimension  $p$ , we generate a sample of size  $n = 1000$  from the mixture model

$$\frac{3}{4}N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.75 & 0 & 0 \\ 0.75 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \right) + \frac{1}{4}N \left( \begin{pmatrix} 3 \\ 0 \\ 0 \\ 0 \end{pmatrix}, I_4 \right).$$



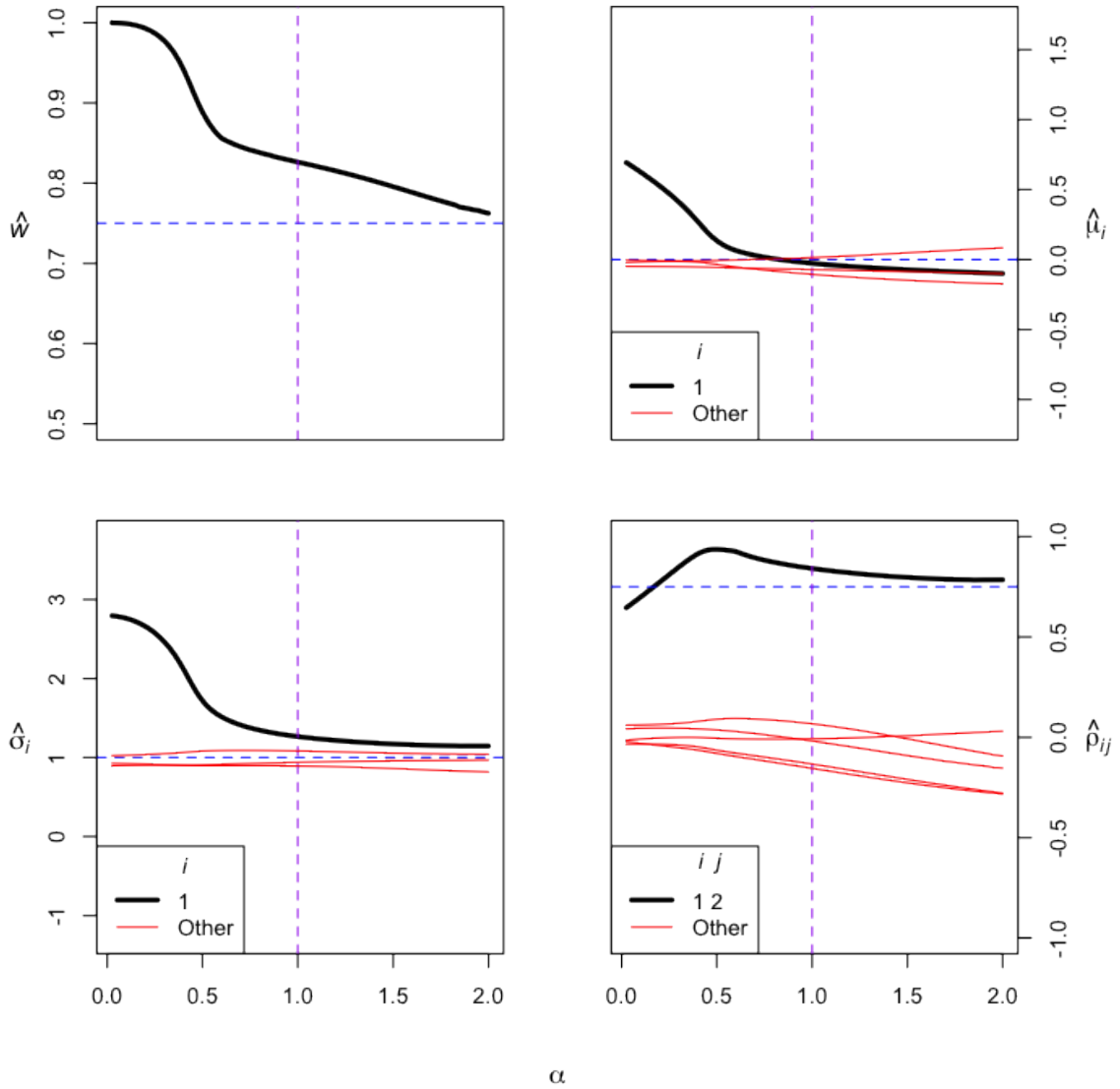


Figure 3.13: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0)'$ ;  $\mu_2 = (3, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_4$ .

Figure 3.13 shows that the parameter estimates begin to get close to their true values around  $\alpha = 0.5$ , converging around  $\alpha = 0.8$ . Once again,  $\hat{w}_\alpha$  is still slightly inflated due to the overlap between clusters. The results are very similar to the analogous case for  $p = 3$ .

### Well-Separated Clusters

Increasing cluster separation, we generate a sample of size  $n = 1000$  from the mixture model

$$\frac{3}{4}N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.75 & 0 & 0 \\ 0.75 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \right) + \frac{1}{4}N \left( \begin{pmatrix} 7 \\ 0 \\ 0 \\ 0 \end{pmatrix}, I_4 \right).$$

Figure 3.14 shows that the parameter estimates converge to their true values around  $\alpha = 0.3$ . Due to the separation between clusters, we do not see much inflation in  $\hat{w}_\alpha$ . The results are very similar to the analogous case for  $p = 3$ .

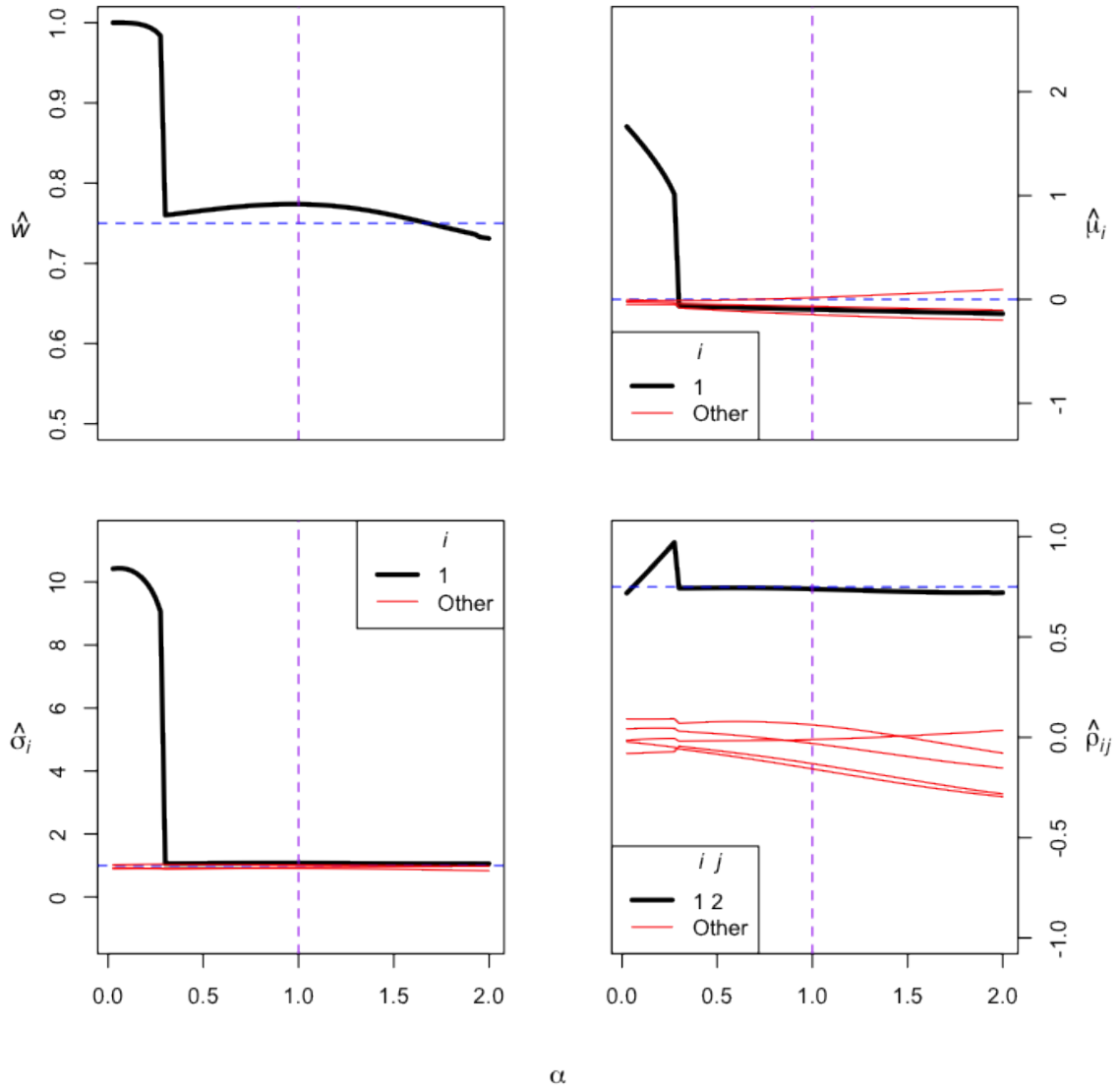


Figure 3.14: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0)'$ ;  $\mu_2 = (7, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_4$ .

### 3.3.4 Simulated Cases for $p = 5$

#### Overlapping Clusters

For a 5-dimensional example, we generate a sample of size  $n = 1000$  from the mixture model

$$\frac{3}{4}N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.75 & 0 & 0 & 0 \\ 0.75 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \right) + \frac{1}{4}N \left( \begin{pmatrix} 3 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, I_5 \right).$$

As we can see in Figure 3.15, the parameter estimates get close to their true values around  $\alpha = 1$ . We continue to see the slight inflation in  $\hat{w}_\alpha$  due to the overlap between the main data and contamination. It seems that the minimum  $\alpha$  required to yield a consistent solution does not decrease monotonically with increasing dimension,  $p$ .

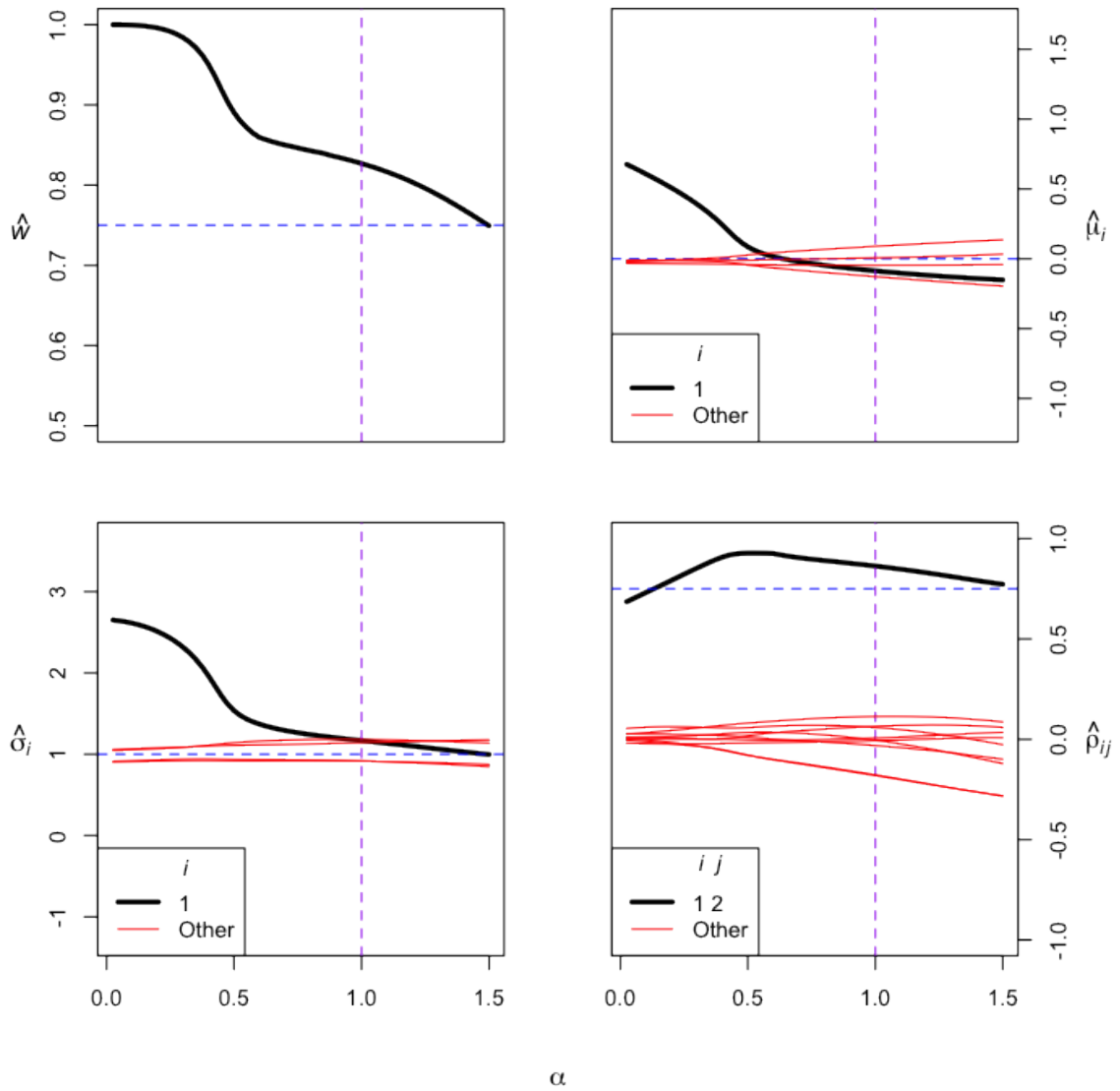


Figure 3.15: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 1.5 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0)'$ ;  $\mu_2 = (3, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_5$ .

## Well-Separated Clusters

Increasing the separation between clusters, we generate a sample of size  $n = 1000$  from the mixture model

$$\frac{3}{4}N \begin{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.75 & 0 & 0 & 0 \\ 0.75 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{pmatrix} + \frac{1}{4}N \begin{pmatrix} \begin{pmatrix} 7 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, I_5 \end{pmatrix}.$$

Figure 3.16 shows essentially the same picture as we have seen in the previous analogous cases where the contamination is centered at a point with first component 7. The estimates converge to their respective true values around  $\alpha = 0.3$ .

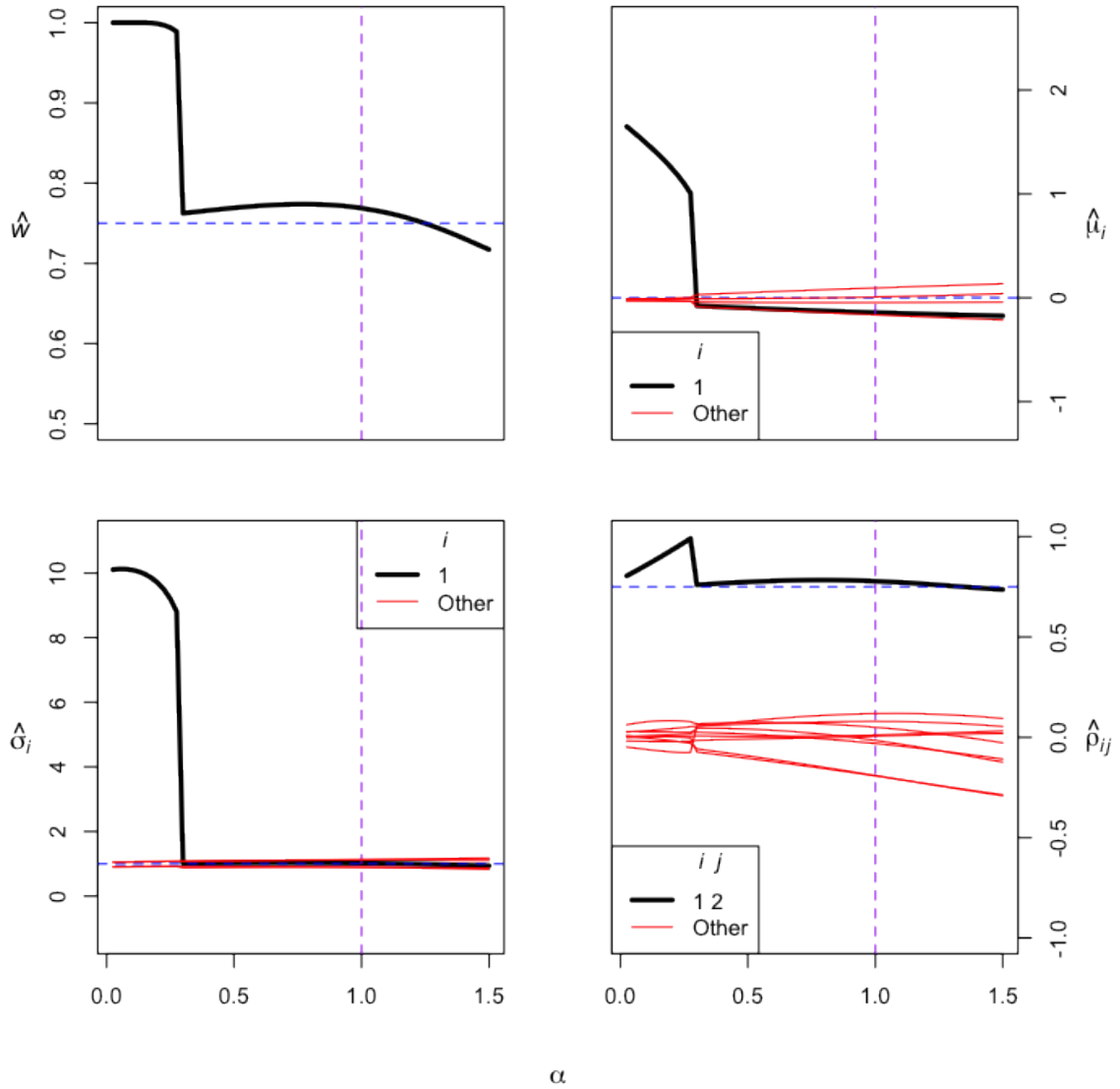


Figure 3.16: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 1.5 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0)'$ ;  $\mu_2 = (7, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_5$ .

### 3.3.5 Simulated Cases for $6 \leq p \leq 10$

The trace plots for the cases of  $p \in [6, 10]$  can be found in the Appendix. We examine both overlapping and well-separated cases for those dimensions. We wish to utilize these simulated cases to help establish an exploratory method of selecting an optimal  $\alpha$  level that will yield consistent robust estimates for a particular dataset. In order to do so, we will have to repeat these sampling procedures multiple times.

### 3.3.6 Selecting $\alpha$

As there is no universally agreed upon method to select  $\alpha$  when utilizing this estimation method, we can take a number of approaches to select the appropriate  $\alpha$  values. Because of the tradeoff between robustness and efficiency as  $\alpha$  increases, we could set a desired level of efficiency we wish to attain and choose  $\alpha$  accordingly. Our approach in the following real data example will be to exploit a prior notion of the extent of contamination in the model, namely the proportion of contamination and its degree of separation from the main data. Thus, we wish to understand how the optimal  $\alpha$  value varies with the level of separation and the dimension,  $p$ .

We will utilize our simulated cases in the previous section, generating  $M = 100$  iterations of samples from the mixture

$$\frac{3}{4} N(\mathbf{0}_p, \boldsymbol{\Sigma}_1) + \frac{1}{4} N((s, \mathbf{0}_{p-1})', I_p) \quad (3.7)$$

where  $(\boldsymbol{\Sigma}_1)_{ij} = 0.75$  for  $\{(i, j) = (1, 2) \text{ or } (2, 1)\}$  and  $(\boldsymbol{\Sigma}_1)_{ij} = \delta_{ij}$  else. The value of  $\alpha$  will range from 0.025 to 3 in increments of 0.025. Our separation variable,  $s$ , will be 3 (overlap), 5 (minimal overlap), or 7 (no overlap). The dimension  $p$  will range from



2 to 10. Thus, for each value of  $p$  we will have 360 estimates  $\hat{\boldsymbol{\theta}}_\alpha$  (120  $\alpha$  values x 3  $s$  values). For each of these  $\hat{\boldsymbol{\theta}}_\alpha$  we compute

$$RMSE(\hat{\boldsymbol{\theta}}_\alpha) = \sqrt{\frac{1}{M} \sum_{i=1}^M (\hat{\boldsymbol{\theta}}_{\alpha,i} - \boldsymbol{\theta}_{\alpha,i})^2}.$$

To provide an exploratory method of  $\alpha$  selection, we examine the RMSE values versus  $\alpha$  for these various values of  $p$  and  $s$ . Thus, for a given dimension and degree of separation, we can select the range of  $\alpha$  which gives us the smallest RMSE. On each plot we will specify the  $\alpha$  value at which the minimum RMSE is attained as well as a range of  $\alpha$  values for which the RMSE is within 10% of its minimum.

Case:  $p=2$

Figure 3.17 displays the relationship between RMSE and  $\alpha$  for the 2-dimensional case. We require values of  $\alpha$  between 0.8 and 1.4 to yield consistent estimates when there is overlap between the main data and contamination ( $s = 3$ ). If the clusters are well-separated, i.e.  $s \in \{5, 7\}$ , then  $\alpha$  values between 0.3 and 0.5 are optimal.

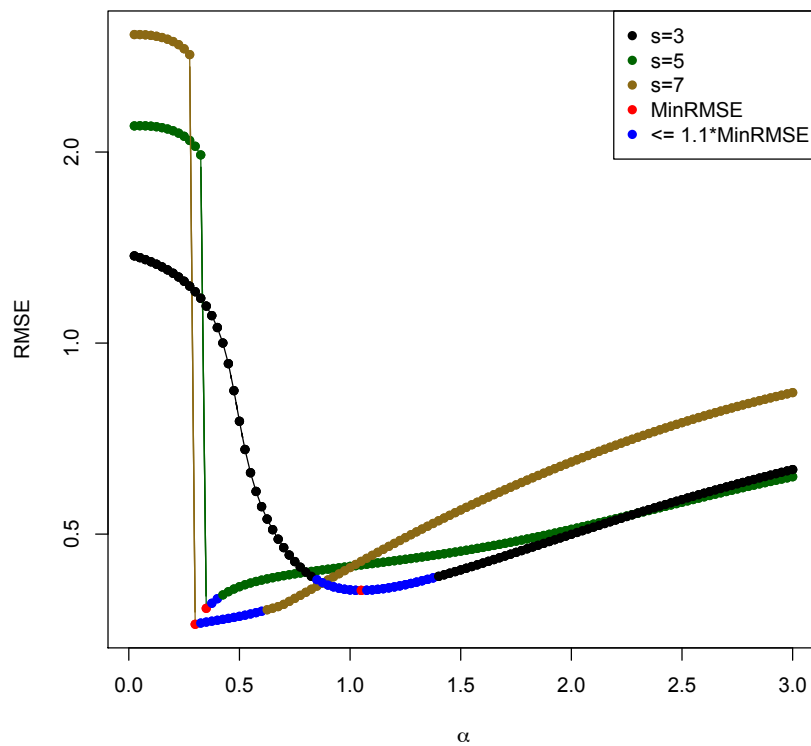


Figure 3.17: Selection of  $\alpha$  for  $p = 2$ . RMSE of MPDC- $\alpha$  estimate versus  $\alpha$  for three different degrees of separation ( $s$ ). Derived from 100 simulations of samples of size  $n = 1000$  from a Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0)'$ ;  $\mu_2 = (s, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_2$ . Points at which minimum RMSE is attained are in red, and points at which RMSE is within 10% of its minimum are in blue.

Case:  $p=3$

Figure 3.18 displays the RMSE plot for the 3-dimensional case. For the low-separation case ( $s = 3$ ), the optimal  $\alpha^*$  is approximately 1.2, with  $\alpha \in [0.8, 1.7]$  yielding fairly consistent solutions. Once again, when we increase separation between clusters,  $\alpha \in [0.3, 0.5]$  are optimal.

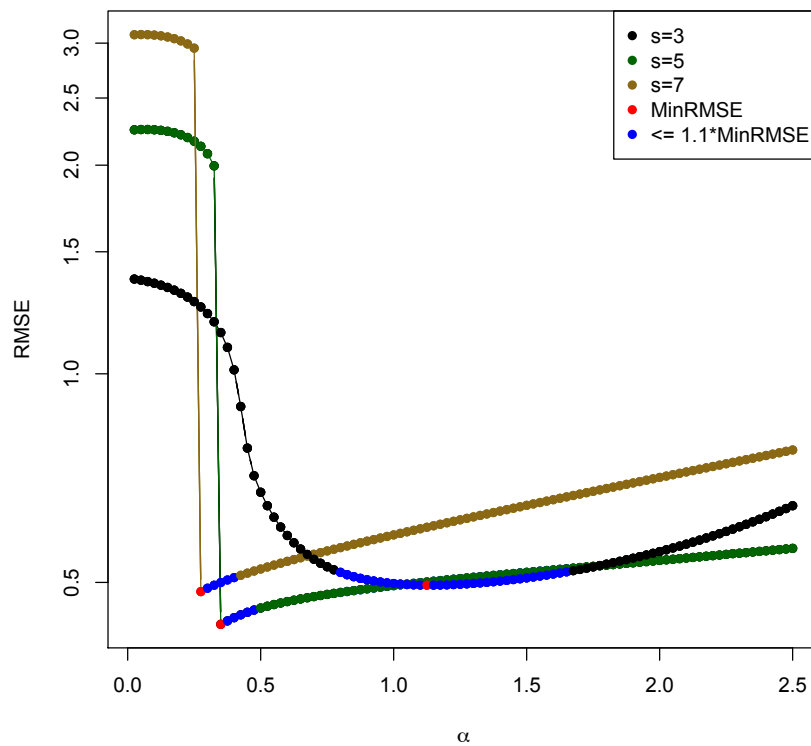


Figure 3.18: Selection of  $\alpha$  for  $p = 3$ . RMSE of MPDC- $\alpha$  estimate versus  $\alpha$  for three different degrees of separation ( $s$ ). Derived from 100 simulations of samples of size  $n = 1000$  from a Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0)'$ ;  $\mu_2 = (s, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_3$ .

Case:  $p=4$

Figure 3.19 displays the RMSE plot for the 4-dimensional case. For the low-separation case ( $s = 3$ ),  $\alpha^*$  is about 0.8, with  $\alpha \in [0.7, 0.9]$  yielding fairly consistent solutions. Once again, when we increase separation between clusters,  $\alpha \in [0.3, 0.5]$  are optimal.

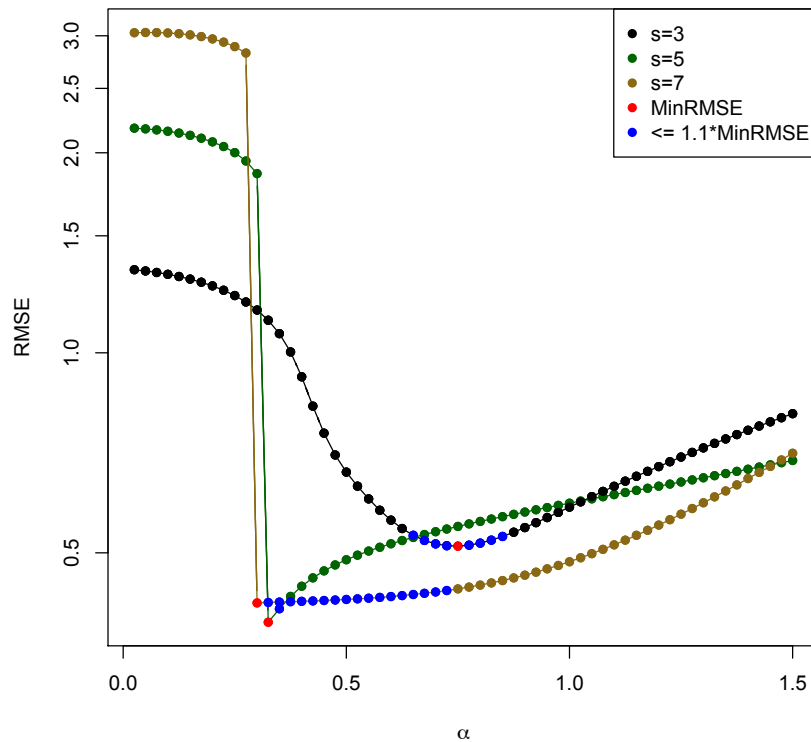


Figure 3.19: Selection of  $\alpha$  for  $p = 4$ . RMSE of MPDC- $\alpha$  estimate versus  $\alpha$  for three different degrees of separation ( $s$ ). Derived from 100 simulations of samples of size  $n = 1000$  from a Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0)'$ ;  $\mu_2 = (s, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_4$ .

Case:  $p=5$

Figure 3.20 displays the RMSE plot for the 5-dimensional case. For low values of separation, we rely on  $\alpha$  values between 0.8 and 1.1 for  $\hat{\theta}_\alpha$  to consistently estimate  $\theta$ , with an optimal value of  $\alpha^* = 0.9$ . An  $\alpha$  value of 0.3 or 0.4 is optimal when  $s = 5$  or 7.

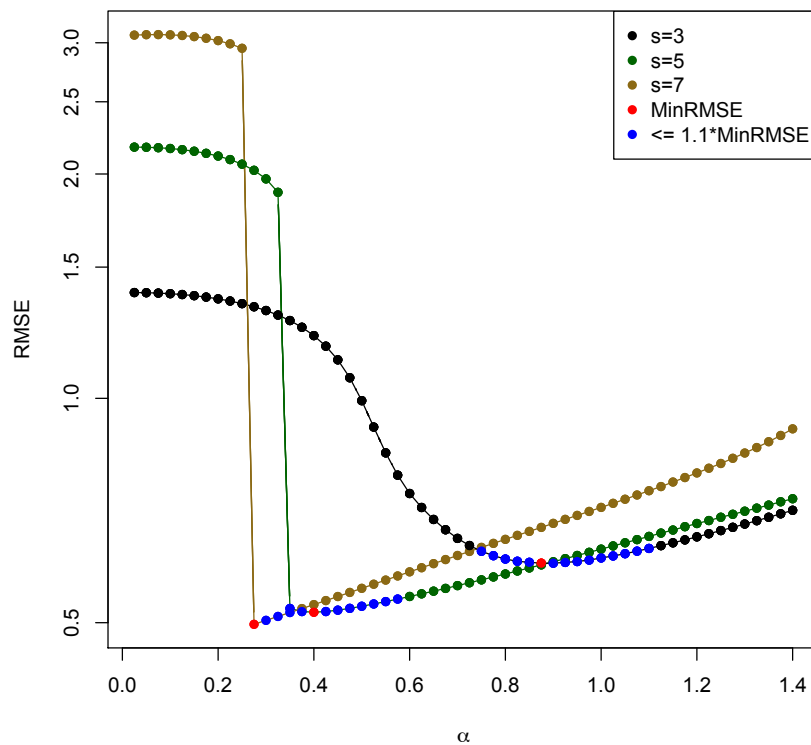


Figure 3.20: Selection of  $\alpha$  for  $p = 5$ . RMSE of MPDC- $\alpha$  estimate versus  $\alpha$  for three different degrees of separation ( $s$ ). Derived from 100 simulations of samples of size  $n = 1000$  from a Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0)'$ ;  $\mu_2 = (s, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_5$ .

**Cases:  $6 \leq p \leq 10$** 

The RMSE plots for the cases of  $6 \leq p \leq 10$  can be found in the Appendix. Rather than describing them all in detail here, we summarize the simulation results in a table.

**Summary**

Table 3.1 lists ranges of  $\alpha$  values that are optimal for situations when the main data and contamination are either overlapping or well-separated. This is for the case of  $n = 1000$  and 25% contamination.

$p$	Overlapping $\alpha^*$	Separated $\alpha^*$	$\alpha_{max}$
2	[0.8, 1.4]	[0.3, 0.5]	3.0
3	[0.7, 1.6]	[0.3, 0.5]	2.5
4	[0.7, 0.9]	[0.3, 0.4]	1.5
5	[0.7, 1.1]	[0.3, 0.4]	1.4
6	[0.6, 0.9]	[0.3, 0.4]	1.3
7	[0.6, 0.9]	[0.3, 0.4]	1.1
8	[0.5, 0.7]	[0.3, 0.4]	0.9
9	[0.5, 0.8]	[0.3, 0.4]	0.8
10	[0.5, 0.7]	[0.3, 0.4]	0.8

Table 3.1: Guidelines for selection of  $\alpha$  for various dimension ( $p$ ) values. The range of optimal  $\alpha$  values ( $\alpha^*$ ) and the maximum recommended  $\alpha$  value ( $\alpha_{max}$ ) are given for cases of overlapping and well-separated clusters.

We should note that the minimum  $\alpha$  value required to attain the true solution, for a fixed level of separation  $s$ , does not decrease monotonically as the dimension  $p$  increases. There is more to be done on understanding this pattern.

We will now seek to apply the MPDC- $\alpha$  divergence estimation procedure to a real data example.

## 3.4 Application: Baseball

The statistical analysis of sports is becoming increasingly prevalent, particularly in baseball since the nature of the game is inherently conducive to such analysis. Recently, there has been more attention given to the theory of Defense Independent Pitching Statistics (DIPS), the idea that we can isolate a pitcher’s influence on the outcome of an at-bat.

Fairly new data has been made available via cameras that track various components of pitches for every at-bat during a Major League Baseball season. We pulled together this data, called PITCHf/x, from every at-bat during the 2009, 2010, 2011 MLB seasons, yielding a dataset consisting of  $N = 132,103$  observations and  $p = 19$  variables. Among the variables we will examine are horizontal movement, vertical movement, pitch speed, break and spin.

### 3.4.1 Robust Parametric Estimation for PITCHf/x Variables

Many of the characteristics of a pitch can be attributed to its type. Naturally, fastballs and curveballs tend to have different speeds, movement, break angles and spin. We can crudely classify a given pitch as a breaking ball (curveball, slider, change-up or knuckleball) or a fastball (four-seam fastball, two-seam fastball, cut fastball or sinker).

#### Horizontal and Vertical Movement

First, we explore the relationship between horizontal and vertical movement (denoted “pfx\_x” and “pfx\_z”, respectively). We seek to estimate the joint distribution of pfx\_x



and  $\text{pfx}_z$ , a bivariate problem. Because we are looking at breaking balls and fastballs as different populations, we will have separate  $\text{pfx}_x$  and  $\text{pfx}_z$  distributions for each of those two pitch types. Before proceeding with our analysis, we discard 7,444 data points with pitch type labels that do not fall into either of the two categories (e.g. pitch-outs), leaving us with a sample size of  $N = 124,659$ . Based on prior research (Stackpole 2012), the distributions of  $\text{pfx}_x$  and  $\text{pfx}_z$  are approximately Normal. Also, the data indicates that 65% of pitches are fastballs. A summary of the data can be found in Table 3.2.

Fastball	Breaking	All
$n_1 = 96,218$	$n_2 = 28,441$	$N = 124,659$
$\bar{\mathbf{x}}_1 = (-2.45, 6.99)'$	$\bar{\mathbf{x}}_2 = (2.03, -0.98)'$	$\bar{\mathbf{x}} = (-1.43, 5.18)'$
$\mathbf{S}_1 = \begin{pmatrix} 44.1 & 2.6 \\ 2.6 & 13.0 \end{pmatrix}$	$\mathbf{S}_2 = \begin{pmatrix} 16.3 & -2.9 \\ -2.9 & 21.8 \end{pmatrix}$	$\mathbf{S} = \begin{pmatrix} 41.3 & -5.0 \\ -5.0 & 26.2 \end{pmatrix}$
$r_1 = 0.1$	$r_2 = -0.2$	$r = -0.2$

Table 3.2: Sample statistics for horizontal and vertical movement of fastballs, breaking pitches, and the full sample.

In order to apply the MPDC- $\alpha$  estimator to this example, we begin by standardizing the values of  $\text{pfx}_x$  and  $\text{pfx}_z$ . The sample mean is  $\bar{\mathbf{x}} = (-1.43, 5.18)'$ , and the sample covariance matrix is  $\mathbf{S} = \begin{pmatrix} 41.3 & -5.0 \\ -5.0 & 26.2 \end{pmatrix}$ . We are trying to estimate the weight parameter, mean and covariance matrix for the fastballs. Because we have the benefit of knowing the pitch types for every sample point, we can utilize these starting values

to attempt to locate the “fastball” component:  $\hat{w} = 0.65$ ,  $\bar{\mathbf{x}}_1 = (-2.45, 6.99)'$ , and  $\mathbf{S}_1 = \begin{pmatrix} 44.1 & 2.6 \\ 2.6 & 13.0 \end{pmatrix}$ . In order to use the MPDC- $\alpha$  divergence estimator, we apply the following transformation to the sample data points:

$$\mathbf{W}_{ij} = \frac{\mathbf{X}_{ij} - \bar{\mathbf{x}}_i}{\sqrt{\mathbf{S}_{jj}}} \quad (3.8)$$

Figure 3.21 shows a contour plot of the estimated distribution of the standardized values of  $(pfx\_x, pfx\_z)$  for straight pitches overlain on the standardized sample points, which are colored by pitch type (green for fastball and blue for breaking ball). We can see that there is considerable overlap between the clusters. We seek to estimate the weight parameter, mean vector and covariance matrix for the fastball group using the MPDC- $\alpha$  divergence estimator.

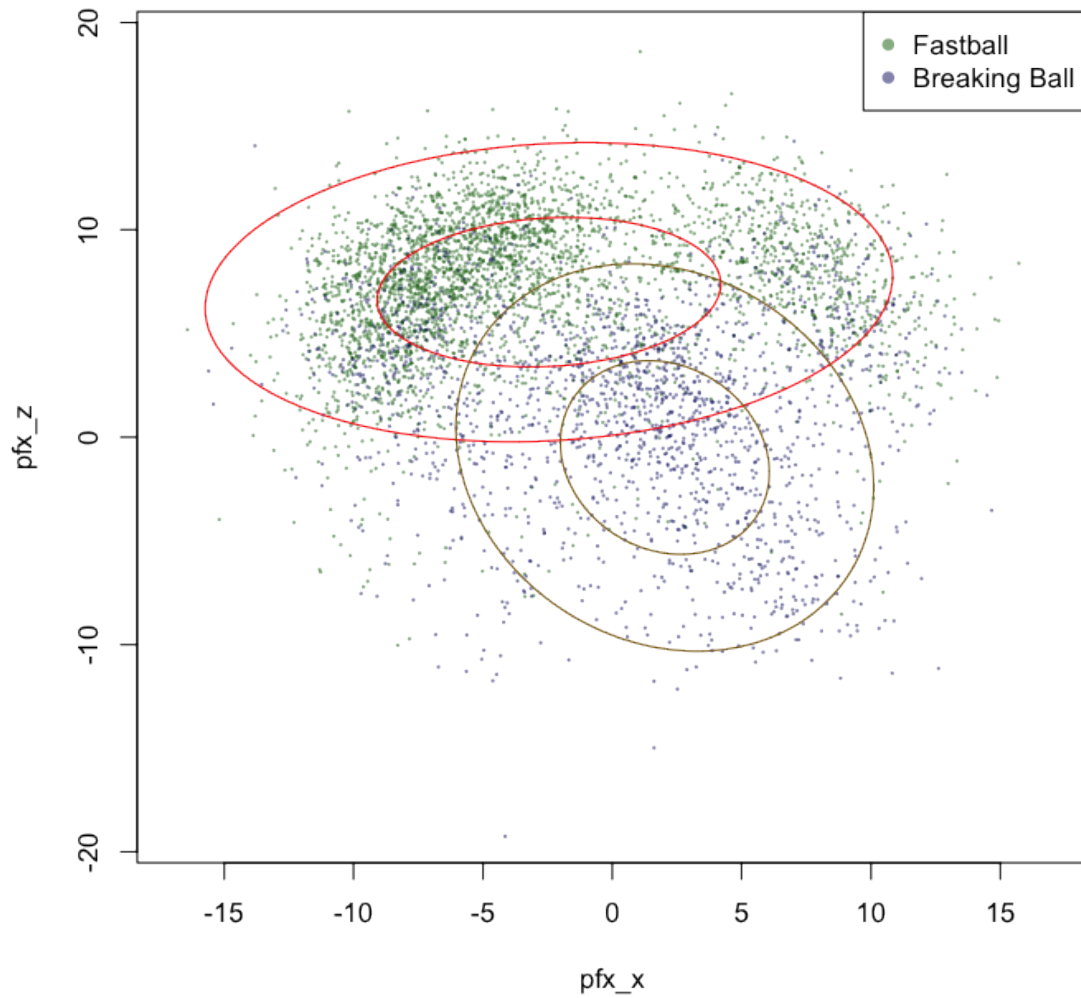


Figure 3.21: Horizontal and vertical movement  $(pfx_x, pfx_z)$  for a random sample of 5,000 pitches, with fastballs in green and breaking balls in blue. The estimated bivariate distribution of  $(pfx_x, pfx_z)$  for fastballs and breaking balls is represented by the 1- and 2-sigma ellipses in red and gold, respectively.  $\hat{w}_{fastball, \alpha} = 0.77$ ;  $\hat{w}_{breaking, \alpha} = 0.41$ .

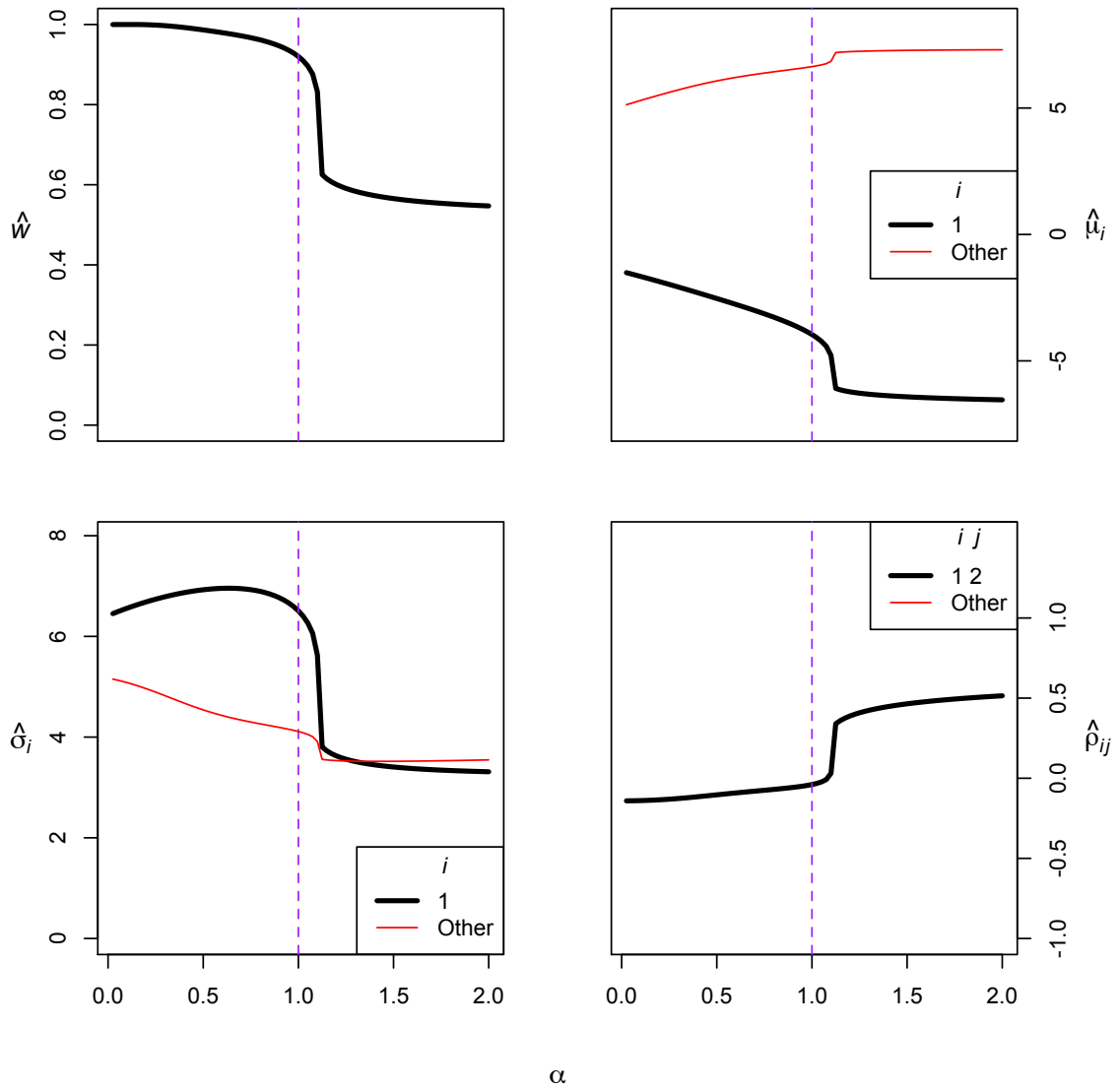


Figure 3.22: Trace plots of MPDC- $\alpha$  estimates for the distribution of  $(pfx_x, pfx_z)$  for fastballs for  $\alpha$  ranging from 0 to 2 in increments of 0.025. The black lines represent parameters of interest, and the red lines show other parameters to track algorithm stability.

Upon applying the transformation in (3.8), we seek to estimate  $\hat{w}_\alpha$ ,  $\hat{\mu}_\alpha$  and  $\hat{\Sigma}_\alpha$ .

Thus, we compute the MPDC- $\alpha$  divergence estimator for  $\alpha$  values in the range  $[0,2]$ .

Figure 3.22 shows trace plots of  $\hat{w}_\alpha$ ,  $\hat{\mu}_{j,\alpha}$ ,  $\hat{\sigma}_{j,\alpha}$  and  $\hat{\rho}_{12,\alpha}$ .

We can see that  $\hat{w}_\alpha$  is 1 for smaller  $\alpha$  values, and it gets closest to  $\hat{w} = 0.65$  around  $\alpha = 1.2$ . Transforming back to the original units, the other parameter estimates also fairly consistently estimate their corresponding true values for that  $\alpha$  level:  $\hat{w}_{(\alpha=1.2)} = 0.77$ ,  $\hat{\boldsymbol{\mu}}_{(\alpha=1.2)} = (-4.58, 6.97)'$ ,  $\hat{\sigma}_{1,(\alpha=1.2)}^2 = 35.5$ ,  $\hat{\sigma}_{2,(\alpha=1.2)}^2 = 15.7$ ,  $\hat{\rho}_{12,(\alpha=1.2)} = -0.01$ . This is a two-dimensional example with overlap, so we would expect that an  $\alpha$  value greater than 1 would be needed. The  $\alpha$  value of 1.2 closely matches the suggested value from the exploratory  $\alpha$  selection ranges we established in Section 3.3.6.

While this example takes advantage of the fact that we know whether every individual pitch in the data is a fastball or breaking ball, we can see how the MPDC- $\alpha$  divergence method would allow us to perform parametric estimation in practical situations where the labels are unknown. We could utilize our prior understanding of the separation between the main data and contamination, and assuming our sample is sufficiently large, we would simply select the appropriate  $\alpha$  value for the particular dimension  $p$  of the problem to yield consistent robust estimates of the parameters of interest.

Returning to the context of this example, we see that fastballs tend to move about 4.6 inches to a catcher's left and have, on average, 6.97 inches less downward movement than pitches without spin. There is also no apparent correlation between horizontal pitch movement (`px_x`) and vertical pitch movement (`px_z`) for straight pitches. Clearly, there are other variables to be considered that could potentially have an effect on the outcome of a pitch, including the pitch speed, break and spin.

## Pitch Speed, Break and Spin

We now add three more variables to the picture: pitch start speed, break angle, and spin direction. Including horizontal and vertical movement, we have a 5-dimensional problem with variables ( $pfx_x, pfx_z, start\_speed, break\_angle, spin\_rate$ ).

Fastball	Breaking
$n_1 = 96,218$	$n_2 = 28,441$
$\bar{\mathbf{x}}_1 = (-2.5, 7.0, 89, 10.0, 1975)'$	$\bar{\mathbf{x}}_2 = (2.0, -1.0, 81, -5.2, 973)'$
$\mathbf{S}_1 = \begin{pmatrix} 44.1 & 2.6 & -5.8 & -164.2 & -84.2 \\ 2.6 & 13.0 & 5.1 & -0.9 & 1133 \\ -5.8 & 5.1 & 19.9 & 28.9 & 931.2 \\ -164.2 & -0.9 & 28.9 & 674.6 & 1290 \\ -84.2 & 1133 & 931.2 & 1290 & 279230 \end{pmatrix}$	$\mathbf{S}_2 = \begin{pmatrix} 16.3 & -2.9 & -1.7 & -34.6 & 639 \\ -2.9 & 21.8 & 13.9 & 0.7 & -1061 \\ -1.7 & 13.9 & 27.6 & -2.9 & -916 \\ -34.6 & 0.7 & -2.9 & 83.1 & -1033 \\ 639 & -1061 & -915.5 & -1033.5 & 244715 \end{pmatrix}$
$r_{12,1} = 0.1$	$r_{12,2} = -0.2$
<hr style="width: 50%; margin: 0 auto;"/> <p style="text-align: center; margin: 0;">All</p> <hr style="width: 50%; margin: 0 auto;"/>	
$N = 124,659$	
$\bar{\mathbf{x}} = (-1.5, 4.9, 87, 6.7, 1714)'$	
$\mathbf{S} = \begin{pmatrix} 40.3 & -5.7 & -11.5 & -140.9 & -737 \\ -5.7 & 28.5 & 20.3 & 22.3 & 2156 \\ -11.5 & 20.3 & 36.0 & 42.5 & 2095 \\ -140.9 & 22.3 & 42.5 & 551.1 & 3387 \\ -737 & 2156 & 2095 & 3387 & 454580 \end{pmatrix}$	
$r_{12} = -0.2$	

Table 3.3: Sample statistics for horizontal and vertical movement, pitch speed, break and spin for fastballs, breaking balls and the full sample.

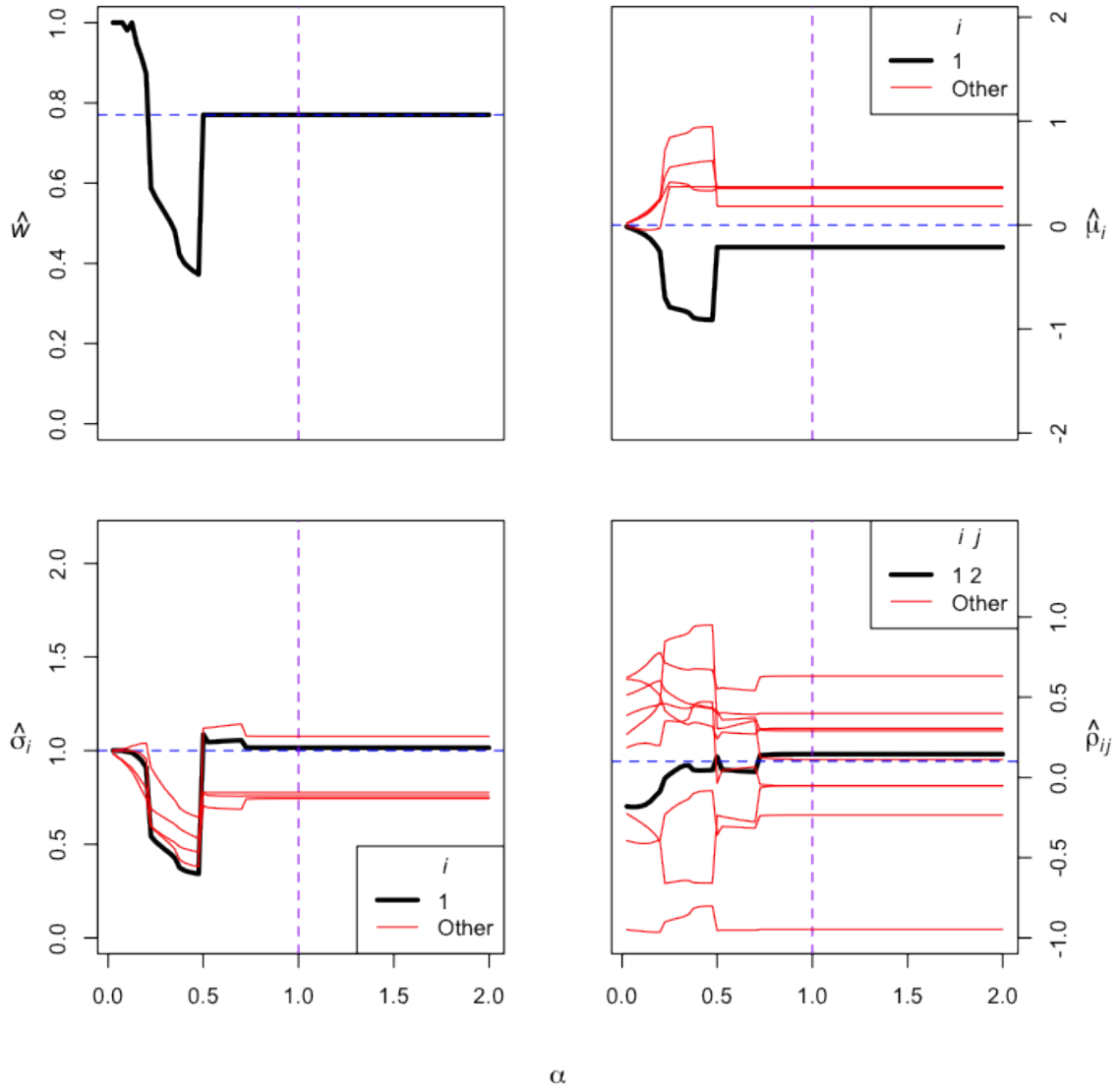


Figure 3.23: Trace plots of MPDC- $\alpha$  estimates for distribution of `pfx_x`, `pfx_z`, `pitch_speed`, `break_angle`, and `spin_dir` for fastballs for  $\alpha$  ranging from 0 to 2 in increments of 0.025. The black lines represent parameters of interest, and the red lines show other parameters to track algorithm stability.

In order to apply the MPDC- $\alpha$  estimator to this example, we standardize the

values of `px_x`, `px_z`, `start_speed`, `break_angle`, and `spin_rate`. Thus, as in the  $p = 2$  case, the MLE solution is inconsistent and insufficiently robust. We apply the same transformation in (3.8) to the sample data points.

Upon applying the transformation, we seek to obtain  $\hat{w}_\alpha$ ,  $\hat{\boldsymbol{\mu}}_\alpha$  and  $\hat{\boldsymbol{\Sigma}}_\alpha$ . Thus, we compute the MPDC- $\alpha$  divergence estimator for  $\alpha$  values in the range  $[0,2]$ .

We can see that  $\hat{w}_\alpha$  is 1 for smaller  $\alpha$  values, decreases to about 0.4 at  $\alpha$  around 0.5, and then gets close to the true value of  $w$  at an  $\alpha$  value of about 0.6. Transforming back to the original units, the other parameter estimates reach their corresponding true values for an  $\alpha$  value around 1. The best solution occurs at  $\alpha = 1.2$ :  $\hat{w}_{(\alpha=1.2)} = 0.77$ ,  $\hat{\boldsymbol{\mu}}_{(\alpha=1.2)} = (-2.7, 6.9, 89, 10.6, 1960)'$ ,  $\hat{\rho}_{(\alpha=1.2),12} = 0.14$ . The  $\alpha$  value of 1.2 closely matches the suggested value from the exploratory  $\alpha$  selection ranges we established in Section 3.3.6. There is considerable overlap between the two components, so we would expect that an  $\alpha$  value around 1 or greater would be needed.



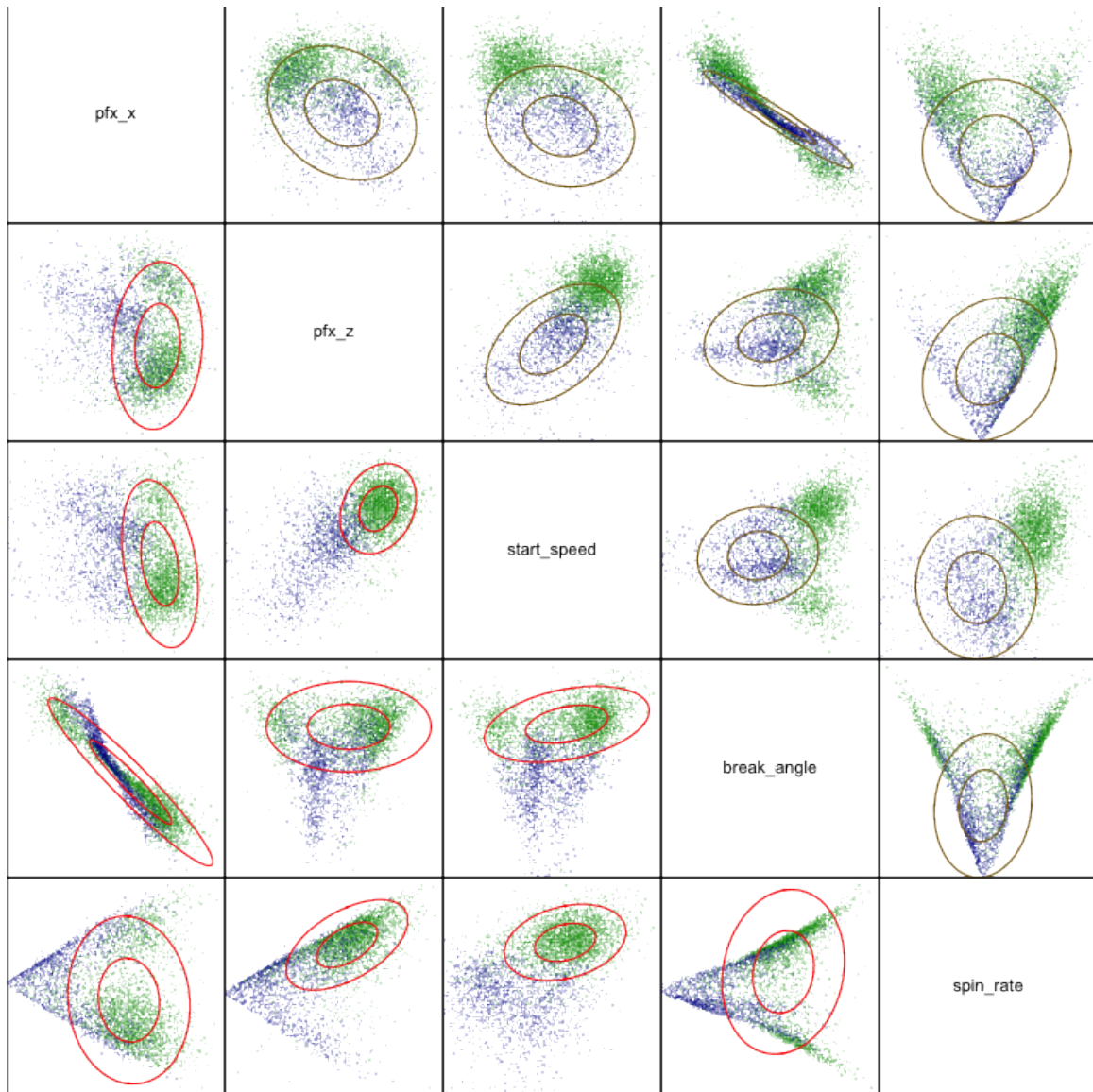


Figure 3.24: Pairwise scatterplots for  $\text{pfx}_x$ ,  $\text{pfx}_z$ ,  $\text{pitch\_speed}$ ,  $\text{break\_angle}$ ,  $\text{spin\_dir}$  for a random sample of 5,000 pitches, with fastballs in green and breaking balls in blue. 2 MPDC- $\alpha$  solutions are shown. Above the diagonal, the solutions for the breaking ball group are shown by the gold 1 and 2-sigma ellipses. Below the diagonal, the solutions for the fastball group are shown by the red 1 and 2-sigma ellipses.  $\hat{w}_{\text{fastball},\alpha} = 0.77$ ;  $\hat{w}_{\text{breaking},\alpha} = 0.41$ .

When considering these 3 additional variables, we see that fastballs and breaking balls tend to have the following characteristics, on average:

Variable	Fastball	Breaking Ball
Horiz. Movement	2.6 in. left	0.7 in. right
Vert. Movement	7.4 in. up	1.1 in. up
Speed	91 mph	81 mph
Break Angle	11 degrees	-1.5 degrees
Spin Rate	2040 rpm	1215 rpm

Table 3.4: Comparison of fastballs and breaking balls with respect to five PITCHf/x characteristics.

# Chapter 4

## Asymptotics

We wish to examine the behavior of our estimator,  $\hat{\boldsymbol{\theta}}_\alpha$ , as  $n \rightarrow \infty$ . Basu et. al. (1998) provided asymptotic results for  $\hat{\boldsymbol{\mu}}_\alpha$  as well as the asymptotic distribution of  $\hat{\sigma}_\alpha^2$ . We will verify these results and extend them to the case of  $\boldsymbol{\Sigma} = \sigma^2 I_p$  ( $\sigma$  unknown) and unknown  $\boldsymbol{\Sigma}$  for a general dimension  $p$ . The asymptotic distribution of the weight parameter,  $\hat{w}_\alpha$ , was computed in Section 2.3.1.

### 4.1 Motivation: $p = 1$

We begin by motivating our asymptotic distribution computations with the one-dimensional cases. These distributions are given in Basu (1998), but the details of their derivations are verified here.

### 4.1.1 Asymptotic Distribution of $\hat{\mu}_\alpha$ , Known $\sigma$

**Lemma 4.1.** *Let  $X_1, X_2, \dots, X_n$  be iid  $wN(\mu, \sigma^2) + (1-w)F^*$ , and let  $\hat{\mu}_\alpha$  be the mean component of the MPDC- $\alpha$  divergence estimator,  $\hat{\boldsymbol{\theta}}_\alpha \equiv (\hat{w}_\alpha, \hat{\mu}_\alpha)$ . Then as  $n \rightarrow \infty$ :*

$$\sqrt{n}(\hat{\mu}_\alpha - \mu) \rightarrow N\left(0, \left(1 + \frac{\alpha^2}{1 + 2\alpha}\right)^{3/2} \frac{\sigma^2}{w}\right). \quad (4.1)$$

**Proof.**

$$\text{We define the model: } f_{\boldsymbol{\theta}} = \frac{w}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

$$\text{Taking the log: } \log f_{\boldsymbol{\theta}} = \log(w) - \frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}(x - \mu)^2.$$

$$\text{The score function is: } u_{\boldsymbol{\theta}} \equiv \frac{\partial \log f_{\boldsymbol{\theta}}}{\partial \mu} = \frac{1}{\sigma^2}(x - \mu).$$

$$\begin{aligned} \text{Thus, } \xi &= \int u_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}^{1+\alpha} \\ &= \int \frac{x - \mu}{\sigma^2} \frac{w^{1+\alpha}}{(2\pi)^{\frac{1+\alpha}{2}} \sigma^{1+\alpha}} e^{-\frac{1+\alpha}{2\sigma^2}(x-\mu)^2} \\ &= 0. \end{aligned}$$

$$\begin{aligned} \text{Also, } J &= \int u_{\boldsymbol{\theta}}^2 f_{\boldsymbol{\theta}}^{1+\alpha} \\ &= \int \frac{(x - \mu)^2}{\sigma^4} \frac{w^{1+\alpha}}{(2\pi)^{\frac{1+\alpha}{2}} \sigma^{1+\alpha}} e^{-\frac{1+\alpha}{2\sigma^2}(x-\mu)^2} \\ &= w^{1+\alpha} (2\pi)^{-\frac{\alpha}{2}} \sigma^{-(2+\alpha)} (1 + \alpha)^{-\frac{3}{2}}. \end{aligned}$$

$$\begin{aligned} \text{Then, } K &= \int u_{\boldsymbol{\theta}}^2 f_{\boldsymbol{\theta}}^{1+2\alpha} \\ &= \int \frac{(x - \mu)^2}{\sigma^4} \frac{w^{1+2\alpha}}{(2\pi)^{\frac{1+2\alpha}{2}} \sigma^{1+2\alpha}} e^{-\frac{1+2\alpha}{2\sigma^2}(x-\mu)^2} \\ &= w^{1+2\alpha} (2\pi)^{-\alpha} \sigma^{-2(1+\alpha)} (1 + 2\alpha)^{-\frac{3}{2}}. \end{aligned}$$

$$\begin{aligned}
\text{Therefore, } \frac{K}{J^2} &= \frac{(1+\alpha)^3 \sigma^2}{(1+2\alpha)^{\frac{3}{2}} w} \\
&= \left[ \frac{(1+\alpha)^2}{(1+2\alpha)^{\frac{3}{2}}} \right]^{\frac{3}{2}} \frac{\sigma^2}{w} \\
&= \left( 1 + \frac{\alpha^2}{1+2\alpha} \right)^{\frac{3}{2}} \frac{\sigma^2}{w}.
\end{aligned}$$

$$\text{Thus, } \sqrt{n}(\hat{\mu}_\alpha - \mu) \rightarrow N \left( 0, \left( 1 + \frac{\alpha^2}{1+2\alpha} \right)^{3/2} \frac{\sigma^2}{w} \right).$$

□

### 4.1.2 Asymptotic Distribution of $\hat{\sigma}_\alpha$ , Unknown $\sigma$

**Lemma 4.2.** *Let  $X_1, X_2, \dots, X_n$  be iid  $wN(\mu, \sigma^2) + (1-w)F^*$ , and let  $\hat{\sigma}_\alpha$  be the standard deviation component of the MPDC- $\alpha$  divergence estimator,  $\hat{\theta}_\alpha \equiv (\hat{w}_\alpha, \hat{\mu}_\alpha, \hat{\sigma}_\alpha)$ .*

*Then as  $n \rightarrow \infty$ :*

$$\sqrt{n}(\hat{\sigma}_\alpha - \sigma) \rightarrow N \left( 0, \frac{(1+\alpha)^5}{(2+\alpha^2)^2} \left[ \frac{2(1+2\alpha^2)}{(1+2\alpha)^{5/2}} - \frac{\alpha^2}{(1+\alpha)^3} \right] \frac{\sigma^2}{w} \right). \quad (4.2)$$

**Proof.**

$$\begin{aligned}
\text{The score function is: } u_\theta &= \left( \frac{\partial \log f}{\partial \mu}, \frac{\partial \log f}{\partial \sigma} \right)^T \\
&= \left( \frac{x - \mu}{\sigma^2}, \frac{(x - \mu)^2 - \sigma^2}{\sigma^3} \right)^T.
\end{aligned}$$

$$\text{Then, } \xi = \int u_\theta f_\theta^{1+\alpha} = \left( 0, \frac{w^{1+\alpha} (2\pi)^{-\alpha/2} \alpha \sigma^{-1-\alpha}}{(1+\alpha)^{3/2}} \right)^T.$$

$$\text{Next, } J = \int u_\theta u_\theta^T f_\theta^{1+\alpha} = \begin{pmatrix} \frac{w^{1+\alpha} (2\pi)^{-\alpha/2} \sigma^{-(2+\alpha)}}{(1+\alpha)^{3/2}} & 0 \\ 0 & \frac{w^{1+\alpha} (2\pi)^{-\alpha/2} (2+\alpha^2) \sigma^{-2-\alpha}}{(1+\alpha)^{5/2}} \end{pmatrix}.$$

$$\begin{aligned}
\text{Also, } K &= \int u_\theta u_\theta^T f_\theta^{1+2\alpha} - \xi \xi^T \\
&= \begin{pmatrix} \frac{w^{1+2\alpha}(2\pi)^{-\alpha} \sigma^{-2(1+\alpha)}}{(1+2\alpha)^{3/2}} & 0 \\ 0 & \frac{w^{1+2\alpha} 2^{1-\alpha} \pi^{-\alpha} (1+2\alpha^2) \sigma^{-2(1+\alpha)}}{(1+2\alpha)^{5/2}} \end{pmatrix} - \\
&\quad \begin{pmatrix} 0 & 0 \\ 0 & \frac{w^{2+2\alpha} (2\pi)^{-\alpha} \alpha^2 \sigma^{-2(1+\alpha)}}{(1+\alpha)^3} \end{pmatrix}.
\end{aligned}$$

Therefore,

$$J^{-1} K J^{-1} = \begin{pmatrix} \frac{(1+\alpha)^3 \sigma^{4+2\alpha-2(1+\alpha)}}{w(1+2\alpha)^{3/2}} & 0 \\ 0 & \frac{(2\pi)^\alpha (1+\alpha)^5 \sigma^{4+2\alpha} \left( \frac{2^{1-\alpha} \pi^{-\alpha} (1+2\alpha^2) \sigma^{-2(1+\alpha)}}{(1+2\alpha)^{5/2}} - \frac{(2\pi)^{-\alpha} \alpha^2 \sigma^{-2\alpha}}{(1+\alpha)^3} \right)}{w(2+\alpha^2)^2} \end{pmatrix}.$$

Thus,

$$\sqrt{n}(\hat{\sigma}_\alpha - \sigma) \rightarrow N \left( 0, \frac{(1+\alpha)^5}{(2+\alpha^2)^2} \left[ \frac{2(1+2\alpha^2)}{(1+2\alpha)^{5/2}} - \frac{\alpha^2}{(1+\alpha)^3} \right] \frac{\sigma^2}{w} \right).$$

□

## 4.2 General $p$

We derive the asymptotic distribution of the MPDC- $\alpha$  estimator  $\hat{\boldsymbol{\theta}}_\alpha \equiv (\hat{w}_\alpha, \hat{\boldsymbol{\mu}}_\alpha, \hat{\boldsymbol{\Sigma}}_\alpha)$  for a general dimension  $p$ .

### 4.2.1 Asymptotic Distribution of $\hat{\boldsymbol{\mu}}_\alpha$ , Known $\boldsymbol{\Sigma}$

**Lemma 4.3.** *Let  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  be iid  $wN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + (1-w)F^*$ , and let  $\hat{\boldsymbol{\mu}}_\alpha$  be the mean component of the MPDC- $\alpha$  divergence estimator,  $\hat{\boldsymbol{\theta}}_\alpha \equiv (\hat{w}_\alpha, \hat{\boldsymbol{\mu}}_\alpha, \hat{\boldsymbol{\Sigma}}_\alpha)$ . Then as  $n \rightarrow \infty$ :*

$$\sqrt{n}(\hat{\boldsymbol{\mu}}_\alpha - \boldsymbol{\mu}) \rightarrow N \left( \mathbf{0}_p, \frac{1}{w} \left( 1 + \frac{\alpha^2}{1+2\alpha} \right)^{p/2+1} \boldsymbol{\Sigma} \right). \quad (4.3)$$

**Proof.**

We observe that:  $\xi = \int u_\theta f_\theta^{1+\alpha} = 0$ .

Integrating by parts, we get that:

$$\begin{aligned} J &= \int u_\theta u_\theta^T f_\theta^{1+\alpha} = \int \frac{w^{1+\alpha} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}}{(2\pi)^{p(1+\alpha)/2} |\boldsymbol{\Sigma}|^{(1+\alpha)/2}} e^{-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \left(\frac{\boldsymbol{\Sigma}}{1+\alpha}\right)^{-1} (\mathbf{x} - \boldsymbol{\mu})} \\ &= \frac{w^{1+\alpha} \boldsymbol{\Sigma}^{-1} (1 + \alpha)^{-p/2}}{(2\pi)^{p\alpha/2} |\boldsymbol{\Sigma}|^{\alpha/2} (1 + \alpha)} \int \frac{e^{-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \left(\frac{\boldsymbol{\Sigma}}{1+\alpha}\right)^{-1} (\mathbf{x} - \boldsymbol{\mu})}}{(2\pi)^{p/2} \left|\frac{\boldsymbol{\Sigma}}{1+\alpha}\right|^{1/2}} \\ &= w^{1+\alpha} (2\pi)^{-p\alpha/2} |\boldsymbol{\Sigma}|^{-\alpha/2} (1 + \alpha)^{-p/2-1} \boldsymbol{\Sigma}^{-1}. \end{aligned}$$

$$\begin{aligned} \text{Similarly, } K &= \int u_\theta u_\theta^T f_\theta^{1+2\alpha} = w^{1+2\alpha} (2\pi)^{-p(2\alpha)/2} |\boldsymbol{\Sigma}|^{-2\alpha/2} (1 + 2\alpha)^{-p/2-1} \boldsymbol{\Sigma}^{-1} \\ &= w^{1+2\alpha} (2\pi)^{-p\alpha} |\boldsymbol{\Sigma}|^{-\alpha} (1 + 2\alpha)^{-p/2-1} \boldsymbol{\Sigma}^{-1}. \end{aligned}$$

$$\begin{aligned} \text{Then, } J^{-1} K J^{-1} &= \frac{1}{w} (1 + \alpha)^{p+2} (1 + 2\alpha)^{-p/2-1} \boldsymbol{\Sigma} \\ &= \frac{1}{w} \left( \frac{1 + \alpha}{\sqrt{1 + 2\alpha}} \right)^{p+2} \boldsymbol{\Sigma} \\ &= \frac{1}{w} \left[ \frac{(1 + \alpha)^2}{1 + 2\alpha} \right]^{p/2+1} \boldsymbol{\Sigma} \\ &= \frac{1}{w} \left( 1 + \frac{\alpha^2}{1 + 2\alpha} \right)^{p/2+1} \boldsymbol{\Sigma}. \end{aligned}$$

Thus,

$$\sqrt{n}(\hat{\boldsymbol{\mu}}_\alpha - \boldsymbol{\mu}) \rightarrow N \left( \mathbf{0}_p, \frac{1}{w} \left( 1 + \frac{\alpha^2}{1 + 2\alpha} \right)^{p/2+1} \boldsymbol{\Sigma} \right).$$

□

#### 4.2.2 Asymptotic Distribution of $\hat{\boldsymbol{\Sigma}}_\alpha$ , $\boldsymbol{\Sigma} = \sigma^2 I_p$ , $\sigma$ Unknown

**Lemma 4.4.** *Let  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  be iid  $wN(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \sigma^2 I_p) + (1 - w)F^*$ , and let  $\hat{\boldsymbol{\Sigma}}_\alpha$  be the covariance component of the MPDC- $\alpha$  divergence estimator,  $\hat{\boldsymbol{\theta}}_\alpha \equiv (\hat{w}_\alpha, \hat{\boldsymbol{\mu}}_\alpha, \hat{\boldsymbol{\Sigma}}_\alpha)$ .*

Then as  $n \rightarrow \infty$ :

$$\begin{aligned} \sqrt{n}(\hat{\sigma}_\alpha - \sigma)\mathbf{1}_p &\rightarrow N\left(\mathbf{0}_p, \frac{(1+\alpha)^{p+4}}{[p^2(1+\alpha)^2 - 2p(1+\alpha) + 3]^2} \left\{ p^2[(1+2\alpha)^{-p/2} - \right. \right. \\ &\quad \left. \left. (1+\alpha)^{-p}] - 2p[(1+2\alpha)^{-p/2-1} - (1+\alpha)^{-p-1}] + \right. \right. \\ &\quad \left. \left. [3(1+2\alpha)^{-p/2-2} - (1+\alpha)^{-p-2}] \right\} \frac{\sigma^2}{w} I_p \right). \end{aligned} \quad (4.4)$$

**Proof.**

$$\text{The model: } f_\theta = wN(\boldsymbol{\mu}, \sigma^2 I_p) = w(2\pi)^{p/2} \sigma^{-p} e^{\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})}.$$

$$\text{Taking the log: } \log f_\theta = \log(w) - p \log \sigma - \frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu}).$$

$$\text{The score function: } u_\theta = \frac{\partial \log f}{\partial \sigma} = \frac{(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu}) - p\sigma^2}{\sigma^3}.$$

$$\begin{aligned} \text{Then, } \xi &= \int u_\theta f_\theta^{1+\alpha} = \int \left[ \frac{(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu}) - p\sigma^2}{\sigma^3} \right] \frac{w^{1+\alpha} e^{\frac{1+\alpha}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})}}{(2\pi)^{p(1+\alpha)/2} \sigma^{p(1+\alpha)}} \\ &= \int \frac{(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu}) w^{1+\alpha} e^{\frac{1+\alpha}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})}}{(2\pi)^{p(1+\alpha)/2} \sigma^{p(1+\alpha)+3}} - \\ &\quad \int \frac{p w^{1+\alpha} e^{\frac{1+\alpha}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})}}{(2\pi)^{p(1+\alpha)/2} \sigma^{p(1+\alpha)+1}} \\ &= w^{1+\alpha} (2\pi)^{-p\alpha/2} \sigma^{-p\alpha-1} [(1+\alpha)^{-p/2-1} - p(1+\alpha)^{-p/2}] \\ &= w^{1+\alpha} (2\pi)^{-p\alpha/2} \sigma^{-p\alpha-1} (1+\alpha)^{-p/2} \left[ \frac{1}{1+\alpha} - p \right] \\ &= w^{1+\alpha} (2\pi)^{-p\alpha/2} \sigma^{-p\alpha-1} (1+\alpha)^{-p/2-1} [1 - p(1+\alpha)]. \end{aligned}$$



$$\begin{aligned}
\text{Next, } J &= \int u_\theta u_\theta^T f_\theta^{1+\alpha} \\
&= \int \left\{ \left[ \frac{(\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu}) - p\sigma^2}{\sigma^3} \right] \left[ \frac{(\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu}) - p\sigma^2}{\sigma^3} \right] \times \right. \\
&\quad \left. \frac{w^{1+\alpha} e^{\frac{1+\alpha}{2\sigma^2} (\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu})}}{(2\pi)^{p(1+\alpha)/2} \sigma^{p(1+\alpha)}} \right\} \\
&= \int \frac{w^{1+\alpha} (\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu}) (\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu}) - 2p\sigma^2 (\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu}) + p^2 \sigma^4}{\sigma^6} \\
&= w^{1+\alpha} [3(1+\alpha)^{-p/2-2} (2\pi)^{-p\alpha/2} \sigma^{-p\alpha-2} - 2p(1+\alpha)^{-p/2-1} (2\pi)^{-p\alpha/2} \sigma^{-p\alpha-2} + \\
&\quad p^2 (1+\alpha)^{-p/2} (2\pi)^{-p\alpha/2} \sigma^{-p\alpha-2}] \\
&= \frac{(1+\alpha)^2 - 2(1+\alpha) + 3}{(1+\alpha)^2} w^{1+\alpha} (2\pi)^{-p\alpha/2} \sigma^{-p\alpha-2} (1+\alpha)^{-p/2} \times \\
&\quad [p^2 - 2p(1+\alpha)^{-1} + 3(1+\alpha)^{-2}] \\
&= w^{1+\alpha} (2\pi)^{-p\alpha/2} \sigma^{-p\alpha-2} (1+\alpha)^{-p/2-2} [p^2 (1+\alpha)^2 - 2p(1+\alpha) + 3].
\end{aligned}$$

$$\begin{aligned}
\text{Similarly, } K &= \int u_\theta u_\theta^T f_\theta^{1+2\alpha} - \xi \xi^T \\
&= w^{1+2\alpha} (2\pi)^{-p\alpha} \sigma^{-2(p\alpha+1)} (1+2\alpha)^{-p/2-2} [p^2 (1+2\alpha)^2 - 2p(1+2\alpha) + 3] \\
&\quad - w^{1+2\alpha} (2\pi)^{-p\alpha} \sigma^{-2(p\alpha+1)} (1+\alpha)^{-p-2} [1 - p(1+\alpha)]^2 \\
&= w^{1+2\alpha} (2\pi)^{-p\alpha} \sigma^{-2(p\alpha+1)} \{ p^2 [-(1+\alpha)^{-p} + (1+2\alpha)^{-p/2}] - \\
&\quad 2p [-(1+\alpha)^{-p-1} + (1+2\alpha)^{-p/2-1}] + \\
&\quad [-(1+\alpha)^{-p-2} + 3(1+2\alpha)^{-p/2-2}] \}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
J^{-1} K J^{-1} &= \frac{(1+\alpha)^{p+4}}{[p^2 (1+\alpha)^2 - 2p(1+\alpha) + 3]^2} \{ p^2 [(1+2\alpha)^{-p/2} - (1+\alpha)^{-p}] - \\
&\quad 2p [(1+2\alpha)^{-p/2-1} - (1+\alpha)^{-p-1}] + \\
&\quad [3(1+2\alpha)^{-p/2-2} - (1+\alpha)^{-p-2}] \} \frac{\sigma^2}{w} I_p.
\end{aligned}$$

Thus,

$$\sqrt{n}(\hat{\sigma}_\alpha - \sigma)\mathbf{1}_p \rightarrow N\left(\mathbf{0}_p, \frac{(1+\alpha)^{p+4}}{[p^2(1+\alpha)^2 - 2p(1+\alpha) + 3]^2} \left\{ p^2[(1+2\alpha)^{-p/2} - (1+\alpha)^{-p}] - 2p[(1+2\alpha)^{-p/2-1} - (1+\alpha)^{-p-1}] + [3(1+2\alpha)^{-p/2-2} - (1+\alpha)^{-p-2}] \right\} \frac{\sigma^2}{w} I_p \right).$$

□

### 4.2.3 Asymptotic Distribution of $(\hat{\sigma}_{1,\alpha}, \dots, \hat{\sigma}_{p,\alpha})'$ , Unknown $\Sigma$

**Lemma 4.5.** *Let  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  be iid  $wN(\boldsymbol{\mu}, \Sigma) + (1-w)F^*$ . Let  $(\hat{\sigma}_{1,\alpha}, \dots, \hat{\sigma}_{p,\alpha})'$  be the vector of standard deviations from  $\hat{\Sigma}_\alpha$  and  $(\sigma_1, \dots, \sigma_p)'$  the vector of standard deviations from  $\Sigma$ . As  $n \rightarrow \infty$ :  $\sqrt{n}((\hat{\sigma}_{1,\alpha}, \dots, \hat{\sigma}_{p,\alpha})' - (\sigma_1, \dots, \sigma_p)') \rightarrow$*

$$N\left(\mathbf{0}_p, \frac{1}{w} \frac{(1+\alpha)^{p+4}}{[p^2(1+\alpha)^2 - 2p(1+\alpha) + 3]^2} \left\{ p^2[(1+2\alpha)^{-p/2} - (1+\alpha)^{-p}] - 2p[(1+2\alpha)^{-p/2-1} - (1+\alpha)^{-p-1}] + [3(1+2\alpha)^{-p/2-2} - (1+\alpha)^{-p-2}] \right\} \Sigma \right).$$

**Proof.** We generalize Lemma 4.4. □

### 4.2.4 Asymptotic Distribution of $\hat{\rho}_\alpha$ , Unknown $\rho$

**Lemma 4.6.** *Let  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  be iid  $wN(\mathbf{0}_p, \Sigma) + (1-w)F^*$ . Let  $\hat{\rho}_\alpha$  be an arbitrary correlation from  $\hat{\Sigma}_\alpha$  and  $\rho$  the corresponding correlation from  $\Sigma$ . As  $n \rightarrow \infty$ :*

$$\sqrt{n}(\hat{\rho}_\alpha - \rho) \rightarrow^d N\left(0, \frac{(1+\alpha)^p}{(1+2\alpha)^{p/2}} \frac{(1-\rho^2)^2}{w}\right)$$

**Proof.** Assume W.L.O.G.  $\Sigma_{jj} = 1 \quad \forall j \in [1, p]$  and  $\rho$  is the only non-zero correlation.

$$\text{We define the model: } f_\theta = w\phi(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}}{(2\pi)^{p/2} |\Sigma|^{1/2}}.$$

Then we obtain an estimate of the divergence function by plugging the model into

(1.1):

$$\widehat{d}_\alpha = \int \frac{w^{1+\alpha} e^{-\frac{1+\alpha}{2} \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}}}{(2\pi)^{p(1+\alpha)/2} (1-\rho^2)^{\frac{1+\alpha}{2}}} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^n \left[ \frac{w^\alpha e^{-\frac{\alpha}{2} \mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_i}}{(2\pi)^{p\alpha/2} (1-\rho^2)^{\frac{\alpha}{2}}} \right].$$

$$\text{Letting } \gamma_1 \equiv \int \frac{w^{1+\alpha} e^{-\frac{1+\alpha}{2} \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}}}{(2\pi)^{p(1+\alpha)/2} (1-\rho^2)^{\frac{1+\alpha}{2}}} = w^{1+\alpha} (2\pi)^{-p\alpha/2} (1+\alpha)^{p/2} (1-\rho^2)^{-\alpha/2}$$

$$\text{and } \gamma_2 \equiv \frac{w^\alpha e^{-\frac{\alpha}{2} \mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_i}}{(2\pi)^{p\alpha/2} (1-\rho^2)^{\frac{\alpha}{2}}},$$

$$\text{we get } E[\gamma_2] = \int \frac{w^{1+2\alpha} e^{-\frac{1+\alpha}{2} \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}}}{(2\pi)^{p(1+\alpha)/2} (1-\rho^2)^{\frac{1+\alpha}{2}}} = \gamma_1.$$

$$\begin{aligned} \text{Thus, } \widehat{d}_\alpha &= \gamma_1 - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^n [(\gamma_2 - \gamma_1) + \gamma_1] \\ &= -\frac{1}{\alpha} \gamma_1 - \frac{(1 + \frac{1}{\alpha})}{\sqrt{n}} Z \sqrt{\Sigma(\theta)}, \quad \text{where } \Sigma(\theta) \equiv \text{Var}(\gamma_2). \end{aligned}$$

To find  $\hat{\rho}$  we differentiate with respect to  $\rho$ :

$$\frac{\partial \widehat{d}_\alpha}{\partial \rho} = -\gamma_1 \frac{\rho}{1-\rho^2} + 2 \frac{1+\alpha}{\sqrt{n}} Z \sqrt{\Sigma(\theta)} \frac{\rho}{1-\rho^2} = 0.$$

$$\begin{aligned} \text{Therefore, } \sqrt{n}(\hat{\rho} - \rho) &\rightarrow^d N\left(0, \left[1 - \frac{\Sigma(\theta)}{\gamma_1^2}\right] (1-\rho^2)^2\right) \quad \text{as } n \rightarrow \infty \\ &\rightarrow^d N\left(0, \frac{(1+\alpha)^p}{(1+2\alpha)^{p/2}} \frac{(1-\rho^2)^2}{w}\right). \end{aligned}$$

□

### 4.3 Simulation to Verify Asymptotic Results

To verify our asymptotic results, we simulate  $M = 1000$  samples of size  $n = 1000$

from the mixture distribution

$$\frac{3}{4} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_2\right) + \frac{1}{4} N\left(\begin{pmatrix} 7 \\ 0 \end{pmatrix}, I_2\right).$$

This corresponds to the two-dimensional case of well-separated clusters with zero correlation from Section 3.3.1. Because we found an  $\alpha$  value of 0.5 to be in the optimal range for that case, we will fix  $\alpha$  to be 0.5 for this simulation.

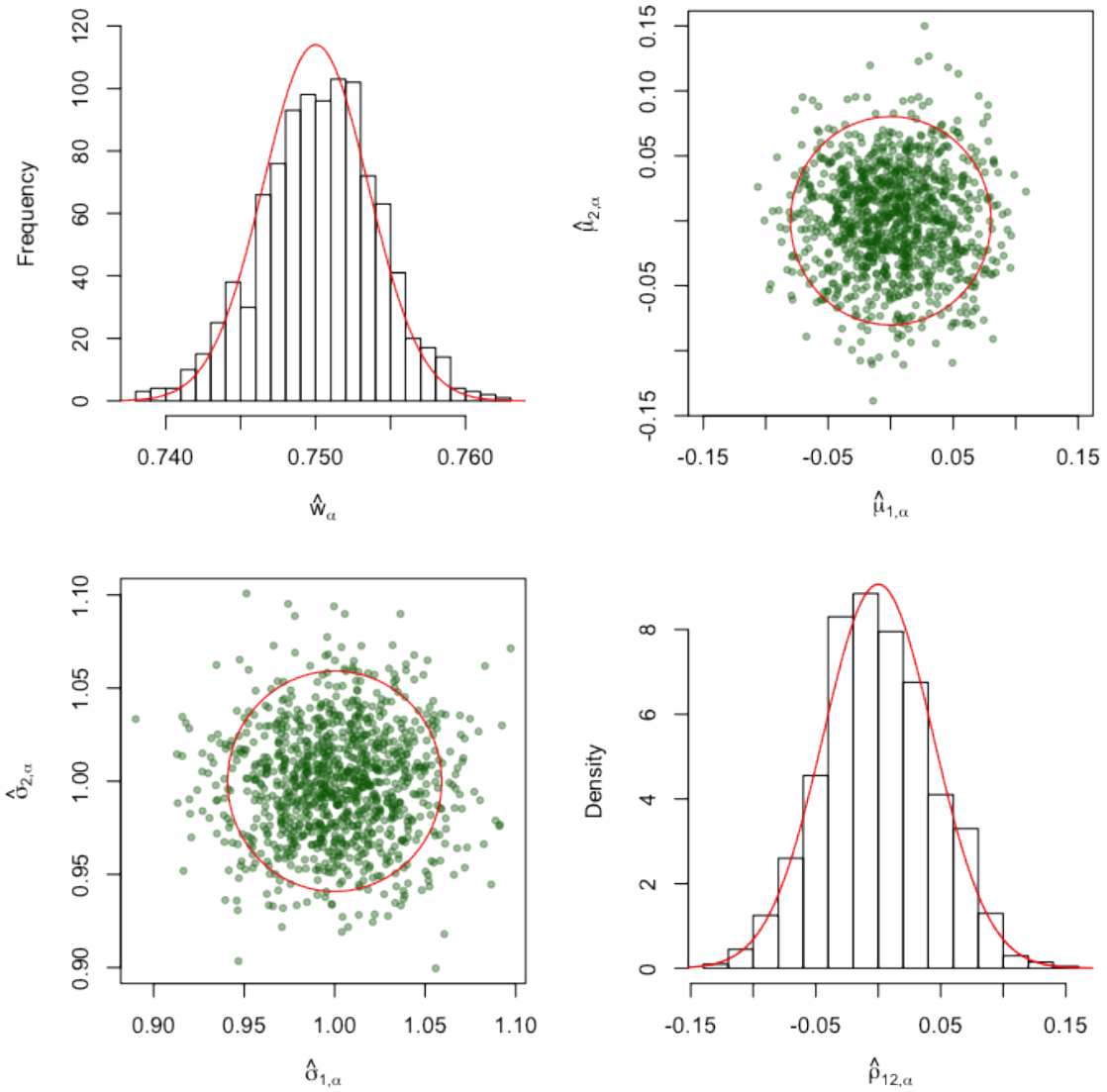


Figure 4.1: Sampling Distributions for Components of  $\hat{\theta}_\alpha$ . Obtained via 1000 simulations of samples of size  $n = 1000$  from a bivariate Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0)'$ ;  $\mu_2 = (7, 0)'$ ;  $\Sigma_1 = \Sigma_2 = I_2$ . Theoretical asymptotic densities are shown in red.

Figure 4.1 shows the resulting sampling distributions for  $\hat{w}_\alpha$ ,  $\hat{\boldsymbol{\mu}}_\alpha$ ,  $\sqrt{\hat{\boldsymbol{\Sigma}}_{jj,\alpha}}$  (where

$j = 1$  or  $2$ ), and  $\hat{\rho}_{12,\alpha}$ . The estimated sampling distributions of  $\hat{w}_\alpha$  (top left) and  $\hat{\rho}_{12,\alpha}$  (bottom right) are approximately Normal, and the estimated bivariate sampling distributions of  $\hat{\boldsymbol{\mu}}_\alpha$  (top right) and  $\hat{\sigma}_{j,\alpha} = \sqrt{\hat{\Sigma}_{jj,\alpha}}$  (bottom left) are approximately bivariate Normal. We check that the parameters we have derived for these asymptotic distributions are accurate.

### 4.3.1 Verification of Asymptotic Distribution of $\hat{w}_\alpha$

Recall from Section 2.3.1 that the asymptotic distribution of  $\hat{w}_\alpha$  is

$$N \left( w, \frac{\frac{1}{w} \left[ \frac{(1+\alpha)^p}{(1+2\alpha)^{p/2}} - 1 \right]}{n} \right).$$

Thus, for  $p = 2$ ,  $n = 1000$  and  $\alpha = 0.5$ ,  $\hat{w}_\alpha$  is asymptotically Normal with mean  $w = 0.75$  and variance  $\frac{\frac{1}{w} \left[ \frac{(1+\alpha)^p}{(1+2\alpha)^{p/2}} - 1 \right]}{n} = \frac{(1.5)^2 - 1}{\frac{2^1}{750}} = 1.667 \times 10^{-5}$ . From our simulation, we yielded  $\hat{E}[\hat{w}_\alpha] = 0.7502$  and  $\widehat{Var}[\hat{w}_\alpha] = 1.551 \times 10^{-5}$ , which are close to the theoretical results.

### 4.3.2 Verification of Asymptotic Distribution of $\hat{\boldsymbol{\mu}}_\alpha$

Recall from Section 4.2.1 that the asymptotic distribution of  $\hat{\boldsymbol{\mu}}_\alpha$  is

$$N \left( \boldsymbol{\mu}, \frac{\frac{1}{w} \left( 1 + \frac{\alpha^2}{1+2\alpha} \right)^{p/2+1}}{n} \boldsymbol{\Sigma} \right).$$

Thus, for  $p = 2$ ,  $n = 1000$  and  $\alpha = 0.5$ ,  $\hat{\boldsymbol{\mu}}_\alpha$  is asymptotically bivariate Normal with mean  $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$  and covariance matrix  $\frac{\frac{1}{w} \left( 1 + \frac{\alpha^2}{1+2\alpha} \right)^{p/2+1}}{n} \boldsymbol{\Sigma} = \frac{\frac{1}{w} \left( 1 + \frac{(0.5)^2}{2} \right)^2}{750} I_2 = \begin{pmatrix} 0.001688 & 0.000000 \\ 0.000000 & 0.001688 \end{pmatrix}$ . From our simulation, we yielded  $\hat{E}[\hat{\boldsymbol{\mu}}_\alpha] = \begin{pmatrix} 0.000360 \\ -0.002038 \end{pmatrix}$  and  $\widehat{Cov}[\hat{\boldsymbol{\mu}}_\alpha] = \begin{pmatrix} 0.001692 & -0.000033 \\ -0.000033 & 0.001705 \end{pmatrix}$ , which are fairly close to the theoretical results.

### 4.3.3 Verification of Asymptotic Distribution of $\hat{\sigma}_{j,\alpha} = \sqrt{\hat{\Sigma}_{jj,\alpha}}$

Recall from Section 4.2.3 that the asymptotic distribution of  $\hat{\sigma}_{j,\alpha}$  has a mean of  $\sigma_j$  and variance

$$\frac{1}{w} \frac{(1+\alpha)^{p+4}}{[p^2(1+\alpha)^2 - 2p(1+\alpha) + 3]^2} \{p^2[(1+2\alpha)^{-p/2} - (1+\alpha)^{-p}] - 2p[(1+2\alpha)^{-p/2-1} - (1+\alpha)^{-p-1}] + [3(1+2\alpha)^{-p/2-2} - (1+\alpha)^{-p-2}]\} \frac{\sigma_j^2}{n}.$$

Thus, for  $p = 2$ ,  $n = 1000$  and  $\alpha = 0.5$ , for  $j \in \{1, 2\}$ ,  $\hat{\sigma}_{j,\alpha} = \sqrt{\hat{\Sigma}_{jj,\alpha}}$  is asymptotically Normal with mean  $\sigma_j = 1$  and variance = 0.004967. From our simulation, we yielded  $\hat{E}[\hat{\sigma}_{1,\alpha}] = 1.0028$ ,  $\hat{E}[\hat{\sigma}_{2,\alpha}] = 0.9933$ ,  $\widehat{Var}[\hat{\sigma}_{1,\alpha}] = 0.004398$ ,  $\widehat{Var}[\hat{\sigma}_{2,\alpha}] = 0.004239$ , which are close to the theoretical results.

### 4.3.4 Verification of Asymptotic Distribution of $\hat{\rho}_{12,\alpha}$

Using the asymptotic result from Section 4.2.4, we know that the asymptotic distribution of  $\hat{\rho}_{12,\alpha}$  is Normal with mean  $\rho_{12}$  and variance  $\frac{\frac{1}{w} \frac{(1+\alpha)^p}{(1+2\alpha)^{p/2}} (1-\rho_{12}^2)^2}{n}$ . Thus, for  $p = 2$ ,  $n = 1000$  and  $\alpha = 0.5$ ,  $\hat{\rho}_{12,\alpha}$  is asymptotically Normal with mean  $\rho_{12} = 0$  and variance = 0.001500. From our simulation, we yielded  $\hat{E}[\hat{\rho}_{12,\alpha}] = 0.0006$  and  $\widehat{Var}[\hat{\rho}_{12,\alpha}] = 0.001738$ , which are close to the theoretical results.

# Chapter 5

## Regression

Now that we have established the theoretical background and applicability of the MPDC- $\alpha$  divergence estimator to parametric estimation, we move on to consider the MPDC- $\alpha$  approach in a robust regression context.

As in the usual setting, we define a regression model of response variable  $y$  on the data matrix  $\mathbf{X}$ :

$$y = \boldsymbol{\beta}' \mathbf{X} + \epsilon \tag{5.1}$$

Scott (2001) extends the MPDC approach to regression. We assume the residuals,  $\epsilon_i$ , come from the model

$$f(\epsilon_i | \boldsymbol{\theta}) = w \phi(\epsilon_i | \boldsymbol{\theta}), \tag{5.2}$$

where  $\boldsymbol{\theta} = (w, \boldsymbol{\beta}, \sigma_\epsilon)$ .

We then apply the  $\alpha$ -divergence function to yield  $\hat{\boldsymbol{\theta}}_\alpha = (\hat{w}_\alpha, \hat{\boldsymbol{\beta}}_\alpha, \hat{\sigma}_{\epsilon,\alpha})$ .



## 5.1 Criterion Definition

Just as we did for parametric estimation, we construct the setting for our simulated regression cases, which will draw residuals from a two-component Normal mixture. It should be noted that the applicability of the MPDC- $\alpha$  divergence estimator is not limited to these cases. We can utilize  $\hat{\boldsymbol{\theta}}_\alpha$  for pure samples as well as samples with outliers, and the contamination can be particular points or a cluster of points. The framework for our study will be defined as follows:

Given an *iid* sample of residuals  $(\epsilon_1, \epsilon_2, \dots, \epsilon_n)'$  from the mixture

$$wN(0, \sigma_{\epsilon_1}^2) + (1 - w)N(0, \sigma_{\epsilon_2}^2)$$

and model

$$f(\epsilon_i | \boldsymbol{\theta}) = w\phi(\epsilon_i | \boldsymbol{\theta})$$

we estimate the parameter vector  $\boldsymbol{\theta} = (w, \boldsymbol{\beta}, \sigma_\epsilon)$  with  $\hat{\boldsymbol{\theta}}_\alpha$  by solving the following optimization problem for  $\alpha > 0$ :

$$\begin{aligned} \hat{\boldsymbol{\theta}}_\alpha &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[ \int f^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^n f(\epsilon_i)^\alpha \right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[ \int \frac{w^{1+\alpha} e^{-\frac{1+\alpha}{2\sigma_\epsilon^2} \epsilon_i^2}}{(\sqrt{2\pi}\sigma_\epsilon)^{1+\alpha}} - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n \phi(\epsilon_i)^\alpha \right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[ \frac{w^{1+\alpha}}{(\sqrt{2\pi}\sigma_\epsilon)^\alpha \sqrt{1+\alpha}} \int \frac{e^{-\frac{\epsilon_i^2}{2\left(\frac{\sigma_\epsilon}{\sqrt{1+\alpha}}\right)^2}}}{\sqrt{2\pi} \left(\frac{\sigma_\epsilon}{\sqrt{1+\alpha}}\right)} - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n \phi(\epsilon_i)^\alpha \right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[ \frac{w^{1+\alpha}}{(\sqrt{2\pi}\sigma_\epsilon)^\alpha \sqrt{1+\alpha}} - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n \frac{e^{-\frac{\alpha\epsilon_i^2}{2\sigma_\epsilon^2}}}{(\sqrt{2\pi}\sigma_\epsilon)^\alpha} \right]. \end{aligned}$$

Thus,

$$\hat{\theta}_\alpha = \begin{cases} \underset{\theta}{\operatorname{argmin}} \left[ \frac{w^{1+\alpha}}{(\sqrt{2\pi\sigma_\epsilon})^\alpha \sqrt{1+\alpha}} - \left(1 + \frac{1}{\alpha}\right) \frac{w^\alpha}{n} \sum_{i=1}^n \frac{e^{-\frac{\alpha\epsilon_i^2}{2\sigma_\epsilon^2}}}{(\sqrt{2\pi\sigma_\epsilon})^\alpha} \right] & \alpha > 0 \\ \hat{\theta}_{LS} & \alpha = 0 \end{cases} .$$

where  $\hat{\theta}_{LS}$  denotes the least-squares estimator.

## 5.2 Parameter Transformations

As we did in the case of parametric density estimation, we will apply transformations to the regression parameters for the purposes of the unconstrained optimization algorithm.

### 5.2.1 $w$ : *logit* transformation

We apply the same *logit* transformation as we did in the parametric density estimation setting to  $\tau(w)$ :

$$\tau(w) = \log\left(\frac{w}{1-w}\right), \quad (5.3)$$

so that when we reverse the transformation, we get:

$$w = \frac{1}{1 + e^{-\tau}}. \quad (5.4)$$

This yields the desired range of (0,1) for values of  $w$ .

### 5.2.2 $\sigma_\epsilon$ : *exp* transformation

In order to keep  $\sigma_\epsilon$  in the range of  $(0, \infty)$ , we exponentiate it:

$$\nu = \exp(\sigma_\epsilon) \tag{5.5}$$

We will optimize over  $\nu$ .

## 5.3 Simulated Cases

The regression settings we consider involve varying degrees of contamination and separation of the contamination from the main data. We will revisit some of our motivating examples from Section 1.1.2.

**Example 5.1:** We simulate a sample of size  $n = 100$ . Let  $x_1, x_2, \dots, x_{80} \sim iid U(-5, 5)$  and  $y_i = 2x_i - 5 + e_i$  for  $1 \leq i \leq 80$ , where  $e_1, e_2, \dots, e_{80} \sim iid N(0, 1)$ . Then, let  $x_{81}, x_{82}, \dots, x_{100} \sim iid U(-5, -2)$  and  $y_{81}, y_{82}, \dots, y_{100} \sim iid U(10, 20)$ .

Trace plots for  $\hat{w}_\alpha$ ,  $\hat{\beta}_\alpha$ , and  $\hat{\sigma}_{\epsilon, \alpha}$  can be found in Figure 5.1. The estimates converge to their true values around  $\alpha = 0.25$ , shown by the dashed red vertical line.

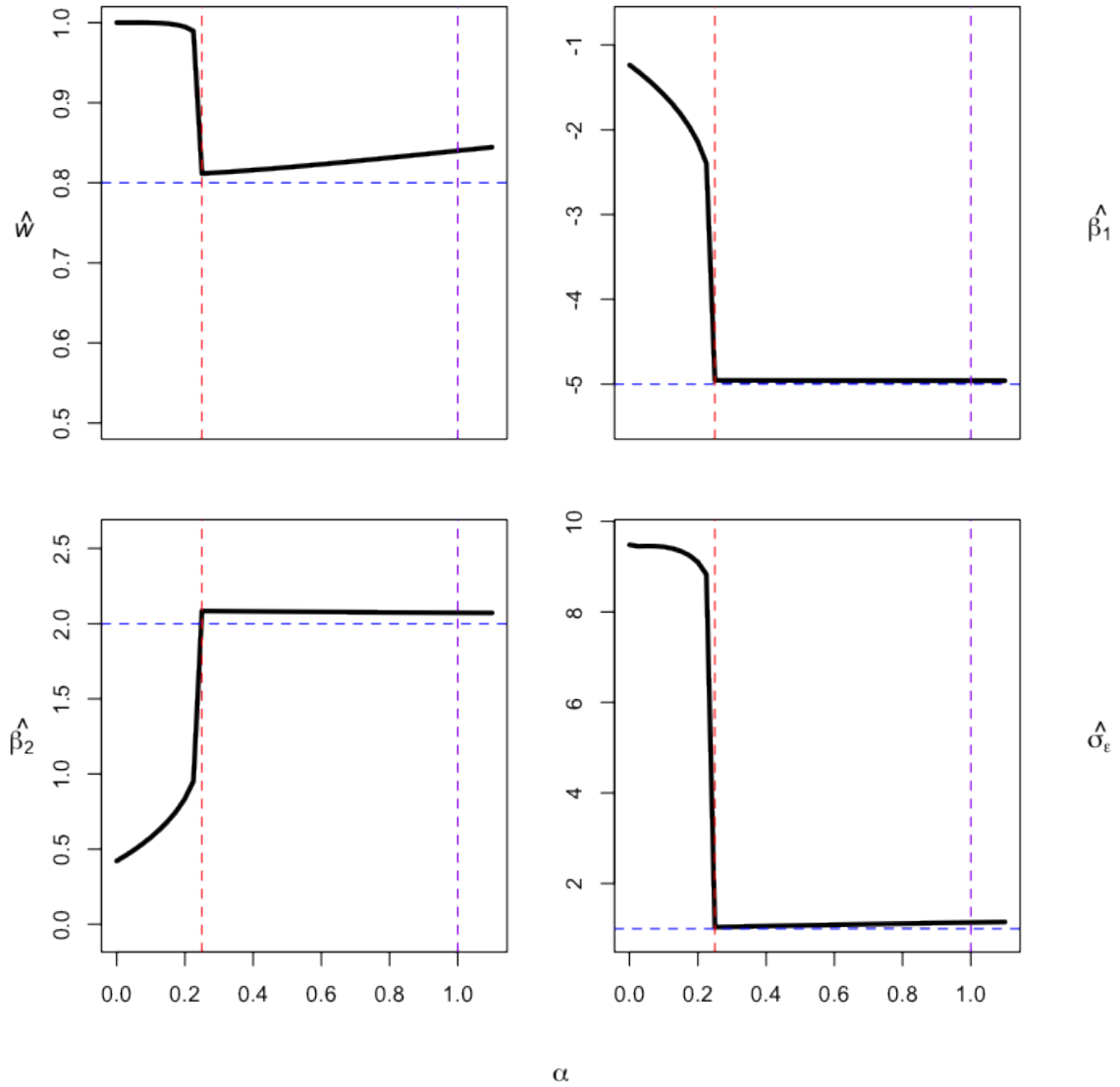


Figure 5.1: Regression Example 1 - Outlying Cluster (20% contamination): Trace plots of MPDC- $\alpha$  divergence estimates (in black) for  $\alpha$  ranging from 0 to 1 by increments of 0.025. Also shown are the targeted parameter values (least-squares estimates computed on just the targeted cluster, in blue), the best  $\alpha$  choice for the example (in red), and the  $L_2E$  mark (in purple) as a reference.

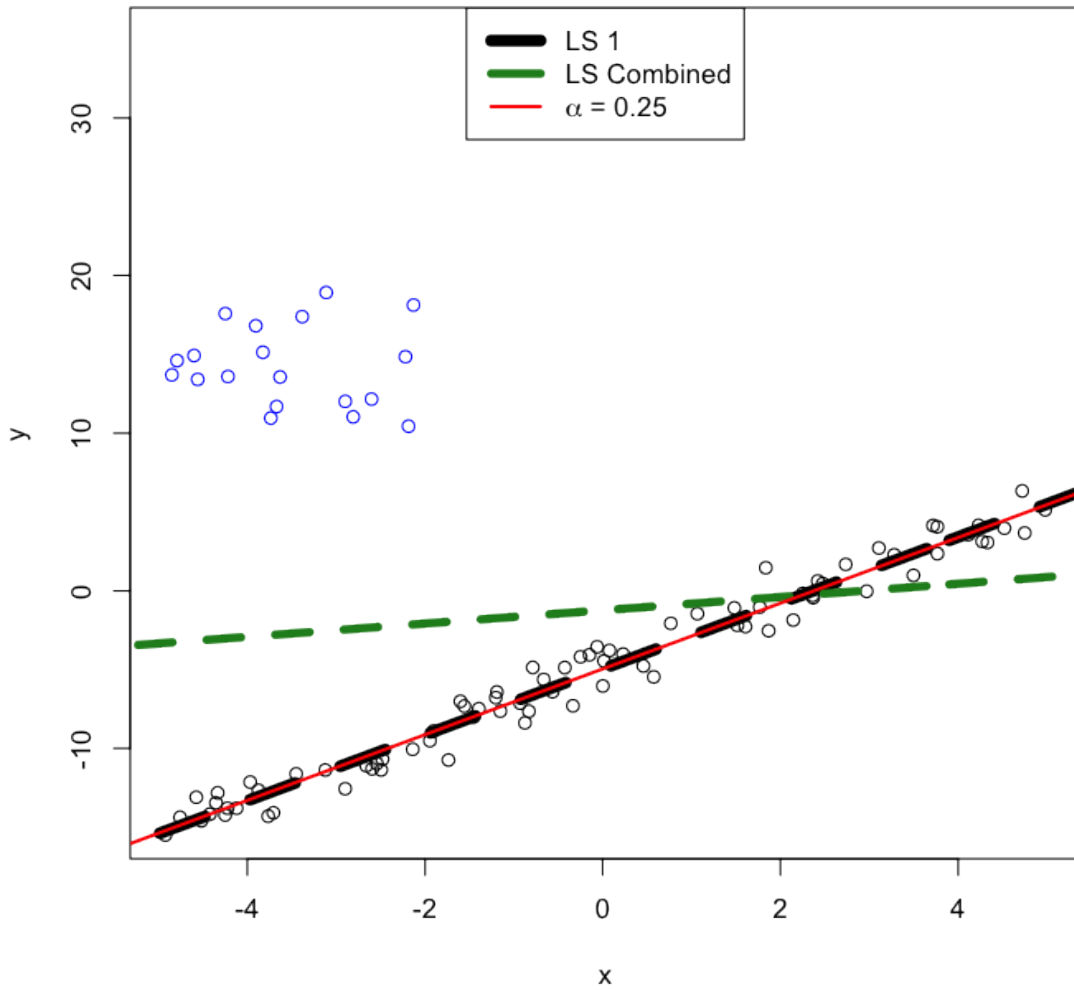


Figure 5.2: Regression Example 1 - Outlying Cluster (20% contamination). MPDC- $\alpha$  estimate for  $\alpha = 0.25$  is shown (red) along with least-squares (LS Combined, in green) estimate is shown and least-squares estimate for only the uncontaminated data (LS 1, in black).

The simulated data for Example 5.1 can be seen in Figure 5.2. The least-squares

line is affected considerably by the outlying cluster, while the MPDC- $\alpha$  estimate with  $\alpha = 0.25$  provides a robust, consistent solution. We investigate the effect on  $\alpha$  of increasing the overlap and level of contamination in the data to 50%.

**Example 5.2:** We simulate a sample of size  $n = 100$ . Let  $x_1, x_2, \dots, x_{50} \sim iid U(-5, 5)$  and  $y_i = 2x_i - 5 + e_i$  for  $1 \leq i \leq 50$ , where  $e_1, e_2, \dots, e_{50} \sim iid N(0, 1)$ . Then, let  $x_{51}, x_{52}, \dots, x_{100} \sim iid U(-5, -2)$  and  $y_{51}, y_{52}, \dots, y_{100} \sim iid U(0, 10)$ .

We have now increased the level of contamination to 50%. As we can see in Figure 5.3, the estimates converge to the true values around an  $\alpha$  value of 1.4. The increase in contamination has led us to rely on a higher  $\alpha$  value to yield an unbiased solution.

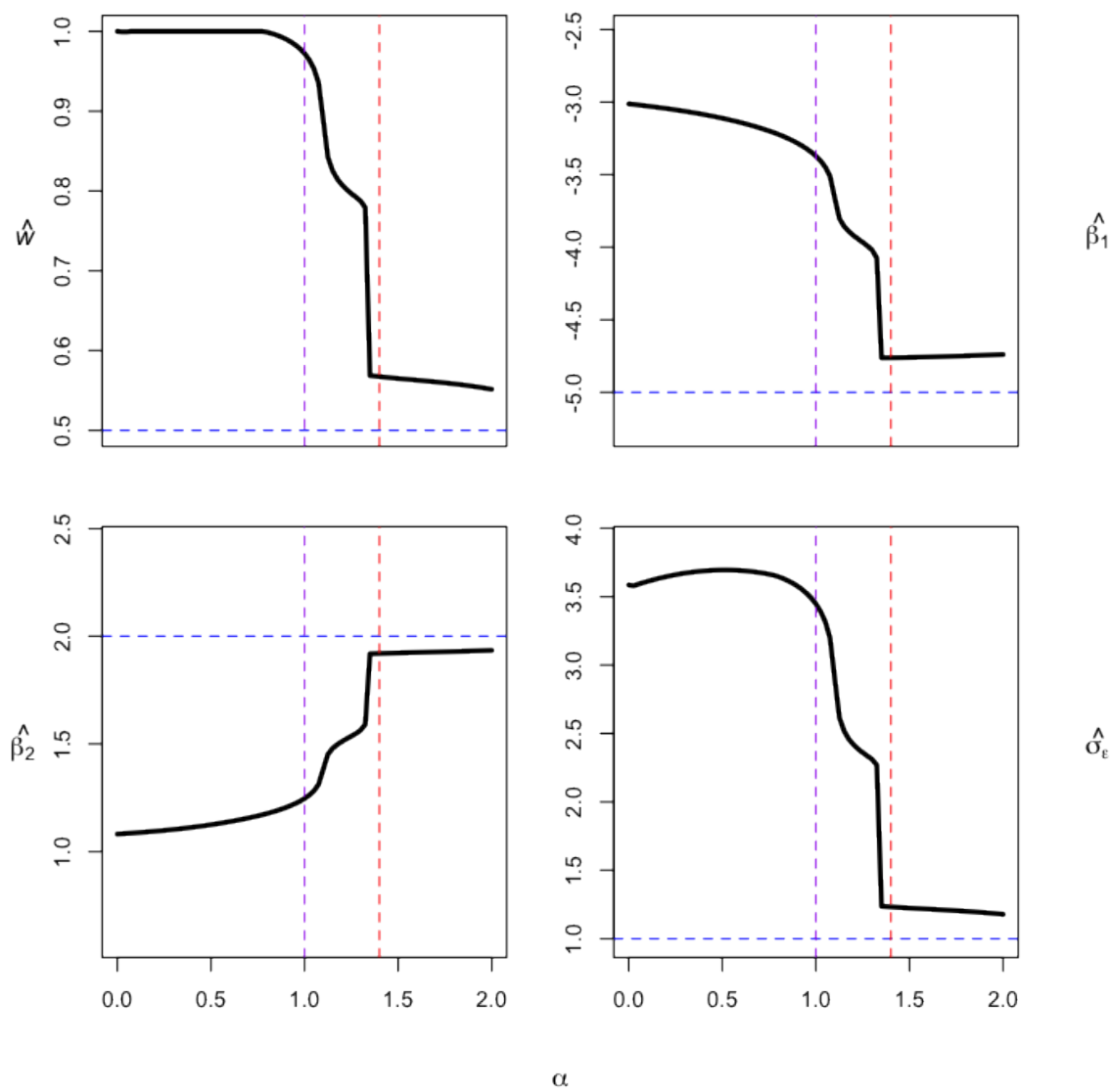


Figure 5.3: Regression Example 2 - Outlying Cluster (50% overlapping contamination): Trace plots of MPDC- $\alpha$  divergence estimates (in black) for  $\alpha$  ranging from 0 to 2.5 by increments of 0.025. Also shown are the targeted parameter values (least-squares estimates computed on just the targeted cluster, in blue), the best  $\alpha$  choice for the example (in red), and the  $L_2E$  mark (in purple) as a reference.

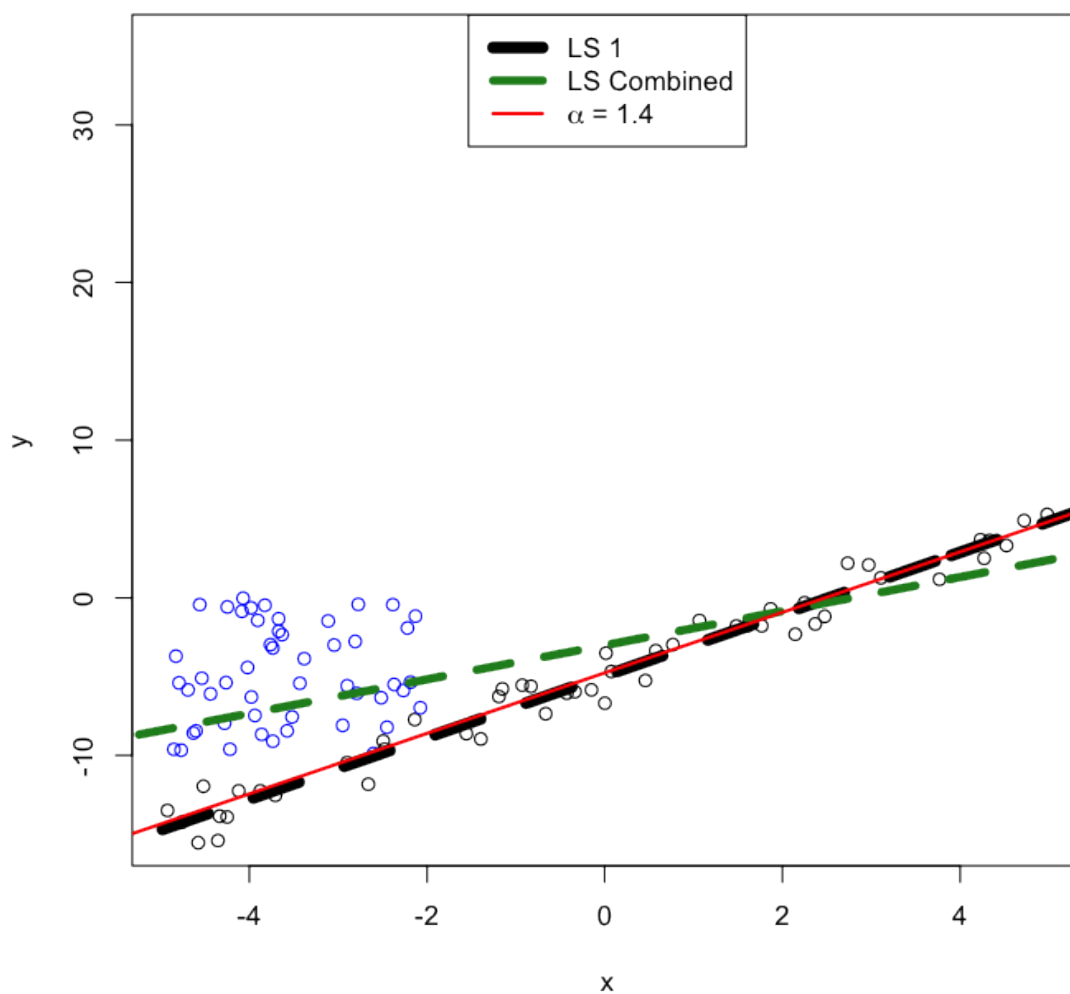


Figure 5.4: Regression Example 2 - Outlying Cluster (50% overlapping contamination). MPDC- $\alpha$  estimate for  $\alpha = 1.4$  is shown (red) along with least-squares (LS Combined, in green) estimate is shown and least-squares estimate for only the uncontaminated data (LS 1, in black).

The simulated data for Example 5.2 can be seen in Figure 5.4. The LS line is



even more greatly affected by the outliers since there are more of them than in the previous example. MPDC- $\alpha$  regression with  $\alpha = 1.4$  provides a consistent solution. Once again,  $L_2E$  fails to reach the desired solution when the least-squares estimates are used as the initial values.

We have shown the need for  $\alpha$  values beyond 1 in particular cases of contamination that comprises a large fraction of the data and/or lies in a particular orientation with respect to the main data. Next we explore a special non-linear regression case for which these high  $\alpha$  values once again become valuable.

## 5.4 Mixed Quadratic Example

We consider an example where we have a mixture of two clusters that can each be modeled by a separate quadratic function:

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \epsilon, \quad (5.6)$$

where  $\epsilon$  has mean 0 and standard deviation  $\sigma_\epsilon$ . We simulate a sample of size  $n = 100$ .

Let  $x_1, x_2, \dots, x_{100} \sim iid U(-2.5, 2.5)$  and  $y_i = 3x_i^2 - 2x_i + 5 + e_i$  for  $1 \leq i \leq 50$  and  $y_i = -2x_i^2 + x_i - 10 + e_i$  for  $51 \leq i \leq 100$ , where  $e_1, e_2, \dots, e_{100} \sim iid N(0, 1)$ .

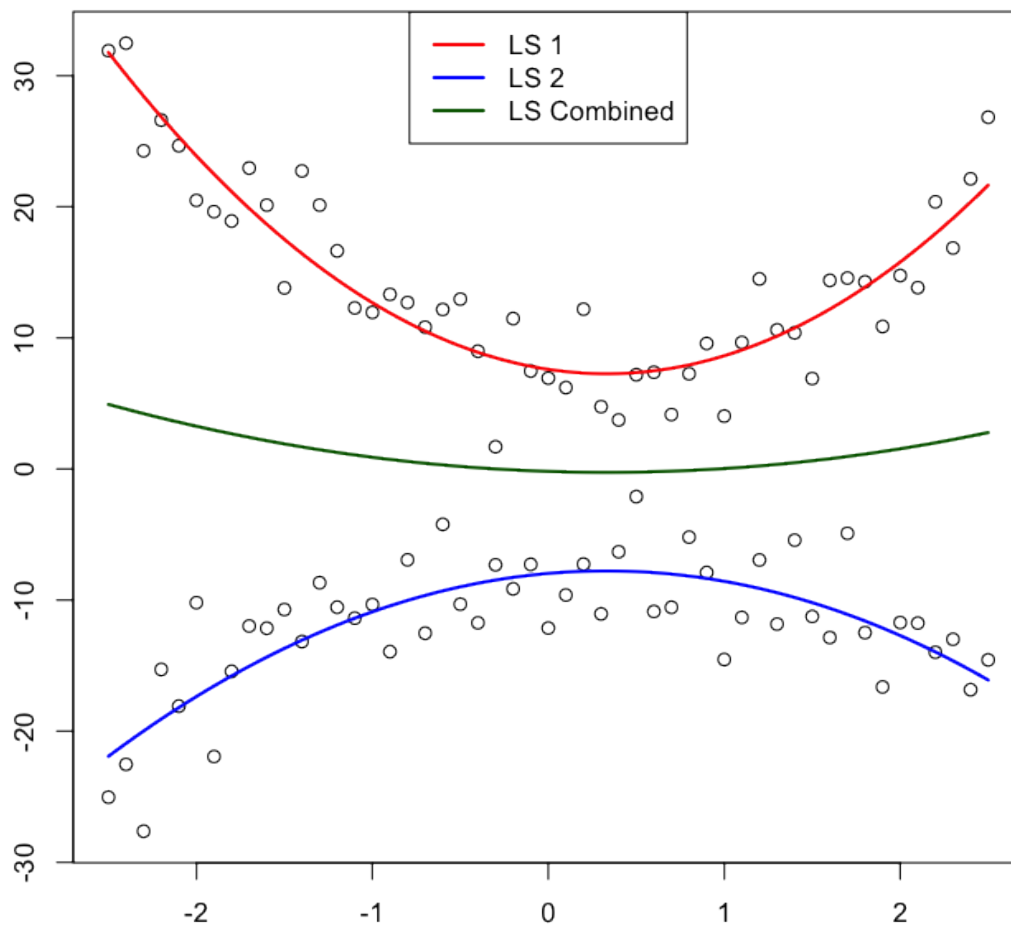


Figure 5.5: Sample of size  $n = 100$  simulated from equal mixture of two quadratic functions with  $N(0, 1)$  noise added to each:  $Y_1 = 3X^2 - 2X + 5$ ,  $Y_2 = -2X^2 + X - 10$ . Also shown are least-squares estimate of  $Y_1$  on  $(X, X^2)'$  (LS 1, shown in red), least-squares estimate of  $Y_2$  on  $(X, X^2)'$  (LS 2, shown in blue), and least-squares estimate of  $(Y_1, Y_2)'$  on  $(X, X^2)'$  (LS Combined, shown in green).

The resulting sample can be seen in Figure 5.5, along with the separate least-squares fits for each cluster and the combined least-squares fit for all the data points. We will put the MPDC- $\alpha$  divergence regression method to the test to see if it can converge to the two separate solutions (red and green lines).

Because of the local nature of our method, it is essential to be mindful of the starting values we use for the algorithm. Thus, we generate  $N = 200$  random starts as follows:

- $\beta_{1,0} \sim U(-30, 30)$
- $\beta_{2,0} \sim U(-20, 20)$
- $\beta_{3,0} \sim U(-10, 10)$
- $\sigma_{\epsilon,0} \sim U(\frac{s_y}{20}, \frac{s_y}{2})$

We then run the MPDC- $\alpha$  divergence algorithm for these 200 random starts for a range of  $\alpha$  values.

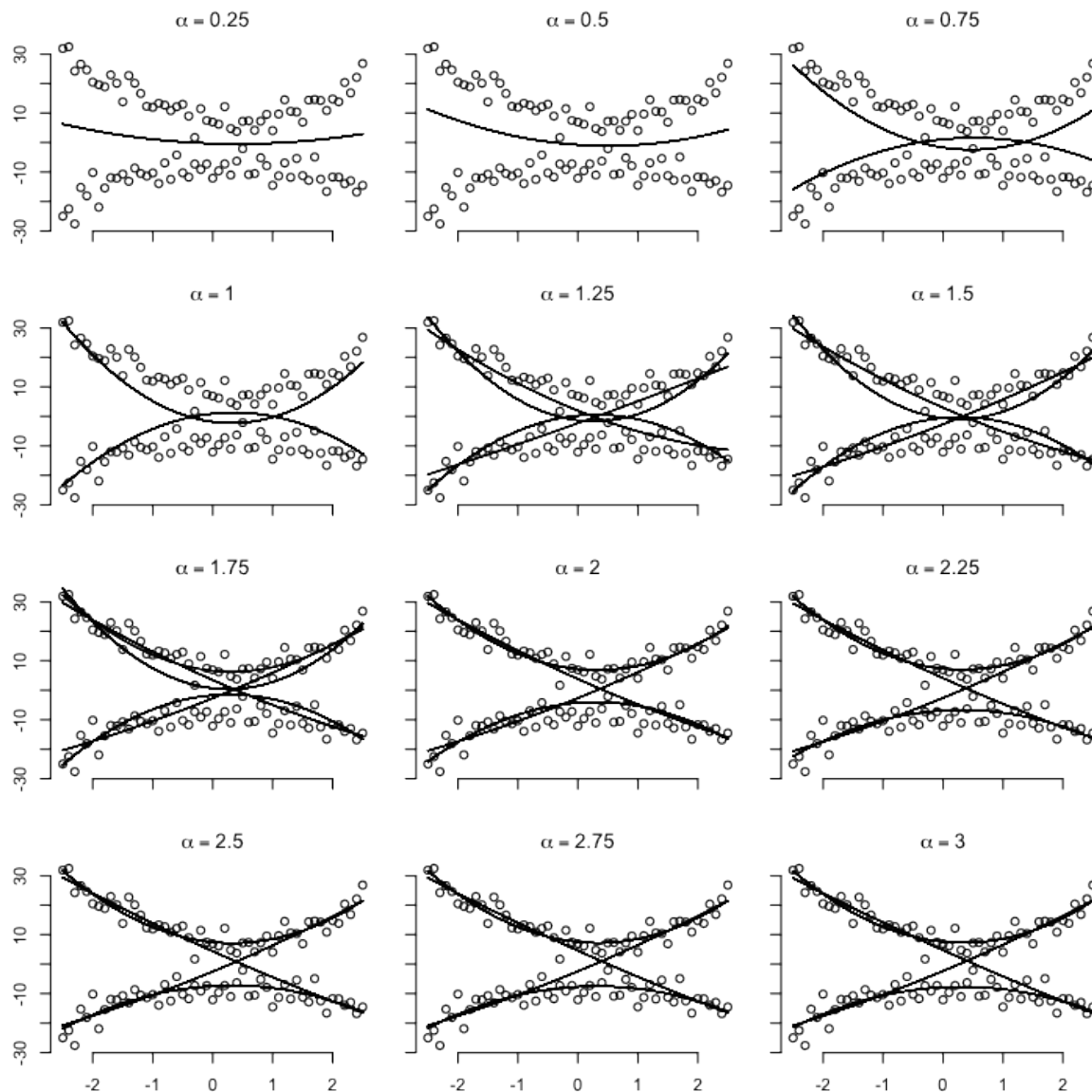


Figure 5.6: Resulting MPDC- $\alpha$  estimates for 200 random starts for sample of size  $n = 100$  generated from mixture of quadratic functions with  $N(0, 1)$  noise added to each:  $Y_1 = 3X^2 - 2X + 5$ ,  $Y_2 = -2X^2 + X - 10$ .

The results are shown in Figure 5.6, with the black lines denoting the solutions yielded by MPDC- $\alpha$  divergence regression. As we would expect, for  $\alpha$  values close to

0, we converge to a solution close to the least-squares fit for all of the data, which misses all of the points. As  $\alpha$  increases, we see that solutions start to take the quadratic shape, but they still miss the bulk of the points. It is not until  $\alpha = 2.5$  that we see consistent convergence to 4 solutions, 2 of which are the least-squares fits for the individual clusters. Thus, the additional robustness provided by  $\alpha$  values beyond 1 allows us to detect these local solutions. There is an upper threshold to this robustness, as we can see by the sporadic solutions for  $\alpha = 3.5$ . The  $\alpha$  range of 2.5 to 3 provides consistent solutions for this example.

# Chapter 6

## Conclusions and Future Work

We have developed the framework for the MPDC- $\alpha$  divergence estimator, which provides a robust procedure for estimating particular functional components. A deeper exploration of exact computing times and comparison with those of other robust algorithms such as the FAST-MCD is needed. The selection of  $\alpha$  should be done with consideration for characteristics of the data, any prior knowledge about the level of contamination, and ensuring that not too much efficiency is sacrificed depending on the dimension  $p$ . Values of  $\alpha$  in the range of  $[0.3, 0.5]$  serve us well when estimating parameters with data that has contamination that is well-separated from the main data. When there is overlap between the contamination and main data, we benefit from  $\alpha$  values between 0.5 and 2, depending on the dimension of the problem. Basu's  $\alpha$ -divergence procedure limited the range of  $\alpha$  values to  $[0,1]$ , and we have found usefulness for values of  $\alpha$  greater than 1, particularly in the case of low-dimensional parametric density estimation as well as for robust regression when there is high contamination or considerable overlap between the main and contaminated data.

Other parameter transformations can also be attempted to help improve computational efficiency, including the Givens parametrization for the covariance matrix. We would also like to explore other parametric models to be used with the MPDC, such as a Beta distribution, and nonparametric approaches such as kernel estimation with the incorporation of the  $\alpha$ -divergence function. While it is not explored here, if  $\alpha$  increases to a certain level (beyond 3 in many of the parametric estimation cases), the estimates yielded by MPDC- $\alpha$  mimic the behavior of modal estimates – this connection could be further investigated. We would also examine other applications for the MPDC- $\alpha$  estimator, including estimating the covariance matrix for a financial time series model.

# References

- Aelst S.V. and Rousseeuw P.J. (2009), “Minimum volume ellipsoid.” *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 1, Issue 1, pp. 71-82.
- Basu, A. et al. (1998), “Robust and efficient estimation by minimising a density power divergence.” *Biometrika*, Vol. 85, No. 3, pp. 549-559.
- Campbell, N.A. (1980), “Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 29, No. 3, pp. 231-237.
- Huber, P. (1964), “Robust Estimation of a Location Parameter.” *Ann. Math. Statistics*, Vol. 35, No. 1, pp. 73-101.
- Huber, P and Ronchetti, E.M. (2009), *Robust Statistics*, 2nd ed.
- Hubert, M. and Debruyne, M. (2010), “Minimum covariance determinant.” *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 2, No. 1, pp. 3643.
- Mardia, K., Kent, J.T. and Bibby, J.M. (1980), *Multivariate Analysis*, 1st ed.
- Maronna, R.A. (1976), “Robust M-Estimators of Location and Scatter.” *Annals*



*of Statistics*, Vol. 4, No.1, pp. 51-67.

- Rousseeuw, P.J. (1984), “Least Median of Squares Regression.” *Journal of the American Statistical Association*, Vol. 79, No. 388, pp. 871-880.
- Rousseeuw, P.J. (1985), “Multivariate Estimation With High Breakdown Point.” *Mathematical Statistics and Applications*, Vol. B, pp. 283-297.
- Rousseeuw, P.J. and Hubert, M. (2011). “Robust statistics for outlier detection.” *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 1, Issue 1, pp. 7379.
- Rousseeuw, P.J. and Van Driessen, K. (1999), “A Fast Algorithm for the Minimum Covariance Determinant Estimator.” *Technometrics*, Vol. 41, No. 3, pp. 212-223.
- Scott, D.W. (2001), “Parametric Statistical Modeling by Minimum Integrated Square Error.” *Technometrics*, Vol. 43, pp. 274-285.
- Scott, D.W. (2004), “Partial Mixture Estimation and Outlier Detection in Data and Regression,” in *Theory and Applications of Recent Robust Methods*, edited by M. Hubert, G. Pison, A. Struyf and S. Van Aelst, Series: Statistics for Industry and Technology, Birkhauser, Basel, pp. 297-306.
- Stackpole, E. and Mahon, D. (2012), “Analyzing a Pitchers Influence of Balls Hit Into Play.” Pending publication. Rice University Vertical Integration of Research and Education: Statistics in Sports Group.

# Appendix

## Optimizer: *nlsminb*

For our optimization we use the *nlsminb()* function in R. Without altering the default, the function performs unconstrained optimization on the objective for the given initial values. Because of some numerical rounding errors, we introduced a small additive adjustment factor,  $\epsilon$ , to our objective in order to keep the algorithm stable.

## Parametric Estimation Trace Plots for $6 \leq p \leq 10$

These are the trace plots of the parameter estimates for the cases of  $p \in [6, 10]$ .

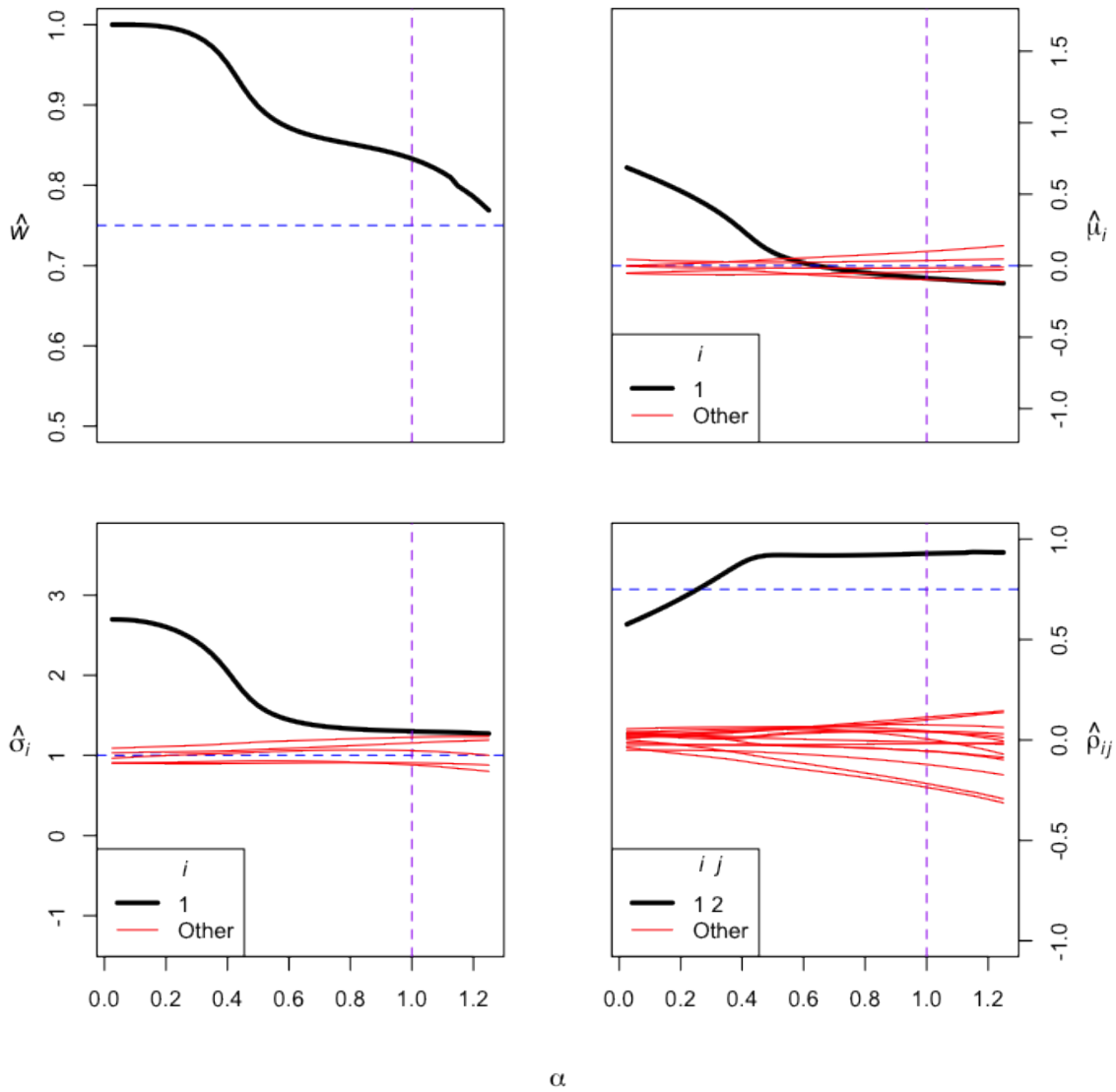


Figure A.1: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (3, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_6$ .

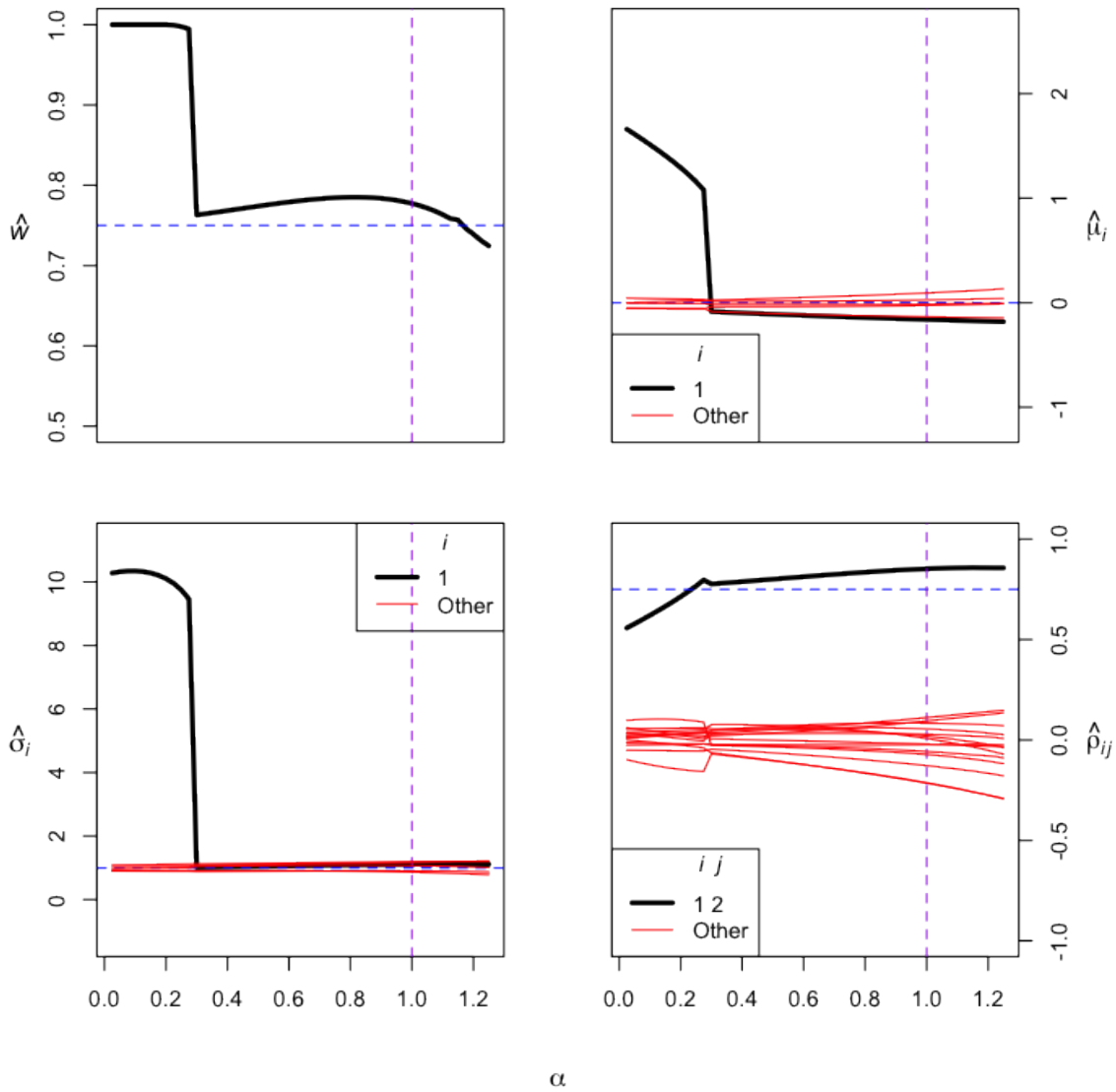


Figure A.2: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (7, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_6$ .

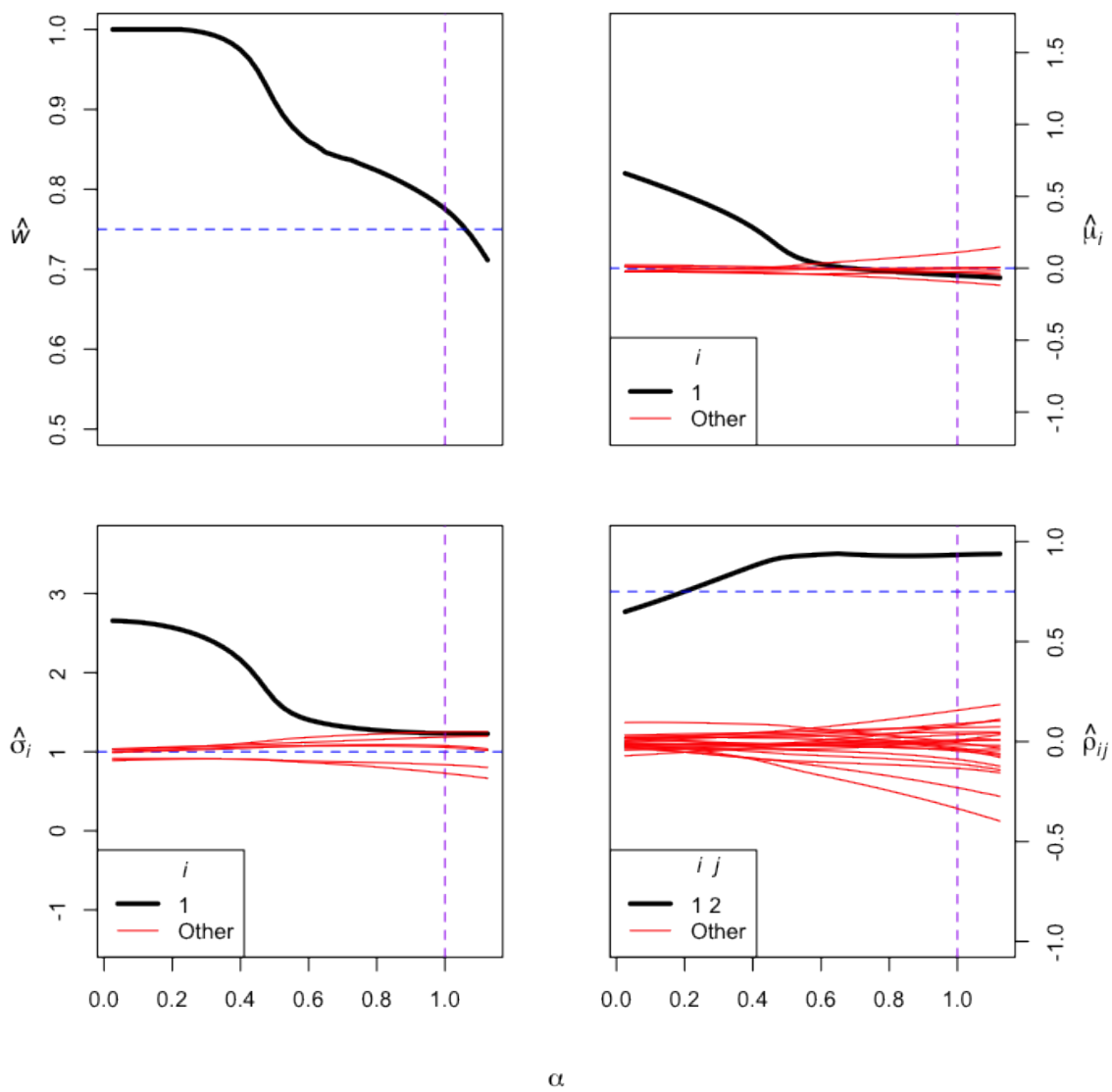


Figure A.3: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (3, 0, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_7$ .

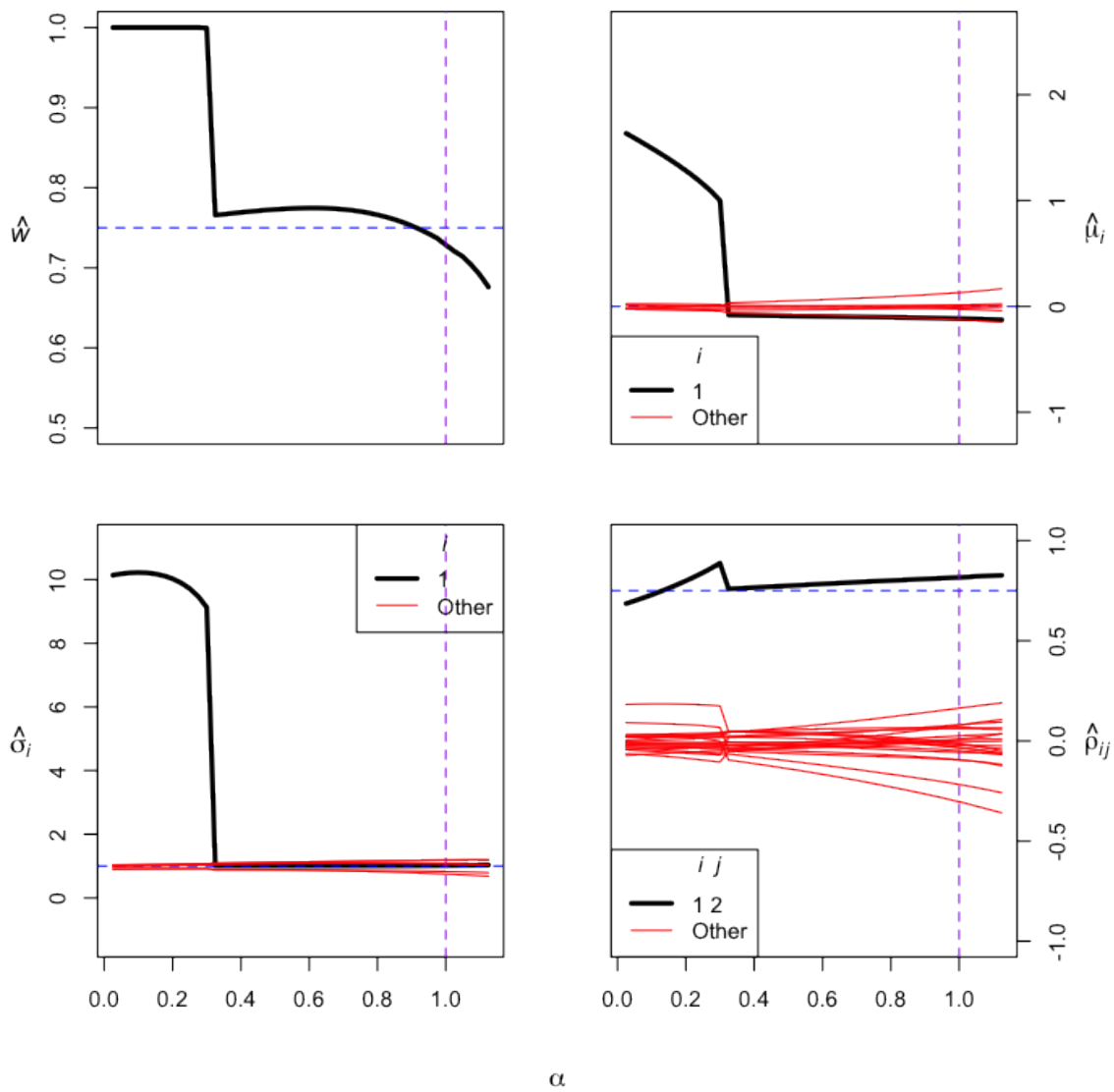


Figure A.4: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (7, 0, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_7$ .

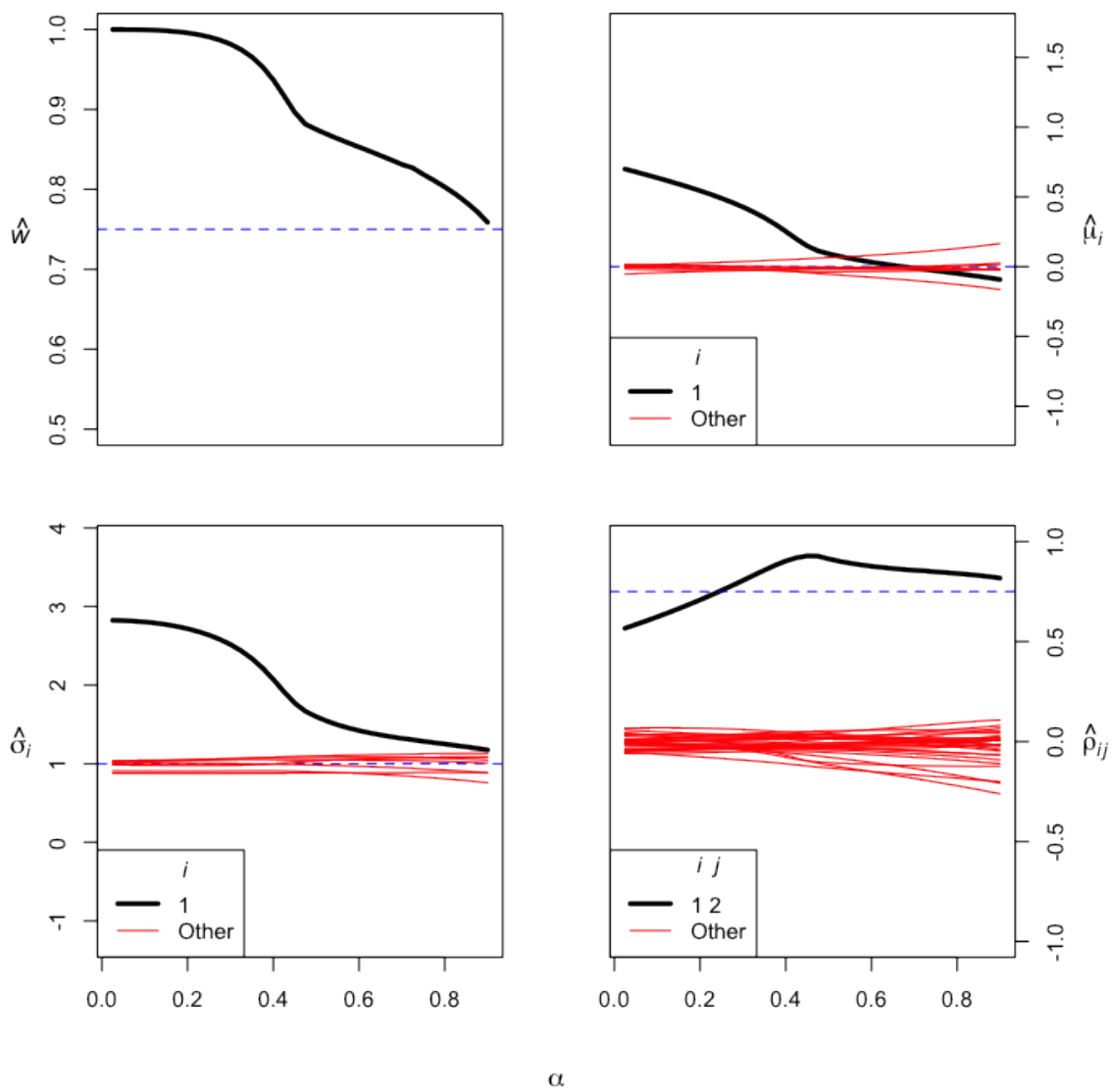


Figure A.5: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (3, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_8$ .

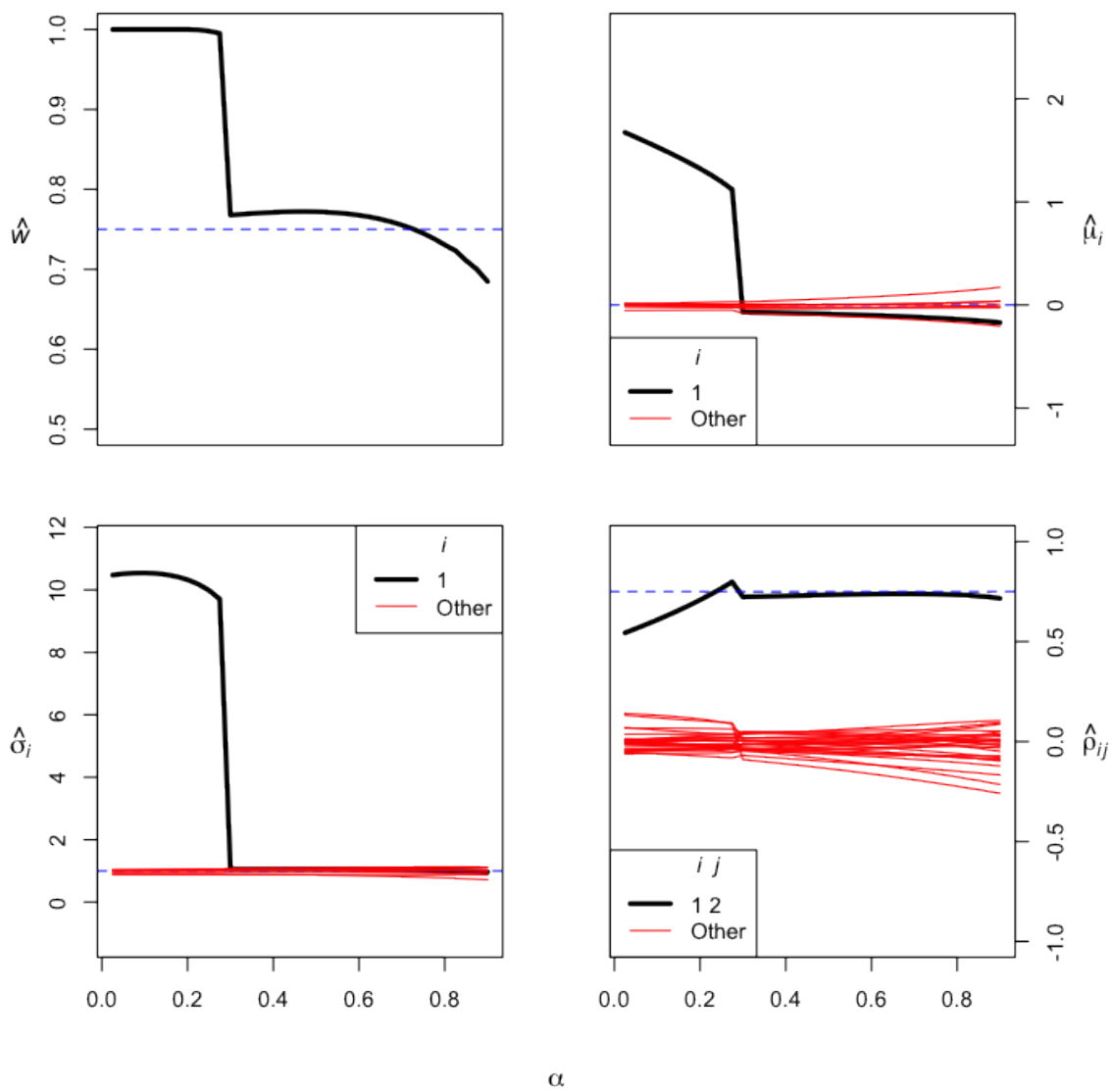


Figure A.6: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (7, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_8$ .



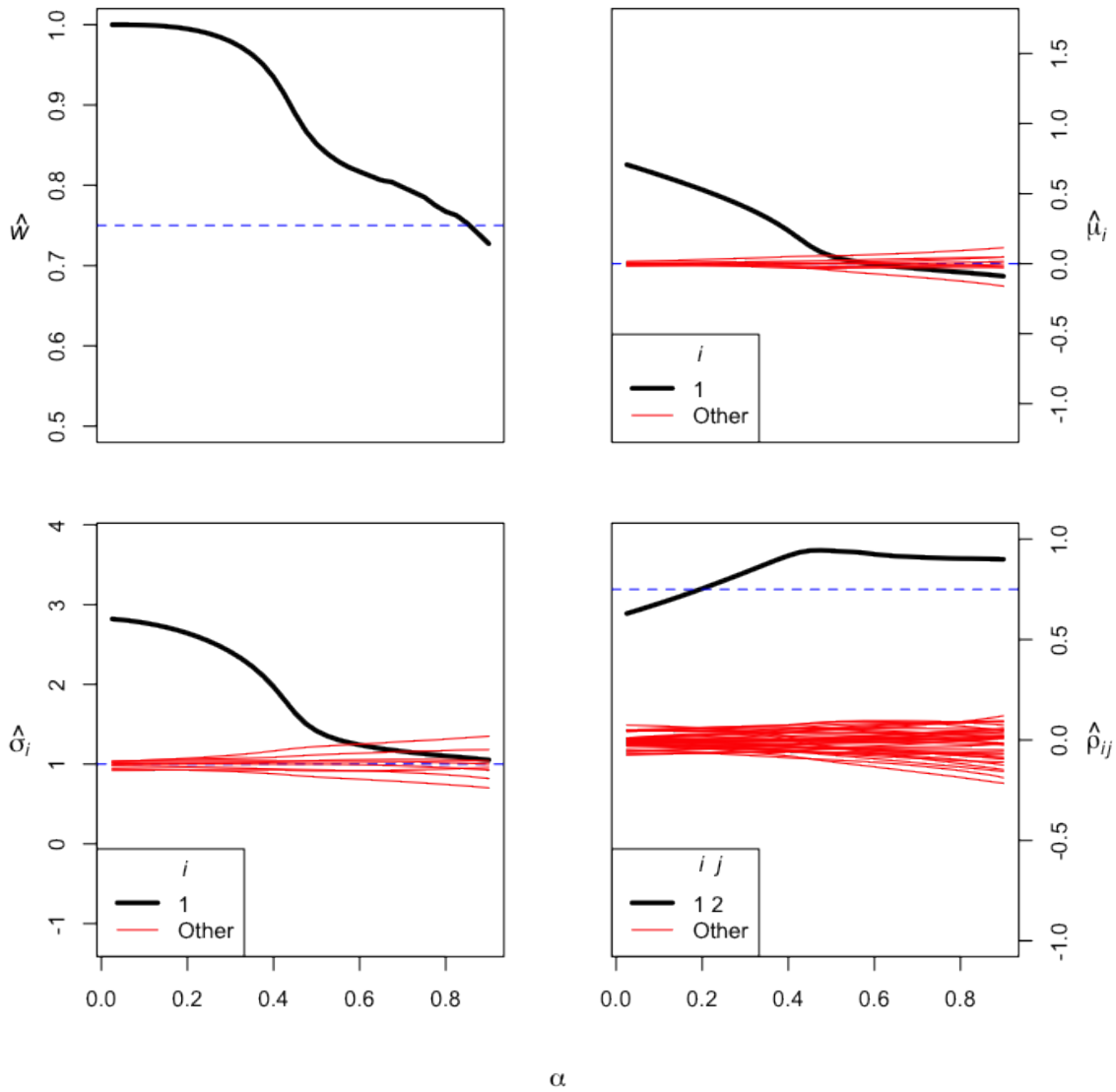


Figure A.7: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (3, 0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_9$ .

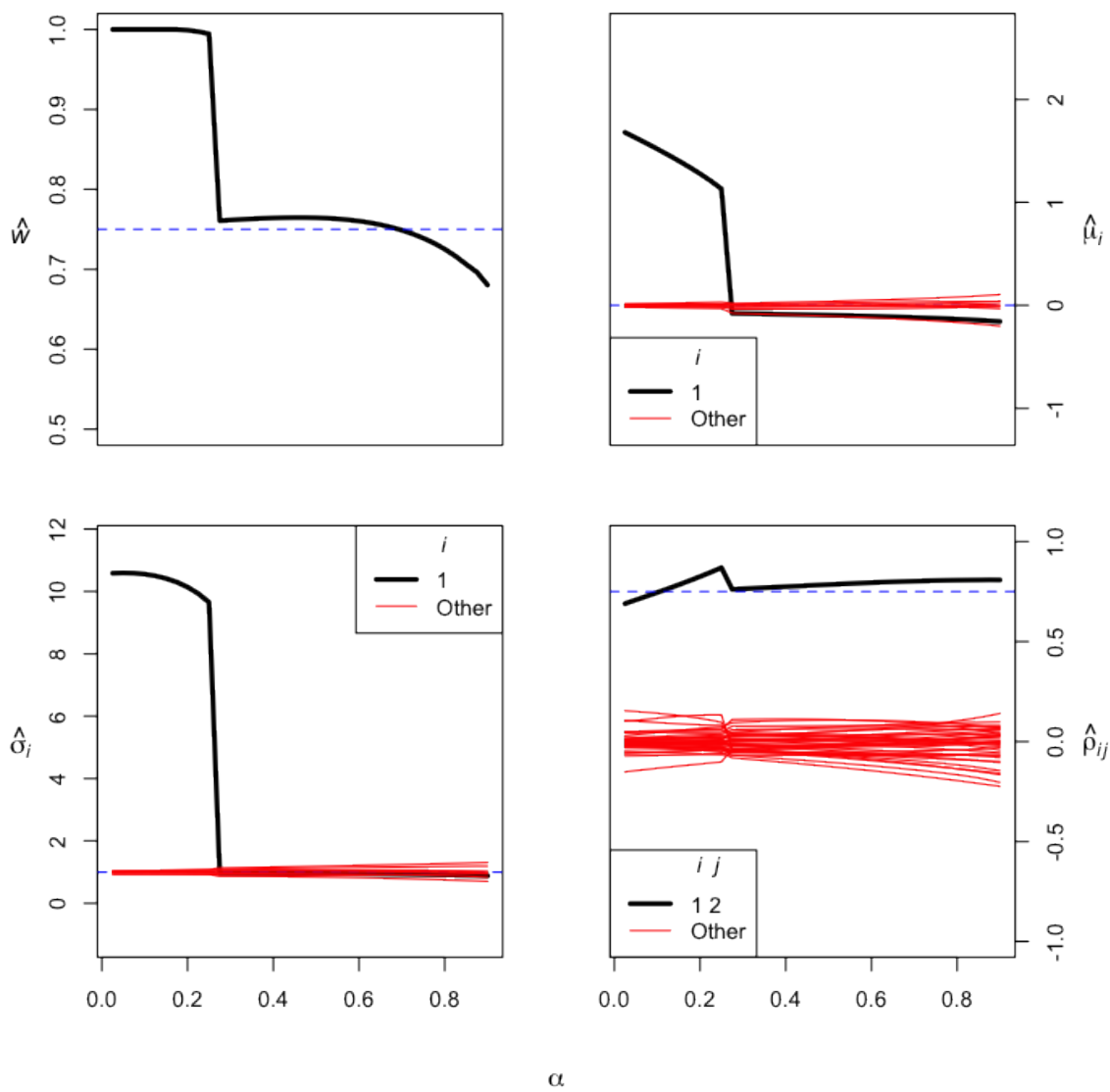


Figure A.8: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (7, 0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_9$ .

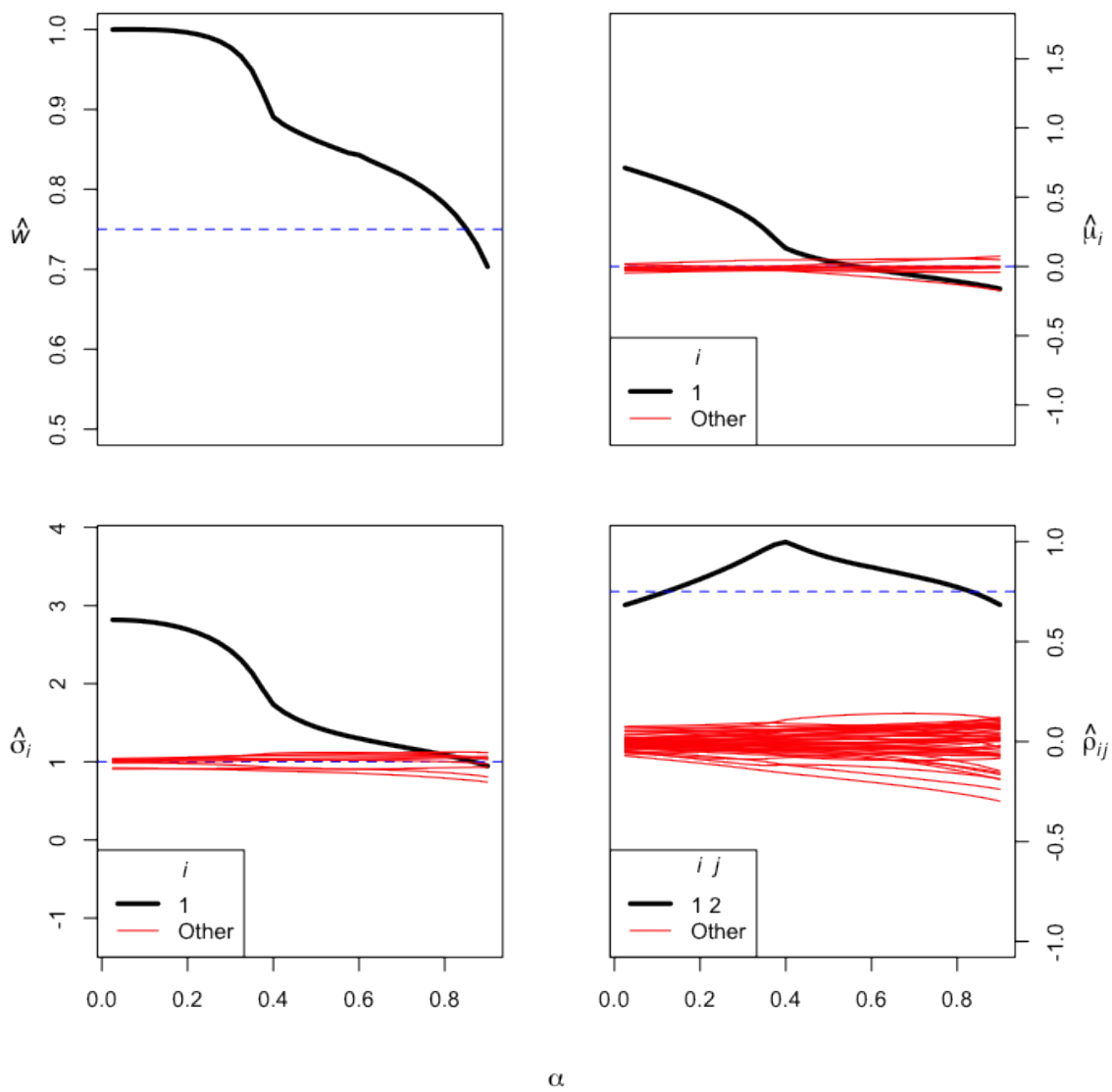


Figure A.9: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (3, 0, 0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_{10}$ .

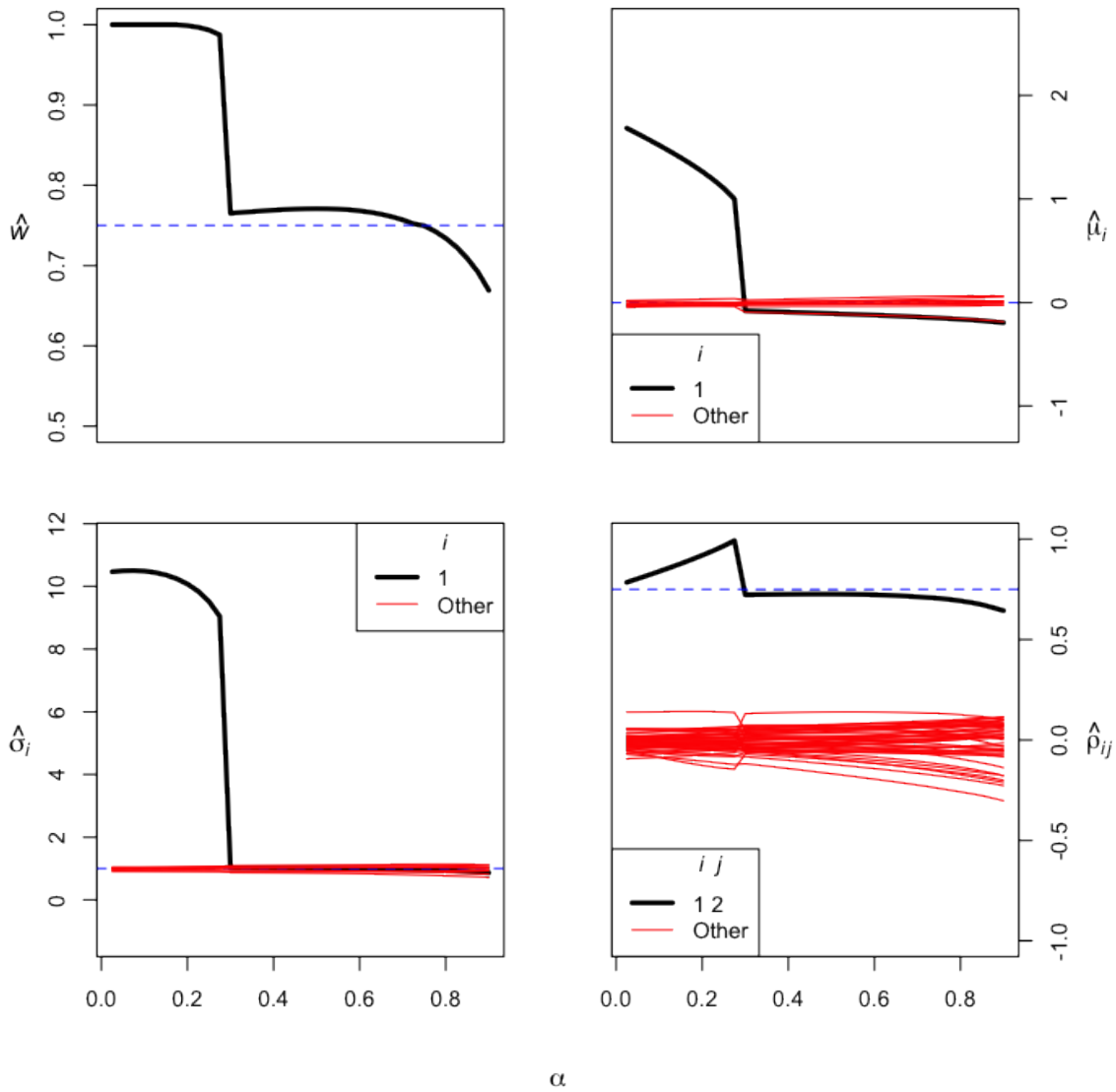


Figure A.10: Trace plots of MPDC- $\alpha$  divergence estimates for  $\alpha$  ranging from 0 to 2 by increments of 0.025. Black lines represent parameters of interest, and red lines indicate other parameters that we track to assure algorithm stability. Estimates based on sample of size  $n = 1000$  simulated from Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (7, 0, 0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_{10}$ .

## Parametric Estimation RMSE Plots for $6 \leq p \leq 10$

These are the RMSE plots for  $\alpha$  selection for the cases of  $p \in [6, 10]$ .

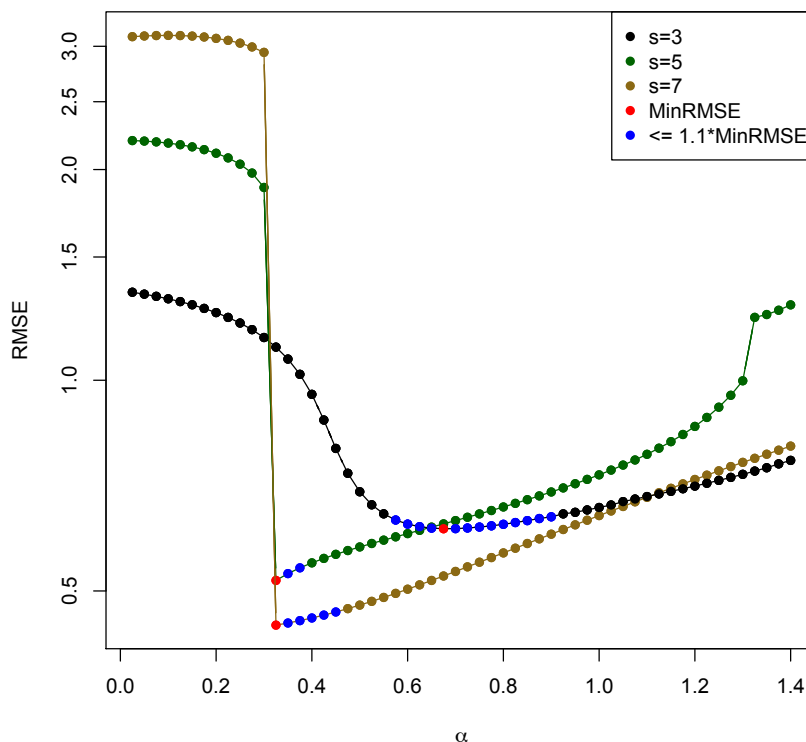


Figure A.11: Selection of  $\alpha$  for  $p = 6$ . RMSE of MPDC- $\alpha$  estimate versus  $\alpha$  for three different degrees of separation ( $s$ ). Derived from 100 simulations of samples of size  $n = 1000$  from a Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (s, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_6$ .

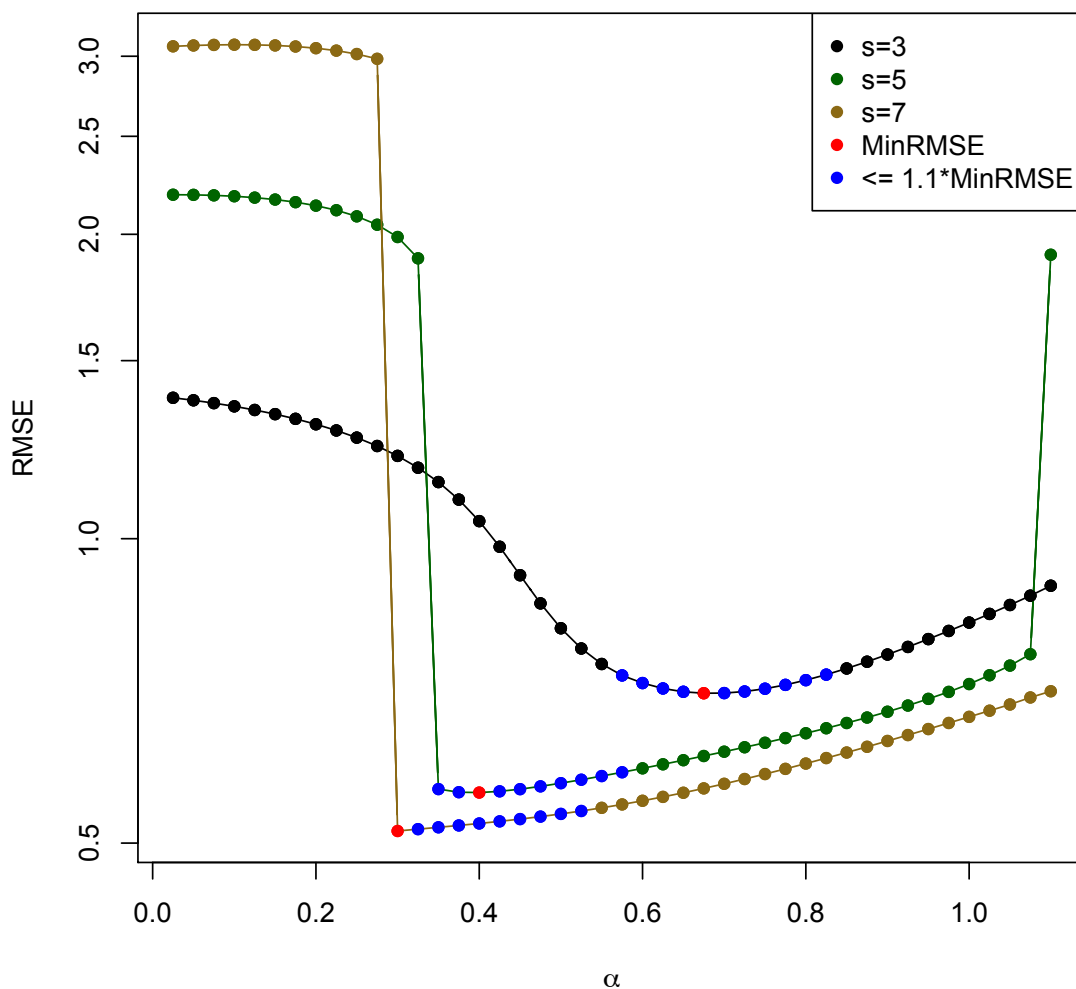


Figure A.12: Selection of  $\alpha$  for  $p = 7$ . RMSE of MPDC- $\alpha$  estimate versus  $\alpha$  for three different degrees of separation ( $s$ ). Derived from 100 simulations of samples of size  $n = 1000$  from a Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (s, 0, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_7$ .

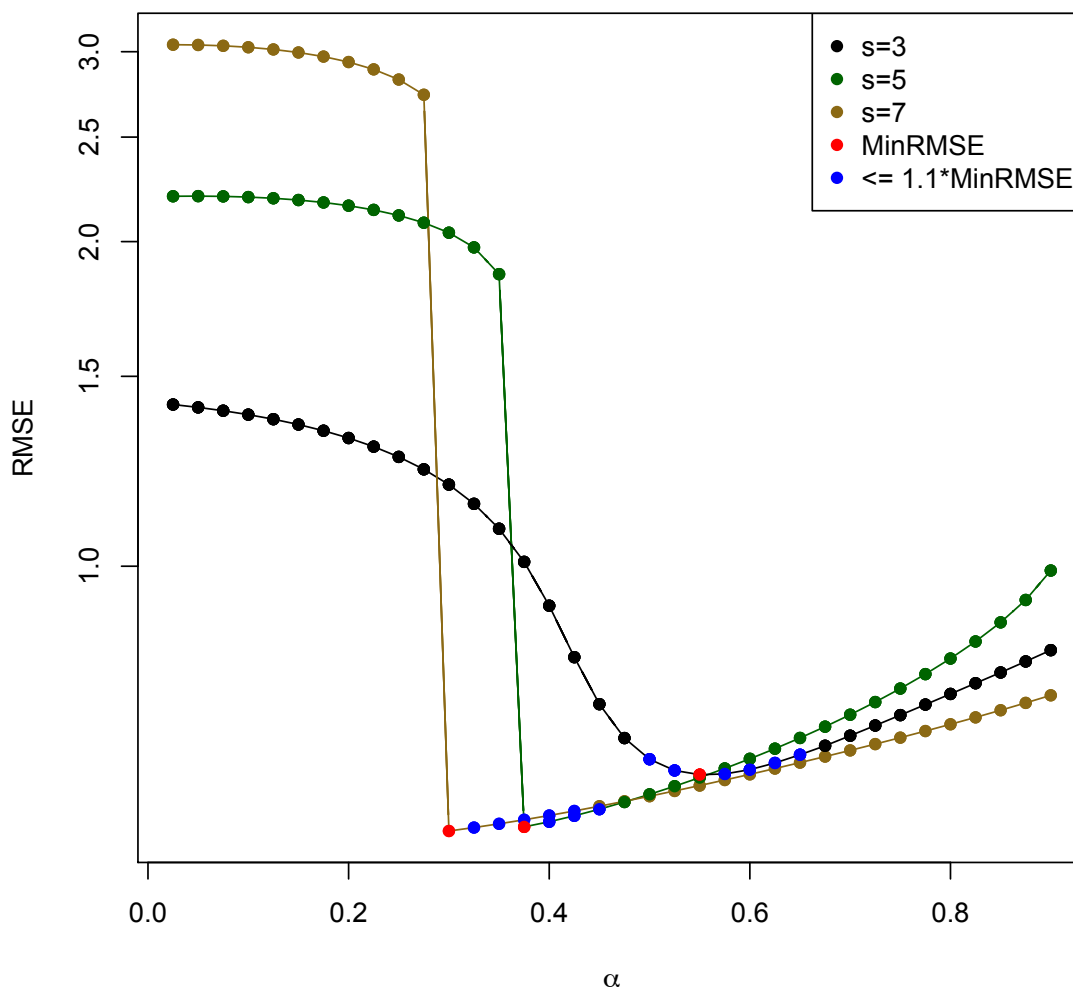


Figure A.13: Selection of  $\alpha$  for  $p = 8$ . RMSE of MPDC- $\alpha$  estimate versus  $\alpha$  for three different degrees of separation ( $s$ ). Derived from 100 simulations of samples of size  $n = 1000$  from a Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (s, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_8$ .

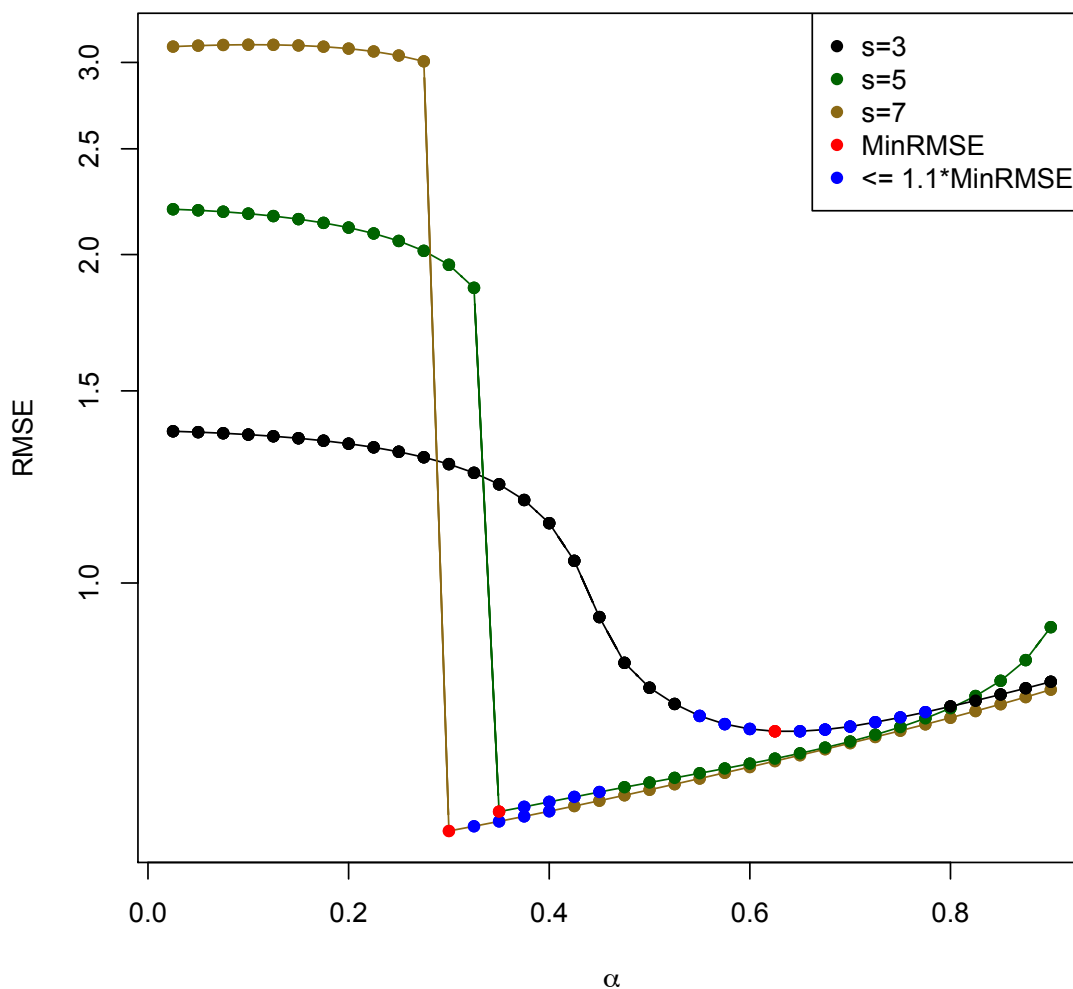


Figure A.14: Selection of  $\alpha$  for  $p = 9$ . RMSE of MPDC- $\alpha$  estimate versus  $\alpha$  for three different degrees of separation ( $s$ ). Derived from 100 simulations of samples of size  $n = 1000$  from a Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (s, 0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_9$ .



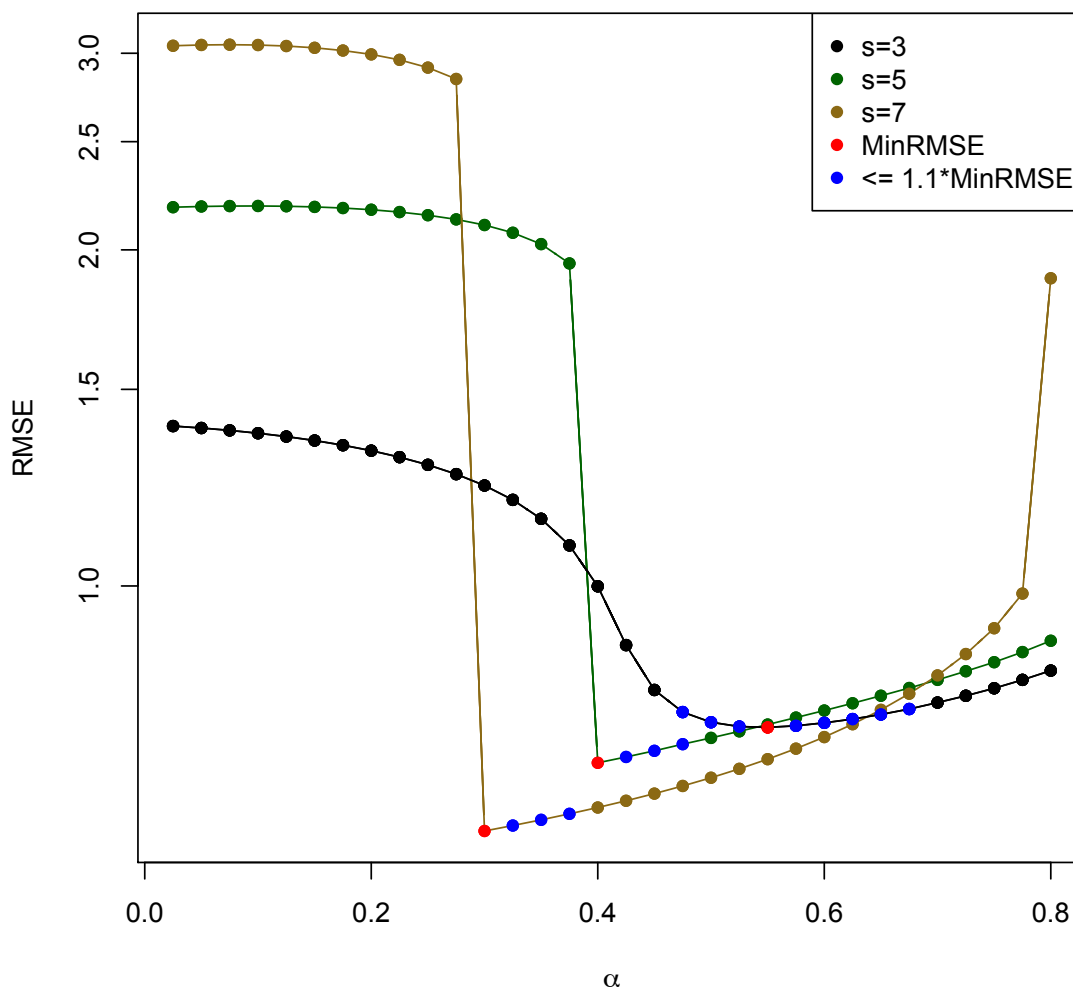


Figure A.15: Selection of  $\alpha$  for  $p = 10$ . RMSE of MPDC- $\alpha$  estimate versus  $\alpha$  for three different degrees of separation ( $s$ ). Derived from 100 simulations of samples of size  $n = 1000$  from a Normal mixture distribution with parameters:  $w = 0.75$ ;  $\mu_1 = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\mu_2 = (s, 0, 0, 0, 0, 0, 0, 0, 0, 0)'$ ;  $\rho_{12} = 0.75$ ;  $\Sigma_2 = I_{10}$ .

## R Code

```
##### MPDC-Alpha Divergence Criterion & Estimation #####
##### Justin Silver #####
##### Rice University #####
##### Department of Statistics #####

# MPDC-alpha criterion, small additive factor epsilon
crit = function(x, alpha=alpha, X=X, eps=eps) {
  n = nrow(X); p = ncol(X) # dimensions of data
  U <- matrix(0,p,p); U[col(diag(p))>=row(diag(p))] <-
    x[1:(p*(p+1)/2)]
  diag(U) = exp(diag(U)) # transformation of Cholesky
  U[col(diag(p))>row(diag(p))]=2/(1+exp(-U[col(diag(p))>
    row(diag(p))]))^2-1
  w = 1/(1+exp(-x[p*(p+1)/2+1])) # logistic transformation of weight
  mu <- rep(NA, p); mu = x[(p*(p+1)/2+2):((p+2)*(p+1)/2)] # mean vec.
  d = prod(diag(U)) # Cholesky determinant
  # computationally efficient criterion
  w^(1+alpha)*(2*pi)^(-p*alpha/2)*(1+alpha)^(-p/2)*d^alpha-
    (1+1/alpha)*w^alpha/n*sum((2*pi)^(-p*alpha/2)*d^alpha*
    exp(-.5*alpha*rep(1,p)%*(U%*(t(X)-mu))^2))-eps
}
```

```

# Function yielding parameter estimates for range of alpha values
alp.div <- function(X, parms, alpha, eps, p) {
  # Reverse parameter transformations
  U <- matrix(0,p,p); U[col(diag(p))>=row(diag(p))] <-
    parms[1:(p*(p+1)/2)]
  diag(U) = log(diag(U))
  U[col(diag(p))>row(diag(p))] <- -log(sqrt(2/(1+U[col(diag(p))>
    row(diag(p))]))-1)
  w = parms[p*(p+1)/2+1]; mu <- rep(NA,p); mu = parms[(p*(p+1)/2+2):
    ((p+2)*(p+1)/2)]

  x0 = c(U[col(diag(p))>=row(diag(p))],log(w/(1-w)), mu) # init. vals
  # unconstrained minimization
  ans = nlminb(x0,crit,alpha=alpha,X=X,eps=eps,control=list(
    iter.max=500,eval.max=500))

  U <- matrix(0,p,p); U[col(diag(p))>=row(diag(p))] <-
    ans$par[1:(p*(p+1)/2)]
  diag(U) = exp(diag(U))
  U[col(diag(p))>row(diag(p))] <- 2/(1+exp(-U[col(diag(p))>
    row(diag(p))]))^2-1

```

```
w = 1/(1+exp(-ans$par[p*(p+1)/2+1]))  
mu <- rep(NA, p); mu = ans$par[(p*(p+1)/2+2):((p+2)*(p+1)/2)]  
obj = ans$obj; conv = ans$conv # examine objective and convergence  
list(w=w, mu=mu, sig=chol2inv(U), obj=obj, conv=conv) # par. est.  
}
```