

A MULTIVARIATE
ADAPTIVE TRIMMED LIKELIHOOD
ALGORITHM

Daniel Dice Schubert

THIS THESIS IS PRESENTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
MURDOCH UNIVERSITY
MURDOCH, WA 6150
AUSTRALIA
2005

I declare that this thesis is my own account of my research and contains as its main content work which has not previously been submitted for a degree at any tertiary education institution.

Daniel Dice Schubert

Acknowledgements

I would like to sincerely thank and acknowledge the following people who helped me with this thesis.

- My supervisor Dr. Brenton Clarke for introducing me to Robust Statistics and Multivariate Data Analysis in my graduate years and for his friendship, support and encouragement.
- Will Stirling, Murdoch University IT Guru, for helping me with all my PC problems.
- Professor Ronald Butler for sending me his personal notes regarding the asymptotic theory relating to Butler et al 1993.
- Dr. Mark Lukas for his many Matlab hints to speed up my programs and his instructions with regard to LaTeX.
- Phd candidate Suzanne Brown and Dr. Martine van de Poll for their helpful ideas with my R code.

Abstract

The research reported in this thesis describes a new algorithm which can be used to robustify statistical estimates *adaptively*. The algorithm does not require any pre-specified cut-off value between inlying and outlying regions and there is no presumption of any cluster configuration. This new algorithm adapts to any particular sample and may advise the *trimming* of a certain proportion of data considered extraneous or may divulge the structure of a multi-modal data set. Its adaptive quality also allows for the confirmation that uni-modal, multivariate normal data sets are outlier free. It is also shown to behave independently of the *type* of outlier, for example, whether applied to a data set with a solitary observation located in some extreme region or to a data set composed of clusters of outlying data, this algorithm performs with a high probability of success.

Contents

Introduction	1
1 Review of Robust Estimation techniques	4
1.1 Statistical Distance	4
1.2 Affine Equivariance and Maximum Likelihood Estimation	9
1.3 M-estimate	10
1.4 Robustification of Univariate Regression	14
1.5 S-estimate	16
1.6 M-estimate for Multivariate Data	17
1.7 S-estimate for Multivariate data	19
1.8 The MVE and MCD estimates	19
1.9 Computational Expense	21
1.10 MCD Algorithm	22
1.11 Outliers	23
1.12 Fixed Threshold Detection Methods	25

1.12.1	Robust fixed threshold	25
1.12.2	Forward Search	28
1.12.3	Standardized distances and simulations	30
1.13	Cluster Techniques	32
1.13.1	K-means	32
1.13.2	Agglomerative Hierarchical	34
1.13.3	EM-Algorithm	38
2	New Proposal	46
2.1	Univariate Adaptive Trimmed Likelihood	46
2.2	Multivariate Adaptive Trimmed Likelihood	49
2.3	Basic constructs for new algorithm	51
2.4	Monte Carlo simulations	53
2.4.1	Instances of multiple minima	57
2.4.2	t -distributed data	62
2.4.3	Correlated transformations	66
2.4.4	T2 vs non-robust estimates	68
2.5	Comparison with Fixed Threshold Methodology	69
2.6	The T2 Algorithm - further deliberations	75
2.6.1	Determinant vs Trace	77

2.7	Gervini comparison	77
2.8	Online data sets	82
2.8.1	Cricket Batting Data	92
2.9	Algorithm for the new Proposal	95
3	New Robustification of Univariate and Multivariate Regression	98
3.1	Univariate Regression	98
3.1.1	MMATLA	99
3.1.2	MMATLA comparison with other robust strategies	102
3.1.3	The new proposal robustifies Univariate Regression	105
3.1.4	2 real data sets revisited	107
3.2	Multivariate Regression	108
3.2.1	Robust Multivariate Regression Algorithms	110
3.2.2	Simulation models	111
3.2.3	New proposals for Multivariate Regression	112
3.2.4	Bias and MSE tests	116
3.2.5	Finite-Sample Efficiencies	118
3.3	Regression with Correlated Variables	121
4	Using an Adaptive Trimmed Likelihood for Cluster Detection	122
4.1	Example using an artificial data set	125

4.2	Simulations involving clustered data	127
4.2.1	Relaxing breakdown restrictions	131
4.3	Example using real data	134
5	Other Diagnostics	137
5.1	Principal Components Analysis	137
5.1.1	New PCA proposal and simulations	140
5.1.2	t_5 -distributed data sets	148
5.2	Discriminant Analysis	154
5.2.1	New Discriminant Analysis (DA) proposal and simulations	155
5.2.2	Examples of robustifying allocation	161
5.3	Canonical Correlation Analysis	162
6	Conclusion	169

List of Figures

1.1	Ellipse representing an equivalent statistical distance.	7
1.2	Ellipse's delineating regions of equivalent probability.	8
1.3	Huber Minimax	12
1.4	Hampel's Psi function	13
1.5	Single outlier displaced $d = 2\sqrt{\chi_{0.975,2}^2}$	27
1.6	Single outlier displaced $d = 4\sqrt{\chi_{0.975,2}^2}$	27
1.7	Thirty outliers displaced about a mean $d = 2\sqrt{\chi_{0.975,2}^2}$ from underlying centroid.	27
1.8	Thirty outliers displaced about a mean $d = 4\sqrt{\chi_{0.975,2}^2}$ from underlying centroid.	27
2.1	$n = 100, \epsilon = 0.1, d = 4\sqrt{\chi_{0.975,2}^2}$	60
2.2	$n = 100, \epsilon = 0.3, d = 4\sqrt{\chi_{0.975,2}^2}$	60
2.3	$n = 100, \epsilon = 0.1, d = 4\sqrt{\chi_{0.975,2}^2}$	60
2.4	$n = 100, \epsilon = 0.3, d = 4\sqrt{\chi_{0.975,2}^2}$	60
2.5	$n = 100, \epsilon = 0.1, d = 2\sqrt{\chi_{0.975,2}^2}$	61
2.6	$n = 100, \epsilon = 0.3, d = 2\sqrt{\chi_{0.975,2}^2}$	61

2.7	Bivariate Cauchy.	63
2.8	Trivariate Cauchy.	63
2.9	Bivariate t_3 -distributed data.	64
2.10	Trivariate t_3 -distributed data.	64
2.11	$p = 20$ dimensional, t_{10} -distributed data.	65
2.12	$\rho_{12} = \rho_{21} \approx -0.95$	67
2.13	$\rho_{12} = \rho_{21} \approx -0.50$	67
2.14	$\rho_{12} = \rho_{21} \approx +0.50$	68
2.15	$\rho_{12} = \rho_{21} \approx +0.95$	68
2.16	$\rho_{12} = \rho_{21} \approx 0$	69
2.17	One outlier no trimming, $n = 100$, $p = 3$	69
2.18	T2 vs Fixed Threshold $n = 100$, $p = 3$, $\epsilon = 0.01$	71
2.19	T2 vs Fixed Threshold $n = 100$, $p = 3$, $\epsilon = 0.1$	71
2.20	T2 vs Fixed Threshold $n = 100$, $p = 3$, $\epsilon = 0.3$	72
2.21	T2 vs Fixed Threshold $n = 500$, $p = 10$, $\epsilon = 0.002$	72
2.22	T2 vs Fixed Threshold $n = 500$, $p = 10$, $\epsilon = 0.1$	72
2.23	T2 vs Fixed Threshold $n = 500$, $p = 10$, $\epsilon = 0.3$	72
2.24	T2 vs Fixed Threshold $n = 100, 200, \dots, 1000$, $p = 3$, $\epsilon = 0$	73
2.25	T2 vs FT3 $n = 100, 200, \dots, 1000$, $p = 3$, $\epsilon = 0$	73

2.26	T2 vs Fixed Threshold $n = 100, p = 2, 3, \dots, 10, \epsilon = 0.$	74
2.27	T2 vs FT3 $n = 100, p = 2, 3, \dots, 10, \epsilon = 0.$	74
2.28	T2 vs FT3 $n = 100, p = 3, \epsilon_d = 0.2, \epsilon_{d/2} = 0.2.$	74
2.29	T2 vs FT3 $n = 500, p = 10, \epsilon_{pth} = 0.2, \epsilon_{(p-1)th} = 0.2.$	74
2.30	T2 vs FT3 $n = 50, p = 10, \epsilon = 0.02, 0.1, 0.3.$	75
2.31	$S_{\min_i(m_i)} \neq S_{m_j}$ $n = 100, p = 3, \epsilon_d = 0.2, \epsilon_{d/2} = 0.2.$	76
2.32	$S_{\min_i(m_i)} \neq S_{m_j}$ $n = 500, p = 10, \epsilon_{pth} = 0.2, \epsilon_{(p-1)th} = 0.2.$	76
2.33	Determinant vs Trace $n = 100, p = 3, d = 0, \dots, 20.$	77
2.34	Acorn data set.	82
2.35	Minima occurring.	82
2.36	CEO data set.	83
2.37	Football's kicked data set.	83
2.38	Massachusetts lunatics 1854.	84
2.39	Minimum occurring.	84
2.40	Quarterback data set.	86
2.41	Babe Ruth data set.	86
2.42	Breast Cancer data set.	86
2.43	New York Police data set.	86
2.44	State Spending data set.	87

2.45	Minimum occurring.	87
2.46	Teachers Pay data set.	87
2.47	Minimum occurring.	87
2.48	TV adds data set.	89
2.49	Multiple minima occurring.	89
2.50	Wages hours perspective 1.	90
2.51	Wages hours perspective 2.	90
2.52	Wages hours perspective 3.	90
2.53	Wages hours perspective 4.	90
2.54	Wages hours perspective 5.	90
2.55	Wages hours perspective 6.	90
2.56	Wages Hours Minima.	91
2.57	Size of (2.7) for subsets chosen by Forward Search.	94
2.58	Excerpt of Figure 2.57 confirming minimum when Bradman's figures expelled.	94
2.59	Innings, Fifties, Runs (1).	95
2.60	Innings, Fifties, Runs (2).	95
2.61	Fifties, Hundreds, Runs (1).	95
2.62	Fifties, Hundreds, Runs (2).	95
2.63	Minimum when Bradman expelled.	96

2.64	Minimum when Bradman expelled.	96
2.65	Runs vs Fifties	96
2.66	(2.7) minimized at $\alpha = 1/90$ when Bradman removed.	96
3.1	Tukey psi function	100
3.2	Simple Regression Low Leverage.	104
3.3	Simple Regression High Leverage.	104
3.4	Multiple Regression Low Leverage.	104
3.5	Multiple Regression High Leverage.	104
3.6	Multiple MMR Regression Low Leverage	107
3.7	Multiple MMR Regression High Leverage	107
3.8	Method A on Salinity.	108
3.9	Method A on Wood Specific Gravity	108
3.10	Diagnostic plots for three contamination levels.	114
3.11	Outlier Level CL2	116
3.12	Outlier Level CL3	116
3.13	Slope MSE CL2	118
3.14	Slope MSE CL3	118
3.15	Intercept MSE CL2	119
3.16	Intercept MSE CL3	119

4.1	3 dimensional perspective showing one outlying cluster.	126
4.2	Perspective revealing exact cluster configuration.	126
4.3	First application.	126
4.4	Second application after cleaning sample.	126
4.5	3 dimensional C622 perspective showing no obvious clustering.	128
4.6	C622 perspective revealing cluster configuration.	128
4.7	C622 first application.	128
4.8	C622 second application after cleaning sample.	128
4.9	3 dimensional C631 perspective showing no obvious clustering.	129
4.10	C631 perspective revealing cluster configuration.	129
4.11	First application.	129
4.12	Second application after cleaning sample.	129
4.13	C532 detection rates.	131
4.14	C541 detection rates.	131
4.15	C433 perspective showing no obvious clustering.	135
4.16	Perspective showing clusters.	135
4.17	C433 First application of T2 detects a minor cluster.	135
4.18	Second application of T2 revealing other two clusters.	135
4.19	C433 First application of T2 isolates main cluster.	136

4.20	Second application after loosening breakdown restrictions.	136
4.21	Cars perspective exposing planted outlier.	136
4.22	Cars perspective exposing outlying cluster.	136
4.23	Multiple Minima	136
4.24	Stray point removed.	136
5.1	Proportion of variability, $n = 100$, $p = 4$, $\epsilon = 1/n$	142
5.2	Proportion of variability, $n = 100$, $p = 4$, $\epsilon = 0.1$	142
5.3	Proportion of variability $n = 100$, $p = 4$, $\epsilon = 0.2$	143
5.4	Maximum angle $n = 100$, $p = 4$, $\epsilon = 0.01$	147
5.5	Maximum angle $n = 100$, $p = 4$, $\epsilon = 0.1$	147
5.6	Maximum angle $n = 100$, $p = 4$, $\epsilon = 0.2$	147
5.7	Proportion of variability explained $n = 20$ 1 outlier.	151
5.8	Maximum angle $n = 20$ 1 outlier.	151
5.9	Proportion of variability explained $n = 50$ 1 outlier.	151
5.10	Maximum angle $n = 50$ 1 outlier.	151
5.11	Proportion of variability explained $n = 100$ 1 outlier.	151
5.12	Maximum angle $n = 100$ 1 outlier.	151
5.13	Proportion of variability explained $n = 20$ 2 outliers.	152
5.14	Maximum angle $n = 20$ 2 outliers.	152

5.15	Proportion of variability explained $n = 50$ 5 outliers.	152
5.16	Maximum angle $n = 50$ 5 outliers.	152
5.17	Proportion of variability explained $n = 100$ 10 outliers.	152
5.18	Maximum angle $n = 100$ 10 outliers.	152
5.19	Proportion of variability explained $n = 20$ 4 outliers.	153
5.20	Maximum angle $n = 20$ 4 outliers.	153
5.21	Proportion of variability explained $n = 50$ 10 outliers.	153
5.22	Maximum angle $n = 50$ 10 outliers.	153
5.23	Proportion of variability explained $n = 100$ 20 outliers.	153
5.24	Maximum angle $n = 100$ 20 outliers.	153
5.25	MP1 case D2	160
5.26	MP2 case D2	160
5.27	MP3 case D2	160
5.28	MP case D2	160
5.29	CCA comparisons for $\tilde{\Sigma} = 10\mathbf{I}_p, \dots, 100\mathbf{I}_p$	168
5.30	Magnified version of Figure 5.29.	168

List of Tables

1.1	subset count	21
1.2	Results of simulations using Rousseeuw and van Zomeren (1990) algorithm.	31
1.3	Results of simulations using Hadi (1992,1994) algorithm.	31
1.4	Results of simulations using Rocke and Woodruff (1996) algorithm	35
1.5	Silhouette cutoffs for K-means.	35
1.6	Simulation results using K-means.	35
1.7	Single linkage vs complete linkage cluster identification.	39
1.8	Simulation results using the Agglomerative Hierarchical single linkage algorithm.	39
1.9	K-means + MINO + iterative EM-algorithm	43
1.10	K-means + MINO + EM-algorithm: The success rate at determining cluster structure.	45
1.11	K-means + Silhouettes + MINO + iterative EM-algorithm.	45
2.1	Establishing T2 cut-off sample size.	54

2.2	Simulation results for sole outlier.	55
2.3	Simulation results one outlying cluster.	56
2.4	Simulation results for one cluster of Point Mass outliers.	56
2.5	Simulation results for two outlying clusters.	57
2.6	Frequency of multiple minima.	59
2.7	t_1 data.	65
2.8	t_3 data.	65
2.9	t_{10} data.	65
2.10	$\rho_{12} = \rho_{21} = \boldsymbol{\rho}$	80
2.11	Errors of location and scatter estimates for shifted normal.	80
2.12	Errors of location and scatter estimates for amplified variance.	81
2.13	Errors in Cauchy estimation with respect to Cauchy MLE.	82
2.14	Top 90 Australian and English batsmen.	93
3.1	Comparison of MMATLA results with Rousseeuw and Leroy (1987).	101
3.2	Results of MMATLA simulations.	103
3.3	Simulation results for MMATLA, method A , B and C applied to Multiple Regression models.	106
3.4	Outlier detection accuracy using R1 and R2 , $p = q = 4$	113
3.5	Method R1 $p=4$, $q=4$	117

3.6	Method R2 $p=4, q=4$	117
3.7	Clean data, no trimming algorithm applied $p=4, q=4$	117
3.8	Method R1 $p=4, q=4$	120
3.9	Method R2 $p=4, q=4$	120
3.10	Clean data sets, no trimming algorithm imposed, $p=4, q=4$	120
3.11	$n = 100, p = 4, q = 4$, Regression with Correlated Variables.	121
4.1	Sample types C622 _{100,3} and C631 _{100,3}	127
4.2	Cluster detection proportions.	130
4.3	Sample types C532 _{500,5} and C541 _{500,5}	130
4.4	Cluster detection proportions.	130
4.5	Sample types C433 _{00,3} and C55 _{100,3}	133
4.6	Simulation results comparing different T2 Forward Search starting points.	134
5.1	Expected proportion of variability explained 0.9333.	144
5.2	Expected proportion of variability explained 0.9333.	144
5.3	Expected proportion of variability explained 0.9333.	146
5.4	Average maximum angle	146
5.5	Average maximum angle	146
5.6	Average maximum angle	149
5.7	Expected proportion of variability explained 0.9000.	149

5.8	Average maximum angle.	149
5.9	Results of simulations for t_5 data sets of size $n = 100$ and dimension $p = 10$	150
5.10	Sample types used for DA simulations.	157
5.11	DA misclassification probabilities.	163
5.12	Group sizes at three stages of allocation.	163
5.13	CCA simulation results $MSE(\rho)$	167
5.14	CCA simulation results $MSE(\mathbf{a})$	167
5.15	CCA results $MSE(\mathbf{b})$	167