

ROBUST STATISTICS

Theory and Methods

Ricardo A. Maronna – *Universidad Nacional de La Plata, Argentina*

R. Douglas Martin – *University of Washington, Seattle, USA*

Victor J. Yohai – *University of Buenos Aires, Argentina*

Classical statistical techniques fail to cope well with deviations from a standard distribution. Robust statistical methods take into account these deviations while estimating the parameters of parametric models, thus increasing the accuracy of the inference. Research into robust methods is flourishing, with new methods being developed and different applications considered.

Robust Statistics sets out to explain the use of robust methods and their theoretical justification. It provides an up-to-date overview of the theory and practical application of robust statistical methods in regression, multivariate analysis, generalized linear models and time series. This unique book:

- Enables the reader to select and use the most appropriate robust method for their particular statistical model.
- Features computational algorithms for the core methods.
- Covers regression methods for data mining applications.
- Includes examples with real data and applications using the S-Plus robust statistics library.
- Describes the theoretical and operational aspects of robust methods separately, so the reader can choose to focus on one or the other.
- Is supported by a supplementary website featuring datasets.

Robust Statistics aims to stimulate the use of robust methods as a powerful tool to increase the reliability and accuracy of statistical modelling and data analysis. It is ideal for researchers, practitioners and graduate students of statistics, electrical, chemical and biochemical engineering, and computer vision. There is also much to benefit researchers from other sciences, such as biotechnology, who need to use robust statistical methods in their work.

 **WILEY**
wiley.com

ISBN 978-0-470-01092-1



9 780470 010921

Contents

Preface	xv
1 Introduction	1
1.1 Classical and robust approaches to statistics	1
1.2 Mean and standard deviation	2
1.3 The “three-sigma edit” rule	5
1.4 Linear regression	7
1.4.1 Straight-line regression	7
1.4.2 Multiple linear regression	9
1.5 Correlation coefficients	11
1.6 Other parametric models	13
1.7 Problems	15
2 Location and Scale	17
2.1 The location model	17
2.2 M-estimates of location	22
2.2.1 Generalizing maximum likelihood	22
2.2.2 The distribution of M-estimates	25
2.2.3 An intuitive view of M-estimates	27
2.2.4 Redescending M-estimates	29
2.3 Trimmed means	31
2.4 Dispersion estimates	32
2.5 M-estimates of scale	34
2.6 M-estimates of location with unknown dispersion	36
2.6.1 Previous estimation of dispersion	37
2.6.2 Simultaneous M-estimates of location and dispersion	37
2.7 Numerical computation of M-estimates	39
2.7.1 Location with previously computed dispersion estimation	39
2.7.2 Scale estimates	40
2.7.3 Simultaneous estimation of location and dispersion	41

2.8	Robust confidence intervals and tests	41
2.8.1	Confidence intervals	41
2.8.2	Tests	43
2.9	Appendix: proofs and complements	44
2.9.1	Mixtures	44
2.9.2	Asymptotic normality of M-estimates	45
2.9.3	Slutsky's lemma	46
2.9.4	Quantiles	46
2.9.5	Alternative algorithms for M-estimates	46
2.10	Problems	48
3	Measuring Robustness	51
3.1	The influence function	55
3.1.1	*The convergence of the SC to the IF	57
3.2	The breakdown point	58
3.2.1	Location M-estimates	58
3.2.2	Scale and dispersion estimates	59
3.2.3	Location with previously computed dispersion estimate	60
3.2.4	Simultaneous estimation	60
3.2.5	Finite-sample breakdown point	61
3.3	Maximum asymptotic bias	62
3.4	Balancing robustness and efficiency	64
3.5	*"Optimal" robustness	65
3.5.1	Bias and variance optimality of location estimates	66
3.5.2	Bias optimality of scale and dispersion estimates	66
3.5.3	The infinitesimal approach	67
3.5.4	The Hampel approach	68
3.5.5	Balancing bias and variance: the general problem	70
3.6	Multidimensional parameters	70
3.7	*Estimates as functionals	71
3.8	Appendix: proofs of results	75
3.8.1	IF of general M-estimates	75
3.8.2	Maximum BP of location estimates	76
3.8.3	BP of location M-estimates	76
3.8.4	Maximum bias of location M-estimates	78
3.8.5	The minimax bias property of the median	79
3.8.6	Minimizing the GES	80
3.8.7	Hampel optimality	82
3.9	Problems	84
4	Linear Regression 1	87
4.1	Introduction	87
4.2	Review of the LS method	91
4.3	Classical methods for outlier detection	94

4.4	Regression M-estimates	98
4.4.1	M-estimates with known scale	99
4.4.2	M-estimates with preliminary scale	100
4.4.3	Simultaneous estimation of regression and scale	103
4.5	Numerical computation of monotone M-estimates	103
4.5.1	The L1 estimate	103
4.5.2	M-estimates with smooth ψ -function	104
4.6	Breakdown point of monotone regression estimates	105
4.7	Robust tests for linear hypothesis	107
4.7.1	Review of the classical theory	107
4.7.2	Robust tests using M-estimates	108
4.8	*Regression quantiles	110
4.9	Appendix: proofs and complements	110
4.9.1	Why equivariance?	110
4.9.2	Consistency of estimated slopes under asymmetric errors	111
4.9.3	Maximum FBP of equivariant estimates	112
4.9.4	The FBP of monotone M-estimates	113
4.10	Problems	114
5	Linear Regression 2	115
5.1	Introduction	115
5.2	The linear model with random predictors	118
5.3	M-estimates with a bounded ρ -function	119
5.4	Properties of M-estimates with a bounded ρ -function	120
5.4.1	Breakdown point	122
5.4.2	Influence function	123
5.4.3	Asymptotic normality	123
5.5	MM-estimates	124
5.6	Estimates based on a robust residual scale	126
5.6.1	S-estimates	129
5.6.2	L-estimates of scale and the LTS estimate	131
5.6.3	Improving efficiency with one-step reweighting	132
5.6.4	A fully efficient one-step procedure	133
5.7	Numerical computation of estimates based on robust scales	134
5.7.1	Finding local minima	136
5.7.2	The subsampling algorithm	136
5.7.3	A strategy for fast iterative estimates	138
5.8	Robust confidence intervals and tests for M-estimates	139
5.8.1	Bootstrap robust confidence intervals and tests	141
5.9	Balancing robustness and efficiency	141
5.9.1	"Optimal" redescending M-estimates	144
5.10	The exact fit property	146
5.11	Generalized M-estimates	147
5.12	Selection of variables	150

5.13	Heteroskedastic errors	153
5.13.1	Improving the efficiency of M-estimates	153
5.13.2	Estimating the asymptotic covariance matrix under heteroskedastic errors	154
5.14	*Other estimates	156
5.14.1	τ -estimates	156
5.14.2	Projection estimates	157
5.14.3	Constrained M-estimates	158
5.14.4	Maximum depth estimates	158
5.15	Models with numeric and categorical predictors	159
5.16	*Appendix: proofs and complements	162
5.16.1	The BP of monotone M-estimates with random X	162
5.16.2	Heavy-tailed x	162
5.16.3	Proof of the exact fit property	163
5.16.4	The BP of S-estimates	163
5.16.5	Asymptotic bias of M-estimates	166
5.16.6	Hampel optimality for GM-estimates	167
5.16.7	Justification of RFPE*	168
5.16.8	A robust multiple correlation coefficient	170
5.17	Problems	171
6	Multivariate Analysis	175
6.1	Introduction	175
6.2	Breakdown and efficiency of multivariate estimates	180
6.2.1	Breakdown point	180
6.2.2	The multivariate exact fit property	181
6.2.3	Efficiency	181
6.3	M-estimates	182
6.3.1	Collinearity	184
6.3.2	Size and shape	185
6.3.3	Breakdown point	186
6.4	Estimates based on a robust scale	187
6.4.1	The minimum volume ellipsoid estimate	187
6.4.2	S-estimates	188
6.4.3	The minimum covariance determinant estimate	189
6.4.4	S-estimates for high dimension	190
6.4.5	One-step reweighting	193
6.5	The Stahel–Donoho estimate	193
6.6	Asymptotic bias	195
6.7	Numerical computation of multivariate estimates	197
6.7.1	Monotone M-estimates	197
6.7.2	Local solutions for S-estimates	197
6.7.3	Subsampling for estimates based on a robust scale	198
6.7.4	The MVE	199
6.7.5	Computation of S-estimates	199

6.7.6	The MCD	200
6.7.7	The Stahel–Donoho estimate	200
6.8	Comparing estimates	200
6.9	Faster robust dispersion matrix estimates	204
6.9.1	Using pairwise robust covariances	204
6.9.2	Using kurtosis	208
6.10	Robust principal components	209
6.10.1	Robust PCA based on a robust scale	211
6.10.2	Spherical principal components	212
6.11	*Other estimates of location and dispersion	214
6.11.1	Projection estimates	214
6.11.2	Constrained M-estimates	215
6.11.3	Multivariate MM- and τ -estimates	216
6.11.4	Multivariate depth	216
6.12	Appendix: proofs and complements	216
6.12.1	Why affine equivariance?	216
6.12.2	Consistency of equivariant estimates	217
6.12.3	The estimating equations of the MLE	217
6.12.4	Asymptotic BP of monotone M-estimates	218
6.12.5	The estimating equations for S-estimates	220
6.12.6	Behavior of S-estimates for high p	221
6.12.7	Calculating the asymptotic covariance matrix of location M-estimates	222
6.12.8	The exact fit property	224
6.12.9	Elliptical distributions	224
6.12.10	Consistency of Gnanadesikan–Kettenring correlations	225
6.12.11	Spherical principal components	226
6.13	Problems	227
7	Generalized Linear Models	229
7.1	Logistic regression	229
7.2	Robust estimates for the logistic model	233
7.2.1	Weighted MLEs	233
7.2.2	Redescending M-estimates	234
7.3	Generalized linear models	239
7.3.1	Conditionally unbiased bounded influence estimates	242
7.3.2	Other estimates for GLMs	243
7.4	Problems	244
8	Time Series	247
8.1	Time series outliers and their impact	247
8.1.1	Simple examples of outliers' influence	250
8.1.2	Probability models for time series outliers	252
8.1.3	Bias impact of AOs	256

8.2	Classical estimates for AR models	257
8.2.1	The Durbin–Levinson algorithm	260
8.2.2	Asymptotic distribution of classical estimates	262
8.3	Classical estimates for ARMA models	264
8.4	M-estimates of ARMA models	266
8.4.1	M-estimates and their asymptotic distribution	266
8.4.2	The behavior of M-estimates in AR processes with AOs	267
8.4.3	The behavior of LS and M-estimates for ARMA processes with infinite innovations variance	268
8.5	Generalized M-estimates	270
8.6	Robust AR estimation using robust filters	271
8.6.1	Naive minimum robust scale AR estimates	272
8.6.2	The robust filter algorithm	272
8.6.3	Minimum robust scale estimates based on robust filtering	275
8.6.4	A robust Durbin–Levinson algorithm	275
8.6.5	Choice of scale for the robust Durbin–Levinson procedure	276
8.6.6	Robust identification of AR order	277
8.7	Robust model identification	278
8.7.1	Robust autocorrelation estimates	278
8.7.2	Robust partial autocorrelation estimates	284
8.8	Robust ARMA model estimation using robust filters	287
8.8.1	τ -estimates of ARMA models	287
8.8.2	Robust filters for ARMA models	288
8.8.3	Robustly filtered τ -estimates	290
8.9	ARIMA and SARIMA models	291
8.10	Detecting time series outliers and level shifts	294
8.10.1	Classical detection of time series outliers and level shifts	295
8.10.2	Robust detection of outliers and level shifts for ARIMA models	297
8.10.3	REGARIMA models: estimation and outlier detection	300
8.11	Robustness measures for time series	301
8.11.1	Influence function	301
8.11.2	Maximum bias	303
8.11.3	Breakdown point	304
8.11.4	Maximum bias curves for the AR(1) model	305
8.12	Other approaches for ARMA models	306
8.12.1	Estimates based on robust autocovariances	306
8.12.2	Estimates based on memory- m prediction residuals	308
8.13	High-efficiency robust location estimates	308
8.14	Robust spectral density estimation	309
8.14.1	Definition of the spectral density	309
8.14.2	AR spectral density	310
8.14.3	Classic spectral density estimation methods	311
8.14.4	Prewhitening	312

8.14.5	Influence of outliers on spectral density estimates	312
8.14.6	Robust spectral density estimation	314
8.14.7	Robust time-average spectral density estimate	316
8.15	Appendix A: heuristic derivation of the asymptotic distribution of M-estimates for ARMA models	317
8.16	Appendix B: robust filter covariance recursions	320
8.17	Appendix C: ARMA model state-space representation	322
8.18	Problems	323
9	Numerical Algorithms	325
9.1	Regression M-estimates	325
9.2	Regression S-estimates	328
9.3	The LTS-estimate	328
9.4	Scale M-estimates	328
9.4.1	Convergence of the fixed point algorithm	328
9.4.2	Algorithms for the nonconcave case	330
9.5	Multivariate M-estimates	330
9.6	Multivariate S-estimates	331
9.6.1	S-estimates with monotone weights	331
9.6.2	The MCD	332
9.6.3	S-estimates with nonmonotone weights	333
9.6.4	*Proof of (9.25)	334
10	Asymptotic Theory of M-estimates	335
10.1	Existence and uniqueness of solutions	336
10.2	Consistency	337
10.3	Asymptotic normality	339
10.4	Convergence of the SC to the IF	342
10.5	M-estimates of several parameters	343
10.6	Location M-estimates with preliminary scale	346
10.7	Trimmed means	348
10.8	Optimality of the MLE	348
10.9	Regression M-estimates	350
10.9.1	Existence and uniqueness	350
10.9.2	Asymptotic normality: fixed X	351
10.9.3	Asymptotic normality: random X	355
10.10	Nonexistence of moments of the sample median	355
10.11	Problems	356
11	Robust Methods in S-Plus	357
11.1	Location M-estimates: function <i>Mestimate</i>	357
11.2	Robust regression	358
11.2.1	A general function for robust regression: <i>lmRob</i>	358
11.2.2	Categorical variables: functions <i>as.factor</i> and <i>contrasts</i>	361

11.2.3	Testing linear assumptions: function <i>rob.linear.test</i>	363
11.2.4	Stepwise variable selection: function <i>step</i>	364
11.3	Robust dispersion matrices	365
11.3.1	A general function for computing robust location–dispersion estimates: <i>covRob</i>	365
11.3.2	The SR- α estimate: function <i>cov.SRocke</i>	366
11.3.3	The bisquare S-estimate: function <i>cov.Sbic</i>	366
11.4	Principal components	366
11.4.1	Spherical principal components: function <i>prin.comp.rob</i>	367
11.4.2	Principal components based on a robust dispersion matrix: function <i>princomp.cov</i>	367
11.5	Generalized linear models	368
11.5.1	M-estimate for logistic models: function <i>BYlogreg</i>	368
11.5.2	Weighted M-estimate: function <i>WBYlogreg</i>	369
11.5.3	A general function for generalized linear models: <i>glmRob</i>	370
11.6	Time series	371
11.6.1	GM-estimates for AR models: function <i>ar.gm</i>	371
11.6.2	F τ -estimates and outlier detection for ARIMA and REGARIMA models: function <i>arima.rob</i>	372
11.7	Public-domain software for robust methods	374
12	Description of Data Sets	377
	Bibliography	383
	Index	397