Kybernetika

ROBUST TIME SERIES ANALYSIS: A SURVEY

NORBERT STOCKINGER, RUDOLF DUTTER

ACADEMIA
PRAHA

INTRODUCTION

This monography gives a survey about recent work on robust time series analysis. After short introductions to different topics, investigations and new results are reported. Although the choice of literature is very subjective and the list of references is far from being complete, we have also put some unpublished entries which might be difficult to obtain from public libraries but they were relevant for our research.

In the first chapter we review different concepts of robustness for independently, identically distributed observations as well as for time series. Min-max robustness, efficiency robustness and qualitative robustness is treated in more details.

Consistency and robustness properties of least squares estimators of autoregressive integrated moving average parameters is dealt with, where the given time series is possibly contaminated by outliers. The least squares estimation behaves differently well in cases of considered two types of outliers. The bad performance of least squares estimators for contaminated data shows the necessity of robust estimation methods, methods which are robust toward outliers and wrong specification of the model.

Chapter II deals with definitions, computational methods and properties of maximum likelihood type estimators (M-estimators) for pure autoregressive models as well as for ARMA models. In contrast to least squares estimators, M-estimators are, in particular, efficiency robust if the given time series is contaminated by innovation outliers. Two estimation methods which can be used advantageously for time series including additive outliers, are outlined.

An appropriate generalization of the maximum likelihood type (M-)method yields more satisfactory estimates of ARMA parameters in the case that the given time series is contaminated by additive outliers. Chapter III deals with definitions, computational methods and properties of generalized maximum likelihood type estimators (GM-estimators) for pure autoregressive models as well as for ARMA models. In additive outlier situations GM-estimators have, in particular, the following properties. GM-estimators do not require independently, identically distributed outliers. GM-estimators have a positive breakdown point, a bounded influence curve, considerable robustness and much smaller bias than M-estimators and least squares estimators.

The properties of M-estimators and GM-estimators of AR parameters can be used to create tests which are able to determine the type of outliers in a time series. Robustified methods for the identification of AR models and ARIMA models are mentioned.

In order to deal with robust filtering and smoothing a vector state-variable representation of ARMA processes is described in Chapter IV. Here, a filtered value is defined to depend only on previous observations while a smoothed value is defined to depend on all given observations. A recursive algorithm for the computation of approximate conditional-mean (ACM) filters which are able to remove outliers from contaminated data, is dealt with.

Maximizing a likelihood function which is approximated (also by an ACM filter), leads to approximate maximum likelihood (AML) estimators. Proceeding further by replacing the negative of the log-likelihood by a loss function which uses a robustifying rho-function, yields approximation maximum likelihood type (AM) estimators. A relatively simple iterative scheme can be used to compute AM-estimators. Conditional-mean M-estimators can be regarded as AM-estimators especially for AR models. Other methods for robust filtering and smoothing are provided, for example, by the robustified Kalman filter, L-smoothers, moving M-estimate smoothers and robustified splines.

Chapter V presents a Monte Carlo investigation of methods for the least squares estimation, M-estimation and GM-estimation of ARMA models. Monte Carlo generally reveals properties which are expected from theory. For outlier-free data the means of the estimated parameters differ scarcely, and the mean square errors of M-estimators and GM-estimators are larger than those for least squares estimators. For the processes chosen here, with innovation outliers, the means of the estimated parameters also differ only slightly, but the sample relative efficiencies of M-estimators are larger than the sample relative efficiencies of GM-estimators and of least squares estimators. In the presence of additive outliers the GM-estimation essentially yields better parameters and substantially smaller mean square errors than the least squares estimation and than the M-estimation.

Several topics for further research concerning identification and estimation of various models, outlier detection, filters and spectral density estimation are discussed.

ACKNOWLEDGEMENTS

We like to thank Peter J. Huber and Frank R. Hampel for introducing us to the challenging field of robust statistics. We are grateful to R. D. Martin (who is the "main contributor" of this manuscript), V. J. Yohai, O. H. Bustos and J. E. Zeh for many stimulating discussions, to Karl Pfeiffer for stimulating the application of robust techniques and to E. Stadlober for the disposition of computer programs for generating pseudo random numbers.

Some results are taken from the doctoral dissertation of the first author. The research was partially supported by the "Fonds zur Förderung der wissenschaftlichen Forschung", project no. 4487 and 4972. Finally, the comments of the referee and the excellent collaboration of the editors has been highly appreciated.

BIBLIOGRAPHY

Abraham B. and G. E. P. Box (1979): Bayesian Analysis of Some Outlier Problems in Time Series. Biometrika *66*, 229–236.

Ahrens J. H. and U. Dieter (1974): Computer Methods for Sampling from Gamma, Beta, Poisson and Binomial Distributions. Computing *12*, 223–246.

Akaike H. (1969): Power Spectrum Estimation through Autoregressive Model Fitting. Ann. Inst. Statist. Math. *21*, 407–419.

Akaike H. (1974): A New Look at the Statistical Model Identification. IEEE Trans. Automat. Control. *AC-19*, 716–722.

Anderson T. W. (1971): The Statistical Analysis of Time Series. John Wiley, New York.

Andrews D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers and J. W. Tukey (1972): Robust Estimates of Location; Survey and Advances. Princeton University Press, Princeton.

Beaton A. E. and J. W. Tukey (1974): The Fitting of Power Series, Meaning Polynomials, Illustrated on Bandspectroscopic Data. Technometrics *16*, 2, 147–197.

Boente G., R. Fraiman and V. J. Yohai (1982): Qualitative Robustness for General Stochastic Processes. Techn. Rep. 26, Dept. Statist., Univ. Washington, Seattle.

Box G. E. P. and G. M. Jenkins (1976): Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Box. G. E. P. and G. C. Tiao (1975): Intervention Analysis with Applications to Economic and Environmental Problems. J. Amer. Statist. Assoc. *70*, 70–79.

Brubacher S. R. (1974): Time Series Outlier Detection and Modeling with Interpolation. Bell Laboratories Techn. Mem.

Bustos O. H. (1981): Qualitative Robustness for General Processes. Informes de Matematica, Serie B-002/81, Instituto de Matematica Pura e Aplicada, Rio de Janeiro.

Bustos O. H. (1982): General M-Estimates for Contaminated pth-Order Autoregressive Processes: Consistency and Asymptotic Normality. Z. Wahrsch. verw. Gebiete *59*, 491–504.

Bustos O., R. Fraiman and V. J. Yohai (1984): Asymptotic Behavior of Estimates Based on Residual Autocovariances for ARMA Models. Informes de Matematica, Serie B-019-Junho/84, Instituto de Matematica Pura e Aplicada, Rio de Janeiro. To appear in Proc. Heidelberg Workshop on Robust and Nonlinear Time Series, Sept. 1983.

Bustos O. H. and V. J. Yohai (1983): Robust Estimates for ARMA Models. Manuscript. Submitted to J. Amer. Statist. Soc.

Chernick M. R., D. J. Downing and D. H. Pike (1982): Detecting Outliers in Time Series Data. J. Amer. Statist. Assoc. *77*, 380.

Cleveland W. S. (1979): Robust Locally Weighted Regression and Smoothing Scatterplots. J. Amer. Statist. Assoc. *74*, 368, 829–836.

Cleveland W. S. (1982): A Reader's Guide to Smoothing Scatterplots and Graphical Methods for Regression. In: Modern Data Analysis (Launer and Siegel, eds.) Acad. Press, New York.

Cox D. D. (1981): Metrics on Stochastic Processes and Qualitative Robustness. Techn. Rep. 3, Dept. Statist., Univ. Washington, Seattle.

Denby L. and W. Larsen (1977): Robust Regression Estimators Compared via Monte Carlo. Commun. Statist. *A 6*, 4, 335–362.

Denby L. and R. D. Martin (1979): Robust Estimation of the First-Order Autoregressive Parameter. J. Amer. Statist. Assoc. *74*, 365, 140–146.

Devlin S. J., R. Gnanadesikan and J. R. Kettenring 1975): Robust Estimation and Outlier Detection with Correlation Coefficients. Biometrika *62*, 531–545.

Donoho D. L. and P. J. Huber (1983): The Notion of Breakdown Point. In: Festschrift for Erich L. Lehmann (Bickel et al. eds.). Wadsworth, Belmont, CA.

Durbin J. (1959): Efficient Estimation of Parameters in Moving Average Models. Biometrika *46*, 306—316.

Dutter R. (1975): Robust Regression: Different Approaches to Numerical Solutions and Algorithms. Res. Rep. 6, Fachgruppe f. Statist., Eidgen. Techn. Hochsch., Zürich.

Dutter R. (1980): Robuste Regression. Bericht 135, Math. Statist. Sektion im Forschungszentrum Graz.

Dutter R. (1983): COVINTER: A Computer Program for Computing Robust Covariances and for Plotting Confidence Ellipses. Res. Rep. 10, Inst. for Statist., Techn. Univ. Graz.

Dutter R. (1983b): Computer Program BLINWDR for Robust and Bounded Influence Regression Res. Rep. 8, Inst. Statist., Techn. Univ. Graz.

Dutter R. and P. J. Huber (1981): Numerical Methods for the Nonlinear Robust Regression Problem. J. Statist. Comput. Simul. *13*, 2, 79—114.

Fox A. J. (1972): Outliers in Time Series. J. Roy. Statist. Soc. B, *34*, 3, 350—363.

Fuller W. A. (1976): Introduction to Statistical Time Series. John Wiley, New York.

Gastwirth J. L. and H. Rubin (1975): The Behavior of Robust Estimators on Dependent Data. Ann. Statist *3*, 5, 1070—1100.

Grenander V. and M. Rosenblatt (1957): Statistical Analysis of Stationary Time Series. John Wiley, New York.

Hampel F. R. (1968): Contributions to the Theory of Robust Estimation. Ph. D. Thesis. University of California, Berkeley.

Hampel F. R. (1971): A General Qualitative Definition of Robustness. Ann. Math. Statist. *42*, 6, 1887—1896.

Hampel F. R. (1973): Robust Estimation: A Condensed Partial Survey. Z. Wahrsch. verw. Gebiete *27*, 87—104.

Hampel F. R. (1974): The Influence Curve and Its Role in Robust Estimation. J. Amer. Statist. Assoc. *69* 346, 382—393.

Hampel F. R. (1975): Beyond Location Parameters: Robust Concepts and Methods. ISI Invited Paper, Proceedings of the 40th Session, Vol. XLVI, Book I, 375—382. Warsaw.

Hampel F. R., A. Marazzi, E. Ronchetti, P. Rousseeuw, W. Stahel and R. E. Welsch (1982): Robust Statistical Methods, Handouts for the Instructional Meeting on. Part IV. Palermo, Italy, Sept. 10—11, 1982.

Hampel F. R., W. A. Stahel, E. M. Ronchetti and P. J. Rousseeuw (1986): Robust Statistics: The Approach Based on Influence Functions. John Wiley, New York.

Hannan E. J. (1970): Multiple Time Series. John Wiley, New York.

Hannan E. J. (1973): The Asymptotic Theory of Linear Time Series Models. J. Appl. Prob. *10*, 130—145.

Hannan E. H. and M. Kanter (1977): Autoregressive Processes with Infinite Variance. J. Appl. Prob. *14*, 411—415.

Huang T. S., G. J. Yang and G. Y. Tang (1979): A Fast Two-Dimensional Median Filtering Algorithm. IEEE Trans. Acoust. Speech Signal Process. *27*, 1, 13—18.

Huber P. J. (1964): Robust Estimation of a Location Parameter. Ann. Math. Statist. *35*, 1, 73—101.

Huber P. J. (1973): Robust Regression: Asymptotics, Conjectures and Monte Carlo. Ann. Statist. *1*, 5, 799—821.

Huber P. J. (1977): Robust Covariances. In: Statistical Decision Theory and Related Topics (Gupta S. and D. Moore, eds.), Vol. II. Academic Press, New York.

Huber P. J. (1979): Robust Smoothing. In: Robustness in Statistics (Launer and Wilkinson, eds.). Academic Press, New York.

Huber P. J. (1981): Robust Statistics. John Wiley, New York.

Huber P. J. (1982): Current Issues in Robust Statistics. In: Some Recent Advances in Statistics (Tiago de Oliveira and Epstein, eds.) Acad. Press, New York.

Jazwinski A. (1970): Stochastic Processes and Filtering Theory. Acad. Press, New York.

Jones R. H. (1980): Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations. Technometrics *22*, 3.

Justusson B. (1977): Statistical Properties of Median Filters in Signal and Image Processing. Unpubl. Rep., Math. Institut., Royal Instit. of Techn., Stockholm. Sweden.

Kailath T. (1968): An Innovations Approach to Linear Least Squares Estimation and Filtering. IEEE Trans. Automat. Control. *AC-13*, 6, 646—655.

Kanter M. and W. L. Steiger (1974): Regression and Autoregression With Infinite Variance. Adv. in Appl. Prob. *6*, 768—783.

Kassam S. A. and H. V. Poor (1985): Robust Techniques for Signal Processing: A Survey. Proc. IEEE *73*, 3.

Kleiner R., R. D. Martin and D. J. Thomson (1979). Robust Estimation of Power Spectra. J. Royal Statist. Soc. B *41*, 3, 313—338.

Krasker W. S. and R. E. Welsch (1982): Efficient Bounded-Influence Regression Estimation. J. Amer. Statist. Soc. *77*, 379, 595—604.

Kuensch H. (1983a): Infinitesimal Robustness for Autoregressive Processes. Res. Rep. 38, Fachgr. Statistik, ETH Zürich.

Kuensch H. (1983b): The Influence Function and Optimal Robust Estimators for Time Series. Subm. to Ann. Statist.

Lee C. H. (1981): M-Estimates for ARMA Processes. Ph. D. Thesis. Dept. Elect. Engin., Univ. Washington. Seattle.

Lee C. H. and R. D. Martin (1982): M-Estimates for ARMA Processes. Techn. Rep. 23, Dept. Statist., Univ. Washington, Seattle.

Lee C. H. and R. D. Martin (1982b): The Information Matrix and Robust M-Estimates for ARMA Processes. Techn. Rep. 24, Dept. Statist., Univ. Washington, Seatle.

Lee C. H. and R. D. Martin (1984): Ordinary and Proper Location M-Estimates for ARMA Models. Techn. Rep. 29, Dept. Statist., Univ. Washington, Seattle.

Lenth R. V. (1977): Robust Splines. Commun. Statist. *A 6*, 847—854.

Mallows C. L. (1976): On Some Topics in Robustness. Bell Laboratories, Techn, Memo., Murray Hill, New Jersey.

Mallows C. L. (1980a): Resistant Smoothing. In: Time Series (Anderson O. D., ed.). North-Holland Publishing Company.

Mallows C. L. (1980b): Some Theory of Nonlinear Smoothers. Ann. Statist. *8*, 4, 695—715.

Mann H. B. and A. Wald (1943): On the Statistical Treatment of Linear Stochastic Difference Equations. Econometrica *11*, 173—220.

Marazzi A. (1980): Robust Affine Invariant Covariances in ROBETH. Res. Rep. 24, Eidgenössische Techn. Hochschule, Zürich.

Maronna R. (1976): Robust M-Estimation of Multivariate Location and Scatter. Ann. Statist. *4* 1, 51—67.

Maronna R., O. Bustos and V. J. Yohai (1979): Bias- and Efficiency-Robustness of General M-Estimators for Regression with Random Carriers. In: Smoothing Techniques for Curve Estimation — Proc. Heidelberg (Gasser Th. and M. Rosenblatt, eds.). Lecture Notes in Math. 757, Springer, Berlin.

Martin R. D. (1978a): Asymptotic Properties of M-estimates for *p*th-order Autoregressions. Techn. Rep. 212, Dept. Electrical Engineering, Univ. Washington, Seattle.

Martin R. D. (1978b): Robust Estimates of the Mean with Autoregressive Errors. Techn. Rep. 211, Dept. Electrical Engineering, Univ. Washington, Seattle.

Martin R. D. (1978c): Asymptotic Properties of Generalized M-estimates for Autoregressive Parameters. Techn. Rep. 213, Dept. Elec. Engng., Univ. Washington, Seattle.

Martin R. D. (1979): Robust Estimation for Time Series Autoregressions. In: Robustness in Statistics (Launer and Wilkinson, eds.) Acad. Press, New York.

Martin R. D. (1979b): Robust Estimation of Location with Autoregressive Errors. Unpubl. manuscript, Dept. Electrical Engineering, Univ. Washington, Seattle.

Martin R. D. (1979c): Approximate Conditional-Mean Type Smoothers and Interpolators. In: Smoothing Techniques for Curve Estimation — Proc. Heidelberg, 1979. (Gasser and Rosenblatt, eds.). Springer-Verlag. New York.

Martin R. D. (1980): Robust Estimation of Autoregressive Models. In: Directions in Time Series. (Brillinger D. R. and G. C. Tiao, eds.). Inst. Math. Statist. Publications, Haywood, CA, pp. 228—254.

Martin R. D. (1980b): Time Series: Model Estimation, Data Analysis and Robust Procedures. In: Proc. Symp. Appl. Math., (R. V. Hogg, ed.), Vol. 23.

Martin R. D. (1981): The Cramer-Rao Bound and Robust M-Estimates for Autoregressions. Techn. Rep. 9, Dept. Statist., Univ. Washington, Seattle.

Martin R. D. (1981b): Robust Methods for Time Series. In: Applied Time Series II (Findley, ed.). Acad. Press, New York.

Martin R. D. (1982): The Cramer-Rao Bound and Robust M-Estimates for Autoregressions. Biometrika *69*, 2, 437—442.

Martin R. D. (1983): Robust-Resistant Spectral Analysis. In: Handbook of Statistics. Vol. 3 (Brillinger and Krishnaiah, eds.). Elsevier Sc. Publ. B. V.

Martin R. D. (1984): Robust-Resistant Spectral Analysis. Techn. Rep. 27, Dept. Statist., Univ. Washington, Seattle.

Martin R. D. and G. De Bow (1976): Robust Filtering with Data-Dependent Covariance. Proc. John Hopkins Conference and Informations Sciences and Systems, March 31 — April 2.

Martin R. D. and J. M. Jong (1976): Asymptotic Properties of Robust Generalized M-Estimates for the First-Order Autoregressive Parameter. Bell Laboratories Techn. Memo., Murray Hill, New Jersey.

Martin R. D. and C. H. Lee (1980): Robust Estimation of Location with Autoregressive Errors. Manuscript. Dept. Statist., Univ. Washington, Seattle.

Martin R. D., A. Samarov and W. Vandaele (1983): Robust Methods for ARIMA Models. In: Applied Time Series Analysis of Economic Data (Zellner, ed.). Econ. Res. Rep. ER-5, Bureau of the Census. Washington, DC.

Martin R. D. and D. J. Thomson (1982): Robust-Resistant Spectrum Estimation. Proc. IEEE *70*, 9.

Martin R. D. and V. J. Yohai (1984a): Robustness in Time Series and Estimating ARMA Models. Techn. Rep. 50, Dept. Statist., Univ. Washington. Seattle.

Martin R. D. and V. J. Yohai (1948b): Influence Curves for Time Series. Techn. Rep. 51, Dept. Statist., Univ. Washington. Seattle.

Martin R. D. and J. E. Zeh (1977): Determining the Character of Time Series Outliers. Proc. Amer. Statist. Assoc., Business and Economics Section.

Martin R. D. and J. E. Zeh (1979): Generalized M-estimates for Autoregressions, Including Small-Sample Efficiency Robustness. Techn. Rep. 214, Department of Electrical Engineering, University of Washington, Seattle.

Masreliez C. J. (1975): Approximate Non-Gaussian Filtering with Linear State and Observation Relations. IEEE Trans. Automat. Control *AC-20*, 107—110.

Masreliez C. J. and R. D. Martin (1977): Robust Bayesian Estimation for the Linear Model and Robustifying the Kalman Filter. IEEE Trans. Automat. Control *AC-22*, 361—371.

Mosteller F. and J. W. Tukey (1977): Data Analysis and Regression. Addison-Wesley, Reading, MA.

Nagel G. and W. Wolff (1974): Ein Verfahren zur Minimierung einer Quadratsumme nichtlinearer Funktionen. Biometrische Zeitschr. *16*, 6, 431—439.

Pagano M. (1974): Estimation of Models of Autoregressive Signal Plus White Noise. Ann. Statist. *2*, 99—108.

Papantoni-Kazakos P. and R. M. Gray (1979): Robustness of Estimators of Stationary Observations. Ann. Probab. *7*, 6, 989—1002.

Parzen E. (1971): Efficient Estimation of Stationary Time Series Mixed Schemes. Proc. 39th Session of ISI, Washington, D. C.

Parzen E. (1974): Some Recent Advances in Time Series Modelling. IEEE Trans. Automat. Control *AC-19*, 723—730.

Polasek W. (1982): Robust Estimation and Resistance Analysis for the Autocorrelation Function. Preprint 47, Univ. Vienna.

Polasek W. (1982b). Exploratory Business-cycle Analysis Using Running Medians. Empirica *1*, 49—70.

Polasek W. and R. Mertl (1983): Robust and Jackknife Estimators of the Autocorrelation Function. Res. Rep., Inst. Statist. and Informatics, Univ. Vienna.

Rabiner L. R., M. R. Sambur and C. E. Schmidt (1975): Applications of a Nonlinear Smoothing Algorithm to Speech Processing. IEEE Trans. Acoust. Speech Signal Process *ASSP-23*, 552—557.

Reinisch C. H. (1967): Smoothing by Spline Functions. Numer. Mathem. *10*, 177—183, and Numer. Mathem. *16*, 451—454.

Relles D. A. (1968): Robust Regression by Modified Least Squares. Ph. D. Thesis. Dept. Statist., Yale University.

Rieder H. (1980): Locally Robust Correlation Coefficients. Commun. Statist. *A9*, 8, 803—819.

Schweppe F. (with E. Handschin, J. Kohlas and A. Fiechter) (1975): Bad Data Analysis for Power System State Estimation. IEEE Trans. Power Apparatus Systems *94*, 2, 329—337.

Serfling R. J. (1980): Approximation Theorems of Mathematical Statistics. John Wiley, New York.

Shibata R. (1976): Selection of the Order of an Autoregressive Model by Akaike's Information Criterion. Biometrika *63*, 1, 117—128.

Siddiqui M. M. (1958): On the Inversion of the Sample Covariance Matrix in a Stationary Autoregressive Process. Ann. Math. Statist. *29*, 585—588.

Stadlober E. and U. Dieter (1985): Computer Methods for Generating Student *t*-Variates. To appear in Computing.

Stockinger N. (1983): Robust Estimation of Autoregressive Moving Average Models. In: Proc. of the 4th Pannonian Symposium on Math. Statist., Bad Tatzmannsdorf, Austria (Grossmann, Konecny, Pflug, Vincze and Wertz, eds.), pp. 299—309. Reidel Publ. Comp., Dordrecht-Holland.

Stockinger N. (1984): Detection of Outliers in Arrhythmic Pressure Pulses by Robust Methods of Time Series Analysis. Res. Rep. 11, Inst. Statist., Techn. Univ. Graz.

Stockinger N. (1985a): Generalized Maximum Likelihood Type Estimation of Autoregressive Moving Average Models. Ph. D. Thesis, Techn. Univ. Graz.

Stockinger N. (1985b): Computer Programs for the Simulation and GM-Estimation of ARMA models. Res. Rep. TS-1985-2; Inst. Statist. and Wahrscheinlichkeitstheorie, Techn. Univ. Vienna.

Stockinger N. and R. Dutter (1983): Robust Time Series Analysis — An Overview. Res. Rep. 9, Inst. Statist., Techn. Univ. Graz.

Stockinger N., K. P. Pfeiffer and R. Dutter (1984): Ausreissererkennung in Arrhythmischen Druckpulsen durch Robuste Methoden der Zeitreihenanalyse. In: Medizinische Informatik

'84 (Gell and Eichtinger, eds.), Schriftenreihe der Oesterr. Computer Ges., Vol. 24, Oldenbourg Verlag, Vienna.

Stuetzle W. (1979): Asymptotics for Running M-Estimates. In: Smoothing Techniques for Curve Estimation — Proc. Heidelberg, 1979 (Gasser and Rosenblatt, eds.). Springer-Verlag, New York.

Thomson D. J. (1977): Spectrum Estimation Techniques for Characterization and Development of WT 4 Waveguide — I. Bell System Techn. J. *56*, 4, 1769—1815.

Tukey J. W. (1960): A Survey of Sampling from Contaminated Distributions. In: Contributions to Probability and Statistics (I. Olkin, ed.). Stanford University Press, Stanford, CA.

Tukey J. W. (1977): Explorative Data Analysis. Addison Wesley, Reading, MA.

Tukey J. W. and T. E. Harris (1949): Development of Large-Sample. Measures of Location and Scale Which Are Relatively Insensitive to Contamination (Sampling from Contaminated Distributions, 3). Memorandum Rep. 31, Statist. Research Group, Princeton University, New Jersey.

Velleman P. F. (1975): Robust Nonlinear Data Smoothers; Theory, Definitions and Applications. Ph. D. Thesis, Dept. Statist., Princeton University.

Velleman P. F. (1980): Definition and Comparison of Robust Nonlinear Data Smoothing Algorithms. J. Amer. Statist. Assoc. *75*, 371, 609—615.

Walker A. M. (1960): Some Consequences of Superimposed Error in Time Series Analysis. Biometrika *47*, 33—43.

Wegman E. J. and R. J. Carrol (1977): A Monte Carlo Study of Robust Estimators of Location. Commun. Statist. — Theor. Meth. *A 6*, 9, 795—812.

Whittle P. (1952): Estimation and Information in Stationary Time Series. Arkiv foer Matematik *2*, 23, 423—434.

Wilson G. T. (1969): Factorization of the Generating Function of a Pure Moving Average Process. SIAM J. Num. Analysis *6*, 1.

Yohai V. J, and R. A. Maronna (1977): Asymptotic Behavior of Least Squares Estimates for Autoregressive Processes with Infinite Variances. Ann. Statist. *5*, 3, 554—560.

Zeh J. E. (1979): Efficiency Robustness of Generalized M-Estimates for Autoregression and Their Use in Determining Outlier Type. Ph. D. Thesis, Univ. Washington, Seattle.

I. MODELS AND CONCEPTS OF ROBUSTNESS

This contribution is thought to be a first and introductory chapter in a series of five chapters.

We will review different concepts of robustness for independently, identically distributed observations as well as for time series. Min-max robustness, efficiency robustness and qualitative robustness will be treated in more details.

Consistency and robustness properties of least squares estimators of autoregressive integrated moving average parameters will be dealt with, where the given time series is possibility contaminated by outliers. The least squares estimation behaves differently well for two types of outliers which will be considered. The bad performance of least squares estimators for contaminated data will show the necessity of robust estimation methods, methods which are robust toward outliers and wrong specification of the model.

I. 1 GENERAL CONCEPTS OF ROBUSTNESS

Loosely speaking, a robust estimator is one whose performance remains quite good if the true distribution of data deviates slightly from the assumed one. Data sets for which often the Gaussian model is assumed, sometimes contain a small fraction of outliers. More realistic models for such data sets are provided by heavy-tailed distributions. A large portion of the literature on robustness, e.g. Dutter (1980), treats location and linear regression models with independently, identically distributed errors. A relatively small number of contributions, e.g. Dutter (1983), Polasek and Mertl (1983), deal with robust estimation of covariances. In this section some concepts of robustness that have been primarily developed in the independent observations context will be discussed.

There are different possibilities to judge the robustness performance of an estimator, namely by the concepts of efficiency robustness, min-max robustness and qualitative robustness.

Efficiency Robustness

Efficiency robustness requires – roughly speaking – high efficiencies of an estimator in a neighborhood of an assumed distribution (Tukey, 1960). Efficiency robustness can be defined more exactly as follows (Martin and Yohai, 1984a):

Let $T_n = T_n(Y_1, \ldots, Y_n)$ be an estimator of a scalar parameter μ in the distribution P_μ^n of $\mathbf{Y} = (Y_1, \ldots, Y_n)$, a sample of size n, and let $EFF\ (T_n, P_\mu^n)$ denote a suitably defined efficiency of T_n at P_μ^n. For example we might have

$$EFF(T_n, P_\mu^n) = \frac{VAR_{P_\mu^n}(\text{known estimator of } \mu \text{ with the smallest variance})}{VAR_{P_\mu^n}(T_n)}$$

or we might have

$$EFF(T_n, P_\mu^n) = \frac{V_{CR}(P_\mu^n)}{VAR_{P_\mu^n}(T_n)}$$

where $V_{CR}(P_\mu^n)$ is the Cramer-Rao lower bound of the variance at P_μ^n.

When the focus is on asymptotic efficiencies, the estimator is denoted by T, the distribution for the process $\{Y_i\}_{,i\geqq 1}$ is denoted by P_μ^∞, and the efficiency of T at P_μ^∞

$$EFF(T, P_\mu^\infty) = \frac{V_{CR}(P_\mu^\infty)}{V_\infty(T)}$$

where $V_\infty(T)$ is the asymptotic variance of $\sqrt{(n)}\ T_n$ at P_μ^∞.

Let P_μ^∞ be the assumed distribution (which is often called nominal distribution) for the data (typically P_μ^∞ is Gaussian), and let $P_{\mu,1}^\infty, P_{\mu,2}^\infty, \ldots, P_{\mu,K}^\infty$ be a set of distributions which are in some sense "near" to P_μ^∞. Then an estimator T is said to be *efficiency robust* if T has high efficiency at P_μ^∞, and also at $P_{\mu,1}^\infty, \ldots, P_{\mu,K}^\infty$. High efficiency at P_μ^∞ will usually mean an efficiency in the range between 90% and 95%.

Min-Max Robustness

Huber's (1964) *min-max robust* location estimates minimize the maximum asymptotic variance over certain uncountably infinite families of distributions. More precisely, this concept of robustness can be formulated as follows: Let $V(T, P)$ denote the asymptotic variance of an estimator T at distribution P, and let $\mathbf{T}$ denote a family of estimators, while $\mathbf{P}$ denotes a family at univariate distributions. A min-max robust estimator T_0 solves the problem

$$\inf_{T\in\mathbf{T}} \sup_{P\in\mathbf{P}} V(T, P)\,.$$

For more min-max theory and results see Serfling (1980) and Huber (1981).

Qualitative Robustness

Hampel's (1968, 1971) concept of qualitative robustness requires equicontinuity of an estimator on a set of distributions of the data. This concept is summarized in the following.

Let $Y_1, \ldots, Y_n, \ldots$ be independently, identically distributed (i.i.d.) random variables with values on a complete and separable metric space (Ω, d) with metric d. In most

cases Ω is a Euclidean space. Let Ω^n and Ω^∞ be the Cartesian product of n copies of Ω and countable copies of Ω, respectively. Let $\mathfrak{B}$ denote the Borel-σ-algebra on Ω and let $\mathfrak{B}^n$ denote the corresponding product σ-algebra on Ω^n.

An *ε-neighborhood* ($\varepsilon > 0$) of $B \in \mathfrak{B}$ is defined by

$$B^\varepsilon: \{x \in \Omega \mid \inf_{y \in B} d(x, y) \leqq \varepsilon\} .$$

For the measurable space $(\Omega, \mathfrak{B})$ let $\mathfrak{p}(\Omega)$ denote the set of all probability measures on $\mathfrak{B}$.

For F and G in $\mathfrak{p}(\Omega)$ the Prohorov distance of these measures is defined by

$$\pi_d(F, G) := \inf \{\varepsilon > 0 \mid \text{ for all } \; B \in \mathfrak{B}, \; F(B) \leqq G(B^\varepsilon) + \varepsilon\} .$$

For a given $F \in \mathfrak{p}(\Omega)$ let F^n denote the corresponding product measure in $\mathfrak{p}(\Omega^n)$. Let $T_n: \Omega^n \to \theta$ be a sequence of estimators where the parameter space (θ, γ) is also a complete and separable metric space.

The sequence of estimators $\{T_n\}_{n \geqq n_0}$ is *qualitatively robust* at $F \in \mathfrak{p}(\Omega)$, if, given $\varepsilon > 0$, there exists $\delta > 0$ such that, for all $n \geqq n_0$ and for all $G \in \mathfrak{p}(\Omega)$

$$\pi_d(F, G) < \delta \Rightarrow \pi_\gamma(\mathfrak{L}(T_n \mid F^n), \mathfrak{L}(T_n \mid G^n)) < \varepsilon ,$$

where $\mathfrak{L}(T_n \mid F^n)$ denotes the law of T_n under F^n.

This definition of qualitative robustness requires, uniformly in sample size n, that the distribution of the estimators does not change much when there is a small change in the marginal distribution of the observations, which might be produced by one or both of

(a) a contamination of a small fraction of observations with gross errors (outliers),
(b) small errors in all the observations (e.g. rounding or grouping errors).

Influence Curve and Breakdown Point

Since qualitative robustness gives no possibility to distinguish between more or less robust estimators, Hampel (1968, 1971, 1974) introduced the influence curve and the breakdown point.

Let T denote a vector valued mapping of a subset of $\mathfrak{p}(\Omega)$ into the k-dimensional Euclidean space $\mathbb{R}^k$ and let F be in the domain of T. Let δ_y denote the degenerated distribution having all its mass in $y \in \Omega$. The *influence curve* of T for F is defined pointwise by

$$IC_{T,F}(y) = \lim_{t \to 0} \frac{T[(1 - t) F + t\delta_y] - T(F)}{t} .$$

The influence curve describes the standardized influence of an infinitesimal term at a certain position, on an estimator.

The *breakdown point* is essentially the largest fraction of contamination, which

does not ruin an estimate (see Donoho and Huber, 1983, for a good and exact definition).

There are also other tools to measure the robustness of estimators, e.g. the gross-error sensitivity, the local shift sensitivity and the rejection point (compare Dutter, 1980). The infinitesimal approach to robustness is already documented in a book (Hampel et al., 1986).

I. 2 CONCEPTS OF ROBUSTNESS FOR TIME SERIES

Efficiency Robustness and Min-Max Robustness

For time series parameter estimation problems, efficiency robustness and min-max robustness are directly applicable concepts, because these concepts do not require independent (possibly vector valued) data. Efficiency robustness for vector parameters can be defined similarly to that of scalar parameters by using an appropriate definition of multivariate efficiency. Zeh (1979) investigated efficiency robustness of estimators of time series models using different measures for multivariate efficiency.

Influence Curve and Breakdown Point

Kuensch (1983b) and Martin and Yohai (1984b) give definitions for influence curves of parameter estimators in time series models. The definition of breakdown points for time series parameter estimators must pay attention to the detailed nature of the failure mechanism. For instance i.i.d. gross errors on the one hand, and highly correlated or patchy gross errors on the other, may yield different breakdown points. Martin and Yohai (1984a) comment on breakdown points for time series parameter estimators.

Qualitative Robustness

The problem which remains is providing an appropriate definition of qualitative robustness in the time series context. One possibility, but with not entirely satisfactory theory, is to use an asymptotic version of qualitative robustness (Martin, 1979), requiring continuity but not equicontinuity and replacing estimator sample distributions with asymptotic distributions. Thus an estimator T is asymptotically qualitatively robust at $F \in \mathfrak{p}(\Omega)$, if, given $\varepsilon > 0$, there exists $\delta > 0$, such that for all $G \in \mathfrak{p}(\Omega)$

$$\pi_d(F, G) < \delta \Rightarrow \pi_\gamma(\mathfrak{L}(T \mid F), \mathfrak{L}(T \mid G)) < \varepsilon$$

where $\mathfrak{L}(T \mid F)$ denotes the asymptotic distribution (the "law") of T for distribution F.

In order to, at least partially, cover non-i.i.d. observations, Hampel (1971) introduced the concept of qualitative π-robustness which is thought for observations

which are dependent in a certain weak sense. In contrast to the definition of qualitative robustness, which is based on marginal distributions F and G, the definition of qualitative π-robustness is based on multivariate probability measures.

Using the qualitative π-robustness as a starting point, Papantoni-Kazakos and Gray (1979) define qualitative robustness of estimators on stationary observations. The authors substitute the generalized Ornstein metric for the Prohorov metric to measure the distance of sample distributions of stationary processes. Cox (1981) thinks that the generalized Ornstein metric is not superior to the Prohorov metric and presents other metrics on distributions of stochastic processes in order to define qualitative robustness for dependent data. Infinitesimal robustness for autoregressive processes was considered by Kuensch (1983a). Bustos (1981) also did some work on qualitative robustness for general processes.

Resistance

Boente, Fraiman and Yohai (1982) propose a new approach to qualitative robustness, based on the concept of resistance (compare Mosteller and Tukey (1977)). This approach has the advantage that it may be applied without special assumptions on the probability model for the observations, e.g. they may be dependent or non-identically distributed. The concept of resistance can be formalized as follows:

Given $\mathbf{x}^n = (x_1, \ldots, x_n)$ and $\mathbf{y}^n = (y_1, \ldots, y_n)$ in Ω^n, define a distance d_n^+ on Ω^n

$$d_n^+(\mathbf{x}^n, \mathbf{y}^n) := \inf\{\varepsilon \mid \text{number of } \{i \mid d(x_i, y_i) \geqq \varepsilon\} \leqq n\varepsilon\} .$$

Therefore two points of Ω^n have a distance smaller or equal than ε, if for one point a fraction not greater than ε of observations are replaced by arbitrary outliers, or if all the observations of one point are perturbed by round-off errors smaller than ε.

A change of T_n, which is caused by a change – characterized by $\delta > 0$ – of $\mathbf{x}^n \in \Omega^n$, is defined by

$$\Delta T_n(\mathbf{x}^n, \delta) = \sup\{|T_n(\mathbf{y}^n) - T_n(\mathbf{z}^n)| \mid d_n^+(\mathbf{y}^n, \mathbf{x}^n) < \delta, d_n^+(\mathbf{z}^n, \mathbf{x}^n) \leqq \delta\} .$$

The following definition formalizes the data oriented concept of resistance. Let $\mathbf{x} = (x_1, \ldots, x_n, \ldots) \in \Omega^\infty$ and $\mathbf{x}^n = (x_1, \ldots, x_n)$. Then $\{T_n\}_{n \geqq n_0}$ is resistant at $\mathbf{x}$ if, given $\varepsilon > 0$, there exists $\delta > 0$ and n_0 such that

$$\Delta T_n(\mathbf{x}^n, \delta) \leqq \varepsilon \quad \text{for all} \quad n \geqq n_0 .$$

The following definitions of strong and weak robustness represent alternatives to Hampel's definition of qualitative robustness. Let $F^\infty \in \mathfrak{P}(\Omega^\infty)$. $\{T_n\}_{n \geqq n_0}$ is *strongly robust* at F^∞, if

$$F^\infty([\{T_n\}_{n \geqq n_0} \text{ is resistant at } \mathbf{x}]) = 1 .$$

$\{T_n\}_{n \geqq n_0}$ is *weakly robust* at F^∞ if, given $\varepsilon > 0$, there exist $\delta > 0$ and n_0 such that

$$F^\infty([\Delta T_n(\mathbf{x}^n, \delta) \leqq \varepsilon]) \geqq 1 - \varepsilon \quad \text{for all} \quad n \geqq n_0 .$$

Although the latter definitions of qualitative robustness are very useful and transparent, it is not trivial to prove qualitative robustness of implicitly defined estimators (e.g. M-estimators).

I. 3 AUTOREGRESSIVE INTEGRATED MOVING AVERAGE MODELS

A widely used method to describe the mechanism that generates and explains a univariate time series or process

$$y_1, \ldots, y_n \tag{I.1}$$

is the estimation of an *autoregressive integrated moving average model of orders p, d and q* (ARIMA (p, d, q) model) (see Box and Jenkins (1976))

$$\Phi(B)(y_i - \mu) = \theta(B)\, a_i \tag{I.2}$$

where $\theta(B)$ and $\Phi(B)$ denote the moving average operator and the nonstationary autoregressive operator, respectively, i.e.

$$\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$$

and

$$\Phi(B) = \phi(B)(1 - B)^d$$

where $\phi(B)$ denotes the autoregressive operator

$$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$$

with the backward shift operator B $(By_i = y_{i-1})$, autoregressive parameters $\phi_1, \ldots, \phi_p$ and moving average parameters $\theta_1, \ldots, \theta_q$. If $d = 0$, then it is reasonable to use a location parameter $\mu \neq 0$. The a_i's are realization of i.i.d. random variables A_i with a symmetric distribution G with mean zero and scale σ. The density of G will be denoted by g. The A_i's are called innovations.

For all subsequent considerations the stationarity of the autoregressive operator and invertibility of the moving average operator is supposed. Therefore the roots of each of the characteristic equations

$$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p = 0$$

and

$$\theta(B) = 1 - \theta_1 B - \ldots - \theta_p B^p = 0\,,$$

B now denoting a variable, must lie outside the complex unit circle.

Box and Jenkins (1976) propose to estimate d by "differencing" the given time series (I.1), i.e. by regarding the differences of subsequent observations as a new time series, until the autocorrelation function of the new time series decays quickly.

If a value for d is determined there only remains the problem of estimating an autoregressive moving average model *of orders p and q* (ARMA (p, q) model)

$$\phi(B)\, w_i = \theta(B)\, a_i \tag{I.3}$$

where $w_i = (1 - B)^d\, y_i$ denotes an observation of the d-times "differenced" time series (I.1).

In the following we will assume $d = 0$ and $w_i = y_i$ and concentrate on the estimation of ARMA (p, q) models.

Special cases of an ARIMA (p, d, q) model (I.2) are the *moving average model of order q* (MA (q) model)

$$y_i - \mu = \theta(B)\, a_i \tag{I.4}$$

and the *autoregressive model of order p* (AR (p) model)

$$\phi(B)\,(y_i - \mu) = a_i\,. \tag{I.5}$$

After defining the intercept

$$\lambda := \mu\Big(1 - \sum_{i=1}^{p} \phi_i\Big) \tag{I.6}$$

the AR (p) model (I.5) can be written as a linear regression model with parameter vector

$$\boldsymbol{\beta} := (\lambda, \phi_1, \ldots, \phi_p)^{\mathrm{T}}\,. \tag{I.7}$$

I.4 TIME SERIES OUTLIER MODELS

We assume that realizations x_i of random variables X_i satisfy the ARIMA (p, d, q) model (1.2) to be estimated. The time series (I.1) is called to be *outlier-free*, if $y_i = x_i$, $i = 1, \ldots, n$, and G, the distribution of the innovations, is Gaussian.

When considering the problem of estimating time series parameters robustly, there is a need of characterizing time series contaminated by outliers in appropriate probabilistic models. Since complete probabilistic models are difficult to formulate (Martin, 1979), it seems imperative to begin with specifying simple outlier generating models, which are able to represent real data with outliers. In practice, outliers behave often as follows (Martin, 1979, 1980):

For a possible outlier behavior, the character of the outliers is consistent with the remainder of the sample path except for an initial jump. A second possibility is that of isolated or gross-error outliers which might be due to various reasons like recording errors. A third possibility is that of patchy type outliers whose behavior appears somewhat or totally unrelated to the behavior of the remainder of the sample. This type might be due to a brief malfunctioning of a recording instrument.

Now we want to capture some of the essence of the above kinds of behavior with appropriate formal models.

The first kind of behavior might be obtained by an *innovation outlier (IO) model*, which is given if the y_i's are equal to the x_i's and if the innovations distribution G is symmetric and heavy-tailed. Outliers generated by an IO model are called *innovation outliers* (IO's). G could be a t-distribution or a contaminated normal

$$CN(v, \sigma_1, \sigma_2) = (1 - v)\, N(0, \sigma_1^2) + vN(0, \sigma_2^2) \tag{I.8}$$

where $N(0, \sigma^2)$ denotes the normal distribution with mean 0 and variance σ^2 and $\sigma_2^2 \gg \sigma_1^2$ and v is small.

For the second and third kind of behavior the following *additive outlier (AO) model* may be the simplest appropriate representation. *Additive outliers* (AO's) (which are generated by AO models) are given if

$$Y_i = X_i + V_i \tag{I.9}$$

where the innovations A_i are normally distributed and the V_i's are random variables, distributed independently of X_i and whose marginal distribution satisfies $P(V_i = 0) = 1 - \gamma$ with γ not too large. For time series occurring in practice γ is in the range from .01 to .25.

Independently and identically distributed V_i's model the gross-error situation. The distribution of the V_i's could be a Gaussian mixture distribution

$$CND(\gamma, \sigma_3) = (1 - \gamma)\, \delta_0 + \gamma N(0, \sigma_3^2) \tag{I.10}$$

where δ_0 denotes the degenerated distribution having all its mass in the origin.

Patchy type additive outliers can be obtained if the independence assumption for the V_i's is dropped.

These types of outliers were first mentioned by Fox (1972). He considers two types of outliers: those which affect only the observation on which they occur (Type I outliers) and those which affect successive observations as well (Type II outliers).

Gastwirth and Rubin (1975) study the behavior of some robust estimators of location for a first-order autoregressive process with a double exponential marginal distribution. This process is a special version of an IO model.

Also Abraham and Box (1979) use both AO and IO models to consider inferences about the parameters of a possibly contaminated autoregressive process. However, they call their outlier generating models "aberrant observation model" and "aberrant innovation model", respectively.

Some outlier-handling techniques require the specification of the data points which have to be treated as outliers, e.g. Brubacher (1974), Jones (1980). Since it is not likely to have this specification (see however Chernick, Downing and Pike, 1982) those techniques will not be discussed in this series of contributions.

I.5 LEAST SQUARES ESTIMATION OF AR MODELS

If $\mu = 0$ the AR (p) model (I.5) can be written in the linear model form

$$\mathbf{y} = Z\boldsymbol{\phi} + \mathbf{a}$$

where $\mathbf{y} = (y_{p+1}, \ldots, y_n)^{\mathrm{T}}$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^{\mathrm{T}}$, $\mathbf{a} = (a_{p+1}, \ldots, a_n)^{\mathrm{T}}$ and $Z = [\mathbf{z}_{p+1}, \ldots, \mathbf{z}_n]^{\mathrm{T}}$ with $\mathbf{z}_i = (y_{i-1}, \ldots, y_{i-p})^{\mathrm{T}}$. The least squares estimator $\boldsymbol{\phi}\hat{}$ is defined by the solution of

$$\sum_{i=p+1}^{n} (y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\phi}')^2 = \min . \tag{I.11}$$

Mann and Wald (1943) show that even without a Gaussian G, if the fourth moments of A_i exist and are finite and $y_i = x_i$, $i = 1, \ldots, n$, then $\sqrt{(n)}\,(\boldsymbol{\phi}\hat{} - \boldsymbol{\phi})$ has a limiting normal distribution and $\sigma^2(Z^{\mathrm{T}}Z)^{-1}$ is the asymptotically correct expression for the covariance matrix of $\boldsymbol{\phi}\hat{}$. Therefore we can treat the problem of estimating autoregressive parameters like the classical regression problem.

Consistency and Robustness Properties for IO Models

It is well known (Martin, 1982) that the least squares estimator $\boldsymbol{\beta}\hat{}^{\mathrm{T}} = (\hat{\lambda}, \boldsymbol{\phi}\hat{}^{\mathrm{T}})$ (I.7) – which can be defined similarly to $\boldsymbol{\phi}\hat{}$ – is asymptotically normal and asymptotically efficient when the innovation distribution G is Gaussian. The same is true for both the innovation scale estimator $\hat{\sigma}$, obtained from the sum of squared residuals, and the "autoregressive-errors" location estimator

$$\hat{\mu} = \hat{\lambda} / \left(1 - \sum_{i=1}^{p} \hat{\phi}_i\right) . \tag{I.12}$$

In classical theory (Anderson, 1971) it is proved that $\boldsymbol{\phi}\hat{}$ is consistent if the variance of the innovation is finite.

But some data may be better represented by AR models with innovations which have infinite variances. This has raised the question of whether the classical estimators are still reliable when innovation variances do not exist. A partial answer was obtained by Kanter and Steiger (1974). They show that $\boldsymbol{\phi}\hat{}$ is consistent if G is a symmetric stable law of index $\alpha \in (0, 2]$, which is defined by

$$\int_{-\infty}^{\infty} \exp(\mathrm{i}tx)\, \mathrm{d}G(x) = \exp(-c|t|^{\alpha})$$

for some $c > 0$.

Yohai and Maronna (1977) have shown, more generally, that a sufficient condition for consistency of $\boldsymbol{\phi}\hat{}$ is

$$E\{[\log |a_i|]^+\} < \infty ,$$

where $[x]^+$ denotes the positive part of x. This condition cannot be weakened since it is necessary for the existence of the stationary autoregressive process.

Hannan and Kanter (1977) have shown that if G belongs to the *domain of attraction* of a stable law of index $\alpha \in (0, 2)$, then $\hat{\phi}$ converges in probability to the true with rate $T^{1/\alpha}$, and therefore faster than $T^{1/2}$ as in the finite variance case.

The results from Kanter and Steiger (1974), Yohai and Maronna (1977) and Hannan and Kanter (1977), however, are based on the assumption that the location parameter μ is known.

$\hat{\phi}$ is asymptotically qualitatively robust in the sense, that its asymptotic covariance matrix $V_{\hat{\phi}}$ (the covariance matrix of the limiting distribution of $\sqrt{(n)}\,(\hat{\phi} - \phi)$) is independent of the innovation distribution, at least provided that the innovations have finite variance. This fact is somewhat obscured by the common practice of writing

$$V_{\hat{\phi}} = \sigma^2 C^{-1}$$

where the elements of C are given by $c_{ij} = \mathit{covariance}\,(Y_i, Y_j)$, $1 \leqq i, j \leqq p$. However, $C = \sigma^2 \bar{C}$ where $\bar{C}$ is the covariance matrix for innovations with unit variance. Thus $V_{\hat{\phi}}$ is better written as

$$V_{\hat{\phi}} = \bar{C}^{-1} \tag{I.13}$$

where $\bar{C}$ depends only upon ϕ. This behavior was pointed out first by Whittle (1952). In fact, the distribution-free property exhibited in (I.13) is an asymptotic analogue of Hampel's (1971) qualitative robustness concept, provided that only innovation outliers are possible and the innovation variance is finite (Martin, 1981).

In sharp contrast to $\hat{\phi}$, the least squares estimators of the location μ and the innovations scale σ are not robust in the above sense.

On the other hand $\hat{\phi}$ is not efficiency robust toward heavy-tailed innovation distributions, i.e. arbitrarily small departures of G from normality may cause arbitrarily large asymptotic variances of $\hat{\phi}$ (Maronna, Bustos and Yohai, 1979). This can be seen easily by computing asymptotic efficiencies. Straightforward calculation (Martin, 1981) shows that the large sample information matrix for $\hat{\phi}$ is

$$I_{\hat{\phi}} = \sigma^2\, i(g)\, \bar{C}$$

where $i(g) = E\{\partial \log g(a, \mu)/\partial \mu\}^2$ is the Fisher information (for location) for an innovation density g with finite variance. The Cramer-Rao lower bound $V_{\phi,CR}$ for the variance of $\hat{\phi}$ is the inverse information matrix.

Taking the pth root of the ratio of determinants as a multivariate measure of efficiency (compare Anderson, 1971) gives

$$EFF(LS, g) = \left(\frac{\det V_{\phi,CR}}{\det V_{\hat{\phi}}}\right)^{1/p} = (\sigma^2\, i(g))^{-1}\,. \tag{I.14}$$

But this is just the p-th power of the asymptotic efficiency of the sample mean for i.i.d. random variables, and the latter is notoriously lacking in efficiency robustness toward heavy-tailed G's. Computing the efficiency of $\hat{\mu}$ yields also the right hand side

expression in (I.14). The efficiency of $\hat{\sigma}$ is the same as that of the sample standard deviation calculated for i.i.d. data. The latter estimator has even less efficiency robustness than the sample mean (Tukey and Harris, 1949, Tukey, 1960). Consideration of the Cramer-Rao lower bound

$$V_{\phi_1,CR} = \frac{1 - \phi_1^2}{\sigma^2} \frac{1}{i(g)}$$

for the first-order AR parameter, where $1 - \phi_1^2$ is the asymptotic variance of $\hat{\phi}_1$ (Martin and Jong, 1976), makes it transparent how heavy-tailed distributions diminish the Cramer-Rao bound and therefore also the efficiency $EFF(LS, g)$ (I.14); for σ^2 can become arbitrarily large in arbitrarily small neighborhoods of the Gaussian distribution while $i(g)$ remains relatively stable.

Consistency and Robustness Properties for AO Models

If a time series contains additive outliers, $\hat{\boldsymbol{\phi}}$ not only lacks efficiency robustness but also suffers from serious bias problems. Martin and Jong (1976) and Denby and Martin (1979) show that the variance of $\hat{\phi}_1$ can be very large. Bias problems for the first-order AR parameter will be explained in the following. The bias for $\hat{\phi}_1$ is

$$B(\hat{\phi}_1) = -\phi_1 \sigma_V^2 / (\sigma_X^2 + \sigma_V^2)$$

assuming finite variance σ_X^2 and σ_V^2 for X_t and i.i.d. V_t, respectively (I.9). This bias vanishes only if $\phi_1 = 0$ or if $\sigma_X^2/\sigma_V^2 \to \infty$, what corresponds to an innovation outlier model. $B(\hat{\phi}_1)$ can be disastrous for rather mild contaminations through V_t's. For example if V_t is CND (.1, 10) distributed (I.10) and $\sigma_X^2 = 1$, then $B(\hat{\phi}_1) = -\phi_1/2$, i.e. the bias is 50%.

Certain additive outliers can produce the effect that some of the roots of the characteristic equation $\hat{\phi}(B) = 0$ lie on the unit circle, therefore $\hat{\boldsymbol{\phi}}$ has a breakdown point of value zero (Martin, 1980).

I.6 LEAST SQUARES ESTIMATION OF ARMA MODELS

In contrast to AR models, the estimation of MA and ARMA models is always a nonlinear problem. Box and Jenkins (1976) treat the estimation of ARMA models for outlier-free time series. Their methods unfortunately give no reliable results if the given time series contains outliers. (Compare Martin and Jong (1976) and Denby and Martin (1979) for the first-order autoregressive parameter.) The authors present a conditional maximum likelihood estimator of $(\boldsymbol{\phi}^T, \boldsymbol{\theta}^T, \sigma)$, where the not observed values $y_{1-p}, \ldots, y_0$ and $a_{1-q}, \ldots, a_0$ must be chosen in advance. For a fixed scale σ this estimator is equivalent to a least squares estimator with the same conditions. A conditional least squares estimator of $\boldsymbol{\alpha} := (\boldsymbol{\phi}^T, \boldsymbol{\theta}^T, \mu)^T$ is, however,

more conveniently computed by solving

$$\sum_{i=p+1}^{n} r_i^2(\boldsymbol{\alpha}') = \min \tag{I.15}$$

where the minimum in $\boldsymbol{\alpha}'$ has to be achieved and what avoids the problem of forecasting or, more simply, choosing observations that were not observed. The residuals

$$r_i(\boldsymbol{\alpha}') = \theta'^{-1}(B)\,\phi'(B)\,(y_i - \mu') \tag{I.16}$$

where an AR or MA operator marked by a prime use arbitrary AR parameters $\boldsymbol{\phi}'$ or MA parameters $\boldsymbol{\theta}'$, respectively, can be computed recursively by the following algorithm:

(1) Assume $y_j = \mu$ and $r_j(\boldsymbol{\alpha}') = 0, j \leqq 0$.
(2) Set $i = 1$.
(3) Compute $r_i(\boldsymbol{\alpha}') = (y_i - \mu') - \phi_1'(y_{i-1} - \mu') - \ldots - \phi_p'(y_{i-p} - \mu') + \theta_1' r_{i-1}(\boldsymbol{\alpha}') + \ldots + \theta_q' r_{i-q}(\boldsymbol{\alpha}')$.
(4) Augment $i = i + 1$.
(5) If $i \leqq n$ go to (3), else stop.

Box and Jenkins (1976) also propose an unconditional maximum likelihood estimator of $(\boldsymbol{\phi}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}}, \sigma)$, where the so-called technique of *back forecasting* is used to estimate values $y_i, i \leqq 0$, that were not observed. If n, the number of observations, is not too small, the unconditional maximum likelihood estimator is well approximated by an unconditional least squares estimator. Since maximum likelihood estimators take into account the dependence of the observations of a time series, usage of least squares estimators is justified.

If the given time series is outlier-free, then the least squares estimator $\boldsymbol{\alpha}^{\wedge}$ is asymptotically efficient (Martin and Yohai, 1984a) and, if the variance of G is finite, than

$$\sqrt{(n)}\,(\boldsymbol{\alpha}^{\wedge} - \boldsymbol{\alpha}) \xrightarrow{\mathrm{d}} N(0, K(\boldsymbol{\phi}, \boldsymbol{\theta}, G))$$

where $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution and the $(p + q + 1) \times (p + q + 1)$ covariance matrix $V_{LS} = K(\boldsymbol{\phi}, \boldsymbol{\theta}, G)$ of the limiting normal distribution is given by

$$K(\boldsymbol{\phi}, \boldsymbol{\theta}, G) = \begin{pmatrix} C^{*-1}(\boldsymbol{\phi}, \boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & VAR(G)\,\dfrac{(1 - \theta_1 - \ldots - \theta_q)^2}{(1 - \phi_1 - \ldots - \phi_p)^2} \end{pmatrix} \tag{I.17}$$

where the matrix $C^*(\boldsymbol{\phi}, \boldsymbol{\theta})$ is symmetric and has the elements

$$\begin{aligned} C^*_{j,k} &= \sum_{l=0}^{\infty} \xi_l \xi_{l+k-j}, \quad \text{if} \quad j \leqq k \leqq p \\ C^*_{j,p+k} &= \sum_{l=0}^{\infty} \zeta_l \xi_{l+k-j}, \quad \text{if} \quad j \leqq p, \quad k \leqq q, \quad j \leqq k \\ C^*_{j,p+k} &= \sum_{l=0}^{\infty} \xi_l \zeta_{l+j-k}, \quad \text{if} \quad j \leqq p, \quad k \leqq q, \quad k \leqq j \end{aligned} \tag{I.18}$$

$$C^*_{p+j,p+k} = \sum_{l=0}^{\infty} \zeta_l \zeta_{l+k-j}\,, \quad \text{if} \quad j \leqq k \leqq q\,,$$

where ξ_l and ζ_l denote coefficients in the inverse AR operator and MA operator, respectively

$$\phi^{-1}(B) \sum_{l=0}^{\infty} \xi_l B^l \quad \text{and} \quad \theta^{-1}(B) = \sum_{l=0}^{\infty} \zeta_l B^l\,. \tag{I.19}$$

Notice that the upper left-hand block of (I.17) gives the asymptotic covariance matrix of $(\boldsymbol{\phi}^{\wedge\mathrm{T}}, \boldsymbol{\theta}^{\wedge\mathrm{T}})$. Thus, if $VAR(G) < \infty$, $(\boldsymbol{\phi}^{\wedge\mathrm{T}}, \boldsymbol{\theta}^{\wedge\mathrm{T}},)$ is asymptotically qualitatively robust. On the other hand the asymptotic distribution of $\hat{\mu}$ depends on G.

The asymptotic efficiency of $\boldsymbol{\alpha}^{\wedge}$ can be measured by the ratio of the trace of the asymptotic covariance matrix of the maximum likelihood estimator of the trace of V_{LS}. A maximum likelihood (ML) estimator can asymptotically be obtained by solving

$$-\sum_{i=p+1}^{n} \ln g(r_i(\boldsymbol{\alpha}')) = \min \tag{I.20}$$

and is asymptotically efficient even in the presence of innovation outliers (Martin and Yohai, 1984a). The asymptotic covariance matrix V_{ML} of this estimator is given by

$$V_{ML} = k(\Psi, G)\, K(\boldsymbol{\phi}, \boldsymbol{\theta}, G) \tag{I.21}$$

where $\Psi = -g'/g$,

$$k(\Psi, G) = V_{loc}(\Psi, G)/VAR(G) = \left[i(G)\, VAR(G)\right]^{-1}$$

with $V_{loc}(\Psi, G) = E_G \Psi^2(A)/E_G^2 \Psi'(A)$ the asymptotic variance of the location ML estimator (Huber, 1964), and $i(G)$ is the Fisher information for G, and $K(\boldsymbol{\phi}, \boldsymbol{\theta}, G)$ is given by (I.17).

Using the above described measure of asymptotic efficiency shows that the efficiency of the least squares estimator is just $k(\Psi, G)$. It is well known (Huber, 1964). that for any v-neighborhood of the $N(0, \sigma^2)$ distribution

$$G_v := \{G \mid (1 - v)\, N(0, \sigma^2) + vF,\, v > 0\}$$

with F symmetric, $k(\Psi, G)$ may be arbitrarily small. Thus $\boldsymbol{\alpha}^{\wedge}$ lacks efficiency robustness in the presence of innovation outliers.

$\boldsymbol{\alpha}^{\wedge}$ is neither efficiency robust nor unbiased, if the given date contain additive outliers (Martin and Yohai, 1984a). Martin (1980b) gives more interesting facts about the estimation of time series models.

II. MAXIMUM LIKELIHOOD TYPE ESTIMATION

Using Chapter I as a starting point this article will continue in treating the estimation of autoregressive moving-average (ARMA) models.

Definitions, computational methods and properties of maximum likelihood type estimators (M-estimators) for pure autoregressive models as well as for ARMA models will be dealt with. In contrast to least squares estimators, M-estimators are, in particular, efficiency robust if the given time series is contaminated by innovation outliers.

Two estimation methods which can be used advantageously for time series including additive outliers, will be outlined.

II.1 MAXIMUM LIKELIHOOD TYPE ESTIMATION OF AR MODELS

We now concentrate on estimating an autoregressive model of order p (AR(p) model) (I.5). The appropriate representation of the AR(p) model for the following considerations is the linear regression model

$$\mathbf{y} = Z\boldsymbol{\beta} + \mathbf{a} \tag{II.1}$$

with parameter vector $\boldsymbol{\beta} = (\lambda, \phi_1, \ldots, \phi_p)^T$ (compare (I.6) and (I.7)) and where $\mathbf{y} = (y_{p+1}, \ldots, y_n)^T$ denotes a vector of observations, $\mathbf{a} = (a_{p+1}, \ldots, a_n)^T$ denotes a vector of realizations of innovations and $Z = (\mathbf{z}_{p+1}, \ldots, \mathbf{z}_n)^T$ with $\mathbf{z}_i = (1, y_{i-1}, \ldots, y_{i-p})^T$.

Section I.5 dealt with the least squares estimation of $\boldsymbol{\beta}$ and the innovation scale σ. In particular, the least squares estimator of $\boldsymbol{\phi}$ is consistent and asymptotically qualitatively robust, even it the innovations distribution is heavy tailed. However, we could miss efficiency robustness of the least squares estimator of $\boldsymbol{\beta}$ in the presence of innovation outliers. Therefore we could be interested in a possibility to obtain an efficiency robust estimator of $\boldsymbol{\beta}$ if the given time series is contaminated by innovation outliers.

The attractive small sample robustness as well as asymptotic efficiency robustness properties of maximum likelihood type estimators (M-estimators) for regression proposed by Huber (1973), naturally suggest that for robust autoregression, one uses the analogue of the regression M-estimators (Martin and Jong, 1976). Thus a first step toward robustness is given which unfortunately still has deficiencies in the additive outliers case.

II.1.1. Definition

An *M-estimator* $\hat{\boldsymbol{\beta}}$ is defined by

$$\sum_{i=p+1}^{n} \varrho\left(\frac{y_i - \mathbf{z}_i^T \boldsymbol{\beta}'}{\hat{\sigma}}\right) = \min \tag{II.2}$$

where $\varrho(\cdot)$ is a symmetric robustifying loss function (Relles, 1968), and $\hat{\sigma}$ denotes an estimate of the innovations scale. $\hat{\sigma}$ is used to ensure the scale-invariance of the minimum problem. The maximum likelihood estimator of $\boldsymbol{\beta}$ can be obtained by using $\varrho(\cdot) = -\log g(\cdot)$, where $g(\cdot)$ denotes the density of the innovations. The ϱ-functions are often given in the form of their first derivatives $\psi(t) = \mathrm{d}\varrho(t)/\mathrm{d}t$. Various ψ-functions are listed in Dutter (1980). Examples are

Huber's monotone psi-function (Huber, 1964)

$$\psi_H(t) = \begin{cases} t & |t| \leqq c \\ c \operatorname{sgn}(t) & |t| > c \end{cases} \tag{II.3}$$

where $\operatorname{sgn}(t) = 1$ for $t > 0$ and $\operatorname{sgn}(t) = -1$ for $t < 0$,

Tukey's redescending bisquare psi-function (Beaton and Tukey, 1974)

$$\psi_B(t) = \begin{cases} t[1 - (t/c)^2]^2 & |t| \leqq c \\ 0 & |t| > c\,, \end{cases} \tag{II.4}$$

and

Hampel's three part redescending psi-function (Hampel, 1968)

$$\psi_{HA}(t) = \begin{cases} t & |t| \leqq a \\ a \operatorname{sgn}(t) & a < |t| \leqq b \\ a[t - d \operatorname{sgn}(t)]/(b - d) & b < |t| \leqq d \\ 0 & d < |t| \end{cases} \tag{II.5}$$

The purpose of a ϱ-function and, equivalently, of a ψ-function, is to bound the influence of a large residual $y_i - \mathbf{z}_i^T \boldsymbol{\beta}'$ on the estimation. According to its purpose, a ψ-function should be odd, bounded and continuous. If innovation outliers are possible, the identity function $\psi(t) = t$ is a bad choice for ψ, because in this case (II.2) defines a least squares estimator.

The scale σ could be estimated from the observations $y_1, \ldots, y_i, \ldots, y_n$. Huber (1973) proposed to estimate σ and $\boldsymbol{\beta}$ simultaneously through solving (II.2) and the side condition

$$\frac{1}{n - 2p - 1} \sum_{i=p+1}^{n} \psi^2\left(\frac{y_i - \mathbf{z}_i^T \boldsymbol{\beta}'}{\hat{\sigma}}\right) = b \tag{II.6}$$

if a monotone psi-function – like ψ_H – is used. The constant b is selected so that $\hat{\sigma}$ is asymptotically consistent for σ if the y_i's are free of outliers and the innovations distribution is Gaussian with mean zero and standard deviation σ, i.e. $b = E_{N(0,1)}\{\psi(A)\}$ where A is a random variable with distribution $N(0, 1)$.

II. 1.2. Computational Methods

The minimum problem (II.2) and the side condition (II.6) can be combined to the more general minimum problem

$$h(\boldsymbol{\beta}', \sigma') = \sum_{i=p+1}^{n} \varrho\left(\frac{y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}'}{\sigma'}\right)\sigma' + c\sigma' = \min \tag{II.7}$$

where the minimum in both $\boldsymbol{\beta}'$ and σ' has to be achieved and $c = (n - 2p - 1)\, b/2$ (see Dutter, 1975).

Differentiating $h(\boldsymbol{\beta}', \sigma')$ with respect to σ' and $\boldsymbol{\beta}'$ and equating the resulting expressions to zero yield

$$\sum_{i=p+1}^{n} \chi\left(\frac{y_i - \mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right) = c \tag{II.8}$$

with $\chi(t) = t\,\psi(t) - \varrho(t)$ and a system of equations defining $\hat{\boldsymbol{\beta}}$ for a known σ,

$$\sum_{i=p+1}^{n} \psi\left(\frac{y_i - \mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\mathbf{z}_i = \mathbf{0}\,. \tag{II.9}$$

Note that (II.9) can be written as follows.

$$\mathbf{0} = \sum_{i=p+1}^{n} \psi\left(\frac{r_i}{\hat{\sigma}}\right)\mathbf{z}_i = \sum_{i=p+1}^{n} \frac{r_i}{\hat{\sigma}}\,\frac{\psi(r_i/\hat{\sigma})}{(r_i/\hat{\sigma})}\,\mathbf{z}_i \tag{II.10}$$

where r_i denotes the residual $r_i = y_i - \mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$.

This shows that M-estimators can be regarded as weighted least squares estimators with weights $w_i = \psi(r_i/\hat{\sigma})/(r_i/\hat{\sigma})$. Unfortunately the weights w_i depend on the residuals and therefore on $\hat{\boldsymbol{\beta}}$, hence (II.10) is only an implicit equation. The following iterated weighted least squares (IWLS) algorithm, however, could be used to estimate $\boldsymbol{\beta}$ and σ simultaneously. A convergence proof for the estimation of linear models is given by Dutter (1975). Of course, the so-called H-algorithm (Dutter, 1980; Dutter and Huber, 1981) could also be used to compute M-estimators of $\boldsymbol{\beta}$ and σ.

IWLS algorithm

Let starting values $\boldsymbol{\beta}^{(0)}$ and $\sigma^{(0)}$, and a tolerance value ε be given.

1. Set the iteration counter $m = 0$.
2. Denote $r_i^{(m)} = y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}^{(m)}$, $i = p + 1, \ldots, n$.
3. Compute a new value for σ using (II.8)

$$(\sigma^{(m+1)})^2 = \frac{1}{c}\sum_{i=p+1}^{n} \chi\left(\frac{r_i^{(m)}}{\sigma^{(m)}}\right)(\sigma^{(m)})^2\,.$$

4. Calculate weights

$$w_i^{(m)} = \begin{cases} \psi(r_i^{(m)}/\sigma^{(m+1)})/(r_i^{(m)}/\sigma^{(m+1)}) & \text{if } r_i^{(m)} \neq 0 \\ 1 & \text{otherwise} \end{cases}$$

where $i = p + 1, \ldots, n$. Define a diagonal matrix $W^{(m)}$ with w_i as its $(i - p)$th diagonal element.

5. Solve

$$\sum_{i=p+1}^{n} \left(r_i^{(m)} - \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\tau}^{(m)}\right)^2 w_i^{(m)} = \min$$

for $\boldsymbol{\tau}^{(m)}$, which could be computed by

$$\boldsymbol{\tau}^{(m)} = (Z^{\mathrm{T}} W^{(m)} Z)^{-1} Z^{\mathrm{T}} W^{(m)} \mathbf{y} - \boldsymbol{\beta}^{(m)}$$

where Z and $\mathbf{y}$ are defined by (II.1).

6. Compute new values for $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \omega \boldsymbol{\tau}^{(m)}$$

where $0 < \omega < 2$ is an arbitrary relaxation factor.

7. Stop, if

$$\left|\sigma^{(m)} - \sigma^{(m+1)}\right| < \varepsilon \sigma^{(m+1)}$$

and if the difference between the parameters is less than ε times their approximate standard deviation, i.e.

$$\left|\omega \tau_k^{(m)}\right| < \varepsilon \sigma^{(m+1)} \sqrt{z^{kk}}, \quad k = 1, \ldots, p$$

where z^{kk} is the kth diagonal element in $(Z^{\mathrm{T}} Z)^{-1}$.

8. Augment $m = m + 1$ and go to 2.

The IWLS algorithm described above can only be used for a monotone ψ. But using redescending psi-functions, e.g. Tukey's ψ_B (II.4), yields higher efficiencies at extremely heavy tailed distributions than the monotone psi-functions, e.g. Huber's ψ_H (II.3) (Andrews et al., 1972; Denby and Larsen, 1977). For a redescending ψ the IWLS algorithm must be modified as follows: An estimated value for the scale must be given and Step 3 must be omitted.

It must be considered that for a redescending ψ the estimating equation (II.9) could have multiple roots. Therefore the following overall computational strategy is advisable when using a redescending ψ:

Step 1: Set $\psi(t) = t$ to obtain least squares estimates $\boldsymbol{\beta}$ and $\hat{\sigma}$ from the IWLS algorithm.

Step 2: Use the least squares estimates as starting values for an IWLS algorithm with a monotone and bounded ψ. Typically 3 or 4 iterations will be sufficient.

Step 3: Use the results of Step 2 as starting values for an IWLS algorithm based on redescending ψ, which does not iterate σ.

The motivation for the above strategy is rather obvious. It is hoped that the estimates based on a monotone ψ are close to the "appropriate" solution of the estimating equation (II.9) based on a non-monotone ψ.

II.1.3 PROPERTIES

Consistency and Robustness Properties for IO Models

Under regularity conditions consistency and asymptotic normality of $\hat{\beta}$ are obtained for time series containing innovation outliers with finite innovations variance (Martin, 1978a).

The asymptotic covariance matrix of $\hat{\beta}$ is found to be (Martin, 1979)

$$V_{\hat{\beta}} = V_{loc}(\psi, g)\, D^{-1} \tag{II.11}$$

where $V_{loc}(\psi, g) = E_G\psi^2(A)/E_G^2\psi'(A)$ is the asymptotic variance of a location M-estimator at innovation density g (G denotes the innovations distribution) (Huber, 1964) and

$$D = E(\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}) = \left[\begin{array}{c|c} 1 & \mu\,\mathbf{1}^{\mathrm{T}} \\ \hline \mu\,\mathbf{1} & C \end{array}\right] \tag{II.12}$$

with $\mathbf{1}$ a $(p \times 1)$ vector of 1's and C the $(p \times p)$ moment matrix with elements $C_{ij} =$ $=$ covariance (Y_i, Y_j), $1 \leqq i, j \leqq p$.

An inversion formula for partitioned matrices yields

$$V_{\hat{\beta}} = V_{loc}(\psi, g)\left[\begin{array}{c|c} 1 + \mu^2\,\mathbf{1}^{\mathrm{T}}C^{-1}\mathbf{1} & -\mu\,\mathbf{1}^{\mathrm{T}}C^{-1} \\ \hline -\mu C^{-1}\mathbf{1} & C^{-1} \end{array}\right]. \tag{II.13}$$

The 1-1 element of (II.13) is the variance of the intercept λ. The lower-right $(p \times p)$ part of (II. 13), $V_{loc}(\psi, g)\, C^{-1}$, is $V_{\hat{\phi}}$, the covariance matrix of $\hat{\phi}$. The covariance matrix of the least squares estimator $\hat{\phi}$ is contained as a special case in (II.13), because $V_{loc}(\psi, g) = \sigma^2$ for $\psi(t) = t$.

Taking – analogously to (I.14) – the pth root of the ratio of the determinants of the Cramer-Rao lower bound $V_{\phi,CR}$ and of the asymptotic covariance matrix $V_{\hat{\phi}}$ as a multivariate measure of efficiency gives

$$EFF(M, g) = \left(\frac{\det V_{\phi,CR}}{\det V_{\hat{\phi}}}\right)^{1/p} = (V_{loc}(\psi, g)\, i(g))^{-1} \tag{II.14}$$

where $i(g)$ denotes the Fisher information (compare (I.14)). But this is just the pth power of the asymptotic efficiency of a location M-estimator based on ψ, at an error density g. Therefore, an M-estimator of ϕ has the same attractive asymptotic efficiency robustness as a corresponding location M-estimator for i.i.d. data. Martin (1982) treats efficiency robustness of $\hat{\phi}$ in more details.

An M-estimator of ϕ can have far greater precision (i.e. smaller variance) than a least squares estimator, because in (II.13) $C = \sigma^2\bar{C}$ where $\bar{C}$ depends only on ϕ (compare Section I.5) and because with a good choice of ψ the value of $V_{loc}(\psi, g)$ is relatively stable while σ^2 takes on arbitrarily large values for arbitrarily small heavy tailed deviations of g from normality. The M-estimation of an AR(1) model

provides a particularly transparent special case because

$$V_{\hat{\phi}_1} = \frac{1 - \phi_1^2}{\sigma^2} V_{loc}(\psi, g) \tag{II.15}$$

(Denby and Martin, 1979).

If the ψ-function which is used in (II.9) is bounded, then for each fixed $\boldsymbol{\beta}$ the function $f(\mathbf{z}, y) = \psi((y - \mathbf{z}^T\boldsymbol{\beta})/\hat{\sigma})\,\mathbf{z}$ is bounded in the scalar y. but unbounded in $\mathbf{z}$. Correspondingly it turns out that the influence curve for $\hat{\boldsymbol{\beta}}$ is bounded in y and unbounded in $\mathbf{z}$. This feature would be appropriate if one could be sure that the $\mathbf{z}$ portion of the model (II.1) is correctly specified. The $\mathbf{z}$ portion is correctly specified for outlier-free time series and for time series with innovation outliers.

An M-estimator of $\boldsymbol{\phi}$ is – in contrast to a least squares estimator $\hat{\boldsymbol{\phi}}$ – not asymptotically qualitatively robust if innovation outliers are possible because the asymptotic covariance of $\hat{\boldsymbol{\phi}}$ depends on the innovations distribution G. However, this is hardly a serious deficiency because an M-estimate has greater precision than a least squares estimate.

Consistency and Robustness Properties for AO Models

In the presence of additive outliers M-estimators can have an inflated variance and finite sample biases and asymptotic biases which can be as catastrophic as those of least squares estimators (Denby and Martin, 1979; Martin and Jong, 1976).

In the presence of additive outliers an M-estimator of $\boldsymbol{\phi}$ is no longer efficiency robust (Martin, 1979) and has a breakdown point of value zero. The latter fact is not surprising since – in terms of regression analysis – additive outliers produce an errors-in-both-variables problem, and M-estimators do nothing to cope with errors in the "independent" variables (Martin, 1980).

For time series observed with additive outliers there is a $\mathbf{z}$ misspecification in the linear model (II.1). When such deviations from the ideal model are possible the influence curve is unbounded. Compare also Dutter (1980) for details of the influence curve for linear regression.

Summarizing we can say that an M-estimation of AR parameters is advisable if the given series is outlier-free or contain innovation outliers, because in these situations clean asymptotics and efficiency robustness can be achieved. For time series contaminated by additive outliers, however, M-estimation is almost worthless and therefore other methods of estimation are needed.

II.2 ROBUST ESTIMATION OF THE LOCATION OF ARMA MODELS

II.2.1. Autoregressive-Errors M-estimator for AR Models

Since the location μ is related to the intercept λ by (I.6), it is appropriate to call

$$\hat{\mu} = \hat{\lambda}/(1 - \sum_{i=1}^{p} \hat{\phi}_i) \tag{II.16}$$

an *autoregressive-errors M-estimator* of μ, if $\hat{\boldsymbol{\beta}} = (\hat{\lambda}, \hat{\phi}_1, \ldots, \hat{\phi}_p)^T$ is an M-estimator of $\boldsymbol{\beta}$. It can be shown (Martin, 1978b) that $\hat{\mu}$ is a consistent and robust M-estimator for innovation outlier situations if $\hat{\boldsymbol{\beta}}$ is a suitably chosen M-estimator.

Let $\hat{\boldsymbol{\alpha}}$ denote an estimator of $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{p+1})^T = (\mu, \phi_1, \ldots, \phi_p)^T$. Let $\boldsymbol{\beta} = h(\boldsymbol{\alpha})$ denote the transformation from $\boldsymbol{\alpha}$ to $\boldsymbol{\beta}$ and let H denote the matrix of partial derivatives of $h(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$, i.e. $h_{ij} = (\partial/\partial\alpha_j)\, h_i(\boldsymbol{\alpha})$.

If $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal with covariance matrix $V_{\hat{\boldsymbol{\beta}}}$, then $\hat{\boldsymbol{\alpha}}$ is also consistent and asymptotically normal with covariance matrix

$$V_{\hat{\boldsymbol{\alpha}}} = H^{-1} V_{\hat{\boldsymbol{\beta}}} (H^T)^{-1} . \tag{II.17}$$

For the following consideration a finite innovations variance is assumed. The asymptotic covariance matrix of an M-estimator $\hat{\boldsymbol{\beta}}$ is given by (II.11) and therefore, the asymptotic covariance matrix of the corresponding $\hat{\boldsymbol{\alpha}} = h^{-1}(\hat{\boldsymbol{\beta}})$ is

$$V_{\hat{\boldsymbol{\alpha}}} = (H^T D H)^{-1}\, V_{loc}(\psi, g) = \left[\begin{array}{c|c} (1 - \sum_{i=1}^{p} \phi_i)^{-2} & \mathbf{0}^T \\ \hline \mathbf{0} & C^{-1} \end{array}\right] V_{loc}(\psi, g) \tag{II.18}$$

where $V_{loc}(\psi, g)$ is defined as in (II.11) and C is defined as in (II.12).

Martin (1981, 1982) shows that the efficiency of the autoregressive-error M-estimator $\hat{\mu}$ is

$$EFF(\hat{\mu}, \psi, g) = [i(g)\, V_{loc}(\psi, g)]^{-1} . \tag{II.19}$$

Expression (II. 19) is equal to the efficiency of an ordinary location M-estimator for i.i.d. errors. The upper-left element of $V_{\hat{\boldsymbol{\alpha}}}$ (II. 18) differs from the usual location M-estimator asymptotic variance $V_{loc}(\psi, g)$ for i.i.d. errors only by the scale factor $(1 - \sum_{i=1}^{p} \phi_i)^{-2}$. It follows that Huber's (1964) min-max robustness results hold for autoregressive-errors M-estimators of location over families of distributions with finite variances.

The autoregressive errors M-estimator $\hat{\mu}$ (II.16) is almost worthless if a time series contains additive outliers, because the same is true for the M-estimator $\hat{\boldsymbol{\phi}}$. One might use some other procedure, such as the generalized M-estimation (which will be described in Chapter III), to obtain robust estimators $\hat{\lambda}$ and $\hat{\boldsymbol{\phi}}$. A difficulty with the latter approach is that although the estimator of μ will be robust, it will not necessarily be consistent unless $\hat{\lambda}$ and $\hat{\boldsymbol{\phi}}$ are consistent as well as robust. (Generalized M-estimators are robust but not consistent for additive outlier models.) However,

it is possible to construct a robust and consistent estimator of μ from robust but not necessarily consistent estimator $\hat{\lambda}$ and $\hat{\phi}$ (Martin, 1979b).

Lee and Martin (1984) give more information about the computation and the properties of location M-estimators for ARMA models.

II.2.2. Robustly Centered Data

It is also possible to estimate autoregressive and moving-average parameters for observations that were robustly centered by an ordinary location M-estimator as defined by Huber (1964). This is analogous to the usual approach for estimating ARMA parameters by least squares where the sample mean is used to center the data. One might hope that the M-estimator $\hat{\mu}$ is efficiency robust for the outlier models considered here, because the sample mean is asymptotically efficient for a wide class of correlated Gaussian processes (Grenander and Rosenblatt, 1957). While some caution is appropriate until the robustness properties of the M-estimator $\hat{\mu}$ are better understood (see, for example, Wegman and Carroll, 1977), Monte Carlo experience (Zeh, 1979) indicates that use of $\hat{\mu}$ will not impair the robustness properties of the ensuing estimator of ϕ.

II.2.3. Prewhitening-Based M-Estimation of the Location

For the following considerations we assume that observations y_i are generated by random variables $Y_i = X_i + V_i$ (compare (I.9)), where the random variables $X_i + V_i$ are identically and symmetrically distributed but are not necessarily independent.

For simplicity we assume that $X_i + V_i$ is an outlier-free or contaminated AR(1) process, i.e. X_i is an AR(1) process and $V_i \neq 0$ for an additive outlier process. (Generalizations to higher-order autoregressions are straightforward.) A robust estimator $\hat{\phi}_1$ of the first-order autoregressive parameter can be used to prewhiten the observations robustly

$$u_i = y_i - \hat{\phi}_1 y_{i-1}, \quad 2 \leqq i \leqq n, \quad u_1 = y_1. \tag{II.20}$$

The u_i's can be used to compute an ordinary location M-estimate $\hat{\mu}$ by solving (compare Martin, 1981b)

$$\frac{u_1 - \hat{\mu}}{\hat{\sigma}(1 - \hat{\phi}_1)^{1/2}} + \sum_{i=2}^{n} \left[\frac{u_i - \hat{\mu}}{\hat{\sigma}} \right] = 0. \tag{II.21}$$

Then a prewhitening-based M-estimate $\hat{\mu}_p$ of the location can be computed

$$\hat{\mu}_p = \hat{\mu}(1 - \hat{\phi}_1). \tag{II.22}$$

If $\hat{\phi}_1 \xrightarrow{P} \phi_0$, $|\phi_0| < 1$, $\hat{\sigma} \xrightarrow{P} \sigma$ and the distribution of $(X_2 + V_2) - \phi_0(X_1 + V_1)$ is symmetric (what is the case if the V_i's are i.i.d. and have a symmetric distribution), then $\hat{\mu}_p$ is, under reasonable conditions, a consistent and asymptotically normal

estimator of μ. Martin and Lee (1970) found high efficiencies of $\hat{\mu}_p$ relative to ordinary M-estimators of μ.

It might be noted that when $\hat{\phi}_1$ is an M-estimator or generalized M-estimator and $V_i = 0$ for all i, then the asymptotic variance of $\hat{\mu}_p$ is

$$(1 - \phi_1)^{-2} V_{loc}(\psi, g)$$

which is exactly the same as the asymptotic variance obtained by estimating μ with an autoregressive-errors M-estimator (II.18) for a time series which is possibly contaminated by innovation outliers. The main advantage of prewhitening appears in additive outlier situations. Both M-estimators and generalized M-estimators of λ and ϕ are asymptotically biased toward additive outliers and these estimators use no constraints that would insure consistency of an autoregressive-errors M-estimator of μ. However, the implied constraint of the prewhitening step in computing $\hat{\mu}_p$ forces consistency.

II.3 MAXIMUM LIKELIHOOD TYPE ESTIMATION OF ARMA MODELS

II.3.1. Definition

Section I.6 illustrated the distribution-free asymptotic behavior, i.e. the asymptotic qualitative robustness, of least squares estimators of autoregressive moving-average parameters ϕ and θ, but Section I.6 also revealed the lack of efficiency robustness of these estimators in the presence of innovation outliers.

The maximum likelihood estimator defined by (I.20) is asymptotically efficient, but it can be computed only when it happens that the innovations density g is known. Since g in general is not known, it is also possible to use the maximum likelihood estimator in practice. However (I.20) suggests to define the following class of *maximum likelihood type estimators* (*M-estimators*) $\boldsymbol{\alpha}^\wedge = (\boldsymbol{\phi}^{\wedge T}, \boldsymbol{\theta}^{\wedge T}, \hat{\mu})^T$ by

$$g(\boldsymbol{\alpha}', \hat{\sigma}) = \sum_{i=p+1}^{n} \varrho\left[\frac{r_i(\boldsymbol{\alpha}')}{\hat{\sigma}}\right] = \min \tag{II.23}$$

where $r_i(\boldsymbol{\alpha}')$ denotes a residual (I.16). The ϱ-function has the same purpose as that that in (II.2), namely to bound the influence of large residuals on the estimation and therefore the ϱ-functions used for an M-estimation of AR parameters can also be used here.

Similarly as in Section II.1.1 $\boldsymbol{\alpha}$ and σ may be found simultaneously through solving (II.23) and the side condition

$$\frac{1}{n - 2p - q - 1} \sum_{i=p+1}^{n} \psi^2\left(\frac{r_i(\boldsymbol{\alpha}')}{\hat{\sigma}}\right) = b \tag{II.24}$$

if a monotone psi-function – like ψ_H (II.3) – is used. The constant b is the same as in (II.6).

The minimum problem (II.23) and the condition (II.24) can be combined to the more general minimum problem

$$h(\boldsymbol{\alpha}', \sigma') = \sum_{i=p+1}^{n} \varrho\left(\frac{r_i(\boldsymbol{\alpha}')}{\sigma'}\right)\sigma' + c\sigma' = \min . \tag{II.25}$$

where the minimum in both $\boldsymbol{\alpha}'$ and σ' has to be achieved and $c = (n - 2p - q - 1) \cdot b/2$.

If $\psi(t)$ denotes $\mathrm{d}\varrho(t)/\mathrm{d}t$, differentiating of (II.25) with respect to $\boldsymbol{\alpha}'$ and setting the resulting expression equal to zero yield the system of estimation equations

$$\sum_{i=p+1}^{n} \psi\left(\frac{r_i(\boldsymbol{\alpha}^\wedge)}{\hat{\sigma}}\right)\mathbf{d}_i^+(\boldsymbol{\alpha}^\wedge) = \mathbf{0} \tag{II.26}$$

where $\mathbf{d}_i^+(\boldsymbol{\alpha}^\wedge)$ denotes the vector of the first derivatives of the residual $r_i(\boldsymbol{\alpha}^\wedge)$

$$\mathbf{d}_i^+(\boldsymbol{\alpha}^\wedge) = (\mathbf{d}_i^{\mathrm{T}}(\boldsymbol{\alpha}^\wedge), \quad -\partial r_i(\boldsymbol{\alpha}^\wedge)/\partial\mu)^{\mathrm{T}} \tag{II.27}$$

with

$$-\partial r_i(\boldsymbol{\alpha}^\wedge)/\partial\mu = (1 - \hat{\phi}_1 - \ldots - \hat{\phi}_p)/(1 - \hat{\theta}_1 - \ldots - \hat{\theta}_q) \tag{II.28}$$

and

$$\mathbf{d}_i(\boldsymbol{\alpha}^\wedge) = (s_{i-1}(\boldsymbol{\alpha}^\wedge), \ldots, s_{i-p}(\boldsymbol{\alpha}^\wedge), \quad t_{i-1}(\boldsymbol{\alpha}^\wedge), \ldots, t_{i-q}(\boldsymbol{\alpha}^\wedge))^{\mathrm{T}},$$

$$s_{i-j}(\boldsymbol{\alpha}^\wedge) = -\partial r_i(\boldsymbol{\alpha}^\wedge)/\partial\phi_j = \hat{\phi}^{-1}(B)\, r_{i-j}(\boldsymbol{\alpha}^\wedge) \tag{II.29}$$

$$t_{i-j}(\boldsymbol{\alpha}^\wedge) = -\partial r_i(\boldsymbol{\alpha}^\wedge)/\partial\theta_j = -\hat{\theta}^{-1}(B)\, r_{i-j}(\boldsymbol{\alpha}^\wedge) . \tag{II.30}$$

Similar to (II.10) equation (II.26) can also be written as a weighted least squares problem with weights $w_i = \psi(r_i(\boldsymbol{\alpha}^\wedge)/\hat{\sigma})/(r_i(\boldsymbol{\alpha}^\wedge)/\hat{\sigma})$; the least squares problem, however, is nonlinear.

II.3.2. Computational Methods

Before an algorithm to compute M-estimates of $\boldsymbol{\alpha}$ and σ is described, algorithms to compute the first derivatives of the residuals with respect to AR parameters (II.29) and with respect to MA parameters (II.30) are given.

Computation of the first derivatives of the residuals with respect to AR parameters

(1) Set $s_j(\boldsymbol{\alpha}') = 0$ for $j = 1 - p, 2 - p, \ldots, 0$.
(2) Set $i = 1$.
(3) Compute $s_i(\boldsymbol{\alpha}') = \phi_1' s_{i-1}(\boldsymbol{\alpha}') + \ldots + \phi_p' s_{i-p}(\boldsymbol{\alpha}') + r_i(\boldsymbol{\alpha}')$.
(4) Augment $i = i + 1$.
(5) If $i \leqq n - 1$ go to (3), else stop.

Computation of the first derivatives of the residuals with respect to MA parameters

(1) Set $t_j(\alpha') = 0$ for $j = p - 2q + 1,\ p - 2q + 2, \ldots, p - q$.
(2) Set $i = p - q + 1$.
(3) Compute $t_i(\alpha') = \theta_1' t_{i-1}(\alpha') + \ldots + \theta_q' t_{i-q}(\alpha') - r_i(\alpha')$.
(4) Augment $i = i + 1$.
(5) If $i \leqq n - 1$ go to (3), else stop.

Residuals can be computed by the recursive algorithm given in Section I.6.

The minimum problem (II.25) can be solved iteratively by adapting the WS-algorithm (W-Sophisticated) which was applied by Dutter and Huber (1981) for the nonlinear robust regression problem. The WS-algorithm consists of the iterated weighted least squares (IWLS- or W-) algorithm and uses the algorithm of Nagel and Wolff (1974) to solve the nonlinear least squares problem. The algorithm of Nagel and Wolff is based on a linear compromise between the Gauss-Newton procedure and the method of steepest descent. The motivation for the step of the WS-algorithm may be seen in Dutter and Huber (1981).

An Algorithm for the M-estimation of ARMA parameters

Let starting values $\alpha^{(0)}$ and $\sigma^{(0)}$, a tolerance value $\varepsilon > 0$ and a constant c as in (II.25) be given.

1. Compute residuals $\mathbf{r}^{(0)} = (r_{p+1}(\alpha^{(0)}), \ldots, r_n(\alpha^{(0)}))^T$.
2. Set the iteration counter $m = 0$.
3. Find an improved scale

$$(\sigma^{(m+1)})^2 = \frac{1}{c} \sum_{i=p+1}^{n} \chi\left(\frac{r_i(\alpha^{(m)})}{\sigma^{(m)}}\right)(\sigma^{(m)})^2$$

where $\chi(t) = t\,\chi(t) - \varrho(t)$.

4. Calculate weights

$$w_i^{(m)} = \psi(r_i(\alpha^{(m)})/\sigma^{(m+1)})/(r_i(\alpha^{(m)})/\sigma^{(m+1)}),$$

if $r_i(\alpha^{(m)}) \neq 0$, otherwise $w_i^{(m)} = 1$, $i = p + 1, \ldots, n$; define a diagonal matrix $W^{(m)}$ with $w_i^{(m)}$ as its $(i - p)$th diagonal element.

5. Compute the first derivatives of the residuals with respect to the parameters

$$s_1(\alpha^{(m)}) = -\partial r_{p+1}(\alpha^{(m)})/\partial\phi_p, \ldots, s_{n-1}(\alpha^{(m)}) = -\partial r_n(\alpha^{(m)})/\partial\phi_1,$$
$$t_{p+1-q}(\alpha^{(m)}) = -\partial r_{p+1}(\alpha^{(m)})/\partial\theta_p, \ldots, t_{n-1}(\alpha^{(m)}) = -\partial r_n(\alpha^{(m)})/\partial\theta_1,$$
$$-\partial r_i(\alpha^{(m)})/\partial\mu = (1 - \phi_1^{(m)} - \ldots - \phi_p^{(m)})/(1 - \theta_1^{(m)} - \ldots - \theta_q^{(m)}),$$

$i = p + 1, \ldots, n$, and for $i = p + 1, \ldots, n$ form the vectors

$$\mathbf{d}_i(\alpha^{(m)}) = (s_{i-1}(\alpha^{(m)}), \ldots, s_{i-p}(\alpha^{(m)}), t_{i-1}(\alpha^{(m)}), \ldots, t_{i-q}(\alpha^{(m)}))^T$$

and

$$\mathbf{d}_i^+(\boldsymbol{\alpha}^{(m)}) = \left(\mathbf{d}_i^{\mathrm{T}}(\boldsymbol{\alpha}^{(m)}),\ -\partial r_i(\boldsymbol{\alpha}^{(m)})/\partial\mu\right)^{\mathrm{T}} .$$

Let $D^{(m)}$ denote the matrix $\left[\mathbf{d}_{p+1}^+(\boldsymbol{\alpha}^{(m)}), \ldots, \mathbf{d}_n^+(\boldsymbol{\alpha}^{(m)})\right]^{\mathrm{T}}$.

6. Solve

$$\sum_{i=p+1}^{n} \left(r_i(\boldsymbol{\alpha}^{(m)}) - \mathbf{d}_i^{+\mathrm{T}}(\boldsymbol{\alpha}^{(m)})\,\boldsymbol{\tau}^{(m)}\right)^2 w_i^{(m)} = \min .$$

for the (approximate) direction $\boldsymbol{\tau}^{(m)}$ of the Gauss-Newton method, i.e. solve

$$H^{(m)}\boldsymbol{\tau}^{(m)} = \gamma^{(m)}$$

where $H^{(m)} = D^{(m)\mathrm{T}}W^{(m)}D^{(m)}$ and $\gamma^{(m)} = D^{(m)\mathrm{T}}W^{(m)}\mathbf{r}^{(m)}$ denotes the vector of steepest descent and $\mathbf{r}^{(m)}$ denotes $(r_{p+1}(\boldsymbol{\alpha}^{(m)}), \ldots, r_n(\boldsymbol{\alpha}^{(m)}))^{\mathrm{T}}$.

7. Calculate $q_1 = g(\boldsymbol{\alpha}^{(m)}, \sigma^{(m+1)})$, new values for $\boldsymbol{\alpha}$ by $\boldsymbol{\alpha}^{(m+1)} = \boldsymbol{\alpha}^{(m)} + \boldsymbol{\tau}^{(m)}$, $\mathbf{r}^{(m+1)}$ and $g_2 = g(\boldsymbol{\alpha}^{(m+1)}, \sigma^{(m+1)})$. If $q_2 < g_1$ go to 8.

7a. Compute

$$\omega = \frac{\gamma^{(m)\mathrm{T}}\gamma^{(m)}}{\gamma^{(m)\mathrm{T}}H^{(m)}\gamma^{(m)}},$$

which is an approximative value so that g takes its minimum in the direction of

$$v\omega\gamma^{(m)} + (1 - v)\,\boldsymbol{\tau}^{(m)} \quad \text{with} \quad 0 < v \leqq 1 .$$

Perform the following steps.

(1) Set $\iota = 0$.
(2) Augment $\iota = \iota + 1$.
(3) Compute new values for $\boldsymbol{\alpha}$ by

$$\boldsymbol{\alpha}^{(m+1)} = \boldsymbol{\alpha}^{(m)} + \iota\omega\gamma^{(m)}/5 + (1 - \iota/5)\,\boldsymbol{\tau}^{(m)} .$$

(4) Compute $\mathbf{r}^{(m+1)}$ and $g_{2+\iota} = g(\boldsymbol{\alpha}^{(m+1)}, \sigma^{(m+1)})$.
(5) If $g_{2+\iota} < g_1$, go to 7b.
(6) If $\iota < 5$ go to (2), else go to 7c.

7b. Compute a linear back-interpolation between $(\iota - 1)/5$ and $\iota/5$ by

$$v^* = \frac{\iota}{5} - \frac{g_1 - g_{2+\iota}}{g_{1+\iota} - g_{2+\iota}}\,\frac{1}{5}.$$

Put $\boldsymbol{\alpha}^{(m+1)} = \boldsymbol{\alpha}^{(m)} + v^*\omega\gamma^{(m)} + (1 - v^*)\,\boldsymbol{\tau}^{(m)}$.
Compute residuals $\mathbf{r}^{(m+1)}$.
If $g(\boldsymbol{\alpha}^{(m+1)}, \sigma^{(m+1)}) < g_1$, go to 8.
Otherwise put $\boldsymbol{\alpha}^{(m+1)} = \boldsymbol{\alpha}^{(m)} + \iota\omega\gamma^{(m)}/5 + (1 - \iota/5)\,\boldsymbol{\tau}^{(m)}$,
Compute residuals $\mathbf{r}^{(m+1)}$ and go to 8.

7c. Put $\omega = \omega/2$.
Compute $\boldsymbol{\alpha}^{(m+1)} = \boldsymbol{\alpha}^{(m)} + \omega\mu^{(m)}$ and residuals $\mathbf{r}^{(m+1)}$. If $g(\boldsymbol{\alpha}^{(m+1)}, \sigma^{(m+1)}) \geqq g_1$, repeat step 7c.

8. Stop, if

$$|\sigma^{(m+1)} - \sigma^{(m)}| < \varepsilon\sigma^{(m+1)}$$

and all the differences between the parameters are less than ε times their approximate standard deviation, i.e. if

$$|\alpha_k^{(m+1)} - \alpha_k^{(m)}| < \varepsilon\sigma^{(m+1)}\sqrt{d^{kk}} \quad \text{for all} \quad k = 1, \ldots, p + q + 1,$$

where α_k denotes the kth element of $\boldsymbol{\alpha}$ and d^{kk} denotes the kth diagonal element of $[D^{(m)\mathrm{T}}D^{(m)}]^{-1}$.

9. Augment $m = m + 1$ and go to 3.

The algorithm which is described above, can only be used with a monotone ψ-function. For a redescending ψ an estimated value for the scale must be given and the improvement of the scale (Step 3) must be omitted. An illustrative application of this algorithm to compute M-estimates of ARMA parameters for simulated data is given in Stockinger (1983).

II.3.3. Properties

Under general regularity conditions it may be proved (Martin and Yohai, 1984a) that for an M-estimator of $\boldsymbol{\alpha}$

$$\sqrt{(n)}\,(\boldsymbol{\alpha}^{\wedge} - \boldsymbol{\alpha}) \xrightarrow{\mathrm{d}} N(\mathbf{0},\, k(\psi, G)\, K(\boldsymbol{\phi}, \boldsymbol{\theta}, G)) \tag{II.31}$$

where $K(\boldsymbol{\phi}, \boldsymbol{\theta}, G)$ is given by (I.17) and the form of $k(\psi, G)$ is the same as in (I.21), except for Ψ being replaced by a general psi-function ψ. More details are given by Lee and Martin (1982).

The ratio of the trace of V_{ML} (I.21) to the trace of the asymptotic covariance matrix of $\boldsymbol{\alpha}^{\wedge}$ gives $[i(G)\, V_{loc}(\psi, G)]^{-1}$ as a measure for multivariate efficiency which is just the asymptotic efficiency expression for a location M-estimator and $\boldsymbol{\alpha}^{\wedge}$ therefore has the same attractive asymptotic efficiency robustness properties as a location M-estimator based on the same psi-function ψ.

If the variance of G is large, then the variance of an M-estimator for $(\boldsymbol{\phi}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})$ for a good ψ, is smaller than the variance of a least squares estimator. In contrast, it follows from (II.31), that the M-estimator for $\boldsymbol{\alpha}$ is not asymptotically qualitatively robust.

Additive outliers however, can cause not only inflated variability of M-estimators, but also considerable bias, even asymptotically.

II.4 METHODS USED IN THE ADDITIVE OUTLIERS CASE

As we have seen in Section II.1.3 M-estimation is not satisfactory if a time series might contain additive outliers. In this section two methods for dealing with i.i.d. additive outliers will be outlined. (The next chapter will treat a more powerful method.)

II.4.1. Pagano's Method

In this section it is assumed that X_i (I.9) is an outlier-free pth order autoregressive process with location $\mu = 0$.

If the V_i's are independent and Gaussian with variance σ_V^2, it is possible to construct parameter estimators of $\boldsymbol{\phi}$, σ^2 and σ_V^2 which are not only consistent but are also asymptotically efficient. Pagano's (1974) method of doing this is as follows:

Apply the X_i-whitening transformation to the observations y_i which yields

$$u_i = y_i - \phi_1 y_{i-1} - \ldots - \phi_p y_{i-p} =$$
$$= (x_i + v_i) - \phi_1(x_{i-1} + v_{i-1}) - \ldots - \phi_p(x_{i-p} + v_{i-p}) =$$
$$= x_i - \phi_1 x_{i-1} - \ldots - \phi_p x_{i-p} + v_i - \phi_1 v_{i-1} - \ldots - \phi_p v_{i-p} =$$
$$= a_i + v_i - \phi_1 v_{i-1} - \ldots - \phi_p v_{i-p}\,.$$

The last line reveals u_i is produced by an MA(p) model. Then it follows that there exists a white noise sequence of random variables η_i with an $N(0, \sigma_\eta^2)$ distribution and there exist constants $\theta_1, \theta_2, \ldots, \theta_p$, so that

$$u_i = \eta_i - \theta_1 \eta_{i-1} - \ldots - \theta_p \eta_{i-p}\,.$$

Thus Y_i is an ARMA(p, p) process with parameters $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_p, \sigma_\eta^2$.

The parameters $\theta_1, \ldots, \theta_p, \sigma_\eta^2$ could be determined by the covariances $c(k) = EU_iU_{i+k}$, $k = 0, 1, \ldots, p$. Thus the process Y_i is equivalently parameterized by $\phi_1, \ldots, \phi_p, c(0), \ldots, c(p)$.

Although consistent and asymptotically efficient estimators of the above equivalent parameter sets are available (Hannan, 1970, 1973; Parzen, 1971), they do not directly provide efficient estimators of the original parameters $\phi_1, \ldots, \phi_p, \sigma^2, \sigma_V^2$. Pagano obtains efficient estimators of these parameters by a least squares regression of the estimates $\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{c}(0), \ldots, \hat{c}(p)$ on $\phi_1, \ldots, \phi_p, \sigma^2, \sigma_V^2$ using the relations

$$c(k) = \sigma^2 \delta_{0,k} + \sigma_V^2 \sum_{j=0}^{p} \phi_j \phi_{j+k}\,, \quad k = 0, 1, \ldots, p$$

where $\delta_{0,k} = 0$ for $k \neq 0$, $\delta_{0,0} = 1$ and $\phi_j = 0$ for $j = 0$ or $j > p$.

II.4.2. A Robust Instrumental Variables Approach

We consider a special case of (I.9) in which X_i is a first-order autoregressive process with location $\mu = 0$ and the V_i's are i.i.d. For the linear model

$$y_i = \phi_1 y_{i-1} + u_i \tag{II.32}$$

we have (because $y_i = x_i + v_i$ and $x_i = \phi_1 x_{i-1} + a_i$)

$$u_i = y_i - \phi_1 y_{i-1} = x_i + v_i - \phi_1 y_{i-1} = \phi_1 x_{i-1} + a_i + v_i - \phi_1 y_{i-1} =$$
$$= a_i + v_i - \phi_1(y_{i-1} - x_{i-1}) = a_i + v_i - \phi_1 v_{i-1}\,.$$

Thus the usual linear regression approach does not yield a consistent estimate of ϕ_1 because $E(U_i \mid Y_{i-1}) \neq 0$. In fact, as mentioned in Section I.5, the asymptotic bias of the least squares estimator of ϕ_1 is $-\phi_1\sigma_V^2/(\sigma_X^2 + \sigma_V^2)$.

However, the least squares *instrumental variable* (IV) approach is appropriate in this case (Walker, 1960, Martin, 1981b). Because $E(U_i \mid Y_{i-2}) = 0$, Y_{i-2} serves as an instrumental variable and

$$\hat{\phi}_1 = \frac{\sum_{i=3}^{n} y_i y_{i-2}}{\sum_{i=3}^{n} y_{i-1} y_{i-2}}$$

is a consistent estimator of ϕ_1, if $\phi_1 \neq 0$. Notice the difference between $\hat{\phi}_1$ and the usual least squares estimator of ϕ_1 which can be computed by

$$\left(\sum_{i=2}^{n} y_i y_{i-1}\right) / \left(\sum_{i=2}^{n} y_{i-1}^2\right).$$

The least squares instrumental variable estimator $\hat{\phi}_1$ can be robustified easily. An *instrumental variable generalized M-estimate* $\hat{\phi}_{IV}$ is obtained by solving

$$\sum_{i=3}^{n} \left(\frac{y_{i-2}}{\hat{\sigma}_X}\right) W\left(\frac{y_{i-2}}{\hat{\sigma}_X}\right) \psi\left(\frac{y_i - y_{i-1}\hat{\phi}_{IV}}{\hat{\sigma}}\right) = 0.$$

The weight function $W(t)$ should be chosen so that $t\,W(t)$ is bounded. $\hat{\sigma}_X$ denotes a robust scale estimate which might be computed directly from the data. The robust scale estimate $\hat{\sigma}$ is computed from an auxiliary equation.

Under regularity conditions and if $\phi_1 \neq 0$ the estimator $\hat{\phi}_{IV}$ is consistent and asymptotically normal, even in the presence of additive outliers.

III. GENERALIZED MAXIMUM LIKELIHOOD TYPE ESTIMATION

An appropriate generalization of the maximum likelihood type (M-) method yields more satisfactory estimates of ARMA parameters in the case that the given time series is contaminated by additive outliers. Definitions, computational methods and properties of generalized maximum likelihood type estimators (GM-estimators) for pure autoregressive models as well as for ARMA models will be dealt with. In additive outlier situations GM-estimators have, in particular, the following properties. GM-estimators do not require i.i.d. outliers. GM-estimators have a positive breakdown point, a bounded influence curve, considerable robustness and much smaller bias than M-estimators and least squares estimators.

The properties of M-estimators and GM-estimators of AR parameters can be used to create tests which are able to determine the type of outliers in a time series.

Robustified methods for the identification of AR models and ARIMA models will be mentioned.

III.1 GENERALIZED MAXIMUM LIKELIHOOD TYPE ESTIMATION OF AR MODELS

We now concentrate on estimating an autoregressive model of order p (AR(p) model) (I.5). First we center the data robustly (compare Section II.2.2) by using an ordinary location maximum likelihood type (M-)estimator $\hat{\mu}$, that is defined by

$$\sum_{i=1}^{n} \varrho\left(\frac{y_i - \mu'}{\hat{\sigma}_y}\right) = \min . \tag{III.1}$$

where $\varrho(\cdot)$ is a symmetric robustifying loss function and $\hat{\sigma}_y$ is an estimate of the scale of the y_i's (Huber, 1964). Some explanations to ϱ-functions and to their first derivatives, denoted by ψ, were already given in Section II.1.1. If a robustly centered observation is – for notational convenience – again denoted by y_i, then the AR(p) model can be written in the linear model form

$$\mathbf{y} = Z\boldsymbol{\phi} + \mathbf{a} \tag{III.2}$$

where $\mathbf{y} = (y_{p+1}, \ldots, y_n)^T$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T$, $\mathbf{a} = (a_{p+1}, \ldots, a_n)^T$ and $Z = [\mathbf{z}_{p+1}, \ldots, \mathbf{z}_n]^T$ with $\mathbf{z}_i = (y_{i-1}, \ldots, y_{i-p})^T$.

M-estimation of AR models (compare Section II.1) is an attractive possibility to obtain asymptotic efficiency robustness in situations where only innovation outliers

are present. However, M-estimators have a breakdown point of value zero and large biases in the case of additive outliers and lack qualitative robustness toward both outlier situations. Since additive outliers occur probably much more frequently than innovation outliers, other methods of estimation are strongly required. Two methods for dealing with i.i.d. additive outliers were mentioned in Section II.4. Here we will be concerned with a more general method. As stated in Section II.1.3 the influence curve of M-estimators is bounded in y, but unbounded in $\mathbf{z}$. This is an undesirable property if additive outliers occur. The possibility of bounding the influence curve in y and also in $\mathbf{z}$ for usual regression problems has been alluded to by Huber (1973), suggested by Mallows (1976) and advocated by Hampel (1973, 1975).

III.1.1. Definition

The basic idea of generalized M-estimators (GM-estimators) is to modify the minimum problem (II.7) so that the summands of the estimating equation (II.9) are bounded and continuous functions of the data. This in turn results in an influence curve which has the same properties. GM-estimators $\hat{\boldsymbol{\phi}}$ and $\hat{\sigma}$ are analogues of bounded-influence regression estimators and are given as an extention of Huber's (1973) proposal for robust regression by the general minimum problem

$$h(\boldsymbol{\phi}', \sigma') = \sum_{i=p+1}^{n} u_i v_i \varrho_1\left(\frac{y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\phi}'}{u_i \sigma'}\right)\sigma' + c\sigma' = \min . \tag{III.3}$$

where the minimum in $\boldsymbol{\phi}'$ and σ' has to be achieved and $\psi_1(t)$, the first derivative of $\varrho_1(t)$, should be monotone, e.g. $\psi_1(t) = \psi_H(t)$ (II.3). $\hat{\sigma}$ is consistent for σ if the y_i's are free of outliers with $N(0, \sigma^2)$-distributed innovations and $\hat{\mu} = \mu$, if $c = (n - 2p)\, E u_i v_i E_{N(0,1)} \psi_1^2(A)/2$ where A is a random variable with an $N(0, 1)$ distribution. The u_i's and v_i's are weights depending on the "largeness" of $\mathbf{z}_i$. Dutter (1983b) uses an equation like (III.3) to compute bounded-influence estimators for linear regression. Differentiating $h(\boldsymbol{\phi}', \sigma')$ with respect to σ' and $\boldsymbol{\phi}'$ and equating the resulting expressions to zero yield

$$\sum_{i=p+1}^{n} u_i v_i \chi_1\left(\frac{y_i - \mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\phi}}}{u_i \hat{\sigma}}\right) = c \tag{III.4}$$

with $\chi_1(t) = t\,\psi_1(t) - \varrho_1(t)$ and a system of equations defining $\hat{\boldsymbol{\phi}}$ for a known σ

$$\sum_{i=p+1}^{n} v_i \psi_1\left(\frac{y_i - \mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\phi}}}{u_i \hat{\sigma}}\right)\mathbf{z}_i = \mathbf{0} . \tag{III.5}$$

Equation (III.5) defines the least squares estimator of $\boldsymbol{\phi}$ if $v_i = u_i = 1$ for all i and if ψ_1 is the identity function. In contrast to the least squares estimator the influence of the residuals $y_i - \mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\phi}}$ and therefore also the bad influence of innovation outliers is bounded for a good choice of ψ_1 in (III.5). For a *Mallows type GM-estimator* (Mallows, 1976) every u_i is equal to 1 and $v_i = \psi_2(b_i)/b_i$, where b_i denotes

the "largeness' of $\mathbf{z}_i$ and ψ_2 is for example one of the ψ-functions (II.3) to (II.5). The v_i's should bound the influence of the $\mathbf{z}_i$'s and therefore also the influence of additive outliers on the estimation. A *Schweppe type GM-estimator* (Schweppe, 1975) uses $u_i = v_i = \psi_2(b_i)/b_i$ to increase the influence of an observation with a small v_i if the residual $y_i - \mathbf{z}_i^T \hat{\boldsymbol{\phi}}$ is also small. Therefore a Schweppe type GM-estimator should be superior to a Mallows type GM-estimator with the same psi-functions if innovation outliers are present.

III.1.2. Computational Methods

First we will describe possibilities to assess the "largeness" b_i of $\mathbf{z}_i$ and then we will explain a method to compute GM-estimates of AR parameters.

The "largeness" b_i of $\mathbf{z}_i$ can be assessed by

$$b_i = (p^{-1}\mathbf{z}_i^T \hat{C}^{-1}\mathbf{z}_i)^{1/2} \tag{III.6}$$

where $\hat{C}^{-1}$ is an estimate of the a priori unknown inverse $p*p$ covariance matrix of the outlier-free process X_i which is the basis of Y_i (compare Section I.4). Martin (1980) estimates C^{-1} in the following way: Suppose that X_i is a Gaussian process (not necessarily a pth-order autoregression) with $p*p$ covariance matrix C_p and let $\phi_{k1}, \ldots, \phi_{kk}$, $k = 1, 2, \ldots, p-1$, be the coefficients of the predictors of X_i based on $X_{i-1}, \ldots, X_{i-k}$ with the minimum mean square error. Denote the corresponding prediction-error variance by $\sigma^2(k)$. Then C_p^{-1} has the factorization (Akaike, 1969)

$$C_p^{-1} = S_p^T S_p$$

where

$$(S_p)_{kj} = \begin{cases} \dfrac{-\phi_{p-k,j-k}}{\sigma(p-k)}, & j > k \\ \dfrac{1}{\sigma(p-k)}, & j = k \\ 0, & j < k \end{cases}$$

with $1 \leqq k, j \leqq p$ and $\sigma(0)$ denoting the scale of the X_i's. $\sigma(0)$ could be estimated by

$$\hat{\sigma}_x = \underset{i}{\text{med}}\, |y_i - \underset{j}{\text{med}}\, y_j| / .6745 . \tag{III.7}$$

Assuming that AR models of order $p = 1, 2, \ldots, p_{max}$ are fitted in succession using GM-estimates, set $\hat{C}_p^{-1} = \hat{S}_p^T \hat{S}_p$ where $\hat{S}_p$ is obtained from S_p by replacing ϕ_{kj} by its GM-estimate and replacing $\sigma(k)$ by the appropriate GM-estimate of scale.

To estimate the first-order AR parameter ϕ_1, the system (III.5) becomes

$$\sum_{i=2}^{n} v_i \psi_1 \left(\frac{y_i - y_{i-1}\hat{\phi}_1}{u_i \hat{\sigma}} \right) y_{i-1} = 0 \tag{III.8}$$

where the "largeness" to compute v_i and u_i could be

(III.9) $$b_i = \hat{\sigma}_x^{-1} y_{i-1}$$

with $\hat{\sigma}_x$ defined by (III.7).

Another possibility to estimate the inverse covariance matrix C^{-1} is to express C^{-1} as a function of ϕ, $C^{-1} = C^{-1}(\phi)$, using Siddiqui's (1958) results and then set $\hat{C}^{-1} = C^{-1}(\phi\hat{})$, where $\phi\hat{}$ denotes a GM-estimator. But this method creates extreme difficulties in establishing existence and uniqueness for solutions of the estimating equations.

A special formulation of equation (III.5) reveals a GM-estimator $\phi\hat{}$ as a weighted least squares estimator whose weights depend on the residuals $r_i = y_i - \mathbf{z}_i^{\mathrm{T}}\phi\hat{}$, $i = p+1, \ldots, n$ (compare (II.9)). It follows that an *iterated weighted least squares* (IWLS) algorithm could be used to estimate ϕ and σ simultaneously. Before starting the IWLS algorithm, the weights u_i and v_i, $i = p+1, \ldots, n$, which are – in terms of linear regression – weights in the factor space (and which are constant for a fixed time series $y_1, \ldots, y_n$), must be determined.

The IWLS algorithms for the M-estimation (compare Section II.1.2) and GM-estimation of AR parameters are in general similar, but the improvement of the scale and the calculation of the weights are different. Thus the IWLS algorithm described in Section II.1.2 can be used here if it uses model (III.2) instead of model (II.1) (β has to be replaced by ϕ) and Step 3 and Step 4 must be newly formulated:

3. Compute a new value for σ using (III.4)

$$(\sigma^{(m+1)})^2 = \frac{1}{c} \sum_{i=p+1}^{n} u_i v_i \chi_1\left(\frac{r_i^{(m)}}{u_i \sigma^{(m)}}\right)(\sigma^{(m)})^2 .$$

4. Calculate weights considering that $u_i = 1$ for a Mallows type estimator and $u_i = v_i$ for the Schweppe type estimator

$$w_i^{(m)} = \begin{cases} v_i \psi_1\left(\dfrac{r_i^{(m)}}{u_i \sigma^{(m+1)}}\right) \Big/ \left(\dfrac{r_i^{(m)}}{\sigma^{(m+1)}}\right), & \text{if } r_i^{(m)} \neq 0, \quad u_i \neq 0 \\ v_i / u_i, & \text{if } r_i^{(m)} = 0, \quad u_i \neq 0 \\ 1, & \text{if } r_i^{(m)} = u_i = v_i = 0 \\ 1, & \text{if } r_i^{(m)} \neq 0, \quad u_i = v_i = 0, \quad \psi_1(t) = t \\ 0, & \text{if } r_i^{(m)} \neq 0, \quad u_i = v_i = 0, \quad \psi_1 \text{ is bounded} \end{cases}$$

where $i = p+1, \ldots, n$. Define a diagonal matrix $W^{(m)}$ with w_i as its $(i-p)$th diagonal element.

For a non-monotone ψ_1-function the IWLS algorithm described above must be modified in the same way as the IWLS algorithm to compute M-estimates, i.e. an estimated value for the scale must be given and Step 3 must be omitted. The overall computational strategy which was described in Section II.1.2, should be used because the estimating equations (III.4), (III.5) could have multiple roots for a non-monotone ψ_1.

III.1.3. Properties

Asymptotics and robustness properties were mainly investigated for the Mallows type solution of (III.3) and will be described for the Mallows type estimator, if not otherwise stated.

The GM-estimator $(\hat{\boldsymbol{\phi}}^{\mathrm{T}}, \hat{\sigma})$ may be represented as a functional in the following manner. Define a multivariate sample by $(Y_i, Y_{i-1}, \ldots, Y_{i-p})$, $i = p+1, \ldots, n$, and let F_n denote the empirical distribution function for the sample. Let F denote the multivariate distribution function for

$$\mathbf{U}_1^{\mathrm{T}} = (Y_{p+1}, \mathbf{Z}_{p+1}^{\mathrm{T}}) = (Y_{p+1}, Y_p, \ldots, Y_1) \tag{III.10}$$

where Y_i denotes a random variable representing an observation centered by the functional $\mu(F)$, i.e. $\mu(F)$ is the true location parameter and $\hat{\mu} = \mu(F_n)$ is a (robust) location estimate used to center time series data.

The GM-estimator $(\hat{\boldsymbol{\phi}}, \hat{\sigma})$ could be defined by the functional $(\boldsymbol{\phi}^{\mathrm{T}}(F), \sigma(F)) = T_{\boldsymbol{\phi},\sigma}(F)$ whose value is the root of

$$E_F\, V(\mathbf{Z}_{p+1})\, \mathbf{Z}_{p+1} \psi_1^2 \left(\frac{Y_{p+1} - \mathbf{Z}_{p+1}^{\mathrm{T}} \hat{\boldsymbol{\phi}}}{\hat{\sigma}} \right) = \mathbf{0} \tag{III.11}$$

and

$$E_F\, V(\mathbf{Z}_{p+1}) \left\{ \psi_1^2 \left(\frac{Y_{p+1} - \mathbf{Z}_{p+1}^{\mathrm{T}} \hat{\boldsymbol{\phi}}}{\hat{\sigma}} \right) - b_1 \right\} = 0 \tag{III.12}$$

where $V(\mathbf{Z}_{p+1})$ denotes a weight depending on $\mathbf{Z}_{p+1}$. Note that equation (III.12) is the GM-estimation version of the side condition (II.6) proposed by Huber (1973) for the M-estimation.

The values of $\boldsymbol{\phi}(F_n)$ and $\sigma(F_n)$ could be obtained by solving (III.3).

Consistency

GM-estimators defined by (III.3) are consistent and asymptotically normal even in innovation outlier situations without Gaussian or finite variance assumptions under reasonable regularity conditions (Martin, 1978c; Bustos, 1982). For well chosen ψ_1- and ψ_2-functions GM-estimators have much smaller biases than least squares estimators or M-estimators at additive outlier models. Evidence in support of this statement may be found in Martin and Zeh (1978) and Zeh (1979). Some Monte Carlo results for the AR(1) model were presented by Denby and Martin (1979).

The Asymptotic Covariance Matrix

If innovation outliers are possible the joint asymptotic covariance matrix of $\hat{\boldsymbol{\phi}}$ and $\hat{\sigma}$ is (Martin, 1980)

$$V_{\boldsymbol{\phi},\sigma} = \left[\begin{array}{c|c} V_{\boldsymbol{\phi}} & \mathbf{0} \\ \hline \mathbf{0}^{\mathrm{T}} & V_{\sigma} \end{array}\right] \tag{III.13}$$

with

(III.14) $$V_{\phi} = B_1^{-1} B_2 B_1^{-1} \, V_{loc}(\psi, g)$$

where

(III.15) $$B_1 = E\{\mathbf{Z}_{p+1} \, V(\mathbf{Z}_{p+1}) \, \mathbf{Z}_{p+1}^{\mathrm{T}}\},$$
$$B_2 = E\{\mathbf{Z}_{p+1} \, V^2(\mathbf{Z}_{p+1}) \, \mathbf{Z}_{p+1}^{\mathrm{T}}\}$$

and $V_{loc}(\psi, g)$ is the asymptotic variance of a location M-estimator. The expression for V_{σ} is

(III.16) $$V_{\sigma} = \frac{\sigma^2(F)}{d^2} \, E_F \, V^2(\mathbf{Z}_{p+1}) \, E_F\{\psi_1^2[A_{p+1}/\sigma(F)] - b_1\}^2$$

where d is defined as in (III.21).

If $V(\mathbf{Z}_{p+1}) = 1$ the GM-estimator reduces to an M-estimator and the resulting asymptotic covariance matrix for $\hat{\phi}$ is just the lower right part of (III.13). If, in addition, ψ_1 is the identity function we get the least squares covariance matrix (I.13).

Efficiency Robustness

On the one hand, GM-estimators have good efficiency robustness relative to least squares estimators in innovation outlier situations, on the other their efficiency robustness can be poor relative to M-estimators. This is to be expected considering the fact that using the weights v_i in (III.3) results in increased variability relative to M-estimators. The asymptotic variances of an M-estimator and a GM-estimator of the first-order autoregressive parameter are, respectively (compare (III.14)),

(III.17) $$V_{\phi_1, M} = \frac{1}{EY_1^2} \, V_{loc}(\psi_1, g)$$

and

(III.18) $$V_{\phi_1, GM} = \frac{E\psi_2^2(Y_1)}{E^2 Y_1 \, \psi_2(Y_1)} \, V_{loc}(\psi_1, g)$$

where $V_{loc}(\psi_1, g)$ is the asymptotic variance of an M-estimator for location at innovations density g. Therefore the efficiency of the Mallows type GM-estimator relatively to the M-estimator is

(III.19) $$EFF(GM, M) = \frac{V_{\phi_1, M}}{V_{\phi_1, GM}} = \varrho^2_{Y_1, \psi_2(Y_1)}$$

where $\varrho_{Y_1, \psi_2(Y_1)}$ is the correlation coefficient for Y_1 and $\psi_2(Y_1)$. The function ψ_2 will typically be chosen so that $\varrho^2_{Y_1, \psi_2(Y_1)}$ is moderately large for an outlier-free time series – say $\varrho^2_{Y_1, \psi_2(Y_1)} = .95$. The value $\varrho_{Y_1, \psi_2(Y_1)}$, however, can be rather small for some innovation outlier model (Denby and Martin, 1979), which results in considerable loss of efficiency.

Schweppe type GM-estimators offer some hope for obtaining better efficiency robustness at innovation outlier situations.

GM-estimators provide considerable robustness toward additive outliers with modest losses of efficiency relative to M-estimators in "gentle" innovation outlier situations (Martin, 1980). Since innovation outlier situations probably occur infrequently, GM-estimation is an attractive possibility. Favorable small sample efficiency robustness of GM-estimators for the first-order autoregressive parameter is reported in the Denby and Martin (1979) Monte Carlo study.

The Influence Curve

Let $\mathbf{U}'^{\mathrm{T}} = (Y'_{p+1}, \mathbf{Z}'^{\mathrm{T}}_{p+1})$ be a dummy variable replacement for $\mathbf{U}_1^{\mathrm{T}}$ defined by (III.10). The influence curve $IC_{T_{\phi,\sigma},F}(\mathbf{U}')$ of the joint GM-estimator of ϕ and σ can be computed in a straightforward manner (Martin, 1980). The calculaton is simplified by replacing $\mu(F)$ by μ, thus acting as if the location parameter were known. Define $R(F) := Y_{p+1} - \mathbf{Z}_{p+1}^{\mathrm{T}}\,\phi(F)$ and do similarly for $R'(F)$. Then the calculation yields

$$\text{(III.20)} \qquad IC_{T_{\phi,\sigma},F}(\mathbf{U}') = \begin{bmatrix} D & \mathbf{e} \\ \mathbf{f}^{\mathrm{T}} & d \end{bmatrix}^{-1} \begin{bmatrix} V(\mathbf{Z}'_{p+1})\,\mathbf{Z}'_{p+1}\,\sigma(F)\,\psi_1\left(\dfrac{R'(F)}{\sigma(F)}\right) \\ V(\mathbf{Z}'_{p+1})\,\sigma(F)\left\{\psi_1^2\left(\dfrac{R'(F)}{\sigma(F)}\right) - b_1\right\} \end{bmatrix}$$

where

$$\text{(III.21)} \qquad \begin{aligned} D &= E_F\left\{V(\mathbf{Z}_{p+1})\,\mathbf{Z}_{p+1}\mathbf{Z}_{p+1}^{\mathrm{T}}\psi'_1\left(\frac{R'(F)}{\sigma(F)}\right)\right\} \\ \mathbf{e} &= E_F\left\{V(\mathbf{Z}_{p+1})\,\mathbf{Z}_{p+1}\frac{R(F)}{\sigma(F)}\,\psi'_1\left(\frac{R(F)}{\sigma(F)}\right)\right\} \\ \mathbf{f} &= 2E_F\left\{V(\mathbf{Z}_{p+1})\,\mathbf{Z}_{p+1}\psi_1\left(\frac{R(F)}{\sigma(F)}\right)\psi'_1\left(\frac{R(F)}{\sigma(F)}\right)\right\} \\ d &= 2E_F\left\{V(\mathbf{Z}_{p+1})\,\frac{R(F)}{\sigma(F)}\,\psi_1\left(\frac{R(F)}{\sigma(F)}\right)\psi'_1\left(\frac{R(F)}{\sigma(F)}\right)\right\} \end{aligned}$$

assuming that the above inverse exists.

If only innovation outliers are possible then $R(F) = a_{p+1}$ and the expectation values $\mathbf{e}$ and $\mathbf{f}$ are equal to the zero-vector, assuming the innovations distribution G is symmetric and ψ_1 is odd. In this case we have

$$\text{(III.22)} \qquad IC_{T_{\phi,\sigma},F}(\mathbf{U}') = \begin{bmatrix} IC_{T_\phi,F}(\mathbf{U}') \\ IC_{T_\sigma,F}(\mathbf{U}') \end{bmatrix} = \begin{bmatrix} D^{-1}\,V(\mathbf{Z}'_{p+1})\,\mathbf{Z}'_{p+1}\,\sigma(F)\,\psi_1\left(\dfrac{R'(F)}{\sigma(F)}\right) \\ d^{-1}\,V(\mathbf{Z}'_{p+1})\,\sigma(F)\left\{\psi_1^2\left(\dfrac{R'(F)}{\sigma(F)}\right) - b_1\right\} \end{bmatrix}$$

where $IC_{T_\phi,F}(\mathbf{U}')$ and $IC_{T_\sigma,F}(\mathbf{U}')$ are the influence curves for the separate estimators $\hat{\phi}$ (with σ known) and $\hat{\sigma}$ (with ϕ known).

Although the above separation of the influence curve does not hold for the additive outliers model, (III.22) will hold approximately if $\mathbf{e}$ and $\mathbf{f}$ are approximately zero. One may expect this to be the case if $\phi(F) = \phi_0 \neq \phi$ provided $\gamma = P(V_i \neq 0)$ (compare Section I.4) is not too large and the bias $\phi_0 - \phi$ is small. A small bias can be obtained if the ψ_1-function and ψ_2-function are well chosen.

Qualitative Robustness

Let F_0 be a fixed distribution of $\mathbf{X} = (X_{p+1}, \ldots, X_1)^T$ for an innovation outlier model where $Y_i = X_i$. Define an additive outliers model on the innovation outlier model where the vector $\mathbf{Y} = (Y_{p+1}, \ldots, Y_1)^T$ has distribution F. Suppose that $\psi_1(\cdot)$ and $\psi_2(\cdot)$ are chosen so that the influence curve (III.20) exists for all F and is a bounded and continuous function of $\mathbf{U}'$. Then it can be shown that

(i) $T_{\phi,\sigma}(F) = (\phi^T(F), \sigma(F))$ defines a functional which is continuous at F_0;
(ii) under additional regularity conditions, $(\hat{\phi}, \hat{\sigma})$ is asymptotically normal with mean $(\phi^T(F), \sigma(F))$;
(iii) the asymptotic covariance matrix $V_{\hat{\phi},\hat{\sigma}}(F)$ of $(\hat{\phi}^T, \hat{\sigma})$ defines a functional which is continuous along the special "directions" for which $\mathbf{X} \overset{d}{\sim} F_0$ with F_0 fixed and $\mathbf{Y} \overset{d}{\sim} F$.

It further follows that $(\hat{\phi}^T, \hat{\sigma})$ is qualitatively robust at a Gaussian F_0 if deviations in the form of additive outliers are allowed.

The Breakdown Point

It may be shown that the breakdown point of GM-estimators for AR(p) models is positive but somewhat unfortunately it is bounded by $1/(p+1)$. The heuristic reason for this is easy to see. A single gross outlier at a fixed time i_0 appears in p consecutive prediction vectors $\mathbf{z}_i = (y_{i-1}, \ldots, y_{i-p})^T$. When a fraction of $1/(p+1)$ gross errors are uniformly spaced, all the predictors $\mathbf{z}_i^T\phi'$ appearing in the residuals $y_i - \mathbf{z}_i^T\phi'$ will be worthless. Of course the situation will be better in the case of patchy outliers, because then the total fraction of outliers can be higher without ruining the GM-estimates.

III.2 GENERALIZED MAXIMUM LIKELIHOOD TYPE ESTIMATION OR ARMA MODELS

M-estimators of ARMA parameters (compare Section II.3) have the advantage of being efficiency robust toward innovation outliers. But the behaviour of M-estimators in the presence of additive outliers should be better. We now will generalize M-estimators ARMA parameters to diminish the variability and bias if additive outliers are present, where the additive outliers are not necessarily i.i.d.

III.2.1. Definition

In contrast to M-estimators, GM-estimators of ARMA parameters can hardly or not at all be defined by a minimum problem, because the weights to be used depend on the parameters to be estimated (compare Martin and Yohai, 1984a). But the following approach could be used to define GM-estimators of ARMA parameters (Stockinger, 1985a) where the definition of least squares estimators is generalized.

Differentiating the minimum problem (I.5) which defines a least squares estimator of ARMA parameters $\boldsymbol{\alpha} := (\boldsymbol{\phi}^T, \boldsymbol{\theta}^T, \mu)^T$, with respect to $\boldsymbol{\alpha}'$ and setting the resulting expression equal to zero yield the estimation equation for the least squares estimator $\boldsymbol{\alpha}^\wedge$

$$\sum_{i=p+1}^{n} r_i(\boldsymbol{\alpha}^\wedge)\, \mathbf{d}_i^+(\boldsymbol{\alpha}^\wedge) = \mathbf{0} \tag{III.23}$$

where $r_i(\boldsymbol{\alpha}^\wedge)$ denotes the residual at time i (I.16) and $\mathbf{d}_i^+(\boldsymbol{\alpha}^\wedge)$ denotes the vector of the first derivatives of $r_i(\boldsymbol{\alpha}^\wedge)$ which is specified by (II.27) to (II.30).

If a given time series $y_1, \ldots, y_n$ contains outliers, it is advisable to use

$$\sum_{i=p+1}^{n} w\left[r_i(\boldsymbol{\alpha}^\wedge), \mathbf{d}_i(\boldsymbol{\alpha}^\wedge), \boldsymbol{\phi}^\wedge, \boldsymbol{\theta}^\wedge, \hat{\sigma}\right] r_i(\boldsymbol{\alpha}^\wedge)\, \mathbf{d}_i^+(\boldsymbol{\alpha}^\wedge) = \mathbf{0} \tag{III.24}$$

instead of (III.23) to estimate $\boldsymbol{\alpha}$, where $w\left[r, \mathbf{d}, \boldsymbol{\phi}^\wedge, \boldsymbol{\theta}^\wedge, \hat{\sigma}\right] r\mathbf{d}^+$ is bounded and $\hat{\sigma}$ is an estimate of the innovations scale. $w\left[r, \mathbf{d}, \boldsymbol{\phi}^\wedge, \hat{\theta}, \hat{\sigma}\right]$ denotes the above mentioned weight which depends on the parameters $\boldsymbol{\alpha}$ and σ to be estimated and which transforms the least squares problem (III.23) into a weighted least squares problem (III.24). If only innovation outliers are possible, it is sufficient to choose the weight function w, so that $w\left[r, \mathbf{d}, \boldsymbol{\phi}^\wedge, \boldsymbol{\theta}^\wedge, \hat{\sigma}\right] r$ is bounded. If $\psi(r/\hat{\sigma})/(r/\hat{\sigma})$ is selected as weight function where ψ should be a bounded ψ-function, e.g. ψ_H (II.3), equation (III.24) reduces to the estimating equation (II.26) of an M-estimator.

The symmetric matrix $C(\boldsymbol{\phi}, \boldsymbol{\theta})$ which is equal to matrix $C^*(\boldsymbol{\phi}, \boldsymbol{\theta})$ (I.18), except that

$$c_{j,p+k} = -c^*_{j,p+k}, \quad \text{if} \quad j \leqq p, \quad k \leqq q \tag{III.25}$$

is the *covariance matrix* of $\mathbf{d}_i(\boldsymbol{\alpha})/\sigma$ (Martin and Yohai, 1984a) if $\boldsymbol{\alpha}$ is the true parameter vector and σ is the true scale of the innovations.

This can be proved for $j \leqq k \leqq p$ as follows (the proofs for other indices j and k are analogous), where a residual $r_i(\boldsymbol{\alpha})$ computed for the true parameter vector has to be set equal to the corresponding realization a_i of an innovation.

$$COV(s_{i-j}(\boldsymbol{\alpha}), s_{i-k}(\boldsymbol{\alpha})) = COV(\xi_0 a_{i-j} + \xi_1 a_{i-j-1} + \ldots + \xi_{k-j} a_{i-k} + \ldots,$$
$$\xi_0 a_{i-k} + \xi_1 a_{i-k-1} + \ldots) = \sigma^2 \sum_{l=0}^{\infty} \xi_l \xi_{l+k-j}.$$

Therefore $COV(s_{i-j}(\boldsymbol{\alpha})/\sigma, s_{i-k}(\boldsymbol{\alpha})/\sigma) = c_{j,k}$.

The "largeness" of $\mathbf{d}_i(\boldsymbol{\alpha}^\wedge)$ can be assessed by

$$b_i(\boldsymbol{\alpha}^\wedge, \hat{\sigma}) = \hat{\sigma}^{-1}[\mathbf{d}_i^T(\boldsymbol{\alpha}^\wedge)\, C^{-1}(\boldsymbol{\phi}^\wedge, \boldsymbol{\theta}^\wedge)\, \mathbf{d}_i(\boldsymbol{\alpha}^\wedge)]^{1/2}, \tag{III.26}$$

because $\hat{\sigma}^2 C(\boldsymbol{\phi}^\wedge, \boldsymbol{\theta}^\wedge)$ estimates the covariance matrix of $\mathbf{d}_i(\boldsymbol{\alpha}^\wedge)$.

Now estimators $\boldsymbol{\alpha}^\wedge$ which satisfy (III.24) where $w[r, \mathbf{d}, \boldsymbol{\phi}^\wedge, \boldsymbol{\theta}^\wedge, \sigma]\, r\mathbf{d}^+$ is bounded can be defined. These estimators are referred to as *generalized M-estimators* (GM-estimators). Note that there is no need to bound the influence of the last component $-\partial r_i(\boldsymbol{\alpha}^\wedge)/\partial\mu$ (II.28) in $\mathbf{d}_i^+(\boldsymbol{\alpha}^\wedge)$ (II.27) because this component does not depend on the given data $y_1, \ldots, y_n$.

Define terms $v_i(\boldsymbol{\alpha}^\wedge, \hat{\sigma})$ which represent the largeness in the factor space for linear regression problems, by

$$v_i(\boldsymbol{\alpha}^\wedge, \hat{\sigma}) = \begin{cases} \psi_2[b_i(\boldsymbol{\alpha}^\wedge, \hat{\sigma})]/b_i(\boldsymbol{\alpha}^\wedge, \hat{\sigma}), & \text{if } b_i(\boldsymbol{\alpha}^\wedge, \hat{\sigma}) \neq 0 \\ \lim\limits_{t\to 0} \psi_2(t)/t, & \text{if } b_i(\boldsymbol{\alpha}^\wedge, \hat{\sigma}) = 0. \end{cases} \tag{III.27}$$

There are various types of GM-estimators according to the selection of the weights $w[r, \mathbf{d}, \boldsymbol{\phi}^\wedge, \boldsymbol{\theta}^\wedge, \hat{\sigma}]$. By the following choice of $w[r, \mathbf{d}, \boldsymbol{\phi}^\wedge, \boldsymbol{\theta}^\wedge, \hat{\sigma}]$ (III.28) a *Mallows type GM-estimator* (Mallows, 1976) is given if $u = 1$ and a *Schweppe type GM-estimator* (Schweppe, 1975) is given if $u = v$, where v is an abbreviation of $v_i(\boldsymbol{\alpha}^\wedge, \hat{\sigma})$ from (III.27).

$$w[r, \mathbf{d}, \boldsymbol{\phi}^\wedge, \boldsymbol{\theta}^\wedge, \hat{\sigma}] = \begin{cases} v\psi_1(r/u\hat{\sigma})/(r/\hat{\sigma}) & \text{if } r \neq 0,\ u \neq 0 \\ v/u, & \text{if } r = 0,\ u \neq 0 \\ 1, & \text{if } r = u = v = 0 \\ 1, & \text{if } r \neq 0,\ u = v = 0,\ \psi_1(t) = t \\ 0, & \text{if } r \neq 0,\ u = v = 0,\ \psi_1 \text{ is bounded}. \end{cases} \tag{III.28}$$

To estimate the first-order AR parameter ϕ_1, equation (III.24) becomes

$$\sum_{i=2}^{n} v_i(\hat{\phi}_1, \hat{\sigma})\, \psi_1\left(\frac{y_i - y_{i-1}\hat{\phi}_1}{u_i(\hat{\phi}_1, \hat{\sigma})\, \hat{\sigma}}\right) y_{i-1} = 0 \tag{III.29}$$

where the "largeness" to compute $v_i(\hat{\phi}_1, \hat{\sigma})$ is given by

$$b_i(\hat{\phi}_1, \hat{\sigma}) = \hat{\sigma}^{-1}(1 - \hat{\phi}_1^2)^{1/2}\, y_{i-1}. \tag{III.30}$$

Equations (III.8) and (III.9) on the one hand and equations (III.29) and (III.30) on the other give alternative possibilities to estimate ϕ_1 using different estimators for σ_x.

To estimate the first-order MA parameter θ_1 equation (III.24) becomes

$$\sum_{i=1}^{n} v_i(\hat{\theta}_1, \hat{\sigma})\, \psi_1\left(\frac{r_i(\hat{\theta}_1)}{u_i(\hat{\theta}_1, \hat{\sigma})\, \hat{\sigma}}\right) t_{i-1}(\hat{\theta}_1) = 0$$

where the "largeness" to compute $v_i(\hat{\theta}_1, \hat{\sigma})$ is given by

$$b_i(\hat{\theta}_1, \hat{\sigma}) = \hat{\sigma}^{-1}(1 - \hat{\theta}_1^2)^{1/2}\, t_{i-1}(\hat{\theta}_1)$$

($t_i(\hat{\theta}_1)$ is defined by (II.30)).

A *Hampel-Krasker-Welsch type GM-estimator* (Krasker and Welsch, 1982) is defined by

$$(\text{III.31}) \qquad w[r, \mathbf{d}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}] = \begin{cases} \psi_1(rb/\hat{\sigma})/(rb/\hat{\sigma}), & \text{if } r \neq 0 \text{ and } b \neq 0 \\ 1, & \text{if } r = 0 \text{ or } b = 0 \end{cases}$$

where b denotes the "largeness" of $\mathbf{d}$ which could be computed by (III.26). The principle of a Hampel-Krasker-Welsch type estimator is similar to that of a Schweppe type estimator and therefore both estimators should have similar properties. If innovation outliers are present a Schweppe type estimator or a Hampel-Krasker-Welsch type estimator should be superior to a Mallows type estimator because a Mallows type estimator does not simultaneously take into account the largeness of the residuals and the largeness of the first derivatives of the residuals.

The influence curves of Mallows type GM-estimators and Hampel-Krasker-Welsch type GM-estimators of first-order AR and MA parameters published by Martin and Yohai (1984b) encourage the implementation and application of an algorithm for the computation of these estimators.

III.2.2 Computational Methods

Unfortunately it might be very difficult to determine a function whose first derivative with respect to $\boldsymbol{\alpha}$ is the left hand side of (III.24). But (III.24) can be solved without knowing the minimum problem by applying a nonlinear iterative least squares algorithm, where the weights are determined using the approximations for the parameters calculated in the preceding iteration.

Algorithms to compute the residuals $r_i(\boldsymbol{\alpha}')$ (I.16) and the first derivatives of the residuals with respect to AR and MA parameters $s_i(\boldsymbol{\alpha}')$ (II.29) and $t_i(\boldsymbol{\alpha}')$ (II.30) were already given in Section I.6 and Section II.3.1, respectively. Next an algorithm to compute inverse AR or MA operators (I.19) will be given.

Computation of the Coefficients of an Inverse AR Operator (Anderson, 1971)

Let M denote the highest index of the coefficients to be computed.

(1) Set $m = 1$, $\xi_0 = 1$ and $\beta_{0,j} = \phi_j$, $j = 1, \ldots, p$.

(2) Set $\xi_m = \beta_{m-1,1}$.

(3) Compute $\beta_{m,j} = \beta_{m-1,j+1} + \beta_{m-1,1}\phi_j$, $j = 1, \ldots, p-1$, and $\beta_{m,p} = \beta_{m-1,1}\phi_p$.

(4) Augment $m = m + 1$.

(5) If $m > M$ stop, else go to (2).

GM-estimates of ARMA parameters could be computed by the algorithm for the M-estimation of ARMA parameters given in Section II.3.2, if it is modified as follows:

- A constant c as defined in (II.25) is not necessary and Step 3 must be omitted, because there is no side condition like (II.24) to improve the scale.
- Before the weights can be calculated (Step 4) the first derivatives of the residuals must be computed (i.e. Step 5 must be performed) and, in addition, the coefficients $\xi_l^{(m)}$, $l = 0, \ldots, L_\xi$ of the inverse AR operator

$$\phi^{(m)-1}(B) = \sum_{l=0}^{\infty} \xi^{(m)} B^l$$

and the coefficients $\zeta_l^{(m)}$, $l = 0, \ldots, L_\zeta$ of the inverse MA operator

$$\theta^{(m)-1}(B) = \sum_{l=0}^{\infty} \zeta_l^{(m)} B^l$$

must be computed, where L_ξ and L_ζ are chosen so that the inverse operators are sufficiently well approximated. Computer programs (Stockinger, 1985b) for the GM-estimation of ARMA models optionally allow the output of inverse AR and MA operators in order to determine L_ξ and L_ζ, respectively. It is reasonable to choose $L_\xi = 50$ if the AR order p is not too large, say $p \leqq 3$.
Furthermore the matrix $C(\phi^{(m)}, \theta^{(m)})$ must be computed according to (I.18) and (III.25) and inverted.
- In Step 4 the weights

$$w_i^{(m)} = w\left[r_i(\alpha^{(m)}), \mathbf{d}_i(\alpha^{(m)}), \phi^{(m)}, \theta^{(m)}, \sigma^{(m)}\right], \quad i = p+1, \ldots, n,$$

have to be computed for a Mallows type GM-estimator or a Schweppe type GM-estimator (III.28) or a Hampel-Krasker-Welsch type GM-estimator (III.31), and a diagonal matrix $W^{(m)} = \operatorname{diag}(w_{p+1}^{(m)}, \ldots, w_n^{(m)})$ has to be defined.
- Step 6 and Step 7 try to diminish the function

$$g(\alpha^{(m)}) = \tfrac{1}{2} \sum_{i=p+1}^{n} w_i^{(m)} r_i^2(\alpha^{(m)})$$

(where the weights $w_i^{(m)}$ are regarded to be constant for a fixed step of iteration procedure) instead of $g(\alpha^{(m)}, \sigma^{(m+1)})$ (II.23). The computational methods, in particular the computation of the Gauss-Newton direction and the vector of the steepest descent, do not change. These steps are explained in more detail by Stockinger (1985a).
- Before Step 8 is performed, the scale should be improved by

$$\sigma^{(m+1)} = \operatorname*{med}_{p+1 \leqq i \leqq n} \left| r_i(\alpha^{(m+1)}) - \operatorname*{med}_{p+1 \leqq j \leqq n} r_j(\alpha^{(m+1)}) \right| / .6745 . \tag{III.32}$$

III.2.3 Properties

A formal Taylor series expansion indicates that under suitable regularity conditions an estimator of α defined by (III.24) has, for time series without additive outliers,

the property (Martin and Yohai, 1984a)

$$\sqrt{n}\,(\boldsymbol{\alpha}^{\wedge} - \boldsymbol{\alpha}) \xrightarrow{\mathrm{d}} N(0, K(\boldsymbol{\phi}, \boldsymbol{\theta}, G))$$

where

(III.33) $$K(\boldsymbol{\phi}, \boldsymbol{\theta}, G) = \sigma^2 U^{-1} S U^{-1}$$

with

$$S = E\boldsymbol{\eta}[A_i, \mathbf{D}_i^+, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma]\,\boldsymbol{\eta}^{\mathrm{T}}[A_i, \mathbf{D}_i^+, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma]$$

$$U = E\boldsymbol{\eta}_1[A_i, \mathbf{D}_i^+, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma]\,\mathbf{D}_i^+$$

where $\mathbf{D}_i^+$ denotes a random vector whose possible realization is $\mathbf{d}_i^+(\boldsymbol{\alpha})$ (II.27),

$$\boldsymbol{\eta}[a, \mathbf{d}^+, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma] = w[a, \mathbf{d}, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma]\, a\mathbf{d}^+\sigma^{-2}$$

and

$$\boldsymbol{\eta}_1[a, \mathbf{d}^+, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma] = \partial\boldsymbol{\eta}[a, \mathbf{d}^+, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma]/\partial a\,.$$

For a careful proof for the estimation of autoregressive models see Bustos (1982).

For an AR(1) model without location $\mathbf{d}_i^+(\phi_1) = \mathbf{d}_i(\phi_1) = y_{i-1}$ and the "largeness" of $\mathbf{d}_i(\phi_1)$ is $b_i(\phi_1, \sigma) = \sigma^{-1}(1 - \phi_1^2)^{1/2}\, y_{i-1}$ (III.30). Let $\bar{a}$ denote a/σ. Note that for an outlier-free process both the innovations divided by σ and the b_i's have a standard normal distribution. Asymptotic variances (III.33) of GM-estimators of the first-order AR parameter ϕ_1 for outlier-free processes can be computed in the following way:

Mallows type estimator

$$S = \left[(1 - \phi_1^2)\,2\pi\right]^{-1} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi_1^2(\bar{a})\,\psi_2^2(b) \exp\left(-\bar{a}^2/2\right) \exp\left(-b^2/2\right) \mathrm{d}\bar{a}\,\mathrm{d}b$$

$$U = \left[(1 - \phi_1^2)\,2\pi\right]^{-1} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} b\,\psi_1'(\bar{a})\,\psi_2^2(b) \exp\left(-\bar{a}^2/2\right) \exp\left(-b^2/2\right) \mathrm{d}\bar{a}\,\mathrm{d}b$$

where $\psi_1'(\bar{a}) = d\psi_1(\bar{a})/\mathrm{d}\bar{a}$.

Schweppe type estimator

$$S = \left[(1 - \phi_1^2)\,2\pi\right]^{-1} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi_1^2\left(\frac{\bar{a}b}{\psi_2(b)}\right) \psi_2^2(b) \exp\left(-\bar{a}^2/2\right) \exp\left(-b^2/2\right) \mathrm{d}\bar{a}\,\mathrm{d}b$$

$$U = \left[(1 - \phi_1^2)\,2\pi\right]^{-1} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} b^2\psi_1'\left(\frac{\bar{a}b}{\psi_2(b)}\right) \exp\left(-\bar{a}^2/2\right) \exp\left(-b^2/2\right) \mathrm{d}\bar{a}\,\mathrm{d}b\,.$$

Hampel-Krasker-Welsch type estimator

$$S = \left[(1 - \phi_1^2)\,2\pi\right]^{-1} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi_1^2(\bar{a}b) \exp\left(-\bar{a}^2/2\right) \exp\left(-b^2/2\right) \mathrm{d}\bar{a}\,\mathrm{d}b$$

$$U = \left[(1 - \phi_1^2)\,2\pi\right]^{-1} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} b^2\,\psi_1'(\bar{a}b) \exp\left(-\bar{a}^2/2\right) \exp\left(-b^2/2\right) \mathrm{d}\bar{a}\,\mathrm{d}b\,.$$

Expression (III.33) can also be used to compute the asymptotic variance of an

M-estimator of ϕ_1. For an outlier-free process

$$S = [(1 - \phi_1^2)\, 2\pi]^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} b^2\, \psi_1^2(\bar{a}) \exp(-\bar{a}^2/2) \exp(-b^2/2)\, \mathrm{d}\bar{a}\, \mathrm{d}b$$

and

$$U = [(1 - \phi_1^2)\, 2\pi]^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} b^2\, \psi_1'(\bar{a}) \exp(-\bar{a}^2/2) \exp(-b^2/2)\, \mathrm{d}\bar{a}\, \mathrm{d}b$$

is obtained.

For the Monte Carlo results that will be presented in Chapter V the asymptotic variances of estimators for the first-order AR parameter were computed by numerical integration of the expressions stated.

Expression (III.33) could also be used to determine the asymptotic variance of a least squares estimator of ϕ_1. This variance is known to be $1 - \phi_1^2$ (Box and Jenkins, 1976).

If MA parameters have to be estimated, GM-estimators are not robust because an outlier at time i' spoils all ensuing residuals r_i, $i \geqq i'$. One possible remedy is to use truncated residuals similar to the idea of estimators based on truncated autocovariances of the residuals (Bustos and Yohai, 1983).

GM-estimators cannot be expected to be unbiased in the presence of additive outliers but the bias will be smaller than for M-estimators. The variance also can be kept smaller than for M-estimators. This will be demonstrated for small samples in Chapter V.

III.3 DETERMINING OUTLIER TYPE

Methods to determine whether an AR process is contaminated by innovation outliers or additive outliers, will be described in the following.

In Section II.1.3 (formula (II.15)) it was shown that innovation outliers lead to an increased precision of M-estimators of AR parameters (compare Stockinger (1985a) for a graphical explanation). It is intuitively clear that such outliers should not be downweighted for forecasting purposes, e.g. in a GM-estimating equation. In contrast, additive outliers need to be downweighted if future values of the unobservable X_i process are predicted.

Since the appropriate treatments, e.g. the selection of an estimator, of the two types of outliers are different, it could be costly to mistake additive outliers for innovation outliers and vice versa. Thus in time series analysis there is a need to distinguish between different types of outliers in order to effectively deal with them.

III.3.1 A Significance Test for Additive Outliers Versus Innovation Outliers

One approach for constructing a significance test to distinguish between innovation outliers and additive outliers is suggested by the fact that although GM-estimators behave moderately well on an overall basis at both outlier situations, M-estimators are clearly superior if only innovation outliers occur (Denby and Martin, 1979; Martin and Zeh, 1978).

M-estimators have unacceptably large biases if additive outliers are present whereas GM-estimators have attractively small biases and variabilities. Hence a significance test for testing the null hypothesis

(III.34) H_{IO}: model IO holds (where the innovations distribution could also be Gaussian)

versus the alternative

(III.35) H_{AO}: model AO holds

based on the difference $\hat{\phi}_M - \hat{\phi}_{GM}$ between an M-estimate and a GM-estimate, suggests itself. It can be shown that under reasonable assumptions the asymptotic distribution of $\delta_n = \sqrt{n}\,(\hat{\phi}_M - \hat{\phi}_{GM})$ is multivariate normal and has under H_{IO} mean zero and covariance matrix

$$V_{\delta,IO} = \left[B_1^{-1} B_2 B_1^{-1} - C^{-1}\right] V_{loc}(\psi, g) \tag{III.36}$$

where B_1 and B_2 are defined by (III.15) and the elements of C are given by $c_{ij} =$ $=$ covariance (Y_i, Y_j), $1 \leqq i, j \leqq p$, and all expectations in computing B_1, B_2 and C are taken under H_{IO} (Martin, 1979).

The asymptotic distribution of

$$T_n = \delta_n^{\mathrm{T}} V_{\delta,IO}^{-1} \delta_n$$

is chi-squared with p degrees of freedom for H_{IO}. A usable test statistic might be obtained by replacing $V_{\delta,IO}$ by a good estimate and using χ_p^2 critical values or perhaps critical values obtained via Monte Carlo. Further details and some encouraging Monte Carlo results are given by Martin and Zeh (1977) and Zeh (1979).

III.3.2 Diagnostic Scatter Plots

Distinctively different characters of the outlier configurations in scatter plots under H_{IO} (III.34) and H_{AO} (III.35) may be used as exploratory indicators of outlier type.

The scatter plot approach for assessing outlier type in an exploratory manner is based on the residuals

$$r_i = y_i - \sum_{k=1}^{p} \hat{\phi}_k y_{i-k}, \quad p+1 \leqq i \leqq n, \tag{III.37}$$

from a GM-estimator fit. If $\hat{\phi}$ is a good estimate then its value will be close to that

of ϕ for not too small sample size and the residual r_i will be almost the same as

$$u_i = y_i - \sum_{k=1}^{p} \phi_k y_{i-k} = a_i + v_i - \sum_{k=1}^{p} \phi_k v_{i-k}, \quad p+1 \leqq i \leqq n. \tag{III.38}$$

Thus for each i the bivariate distribution of (R_i, R_{i+1}) should be close to that of (U_i, U_{i+1}).

The scatter plot of the pairs (r_i, r_{i+1}) will resemble that of (a_i, a_{i+1}) under H_{IO} (III.34), because in this case $V_i = 0$ for all i. If the A_i's are Gaussian the residuals r_i will produce a circular scatter plot. Outliers resulting from a heavy-tailed innovation distribution will be mainly along the abscissa and the ordinate of a scatter plot.

If on the other hand $V_i \neq 0$ due to additive outliers then U_i and U_{i+1} will usually be dependent. In this case the outliers generally no longer lie mainly along the abscissa and the ordinate in the scatter plot.

III.3.3 Robustified Fox Tests

Fox (1972) considered the problem of detecting a single outlier at an unknown time i assuming that either model I (which is analogous to the additive outliers model) or model II (which is analogous to the innovation outlier model) is the true model. In particular, Fox assumes that the innovations are Gaussian with mean zero. In model II a_i is replaced by $a_i + \Delta$ with Δ unknown and model I produces only an additive outlier at time i.

In addition to studying likelihood ratio tests, Fox considered simplified criteria which, for unknown i, would have the form

$$\max_i \lambda_i \tag{III.39}$$

where

$$\lambda_i = \hat{\Delta}_i / \widehat{VAR}^{1/2}(\hat{\Delta}_i). \tag{III.40}$$

Under modell II Fox defines

$$\hat{\Delta}_i = y_i - \sum_{k=1}^{p} \hat{\phi}_k y_{i-k}, \quad \widehat{VAR}(\hat{\Delta}_i) = \hat{\sigma}^2. \tag{III.41}$$

The estimates $\hat{\phi}$ and $\hat{\sigma}$ are approximate maximum likelihood estimates computed under the assumptions that $\Delta_i \neq 0$ and the null hypothesis $\Delta_i = 0$ holds, respectively. For model I

$$\hat{\Delta}_i = y_i + \Big[\sum_{k=1}^{p} \hat{W}^{i,i+k}(y_{i-k} + y_{i+k})\Big] / \hat{W}^{i,i}, \tag{III.42}$$

$$\widehat{VAR}(\hat{\Delta}_i) = \hat{\sigma}^2 / \sum_{k=0}^{p} \hat{\phi}_k^2$$

where

$$\hat{W}^{i,i+k} = \sum_{j=k}^{p} (-\hat{\phi}_j)(-\hat{\phi}_{j-k})$$

with $-\hat{\phi}_0 = 1$ and $\hat{\phi}_j$ for $1 \leqq j \leqq p$ computed assuming $V_i = \Delta \neq 0$.

Before extending the above technique to the more realistic multiple outlier case, Martin and Zeh (1977) perform a robustification of (III.41) and (III.42) by using M-estimators and GM-estimators of ϕ, respectively. The reason is, that M-estimators behave well under model II and GM-estimators behave well under model I. A robust scale estimator is used for σ.

The statistic for testing H_{IO} (III.34) is (robustified FOX criterion)

$$RFOX = \log\left(\max_i \lambda_{i,I}^2 / \max_i \lambda_{i,II}^2\right), \quad p+1 \leqq i \leqq n-p \tag{III.43}$$

where $\lambda_{i,I}$ and $\lambda_{i,II}$ are the versions of (III.40) obtained from the robustified expressions (III.42) and (III.41), respectively.

III.4 MODEL SELECTION

III.4.1 Robust Estimation of Autoregression Order

For outlier-free time series the minimization with respect to p of either Parzen's (1974) $CAT(p)$ or the Gaussian autoregression version of Akaike's (1974) $AIC(p)$ function provides an estimate of the order p of an AR model.

For perfectly observed Gaussian or non-Gaussian autoregressions Akaike's function is

$$AIC(p) = -2\log f(\mathbf{y}; \hat{\phi}, \hat{\mu}, \hat{\sigma}, p) + 2(p+2) \tag{III.44}$$

where $\hat{\phi}$, $\hat{\mu}$ and $\hat{\sigma}$ denote maximum likelihood estimates, $\mathbf{y}^T = (y_1, y_2, \ldots, y_n)$, $f(\mathbf{y}; \hat{\phi}, \hat{\mu}, \hat{\sigma}, p)$ denotes the maximized likelihood for an AR(p) process and $p+2$ is the number of parameters estimated. In the Gaussian case and if the sample size n is reasonably large, $AIC(p)$ is approximatively equivalent to

$$AAIC(p) = \log \hat{\sigma}^2(p) + 2(p+2)/n \tag{III.45}$$

where $\hat{\sigma}^2(p)$ is an estimate of the variance of the innovations

$$\hat{\sigma}^2(p) = \frac{1}{n-2p-1} \sum_{i=p+1}^{n} (y_i - \mathbf{Z}_i^T \hat{\phi})^2 \tag{III.46}$$

with $\hat{\phi}$ denoting a least squares estimate.

It is well known that, given an i.i.d. sample, the variance estimator based on the sum-of-squared residuals it notoriously non-robust toward heavy-tailed distributions (see, e.g. Tukey, 1960). The same is true of the estimator $\hat{\sigma}^2(p)$ in innovation outlier situations (even if an M-estimator $\hat{\phi}$ is used instead of a least squares estimator) and in additive outlier situations (even if a least squares estimator $\hat{\phi}$ is replaced by a GM-estimator). Thus stoping rules based on such estimators would not be very reliable for either innovation outlier or additive outlier situations, and therefore a robust alternative to $AAIC(p)$ is needed.

Using robustly centered data, a Mallows type GM-estimating equation (III.5) could be obtained by differentiating the loss function

$$L(\mathbf{y}; \boldsymbol{\phi}^{\wedge\prime}, \sigma', p) = \sum_{i=p+1}^{n} v_i \varrho_1\left(\frac{y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\phi}'}{\sigma'}\right) \tag{III.47}$$

with respect to $\boldsymbol{\phi}'$ and setting it equal to zero with $\psi_1 = \mathrm{d}\varrho_1(t)/\mathrm{d}t$. The fact that equation (III.4) cannot be obtained by differentiation of $L(\mathbf{y}; \boldsymbol{\phi}', \sigma', p)$ with respect to σ', will be ignored.

If $\varrho_1(t) = -2 \log g(t)$ where g is the innovations density and $v_i = 1$, then the following approximation is possible (compare Stockinger and Dutter, 1983).

$$-2 \log f(\mathbf{y}; \boldsymbol{\phi}^{\wedge}, \hat{\mu}, \hat{\sigma}, p) \approx -2 \log \prod_{i=p+1}^{n} \frac{1}{\hat{\sigma}} g\left(\frac{y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\phi}^{\wedge}}{\hat{\sigma}}\right) = \tag{III.48}$$

$$= -2 \sum_{i=p+1}^{n} \log \frac{1}{\hat{\sigma}} g\left(\frac{y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\phi}^{\wedge}}{\hat{\sigma}}\right).$$

The right-hand side expression in (III.48) can be transformed to the representation

$$\sum_{i=p+1}^{n} \log\left(\frac{1}{\hat{\sigma}}\right)^{-2} + \sum_{i=p+1}^{n} \log g^{-2}\left(\frac{y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\phi}^{\wedge}}{\hat{\sigma}}\right) = \tag{III.49}$$

$$= 2(n - p) \log \hat{\sigma} + L(\mathbf{y}; \boldsymbol{\phi}^{\wedge}, \hat{\sigma}, p).$$

If only innovation outliers are possible the above equations suggest to construct a robust M-order-selection criterion by approximating $AIC(p)$ (III.44) using $\varrho(\cdot)$ instead of $-2 \log(\cdot)$ (i.e. maximum likelihood estimates are replaced by M-estimates) and by using (III.48) and (III.49), what results into

$$M(p) = \frac{1}{n - p} \sum_{i=p+1}^{n} \varrho_1\left[\frac{y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\phi}^{\wedge}}{\hat{\sigma}}\right] + 2 \log \hat{\sigma} + \frac{2(p + 2)}{n - p} \tag{III.50}$$

where $\boldsymbol{\phi}^{\wedge}$ and $\hat{\sigma}$ denote M-estimates. M-estimators, however, are not robust toward additive outlier situations and, therefore, instead of minimizing $M(p)$ with respect to p in this case, p should be estimated by minimizing a function which uses GM-estimates $\boldsymbol{\phi}^{\wedge}$ and $\hat{\sigma}$

$$GM(p) = \frac{1}{n - p} \sum_{i=p+1}^{n} v_i \varrho_1\left[\frac{y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\phi}^{\wedge}}{\hat{\sigma}}\right] \tag{III.51}$$

$$+ 2 \log \hat{\sigma} + \frac{2(p + 2)}{n - p}.$$

III.4.2 Identification of ARIMA Models

For outlier-free time series, Box and Jenkins (1976) suggested a procedure based on the sample autocorrelation function and partial autocorrelation function to identify an appropriate subclass of ARIMA models.

However, the use of the standard autocorrelation function estimate and partial autocorrelation estimate can be very misleading in the case of contaminated data, because these estimators lack robustness (compare e.g. Polasek, 1982). One possibility of putting an end to these problems could be to adapt correlation and covariance methods for i.i.d. multivariate samples (see Devlin et al., 1975; Maronna, 1976; Huber, 1977; Marazzi, 1980; Rieder, 1980) to the time series setting. Polasek and Mertl (1983) treat robust estimators of the autocorrelation function.

Martin, Samarov and Vandaele (1983) suggested an iterative procedure for the identification of an ARIMA model. The usual Box-Jenkins approach based on the initial unfiltered data is used to specify an initial model. Next, the initial model is used to clean the data by robust filtering. A new model identification pass is based on the cleaned data. If for the raw data and the cleaned data the same model is identified and if the diagnostic checks on the estimation results (e.g. checks on over-and underspecification, residual analysis) do not reveal a model misspecification, we have finished. Otherwise the robust filtering has to be carried out on the cleaned data, and the same diagnostic checks have to be applied.

IV. ROBUST FILTERING AND ROBUST SMOOTHING

In order to deal with robust filtering and smoothing a vector state-variable representation of ARMA processes will be described. Here, a filtered value is defined to depend only on previous observations, while a smoothed value is defined to depend on all given observations.

A recursive algorithm for the computation of approximate conditional-mean (ACM) filters which are able to remove outliers from contaminated data, will be dealt with.

Maximizing a likelihood function which is approximated (also by an ACM filter), leads to approximate maximum likelihood (AML) estimators. Proceeding further by replacing the negative of the log-likelihood by a loss function which uses a robustifying rho-function yields approximation maximum likelihood type (AM) estimators. A relatively simple iterative scheme can be used to compute AM-estimators. Conditional-mean M-estimators can be regarded as AM-estimators especially for AR models.

Other methods for robust filtering and smoothing are provided, for example, by the robustified Kalman filter, L-smoothers, moving M-estimate smoothers and robustified splines.

IV. 1 APPROXIMATE CONDITIONAL-MEAN (ACM) FILTERING AND SMOOTHING

IV.1.1 State-variable Representation of Time Series Models

An ARMA (p, q) process $x_1, \ldots, x_n$ (compare Section I.3) which has mean of value zero and which is free of additive outliers, could be represented in the vector state-variable form

$$\mathbf{x}_i = \Phi \mathbf{x}_{i-1} + \mathbf{a}_i \tag{IV.1}$$

where the first coordinate $(\mathbf{x}_i)_1$ of $\mathbf{x}_i$ is the value of the ARMA process at time i. But the second coordinate $(\mathbf{x}_i)_2$ is not necessarily equal to x_{i-1}! Thus an ARMA (p, q) process contaminated by additive outliers can be represented by (IV.1) together with the equation

$$y_i = \mu + H\mathbf{x}_i + v_i \tag{IV.2}$$

where $H = (1, 0, \ldots, 0)$.

We consider here only one particular state-variable representation for ARMA

(p, q) processes. Assume $p > q$ for the moment and let

$$(\mathbf{x}_i)_1 = \phi_1(\mathbf{x}_{i-1})_1 + (\mathbf{x}_{i-1})_2 - a_i \tag{IV.3}$$

where

$$(\mathbf{x}_{i-1})_2 = \phi_2(\mathbf{x}_{i-2})_1 + \ldots + \phi_p(\mathbf{x}_{i-p})_1 - \theta_1 a_{i-1} - \ldots - \theta_q a_{i-q}\,. \tag{IV.4}$$

Then continue in this manner:

$$\begin{aligned}
(\mathbf{x}_i)_2 &= \phi_2(\mathbf{x}_{i-1})_1 + (\mathbf{x}_{i-1})_3 - \theta_1 a_i \\
(\mathbf{x}_i)_3 &= \phi_3(\mathbf{x}_{i-1})_1 + (\mathbf{x}_{i-1})_4 - \theta_2 a_i \\
&\vdots \qquad\quad \vdots \qquad\qquad \vdots \\
(\mathbf{x}_i)_{p-1} &= \phi_{p-1}(\mathbf{x}_{i-1})_1 + (\mathbf{x}_{i-1})_p - \theta_{p-2} a_i \\
(\mathbf{x}_i)_p &= \phi_p(\mathbf{x}_{i-1})_1 - \theta_{p-1} a_i
\end{aligned} \tag{IV.5}$$

with the stipulation that $\theta_i = 0$ for $i > q$. Here the coordinates $(\mathbf{x}_i)_2, \ldots, (\mathbf{x}_i)_p$ are chosen so that it is possible to construct ensuing ARMA process values $(\mathbf{x}_{i+1})_1$, $(\mathbf{x}_{i+2})_1, \ldots$.

The state transition matrix for representing ARMA models is

$$\Phi = \begin{bmatrix}
\phi_1 & 1 & 0 & 0 & \ldots & 0 & 0 \\
\phi_2 & 0 & 1 & 0 & \ldots & 0 & 0 \\
\phi_3 & 0 & 0 & 1 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
\phi_{p-2} & 0 & 0 & 0 & \ldots & 1 & 0 \\
\phi_{p-1} & 0 & 0 & 0 & \ldots & 0 & 1 \\
\phi_p & 0 & 0 & 0 & \ldots & 0 & 0
\end{bmatrix} \tag{IV.6}$$

and $\mathbf{a}_i^{\mathrm{T}} = -a_i(1, \theta_1, \theta_2, \ldots, \theta_{p-1})$. Correspondingly the covariance matrix Q of the $\mathbf{a}_i$'s has elements

$$Q_{ij} = \begin{cases} \sigma^2 \theta_{i-1}\theta_{j-1}\,, & \text{if } \max(i, j) \leqq q + 1 \\ 0 & \text{otherwise} \end{cases} \tag{IV.7}$$

where $\theta_0 = 1$.

In the case of $q \geqq p$ the above procedure leads to a state equation of dimension $q + 1$ and the first column of Φ contains $\phi_1, \phi_2, \ldots, \phi_{q+1}$ where $\phi_k = 0$ for $k > p$.

Also an ARIMA (p, d, q) model with a mean of value zero can be represented in the vector state-variable form. Instead of $\phi(B)$ as in the ARMA (p, q) model, now $\varphi(B) = \phi(B)(1 - B)^d$ is operating on x_i. The order of $\varphi(B)$ is $p + d$.

The coefficients of the polynomial $\varphi(B)$ in an ARIMA $(1, 1, q)$ model, for example, can be shown to be equal to

$$\begin{aligned}
\varphi_1 &= 1 + \phi_1 \\
\varphi_2 &= -\phi_1\,.
\end{aligned} \tag{IV.8}$$

Therefore the state-variable representation of an ARMA model can be carried over after replacing p by $p + d$, and thus, the dimension of the state transition matrix is $\max(p + d, q + 1)$. Note, however, that an autoregression with parameters $\varphi_1, \varphi_2, \ldots$ does not yield stationary observations x_i.

IV.1.2 ACM Filters

We will use the following terminology:

(IV.9) $\mathbf{y}^i = (y_1, \ldots, y_i)$ — the first i observations

$f(\mathbf{x}_i \mid \mathbf{y}^{i-1})$, $i > 1$ } "state-prediction" density = conditional density of

(IV.10) $f(\mathbf{x}_i \mid \mathbf{y}^0) = f(\mathbf{x}_i)$, $i = 1$ } $\mathbf{x}_i$ given $\mathbf{y}^{i-1}$

(IV.11) $f_y(y_i \mid \mathbf{y}^{i-1})$, $i > 1$ } "observation-prediction" density
$f_y(y_i \mid \mathbf{y}^0) = f_y(y_i)$, $i = 1$ }

(IV.12) $f_x(x_i \mid \mathbf{y}^{i-1})$ — prediction density for the first coordinate $x_i = (\mathbf{x}_i)_1$ of $\mathbf{x}_i$

(IV.13) $\mathbf{X}\hat{}_i = E\{\mathbf{X}_i \mid \mathbf{Y}^i\}$ — conditional mean estimate of $\mathbf{X}_i$ given $\mathbf{Y}^i$

(IV.14) $\mathbf{X}\hat{}_i^{i-1} = E\{\mathbf{X}_i \mid \mathbf{Y}^{i-1}\}$ — conditional mean estimate of $\mathbf{X}_i$ given $\mathbf{Y}^{i-1}$

In the engineering literature $\mathbf{X}\hat{}_i$ is called a "filter" estimate; $\mathbf{X}\hat{}_i^{i-1}$ is called the one-step-ahead predictor.

The filter and the one-step-ahead predictors of an ARIMA process itself $x_i = (\mathbf{x}_i)_1$ are $\hat{X}_i = E\{X_i \mid \mathbf{Y}^i\} = (\mathbf{X}\hat{}_i)_1$ and $\hat{X}_i^{i-1} = E\{X_i \mid \mathbf{Y}^{i-1}\} = (\mathbf{X}\hat{}_i^{i-1})_1$. Under the assumption that the X_i's and V_i's are mutually independent time series with $\{V_i\}$ an i.i.d. sequences, we have

$$\hat{Y}_i^{i-1} = E\{Y_i \mid \mathbf{Y}^{i-1}\} = E\{X_i \mid \mathbf{Y}^{i-1}\} = \hat{X}_i^{i-1}.$$

Thus the one-step-ahead predictors of x_i and y_i are identical, and we shall use $\hat{Y}_i^{i-1}$ and $\hat{X}_i^{i-1}$ interchangeably.

Computation of the exact conditional-mean $\mathbf{X}\hat{}_i$ is difficult for non-Gaussian distributions F_V of the V_i's. Masreliez (1975), however, made the simplifying assumption that the state-prediction density (IV.10) may be well approximated by a Gaussian density

(IV.15) $$f(\mathbf{x}_i \mid \mathbf{y}^{i-1}) \approx N(\mathbf{x}_i; \mathbf{X}\hat{}_i^{i-1}, M_i)$$

to establish a recursive computational algorithm for *approximate conditional-mean* (ACM) filters. The covariance matrix M_i in (IV. 15) is the conditional error covariance matrix for the prediction of $\mathbf{X}_i$, i.e.

(IV.16) $$M_i = E\{(\mathbf{X}_i - \mathbf{X}\hat{}_i^{i-1})(\mathbf{X}_i - \mathbf{X}\hat{}_i^{i-1})^{\mathrm{T}} \mid \mathbf{Y}^{i-1}\}.$$

For the definition of the ACM filter also the conditional filtering error covariance

(IV.17) $$P_i = E\{(\mathbf{X}_i - \mathbf{X}\hat{}_i)(\mathbf{X}_i - \mathbf{X}\hat{}_i)^{\mathrm{T}} \mid \mathbf{Y}^i\}$$

is needed.

In the pure Gaussian situation $\mathbf{X}\hat{}_i$, M_i and P_i, $1 \leqq i \leqq n$, are obtained by the *Kalman filter* recursion (see, for example, Jazwinski, 1970) and M_i and P_i *do not* depend upon the given data $y_1, \ldots, y_n$, what is a rather special feature of the Gaussian case.

For the following ACM filter theorem it is assumed that the observations y_i are generated by (IV.1) and (IV.2) with location parameter $\mu = 0$ and with Φ, F_V and the covariance matrix Q known.

Theorem (Masreliez). If (IV. 15) holds for $i \geqq 1$, then $\mathbf{X}\hat{}_i$ is generated by the recursions

$$\mathbf{X}\hat{}_i = \mathbf{X}\hat{}_i^{i-1} + \mathbf{m}_i \psi_i(y_i) \tag{IV.18}$$

$$M_{i+1} = \Phi P_i \Phi^{\mathrm{T}} + Q \tag{IV.19}$$

$$P_i = M_i - \psi_i'(y_i)\, \mathbf{m}_i \mathbf{m}_i^{\mathrm{T}} \tag{IV.20}$$

and

$$\mathbf{X}\hat{}_i^{i-1} = \Phi \mathbf{X}\hat{}_{i-1} \tag{IV.21}$$

where $\mathbf{m}_i$ is the first column of M_i,

$$\Psi_i(y_i) = -(\partial/\partial y_i) \log f_y(y_i \mid \mathbf{y}^{i-1}) \tag{IV.22}$$

is the scalar-valued score function for the observation-prediction density and

$$\Psi_i'(y_i) = (\partial/\partial y_i)\, \Psi_i(y_i)\,. \tag{IV.23}$$

Martin (1981b) specifies initial conditions for the above recursions. The approximate $\mathbf{X}\hat{}_0$ and M_1 are $\mathbf{X}\hat{}_0 = E\{\mathbf{X}_0\} = \mathbf{0}$ and $M_1 = E\{\mathbf{X}_1 \mathbf{X}_1^{\mathrm{T}}\} = C_x$, i.e. the unconditional mean and covariance of $\mathbf{X}_1$. In the case of stationarity, the latter satisfies the equation $C_x = \Phi C_x \Phi^{\mathrm{T}} + Q$.

From (IV.15) it follows that in particular

$$f_x(x_i \mid \mathbf{y}^{i-1}) \approx N(x_i;\, \hat{X}_i^{i-1},\, m_{1i}) \tag{IV.24}$$

where $\hat{X}_i^{i-1} = (\mathbf{X}\hat{}_i^{i-1})_1$ and m_{1i} is the 1-1 element of M_i.

The observation-prediction density $f_y(y_i \mid \mathbf{y}^{i-1})$ could be obtained by convoluting the prediction density $f_x(x_i \mid \mathbf{y}^{i-1})$ with the noise distribution F_V (Martin, 1981b). Unfortunately, in non-Gaussian situations it generally is difficult to proceed further, because the form of $f(\mathbf{x}_i \mid \mathbf{y}^{i-1})$ is typically quite intractable. The simplifying assumption (IV.24), however, helps.

Because $\hat{Y}_i^{i-1} = \hat{X}_i^{i-1}$ we have

$$f_y(y_i \mid \mathbf{y}^{i-1}) = [N(\hat{Y}_i^{i-1}, m_{1i}) * F_V](y_i) = gd_i(y_i - \hat{Y}_i^{i-1}) \tag{IV.25}$$

where the density gd_i is obtained by convolution

$$gd_i = N(0, m_{1i}) * F_V\,. \tag{IV.26}$$

We could go one step further and represent gd_i in the form

$$gd_i(r) = \frac{1}{s_i}\, gd\left(\frac{r}{s_i}\right) \tag{IV.27}$$

where

$$gd = N(0, c_1) * F_{V,c_2}$$

with

(IV.29) $$F_{V,c_2}(r) = F_V(r/c_2)$$

and s_i, c_1, c_2 are approximately specified. This is not possible in general, if F_V is non-Gaussian. However, if the V_i's are distributed according to a contaminated normal distribution

(IV.30) $$CN(v, \sigma_1, \sigma_2) = (1 - v)\, N(0, \sigma_1^2) + vN(0, \sigma_2^2)\,,$$

it is reasonable to set

(IV.31) $$s_i = (m_{1i} + \sigma_1^2)^{1/2}\,,$$

(IV.32) $$c_1 = m_{1i}/s_i^2\,, \quad c_2 = \sigma_1/s_i$$

and to use (IV.27) as an approximation. The approximations (IV.27), (IV.28), (IV.29), (IV.31), (IV.32) should behave reasonably well for any heavy tailed distribution F_V which is nearly Gaussian in the middle. Applying these approximations in Masreliez's theorem gives

(IV.33) $$\Psi_i(y_i) \approx \frac{1}{s_i}\,\psi\left[\frac{y_i - \hat{Y}_i^{i-1}}{s_i}\right]$$

and

(IV.34) $$\Psi_i'(y_i) \approx \frac{1}{s_i^2}\,\psi'\left[\frac{y_i - \hat{Y}_i^{i-1}}{s_i}\right]$$

where

(IV.35) $$\psi(r) = -(\partial/\partial r) \log gd(r)\,.$$

Usage of (IV.33) to (IV.35) transforms Masreliez's filter into the following filter:

(IV.36) $$\mathbf{X}\hat{}_i = \Phi \mathbf{X}\hat{}_{i-1} + (\mathbf{m}_i/s_i^2)\, s_i\, \psi(r_i/s_i)$$

with prediction residuals

(IV.37) $$r_i = y_i - \hat{Y}_i^{i-1} = y_i - (\Phi \mathbf{X}\hat{}_{i-1})_1$$

and the prediction residual scale s_i given by (IV.31). The recursion for P_i is

(IV.38) $$P_i = M_i - (\mathbf{m}_i \mathbf{m}_i^{\mathrm{T}}/s_i^2)\, \psi'(r_i/s_i)\,.$$

A filter which is given by (IV.36) to (IV.38), (IV.31) and (IV.19) is referred to as an *approximative conditional-mean* (ACM) *filter.*

IV.1.3 ACM Smoothers

The conditional-mean $\mathbf{X}\hat{}_i = E\{\mathbf{X}_i \mid \mathbf{Y}^i\}$ might well be replaced by the conditional-mean $\mathbf{X}\hat{}_i^n = E\{\mathbf{X}_i \mid \mathbf{Y}^n\}$, $1 \leqq i < n$. For $i = n$ we have $\mathbf{X}\hat{}_n^n = \mathbf{X}\hat{}_n$ which is a filtered value. $\mathbf{X}\hat{}_i^n$ depends upon all observed data and is called a smoother.

It turns out to be rather easy to construct ACM smoothers using the ACM filter described in the previous section.

Theorem (see, e. g., Martin, 1979c). Suppose that $f(\mathbf{x}_i \mid \mathbf{y}^{i-1}) = N(\mathbf{x}_i; \mathbf{X}\hat{}_i^{i-1}, M_i)$ where $\mathbf{X}\hat{}_i^{i-1} = \Phi\mathbf{X}\hat{}_{i-1}$ and $\mathbf{X}\hat{}_i = E\{\mathbf{X}_i \mid \mathbf{Y}^i\}$, $1 \leqq i \leqq n$, is the ACM filter of the previous section with the approximate initial conditions. Then assuming that M_{i+1}^{-1} exists, $\mathbf{X}\hat{}_i^n$ satisfies the backward recursion

$$\mathbf{X}\hat{}_i^n = \mathbf{X}\hat{}_i + P_i\Phi^{\mathrm{T}}M_{i+1}^{-1}(\mathbf{X}\hat{}_{i+1}^n - \mathbf{X}\hat{}_{i+1}^i), \quad 1 \leqq i \leqq n-1, \tag{IV.39}$$

with the initial condition $\mathbf{X}\hat{}_n^n = \mathbf{X}\hat{}_n$. The smoothing-error covariance matrix

$$P_i^n = E\{(\mathbf{X}\hat{}_i - \mathbf{X}\hat{}_i^n)(\mathbf{X}\hat{}_i - \mathbf{X}\hat{}_i^n)^{\mathrm{T}} \mid \mathbf{Y}^n\} \tag{IV.40}$$

satisfies the backward recursion

$$P_i^n = P_i + A_i(P_{i+1}^n - M_{i+1})A_i^{\mathrm{T}} \tag{IV.41}$$

with the initial conditions $P_n^n = P_n$ and

$$A_i = P_i\Phi^{\mathrm{T}}M_{i+1}^{-1}. \tag{IV.42}$$

IV.2 APPROXIMATE MAXIMUM LIKELIHOOD TYPE (AM) ESTIMATES

IV.2.1 Approximate Maximum Likelihood (AML) Estimates

In this section the terminology (IV.9) to (IV.14) of Section IV.1.2 will be used. Since it is assumed that the observations y_i can be represented by the equations (IV.2) and (IV.1), the conditional densities and expectations (IV.10) to (IV.14) depend on the ARMA parameters $\boldsymbol{\alpha}^{\mathrm{T}} = (\boldsymbol{\phi}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}}, \sigma)$ and on the distribution F_V of the V_i's. The notation will sometimes (but not always) make explicit the dependence on $\boldsymbol{\alpha}$.

The exact log-likelihood may be expressed in the form

$$\log f(\mathbf{y}; \boldsymbol{\alpha}) = \log f_y(y_1; \boldsymbol{\alpha}) + \sum_{i=2}^{n} \log f_y(y_i \mid \mathbf{y}^{i-1}; \boldsymbol{\alpha}) \tag{IV.43}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$.

As mentioned in Section IV.1.2 it is not easy to evaluate $f_y(y_i \mid \mathbf{y}^{i-1}, \boldsymbol{\alpha})$ exactly. However, using the approximations (IV.27) to (IV.29), (IV.31) and (IV.32), which are based on Masreliez's simplifying assumptions (IV. 15), to evaluate expression (IV.25), and noting that s_i and $\hat{Y}_i^{i-1}$ depend on $\boldsymbol{\alpha}$, gives the following approximation for the log-likelihood:

$$\log f(\mathbf{y} \mid \boldsymbol{\alpha}) \approx -\sum_{i=1}^{n} \log s_i(\boldsymbol{\alpha}) + \sum_{i=1}^{n} \log gd\left[\frac{y_i - \hat{Y}_i^{i-1}(\boldsymbol{\alpha})}{s_i(\boldsymbol{\alpha})}\right]. \tag{IV.44}$$

It is easy to check that $\mathbf{X}\hat{}_i^{i-1} = \Phi\mathbf{X}\hat{}_{i-1}$ and thus the values $\hat{Y}_i^{i-1} = \hat{X}_i^{i-1} = (\hat{X}_i^{i-1})_1$ could be obtained from the conditional-mean values $\mathbf{X}\hat{}_i$, $1 \leqq i \leqq n$. The $\mathbf{X}\hat{}_i$

values can be obtained by the methods described in Section IV.1.2. Application of an ACM filter in (IV.44) results in our final form of approximate log-likelihood. Maximizing this approximate log-likelihood with respect to α yields an *approximate maximum likelihood (AML) estimate.*

IV.2.2 Definition of AM-estimates

Since the distribution F_V of the V_i will rarely be known in practice, ACM filtering (Section IV.1.2) and AML estimation (Section IV.2.1) cannot be performed. Thus Martin (1981b) follows the usual M-estimation route by replacing the score function Ψ (IV.21) by a good robustifying psi-function ψ (compare (II.3)–(II.5)) and by replacing the negative of the log-likelihood (IV. 44) by a loss function which uses a robustifying rho-function ϱ whose derivative is ψ.

We shall call the filter by (IV.19), (IV.31), (IV.36) to (IV.38) an ACM-filter even if Ψ is replaced by some good psi-function ψ and the ACM-filter could be named "robust filter". The negative of the log-likelihood (IV.44) is replaced by the loss function

$$L(\alpha) = \sum_{i=1}^{n} \log s_i(\alpha) + \sum_{i=1}^{n} \varrho\left[\frac{y_i - \hat{Y}_i^{i-1}(\alpha)}{s_i(\alpha)}\right] \tag{IV.45}$$

where $\hat{Y}_i^{i-1}$ and $s_i(\alpha)$ are obtained from the ACM filter recursions. If $\varrho(r) = -\log gd(r)$ and $\psi(r) = \Psi(r) = (\partial/\partial r) \log gd(r)$ then the minimization of $L(\alpha)$ is equivalent to the maximization of the approximate likelihood given by the right-hand side of (IV.44). If, in addition, $gd = N(0, 1)$ the above approximation (IV.44) becomes exact, yielding the Gaussian likelihood and $\hat{Y}_i^{i-1} = (\Phi \mathbf{X}\hat{}_{i-1})_1$ where $\mathbf{X}\hat{}_{i-1}$ are Kalman filter estimates (compare Kailath, 1968).

An *approximate maximum likelihood type (AM) estimate* of α is defined by any $\hat{\alpha}$ which minimizes the loss function $L(\alpha)$. For additive outliers models AM-estimates appear to be the most reasonable analogues of Huber's (1964, 1973) M-estimates for location and ordinary regression (Martin, 1981b). With ample smoothness conditions an AM-estimate is a solution of

$$(\partial/\partial\alpha)\, L(\alpha) = \sum_{i=1}^{n} \frac{(\partial/\partial\alpha)\, s_i}{s_i} - \sum_{i=1}^{n} \frac{y_i - \hat{Y}_i^{i-1}}{s_i^2}\, \psi\left[\frac{y_i - \hat{Y}_i^{i-1}}{s_i}\right](\partial/\partial\alpha)\, s_i - \tag{IV.46}$$

$$- \sum_{i=1}^{n} \frac{(\partial/\partial\alpha)\, \hat{Y}_i^{i-1}}{s_i}\, \psi\left[\frac{y_i - \hat{Y}_i^{i-1}}{s_i}\right] = \mathbf{0}\,.$$

IV.2.3 Computation of AM-estimates

An optimization algorithm for minimization of $L(\alpha)$ is not yet implemented. The reason is that things are more complicated than in the case of the Kalman filter and Gaussian likelihood. Instead, a relatively simple iterative scheme can be

used. The details of this scheme will be given after the following comments, which indicate that the simple iterative scheme will yield parameter estimates which bear reasonable resemblance to AM-estimates while are obtained by direct minimization of $L(\boldsymbol{\alpha})$.

One-sided Outlier-interpolator Mode

If we believe that the assumptions of an additive outliers model (compare Section I.4) are reasonable, what appears to be in many situations, then we set $\sigma_1^2 = 0$ in the contaminated normal distribution (IV.30), in the ACM filter recursions and in the AM-estimation equations. This results into

$$s_i = m_{1i}^{1/2} \tag{IV.47}$$

instead of (IV.31). The difficulty of the estimation problem is reduced by this assumption, because it eliminates the need to estimate σ_1^2.

On the other hand there are problems in which the nominal distribution for the additive noise V_i is a non-degenerate Gaussian distribution with variance σ_1^2 which is positive and unknown. In such cases we will be forced to estimate σ_1^2 as well as $\boldsymbol{\phi}, \boldsymbol{\theta}$ and σ^2. The optimization problem of minimizing $L(\boldsymbol{\alpha})$ (IV.45) appears then to be more difficult than if we set $\sigma_1^2 = 0$.

If σ_1^2 is chosen to be zero, if the parameter $\boldsymbol{\alpha}$ is known and if $\psi = \psi_{HA}$ (II.5), Martin (1981b) prefers to call the ACM filter a *one-sided outlier-interpolator*. The reason is that most of the data will be unaltered (i.e. $\hat{X}_i = Y_i$), while large outliers will be replaced by one-sided predictions (i.e. $\hat{X}_i = Y_i^{i-1}$) (compare Martin (1979c) for a more detailed description). This behavior should be unaltered if $\boldsymbol{\alpha}$ is replaced by a good estimate $\hat{\boldsymbol{\alpha}}$ like the AM-estimate obtained by solving (IV.46).

The Simplified Algorithm

Under the following moderate assumptions it is possible to rationalize a simple alternative to direct minimization of the AM loss function $L(\boldsymbol{\alpha})$:

A 1) At a solution point $\hat{\boldsymbol{\alpha}}$ of (IV.46) the ACM filter uses $\hat{\boldsymbol{\alpha}}$ in place of the true value $\boldsymbol{\alpha}$, the filter is operating in the one-sided outlier-interpolator mode with $\hat{X}_i = Y_i$ most of the time.

A 2) $(\partial/\partial\boldsymbol{\phi})\, s_i \approx \mathbf{0}$, $(\partial/\partial\boldsymbol{\theta})\, s_i \approx \mathbf{0}$.

A 3) $(\partial/\partial\sigma)\, \hat{Y}_i^{i-1} = 0$.

Usage of A 2 and A 3 results in the following significant simplification of the AM-estimation equation (IV.46)

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{Y}_i^{i-1}}{s_i} \psi \left[\frac{y_i - \hat{Y}_i^{i-1}}{s_i} \right] - 1 \right) \frac{1}{s_i} \frac{\partial s_i}{\partial \sigma} = 0 \tag{IV.49}$$

and

$$\sum_{i=1}^{n} \frac{(\partial/\partial(\boldsymbol{\phi}, \boldsymbol{\theta}))\, \hat{Y}_i^{i-1}}{s_i} \psi\left[\frac{y_i - \hat{Y}_1^{i-1}}{s_i}\right] = \mathbf{0}\,. \tag{IV.50}$$

These equations have the same form as the maximum likelihood estimating equations for nonlinear regression with error density g and time-varying scale parameter $s_i = (m_{1i})^{1/2} \approx \hat{\sigma}$.

Using (IV.47) in the first row of (IV.36), noting that $\hat{Y}_i^{i-1} = \hat{X}_i^{i-1}$ and using the simplifying assumption that $s_i \approx \hat{\sigma}$, allows us to write the estimating equation (IV.50) as

$$\sum_{i=1}^{n} (\partial/\partial(\boldsymbol{\phi}, \boldsymbol{\theta}))\, \hat{X}_i^{i-1}(\hat{X}_i - \hat{X}_i^{i-1}) = 0\,. \tag{IV.51}$$

But if $\hat{X}_i = Y_i$ most of the time, this equation provides an approximate solution to the least squares problem

$$g(\boldsymbol{\phi}', \boldsymbol{\theta}') = \sum_{i=1}^{n} (\hat{X}_i - \hat{X}_i^{i-1}(\boldsymbol{\phi}', \boldsymbol{\theta}'))^2 = \min \tag{IV.52}$$

because Y_i is independent of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. This is essentially the usual least squares formulation of ARMA model fitting, except that the y_i's are occasionally replaced by one-sided interpolates and the usual approximations to the residuals when MA terms are present are replaced by approximations based on occasionally altered y_i's.

The foregoing explanations suggest the following iterative technique. Start with initial crude but robust estimates of $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, σ and use the estimates to process the data y_i, $i = 1, \ldots, n$, by an ACM filter. Use the resulting $\hat{X}_i$'s in a nonlinear least squares ARMA estimation program (use e.g. the algorithm described in Section II.3.2 with $\psi(t) = t$) to solve (IV.52) with the $\hat{X}_i$'s fixed. Iterate this procedure with care until there is little change in the estimates. Here is a more detailed description:

Preliminary Estimates:

P 0. Center the data with an ordinary location M-estimator.

P 1. Fit robustly a longish autoregression using the GM-estimation method (Section III.1) to compute $\boldsymbol{\phi}^{\sim} = (\phi_1^{\sim}, \ldots, \phi_L^{\sim})$ and $\hat{\sigma}$.

P 2. Use $\boldsymbol{\phi}^{\sim}$ to compute preliminary ARMA parameter estimates

$$\boldsymbol{\phi}^{(0)} = (\phi_1^{(0)}, \ldots, \phi_p^{(0)})^{\mathrm{T}}, \quad \boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_q^{(0)})^{\mathrm{T}}, \quad \sigma^{(0)}$$

using Durbin's (1959) idea; an alternative reference for Durbin's technique is Fuller (1976), pp. 281–283. These estimates in turn supply preliminary estimates $\Phi^{(0)}$, $Q^{(0)}$ of the state transition matrix and innovations covariance matrix, respectively.

P 3. Use the initial estimates $\boldsymbol{\phi}^{(0)}$, $\boldsymbol{\theta}^{(0)}$, $\sigma^{(0)}$ to compute an initial estimate $C_x^{(0)}$ of the covariance matrix for the state vector $\mathbf{x}_1$; this is done by solving $C_x^{(0)} = \hat{\Phi} C_x^{(0)} \hat{\Phi}^{\mathrm{T}} + \hat{Q}$.

Basic Iteration:

0. Set $j = 0$ and $\sigma_1^2 = 0$, so that $s_i = m_{1i}^{1/2}$.
1. Run a good ARMA ACM outlier interpolator based on $\phi^{(J)}$, $\theta^{(J)}$, $\sigma^{(J)}$, $C_x^{(J)}$ backward in time using initial conditions for Masreliez's theorem (Section IV.1.2). Thus we first compute
$$\mathbf{X}_{R,n}^{(j)} = (\mathbf{m}_{R,N}^{(j)} / S_{R,n}^{(j)})\, \psi(y_n / S_{R,n}^{(j)})$$
where $\mathbf{m}_{R,n}^{(j)}$ is the first column of $C_x^{(j)}$ and $S_{R,n}^{(j)}$ is the 1-1 element of $C_x^{(j)}$. The $\mathbf{X}_{R,n-1}^{(j)}, \ldots, \mathbf{X}_{R,n-2}^{(j)}, \ldots, \mathbf{X}_{R,1}^{(j)}$ are computed by running the recursions (IV.36) to (IV.38), (IV.31) and (IV.19) backward in time.
2. Now run the ACM filter in the forward direction using
$$\mathbf{X}_1^{(j)} = \mathbf{X}_{R,1}^{(j)}, \quad \mathbf{m}_1^{(j)} = \mathbf{m}_{R,1}^{(j)}, \quad S_1^{(j)} = S_{R,1}^{(j)}$$
as initial conditions. The "outlier-interpolated" or filtered series at iteration j is $X_i^{(j)} = (\mathbf{X}_i^{(j)})_1$, $1 \leqq i \leqq n$.
3. Use $X_1^{(j)}, \ldots, X_n^{(j)}$ as input to an ARMA model fitting routine and compute $\phi^{(j+1)}$, $\theta^{(j+1)}$, $\sigma^{(j+1)}$).
4. Compute $C_x^{(j+1)}$ from $\phi^{(j+1)}$, $\theta^{(j+1)}$, $\sigma^{(j+1)}$.
5. Let $\Delta^{(j+1)} = (\phi^{(j+1)\mathrm{T}} - \phi^{(j)\mathrm{T}}, \theta^{(j+1)\mathrm{T}} - \theta^{(j)\mathrm{T}})$.
If $\|\Delta^{(j+1)}\| < \varepsilon \hat{\Sigma}_{kk}^{(j+1)}$ where ε is a tolerance value and $\hat{\Sigma}_{kk}^{(j+1)}$, $1 \leqq k \leqq p + q$, is the estimated standard error for the coefficient estimates, then go to 7, else go to 6.
6. Augment $j = j + 1$ and go to Step 1.
7. Stop.

IV.2.4 Conditional-mean M-estimates

Martin (1979) defines conditional-mean M-estimates for autoregressive parameters. These estimates can be regarded as AM-estimates for AR models. Since things behave more clearly than in the case of ARMA models and since the idea is slightly different, this method of estimating AR parameters will be presented.

Let an AR process with a location of value zero be given, where the process is possibly contaminated by additive outliers. Let $\mathbf{y}^i = (y_1, \ldots, y_i)$ and $\mathbf{X}\hat{}_i = E\{\mathbf{X}_i \mid \mathbf{Y}^i\}$ be defined as in (IV.9) and (IV.13), respectively. Let $\mathbf{X}_i$ denote $(X_{i-1}, \ldots, X_{i-p})^{\mathrm{T}}$.

A conditional-mean M-estimate (*CMM-estimate*) is a solution of the minimization problem

$$L(\phi) = \sum_{i=p+1}^{n} \varrho\left[\frac{y_i - \mathbf{X}\hat{}_i^{\mathrm{T}}\phi}{\hat{\sigma}}\right] = \min . \tag{IV.53}$$

where ϱ is a symmetric robustifying loss function and the scale estimate $\hat{\sigma}$ is yet to be specified.

It can be checked that

$$\text{(IV.54)} \qquad \hat{Y}_i^{i-1} = E\{Y_i \mid \mathbf{Y}^{i-1}\} = E\{X_i \mid \mathbf{Y}^{i-1}\} = \hat{X}_i^{i-1} = (\hat{\mathbf{X}}_{i+1}^{i-1})_1 = \hat{\mathbf{X}}_i^{\mathrm{T}} \boldsymbol{\phi}$$

In Section IV.1.2 $f_y(y_i \mid \mathbf{y}^{i-1})$ is approximated by $gd_i(y_i - \hat{Y}_i^{i-1})$ (IV.25) and the latter expression further by $(1/s_i)\, gd((y_i - \hat{Y}_i^{i-1})/s_i)$ (IV.27). In order to obtain $L(\boldsymbol{\alpha})$ (IV.45) $-\log gd(r)$ was replaced by $\varrho(r)$. If in (IV.43) $f_y(y_i \mid \mathbf{y}^{i-1})$ is approximated by $gd((y_i - \hat{Y}_i^{i-1})/s_i)$ and if $-\log gd(r)$ is set equal to $\varrho(r)$ and $s_i = \hat{\sigma}$, then it can be seen that minimizing $L(\boldsymbol{\phi})$ is equivalent to maximizing the autoregressive version of the log-likelihood (IV.43) approximately and minimizing $L(\boldsymbol{\phi})$ corresponds to the minimization of $L(\boldsymbol{\alpha})$ defined in (IV.45).

Since the solution of the minimum problem (IV.53) is a stationary point we have

$$\text{(IV.55)} \qquad \sum_{i=p+1}^{n} [\hat{\mathbf{X}}_i + D_i(\hat{\boldsymbol{\phi}})\, \hat{\boldsymbol{\phi}}]\, \psi\left[\frac{y_i - \hat{\mathbf{X}}_i^{\mathrm{T}} \hat{\boldsymbol{\phi}}}{\hat{\sigma}}\right] = \mathbf{0}$$

where

$$\text{(IV.56)} \qquad [D_i(\hat{\boldsymbol{\phi}})]_{kj} = \left.\frac{\partial \hat{X}_{i-j}}{\partial \phi_k'}\right|_{\boldsymbol{\phi}' = \hat{\boldsymbol{\phi}}}, \quad k, j = 1, \ldots, p\,.$$

The estimating equation (IV.55) is rather hard to solve due to the presence of $D_i(\hat{\boldsymbol{\phi}})$. There is some evidence in the form of both heuristic arguments and Monte Carlo (Martin, 1979), that the term $D_i(\hat{\boldsymbol{\phi}})\, \hat{\boldsymbol{\phi}}$ may be dropped without seriously degrading the estimate. Thus we turn to the simpler approximate version

$$\text{(IV.57)} \qquad \sum_{i=p+1}^{n} \hat{\mathbf{X}}_i\, \psi\left[\frac{y_i - \hat{\mathbf{X}}_i^{\mathrm{T}} \hat{\boldsymbol{\phi}}}{\hat{\sigma}}\right] = \mathbf{0}\,.$$

One method for obtaining the estimate $\hat{\sigma}$ could be to use the side condition

$$\text{(IV.58)} \qquad \frac{1}{n-2p} \sum_{i=p+1}^{n} \psi^2\left[\frac{Y_i - \hat{\mathbf{X}}_i^{\mathrm{T}} \hat{\boldsymbol{\phi}}}{\hat{\sigma}}\right] = b$$

corresponding to Huber's (1973) proposal for estimating regression coefficients and scale simultaneously (compare (II.6)).

Another method to obtain $\hat{\sigma}$ is provided by the filter algorithms discussed in Section IV.1. In order to solve (IV.57) we need to express $\hat{\mathbf{X}}_i$ as a function of $\boldsymbol{\phi}$ and $\mathbf{y}^i$ for the pth order autoregressive additive outliers model. Good approximate versions of the estimates $\hat{\mathbf{X}}_i$ could be computed by Masreliez's filter theorem (Section IV.1.2). Note that an AR (p) process can be written in the state-variable form (IV.1) and (IV.2) by setting

$$\text{(IV.59)} \qquad \Phi = \begin{bmatrix} \phi_1 & \phi_2 & \ldots & \phi_p \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots\ 1 & 0 \end{bmatrix},$$

$$\text{(IV.60)} \qquad a_i = (a_{i-1}, 0, \ldots, 0)^{\mathrm{T}}$$

and $\mathbf{x}_i$ defined as in this Section, i.e. $(\mathbf{x}_i)_1 = x_{i-1}, \ldots, (\mathbf{x}_i)_p = x_{i-p}$.

Because we will not know the distribution F_V of the V_i's and therefore the distribution $(Y_i \mid \mathbf{Y}^{i-1})$ is also unknown, we have to replace the score function Ψ_i (IV.21) for the observation-prediction density with a bounded and continuous function. We can use (IV.33) to approximate Ψ_i and we can further replace $\psi(r) = -(\partial/\partial r) \cdot \log gd(r)$ by $(\mathrm{d}/\mathrm{d}r)\, \varrho(r) = \psi(r)$, where ψ is a usual psi-function. If V_i has the contaminated normal distribution (IV.30) with $\sigma_1^2 = 0$, then we can set $\hat{s}_i^2 = m_{1i}$ (see also (IV.47)) where m_{1i} is the 1-1 element of M_i (IV.16). Further details including some Monte Carlo results for various ψ shapes are given by Martin and De Bow (1976).

With the above approximations

$$\Psi_i(y_i) \approx \frac{1}{\hat{s}_i} \psi\left[(y_i - \hat{Y}_i^{i-1})/\hat{s}_i\right] \tag{IV.61}$$

and

$$m_{1i} \approx \hat{s}_i^2 \tag{IV.62}$$

in mind, an attractive simplification of Masreliez's filter (IV.18) is

$$\hat{X}_i = \mathbf{X}\hat{}_i^{\mathrm{T}} \boldsymbol{\phi} + \hat{s}_i \psi\left[\frac{y_i - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\phi}}{\hat{s}_i}\right] \tag{IV.63}$$

with $\hat{s}_i$ obtained from the data-dependent auxiliary recursion (IV.19) for M_i. It turns out that (IV.63) is a special case of (IV.36).

Thomson (1977) and Kleiner, Martin and Thomson (1979) (compare also Huber, 1982) used – in connection with spectral density estimation – a robust AR filter which is a further simplification of (IV.63)

$$\hat{X}_i = \mathbf{X}\hat{}_i^{\mathrm{T}} \boldsymbol{\phi} + \hat{s} \psi\left[\frac{y_i - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\phi}}{\hat{s}_i}\right] \tag{IV.64}$$

where $\hat{s}$ is a data-dependent but time-invariant estimate of the scale for the prediction residuals $y_i - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\phi}$. For example, $\hat{\sigma}$ might be determined by (IV.57) and (IV.58). The above filter is referred to as a fixed-scale filter. See also Masreliez and Martin (1977) for some theory about such filters when $\boldsymbol{\phi}$ and are known. Martin (1979) prefers (IV.63) instead of (IV.64) because the scale factor $\hat{s}_i$ depends on the local character of the data and if ψ is redescending then the version (IV.64) is unsafe, because it can then loose track of the data never to regain it.

Notice that if ψ and $\hat{\sigma}$ are chosen to be the same in (IV.57) and (IV.64), what is hardly unreasonable, then it is not necessary to solve equation (IV.57) directly. For multiplying both sides of (IV.64) by $\mathbf{X}\hat{}_i = \mathbf{X}\hat{}_i(\boldsymbol{\phi})$ and summing over i shows that (IV.57) is equivalent to the Yule-Walker type normal equations (compare Box and Jenkins, 1976)

$$\sum_{i=p+1}^{n} \mathbf{X}\hat{}_i(\boldsymbol{\phi}\hat{}) \left[\hat{X}_i(\boldsymbol{\phi}\hat{}) - \mathbf{X}\hat{}_i^{\mathrm{T}}(\boldsymbol{\phi}\hat{})\, \boldsymbol{\phi}\hat{}\right] = \mathbf{0}. \tag{IV.65}$$

The above equation invites the iterative solution

$$\sum_{i=p+1}^{n} \mathbf{X}\hat{}_i(\phi^j)\left[\hat{X}_i(\phi^j) - \mathbf{X}\hat{}_i^{\mathrm{T}}(\phi^j)\,\phi^{j+1}\right] = \mathbf{0}, \quad j = 1, 2, \ldots \tag{IV.66}$$

where $\mathbf{X}\hat{}_i(\phi^j)$ is obtained from (IV.64) with $\phi = \phi^j$ and ϕ^1 is the least squares estimate.

When the observed series $y_1, \ldots, y_n$ contains a relatively small fraction of outliers, the properly calibrated robust scale estimate $\hat{\sigma}$ computed from (IV.58) should differ relatively little from the square root of the usual σ^2 computed from the residuals of the final iteration of (IV.66). Thus the latter simpler method which is applied in conventional least squares procedures, might be adequate.

Some exploratory Monte Carlo results yielded smaller biases for CMM-estimates than for GM-estimates at non-Gaussian additive outlier situations. The corresponding variances were also typically smaller. Efficiences at the Gaussian situation were reasonably high. The Monte Carlo investigations also showed (not unexpectedly) that the performance of CMM-estimates is quite poor at heavy-tailed innovation outlier situations.

IV.3 ROBUST FILTERING AND ROBUST SMOOTHING

A remarkable method to estimate parameters of time series models is to perform a robust filtering algorithm or robust smoothing algorithm on the time series observations which could be contaminated by outliers, so that the outliers are replaced by reasonable values and then to estimate parameters by usual least squares.

A fact to be considered in this context is that most filtering and smoothing algorithms do not fully exploit the information in the data, e.g. they may neglect the correlations between neighboring points. Also the literature in general reveals no attempts to design robust smoothers and filters which are optimal for particular non-Gaussian model specifications. Nevertheless, methods which do not fully exploit the information in the data will be mentioned shortly below. An exception are the ACM filters and ACM smoothers described in Section IV.1, because the filtering and smoothing algorithms assume that the series of interest satisfies an ARMA (p, q) model with additive outliers. Another exception is the robustified Kalman filter introduced by Masreliez and Martin (1977). The essence of this filter will be described in the next Section.

IV.3.1 The Robustified Kalman Filter

In order to construct a robustified Kalman filter, Masreliez and Martin (1977) begin to obtain robust Bayesian estimates $\mathbf{x}\hat{}$ of a vector $\mathbf{x}$ in the linear model

$$\mathbf{y} = H\mathbf{x} + \mathbf{v} \tag{IV.67}$$

for the following two distinct situations:

(i) the state $\mathbf{X}$ is Gaussian and the observation error $\mathbf{V}$ is (heavy-tailed) non-Gaussian (this situation is similar to a time series with additive outliers)

(ii) the state $\mathbf{X}$ is heavy-tailed non-Gaussian and the observation error is Gaussian (this situation can be compared with a time series contaminated by innovation outliers).

In order to apply the estimation procedure it is necessary to transform the linear model (IV.67), so that two certain distributional properties are fulfilled. A lemma insures the existence of an approximate transformation whenever $\mathbf{V}$ has a contaminated normal or elliptical distribution.

Estimating $\mathbf{x}$ requires the knowledge of the covariance matrix of $\mathbf{X}$, what is natural in the Bayesian context, but appears to be a strong restriction for practical computations.

With the results for the linear model (IV.67) in hand it is possible to construct a dynamic filter type estimator through step by step implementation of single step robust Bayesian estimators. The model used now is

$$\mathbf{x}_i = \Phi_i \mathbf{x}_{i-1} + \mathbf{a}_i \tag{IV.68}$$

$$\mathbf{y}_i = H_i \mathbf{x}_i + \mathbf{v}_i \tag{IV.69}$$

which is closely related to the state-variable representation (IV.1), (IV.2) of an ARMA model. Clearly, the simplifications $\Phi_i = \Phi$, $H_i = H$, $\mathbf{Y}_i = Y_i$ and $\mathbf{V}_i = V_i$ would lead to the construction of a filtered ARMA process.

IV.3.2 L-Smoothers

Perhaps the currently best-known type of robust smoothers are those based on moving order statistics as introduced by Tukey (1977). The most simple example of such a smoother would be a moving median of prescribed span. Often odd-span running medians are used. In contrast, Velleman (1975) proposed even-span running medians to reduce difficulties found in odd-span medians. Running medians are often combined with each other and with simple linear filters to improve their performance. Velleman's (1980) article lists a collection of non-linear smoothers based upon running medians and presents methods for describing and comparing their performance, what is not quite easy in face of the non-linearity. A device which is often effective is called "twicing". To understand this device we denote the smoothed value of y_i by $Sm(y_i)$, and remark that a data smoother separates the sequence $\{y_i\}$ into the smooth $\{z_i\} = Sm\{y_i\}$ and the rough $\{r_i\} = \{y_i - z_i\}$. The iterative improvement $\{z_i\} = Sm\{y_i\} + Sm\{r_i\}$ is used to recover patterns from the residuals r_i and is called "twicing".

By analogy to the use of the term "L-estimator" to describe any of a broad class of location parameter estimators based on order statistics we shall refer to smoothers based on moving order statistics as L-smoothers.

Papers on L-smoothers appeared in the engineering literature (Rabiner, Sambur and Schmidt, 1975; Justusson, 1977; Huang, Yang and Tang, 1979) and in econometrics (Polasek, 1982b). This is no doubt due to the real need for some kind of robust smoothing to deal with outliers in time and space series, along with the fact that L-smoothers have rather obvious and intuitively appealing resistance properties.

IV.3.3 Moving M-estimate Smoothers

If L-smoothers are good robust smoothers, then it would come without surprise to find that moving maximum likelihood type estimates of location (Huber, 1964) provide useful robust smoothers. On can find pertinent discussions in the papers of Cleveland (1979, 1982) and Stuetzle (1979).

IV.3.4 Robustified Splines

Let $\{y_i\}$, $1 \leqq i \leqq n$, be the series to be smoothed, let Sm be a smoothing operator and let $\{z_i\} = Sm\{y_i\}$ be the smoothed series.

The theoretically cleanest approach to linear smoothing is through splines (Reinisch, 1967): minimize the mean square of the second (or of a higher order) derivative of z

$$ave\{(z_i'')^2\} \to \min \tag{IV.70}$$

subject to a side condition of the form

$$ave\{(y_i - z_i)^2\} \leqq \text{const}. \tag{IV.71}$$

The means are taken over a suitable range of i-values.

This approach can be robustified very easily (Huber's (1979) paper is a basic reference to this approach): we simply replace the square in (IV. 70) by a less rapidly increasing function ϱ. Past experience with location and regression estimates suggests that ϱ should be chosen convex with a bounded derivative $\psi = \varrho'$, for example

$$\varrho(x) = \begin{cases} \frac{1}{2}x^2 & \text{for } |x| \leqq c \\ c|x| - \frac{1}{2}c^2 & \text{for } |x| > c \end{cases} \tag{IV.72}$$

where the constant c regulates the degree of robustness. As Huber (1979) mentions, robustifying splines have been considered often but little has appeared in the literature (see however Lenth, 1977).

IV.3.5 Problems with Robust Filters and Smoothers

It should be noted that, in general, literature does not distinguish between filters and smoothers in the sense of Section IV.1.2 and IV.1.3, respectively. The terms "filter" and "smoother" are used interchangeably.

Although robust linear filters exist, robust smoothers are inherently nonlinear (see Kassam and Poor, 1985). Nonlinearity causes more problems. Nonlinear smoothers fall outside the classical framework of linear filter theory and are difficult to analyze mathematically. One difficulty is that it is not possible to characterize a nonlinear filter by its transfer function, which is a well known advantage of linear filters. Nonlinearity can also cause transfer of power from one frequency to another.

However, nonlinear data smoothers provide a practical method of finding general smooth patterns for sequenced data confounded with heavy-tailed noise.

The various approaches to robust filtering and smoothing described in this chapter all share the common property of being resistant toward outliers. A detailed understanding of their features in probabilistic terms, however, has been lacking for a long period, because there has been a scarcity of tools which are necessary for the careful statistical analysis of the behavior of nonlinear smoothers. Thus it has been difficult for potential users to determine which of several approaches, and which particular smoother within a given class, will be a good one for his problems.

Mallows (1980a, 1980b) contributes significantly to the theory of nonlinear smoothers what should greatly enhance our ability to analyze proposed robust smoothers of many varieties. A very important aspect of his work is a theorem which characterizes the "linear part" of a nonlinear smoother, and provides an additive orthogonal decomposition of the smoothers into the linear part and a residual process. Presumably a good robust smoother would have a linear part which is "close" to the linear smoother which the user would prescribe for an outlier-free process, and a residual process which is relatively "small". It should be noted that Mallow's decomposition theorem is primarily of use for the analysis but not for the design of robust smoothers.

V. SOME RESULTS CONCERNING APPLICATION AND FURTHER RESEARCH

A Monte Carlo investigation of methods for the least squares estimation, M-estimation and GM-estimation of ARMA models will be presented. Monte Carlo generally reveals properties which are expected from theory. For outlier-free data the means of the estimated parameters differ scarcely, and the mean square errors of M-estimators and GM-estimators are larger than those for least squares estimators. For the processes chosen here, with innovation outliers, the means of the estimated parameters also differ only slightly, but the sample relative efficiencies of M-estimators are larger than the sample relative efficiencies of GM-estimators and of least squares estimators. In the presence of additive outliers the GM-estimation essentially yields better parameters and substantially smaller mean square errors than the least squares estimation and than the M-estimation.

Several topics for further research concerning identification and estimation of various models, outlier detection, filters and spectral density estimation will be mentioned.

V.1 SOME MONTE CARLO RESULTS FOR GM-ESTIMATORS OF AR MODELS

In order to study the behavior of various estimators of AR models, which were discussed in Section II.1 and Section III.1, AR(1) processes with location $\mu = 0$ were simulated. The number of observations for each process is 100. The number of replications for each process is 50. The V_i's that cause additive outliers have a Gaussian mixture distribution $CND(\kappa, \sigma_3) = (1 - \kappa)\,\delta_0 + \kappa N(0, \sigma_3^2)$ with $\sigma_3^2 = 9\ VAR\ X_i$ (compare Section I.4). (For an AR(1) model, $VAR\ X_i = \sigma^2/(1 - \phi_1^2)$,

Table V.1. Simulated AR(1) processes.

Abbreviation	ϕ_1	ν	κ
ARGP 5	·5	0·	0·
ARGP 8	·8	0·	0·
ARIOCNP 5	·5	·1	0·
ARIOCNP 8	·8	·1	0·
ARAOP 5	·5	0·	·1
ARAOP 8	·8	0·	·1

where σ denotes the scale of the innovations.) Abbreviations and values for ϕ_1, v and κ for processes with $CN(v, 1, 11) = (1 - v)\,N(0, 1) + vN(0, 121)$ – distributed innovations are given in Table V.1.

Furthermore, processes with t_4-distributed innovations without additive outliers were simulated for $\phi_1 = \cdot 5$ (ARIOTP 5) and $\phi_1 = \cdot 8$ (ARIOTP 8). Note that the $CN(\cdot 1, 1, 11)$ distribution and the t_4-distribution have both variance 2.

Pseudo random numbers from the normal distribution with mean 0 and variance 1 are generated by a comparison method implemented in the algorithm FL (Forsythe, Ahrens-Dieter) given by Ahrens and Dieter (1974). The algorithm FL uses a multiplicative congruental generator with factor $a = 5\,308\,871\,541$ and module $m = 2^{35}$ to generate pseudo random numbers U_i, $i = 1, 2, \ldots$, uniformly distributed between 0 and 1.

Pseudo random numbers A_i, $i = 1, 2, \ldots$, with a $CN(v, \sigma_1, \sigma_2)$ distribution are generated in the following way.

(1) Set $i = 1$.
(2) Generate a pseudo random number U_i from a uniform distribution between 0 and 1.
(3) Set A_i equal to a pseudo random number from a $N(0, \sigma_1^2)$ distribution, if $U_i > v$.
(4) Set A_i equal to a pseudo random number from a $N(0, \sigma_2^2)$ distribution, if $U_i \leqq v$.
(5) Stop, if enough A_i's are generated.
(6) Augment $i = i + 1$ and go to (2).

Pseudo random numbers V_i, $i = 1, 2, \ldots$, from a $CND(\kappa, \sigma_3)$ distribution are generated by the following algorithm.

(1) Set $i = 1$.
(2) Generate a pseudo random number U_i from a uniform distribution between 0 and 1.
(3) Set $V_i = 0$, if $U_i > \kappa$.
(4) Set V_i equal to a pseudo random number from a $N(0, \sigma_3^2)$ distribution, if $U_i \leqq \kappa$.
(5) Stop, if enough A_i's are generated.
(6) Augment $i = i + 1$ and go to (2).

Pseudo random variables with a t-distribution are generated by a modified rejection method given by Stadlober and Dieter (1985).

For the estimation of AR models for the simulated processes it was assumed that the order of the model to be fitted were known, but no information about the parameters to be estimated would be given. Therefore starting values $\hat{\phi}$ were determined by the Yule-Walker equations (Box and Jenkins, 1976) and the starting value $\hat{\sigma}^2$ was $\hat{\gamma}_0 - \hat{\phi}_1\hat{\gamma}_1 - \ldots - \hat{\phi}_p\hat{\gamma}_p$, where $\hat{\gamma}_k$ denotes an estimate of the autocovariance of the lag k for the given time series. These starting values were used to compute least squares estimates by the IWLS algorithm (compare Section II.1.2) in the Monte Carlo study.

Table V.2 lists the methods that were used to fit AR(1) models to the simulated

AR(1) processes. In the abbreviations the letters H, HA and B stand for Huber's psi, Hampel's psi and Tukey's bisquare psi, respectively (compare Section II.1.1). The letters M, MA and S stand for M-estimators, Mallows type GM-estimators and Schweppe type GM-estimators, respectively (compare Section III.1.1.). Clearly, LS stands for least squares. The starting values for one estimation in general are the results of the preceding estimation except for SH, where the starting values are the results of MB. The tolerance value for the estimations is $\varepsilon = \cdot 001$. The locations μ of the given time series is assumed to be zero. The estimating equation is (III.8), where the "largeness" of y_{t-1} is determined by (III.9).

Table V.2. Types of estimations.

Abbreviation	ψ_1	ψ_2
LS	Identity	Identity
MH	$\psi_{H'}$ $c = 1{\cdot}345$	Identity
MHA	$\psi_{HA'}$ $a = 1{\cdot}4$, $b = 2{\cdot}8$, $d = 4{\cdot}75$	Identity
MB	$\psi_{B'}$ $c = 4{\cdot}685$	Identity
MAH	$\psi_{H'}$ $c = 1{\cdot}65$	equal to ψ_1
MAHA	$\psi_{HA'}$ $a = 1{\cdot}7$, $b = 3{\cdot}4$, $d = 5{\cdot}0$	equal to ψ_1
MAB	$\psi_{B'}$ $c = 5{\cdot}58$	equal to ψ_1
SH	$\psi_{H'}$ $c = 1{\cdot}6$	equal to ψ_1
SHA	$\psi_{HA'}$ $a = 1{\cdot}7$, $b = 3{\cdot}4$, $d = 5{\cdot}5$	equal to ψ_1
SB	$\psi_{B'}$ $c = 6{\cdot}0$	equal to ψ_1

For each estimator, except for the least squares estimator, the constants of the ψ-functions were chosen so that the asymptotic efficiency of the estimator relative to the least squares estimator is ·95 for outlier-free data, where this efficiency is the ratio of the asymptotic variance of the least squares estimator and the estimator in question. With this setting of the constants of the psi-functions comparisons of the estimators make sense. The computation of the asymptotic variances is explained in Section III.2.3.

In the case of $\psi_1 = \psi_{HA}$ the IWLS algorithm is first run with $\psi_1 = \psi_H$, where c is equal to the constant a of ψ_{HA}, to obtain an estimate for σ. If $\psi_1 = \psi_B$ the IWLS algorithm is first run with $\psi_1 = \psi_H$, where the constant is equal to the constant c of ψ_B divided by $\sqrt{5}$, because this $\psi_H(t)$ with $-c/\sqrt{(5)} \leqq t \leqq c/\sqrt{(5)}$ is similar to the increasing part of $\psi_B(t)$.

For each type of simulated processes the mean (MEAN), the mean square error (MSE) of the estimates of ϕ_1 and the mean of the averages of the final weights $w_i^{(m)}$ in the IWLS algorithm (MAVW) were computed. The MSE is a measure for the variability of the method of estimation and is defined by

$$\text{MSE} = \text{REP}^{-1} \sum_{k=1}^{\text{REP}} (\hat{\phi}_{1,k} - \phi_1)^2$$

where REP denotes the number of replications for one type of simulated processes and $\hat{\phi}_{1,k}$ is the estimate of the true parameter value ϕ_1 for the kth simulated process. The MAVW tells about the portion for the kth simulated process. The MAVW for others than the least squares estimator can be expected to be smaller in the presence of additive outliers than in the presence of innovation outliers.

The Table V.3 to V.6 summarize the results of the estimations for the simulated processes. EFF denotes the sample relative efficiency of the estimate with respect to the least squares estimate, i.e. the ratio between the MSE of the least squares

Table V.3. Results of estimations (described in Section II.1 and Section III.1) of ϕ_1 for outlier-free processes.

ESTIMATOR	Simulated processes							
	ARGP 5				ARGP 8			
	MEAN	MSE	EFF	MAVW	MEAN	MSE	EFF	MAVW
LS	·462	1·23	1·00	1·00	·772	·572	1·00	1·00
MH	·459	1·35	·911	·961	·770	·666	·858	·960
MHA	·460	1·34	·919	·965	·770	·659	·867	·965
MB	·459	1·33	·921	·917	·770	·661	·865	·917
MAH	·460	1·35	·909	·962	·769	·679	·842	·960
MAHA	·461	1·33	·921	·966	·769	·672	·850	·964
MAB	·460	1·37	·895	·879	·769	·690	·829	·875
SH	·460	1·36	·901	·977	·770	·691	·827	·977
SHA	·461	1·32	·934	·983	·770	·674	·848	·982
SB	·461	1·31	·940	·945	·769	·665	·859	·945

Table V.4. Results of estimations (described in Section II.1 and Section III.1) of ϕ_1 for processes with *CN*-distributed innovations.

ESTIMATOR	Simulated processes							
	ARIOCNP 5				ARIOCNP 8			
	MEAN	MSE	EFF	MAVW	MEAN	MSE	EFF	MAVW
LS	·470	1·11	1·00	1·00	·776	·750	1·00	1·00
MH	·478	·600	1·85	·935	·780	·388	1·94	·935
MHA	·482	·532	2·08	·932	·782	·330	2·27	·932
MB	·482	·527	2·10	·896	·783	·323	2·33	·897
MAH	·472	·935	1·18	·917	·777	·487	1·54	·923
MAHA	·474	·923	1·20	·914	·778	·485	1·55	·922
MAB	·470	1·02	1·09	·845	·777	·498	1·51	·851
SH	·472	·798	1·39	·951	·780	·414	1·81	·950
SHA	·468	·925	1·20	·944	·781	·410	1·83	·949
SB	·468	·880	1·26	·927	·780	·393	1·91	·931

Table V.5. Results of estimations (described in Section II.1 and Section III.1) of ϕ_1 for processes with t-distributed innovations.

ESTIMATOR	Simulated processes							
	ARIOTP 5				ARIOTP 8			
	MEAN	MSE	EFF	MAVW	MEAN	MSE	EFF	MAVW
LS	·493	·859	1·00	1·00	·777	·500	1·00	1·00
MH	·495	·731	1·18	·940	·779	·420	1·19	·940
MHA	·496	·746	1·15	·941	·780	·415	1·20	·942
MB	·496	·729	1·18	·906	·781	·411	1·22	·906
MAH	·496	·779	1·10	·934	·780	·459	1·09	·933
MAHA	·497	·827	1·04	·935	·779	·475	1·05	·935
MAB	·499	·820	1·05	·861	·779	·490	1·02	·859
SH	·495	·794	1·08	·957	·780	·468	1·09	·956
SHA	·499	·846	1·02	·958	·779	·501	·998	·958
SB	·499	·821	1·05	·934	·779	·498	1·00	·935

Table V.6. Results of estimations (described in Section II.1 and Section III.1) of ϕ_1 for processes with additive outliers.

ESTIMATOR	Simulated processes							
	ARAOP 5				ARAOP 8			
	MEAN	MSE	EFF	MAVW	MEAN	MSE	EFF	MAVW
LS	·211	9·47	1·00	1·00	·348	21·7	1·00	1·00
MH	·212	9·23	1·03	·935	·405	17·5	1·24	·922
MHA	·208	9·40	1·01	·932	·427	16·7	1·30	·915
MB	·207	9·45	1·00	·897	·403	18·0	1·21	·889
MAH	·297	5·51	1·72	·918	·545	8·06	2·70	·900
MAHA	·329	4·50	2·11	·912	·602	5·22	4·16	·890
MAB	·333	4·42	2·14	·844	·597	5·38	4·04	·832
SH	·282	6·33	1·50	·941	·553	7·73	2·81	·910
SHA	·334	4·67	2·03	·930	·657	3·36	6·47	·895
SB	·321	5·19	1·83	·915	·613	5·04	4·31	·898

estimate and the MSE of the estimate in question. The columns denoted by MSE contain 100 times the mean square error.

The following comments are referring to the results shown in Tables V.3 to V.6: The means of the estimated parameters differ only slightly for the outlier-free processes and the processes with innovation outliers (Tables V.3 to V.5), but substantially for processes with additive outliers. This results from the fact that the contamination by innovation outliers is rather mild. The sample relative efficiency of all estimators, except of the least squares estimator, for the outlier-free data

is a little bit less than the desired asymptotic relative efficiency of ·95. Taking the sample relative efficiency as a measure for the performance of an estimator, the M-estimators outperform the GM-estimators, if the data contain innovation outliers. The Schweppe type estimators are superior to the Mallows type estimators with the same psi-functions for data with *CN*-distributed innovations. Data with *t*-distributed innovations – in contrast to the theoretical expectation – are slightly better estimated by Mallows type estimators than by Schweppe type estimators and the mean square error for the Schweppe type estimator with Hampel's psi ARIOTP 8 is even larger than the mean square error of the least squares estimator. If additive outliers are present, GM-estimators give better results, or more precisely speaking, estimated parameters closer to the true parameters and smaller mean square errors. In particular, it is demonstrated that GM-estimators using redescending psi-functions have high efficiencies. For $\phi_1 = \cdot 5$ the Mallows type estimators are superior to Schweppe type estimators with the same psi-functions, as expected from the theory. In the case of $\phi_1 = \cdot 8$, however, the Schweppe type estimators are superior to the corresponding Mallows type estimators.

Some interesting sample relative efficiencies and means of estimates condensed in Tables V.3 to V.6 are graphically presented by Stockinger (1985a) in his Figures 2.1 to 2.5 which offer a clear optical survey.

V.2 SOME MONTE CARLO RESULTS FOR GM-ESTIMATORS OF ARMA MODELS

The GM-estimators presented in Section II.2 and Section III.2 were applied to estimate the first-order AR parameter ϕ_1 ant the first-order MA parameter θ_1. ϕ_1 was estimated for the 8 types of simulated AR(1) processes which were described in Section V.1. θ_1 was estimated for 10 types of MA(1) processes with location $\mu = 0$. The numbers of observations and replications are the same as for the AR(1) processes, namely 100 and 50, respectively. The V_i's that cause additive outliers have a Gaussian mixture distribution $CND(\kappa, \sigma_3) = (1 - \kappa)\,\delta_0 + \kappa N(0, \sigma_3^2)$ (compare Section I.4), where σ_3 is a multiple of the variance of the outlier-free process. (For an MA(1) model $VAR\, X_i = \sigma^2(1 + \theta_1^2)$.) Abbreviations and values for θ_1, v, κ and σ_3 for processes with $CN(v, 1, 11) = (1 - v)\,N(0, 1) + vN(0, 121)$ – distributed innovations are given in Table V.7.

Not only AR(1) processes with t_4-distributed innovations without additive outliers were simulated, but also MA(1) processes which have the abbreviations MAIOTM 5 for $\theta_1 = -\cdot 5$ and MAIOTM 8 for $\theta_1 = -\cdot 8$.

Starting values for AR parameters were computed by the Yule-Walker equations as it was also described in Section V.1. Starting values for MA parameters were computed by a Newton-Raphson algorithm which was given by Wilson (1969) and which was also described by Box and Jenkins (1976). For this Newton-Raphson

Table V.7. Simulated MA(1) processes.

Abbreviation	θ_1	ν	κ	σ_3^2
MAGM 5	−·5	0·	0·	—
MAGM 8	−·8	0·	0·	—
MAIOCNM 5	−·5	·1	0·	—
MAIOCNM 8	−·8	·1	0·	—
MAAO 1 M	−·5	0·	·05	9 $VAR\, X_i$
MAAO 1 M 8	−·8	0·	·05	9 $VAR\, X_i$
MAAO 2 M 5	−·5	0·	· 05	100 $VAR\, X_i$
MAAO 2 M 8	−·8	0·	·05	100 $VAR\, X_i$

algorithm the order of the model to be estimated must be chosen, but nothing needs to be known about the parameters to be estimated. The algorithm of Wilson also gives a starting value for the innovations scale. For some time series the algorithm of Wilson does not give MA parameters which define an invertible MA process. In these cases the starting values for MA parameters were set equal to the true parameters.

In addition to the estimators described in Table V.2 Hampel-Krasker-Welsch type GM-estimators (Section III.2.1.) given in Table V.8 were used to estimate AR(1) and MA(1) models. Again the constants of the ψ-functions were chosen so that the asymptotic relative efficiency of all estimators, except the least squares estimator, of the first-order AR parameter is ·95. If ψ_1 is redescending, the algorithm for the GM-estimation of ARMA models (Section III.2.2) was not run first with a monotone ψ_1, like the IWLS algorithm (Section II.1.2), because the scale now is improved by the medmed estimator (III.32) but not by using an equation like (III.4) for pure AR models. The starting values for a certain type of estimation in general are the results of the preceding estimation except that an estimation with a redescending ψ_1-function which is based on the estimation of the same type with Huber's ψ_1 and the estimations MAH, SH and HKWH are based on MB to make the GM-estimators comparable.

Table V.8. Types of estimations.

Abbreviation	ψ_1
HKWH	$\psi_{H'}$ $c = 2{\cdot}5$
HKWHA	$\psi_{HA'}$ $a = 2{\cdot}7$, $b = 5{\cdot}4$, $d = 10{\cdot}$
HKWB	$\psi_{B'}$ $c = 9{\cdot}5$

Similar to Section V.1, the mean (MEAN) 100 times the mean square error (MSE), the sample relative efficiency (EFF) and the mean of the averages of the final weights $w_i^{(m)}$, $i = p + 1, \ldots, n$, (MAVW) for various estimates $\hat{\phi}_1$ or $\hat{\theta}_1$ for various types

of time series are given in Tables V.9 to V.17. When the algorithm to compute GM-estimates of ARMA parameters given in Section III.2.2. did not reach the required precision after 30 iterations, the computed estimate was excluded from further analysis. The algorithm failed in fairly few cases, namely a GM-estimator did not reach the required precision for about 1 percent of its applications. The reasons for the failure could be that the algorithm solves equations instead of a minimum problem and that the equations are nonlinear. Unfortunately it seems to be impossible to formulate a minimum problem.

Table V.9. Results of estimations (described in Section II.2 and Section III.2) of ϕ_1 for outlier-free processes.

ESTIMATOR	Simulated processes							
	ARGP 5				ARGP 8			
	MEAN	MSE	EFF	MAVW	MEAN	MSE	EFF	MAVW
LS	·461	1·23	1·00	1·00	·772	·575	1·00	1·00
MH	·459	1·35	·916	·958	·770	·655	·878	·958
MHA	·459	1·35	·914	·963	·770	·658	·875	·962
MB	·459	1·35	·911	·914	·770	·660	·872	·913
MAH	·460	1·36	·905	·962	·771	·657	·876	·965
MAHA	·460	1·36	·906	·966	·771	·656	·878	·969
MAB	·459	1·40	·883	·879	·771	·672	·856	·881
SH	·459	1·39	·890	·974	·769	·676	·851	·974
SHA	·459	1·38	·896	·979	·769	·690	·834	·980
SB	·459	1·38	·894	·938	·769	·696	·827	·938
HKWH	·463	1·32	·937	·991	·770	·657	·876	·991
HKWHA	·459	1·42	·866	·993	·770	·657	·876	·993
HKWB	·458	1·44	·857	·978	·770	·653	·881	·979

The estimates for the first-order AR parameter behave similarly to those listed in Section V.1., giving evidence that the algorithm to compute GM-estimates of ARMA parameters introduced in Section III.2.2 is useful in the AR(1) case. Of course, in this section a Hampel-Krasker-Welsch type estimator is included in addition to the estimators already treated in Section V.1. For *CN*-distributed innovations the Hampel-Krasker-Welsch type estimator tends to be better than the Mallows type estimator but worse than the Schweppe type estimator. For *t*-distributed innovations the Mallows and Schweppe estimators seem to be superior to the Hampel-Krasker-Welsch type estimator. The first-order AR parameter is best estimated by the Mallows type estimator for ARAOP 5 processes. In contrast, the Hampel-Krasker-Welsch type estimator gives on the average better parameter values than a Mallows type estimator for ARAOP 8 processes. The efficiency of Hampel-Krasker-Welsch type estimators is somewhere in the middle of the efficiencies of the two other

Table V.10. Results of estimations (described in Section II.2 and Section III.2) of ϕ_1 for processes with *CN*-distributed innovations.

ESTIMATOR	Simulated processes							
	ARIOCNP 5				ARIOCNP 8			
	MEAN	MSE	EFF	MAVW	MEAN	MSE	EFF	MAVW
LS	·470	1·11	1·00	1·00	·776	·754	1·00	1·00
MH	·478	·609	1·82	·929	·780	·400	1·88	·929
MHA	·481	·543	2·04	·924	·782	·334	2·25	·924
MB	·481	·535	2·07	·877	·782	·326	2·31	·878
MAH	·471	·934	1·19	·904	·778	·493	1·53	·902
MAHA	·471	·985	1·13	·896	·778	·511	1·48	·898
MAB	·469	1·02	1·09	·815	·776	·512	1·47	·813
SH	·471	·818	1·36	·936	·780	·425	1·77	·936
SHA	·469	·959	1·16	·928	·778	·495	1·52	·931
SB	·467	·955	1·16	·888	·777	·479	1·57	·892
HKWH	·468	·926	1·20	·967	·779	·450	1·67	·964
HKWHA	·467	·974	1·14	·969	·780	·466	1·62	·965
HKWB	·468	·994	1·12	·952	·779	·482	1·56	·947

Table V.11. Results of estimations (described in Section II.2 and Section III.2) of ϕ_1 for processes with *t*-distributed innovations.

ESTIMATOR	Simulated processes							
	ARIOPTP 5				ARIOTP 8			
	MEAN	MSE	EFF	MAVW	MEAN	MSE	EFF	MAVW
LS	·493	·862	1·00	1·00	·776	·499	1·00	1·00
MH	·495	·713	1·21	·932	·779	·421	1·19	·931
MHA	·494	·737	1·17	·932	·780	·415	1·20	·931
MB	·495	·727	1·19	·885	·780	·403	1·24	·884
MAH	·496	·805	1·07	·917	·780	·452	1·10	·914
MAHA	·499	·831	1·04	·917	·781	·459	1·09	·914
MAB	·500	·818	1·05	·831	·781	·459	1·09	·827
SH	·495	·780	1·11	·943	·779	·461	1·08	·942
SHA	·499	·790	1·09	·942	·778	·531	·940	·941
SB	·498	·786	1·10	·902	·781	·453	1·10	·901
HKWH	·495	·821	1·05	·971	·779	·502	·993	·969
HKWHA	·498	·859	1·00	·973	·779	·529	·943	·972
HKWB	·498	·847	1·02	·956	·779	·525	·951	·953

types of GM-estimators, and the Schweppe type estimator has the highest sample relative efficiency for the processes with additive outliers.

Analogous to Section V.1 some interesting results of the estimations are graphically presented in Figures 3.1 to 3.5 in Stockinger (1985a).

Table V.12. Results of estimations (described in Section II.2 and Section III.2) of ϕ_1 for processes with additive outliers.

ESTIMATOR	Simulated processes							
	ARAOP 5				ARAOP 8			
	MEAN	MSE	EFF	MAVW	MEAN	MSE	EFF	MAVW
LS	·209	9·55	1·00	1·00	·347	21·8	1·00	1·00
MH	·212	9·23	1·04	·927	·413	17·0	1·28	·909
MHA	·208	9·39	1·02	·924	·427	14·5	1·51	·895
MB	·208	9·39	1·02	·877	·484	13·8	1·58	·850
MAH	·295	5·57	1·71	·912	·560	7·38	2·96	·887
MAHA	·332	4·33	2·21	·904	·624	4·55	4·80	·878
MAB	·331	4·38	2·18	·827	·627	4·36	5·01	·802
SH	·282	6·33	1·51	·932	·565	7·26	3·01	·900
SHA	·332	4·66	2·05	·922	·661	3·36	6·50	·885
SB	·330	4·71	2·03	·885	·660	3·36	6·50	·851
HKWH	·287	6·36	1·50	·963	·572	7·16	3·05	·931
HKWHA	·315	5·80	1·65	·964	·649	4·31	5·06	·930
HKWB	·318	5·73	1·67	·948	·569	4·00	5·46	·917

Table V.13. Results of estimations (described in Section II.2 and Section III.2) of θ_1 for outlier-free processes.

ESTIMATOR	Simulated processes							
	MAGM 5				MAGM 8			
	MEAN	MSE	EFF	MAVW	MEAN	MSE	EFF	MAVW
LS	−·484	1·18	1·00	1·00	−·799	·522	1·00	1·00
MH	−·482	1·31	·901	·958	−·797	·607	·860	·958
MHA	−·483	1·32	·899	·963	−·797	·609	·858	·963
MB	−·483	1·32	·897	·914	−·798	·612	·853	·913
MAH	−·486	1·29	·919	·963	−·798	·581	·899	·963
MAHA	−·487	1·29	·916	·967	−·799	·580	·901	·967
MAB	−·486	1·34	·887	·882	−·798	·605	·864	·882
SH	−·483	1·31	·902	·974	−·798	·593	·881	·974
SHA	−·485	1·33	·892	·980	−·798	·585	·893	·980
SB	−·485	1·35	·878	·938	−·798	·595	·879	·938
HKWH	−·484	1·34	·882	·991	−·798	·583	·896	·991
HKWHA	−·484	1·34	·883	·993	−·799	·568	·919	·993
HKWB	−·484	1·35	·879	·978	−·798	·573	·911	·978

The estimated first-order MA parameters differ only very slightly for outlier-free processes and processes with innovation outliers. For processes with innovation outliers the least squares estimator in some cases has, which shows up as an unexpected phenomenon, a smaller mean square error than GM-estimators; the M-

Table V.14. Results of estimations (described in Section II.2 and Section III.2) of θ_1 for processes with *CN*-distributed innovations.

ESTIMATOR	Simulated processes							
	MAIOCNM 5				MAIOCNM 8			
	MEAN	MSE	EFF	MAVW	MEAN	MSE	EFF	MAVW
LS	−·494	·804	1·00	1·00	−·800	·455	1·00	1·00
MH	−·489	·606	1·33	·931	−·796	·340	1·34	·930
MHA	−·490	·643	1·25	·927	−·795	·365	1·25	·926
MB	−·490	·626	1·28	·880	−·795	·362	1·26	·879
MAH	−·489	·921	·872	·908	−·792	·391	1·16	·907
MAHA	−·487	1·20	·671	·901	−·785	·540	·844	·904
MAB	−·487	1·25	·641	·822	−·784	·531	·857	·819
SH	−·486	·853	·942	·938	−·792	·386	1·18	·940
SHA	−·475	1·02	·790	·930	−·787	·487	·935	·934
SB	−·479	1·08	·746	·892	−·787	·493	·923	·893
HKWH	−·477	·947	·848	·968	−·790	·372	1·22	·968
HKWHA	−·482	1·17	·686	·969	−·789	·393	1·16	·969
HKWB	−·481	1·22	·660	·952	−·788	·397	1·15	·951

Table V.15. Results of estimations (described in Section II.2 and Section III.2) of θ_1 for processes with *t*-distributed innovations.

ESTIMATOR	Simulated processes							
	MAIOTM 5				MAIOTM 8			
	MEAN	MSE	EFF	MAVW	MEAN	MSE	EFF	MAVW
LS	−·509	·887	1·00	1·00	−·811	·350	1·00	1·00
MH	−·510	·766	1·16	·930	−·809	·309	1·13	·930
MHA	−·508	·858	1·03	·930	−·807	·353	·990	·930
MB	−·508	·823	1·08	·883	−·807	·354	·988	·883
MAH	−·509	·740	1·20	·914	−·807	·320	1·09	·914
MAHA	−·512	·815	1·09	·913	−·808	·375	·932	·915
MAB	−·512	·792	1·12	·828	−·809	·363	·965	·824
SH	−·508	·693	1·28	·942	−·809	·304	1·15	·943
SHA	−·513	·723	1·23	·940	−·809	·393	·890	·942
SB	−·513	·732	1·21	·900	−·809	·388	·902	·901
HKWH	−·508	·688	1·29	·971	−·810	·303	1·15	·970
HKWHA	−·507	·727	1·22	·973	−·808	·375	·933	·973
HKWB	−·508	·752	1·18	·955	−·807	·390	·898	·954

estimators have higher sample relative efficiencies than the least squares estimators except for MAIOTM 8 processes. The quality of GM-estimators expressed in well estimated parameters and high efficiencies, is revealed for processes with additive outliers, especially for MAAO 2 processes.

Table V.16. Results of estimations (described in Section II.2 and Section III.2) of θ_1 for processes with additive outliers.

ESTIMATOR	Simulated processes							
	MAAO 1 M 5				MAAO 1 M 8			
	MEAN	MSE	EFF	MAVW	MEAN	MSE	EFF	MAVW
LS	−·312	4·43	1·00	1·00	−·431	14·2	1·00	1·00
MH	−·322	3·97	1·11	·949	−·458	12·2	1·16	·947
MHA	−·324	3·92	1·13	·951	−·462	11·9	1·19	·948
MB	−·326	3·85	1·15	·902	−·465	11·8	1·20	·900
MAH	−·367	2·66	1·67	·942	−·538	7·29	1·95	·931
MAHA	−·383	2·31	1·92	·940	−·578	5·48	2·59	·920
MAB	−·388	2·27	1·95	·857	−·582	5·29	2·68	·840
SH	−·361	2·85	1·55	·958	−·531	7·73	1·83	·950
SHA	−·394	2·32	1·91	·956	−·587	5·17	2·74	·940
SB	−·386	2·34	1·89	·917	−·585	5·30	2·67	·902
HKWH	−·365	2·83	1·57	·978	−·540	7·31	1·94	·968
HKWHA	−·381	2·54	1·74	·978	−·580	5·61	2·53	·965
HKWB	−·387	2·45	1·81	·963	−·581	5·53	2·56	·949

Table V.17. Results of estimations (described in Section II.2 and Section III.2) of θ_1 for processes with additive outliers.

ESTIMATOR	Simulated processes							
	MAAO 2 M 5				MAAO 2 M 5			
	MEAN	MSE	EFF	MAWV	MEAN	MSE	EFF	MAVW
LS	−·073	18·6	1·00	1·00	−·097	49·8	1·00	1·00
MH	−·085	17·5	1·06	·938	−·121	46·4	1·07	·936
MHA	−·096	16·9	1·10	·936	−·144	43·9	1·14	·935
MB	−·096	16·9	1·10	·889	−·147	43·6	1·14	·887
MAH	−·256	6·54	2·85	·905	−·361	19·7	2·53	·892
MAHA	−·386	2·19	8·53	·868	−·569	5·88	8·48	·805
MAB	−·389	2·17	8·60	·796	−·568	5·94	8·40	·740
SH	−·245	7·10	2·63	·927	−·352	20·5	2·43	·914
SHA	−·397	2·26	8·25	·887	−·567	5·98	8·33	·833
SB	−·397	2·31	8·08	·852	−·565	6·13	8·13	·802
HKWH	−·252	6·84	2·73	·948	−·364	19·7	2·54	·932
HKWHA	−·346	3·73	5·00	·932	−·527	8·16	6·11	·892
HKWB	−·356	3·34	5·58	·916	−·524	8·36	5·96	·879

Some interesting results listed in the Tables V.13 to V.17 are graphically presented in the Figures 3·6 to 3·13 in Stockinger (1985a).

Comparison of the Results in Section V.1 and Section V.2

The means of the first-order AR parameter estimated by the method of Martin (1980), which was described in Section III.1, are in general very similar to those estimated by the method presented in Section III.2. But the means for the estimated parameters for ARAOP 8 processes lie closer to the true parameter in Section V.2 than in Section V.1. The sample relative efficiencies for parameters estimated from processes with t-distributed innovations or from processes with additive outliers tend to be larger for the method of Section III.2 than for the method of Section III.1. The sample relative efficiencies for parameters estimated from processes with CN-distributed innovations, in contrast, tend to be less for the method of Section III.2 than for the method of Section III.1.

V.3 TOPICS FOR FURTHER RESEARCH

The algorithms for a GM-estimation of ARMA parameters presented in Chapter III were successfully applied in a Monte Carlo study (compare Section V.1 and Section V.2). The estimation of the first-order AR parameter and the first-order MA parameter was investigated by Monte Carlo because it is difficult to compare higher order models. Nevertheless it would be interesting to compare estimated higher order models. Computer programs (Stockinger, 1985b) already allow the GM-estimation of ARMA (p, q) models with locations and with arbitrary orders p and q. These computer programs offer good hope for interesting research also for data from practical problems. In fact, GM-estimation of AR models was applied for the detection of outliers in arrhythmic pressure pulses (Stockinger, 1984; Stockinger, Pfeiffer and Dutter, 1984).

The methods for the GM-estimation of ARMA models presented in Chapter II and Chapter III could be improved by incorporating backforecasting routines (Box and Jenkins, 1976).

ARIMA model parameter estimates may be obtained similar as ARMA model parameter estimates. One computational method is to express the nonstationary, generalized autoregressive operator $\Phi(B) = \phi(B)(1 - B)^d$ in closed form as autoregressive operator of order $(p + d)$. However, it is not entirely clear to us in which way instationarity affects parameter estimates. Another conventional method for dealing with ARIMA models is to take appropriate differences to get an ARMA model. If the time series contains outliers, however, this procedure becomes less attractive. The reason is that differencing increases the number of outliers. For example, first differences produce two outliers for every isolated outlier in the original series. If the fraction of outliers is very small, we may well get away with taking differences and then applying a robust fitting procedure. Such an approach, however, becomes unattractive as the fraction of outliers increases and alternative robust methods are then needed for dealing with ARIMA models.

A proof of the robustness of autoregressive-errors M-estimates (Section II.2.1) for the location of ARMA (p, q) models is outstanding.

The exact computation of the asymptotic Cramer-Rao lower bound of prewhitening-based location M-estimates (Section II.2.3) at additive outlier models did not yet succeed, but it is hoped that these estimates provide high absolute efficiencies.

The key to Masreliez's filter theorem in Section IV.1.2 is the assumption that the state-prediction density is a Gaussian density with appropriate mean and covariance. This assumption will rarely, if ever, be satisfied exactly. Martin (1979c), however, presents a continuity theorem which lends support to the intuitive notion that the conditional density in question will nearly be Gaussian in a strong sense when the additive noise is nearly Gaussian in a comparatively weak sense. Note however, that here a difficult problem area is presented in which clean theoretical results seem to be unlikely. It is not yet entirely clear whether or not the simplifications (IV.63) and (IV.64) of Masreliez's filter are good ideas.

Other methods for minimization of the loss function $L(\alpha)$ (IV.45) could be established, e.g. direct minimization of $L(\alpha)$ could be tried. Another possibility is to extend the approximate M-estimates from one-sided filter based estimates to two-sided filter ("smoother") based estimates.

Since time series analysis based on a wrong model is worthless, it is very important to identify the correct model. Thus further investigation of robust model selection which often uses the autocorrelation function and the partial autocorrelation function, seems to be valuable. Careful study of order-selection rules, e.g., of those described in Section III.4.1, is clearly needed (compare Shibata, 1976). It is not yet known how many iterations of the identification procedure described in Section III.4.2 are sufficient in general.

It is not entirely clear how the robustified Fox test (Section III.3.3) could be applied in practice. Methods of determining the outlier type if more general models than AR models are used, are urgently called for.

Robust estimates of parameters for time series models could help to detect certain failures of time series models. One possible model failure would declare a "normal" observation to be atypical. Thus we are faced with a model failure if it is known that a certain observation is not an outlier, but in the sense of the fitted time series model it is an outlier. Methods to diagnose the possible inadequacy of the model contemplated would be very important.

Some good methods to detect outliers in time series (e.g. by residual analysis) by robust parameter estimates should be found out. A possibility to detect outliers would be to compare real data with simulated data in an appropriate manner.

Since missing data which are a frequently emerging problem in time series, can be regarded as a special version of outliers, outlier-handling techniques could be modified to behave very well on missing data situations.

Once an appropriate, robustly estimated time series model is found, it should be

relatively easy to forecast future values and/or to replace outliers by reasonable values. Additionally, backforecasting routines would become more reliable.

Box and Tiao (1975) introduced an "intervention-analysis" technique for time series model fitting and analysis if the starting time of a potential change in model structure is known. In situations where intervention analysis is an appropriate tool, robust model fitting procedures may play a useful role which remains to be investigated. The residuals from a robust filter or smoother, for example, may provide guidance for selecting the form of the potential change.

With robust time series model parameter estimates available it is possible to estimate spectral densities robustly. Large progress in this direction has already been obtained (Kleiner, Martin and Thomson, 1979; Martin and Thomson, 1982; Martin, 1983; Martin, 1984), but detailed investigations could still be attempted.

In this chapter and in the foregoing chapters the GM-estimation and techniques of robust filtering and smoothing were treated in order to bound the influence of outliers. Of course, other possibilities of estimation exist. Estimation based on the autocovariance of the residuals was investigated by Bustos and Yohai (1983), and the asymptotic normality and consistency of these estimators are proved in Bustos, Fraiman and Yohai (1984). One-step maximum likelihood type estimators were investigated by Lee (1981), Lee and Martin (1982) and Lee and Martin (1982b).

Much more theoretical robustness properties, thorough studies and comparisons of various methods of robust estimation are required before firm conclusions may by drawn, although some Monte Carlo studies have been in the expected direction.

Obviously more complex outlier-generating models than those given in Section I.4 will be more appropriate for many time series occurring in practice. Things are complicated enough, however, with just the innovations outlier model and additive effects outlier model.

Other time series models than ARIMA models are possible and perhaps sometimes more adequate. Some examples of other models may be found in Hampel et al. (1982). Most work in literature, however, concentrates on ARIMA models.