DEPARTMENT OF MATHEMATICS AND STATISTICS

ROBUST INFERENCE FOR LOG-LINEAR COUNT TIME SERIES MODELS

DOCTOR OF PHILOSOPHY DISSERTATION

STELLA KITROMILIDOU

2015

DEPARTMENT OF MATHEMATICS AND STATISTICS

ROBUST INFERENCE FOR LOG-LINEAR COUNT TIME SERIES MODELS

STELLA KITROMILIDOU

A Dissertation Submitted to the University of Cyprus in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

OCTOBER 2015

To my Athena

Abstract

The aim of this contribution is the investigation of robust estimation methods and their development of statistical inference for count time series models. Sequences of time dependent counts appear in applications in many scientific fields. However, the sensitivity that classical statistical methods reveal when exposed to extreme events establishes the need for a more robust approach.

Our focal point is the study of a log-linear model for counts based on the Poisson distribution and a feedback structure. Firstly, we consider several robust estimation methods and examine their properties and behavior under a variety of intervention type effects, for the case of a simpler model which does not include the feedback mechanism. Namely, we consider level shifts, transient shifts and additive outliers. The Mallows' quasi likelihood estimation method is the one that deserves more focus since it is the estimation method that behaves the most adequately among all other estimation methods.

We advance the study of the Mallows' quasi likelihood estimator in the more complicated case of the log-linear model with feedback. The asymptotic behavior of the proposed estimation method is examined by employing the so called perturbation technique. We find that the robustly weighted Mallows' quasi likelihood estimator is asymptotically normally distributed and a robust score type testing procedure is proposed to examine whether the model can be deduced to a model without feedback.

Additionally, we discuss ways of approximating the autocovariance function of count time series by considering orthogonal polynomial expansions.

Περίληψη

Στόχος αυτής της εργασίας είναι η έρευνα εύρωστων μεθόδων εκτίμησης και η ανάπτυξη στατιστικής συμπερασματολογίας για μοντέλα χρονοσειρών τα οποία λαμβάνουν ακέραιες τιμές. Ακολουθίες χρονικά εξαρτημένων ακέραιων μετρήσεων εμφανίζονται σε πολλά επιστημονικά πεδία. Ωστόσο, η ευαισθησία την οποία παρουσιάζουν οι κλασσικές στατιστικές μέθοδοι όταν εκτίθενται σε ακραία φαινόμενα, δηλώνει την ανάγκη για μια πιο εύρωστη προσέγγιση.

Εστιάζουμε στην μελέτη ενός λογαριθμικού γραμμικού μοντέλου για ακέραιες μετρήσεις βασισμένο στην κατανομή Poisson και μια δομή επανατροφοδότησης. Αρχικά, θεωρούμε διάφορες μεθόδους εύρωστης εκτίμησης και εξετάζουμε τις ιδιότητες και τη συμπεριφορά τους κάτω από διάφορες μορφές παρεμβάσεων, για την περίπτωση ενός απλούστερου μοντέλου που δεν συμπεριλαμβάνει το μηχανισμό επανατροφοδότησης. Ονομαστικά, θεωρούμε μετατοπίσεις επιπέδου, παροδικές μετατοπίσεις και πρόσθετες ακραίες τιμές. Η εκτιμήτρια ημιπιθανοφάνειας του Mallows είναι η μέθοδος η οποία αξίζει περισσότερη προσοχή καθότι είναι η μέθοδος εκτίμησης η οποία συμπεριφέρεται πιο κατάλληλα ανάμεσα σε όλες τις υπόλοιπες μεθόδους εκτίμησης.

Εμβαθύνουμε στη μελέτη της εκτιμήτριας ημιπιθανοφάνειας του Mallows στην πιο περίπλοκη περίπτωση του λογαριθμικού γραμμικού μοντέλου με επανατροφοδότηση. Η ασυμπτωτική συμπεριφορά της προτεινόμενης μεθόδου εκτίμησης εξετάζεται χρησιμοποιώντας τη λεγόμενη τεχνική της διαταραχής. Βρίσκουμε ότι η εκτιμήτρια ημιπιθανοφάνειας του Mallows, εύρωστα σταθμισμένη, είναι ασυμπτωτικά κανονικά κατανεμημένη και προτείνουμε μια εύρωστη διαδικασία ελέγχου τύπου score για να εξεταστεί κατά πόσο το μοντέλο μπορεί να περιορισθεί σε ένα μοντέλο χωρίς επανατροφοδότηση.

Επιπρόσθετα, συζητούμε τρόπους προσέγγισης της συνάρτησης αυτοδιακύμανσης χρονοσειρών που λαμβάνουν ακέραιες τιμές χρησιμοποιώντας επεκτάσεις ορθογώνιων πολυωνύμων.

Acknowledgments

I would like, first and foremost, to thank my supervisor, Professor Konstantinos Fokianos, for his constant support, trust and encouragement throughout the years of my postgraduate studies. He is always patient and willing to help and his guidance, motivational comments and suggestions have been most contributory to the completion of this work. Words are truly very poor to express my appreciation, gratitude and respect to Professor Fokianos.

I also like to thank the members of the examining committee, Professors Roland Fried, Tasos Christofides, Associate Professor Ioannis Kasparis and Assistant Professor Anastassia Baxevani for their useful and constructive comments that improved this thesis.

Additionally, I would like to express my gratitude to Tobias Liboschik for useful discussions.

Moreover, I wish to acknowledge financial support from the Cyprus Research Promotion Foundation TEXNOLOGIA/THEPIS/0609(BE)/02 during my Ph.D studies.

Furthermore, I acknowledge the project "eMammoth - Compute and Store on Grids and Clouds infrastructure"(ANABATHMISI/06609/09), which is co-funded by the Republic of Cyprus and the European Regional Development Fund of the EU. Many thanks to Ms. Andoenna Balla for her help with the computing program.

Finally, I want to express my deepest gratitude to my husband, for his continuous encouragement and unconditional love, and to my parents for all their help and support over the years. Without the help of my family this achievement would not have been possible.

Contents

Abstract **i**

Περίληψη **ii**

Acknowledgments **iii**

List of Figures **vii**

List of Tables **x**

1 Introduction **1**

2 Literature Review **7**

2.1 Linear Poisson Autoregressive Model 8

2.2 Log-Linear Poisson Autoregressive Model 12

2.3 Interventions in Count Time Series 15

2.4 Robust Estimation for Independent Data 21

2.4.1 The Conditionally Unbiased Bounded-Influence Estimator (CUBIF) 24

2.4.2 The Mallows' Quasi Likelihood Estimator (MQLE) 25

2.4.3 L_q - Estimator 26

2.4.4 Weighted Maximum Likelihood Estimates for Poisson Regression . . 27

3 Log-Linear Poisson Model Without Feedback **29**

3.1 Intervention Effects 30

3.2 Methods of Estimation 32

3.2.1 Maximum Likelihood Estimator (MLE) 33

3.2.2 Conditionally Unbiased Bounded–Influence Estimator (CUBIF) . . . 33
3.2.3 Mallows' Quasi-Likelihood Estimator (MQLE) . . . 34
3.3 Empirical Results . . . 36
3.3.1 Level Shift and Transient Shift . . . 37
3.3.2 Additive Outliers . . . 42
3.3.3 Choice of the Tuning Constant . . . 50
3.3.4 Summary . . . 52
3.4 Real Data Examples . . . 52
3.4.1 Polio Data . . . 52
3.4.2 Hyde Park Data . . . 54

4 Log-Linear Poisson Model With Feedback **58**
4.1 Mallows' Quasi Likelihood Estimation . . . 60
4.1.1 Estimation . . . 63
4.1.2 A Simulated Example . . . 67
4.2 Robust Score Test . . . 80
4.2.1 Results . . . 80
4.2.2 Simulation Results for the Robust Score Test . . . 81
4.3 Real Data Examples . . . 84
4.3.1 Measles Data . . . 84
4.3.2 Firearm Homicides Data . . . 85
4.4 Appendix . . . 87

5 Orthogonal Polynomial Approximation **113**
5.1 Orthogonal Polynomials . . . 114
5.1.1 Hermite Polynomials . . . 114
5.1.2 Laguerre Polynomials . . . 119
5.1.3 Poisson-Charlier Polynomials . . . 122
5.2 Representation by ARMA . . . 127
5.2.1 Hermite Expansion . . . 129
5.2.2 Log-Normal Representation . . . 130

5.2.3 Poisson-Charlier Expansion . . . 131
5.3 Discussion . . . 132

6 Discussion and Further Research 134
6.1 Discussion . . . 134
6.2 Further Research . . . 136

A R Code for Chapter 3 140

B R Code for Chapter 4 146

Bibliography 168

List of Figures

1.1 Time series plot of the polio data. The data consist of the monthly number of poliomyelitis cases in USA from 1970 to 1983. 4

2.1 Example of a time series with various forms of interventions, based on the linear Poisson model. 22

2.2 Example of a time series with various forms of interventions, based on the log-linear Poisson model. 23

3.1 MSE for the case of a Level Shift at $\tau = n/4$ for b of the model with $q = 1$ with (a): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = 0.3$, (b): $\theta = (d, b) = (0.2, -0.5)$, $\zeta = 0.3$ (c): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = -0.3$. 37

3.2 Bias for the case of a Level Shift at $\tau = n/4$ for b of the model with $q = 1$ with (a): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = 0.3$, (b): $\theta = (d, b) = (0.2, -0.5)$, $\zeta = 0.3$ (c): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = -0.3$. 38

3.3 MSE for the case of a Transient Shift at $\tau = n/4$ for b of the model with $q = 1$ with (a): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = 0.3$, (b): $\theta = (d, b) = (0.2, -0.5)$, $\zeta = 0.3$ (c): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = -0.3$ and (d): $\theta = (d, b) = (0.2, -0.5)$, $\zeta = -0.3$. . 39

3.4 Bias for the case of a Transient Shift at $\tau = n/4$ for b of the model with $q = 1$ with (a): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = 0.3$, (b): $\theta = (d, b) = (0.2, -0.5)$, $\zeta = 0.3$ (c): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = -0.3$ and (d): $\theta = (d, b) = (0.2, -0.5)$, $\zeta = -0.3$. . 40

3.5 QQ-plots for $\hat{b}$ for the case of a Level Shift with $\theta = (d, b) = (0.2, 0.5)$, $\zeta = 0.3$ and $c = 3$. 41

3.6 QQ-plots for $\hat{b}$ for the case of a Transient Shift with $\theta = (d, b) = (0.2, 0.5)$, $\zeta = 0.3$ and $c = 3$. 41

3.7 MSE for the case of a single outlier for b_1 of the model with $q = 1$ for the first two plots and $q = 2$ for the last two plots. 43
3.8 MSE for the case of a patch of outliers for the model with $q = 1$ and $\zeta = 10$. . . 44
3.9 MSE for the case of a patch of k outliers for the model with $q = 3$ for b_1. 45
3.10 MSE for the case of a patch of k outliers for the model with $q = 3$ for b_1. 46
3.11 Boxplots for $\hat{b}_1$ for the case of the second order model with 10 consecutive outliers of size 10 and for $c = 3$. 47
3.12 QQ-plots for $\hat{b}_1$ for the case of the second order model with 10 consecutive outliers of size 10 and for $c = 3$. 47
3.13 MSE for the case of k isolated outliers for the model with $q = 1$ for b_1. 48
3.14 MSE for the case of k isolated outliers for the model with $q = 3$ for b_1 and $\zeta = 10$. 49
3.15 MSE for the case of k isolated outliers for the model with $q = 3$ for b_1 and $\zeta = 30$. 50
3.16 Efficiency and choice of c. 51
3.17 Estimated Mean Square Error (MSE) and Mean Absolute Error (MAE) of the predicted values for the Polio data. 54
3.18 Hyde Park Time Series Plot. 56
3.19 Estimated Mean Square Error (MSE) and Mean Absolute Error (MAE) of the predicted values for the Hyde Park data. 57

4.1 MSE of $\hat{a}_{\mathrm{MQLE}}$ as a function of the tuning constant c when using for estimation the Huber function (4.6). Data are generated by model (4.1) with $(d, a, b) = (0.2, 0.3, 0.5)$ and with a patch of outliers–see (4.3)–of size $\zeta = 20$. Results are based on 1000 simulations. 72
4.2 MSE values of $\hat{b}_{\mathrm{MQLE}}$ as a function of the tuning constant c when using for estimation the Huber function (4.6). Data are generated by model (4.1) with $(d, a, b) = (0.2, 0.3, 0.5)$ and with a patch of outliers–see (4.3)–of size $\zeta = 20$. Results are based on 1000 simulations. 73
4.3 QQ-plots for the various estimating methods for $\hat{a}_{\mathrm{MQLE}}$. A patch of 15 consecutive outliers of size $\zeta = 20$–see (4.3)–is considered and the Huber function (4.6) is used for the calculation of MQLE with tuning constant equal to $c = 1.571$. Data are generated by model (4.1) with $(d, a, b) = (0.2, 0.3, 0.5)$. Results are based on 1000 simulations. 74

4.4 MSE of $\hat{a}_{\text{MQLE}}$ as a function of the tuning constant c when using for estimation the Huber function (4.6). Data are generated by model (4.1) with $\boldsymbol{\theta} = (d, a, b) = (0.2, 0.3, 0.65)$ and with a patch of outliers–see (4.3)–of size $\zeta = 20$. Results are based on 1000 simulations. 75
4.5 MSE of $\hat{b}_{\text{MQLE}}$ as a function of the tuning constant c when using for estimation the Huber function (4.6). Data are generated by model (4.1) with $\boldsymbol{\theta} = (d, a, b) = (0.2, 0.3, 0.65)$ and with a patch of outliers–see (4.3)–of size $\zeta = 20$. Results are based on 1000 simulations. 76
4.6 MSE of $\hat{a}_{\text{MQLE}}$ (right column) and $\hat{b}_{\text{MQLE}}$ (left column) as a function of the tuning constant c when using for estimation the Huber function (4.6). Data are generated by model (4.4) with $(d, a, b) = (0.2, 0.3, 0.5)$ and with either a LS of size $\zeta = 0.2$ or a TS of size $\zeta = 1$. Results are based on 1000 simulations. 77
4.7 MSE of $\hat{a}_{\text{MQLE}}$ (right column) and $\hat{b}_{\text{MQLE}}$ (left column) as a function of the tuning constant c when using for estimation the Huber function (4.6). Data are generated by model (4.4) with $(d, a, b) = (0.2, 0.3, 0.65)$ and with either a LS of size $\zeta = 0.2$ or a TS of size $\zeta = 1$. Results are based on 1000 simulations. 78
4.8 Efficiency and choice of c. 79
4.9 Power of the test statistic (4.8) for the cases of a patch of outliers and isolated outliers. Robust weighting is considered (MVE and MCD). The tuning constant is $c = 1.571$ and the nominal significance level is $\alpha = 0.05$. 83
4.10 Weekly number of measles infections reported. 85
4.11 Weekly number of firearm homicides reported at the Salt River state mortuary in Cape Town during 1986-1999. 86

List of Tables

3.1 Table of the AIC of different fitted models for the polio data 53
3.2 Minimum MSE and MAE of the estimators and the corresponding value of the tuning constant c for the prediction of the last ten observations of the polio data. 54
3.3 Estimates (standard errors) of the parameters of model (3.9) based on the proposed robust estimation methods. The tuning constant c has been chosen according to Table 3.2. 55
3.4 Table of the AIC of different fitted models for the Hyde Park data 55
3.5 Minimum MSE and MAE of the estimators and the corresponding value of the tuning constant c for the prediction of the last ten observations of the Hyde Park data. 56
3.6 Estimates (standard errors) of the parameters of model (3.10) based on the proposed robust estimation methods. The tuning constant c has been chosen according to the MSE in Table 3.5. 57
4.1 Maximum likelihood estimation of the true value $\boldsymbol{\theta} = (0.2, 0.3, 0.5)$ for the log-linear model (4.1), under increasing number of additive outliers, a level shift and a transient shift intervention. Results are based on 1000 simulations of length 500. Standard errors are reported in parentheses. 62
4.2 Empirical size of the test for the case of a patch of outliers and the case of isolated outliers. 82
4.3 Estimates and standard errors (in parentheses) of the parameters of model (4.1) when applied to the measles infection time series. Fitting is done by employing (4.6) with $c = 3$. 85

4.4 Estimates of the parameters of model (4.1) for the firearm homicides time series, $c = 2.429$. . . . 87

Chapter 1

Introduction

In the last decades, there has been an increasing interest in the concept of robustness for parametric model estimation, especially when the data indicate deviations from the postulated assumptions. Although several studies on robust estimation have been developed for independent data, analogous efforts for dependent data have been delayed. In particular, robust methods for time series of counts have been studied only recently.

Time series of counts of events arise in plentiful contexts in all areas of science, such as medicine, economics, social sciences etc. Some examples may consist of the monthly incidences of some disease, the daily number of transactions of some stock, the yearly number of fatalities in road accidents or the monthly number of claims to an insurance agency. This plethora of applications has led to a growing development of statistical methods for analyzing count time series data.

Our interest focuses on robust inference for count time series in the presence of various types of interventions, and in particular, in the presence of outliers. The aim of this dissertation is to extend the application of certain robust estimation procedures that originally were designed for independent data to the time series data.

Classical statistical procedures heavily rely on a number of prior assumptions, which reflect what the statistician suspects or already knows about the data. The most commonly assumed notions are those of normality and independence. These assumptions are not, in fact, met in real life and classical techniques may be dreadfully vulnerable to minor distortions from the model assumptions. Consequently, a new statistical concept that combined both insensitivity to deviations and credible results was essential.

This new approach is robust statistics and the foundations were introduced in the innovative contributions of Tukey (1960), Huber (1964) and Hampel (1968). Tukey had ever been skeptical about the assumption of normality and in Tukey (1960) he demonstrated the inefficiency and nonrobustness of the sample mean and sample standard deviation under contaminated Normal distributions. Huber (1964) considers a contaminated model and suggests a minimax approach that aims at constructing an estimate that minimizes the maximum asymptotic variance over all neighborhoods of the model. Hampel (1968) on the contrary, considers an exact model and suggests an infinitesimal approach in which the chosen estimator has the minimum asymptotic variance over an infinitesimal neighborhood of the model. In the years that have followed, the interest in the concept of robustness has grown substantially and the field of robust statistics has gained great importance. It is now well acknowledged that contemporary robust statistical methods are in need.

One of the most important aims of the robust procedures is to identify possible outlying observations that are often encountered in the data and decrease their impact on estimation and testing. Outliers, are aberrant observations that do not fit the structure recommended by the bulk of the data. These "bad" observations appear quite frequently in real data sets and are usually a result of gross errors or otherwise false measurements. For example, copying, reading or transmission errors and rare phenomena as is an earthquake (Hampel et al. (1986)). Hampel et al. (1986) suggest that there exists a chance of $1\% - 10\%$ of gross errors in routine data whereas high-quality data do not contain any gross errors. If a potential outlier can be ruled out as a gross error, then it is a true outlier - a correct measurement, but deviant.

Robust methods are characterized by their ability to fit well to the majority of the data. This means that if the data do not contain any outliers, then robust procedures behave approximately as well as the classical procedures. If the data contain a small percentage of outliers, then the robust procedures behave approximately as well as the classical procedures would behave to the clean data.

The outlier problem has been an important aspect in data analysis long before the concept of robustness. The typical treatment was to remove these extreme values and then apply an appropriate stochastic model. However, treating outliers in this manner may conceal very important information. As an example, we refer to the discovery of the ozone hole (Maronna et al. (2006, p. 2)):

> The discovery of the ozone hole was announced in 1985 by a British team working on the ground with ”conventional” instruments and examining its observations in detail. Only later, after reexamining the data transmitted by the TOMS instrument on NASA’s Nimbus 7 satellite, was it found that the hole had been forming for several years. Why had nobody noticed it? The reason was simple: the systems processing the TOMS data, designed in accordance with predictions derived from models, which in turn were established on the basis of what was thought to be ”reasonable”, had rejected the very (”excessively”) low values observed above the Antarctic during the Southern spring. As far as the program was concerned, there must have been an operating defect in the instrument.

Therefore, it is apparent that outliers are of the most valuable observations and should not be discarded but rather they should be interpreted.

Dependent data are not of course an exception and discrepant observations arise in the time series framework as well. Outliers in time series may result either from gross errors or from occasional exogenous interventions (Tsay (1986)). Examples of such exogenous interventions may consist of a financial crisis, an oil crisis or the outbreak of some disease. Fox (1972) was the first to suggest outlier types for time series. He assumed an autoregressive model with Gaussian noise and proposed two types of outliers which he named type I and type II outliers. Type I outliers affect only a single observation and type II outliers also affect succeeding observations. These two types were renamed later on to Additive Outliers (AO) and Innovation Outliers (IO). Tsay (1988) considered another two forms of interventions; namely level shifts (LS) which he classifies further to permanent level change (LC) and transient level change (TC), and variance change (VC). A permanent level change changes the level of the series from the time point that the intervention occurs onwards, whereas a transient change is not permanent but decays exponentially. Finally, a variance change type of outliers affects the variance of the time series through the addition of a random variable. In the following years, permanent level change is referred to as a level shift (LS) and a transient level change is referred to as a transient shift (TS).

For an example of a time series with intervention effects, see Figure 1.1 which exhibits the polio data time series. This time series consists of monthly number of poliomyelitis cases reported in USA between the years 1970 and 1983 and it contains both additive outliers and level shifts, as we will see in detail in Section 3.4.1.

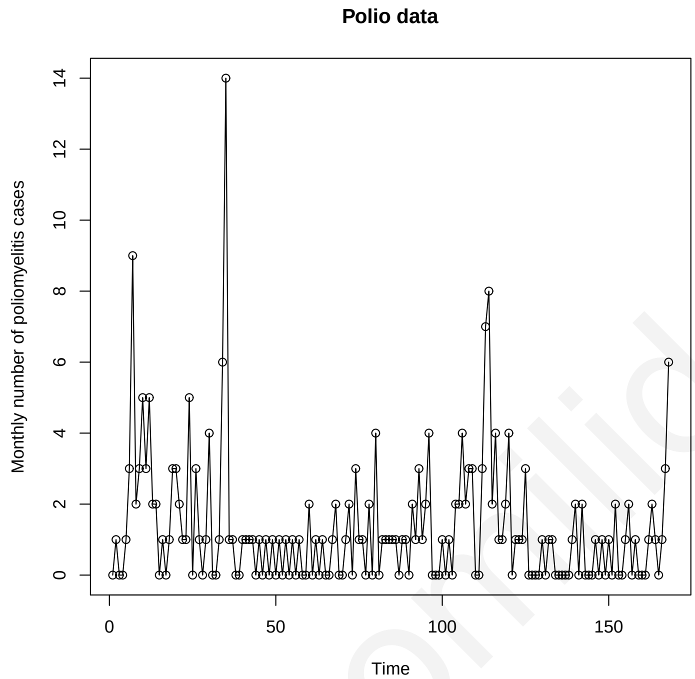

Figure 1.1: Time series plot of the polio data. The data consist of the monthly number of poliomyelitis cases in USA from 1970 to 1983.

Interventions in time series data can cause severe consequences on model identification, estimation and forecasting, which depend on the number, type, size and position of the interventions. The detection of outliers is essential. However, methods for detecting outliers in independent data cannot be used directly for time series. Temporal dependence cannot be ignored since estimates for variance are biased when autocorrelations are not taken into account and additionally, some outliers have effects on more than one observation (Deutsch et al. (1990)). Outlier detection procedures are mostly based on the framework of ARMA models. This is because ARMA models are very popular and widely applicable due to Wold's decomposition and additionally are extensively studied and their properties are well understood, see Brockwell and Davis (1991) for instance. Procedures for the identification and estimation of time series outliers have been introduced by several authors. We note, among others, the works by Fox (1972) who employ maximum likelihood to study and estimate outlier effects, Box and Tiao (1975) who propose an intervention analysis technique, Abraham and Box (1979) who propose a Bayesian method for the estimation of both the magnitude and location of outliers and Martin (1983) who developed a robust procedure based on the spectral domain. In a different approach, Chan (1992) considers sample autocorrelation and partial autocorrelation as tools for outlier detection. His analysis has shown that sample autocorrelation and partial autocorrelation are significantly affected by additive outliers, although they are not affected by innovational outliers. Deutsch et al. (1990) used autocorrelation and partial autocorrelation estimates to demonstrate the effects of a single outlier to ARMA models. Tsay (1986, 1988), Chang et al. (1988) and Chen and Liu (1993) on the other hand, suggest iterative procedures for the detection of outliers. More specifically, Chen and Liu (1993) developed an iterative outlier detection procedure that jointly estimates the model parameters and outlier effects. In two recent contributions, Fokianos and Fried (2010, 2012) develop an iterative detection procedure for time series of counts through testing for intervention effects where both the type and time of the intervention are unknown using maximum likelihood based methodology.

Outline of the Thesis

In the next Chapter we review a linear and a log-linear Poisson model for count time series data when intervention effects are included. Also, some robust estimation procedures for independent data are discussed. The following two chapters concentrate on robust estimation of the log-linear model. In Chapter 3 two robust estimators - the Mallows' Quasi-Likelihood Estimator and the Conditionally Unbiased Bounded Influence Estimator - are studied, suitably adjusted in the count time series context. A log-linear Poisson model without feedback is considered, and the two estimators are compared to the maximum likelihood estimator when interventions are included, particularly level shift, transient shift and additive outliers. In Chapter 4 we focus on the Mallows' Quasi Likelihood Estimator which turned to be the most prominent method for estimating robustly in the presence of additive outliers. These type of outliers prove to be the most harmful ones when fitting a regression model to count data. The asymptotic properties of the Mallows' quasi likelihood estimator are studied under the log-linear Poisson model with feedback. Both chapters are complemented by simulation and real data examples. Chapter 5 focuses on the autocovariance of the log-linear model and discusses possible ways of approximation using orthogonal polynomials. Finally, we conclude this thesis summarizing the results of our work and stating further research possibilities.

Chapter 2

Literature Review

The abundant occurrence of count time series data in various applications in diverse fields of science has caused a surge towards the development of appropriate statistical models that take into account the fact that the response is integer valued. The direction we follow to the analysis of count time series is through the theory of generalized linear models (see McCullagh and Nelder (1989) for independent data and Kedem and Fokianos (2002) for time series).

Most of the models designed for count data assume a Poisson distributed response variable conditioned on the past of the time series although alternatives have been proposed in the literature. Examples are the negative binomial distribution and the double Poisson distribution (Fokianos (2012)). The Negative Binomial distribution is suggested by several authors (see Davis and Wu (2009), Zhu (2011) and Davis and Liu (2015)). Finally, in a more recent contribution Christou and Fokianos (2014) study a class of autoregressive models that include a feedback mechanism based on the mixed Poisson process and prove consistency and asymptotic normality of quasi-likelihood parameter estimates.

The Poisson assumption is the natural distributional assumption used to explain and model events that take place in a set time interval. Studies on models that incorporate the Poisson assumption are given among others by Davis et al. (2003), Fokianos et al. (2009), Fokianos and Tjøstheim (2011), Neumann (2011), Fokianos (2012), Fokianos and Tjøstheim (2012), Doukhan et al. (2012) and Douc et al. (2013).

Neumann (2011) considers a conditional Poisson model in which the intensity process $\{\lambda_t\}$ is a function of its previous value λ_{t-1} and the previous value of the response Y_{t-1}, i.e. $\lambda_t = f(\lambda_{t-1}, Y_{t-1})$ for some function f. It is shown that the joint process $\{Y_t, \lambda_t\}$ is stationary and ergodic under some contractive condition on the function f.

Fokianos (2012) reviewed a variety of regression models suggested for count time series based on the Poisson assumption, including the so called linear Poisson autoregressive and the log-linear Poisson autoregressive model. In the case of the linear Poisson autoregressive model the conditional mean is linearly linked to past values of the response as well as past values of itself. In the case of the log-linear Poisson autoregressive model the logarithm of the conditional mean is linearly linked to its past values and to a logarithmic function of past responses. Both models are discussed below although the log-linear model is studied extensively in the next chapters.

2.1 Linear Poisson Autoregressive Model

Let $\{Y_t\}$ be a count time series. Conditioned on the "past", this series is Poisson distributed with mean process $\{\lambda_t\}$. The simplest model which resembles the AR(1) model but for integer valued response, is the Poisson autoregressive model

$$Y_t \parallel \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = d + bY_{t-1}, \quad t \geq 1$$

where t is an integer and $\mathcal{F}_t = \sigma(Y_s, s \leq t)$ is the σ-field up to and including time t. The parameters d and b take non-negative values, something that ensures the non-negativity of λ_t. As stated in Fokianos (2012), as a general idea when the autocorrelation function of the process is available, if high values are found for large lags of the observations then this signifies that the above model can be used by considering a large number of lagged responses. A more parsimonious approach however, is to employ a feedback mechanism. The linear Poisson autoregressive model with feedback is

$$Y_t \parallel \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = d + a\lambda_{t-1} + bY_{t-1}, \quad t \geq 1. \tag{2.1}$$

The σ-field in this case is given by $\mathcal{F}_t = \sigma(Y_s, \lambda_0, s \leq t)$. The parameters d, a and b take positive values and further restrictions on those values which ensure stationarity will be discussed later. In (2.1), λ_0 is some starting value. The inclusion of past values of the mean process $\{\lambda_t\}$ distinguishes the two models; if $a = 0$ then (2.1) reduces to the previous model. Generally, the model without a feedback mechanism (i.e. equation (2.1) with $a = 0$) is employed for fitting when the autocorrelation function decays quickly after a few lags. On the contrary, the model that includes the feedback mechanism is preferred when the autocorrelation function decays slowly.

The general linear Poisson autoregressive model of order (p, q) is shown in the following representation (Ferland et al. (2006)):

$$Y_t \parallel \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \qquad \lambda_t = d + \sum_{i=1}^{p} a_i \lambda_{t-i} + \sum_{j=1}^{q} b_j Y_{t-j}, \qquad t \geq 1. \tag{2.2}$$

This model is an analogue to the GARCH models for volatility, in which there exists a feedback mechanism of the volatility (Bollerslev (1986)), because in the case of the Poisson distribution the conditional mean is equal to the conditional variance: $\text{E}(Y_t|\mathcal{F}_{t-1}) = \text{Var}(Y_t|\mathcal{F}_{t-1}) = \lambda_t$. Model (2.2), can therefore be considered as an Integer GARCH model of order (p, q), namely an INGARCH(p, q) model. The above model has been studied by several authors: we cite the works by Rydberg and Shephard (2000), Streett (2000), Heinen (2003), and Ferland et al. (2006). Streett (2000) has obtained some stationarity results and Ferland et al. (2006) have proven finite moments and second-order stationarity of the process under the condition $0 < \sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_j < 1$. For the first order model, the stationarity condition becomes $0 < a + b < 1$.

Examining the first order linear Poisson autoregressive model with feedback (2.1), the following representation for the response process can be obtained

$$\begin{aligned} Y_t &= \lambda_t + (Y_t - \lambda_t) = d + a\lambda_{t-1} + bY_{t-1} + \epsilon_t \\ &= d + a(Y_{t-1} - \epsilon_{t-1}) + bY_{t-1} + \epsilon_t \\ &= d + (a + b)Y_{t-1} + \epsilon_t - a\epsilon_{t-1} \end{aligned}$$

by considering that $\epsilon_t = Y_t - \lambda_t$. Therefore, the process $\{Y_t\}$ can be represented as an ARMA(1,1) process with mean $\mu = d/\{1 - (a + b)\}$,

$$\left(Y_t - \frac{d}{1 - (a + b)}\right) = (a + b)\left(Y_{t-1} - \frac{d}{1 - (a + b)}\right) + \epsilon_t - a\epsilon_{t-1}$$

and autocovariance function

$$\text{Cov}(Y_t, Y_{t+h}) = \begin{cases} \dfrac{(1-(a+b)^2+b^2)\mu}{1-(a+b)^2}, & h=0 \\ \\ \dfrac{b(1-a(a+b))(a+b)^{h-1}\mu}{1-(a+b)^2}, & h \geq 1 \end{cases}$$

see Ferland et al. (2006), Weiß (2009) and Fokianos (2012). The above representation for $h = 0$ demonstrates that the variance of the process is greater than the expectation, unless $b = 0$, which means that the time series is over-dispersed unless the conditional mean λ_t is not modeled upon previous observations Y_t.

Repeated substitution of the intensity process $\{\lambda_t\}$ for a fixed starting value λ_0 gives

$$\begin{aligned}
\lambda_t &= d + a\lambda_{t-1} + bY_{t-1} \\
&= d + a(d + a\lambda_{t-2} + bY_{t-2}) + bY_{t-1} \\
&= d(1+a) + a^2\lambda_{t-2} + abY_{t-2} + bY_{t-1} \\
&= d(1+a) + a^2(d + a\lambda_{t-3} + bY_{t-3}) + abY_{t-2} + bY_{t-1} \\
&= d(1+a+a^2) + a^3\lambda_{t-3} + b(a^2Y_{t-3} + aY_{t-2} + Y_{t-1}) \\
&= \cdots \\
&= d\frac{1-a^t}{1-a} + a^t\lambda_0 + b\sum_{i=1}^{t-1} a^i Y_{t-i-1}.
\end{aligned}$$

The above representation demonstrates that the intensity of the process depends solely on lagged values of the response. Therefore, the linear Poisson autoregressive model belongs to the class of observation-driven models, in the sense of Cox (1981). In particular, in Cox (1981) two categories were described, observation-driven and parameter-driven models. In the case of observation-driven models the mean function of the process depends exclusively on lagged values of the dependent variable, whereas in the case of parameter-driven models, a hidden process controls the mean function of the process.

Recall (2.1) and denote by $\boldsymbol{\theta}$ the vector of parameters $\boldsymbol{\theta} = (d, a, b)^T$. Then, given observations $Y_1, Y_2, ...Y_n$, the likelihood function for $\boldsymbol{\theta}$ is given by the expression

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{n} \frac{\exp(-\lambda_t(\boldsymbol{\theta}))\lambda_t^{Y_t}(\boldsymbol{\theta})}{Y_t!}.$$

The log-likelihood is then represented, up to a constant, by

$$\ell_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(\boldsymbol{\theta}) = \sum_{t=1}^{n} (Y_t \log \lambda_t(\boldsymbol{\theta}) - \lambda_t(\boldsymbol{\theta})).$$

The conditional maximum likelihood estimator is the solution of the score equation $S_n(\boldsymbol{\theta}) = 0$ where the score function is

$$S_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\boldsymbol{\theta})} - 1 \right) \frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

In the above, $\partial \lambda_t(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is a three dimensional vector whose components are

$$\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial d} = 1 + a\frac{\partial \lambda_{t-1}(\boldsymbol{\theta})}{\partial d}, \quad \frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial a} = \lambda_{t-1} + a\frac{\partial \lambda_{t-1}(\boldsymbol{\theta})}{\partial a}, \quad \frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial b} = Y_{t-1} + a\frac{\partial \lambda_{t-1}(\boldsymbol{\theta})}{\partial b},$$

see Fokianos et al. (2009) who studied the first order linear Poisson model with feedback (2.1) and proved geometric ergodicity using a perturbation technique. More specifically, a perturbed model is introduced

$$Y_t^m = N_t(\lambda_t^m), \quad \lambda_t^m = d + a\lambda_{t-1}^m + bY_{t-1}^m + \epsilon_{t,m}, \tag{2.3}$$

where $N_t(\cdot)$ is a Poisson process and the response Y_t^m reflects the number of events $N_t(\lambda_t^m)$ in the time interval $[0, \lambda_t]$ and

$$\epsilon_{t,m} = c_m \mathbb{1}(Y_{t-1}^m = 1)\mathcal{U}_t, \quad c_m > 0, \quad c_m \to 0 \text{ as } m \to \infty.$$

Note that $\mathcal{U}_t$ is a sequence of iid uniform random variables on $(0, 1)$ such that $\mathcal{U}_t$ is independent of $N_t(\cdot)$, and $\mathbb{1}(A)$ is the indicator function of a set A. This representation enables to prove geometric ergodicity of the perturbed joint process $\{Y_t^m, \lambda_t^m\}$ (see Robert and Casella (2005, Def. 4.6.8), among others). Then, the consistency and asymptotic normality of the maximum likelihood estimator for the perturbed model is established through geometric ergodicity and by letting $c_m \to 0$, that is, as the perturbed model approximates the non perturbed one, consistency and asymptotic normality for the non perturbed model is established.

Davis and Liu (2015) consider a more general class where the response conditionally on the past follows a distribution belonging to the one-parameter exponential family given a process that is a function of lagged responses. The linear Poisson model is a special case of the model studied by Davis and Liu (2015).

2.2 Log-Linear Poisson Autoregressive Model

Even though the linear Poisson autoregressive model described in Section 2.1 is a competent choice to model count time series data, it has two major disadvantages. First, it can only be employed when the observations are positively correlated. This is a consequence of the stationarity condition $0 < a + b < 1$. Moreover, covariates cannot be additively included in the model unless they result in a positively valued term, otherwise λ_t becomes negative. These problems are not encountered when a log-linear model is assumed. Consequently, a different approach towards modeling count time series should be considered and we resort to log-linear models, taking into account that the logarithm is the most popular link function for modeling count data. Log-linear models for time series of counts have been studied in many contributions, Zeger and Qaqish (1988), Li (1994), MacDonald and Zucchini (1997), Kedem and Fokianos (2002), Davis et al. (2003), Jung et al. (2006), Fokianos and Tjøstheim (2011) and Fokianos (2012).

The log-linear Poisson autoregressive model of order (p, q) is given by

$$Y_t \parallel \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \quad \nu_t = d + \sum_{i=1}^{p} a_i \nu_{t-i} + \sum_{j=1}^{q} b_j \log(1 + Y_{t-j}), \quad t \geq 1 \tag{2.4}$$

where $\nu_t \equiv \log \lambda_t$ is the canonical link process and the σ-field $\mathcal{F}_t$ is generated by $\sigma(Y_{1-q}, \dots, Y_t, \nu_{1-p}, \dots, \nu_0)$. The first order log-linear Poisson autoregressive model is expressed by

$$Y_t \parallel \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \quad \nu_t = d + a\nu_{t-1} + b \log(1 + Y_{t-1}), \quad t \geq 1. \tag{2.5}$$

As opposed to the linear Poisson autoregressive model described in the previous section, this model can accommodate both positive and negative correlation and additionally it can include time dependent covariates in a straightforward way by enlarging the σ-field and adding the covariate to the second equation of (2.5) (see Fokianos and Tjøstheim

(2011) and Fokianos (2012)).

The log-linear model belongs to the class of observation-driven models, since repeated substitution in the log-intensity process shows that the hidden process $\{\nu_t\}$ is determined by past functions of lagged responses,

$$
\begin{aligned}
\nu_t &= d + a\nu_{t-1} + b\log(1 + Y_{t-1}) \\
&= d + a(d + a\nu_{t-2} + b\log(1 + Yt - 2)) + b\log(1 + Y_{t-1}) \\
&= d(1 + a) + a^2\nu_{t-2} + ab\log(1 + Y_{t-2}) + b\log(1 + Y_{t-1}) \\
&= d(1 + a + a^2) + a^3\nu_{t-3} + b(a^2\log(1 + Y_{t-3}) + a\log(1 + Y_{t-2}) + \log(1 + Y_{t-1})) \\
&= \cdots \\
&= d\frac{1 - a^t}{1 - a} + a^t\nu_0 + b\sum_{i=1}^{t-1} a^i \log(1 + Y_{t-i-1}).
\end{aligned}
$$

As in the case of the linear model, the inclusion of the feedback process in (2.5) makes it a more parsimonious model than one in which higher lags of $\log(1 + Y_t)$ are included but not a feedback mechanism.

There are several reasons that support the inclusion of the term $\log(1+Y_{t-1})$ in model (2.5). It is a one-to-one transformation of Y_{t-1} which transforms both λ_t and Y_t into the same scale. Additionally, when compared to a model that includes the term $\log(c+Y_{t-1})$ instead, where c is a constant varying from 1 to 10 with step 0.5, there are not significant differences between the two models in terms of mean square error of the obtained MLE; see Fokianos and Tjøstheim (2011) who also examine a modification of model (2.5) that includes the term Y_{t-1} instead of $\log(1 + Y_{t-1})$. However, this means that the mean λ_t is given by

$$\lambda_t = \exp(d)\lambda_{t-1}^a \exp(bY_{t-1})$$

and stability can only be guaranteed when $b < 0$ because otherwise λ_t will increase in an exponential rate. Consequently, the above modified model that includes the term Y_{t-1} instead of $\log(1 + Y_{t-1})$ can only be considered to model negatively correlated data. In a different approach, Zeger and Qaqish (1988) considered the term $\log(\max(Y_{t-1}, c))$ for $c \in (0, 1]$ rather than $\log(1 + Y_{t-1})$. The authors have proved that the restriction $b < 1$ is a sufficient condition to show stability of the model given that $a = 0$. However, neither ergodicity nor asymptotic inference is examined.

Davis et al. (2003) consider the representation

$$\nu_t = \beta_0 + \beta_1 \frac{Y_{t-1} - \exp(\nu_{t-1})}{\exp(a\nu_t - 1)}$$

for the log-intensity process, where β_0 and β_1 are regression parameters and $a \in (0,1]$ and show stationarity of $\{\nu_t\}$ under the condition $\frac{1}{2} \leq a \leq 1$.

Fokianos and Tjøstheim (2011) study the model (2.5) and derive ergodicity using the perturbation idea, along the lines of Fokianos et al. (2009). Accordingly, an equivalent form of model (2.5) is given by the formulation

$$Y_t = N_t(\lambda_t), \quad \nu_t = d + a\nu_{t-1} + b\log(1 + Y_{t-1})$$

in terms of Poisson processes $N_t(\cdot)$ of unit intensity where Y_t given λ_t equals the number of events $N_t(\lambda_t)$ of the process $N_t(\cdot)$ in the interval $[0, \lambda_t]$. The perturbed chain (Y_t^m, ν_t^m) is defined by

$$Y_t^m = N_t(\lambda_t^m), \quad \nu_t^m = d + a\nu_{t-1}^m + b\log(1 + Y_{t-1}^m) + \epsilon_{t,m}, \tag{2.6}$$

where $\epsilon_{t,m} = c_m \mathbb{1}(Y_{t-1}^m = 1)\mathcal{U}_t$, $c_m > 0$, $c_m \to 0$ as $m \to \infty$, $\mathcal{U}_t$ is a sequence of iid uniform random variables on $(0,1)$ such that $\mathcal{U}_t$ is independent of $N_t(\cdot)$, and $\mathbb{1}(A)$ is the indicator function of a set A.

In the case of the log-linear Poisson model (2.5), the log-likelihood is given, up to a constant, by

$$\ell_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(\boldsymbol{\theta}) = \sum_{t=1}^{n} (Y_t \nu_t(\boldsymbol{\theta}) - \exp(\nu_t(\boldsymbol{\theta})))$$

and the conditional maximum likelihood score function is defined as

$$S_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} (Y_t - \exp(\nu_t(\boldsymbol{\theta}))) \frac{\partial \nu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

where $\partial \nu_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ is a three dimensional vector whose components are

$$\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial d} = 1 + a\frac{\partial \nu_{t-1}(\boldsymbol{\theta})}{\partial d}, \quad \frac{\partial \nu_t(\boldsymbol{\theta})}{\partial a} = \nu_{t-1} + a\frac{\partial \nu_{t-1}(\boldsymbol{\theta})}{\partial a}, \quad \frac{\partial \nu_t(\boldsymbol{\theta})}{\partial b} = Y_{t-1} + a\frac{\partial \nu_{t-1}(\boldsymbol{\theta})}{\partial b}.$$

Fokianos and Tjøstheim (2011) provide conditions to prove geometric ergodicity. Particularly, the authors prove that under the condition that $|a+b| < 1$ when $|a| < 1$ and $b > 0$, and $|a||a+b| < 1$ when $b < 0$ then the perturbed process is stationary and geometrically ergodic with finite moments of any order. Geometric ergodicity is used to deduce the asymptotic properties of the maximum likelihood estimator of the perturbed model parameters. It was shown that if $|a+b| < 1$, whenever a and b have the same sign, and $a^2 + b^2 < 1$ whenever a and b have different signs, then the conditional maximum likelihood estimator of the parameter vector (d, a, b) is consistent and asymptotically normally distributed. Finally, as $c_m \to 0$ asymptotic properties are derived for the non-perturbed model. More specifically, Fokianos and Tjøstheim (2011) show that the maximum likelihood estimator is consistent and asymptotically normal.

Douc et al. (2013) consider the general class of observation driven models and derive ergodicity conditions based on the theory of Markov chains. The log-linear Poisson model is included in their work as an example. Some further properties of the model and relaxation of the above stationarity conditions are discussed. The condition acquired to ensure ergodicity is $\max\{|a+b|, |a|, |b|\} < 1$.

2.3 Interventions in Count Time Series

Fokianos and Fried (2010, 2012) study the problem of estimation and detection of various types of intervention effects on time series of counts for models (2.1) and (2.5).

In the case of the linear model (2.1), Fokianos and Fried (2010) argue that it is more sensible to import intervention effects through the mean process which governs the dynamics of the model. Additionally, if intervention effects are added directly to the response variable then they would have to be integer valued in order for the response variable to remain integer valued. Therefore, a sequence of covariates of the form

$$X_t = \xi(\mathcal{B})\mathbb{1}(t = \tau), \quad t \geq 1$$

is introduced to the mean process that indicates an intervention happening at the time point τ. The response process that is observed is henceforth a contaminated response time series. The contaminated model is given by

$$Z_t \parallel \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t^c), \quad \lambda_t^c = d + \sum_{i=1}^{p} a_i \lambda_{t-i}^c + \sum_{j=1}^{q} b_j Z_{t-j} + \zeta X_t, \quad t \geq 1. \tag{2.7}$$

The σ-field $\mathcal{F}_{t-1}$ is generated by $\{Z_{1-q,\ldots}, Z_t, \lambda_{1-p}, \ldots, \lambda_0\}$ and

$$\mathbb{1}(t = \tau) = \begin{cases} 1 & \text{if } t = \tau \\ 0 & \text{else} \end{cases}$$

is an indicator function that takes the value 1 is $t = \tau$ and 0 otherwise. The term $\xi(\mathcal{B})$ in the covariate sequence is a polynomial which categorizes intervention effects according to which observations they affect. More specifically,

$$\xi(\mathcal{B}) = (1 - \delta\mathcal{B})^{-1}, \quad \delta \in [0, 1]$$

and $\mathcal{B}$ is a shift operator such that $\mathcal{B}^i X_t = X_{t-i}$. The value of δ categorizes the type of the intervention:

- $\delta = 0$ corresponds to a Spiky Outlier (SO), an intervention which affects only a single observation,
- $\delta = 1$ categorizes the intervention as a Level Shift (LS) which affects the entire series from the time of the intervention and on and
- if $\delta \in \{0.7, 0.8, 0.9\}$ then it is called a Transient Shift (TS). In this case the effect of the intervention affects a few observations and decays exponentially with rate δ.

For the intervention types described above, the formulation of the covariate X_t is deduced to $X_t = \delta^{t-\tau}\mathbb{1}(t = \tau)$ where $\mathbb{1}(\cdot)$ denotes the indicator function.

We note that in the case of a Spiky Outlier, where the outlier affects only one value of λ_t, still a few observations Y_t are affected since the outlier is imported on the mean process λ_t which is fed back into the next value.

The authors in this contribution propose an iterative detection and estimation procedure for intervention effects by using the maximum likelihood in three scenarios, when the type and time of the intervention are both known, both unknown and the scenario of multiple interventions. The proposed procedure can be considered as an extension to Poisson time series of the procedure proposed by Chen and Liu (1993). Below we give the algorithms proposed by Fokianos and Fried (2010) and Fokianos and Fried (2012) for the linear and log-linear Poisson models respectively.

Algorithm for the linear Poisson model:

For positive intervention effect $\zeta > 0$, the observed process $\{Z_t\}$ is decomposed as $Z_t = Y_t + C_t$ where Y_t is the unobserved uncontaminated process and C_t is a sequence of Poisson random variables whose mean depends on ζ and also on the type of the intervention, that is on the choice of $\xi(\mathcal{B})$. Set $Z_t^{(1)} = Z_t$, $t = 1, \ldots, n$, and $k = 1$ for initialization:

1. Fit a linear Poisson model (2.2) to the data $\{Z_t^{(k)}, t = 1, \ldots, n\}$.

2. Test the hypothesis $H_0^{(\tau)} : \zeta = 0$ against $H_1^{(\tau)} : \zeta \neq 0$ for a single intervention of any type at any time point by employing (2.7) and using the maximum of the score test statistics $\tilde{T}_n = \max_\tau T_n(\tau)$ where

$$T_n(\tau) = S_{n\tau}^T(\tilde{d}, \ldots, \tilde{a}_p, 0) G_{n\tau}^{-1}(\tilde{d}, \ldots, \tilde{a}_p, 0) S_{n\tau}(\tilde{d}, \ldots, \tilde{a}_p, 0)$$

   and $S_{n\tau}(\tilde{d}, \ldots, \tilde{a}_p, 0)$ and $G_{n\tau}^{-1}(\tilde{d}, \ldots, \tilde{a}_p, 0)$ are the score function and conditional information matrix for the contaminated model (2.7).

3. If there is no significant result, then stop; the data $Z_1^{(k)}, \ldots, Z_n^{(k)}$ are considered as clean. Otherwise:

   (a) Fit a contaminated model (2.7) by choosing $\xi(\mathcal{B})$ according to the type of intervention identified in the previous step. Let $\hat{\zeta}$ be the estimated size of the intervention effect and τ its point in time, which is estimated as the time point maximizing the corresponding test statistic $\tilde{T}_n$.

   (b) Estimate the effect of the intervention on the observation $Z_t^{(k)}$ by the rounded value

$$\hat{C}_t = \left[\frac{\hat{\mu}_t}{\hat{\lambda}_t^c} Z_t^{(k)} \right]$$

   where $\hat{\lambda}_t^c$ is obtained from equation (2.7) by plugging in the estimates of the model parameters and

$$\hat{\mu}_t = \sum_{i=1}^{q} \hat{b}_i \hat{C}_{t-i} + \sum_{j=1}^{p} \hat{\alpha}_j \hat{\mu}_{t-j} + \hat{\zeta} X_t \ , t = \tau, \tau+1, \ldots,$$

with $\hat{C}_t = \hat{\mu}_t = 0$ for $t < \tau$.

(c) Correct the time series for the estimated intervention effects by setting

$$Z_t^{(k+1)} = \begin{cases} Z_t^{(k)} - \hat{C}_t, & t \geq \tau, \\ Z_t^{(k)}, & t < \tau, \end{cases}$$

increase k by 1 and return to step 1.

The iterative procedure is continued until no further interventions are detected.

In a recent contribution, Liboschik et al. (2014) consider a similar contaminated model but in their approach the intervention influences the observation and not the underlying mean at the time it occurs and afterwards enters the dynamics of the process.

Also Fried et al. (2014) discuss the model $\eta(\lambda_t) = \beta_0 + \beta_1 \eta(Y_{t-1} + c) + a_1 \eta(\lambda_{t-1})$ following generalized linear models, where $\eta(\cdot)$ is the link function. By considering the known constant c to be equal to 1, the link functions identity and logarithm correspond to the linear and log-linear Poisson models (2.1) and (2.5) respectively.

A different approach to the linear Poisson model within the Bayesian framework is proposed by Fried et al. (2015). The authors define additive outliers in the context of the linear Poisson model that only affect the observations at which they occur but do not contaminate subsequent observations, and suggest a Bayesian retrospective detection and estimation procedure of outlier effects.

We also point out the work by Elsaied and Fried (2014) which is related to our work and examines robust M-estimation for the parameters of the linear Poisson model. More specifically, Elsaied and Fried (2014) focus on the linear Poisson model without feedback, namely an INARCH(q) model, and propose M-estimation using a modification of the conditional log-likelihood with weighting and bias correction. The proposed estimator is a solution to the estimating equations

$$\sum_{t=2}^{n} \psi\left(\frac{y_t - \lambda_t}{\sqrt{\lambda_t}}\right)\frac{1}{\sqrt{\lambda_t}}\begin{pmatrix} 1 \\ \sigma\psi\left(\frac{y_{t-1}-\lambda}{\sigma}\right) + \lambda \end{pmatrix} - (n-1)\begin{pmatrix}\alpha_0 \\ \alpha_1\end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

In the above representation, the function $\psi(\cdot)$ is either the Huber's or Tukey's ψ function, λ and σ are the marginal mean and variance given by $\lambda = d/(1-b)$ and $\sigma^2 = d/((1-b)(1-b^2))$ respectively and $(\alpha_0, \alpha_1)^T$ is the bias correction term. The authors follow Cantoni and Ronchetti (2001) and approximate the bias correction by

$$\sum_{j=0}^{\infty}\sum_{i=0}^{\infty} \psi\left(\frac{j - \lambda_t}{\sqrt{\lambda_t}}\right)\frac{1}{\sqrt{\lambda_t}}\begin{pmatrix} 1 \\ \sigma\psi\left(\frac{i-\lambda}{\sigma}\right) + \lambda \end{pmatrix} \times \mathrm{P}(Y_t = j|Y_{t-1} = i) \times \mathrm{P}(Y_{t-1} = i)$$

where the conditional probability $\mathrm{P}(Y_t = j|Y_{t-1} = i)$ is derived from a Poisson distribution with rate $\lambda_t = d + bi$ and the probability $\mathrm{P}(Y_{t-1} = i)$ is empirically estimated by Monte Carlo simulation.

All of the intervention effects that are considered in Fokianos and Fried (2010) are modeled through the underlying mean process and therefore influence future observations since the mean depends on past observations and past mean values. However, this framework cannot be used to model purely additive outliers. An Additive Outlier (AO) is another form of intervention given by

$$Z_t = \begin{cases} Y_t + \zeta, & \text{when } \; t = \tau \\ Y_t, & \text{otherwise} \end{cases}$$

where ζ is the size of the outlier at time τ and ζ is an integer.

Fried et al. (2011) extend the work of Fokianos and Fried (2010) to include Additive Outliers, proposing M-estimators for robustly estimating the model parameters. Even though Fokianos and Fried (2010) consider maximum likelihood estimation, in the presence of additive outliers it is quite ambitious to estimate the model parameters using maximum likelihood estimation. This is because in the construction of the likelihood, $Z_{\tau+1}$ needs to be conditioned on Y_τ instead of Z_τ, but Y_τ is unobserved.

The case of the log-linear Poisson model (2.5) however requires a different modeling approach due to the logarithmic structure. Fokianos and Fried (2012) assume a contaminated logarithmic model

$$Z_t \parallel \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t^c), \quad \nu_t^c = d + \sum_{i=1}^{p} a_i \nu_{t-i}^c + \sum_{j=1}^{q} b_j \log(Z_{t-j}+1) + \zeta X_t, \quad t \geq 1 \quad (2.8)$$

where in this case the σ-field is generated by $\{Z_{1-q,\ldots}, Z_t, \nu_{1-p}, \ldots, \nu_0\}$. In contrast to the contaminated linear Poisson model described above where the intervention affects additively the mean process, in the contaminated log-linear Poisson model the intervention affects the logarithm of the mean process, resulting in a multiplicative effect $\exp\{\zeta\}$ on the mean process λ_t.

Algorithm for the log-linear Poisson model:

Set $Z_t^{(1)} = Z_t$, $t = 1, \ldots, n$, and the iteration index k to 1 for initialization:

1. Fit a log linear autoregressive model (2.4) to the data $\{Z_t^{(k)}, t = 1, \ldots, n\}$.

2. Test for a single intervention of any type at any time point by employing (2.8) and using the maximum of the score test statistics $\tilde{T}_n$

3. If there is no significant result, stop; regard the data $Z_1^{(k)}, \ldots, Z_n^{(k)}$ as clean. Otherwise:

   (a) Fit a contaminated log linear Poisson autoregressive model (2.8) by choosing $\xi(\mathcal{B})$ according to the type of intervention identified in the previous step. Let $\hat{\zeta}$ be the estimated size of the intervention effect and $\hat{\tau}$ its estimated point in time, which is the time point maximizing the corresponding test statistic.

   (b) Correct the time series for the estimated intervention effects by setting

$$Z_t^{(k+1)} = \begin{cases} \left[Z_t^{(k)} / \exp(\hat{\mu}_t) \right], & t \geq \tau, \\ Z_t^{(k)}, & t < \tau, \end{cases}$$

   where $\hat{\mu}_t$ is obtained recursively for $t = \tau, \tau+1, \ldots$ from the following equation by plugging in the estimates of the model parameters,

$$\hat{\mu}_t = \sum_{i=1}^{p} \hat{a}_i \hat{\mu}_{t-i} + \sum_{j=1}^{q} \hat{b}_j \log\left(1 + \frac{\hat{C}_{t-j}}{Z_{t-j}^{(k)} + 1}\right) + \hat{\omega} x_t \ , \ t = \tau, \tau+1, \ldots,$$

with $\hat{C}_t = \hat{\mu}_t = 0$ for $t < \tau$ and the estimate of the intervention effect on $Z_t^{(k)}$ being

$$\hat{C}_t = Z_t^{(k)} - Z_t^{(k+1)} \ , \ t \geq \tau \ .$$

Then increase k by 1 and return to step 1.

This iterative procedure is continued until no further interventions are detected.

Figures 2.1 and 2.2 demonstrate how the various types of interventions described above affect a time series in which the linear model (2.1) and log-linear model (2.5) are employed respectively.

Fokianos and Fried (2012) propose a stepwise procedure using a score test statistic to detect and jointly estimate the model parameters as well as the size of interventions using maximum likelihood. A notable remark is that the detection procedure that Fokianos and Fried (2010, 2012) propose does not seem to be strongly influenced by the link function since in the real data examples considered the same intervention effects are detected using both the identity and the logarithmic link function.

2.4 Robust Estimation for Independent Data

In this section we discuss several robust estimators for independent data within the framework of generalized linear models. The response variable follows a distribution F belonging to the exponential family of distributions where the mean $\mathrm{E}(Y_i) = \mu_i$ depends on the explanatory variables x_i through a link function $g(.)$ such that $\eta_i = g(\mu_i) = x_i^T \boldsymbol{\beta}$, $i = 1, \ldots, n$, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of parameters and $\mathrm{Var}(Y_i) = V(\mu_i)$.
An M-estimator is the solution of the estimating equations

$$\sum_{i=1}^{n} \boldsymbol{\psi}(y_i, \mu_i) = \mathbf{0}$$

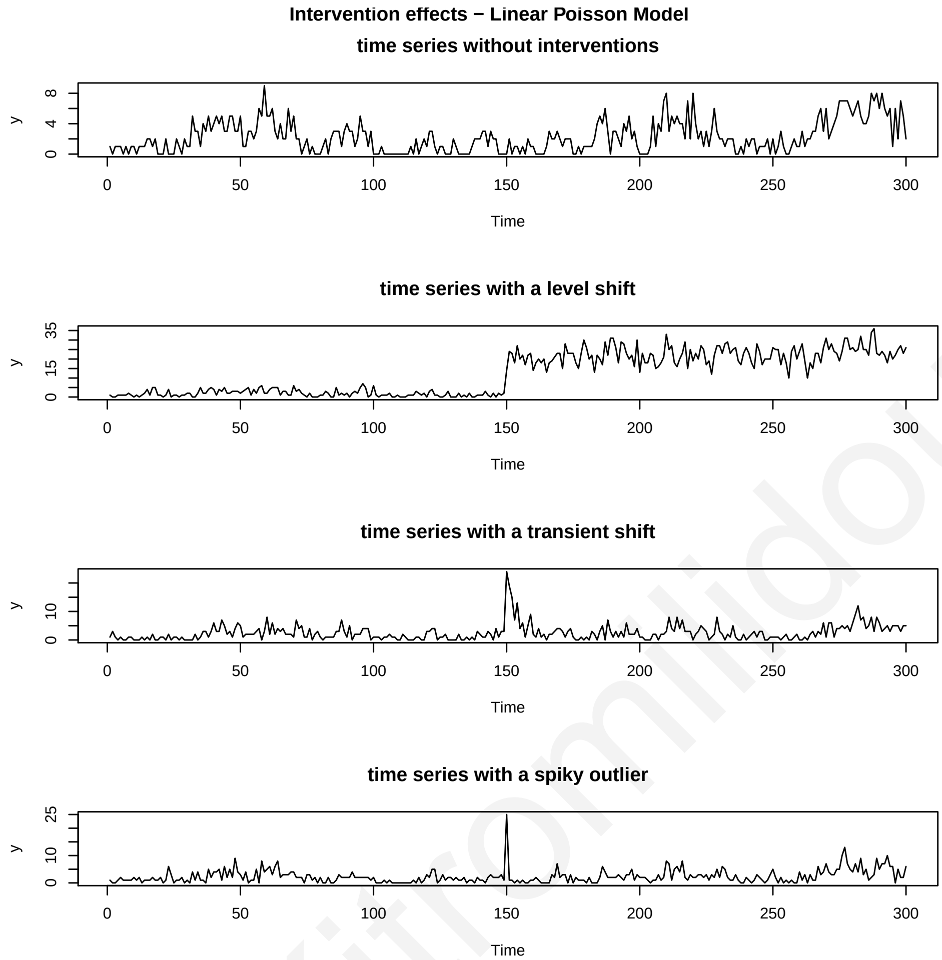

Figure 2.1: Example of a time series with various forms of interventions, based on the linear Poisson model.

where $\boldsymbol{\psi}$ is a function with certain properties (see Section 5, van der Vaart (1998)). Hampel et al. (1986) define the Influence Function of an M-estimator as

$$\mathrm{IF}(y; \boldsymbol{\psi}, F) = M(\boldsymbol{\psi}, F)^{-1} \boldsymbol{\psi}(y, \mu)$$

where the matrix M is given by

$$M(\boldsymbol{\psi}, F) = -\mathrm{E}\left[\partial \boldsymbol{\psi}(y, \mu) / \partial \boldsymbol{\beta}\right].$$

The Influence Function is a measure of robustness which measures the asymptotic bias that is caused by an infinitesimal contamination. Furthermore, the estimator defined above is asymptotically normally distributed and its asymptotic covariance matrix is given by the

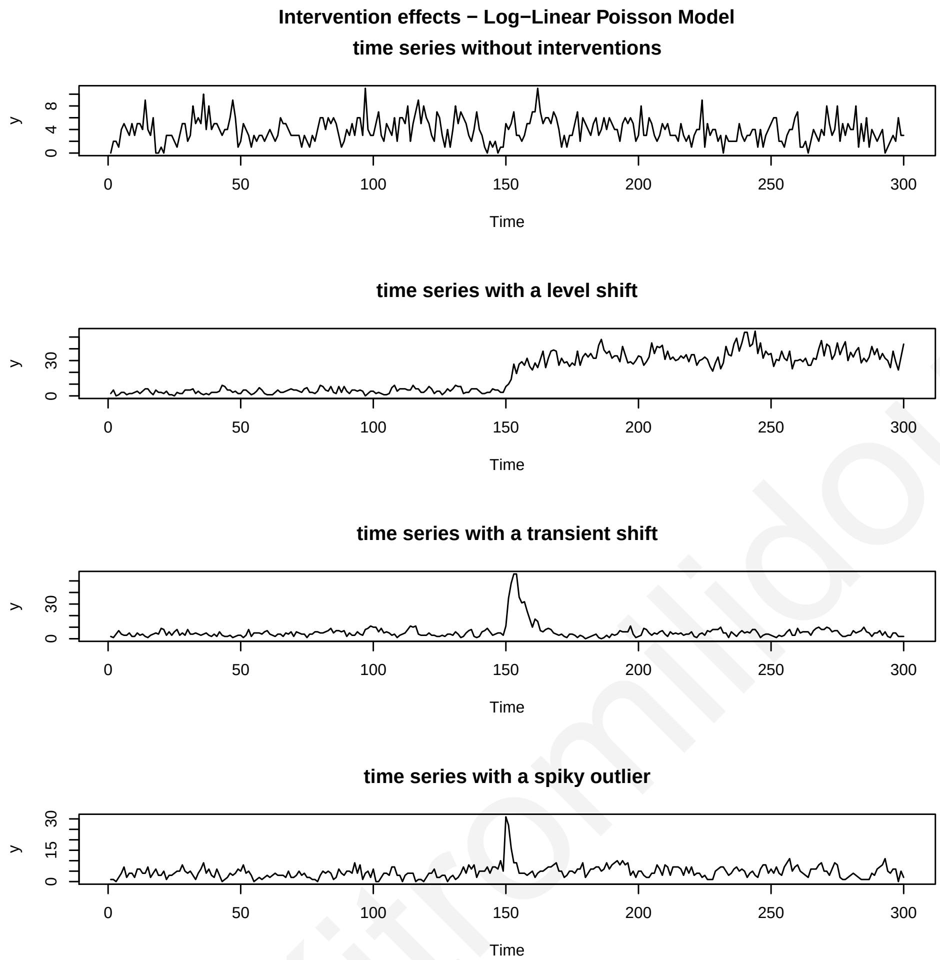

Figure 2.2: Example of a time series with various forms of interventions, based on the log-linear Poisson model.

formulation

$$\Omega = M(\boldsymbol{\psi}, F)^{-1} Q(\boldsymbol{\psi}, F) M(\boldsymbol{\psi}, F)^{-1}$$

where the matrix $Q(\boldsymbol{\psi}, F)$ is

$$Q(\boldsymbol{\psi}, F) = \mathrm{E}\left[\boldsymbol{\psi}(y,\mu)\boldsymbol{\psi}(y,\mu)^T\right].$$

Since the Influence Function is proportional to $\boldsymbol{\psi}(y,\mu)$, it is unbounded. Therefore, a bound on the Influence Function will ensure the robustness of the estimator. Because the Influence Function of the estimator is a vector, a scalar measure of the Influence Function is given by the self-standardized sensitivity (see Krasker and Welsch (1982))

$$s^2(\boldsymbol{\psi}) = \sup_{y,\mu} \boldsymbol{\psi}(y,\mu)^T Q(\boldsymbol{\psi}, F) \boldsymbol{\psi}(y,\mu).$$

2.4.1 The Conditionally Unbiased Bounded-Influence Estimator (CUBIF)

The Conditionally Unbiased Bounded Influence Estimator (CUBIF) was proposed by Künsch et al. (1989) to robustly estimate the parameters in a regression model. CUBIF is a conditionally Fisher consistent M-estimator, made robust by bounding its influence function, more specifically its scalar measure, the self-standardized sensitivity and has minimum variance subject to this bound.

We note that Fisher-consistency depends on the explanatory variables being independent and additionally relies on their distribution. This is a disadvantage because the explanatory variables sometimes are not random. Thus, conditional Fisher-consistency is a more desirable property since not only it does not involve the distribution of the predictors but also it does not depend on them, being random.

Recall again that we discuss a generalized linear model where the response variable follows a distribution F belonging to the exponential family of distributions where the mean $\mathrm{E}(Y_i) = \mu_i$ depends on the explanatory variables x_i through a link function $g(.)$ such that $\eta_i = g(\mu_i) = x_i^T\boldsymbol{\beta}$, $i = 1, \ldots, n$, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of parameters and $\mathrm{Var}(Y_i) = V(\mu_i)$.

The CUBIF score function is given by

$$\psi_{\text{cond}}(y, x, \boldsymbol{\beta}, c, B) = d(y, x, \boldsymbol{\beta}, c, B) W_c(|d(y, x, \boldsymbol{\beta}, c, B)|(x^T B x)^{1/2}) x.$$

In the above representation,

$$d(y, x, \boldsymbol{\beta}, c, B) = y - \mu - C(x^T\boldsymbol{\beta}, \frac{c}{(x^T B^{-1} x)^{1/2}})$$

and

$W_c(\alpha) = \psi_c(\alpha)/\alpha$ where $\psi_c(\alpha)$ is the Huber function

$$\psi_c(\alpha) = \begin{cases} \alpha, & |\alpha| \leq c \\ c\,\mathrm{sign}(\alpha), & |\alpha| > c. \end{cases}$$

The scalar function $C(.)$ is a bias correction that ensures Fisher consistency. The matrix

B is chosen so that a bound is appointed on the sensitivity, $s(\psi_{\text{cond}}) = c$, and is given by $B = \mathrm{E}\{\psi_{\text{cond}}\psi_{\text{cond}}^T\}$.

2.4.2 The Mallows' Quasi Likelihood Estimator (MQLE)

In another contribution, Cantoni and Ronchetti (2001) propose the Mallows' Quasi Likelihood Estimator (MQLE). The MQLE is an M-estimator, based on the quasi likelihood, of Mallows' type. That is, the observations and the explanatory variables are bounded separately. A bound on the influence function ensures the robustness of the estimator and appropriate weights are used to downsize leverage points on the explanatory variables. The MQLE estimator is the solution of the estimating equations

$$\sum_{i=1}^{n} \left[\psi_c(r_i) w(\boldsymbol{x}_i) \frac{1}{V^{1/2}(\mu_i)} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} - \alpha(\boldsymbol{\beta}) \right] = 0.$$

Here, the term

$$\alpha(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E} \left[\psi_c(r_i) w(\boldsymbol{x}_i) \frac{1}{V^{1/2}(\mu_i)} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right]$$

is a bias correction term which ensures Fisher consistency and can be calculated explicitly under a Poisson or a Binomial assumption. In addition,

$$r_i = \frac{Y_i - \mu_i}{V^{1/2}(\mu_i)}$$

are the Pearson residuals. Finally, $w(\boldsymbol{x}_i)$ are weights and the authors discuss several weighting options, including weights defined from the hat matrix and robust weights defined using the inverse of the robust Mahalanobis distance when center and scale are estimated to have high breakdown properties. For more information on breakdown points see Maronna et al. (2006). In the particular cases where the response follows a conditional Binomial or Poisson distribution then the bias correction term can be calculated exactly in a closed form expression. We provide more details on the weighting options for the MQLE estimator in the next chapter where this estimator is studied within the framework of count time series.

2.4.3 L_q - Estimator

Morgenthaler (1992) suggests to use the L_q-norm, $q \geq 1$, in the quasi-likelihood, instead of the L_2-norm. He proposed a conditionally Fisher consistent estimator, the L_q estimator of $\boldsymbol{\beta}$, which is determined as the solution to the estimating equations

$$\sum_{i=1}^{n} \{|y_i - \mu_i|^{q-1} \mathbb{1}(y_i - \mu_i) - c(\mu_i)\} \frac{1}{(V(\mu_i))^{q/2}} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

where the term $c(\mu_i)$ is a bias correction term to ensure Fisher consistency given by

$$c(\mu_i) = \mathrm{E}_{Y|X}\{|y_i - \mu_i|^{q-1} \mathbb{1}(y_i - \mu_i)\}.$$

We note however, that the bias correction term cannot be displayed by a closed form, hence it cannot be easily computed. Therefore, it is computed by Monte Carlo simulation. More specifically, the term $c(\mu_i)$ is obtained as a sample mean: m samples z_{ij}, $j = 1, \ldots, m$, are generated from the distribution that we postulate the observations to follow and calculate the term $c(\mu_i)$ by

$$\hat{c}(\mu_i) = \frac{1}{m} \sum_{i=1}^{m} |z_{ij} - \mu_i|^{q-1} \mathbb{1}(z_{ij} - \mu_i).$$

The L_q estimate $\hat{\boldsymbol{\beta}}$ is computed using the Newton-Raphson method. However, the computation of the L_q estimate presents numerical difficulties because when the estimated parameters $\hat{\mu}_i$ are close to the responses y_i, the Newton-Raphson algorithm does not converge.

Hosseinian (2009) studies the L_q estimator for Poisson regression models. In this case, the bias correction term $c(\mu_i)$ can be calculated by the formulation

$$c(\mu_i) = -\sum_{i=1}^{\lfloor \mu_i \rfloor} (\mu_i - y_i)^{q-1} \mathrm{P}(Y_i = y_i) + \sum_{y_i = \lceil \mu_i \rceil}^{\infty} (y_i - \mu_i)^{q-1} \mathrm{P}(Y_i = y_i).$$

The infinite sum in the above expression is a convergent sum and therefore it can be approximated by a finite sum of the first N terms. To define the proper value of the constant N, Hosseinian (2009) has computed the above approximation for values of $N = 10, 100, 1000, 10000$ in order to find the best value of N for which the maximum error is bounded. Her simulation study indicates that taking $N = 4\bar{y}$ establishes sufficient accuracy.

Hosseinian (2009) also proposes a modification of the L_q-estimator for binary regression, namely the BL_q estimator. In this contribution we do not study binary regression so the BL_q estimator will not be discussed although this could be a subject for further research, as is commented in Chapter 6.

2.4.4 Weighted Maximum Likelihood Estimates for Poisson Regression

Hosseinian (2009) introduces two weighted maximum likelihood estimates for Poisson regression. Both estimates are based on weighting the maximum likelihood estimating equation

$$\sum_{i=1}^{n} \frac{1}{\mu_i} W(\mu_i)(y_i - \mu_i) \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}.$$

The first Weighted Maximum Likelihood Estimate (WMLE) corresponds to the weight function

$$W(\mu_i) = \exp\left(-\frac{(\mu_i - m)^2}{2s^2}\right)$$

where $m = \sum_{i=1}^{n} y_i/n$ is the arithmetic mean and $s^2 = \sum_{i=1}^{n}(y_i - m)^2/n$ is the variance of the observations, although m and s can be replaced by robust alternatives using the MCD algorithm described in Rousseeuw and Driessen (1999). However, the above weight function has a few drawbacks. First, the resulting estimate is biased because $\mathrm{E}(W(\mu_i)(Y_i - \mu_i)) \neq 0$ and an additional drawback is that it exhibits numerical instabilities when the sample size is small. The second weighted maximum likelihood estimate $\mathrm{WMLE}^{\mathrm{MH}}$ corresponds to the weight function

$$W^{MH}(\mu_i) = \begin{cases} 1, & \frac{\upsilon}{c_1} < \mu_i < c_1\upsilon \\ \dfrac{c_1\mu_i}{\upsilon}, & \mu_i < \dfrac{\upsilon}{c_1} \\ \dfrac{c_2\upsilon - \mu_i}{\upsilon}, & c_1\upsilon < \mu_i < c_2\upsilon \\ 0, & \mu_i \geq c_2\upsilon. \end{cases}$$

The constants c_1 and c_2 are arbitrarily chosen and υ is the median of μ. Hosseinian (2009) suggests the arbitrary values of $c_1 = 2$ and $c_2 = 3$.

Chapter 3

Log-Linear Poisson Model Without Feedback

We study robust inference for the log-linear Poisson model without feedback for count time series under three different forms of interventions: additive outliers, transient shifts and level shifts. Some robust estimation procedures, which have been suggested for the analysis of independent and not identically distributed data, are applied to this setting and their performance is investigated by an empirical study. In fact, the aim of this chapter is to extend those estimation procedures to the setting of count time series and to examine empirically their behavior.

We estimate the parameters of the log-linear Poisson model without feedback using the maximum likelihood estimator (MLE), the conditionally unbiased bounded-influence estimator (CUBIF) and the Mallows' quasi-likelihood estimator (MQLE) and compare all three estimators in terms of their mean square error, bias and mean absolute error.

Our empirical results illustrate that under a level shift or a transient shift there are no significant differences among the three estimators and the most interesting results are obtained in the presence of additive outliers. The results are complemented by real data examples.

We will assume that $\{Y_t, t = 1, \ldots, n\}$ is a count time series and we will denote by $\mathcal{F}_t$ the history of the process up to and including time t. An approach to analyze count time series, is based on the framework of generalized linear models as described in the previous chapter.

Elsaied (2012) addresses partially some of the questions that we study in this chapter but from a different perspective, in the context of the linear Poisson model (2.1). More specifically, Elsaied (2012) suggests a new bias-corrected M-estimator which depends on the Tukey's bisquare function - as opposed to the Huber's function. As discussed in Section 2.1, a linear count time series, such as (2.1), is related, in some sense, with the ordinary GARCH models. We note that for this class of processes studies on robust estimation have been advancing over the last years; see Muler and Yohai (2002, 2008) and more recently Mukherjee (2008). Muler and Yohai (2002, 2008) propose two robust M-estimates for ARCH and GARCH models respectively, based on a modification of the likelihood function. The second estimate is adjusted by bounding the effect of one outlier on subsequent observations. The proposed estimates are also consistent and asymptotically normal. Mukherjee (2008) defines a more general class of M-estimates for GARCH models that includes among others the quasi maximum likelihood estimate.

However, in this chapter, we focus on model (2.2) and in particular on models that do not contain a feedback mechanism. Based on this model, we advance robust estimation of the regression parameters in the presence of sudden events. To the best of our knowledge, studies on robust estimation for count time series are missing from the literature. In Section 3.1, we discuss the notion of interventions and their effects on the estimated model parameters. We describe three methods of estimation in Section 3.2, which we compare empirically in Section 3.3. Finally, in Section 3.4 we apply these methods to real data sets.

3.1 Intervention Effects

As reviewed in the previous chapter, Fokianos and Fried (2010, 2012) studied the problem of detection and testing for outliers and intervention effects in the linear count time series model (2.1) and in the log-linear count time series model (2.5), respectively. Our goal is to complement this research by developing robust estimation procedures. Following the work of Fokianos and Fried (2010, 2012), recall that we observe a contaminated process

$\{Z_t\}$ given by

$$Z_t \parallel \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t^c), \quad \nu_t^c = d + \sum_{i=1}^{p} a_i \nu_{t-i}^c + \sum_{j=1}^{q} b_j \log(1 + Z_{t-j}) + \zeta X_t, \quad t \geq 1,$$

where the sequence $\{X_t\}$ is a deterministic covariate sequence modeling the intervention happening at time τ as

$$X_t = \xi(\mathcal{B})\mathbb{1}(t = \tau), \quad t \geq 1.$$

In the above display, $\xi(\mathcal{B}) = (1 - \delta\mathcal{B})^{-1}, \quad \delta \in [0, 1)$, $\mathcal{B}$ is a shift operator such that $\mathcal{B}^i X_t = X_{t-i}$, $\mathbb{1}(t = \tau)$ is an indicator function that is equal to 1 if $t = \tau$ and 0 otherwise and ζ is the size of the intervention. We consider two forms of interventions as follows:

- Transient Shift (TS) which corresponds to the case $\delta \in \{0.7, 0.8, 0.9\}$,
- Level Shift (LS) which corresponds to $X_t = \mathbb{1}(t \geq \tau)$.

For the above cases, the covariate X_t becomes $\delta^{t-\tau}\mathbb{1}(t \geq \tau)$ but other cases can be included in the general framework. We note that the parameter δ can be estimated from the data should we already know or have estimated the time that the intervention occurred. If, however, the time of the intervention is unknown, we consider a fixed value of δ, as discussed in Fokianos and Fried (2010, 2012). Recall that another interesting form of outlier modeling is given by

$$Z_t = \begin{cases} Y_t + \zeta, & \text{when} \quad t = \tau, \\ Y_t, & \text{otherwise,} \end{cases} \tag{3.1}$$

where Y_t follows (2.4). This corresponds to the case of an Additive Outlier (AO) of size ζ at time τ. All these different forms of outliers have been considered by Fokianos and Fried (2012) in the context of the log-linear model (2.5). It was shown that a TS type of outlier yields a sudden effect to the observation which eventually decays, a LS type of intervention changes the mean of the process (either upwards or downwards depending on the sign of the coefficient ζ) and an AO affects the observation Z_τ and its values a few lags after. We note that Innovation Outliers (IO) are not considered in our work. An innovative outlier in the context of ARMA modeling, affects the observations from the time of occurrence henceforth through a polynomial. To be more precise, an innovative outlier of size ζ occurring at time τ has an effect equal to $\zeta\phi_k$ where ϕ_k is the kth coefficient of the polynomial $\phi_0 + \phi_1 B + \phi_2 B^2 + ...$ (Chen and Liu (1993)). However, in the context of the log-linear Poisson model there is no direct equivalent to innovative outliers because the model is not defined in terms of innovations (Fokianos and Fried (2012)).

For the following, we will confine on the special case of (2.4) with $p = 0$; that is of the log–linear model without feedback, which is given by

$$Y_t \parallel \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \quad \nu_t = d + \sum_{j=1}^{q} b_j \log(Y_{t-j} + 1). \tag{3.2}$$

In the presence of interventions we observe a contaminated version $\{Z_t\}$ of the clean process $\{Y_t\}$ of the following form,

$$Z_t \parallel \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t^c), \quad \nu_t^c = d + \sum_{j=1}^{q} b_j \log(Z_{t-j} + 1) + \zeta X_t. \tag{3.3}$$

Consider now model (3.3) and suppose that the goal is to estimate the parameters d and $\{b_j, j = 1, 2, \ldots, q\}$ using maximum likelihood estimation. We will see that the MLE is affected by the contamination in some cases. Hence, employing results that have been obtained for the case of independent and not identically distributed data, we consider robust estimation methods in the context of (3.3) using the following two procedures:

- the Conditionally Unbiased Bounded-Influence estimator (CUBIF), proposed by Künsch et al. (1989),
- the Mallows' Quasi-Likelihood estimator (MQLE), proposed by Cantoni and Ronchetti (2001).

3.2 Methods of Estimation

In this section, we will explain briefly the computation of all estimators considered and we will compare them empirically in the following section.

3.2.1 Maximum Likelihood Estimator (MLE)

The standard approach for estimating the parameters d and $\{b_j, j = 1, \ldots, q\}$ is maximum likelihood estimation. Recall model (3.2). Let $\boldsymbol{\theta}$ be the $q+1$ dimensional vector of unknown parameters $\boldsymbol{\theta} = (d, b_1, ..., b_q)^T$. Then the conditional likelihood function for $\boldsymbol{\theta}$ based on model (3.2) is given by

$$\prod_{t=1}^{n} \frac{\exp(-\lambda_t(\boldsymbol{\theta}))\lambda_t(\boldsymbol{\theta})^{Y_t}}{Y_t!}.$$

Under the model assumptions $\nu_t \equiv \log \lambda_t$, so $\lambda_t(\boldsymbol{\theta}) = \exp(\nu_t(\boldsymbol{\theta}))$. Hence, the log-likelihood function is given up to a constant by

$$\ell_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(\boldsymbol{\theta}) = \sum_{t=1}^{n}(Y_t \log(\lambda_t(\boldsymbol{\theta})) - \lambda_t(\boldsymbol{\theta})) = \sum_{t=1}^{n}(Y_t\nu_t(\boldsymbol{\theta}) - \exp(\nu_t(\boldsymbol{\theta}))).$$

The score function is defined by

$$S_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n}\{Y_t - \exp(\nu_t(\boldsymbol{\theta}))\}\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \qquad (3.4)$$

where $\partial\nu_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is the $q+1$ dimensional vector

$$\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = (\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial d}, \frac{\partial \nu_t(\boldsymbol{\theta})}{\partial b_1}, \ ... \ \frac{\partial \nu_t(\boldsymbol{\theta})}{\partial b_q})^T = (1, \ \ \log(1 + Y_{t-1}), \ ..., \ \log(1 + Y_{t-q}))^T \equiv \tilde{X}_{t-q}.$$

For the case $q = 1$, Fokianos and Tjøstheim (2011) have shown that if $|b_1| < 1$ then a perturbed version of the process (Y_t, ν_t) is geometrically ergodic with moments of any order. In addition, the maximum likelihood estimator given as solution to $S(\boldsymbol{\theta}) = 0$, is consistent and asymptotically normal. These facts are needed for studying the large sample behavior of the proposed estimators, at least in the case $q = 1$.

3.2.2 Conditionally Unbiased Bounded–Influence Estimator (CUBIF)

The idea behind the conditionally unbiased bounded–influence estimator is to find a conditionally Fisher–consistent estimator that has small variance subject to a chosen bound on its influence function; see Künsch et al. (1989). Consider M–estimates of the form

$$\psi(\boldsymbol{\theta}) = \sum_{t=1}^{n} \psi_t(\boldsymbol{Y}_t^{(q)}; \boldsymbol{\theta}, \mathbf{B}) = 0.$$

where $\boldsymbol{Y}_t^{(q)} = (Y_t, Y_{t-1}, ..., Y_{t-q})^T$. In the time series context, the score function for the CUBIF estimator is specified as

$$\begin{aligned} \psi_t(\boldsymbol{Y}_t^{(q)}; \boldsymbol{\theta}, c, \mathbf{B}) &= d(\boldsymbol{Y}_t^{(q)}, \boldsymbol{\theta}, c, \mathbf{B}) \\ &\quad \times W_c(|d(\boldsymbol{Y}_t^{(q)}, \boldsymbol{\theta}, c, \mathbf{B})|(\tilde{X}_{t-q}^T \mathbf{B}^{-1} \tilde{X}_{t-q})^{-1/2}) \tilde{X}_{t-q}^T \end{aligned} \tag{3.5}$$

with

$$d(\boldsymbol{Y}_t^{(q)}, \boldsymbol{\theta}, c, \mathbf{B}) = Y_t - \lambda_t(\boldsymbol{\theta}) - C\left(\nu_t(\boldsymbol{\theta}), \frac{c}{\left(\tilde{X}_{t-q}^T \mathbf{B}^{-1} \tilde{X}_{t-q})^{-1/2}\right)}\right)$$

and $W_c(\alpha) = \psi_c(\alpha)/\alpha$ where $\psi_c(\alpha)$ is the Huber function

$$\psi_c(\alpha) = \begin{cases} \alpha, & |\alpha| \leq c, \\ c\text{sign}(\alpha), & |\alpha| > c, \end{cases} \tag{3.6}$$

with c a tuning constant. The scalar function $C(\cdot)$ and the matrix $\mathbf{B}$ in (3.5) are chosen so that the sensitivity function is bounded and the estimating function to be unbiased, that is $\mathrm{E}(\psi_t(\boldsymbol{Y}_t^{(q)}; \boldsymbol{\theta}, c, \mathbf{B})\psi_t(\boldsymbol{Y}_t^{(q)}; \boldsymbol{\theta}, c, \mathbf{B})^T) = \mathbf{B}$ and $\sum_{t=1}^{n} \mathrm{E}(\psi_t(\boldsymbol{Y}_t^{(q)}; \boldsymbol{\theta}, c, \mathbf{B})||\mathcal{F}_{t-1}) = 0$, respectively - recall the discussion in Section 2.4.1.

3.2.3 Mallows' Quasi-Likelihood Estimator (MQLE)

Cantoni and Ronchetti (2001) robustified the quasi-likelihood approach for estimating the regression coefficient of generalized linear models. Their approach is based on robust deviances which are natural generalizations of the quasi-likelihood functions. The robustification proposed by Cantoni and Ronchetti (2001) is performed by bounding and centering the quasi score function. In the context of model (3.2), the MQLE is given as a solution of the following equation

$$Q(\boldsymbol{\theta}) = \sum_{t=1}^{n}\Big\{\psi_c\Big(\frac{Y_t-\lambda_t(\boldsymbol{\theta})}{\sqrt{\lambda_t(\boldsymbol{\theta})}}\Big)w_t\frac{1}{\sqrt{\lambda_t(\boldsymbol{\theta})}}\frac{\partial\lambda_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \alpha(\boldsymbol{\theta})\Big\} = 0, \tag{3.7}$$

where $\psi_c(\alpha)$ is the Huber function (4.6) and the tuning constant c is chosen to ensure a given level of asymptotic efficiency. The sequence $\{w_t\}$ consist of suitable weights. Some choices for the weights $\{w_t\}$ are $w_t = \sqrt{1-h_{tt}}$ where h_{tt} is the t-th diagonal element of the hat matrix $H = X(X^TX)^{-1}X^T$. Here X is the $\{n-(q+1)\}\times(q+1)$ matrix with the t-th column given by $\tilde{X}_{t-q}$. Following those authors we note that weights defined on the hat matrix do not have high breakdown properties. Therefore it is preferable to consider other weighting schemes that are robust. Such choices are the inverse of the robust Mahalanobis distance where location and scatter are robustly estimated to have high breakdown properties using either the minimum volume ellipsoid (MVE) estimator or the minimum covariance determinant (fast MCD) algorithm (see Rousseeuw and van Zomeren (1990) and Rousseeuw and Driessen (1999), for more). The term $\alpha(\boldsymbol{\theta})$ appearing in (3.7) is a bias correction term which is used to ensure Fisher–consistency. It is given by

$$\alpha(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}\mathrm{E}\left\{\psi_c\Big(\frac{Y_t-\lambda_t(\boldsymbol{\theta})}{\sqrt{\lambda_t(\boldsymbol{\theta})}}\Big)w_t\frac{1}{\sqrt{\lambda_t(\boldsymbol{\theta})}}\frac{\partial\lambda_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|\mathcal{F}_{t-1}\right\},$$

where

$$\begin{aligned}\mathrm{E}\Big(\psi_c\Big(\frac{Y_t-\lambda_t(\boldsymbol{\theta})}{\sqrt{\lambda_t(\boldsymbol{\theta})}}\Big)\Big|\mathcal{F}_{t-1}\Big) &= c\{P(Y_t \geq j_2+1\|\mathcal{F}_{t-1}) - P(Y_t \leq j_1\|\mathcal{F}_{t-1})\}\\ &+ \sqrt{\lambda_t(\boldsymbol{\theta})}\{P(Y_t = j_1\|\mathcal{F}_{t-1}) - P(Y_t = j_2\|\mathcal{F}_{t-1})\},\end{aligned}$$

with j_1 and j_2 are defined as $j_1 = \lfloor\lambda_t(\boldsymbol{\theta}) - c\sqrt{\lambda_t(\boldsymbol{\theta})}\rfloor$ and $j_2 = \lfloor\lambda_t(\boldsymbol{\theta}) + c\sqrt{\lambda_t(\boldsymbol{\theta})}\rfloor$. This can be proved along the lines of Cantoni and Ronchetti (2001, App. A). Notice that as $c\to\infty$, then $MQLE \simeq MLE$, when $w_t \equiv 1$. However, this approximation is no longer true when robust weights are employed for solving (3.7).

3.3 Empirical Results

In this section we compare empirically all three estimators when the data contain different forms of interventions. In particular, the MQLE is computed based on the following weighting schemes (recall (3.7)):

- with no weights,
- with weights defined on the hat matrix,
- with weights obtained by robust Mahalanobis distance by using the minimum volume ellipsoid method (MVE),
- with weights obtained by robust Mahalanobis distance by using the minimum covariance determinant method (MCD).

In all examples below, we run the experiment 1000 times and we consider sample sizes of $n = 200, 500$. We report results with 500 observations since the results from both settings were quite analogous. We are working with model (3.3) with $q = 1, 2, 3$. The estimators are compared empirically in terms of mean square error (MSE), mean absolute error (MAE) and bias. All results have been obtained using R (R Core Team (2014)).

Definition 3.3.1 Mean Square Error (MSE): The MSE of an estimator $\hat{\theta}$ of a parameter θ is $\mathrm{E}\left[(\hat{\theta}-\theta)^2\right]$. The sample MSE is given by $\hat{\mathrm{MSE}} = \sum_{i=1}^{n}(\hat{\theta}_{(i)} - \theta_0)^2/n$, where θ_0 is the true value and $\hat{\theta}_{(i)}$ is a simulated estimate obtained by the simulation.

Definition 3.3.2 Mean Absolute Error (MAE): The MAE of an estimator $\hat{\theta}$ of a parameter θ is $\mathrm{E}|\hat{\theta}-\theta|$. The sample MAE is given by $\hat{\mathrm{MAE}} = \sum_{i=1}^{n}|\hat{\theta}_{(i)} - \theta_0|/n$.

However, to save space we will report only results for the MSE and for some combinations of sample sizes and interventions, since all other quantities yield identical conclusions as those we will present. All plots are constructed by assuming that the intervention occurred at time $\tau = n/4$, but other time points have been considered as well.

3.3.1 Level Shift and Transient Shift

Recall model (3.3). In the examples in this section we focus on the first order model with either a Level Shift or a Transient Shift type of intervention. In the case of the transient shift we consider $\delta = 0.9$.

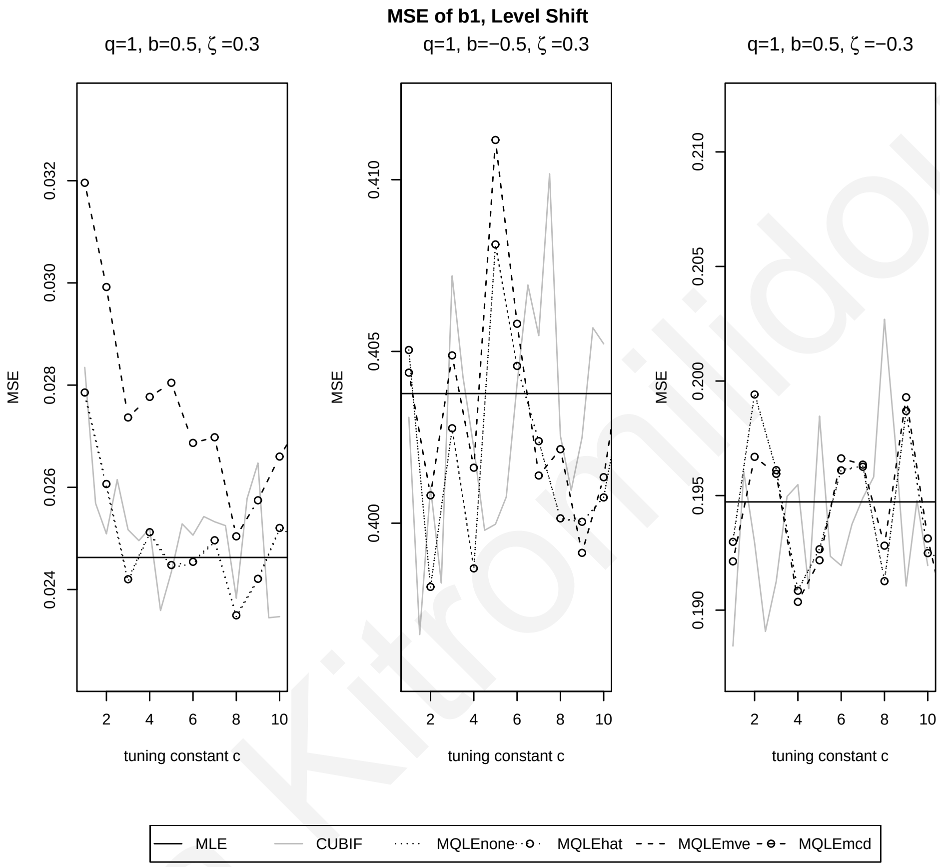

Figure 3.1: MSE for the case of a Level Shift at $\tau = n/4$ for b of the model with $q = 1$ with (a): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = 0.3$, (b): $\theta = (d, b) = (0.2, -0.5)$, $\zeta = 0.3$ (c): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = -0.3$.

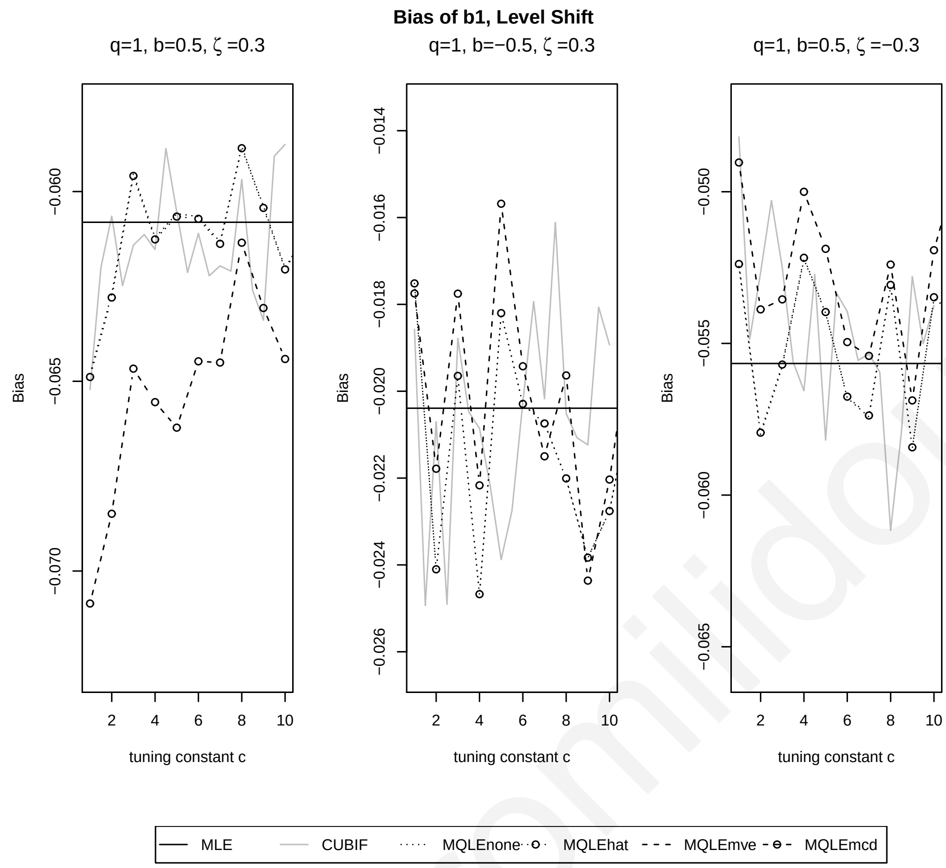

Figure 3.2: Bias for the case of a Level Shift at $\tau = n/4$ for b of the model with $q = 1$ with (a): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = 0.3$, (b): $\theta = (d, b) = (0.2, -0.5)$, $\zeta = 0.3$ (c): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = -0.3$.

Our simulation study on the effect of a LS or TS intervention illustrates that in these cases there do not exist considerable dissimilarities among the proposed estimators in terms of mean square error (MSE).

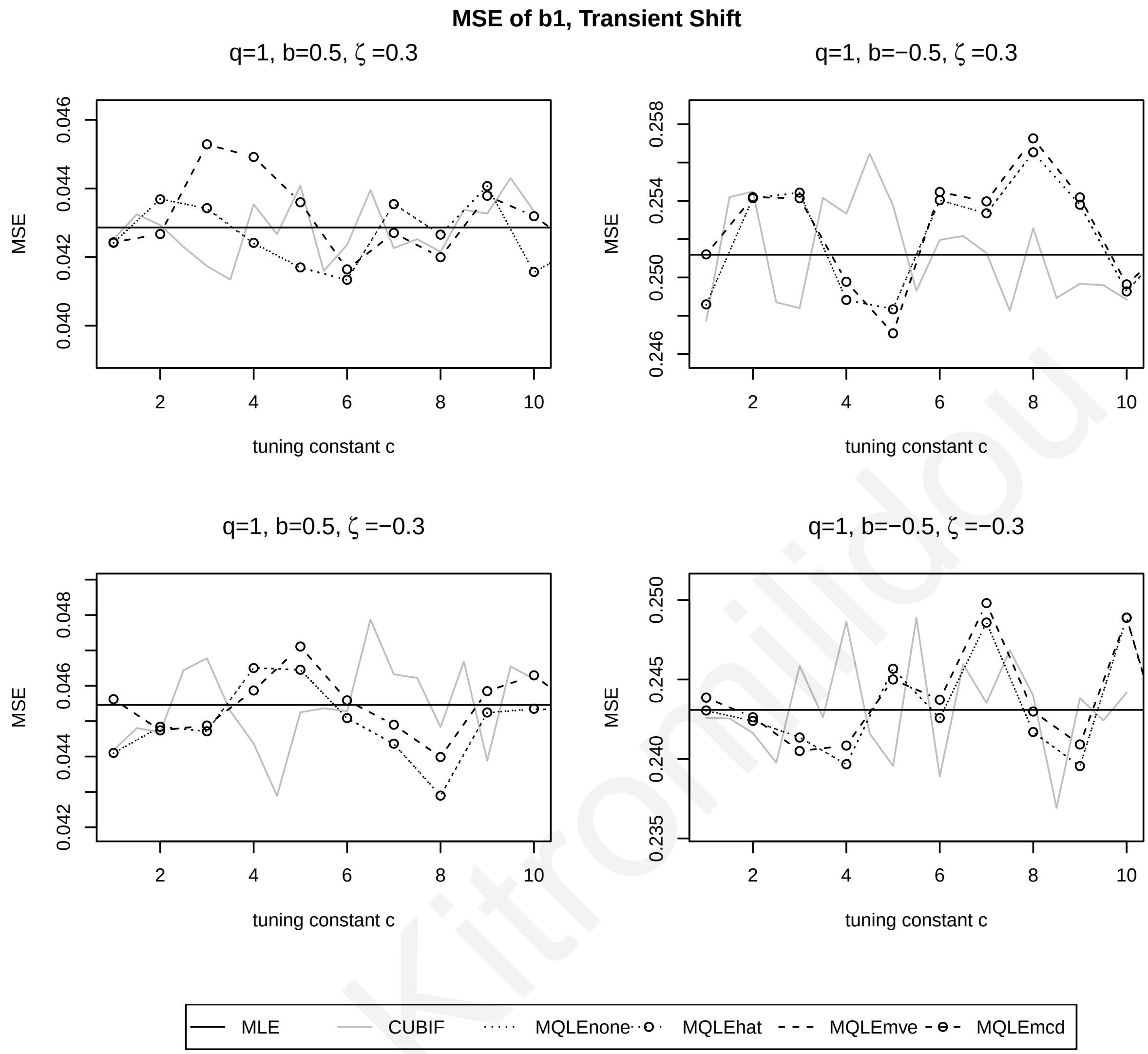

Figure 3.3: MSE for the case of a Transient Shift at $\tau = n/4$ for b of the model with $q = 1$ with (a): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = 0.3$, (b): $\theta = (d, b) = (0.2, -0.5)$, $\zeta = 0.3$ (c): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = -0.3$ and (d): $\theta = (d, b) = (0.2, -0.5)$, $\zeta = -0.3$.

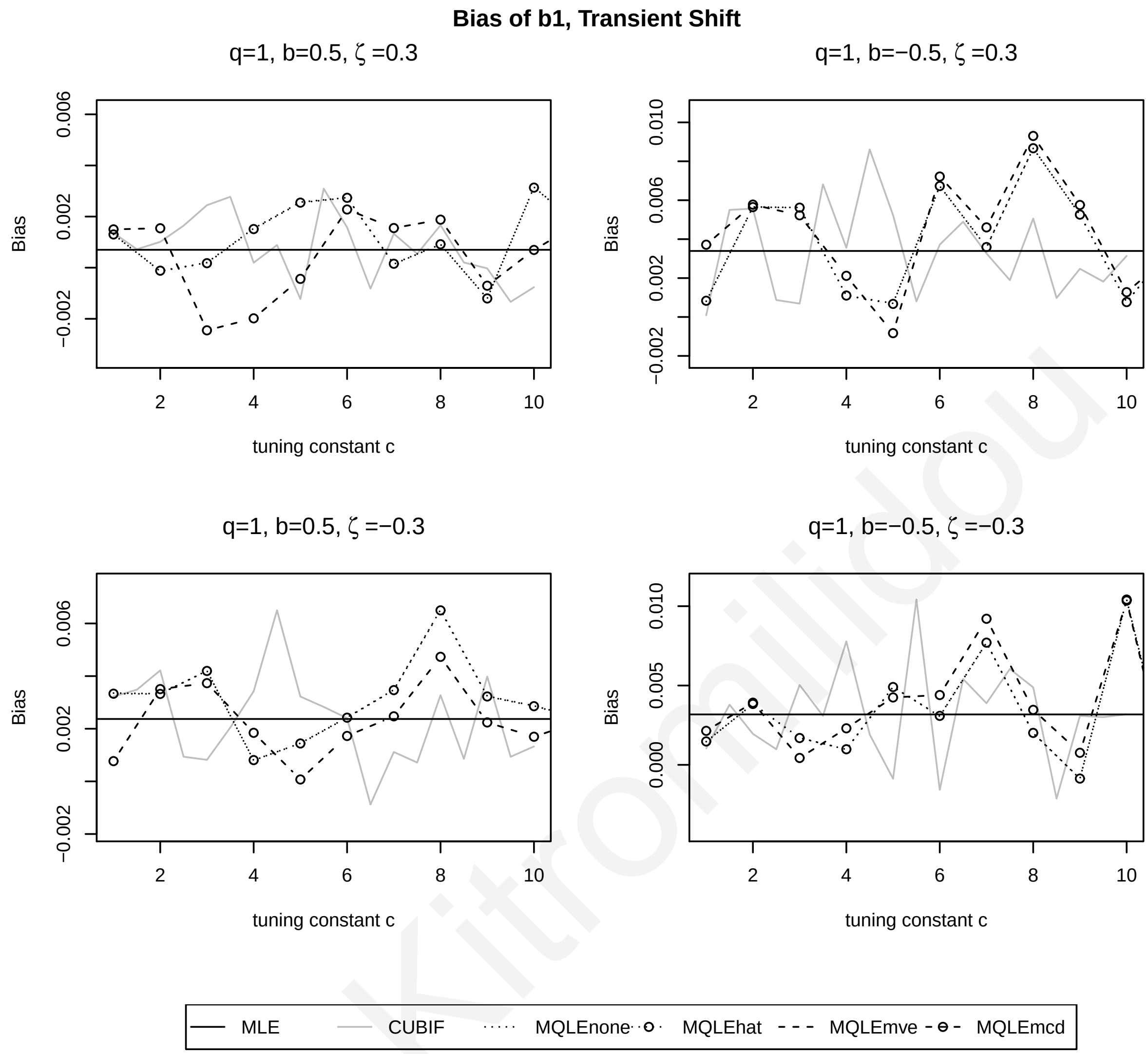

Figure 3.4: Bias for the case of a Transient Shift at $\tau = n/4$ for b of the model with $q = 1$ with (a): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = 0.3$, (b): $\theta = (d, b) = (0.2, -0.5)$, $\zeta = 0.3$ (c): $\theta = (d, b) = (0.2, 0.5)$, $\zeta = -0.3$ and (d): $\theta = (d, b) = (0.2, -0.5)$, $\zeta = -0.3$.

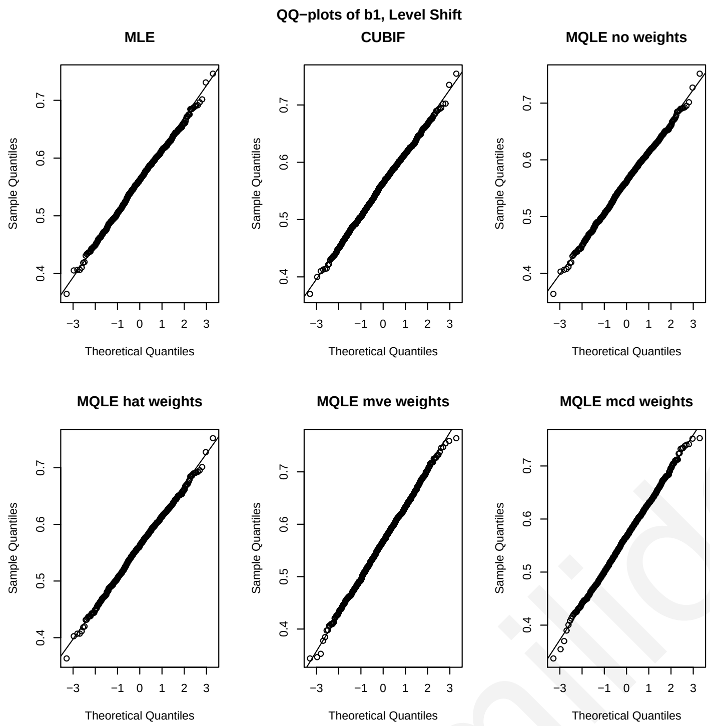

Figure 3.5: QQ-plots for $\hat{b}$ for the case of a Level Shift with $\theta = (d, b) = (0.2, 0.5)$, $\zeta = 0.3$ and $c = 3$.

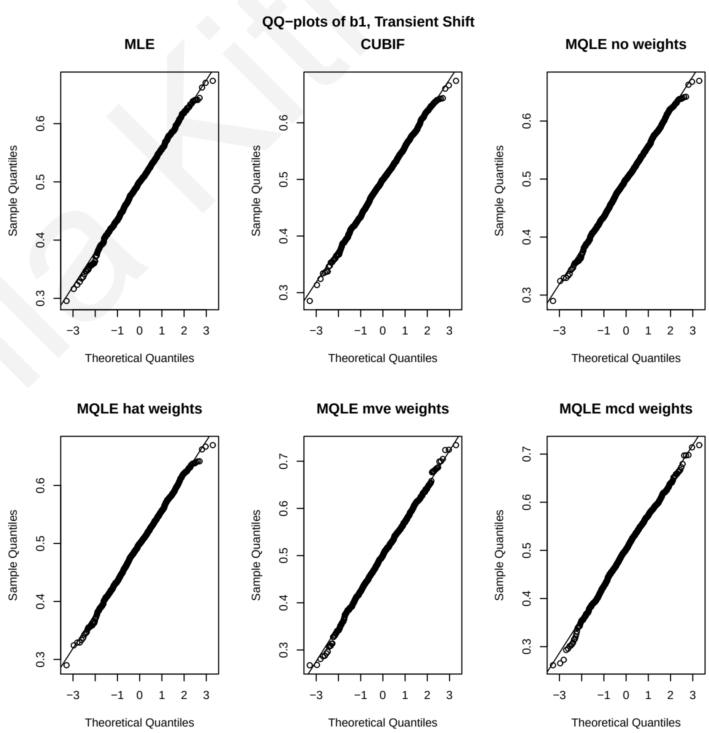

Figure 3.6: QQ-plots for $\hat{b}$ for the case of a Transient Shift with $\theta = (d, b) = (0.2, 0.5)$, $\zeta = 0.3$ and $c = 3$.

3.3.2 Additive Outliers

In the following examples, we examine the effect of large additive outliers in count time series under various scenarios. Note that it is not interesting to test the case when the size of the additive outlier is small because in this case the resulting value of the response could actually be realized by a Poisson random variable. We examine the following three cases:

- Single outlier
- Patch of outliers
- Isolated outliers

Generally the observed contaminated series takes on the form

$$Z_t = \begin{cases} Y_t + \zeta, & t = \tau_1, \tau_2, ..., \tau_k, \\ Y_t, & \text{otherwise}, \end{cases} \tag{3.8}$$

at times $\tau_1, \tau_2, ..., \tau_k$. Note that (3.1) is a special case of (3.8) when $k = 1$.

Single outlier

Recall model (3.8) for an AO type of intervention. When $k = 1$, a single large additive outlier of size ζ is added to the time series. This has an effect on the observed series at time τ and a few lags after that time point. In the following example we consider the first and second order model. The values of the parameters are $\boldsymbol{\theta} = (d, b_1) = (0.2, 0.5)$ for the upper two figures and $\boldsymbol{\theta} = (d, b_1, b_2) = (0.2, 0.3, 0.4)$ for the lower two figures. The size of the outlier ζ takes the values 10 and 30. The first two figures show that in the case of the first order model there are no significant dissimilarities among the estimators and in fact the MLE performs adequately. As the lag increases (see the lower two figures), we observe that the MSE values of the estimates decrease and still no noteworthy differences among the estimators are reported. Similar findings have been obtained for b_2.

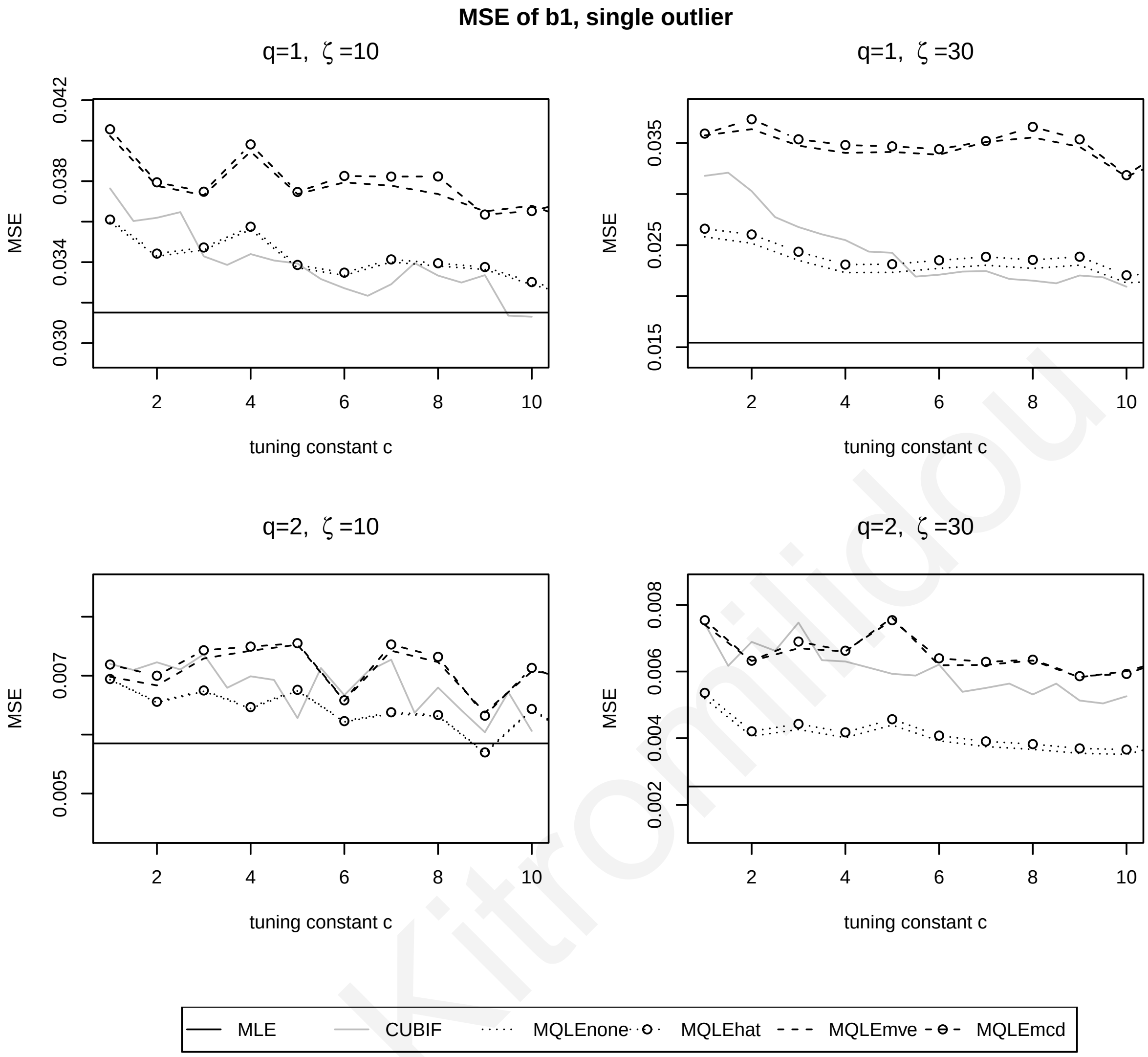

Figure 3.7: MSE for the case of a single outlier for b_1 of the model with $q = 1$ for the first two plots and $q = 2$ for the last two plots.

Patch of outliers

In the case of consecutive interventions, the times $\tau_1, \tau_2, ..., \tau_k$ are contiguous. Hence, if for instance we have a time series of size $n = 500$, $k = 15$, $\tau = 125$ and $\zeta = 10$, then we will observe fifteen consecutive additive outliers of size $\zeta = 10$ occurring at times $t = 125, 126, ..., 139$. Figure 3.8 shows results from the first order model with $\theta = (d, b_1) = (0.2, 0.5)$. The size of the intervention is $\zeta = 10$ but the same conclusions can be drawn for larger outlier sizes, particularly for $\zeta = 20, 30$. It is clear that the best estimator, in terms of MSE, is the robustly weighted MQLE regardless of the number of consecutive outliers.

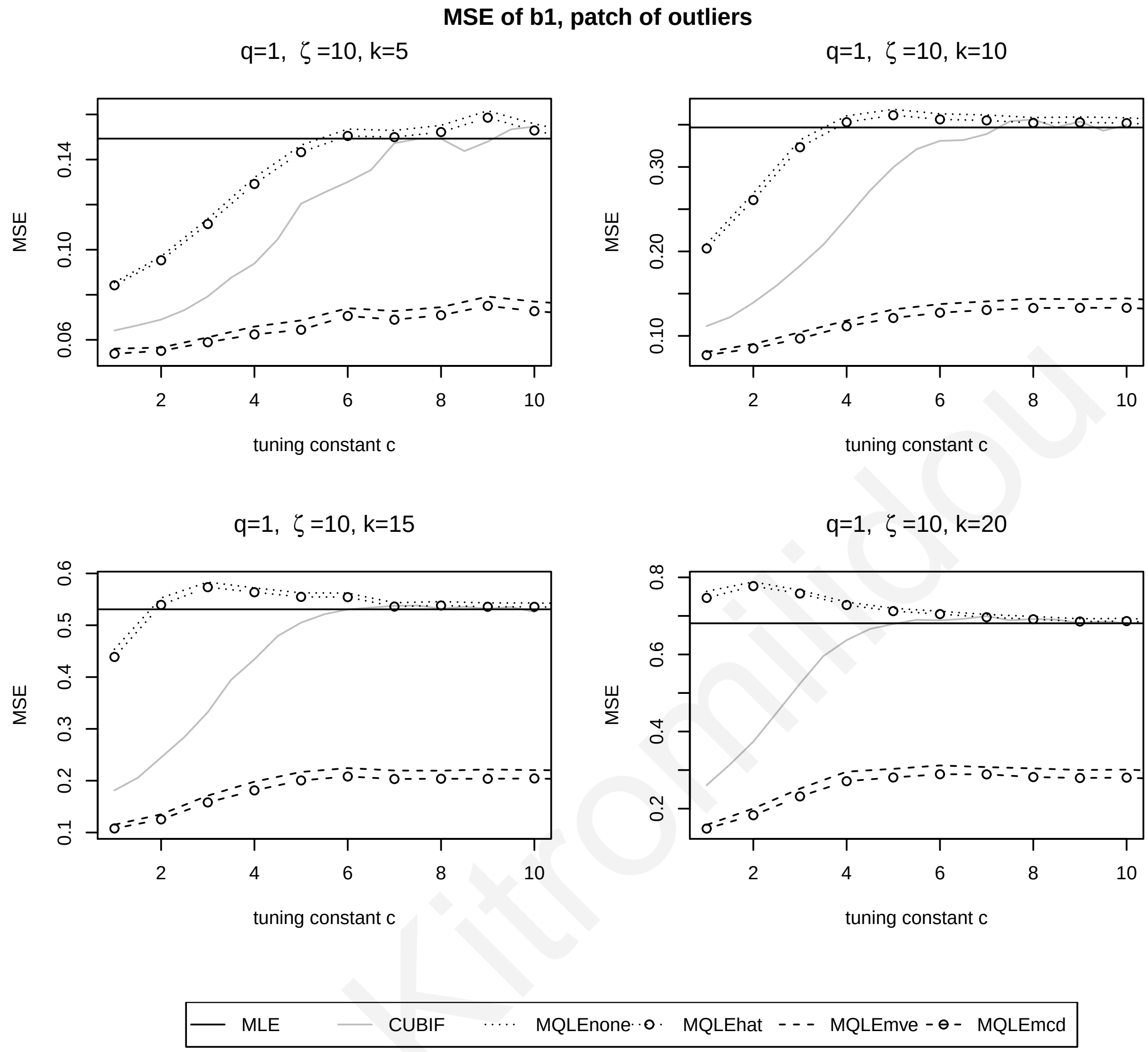

Figure 3.8: MSE for the case of a patch of outliers for the model with $q = 1$ and $\zeta = 10$.

In the case of the second order model, we have discovered that the same results hold as in the case of the first order model for both b_1 and b_2 as well as for various intervention sizes. Additionally, it is observed that the MSE values decrease as the lag increases.

Figures 3.9 and 3.10 correspond to the third order model with $\theta = (d, b_1, b_2, b_3) = (0.2, 0.2, 0.3, 0.4)$ and we report the results for two outlier sizes, $\zeta = 10$ and $\zeta = 30$ and for b_1, although the same results hold for b_2 and b_3. For this example, note that when the size of the intervention is relatively small, that is $\zeta = 10$, then no remarkable contrasts are revealed. However, increasing the size of the intervention and as the number of consecutive outliers increases, then the estimator that performs better, in terms of MSE, than all others is the robustly weighted MQLE. In particular, we note that the

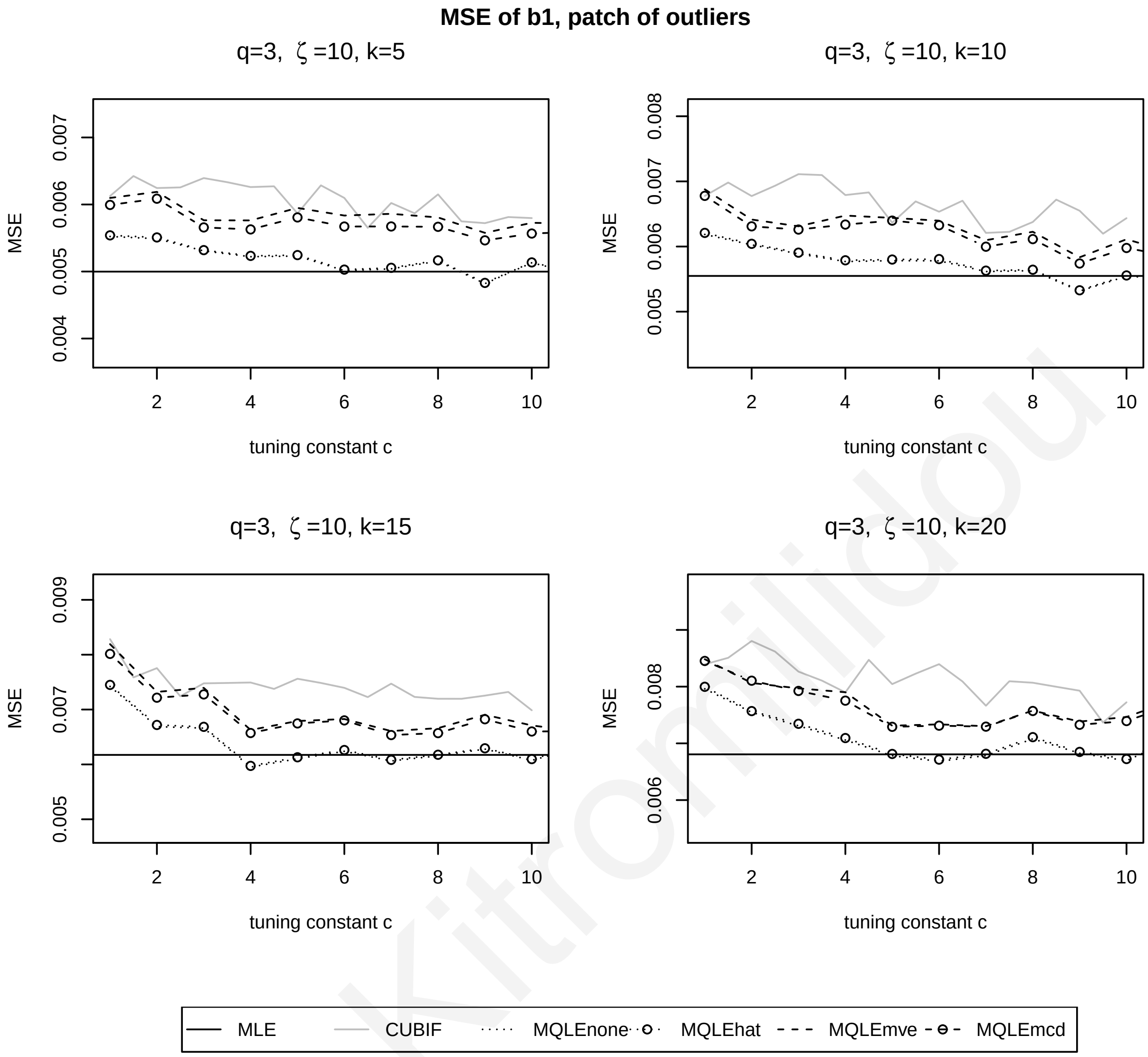

Figure 3.9: MSE for the case of a patch of k outliers for the model with $q = 3$ for b_1.

weights based on the MVE and the fast MCD algorithms yield an estimator with smaller MSE. An additional conclusion in this case is that the CUBIF and the MQLE estimators are competitive for small values of c but only in the case of adding a few outliers to the series. Increasing further the number of consecutive outliers, the differences between these three estimators increase substantially. Furthermore, we point out that as the number of consecutive outliers increases the regression coefficient of MLE, CUBIF and the non-robustly weighted MQLE's increase towards one.

Figures 3.11 and 3.12 exhibit boxplots and qq-plots for $\hat{b}_1$ in the case of the second order model with 10 consecutive outliers of size 10 and for $c = 3$. Both from the boxplots as well as from the qq-plots we conjecture the asymptotic normality of all estimates.

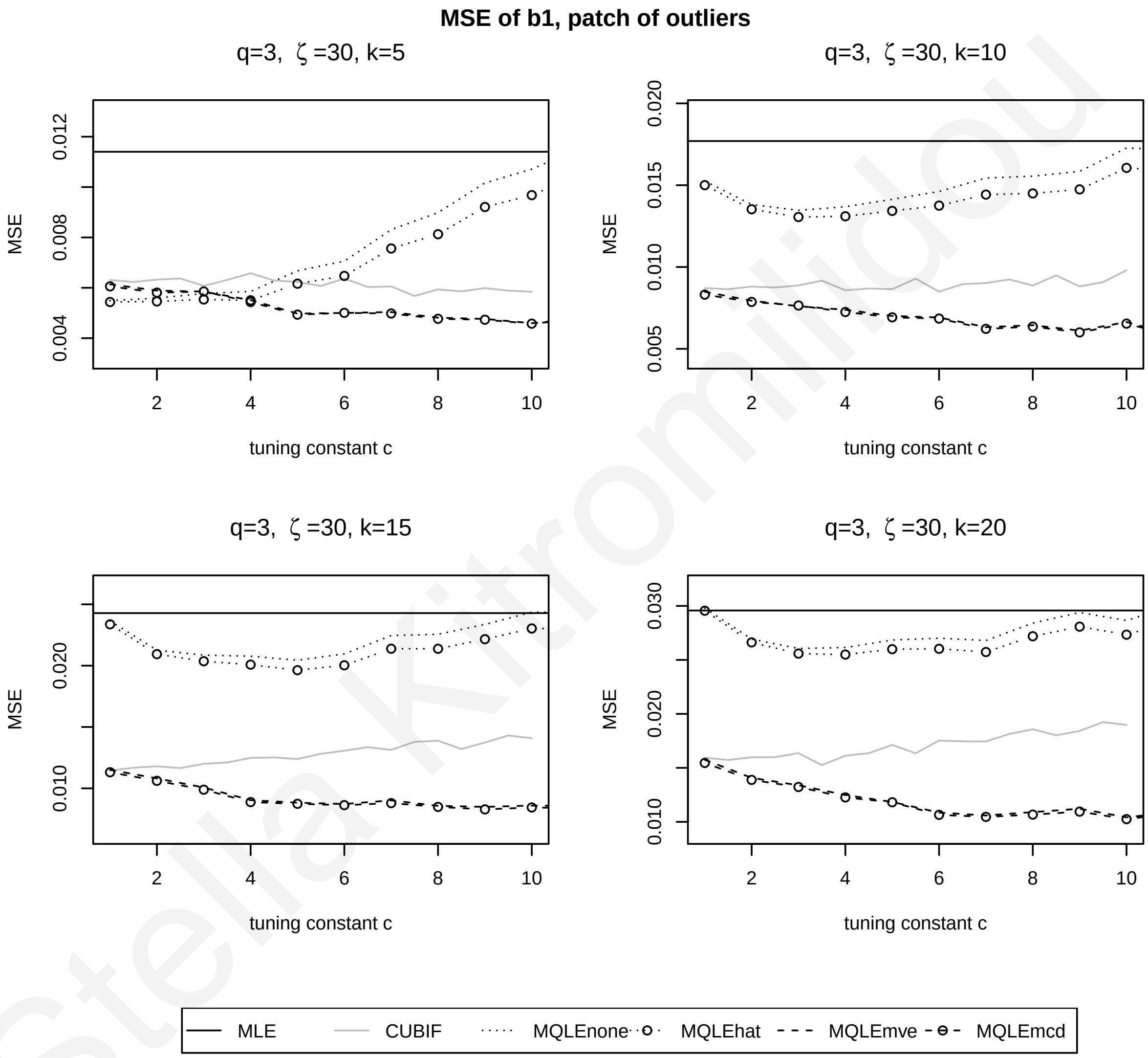

Figure 3.10: MSE for the case of a patch of k outliers for the model with $q = 3$ for b_1.

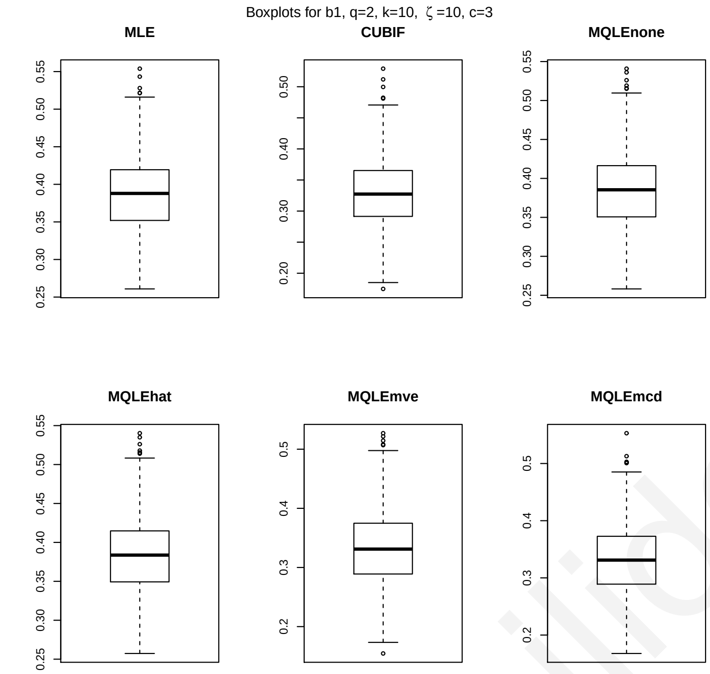

Figure 3.11: Boxplots for $\hat{b}_1$ for the case of the second order model with 10 consecutive outliers of size 10 and for $c = 3$.

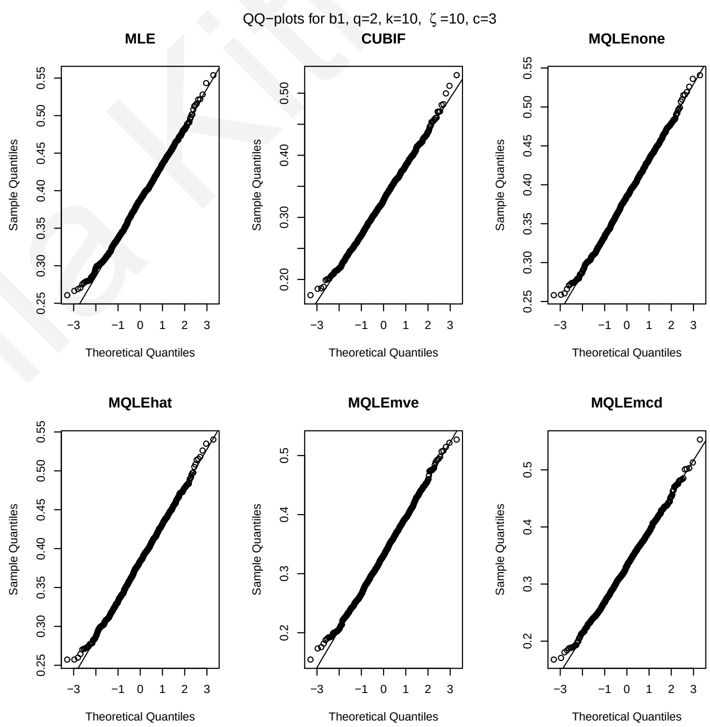

Figure 3.12: QQ-plots for $\hat{b}_1$ for the case of the second order model with 10 consecutive outliers of size 10 and for $c = 3$.

Isolated outliers

Next we examine the case of several isolated large additive outliers. Recall model (3.8) for the additive interventions and set $k = 5, 10, 15, 20$ and suppose that the size ζ are equal to $10, 20, 30$. The outliers are added to the observed data in randomly chosen positions. Figure 3.13 corresponds to the case of the first order model and Figures 3.14 and 3.15 correspond to the case of the third order model. In the case of the first order model, see Figure 3.13, it is clear that when MQLE is robustly weighted, and in particular using the MVE or MCD algorithm, then it is the best estimate as the number of outliers increases. In addition, the difference between robustly weighted MQLE and MLE is of substantial order.

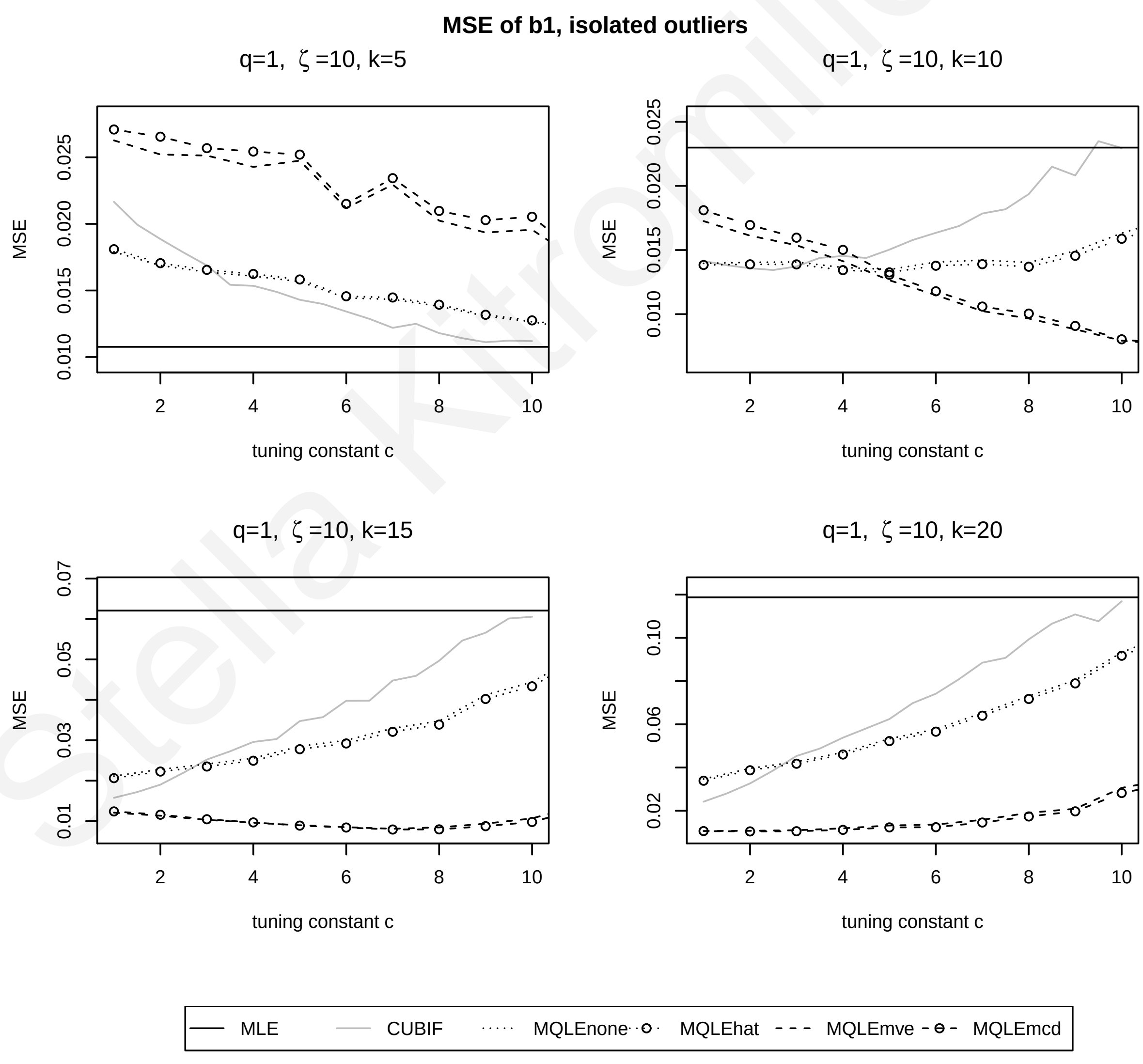

Figure 3.13: MSE for the case of k isolated outliers for the model with $q = 1$ for b_1.

The case of the second order model is not presented though our findings suggest that robustly weighted MQLE performs better. Figure 3.14 shows the MSE of the estimation of b_1 of the third order model when the size of the outliers is $\zeta = 10$ and Figure 3.15 the corresponding when the size of the outliers is increased to $\zeta = 30$. Firstly, it is noticed that increase of the lag yields a decrease in the values of the MSE. Also, we note that when the size of the interventions is $\zeta = 10$ then the differences between the estimators are not substantial. However, increasing the size of the outliers, see Figure 3.15, shows that MQLE when robustly weighted dominates the other estimators, although CUBIF is also competitive.

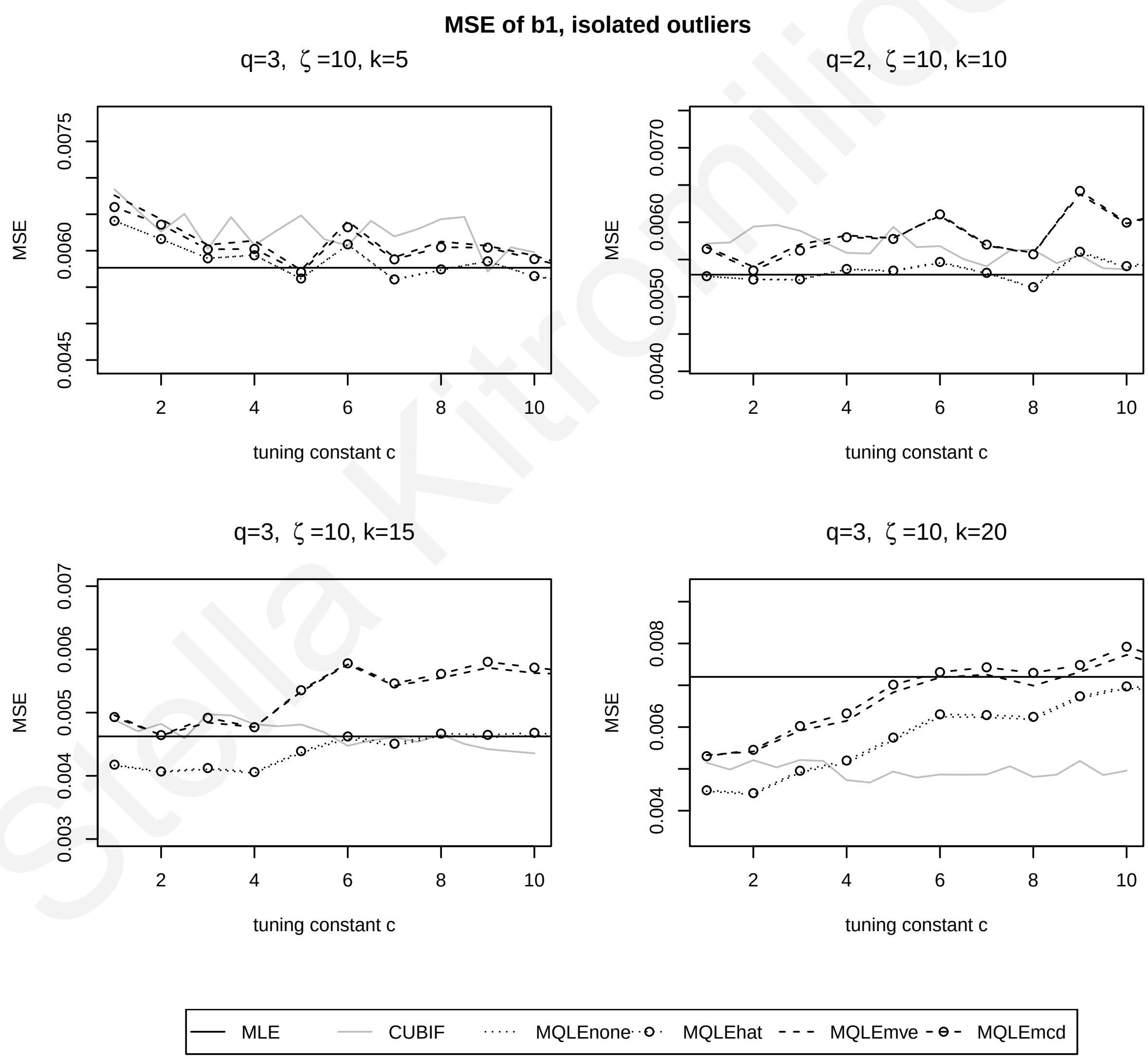

Figure 3.14: MSE for the case of k isolated outliers for the model with $q = 3$ for b_1 and $\zeta = 10$.

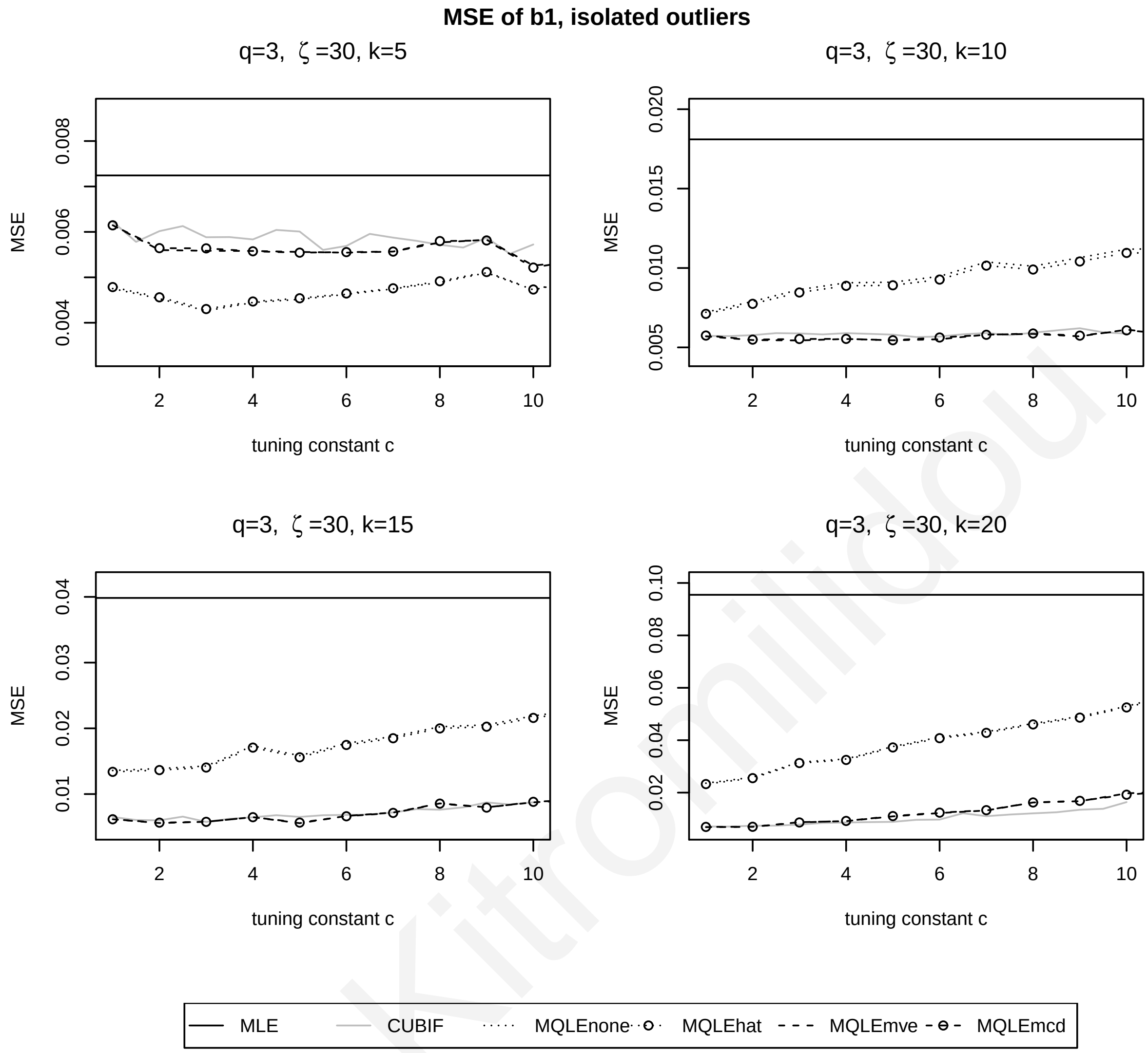

Figure 3.15: MSE for the case of k isolated outliers for the model with $q = 3$ for b_1 and $\zeta = 30$.

3.3.3 Choice of the Tuning Constant

A fundamental aspect in implementing robust method is to calculate a proper value of c to achieve a certain optimality criteria. The tuning constant c is chosen so that a predefined level of relative efficiency is ensured. The MSE of the robust estimate is thus compared to the MSE of MLE and this should yield

$$\frac{\text{MSE}(\hat{\theta}_{\text{MLE}})}{\text{MSE}(\hat{\theta}_{\text{robust}})} \simeq 0.95$$

where $\hat{\theta}_{\text{robust}}$ is any estimator defined by the estimating equations (3.5) and (3.7). In the following simulation example, 1000 samples of clean data, that is data without outliers, are generated according to model (3.2). The model parameters are estimated for numerous values of c varying between 1 and 10 and the relative efficiency is computed for all estimation procedures.

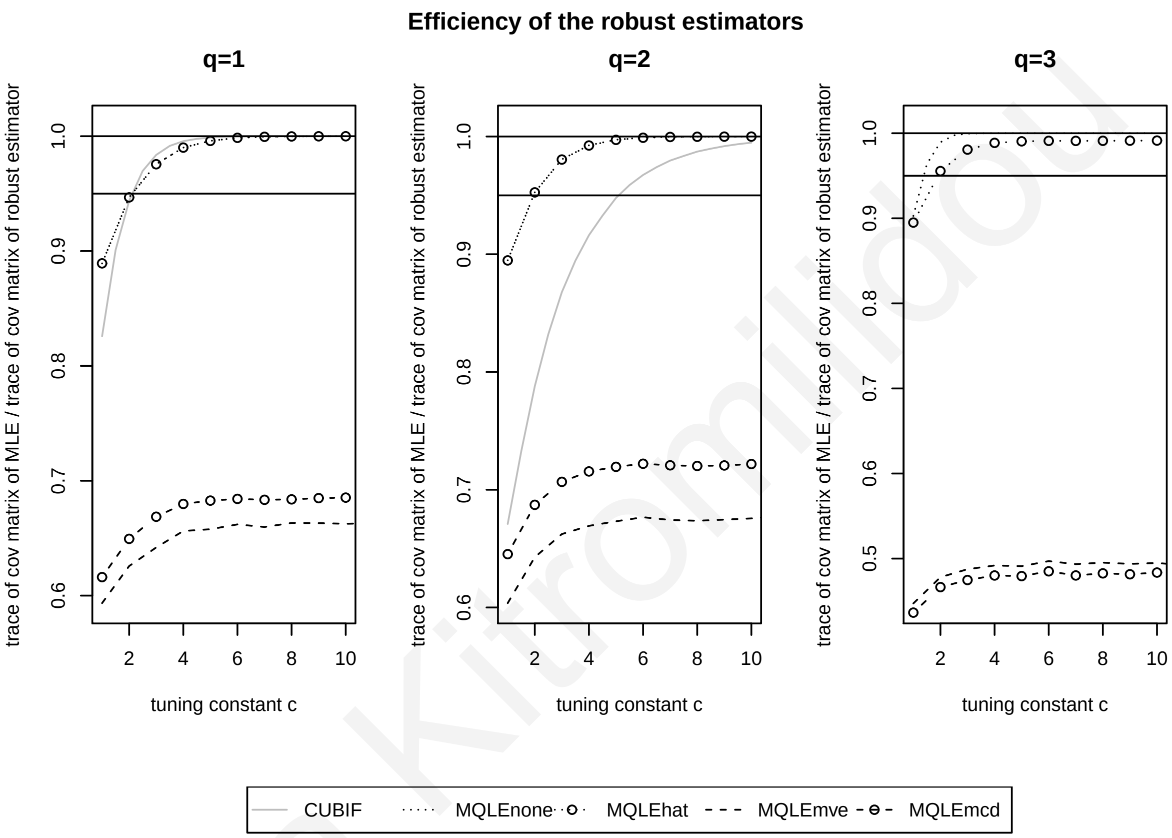

Figure 3.16: Efficiency and choice of c.

Figure 3.16 displays the relative efficiency of all estimation methods in three cases corresponding to the first, second and third order model. In all cases, MQLE robustly weighted is significantly less efficient than MLE. On the contrary, the non robustly weighted MQLE possesses an extremely high level of efficiency, that actually reaches the efficiency of MLE as the value of c increases. This is somewhat anticipated since the MQLE which is not robustly weighted converges to the MLE as the value of the tuning constant increases. It is recommended that the most appropriate c value that ensures a 95% level of efficiency is between 1.5 and 2 for the first order model and between 1 and 1.5 for the higher order models. In the next section, we study an additional method of obtaining c by splitting the data into a training and a testing dataset.

3.3.4 Summary

We have compared several robust estimators for various form of interventions for log-linear models of the form (3.2). We note that when a LS or a TS type of intervention is detected, then all estimators behave in the same way, in terms of MSE, MAE and bias. However, in the case of large AO which occur either isolated or consecutively, then, in general, the robustly weighted MQLE will dominate all the other estimators in terms of MSE.

3.4 Real Data Examples

3.4.1 Polio Data

The polio data is a well known count time series example which has been used in several applied works. It consists of monthly number of incidents of poliomyelitis in USA during the years 1970 to 1983. The data have been released by the U.S. Center for Disease and Control and there are total of $n = 168$ observations. A plot of the data is displayed in Chapter 1. These data have been previously analyzed by Zeger (1988), Fokianos (2001) and Fokianos and Fried (2012), among others. Fokianos and Fried (2012) studied the data by fitting a log-linear model and investigated for possible intervention effects due to unusual events. Their analysis revealed several intervention effects including three spiky outliers and a level shift. We consider a different approach by applying the log-linear model (3.2) of order $q = 6$. A long-term decrease of the incidence rate might exist, and so a trend of the form t/n is also included in the model. Moreover, since the observations are recorded monthly, we include sinusoid terms to model annual seasonality. The fitted model is given by

$$\nu_t = d + \sum_{j=1}^{q} b_j \log(1 + Y_{t-j}) + \beta t/n + \sum_{s=1}^{S} \{\beta_{1;s} \sin(\omega_s t) + \beta_{2;s} \cos(\omega_s t)\}$$

where S is the number of harmonics and $\omega_s = 2\pi s/12$ are the Fourier frequencies. We fit different models in the data using the maximum likelihood estimate. The first six observations are excluded from the fit to ensure comparability among the models and the last ten observations are excluded for prediction purposes. The models are compared in terms of AIC; see Table 3.1.

Model	1	2	3	4	5	6
q	1	1	2	2	3	3
S	1	2	1	2	1	2
AIC	506.661	498.756	506.458	497.277	502.224	495.266
Model	7	8	9	10	11	12
q	4	4	5	5	6	6
S	1	2	1	2	1	2
AIC	500.206	493.941	495.109	490.965	495.653	492.59

Table 3.1: Table of the AIC of different fitted models for the polio data

As we can see from Table 3.1, the chosen model according to the Akaike information criterion is model 10; i.e.

$$\nu_t = d + \sum_{j=1}^{5} b_j \log(1 + Y_{t-j}) + \beta t/n + \sum_{s=1}^{2} \{\beta_{1;s} \sin(\omega_s t) + \beta_{2;s} \cos(\omega_s t)\}. \tag{3.9}$$

The next step in our analysis is to fit the chosen model by employing the proposed robust estimation methods. The estimators are compared in terms of their mean square error (MSE) and mean absolute error (MAE). In the cases of CUBIF and MQLE, we fit the model for fifty different values of the tuning constant c taking values from 1 to 3.5. Figure 3.17 demonstrates the estimated MSE and MAE of the predicted values for the estimation methods considered. From both graphs it is clear that CUBIF is the estimator whose predicted values posses the highest MSE and MAE values. In fact, these values are much greater than the corresponding ones of the MLE, making CUBIF the least favorable estimator among the proposed estimating procedures, even though it is a robust estimation method. Additionally, among the MQLE alternatives, the two robust weighting options of MVE and MCD based weights are preferred although the procedures do not reflect significant differences. We choose the tuning constant c for the Huber function (4.6) by predicting the last ten observations which were excluded from the analysis and calculating the associated MSE or MAE; the value of c that minimizes the MSE or MAE is the choice for the tuning constant; see Table 3.2. The results illustrate that the MQLE based method performs better, especially when it is robustly weighted using the MCD based weights.

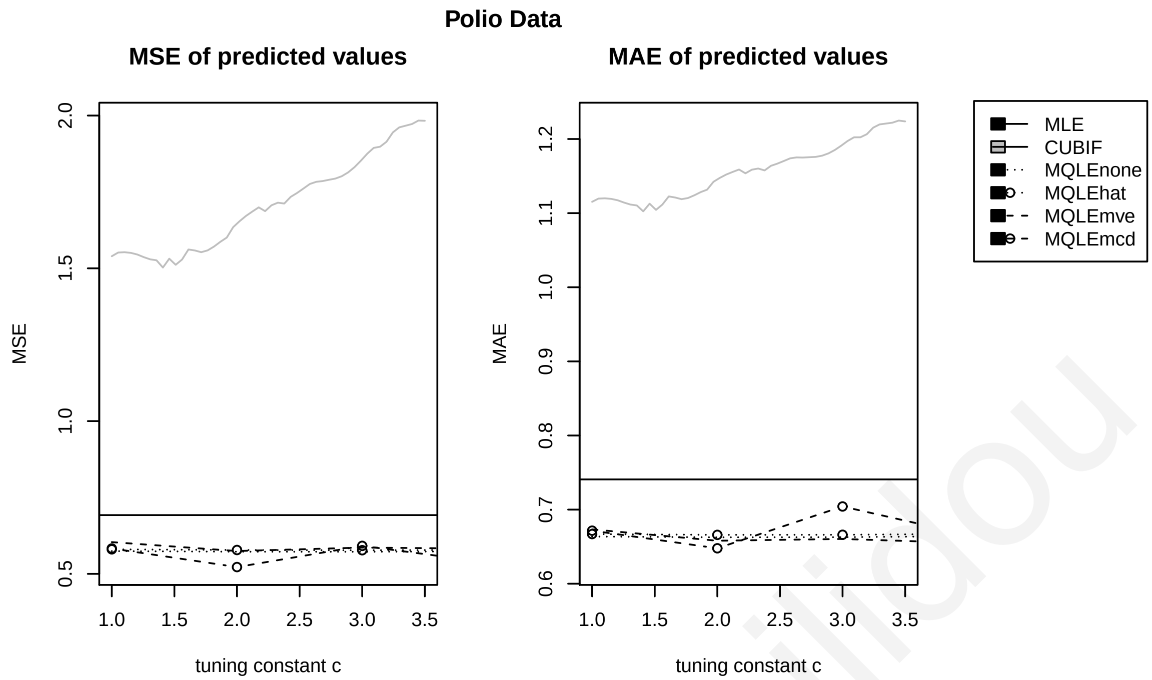

Figure 3.17: Estimated Mean Square Error (MSE) and Mean Absolute Error (MAE) of the predicted values for the Polio data.

	MLE	CUBIF	MQLE	MQLEhat	MQLEmve	MQLEmcd
c	–	1.408	1.051	1.102	3.041	1.051
MSE	0.692	1.503	0.572	0.577	0.535	0.522
c	–	1.408	1.051	1.051	3.041	1.664
MAE	0.741	1.102	0.663	0.666	0.622	0.629

Table 3.2: Minimum MSE and MAE of the estimators and the corresponding value of the tuning constant c for the prediction of the last ten observations of the polio data.

Table 3.3 reports the estimated values of the parameters of model (3.9) under the various proposed methods. The tuning constant c has been chosen according to Table 3.2 and the number in the parentheses is the standard error of the corresponding estimate.

3.4.2 Hyde Park Data

The Hyde Park data consist of the number of purse snatching in the neighborhood Hyde Park of Chicago, Illinois, recorded every 28 days during the period from January 1969 to May 1974. There is a total of $n = 71$ observations. During the 42nd period a community crime prevention program was launched, namely the WhistleStop Operation (see McCleary et al. (1980)). Previous analyses of the data, mostly in the Social Sciences, were

	MLE	CUBIF	MQLEnone	MQLEhat	MQLEmve	MQLEmcd
d	-0.121(0.239)	-0.164(0.316)	-0.192(0.268)	-0.201(0.269)	-0.230(0.296)	-0.136(0.397)
b_1	0.421(0.118)	-0.036(0.164)	0.112(0.137)	0.102(0.138)	0.107(0.147)	0.223(0.192)
b_2	0.289(0.131)	0.427(0.163)	0.333(0.143)	-0.034(0.145)	0.406(0.154)	0.276(0.198)
b_3	-0.336(0.131)	0.067(0.168)	-0.040(0.145)	-0.034(0.145)	-0.099(0.161)	0.301(0.205)
b_4	0.164(0.120)	0.283(0.169)	0.241(0.141)	0.255(0.142)	0.250(0.154)	0.429(0.209)
b_5	0.270(0.120)	-0.197(0.171)	0.027(0.142)	0.016(0.142)	0.124(0.155)	-0.121(0.219)
β	-0.585(0.272)	-0.421(0.369)	-0.454(0.310)	-0.444(0.311)	-0.684(0.336)	-0.454(0.431)
$\beta_{1;1}$	-0.441(0.123)	-0.242(0.149)	-0.419(0.136)	-0.415(0.136)	-0.579(0.172)	-0.224(0.185)
$\beta_{1;2}$	-0.015(0.102)	0.066(0.141)	0.051(0.121)	0.065(0.122)	0.152(0.138)	0.033(0.162)
$\beta_{2;1}$	-0.122(0.107)	0.028(0.140)	0.032(0.123)	0.041(0.124)	0.171(0.141)	0.127(0.164)
$\beta_{2;2}$	0.283(0.109)	0.185(0.138)	0.245(0.123)	0.238(0.123)	0.303(0.135)	0.029(0.175)

Table 3.3: Estimates (standard errors) of the parameters of model (3.9) based on the proposed robust estimation methods. The tuning constant c has been chosen according to Table 3.2.

concentrated in evaluating whether the WhistleStop program resulted in a reduction of purse snatching in the area. The direction we are interested in is the analysis of the data using the log-linear model with intervention effects. The data are shown in Figure 3.18.

We use the outlier detection method of Chen and Liu (1993) and Fokianos and Fried (2010, 2012) to detect intervention effects in the data. The method of Chen and Liu (1993) revealed five interventions and outliers. In particular, four transient shifts at times $15, 27, 33, 35$ and an additive outlier at time 43 are found. The method of Fokianos and Fried (2010, 2012) has detected two transient shifts at times 23 and 35 and a spiky outlier at time 33 when a log-linear model without feedback of order $q = 4$ is applied. We apply the log-linear model without feedback (3.2) of order $q = 4$ and exclude the last ten observations for prediction. To choose the best model that fits the data we compare the models using the Akaike information criteria (AIC); see Table 3.4. The table suggests that the chosen model based on the AIC criterion is the log-linear model of order 2, that is model

$$\nu_t = d + a_1 \log(1 + Y_{t-1}) + a_2 \log(1 + Y_{t-2}). \quad (3.10)$$

	q			
Model	1	2	3	4
AIC	430.255	410.144	411.334	412.768

Table 3.4: Table of the AIC of different fitted models for the Hyde Park data

Next, the model (3.10) is fitted to the data and the parameters of the model are

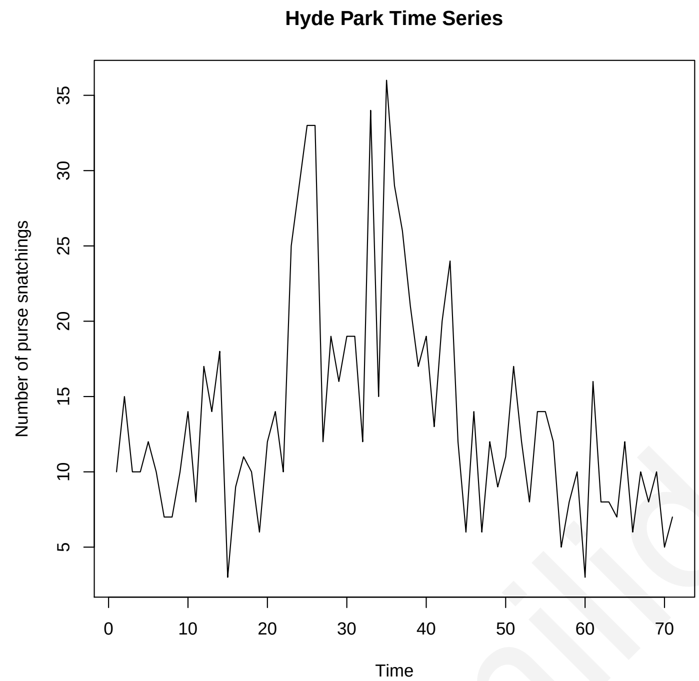

Figure 3.18: Hyde Park Time Series Plot.

estimated using the proposed robust estimates. The CUBIF and MQLE estimates are calculated for values of the tuning constant c of the Huber function from 1 to 3.5. The last test observations that were excluded are then predicted using the fitted model. We calculate the mean square error (MSE) and mean absolute error (MAE) of the predicted values and choose the value of c which minimizes these quantities. The MSE and MAE of the predicted observations are reported in Table 3.5.

	MLE	CUBIF	MQLE	MQLEhat	MQLEmve	MQLEmcd
c	–	1	1.765	1.765	2.327	2.327
MSE	16.170	15.659	15.489	15.085	13.757	14.105
c	–	1	1.663	2.378	3.092	3.449
MAE	3.352	3.353	3.319	3.265	2.999	3.117

Table 3.5: Minimum MSE and MAE of the estimators and the corresponding value of the tuning constant c for the prediction of the last ten observations of the Hyde Park data.

The estimated MSE and MAE of the predicted values are shown in Figure 3.19. Table 3.5 and Figure 3.19 indicate that the robustly weighted MQLE methods outperform both the maximum likelihood and CUBIF estimates. Also, the weighting procedure of the MVE algorithm is somewhat better than the alternative of the MCD algorithm in terms of MSE and MAE.

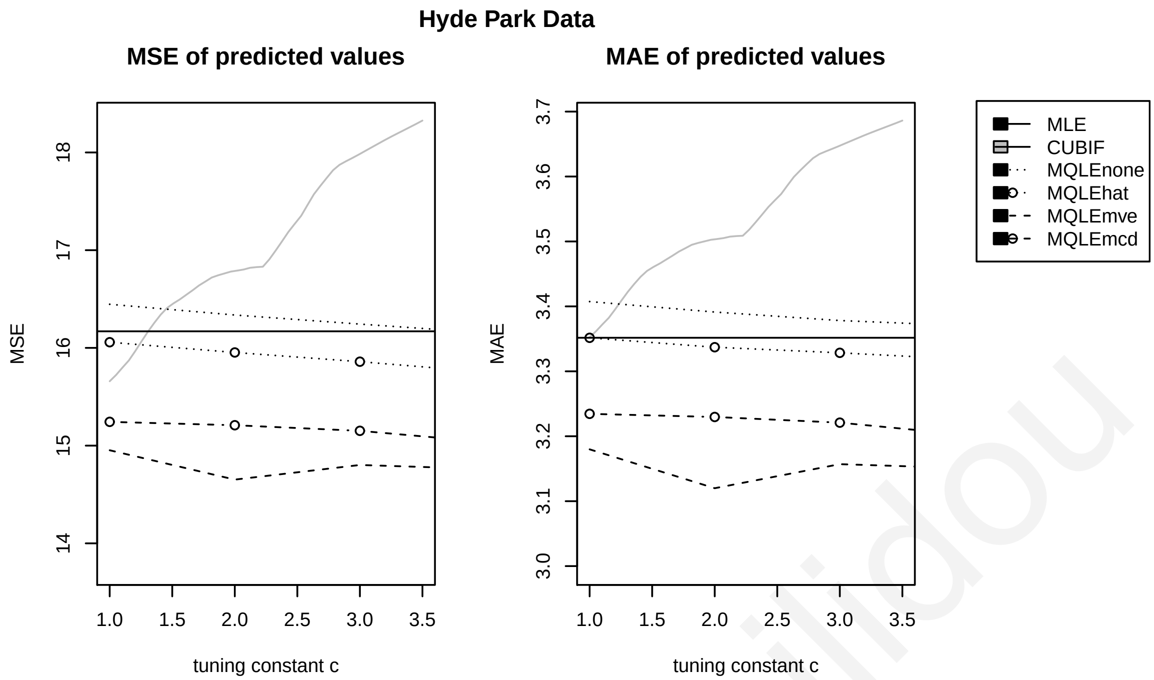

Figure 3.19: Estimated Mean Square Error (MSE) and Mean Absolute Error (MAE) of the predicted values for the Hyde Park data.

Finally, Table 3.6 shows the parameter estimates of the fitted model and their standard errors in parenthesis. The value of the tuning constant c has been chosen according to the MSE values in Table 3.5. In Appendix A, we give the details of this analysis based on the software R (R Core Team (2014)).

	MLE	CUBIF	MQLEnone	MQLEhat	MQLEmve	MQLEmcd
d	0.849(0.230)	0.789(0.315)	0.824(0.234)	0.816(0.237)	0.814(0.277)	0.795(0.265)
b_1	0.295(0.079)	0.218(0.110)	0.333(0.081)	0.325(0.082)	0.274(0.104)	0.324(0.096)
b_2	0.389(0.083)	0.450(0.115)	0.354(0.085)	0.362(0.086)	0.399(0.102)	0.359(0.098)

Table 3.6: Estimates (standard errors) of the parameters of model (3.10) based on the proposed robust estimation methods. The tuning constant c has been chosen according to the MSE in Table 3.5.

Chapter 4

Log-Linear Poisson Model With Feedback

In the previous chapter we studied robust estimation for the log-linear Poisson model without feedback, for contaminated data with level shifts, transient shifts and additive outliers. It turns out that the case of additive outliers deserves special attention. Additionally, among the robust estimators we have examined, the Mallows' Quasi Likelihood estimator (MQLE) appears to be the most prominent, especially when robustly weighted.

In this chapter, we consider the problems of robust estimation and testing for the log-linear Poisson model with feedback for the analysis of count time series. We focus on the first order model with feedback and we propose a robust method for estimating the regression coefficients in the presence of interventions. The resulting robust estimators are asymptotically normally distributed under some regularity conditions. A robust score type test statistic is also examined. We apply the methodology to real and simulated data.

The log-linear Poisson model with feedback studied in this chapter is expected to be more parsimonious than a model that does not include the feedback mechanism, like the one studied in the previous chapter. Additionally, the setup behind the log-linear Poisson model makes it suitable for modeling both negatively and positively correlated count data and allows the inclusion of time dependent covariates. Fokianos and Tjøstheim (2011) have studied extensively the log-linear Poisson model and have examined in detail maximum likelihood estimation by employing a perturbation technique. See also Woodard et al. (2011) and Douc et al. (2013) for related studies. However, the maximum likelihood estimator is highly affected by interventions which enter the log-mean of the process as "unusual" observations, in a variety of ways. These interventions are typically modeled as covariates - as suggested by Fokianos and Fried (2010, 2012) - and can affect the entire time series or a few points, as we have seen in previous sections. Intervention types of this kind are, for example, level shifts (LS) and transient shifts (TS), see Fokianos and Fried (2010, 2012) and more recently Elsaied and Fried (2014). We have seen however that for these types of intervention effects there are no considerable differences among the robust estimation methods and maximum likelihood when a log-linear model without feedback is employed. Adversely, the case of additive outliers (AO) to the process requires more attention. In a different context than what we will be studying, Barczy et al. (2012) consider AO for a first order integer autoregressive model (INAR(1)).

In this chapter, we consider the Mallows' Quasi Likelihood Estimator (MQLE) proposed by Cantoni and Ronchetti (2001). For alternative robust estimation procedures in the context of generalized linear models, see Morgenthaler (1992), Lô and Ronchetti (2009) and Valdora and Yohai (2014), among others. Lô and Ronchetti (2009) propose a robust test statistic for hypothesis testing and variable selection. They show that the test statistic they propose maintains the level of the test and additionally it is asymptotically χ^2. Valdora and Yohai (2014) suggest a class of robust M-estimators for generalized linear models based on transformations of the response that aim at stabilizing the variance of the response to a nearly constant value. The proposed estimates are consistent and asymptotically normally distributed. A general theory of robust statistics is discussed by Maronna et al. (2006) and Huber and Ronchetti (2009). We follow the approach by Cantoni and Ronchetti (2001) because it is based on the notion of quasi-likelihood function (or estimating equations). Additionally, the empirical study in the previous chapter about the log-linear Poisson model without the feedback mechanism has revealed that the MQLE estimator produces better results than the MLE. We show that this method, suitably adjusted for count time series, estimates consistently the regression coefficients in the presence of additive outliers and other types of interventions. We also complement the works by Fokianos and Fried (2010, 2012) who studied detection and testing for intervention effects in count time series. Furthermore, we develop a robust test statistic for testing the existence of the feedback mechanism.

Since the log-linear model is closely related to the standard GARCH model, our work is related to Muler and Yohai (2002, 2008) who propose two robust estimators for ARCH and GARCH models respectively; see also Mukherjee (2008) for other results.

Our interest is to extend the MQLE framework to time series setup and examine its asymptotic properties. We employ a perturbation technique following Fokianos and Tjøstheim (2011). Section 4.1 focuses on estimation and inference of MQLE and in Section 4.2 a robust procedure of the score test is employed for testing the existence of the feedback mechanism. In both sections we complement the study with a simulation example, while Section 4.3 provides two real data examples. Appendix 4.4 contains proofs of the theoretical results and Appendix B shows our R code for the empirical and real data study.

4.1 Mallows' Quasi Likelihood Estimation

Recall that the log–linear Poisson model with feedback is given by

$$Y_t \parallel \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \quad \nu_t = d + a\nu_{t-1} + b\log(1 + Y_{t-1}), \tag{4.1}$$

where $\{Y_t,\ t = 1, \ldots, n\}$ is a time series of counts, $\mathcal{F}_t$ denotes the σ–field $\sigma(Y_s, s \leq t)$ and $\nu_t \equiv \log \lambda_t$. Following the work of Fokianos and Tjøstheim (2011), introduce at each time point t, a Poisson process $N_t(\cdot)$ of unit intensity. Then (4.1) can be restated in terms of these Poisson processes by assuming that Y_t given λ_t is equal to the number of events $N_t(\lambda_t)$ in the time interval $(0, \lambda_t]$. Let therefore $\{N_t(\cdot), t = 1, 2, 3...\}$ be a sequence of independent Poisson processes of unit intensity and rewrite model (4.1) as

$$Y_t = N_t(\lambda_t), \quad \nu_t = d + a\nu_{t-1} + b\log(1 + Y_{t-1}).$$

To advance the theory, we introduce a perturbed chain (Y_t^m, ν_t^m) defined by

$$Y_t^m = N_t(\lambda_t^m), \quad \nu_t^m = d + a\nu_{t-1}^m + b\log(1 + Y_{t-1}^m) + \epsilon_{t,m} \tag{4.2}$$

where $\epsilon_{t,m} = c_m \mathbb{1}(Y_{t-1}^m = 1)\mathcal{U}_t$, $\quad c_m > 0$, $\quad c_m \to 0$ as $m \to \infty$, $\mathcal{U}_t$ is a sequence of iid uniform random variables on $(0,1)$ such that $\mathcal{U}_t$ is independent of $N_t(\cdot)$, and $\mathbb{1}(A)$ is the indicator function of a set A. This perturbation idea which is employed for studying properties of (4.1) is advanced in Fokianos et al. (2009) and Fokianos and Tjøstheim (2011) and is crucial for proving geometric ergodicity of the process. More specifically, the ergodic properties of the process (Y_t, ν_t) are based on the property of irreducibility and the introduction of the sequence $\{U_t\}$ makes possible to prove irreducibility and therefore geometric ergodicity.

Our interest focuses on estimating the parameters of the log-linear model (4.1) using the robust method of estimation proposed by Cantoni and Ronchetti (2001) – the so called Mallows' quasi likelihood estimation (MQLE). This estimation is then compared to the maximum likelihood estimation, as studied in Fokianos and Tjøstheim (2011).

To motivate the study of robust estimation, we show a simulated example where maximum likelihood estimation does not perform satisfactorily. We consider Additive Outliers (AO) of size ζ observed at times $\tau_1, \tau_2, \ldots, \tau_k$. In this case, the time series is given by

$$Z_t = \begin{cases} Y_t + \zeta, & \text{when} \quad t = \tau_1, \tau_2, \ldots, \tau_k, \\ Y_t, & \text{otherwise.} \end{cases} \tag{4.3}$$

We also consider a Level Shift (LS) and a Transient Shift (TS) type of intervention. Both types of interventions enter the dynamics of the model through the log-intensity process $\{\nu_t\}$ as a sequence of deterministic covariates $\{X_t\}$ that models the intervention occurring at time τ. Recall that in this case, we observe a contaminated process $\{Z_t\}$ which is given by

$$Z_t \parallel \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t^c), \quad \nu_t^c = d + a\nu_{t-1}^c + b\log(1 + Z_{t-1}) + \zeta X_t. \tag{4.4}$$

The sequence $\{X_t\}$ is given by

$$X_t = \xi(\mathcal{B})\mathbb{1}(t = \tau),$$

where $\xi(\mathcal{B}) = (1 - \delta\mathcal{B})^{-1}$, $\ \delta \in [0,1)$, $\mathcal{B}$ is a shift operator such that $\mathcal{B}^i X_t = X_{t-i}$, $\mathbb{1}_t(\tau)$

is an indicator function that is equal to 1 if $t = \tau$ and 0 otherwise and ζ is the size of the intervention. A LS corresponds to $\delta = 1$ and a TS corresponds to $\delta \in \{0.7, 0.8, 0.9\}$.

Table 4.1 shows empirical means of maximum likelihood estimates based on 1000 simulated samples, and their sample standard deviation (in parentheses), for a time series of length 500 from model (4.1) with $\boldsymbol{\theta} = (d, a, b) = (0.2, 0.3, 0.5)$ that contains a patch of consecutive outliers(recall (4.3)), a level shift and a transient shift type of intervention. The results demonstrate that the MLE is affected by interventions and does not estimate model's parameters satisfactorily. In fact, we note that the parameter d is generally underestimated but the parameter b is overestimated. A comprehensive discussion of interventions in count time series has been given by Fokianos and Fried (2010). The effect of interventions to count data has been investigated by Elsaied (2012), Fried et al. (2014) and Kitromilidou and Fokianos (2015). It has been noticed empirically that when the data are generated by a model which does not include $\{\nu_t\}$ and contains moderate LS or a TS type of intervention, then the MLE is not affected considerably. However, additive outliers still have an impact on estimation. Table 4.1 shows that the MLE does not estimate consistently the model parameters when the hidden process is included in the model in all the above cases.

Patch of Additive Outliers	$\hat{d}$	$\hat{a}$	$\hat{b}$
5 outliers	0.146 (0.073)	0.224 (0.067)	0.609 (0.050)
10 outliers	0.014 (0.054)	0.283 (0.056)	0.640 (0.048)
15 outliers	-0.043 (0.046)	0.312 (0.051)	0.652 (0.048)
20 outliers	-0.073 (0.042)	0.327 (0.049)	0.656 (0.047)
Level Shift	0.007 (0.042)	0.438 (0.042)	0.552 (0.042)
Transient Shift	0.036 (0.059)	0.273 (0.060)	0.633 (0.050)

Table 4.1: Maximum likelihood estimation of the true value $\boldsymbol{\theta} = (0.2, 0.3, 0.5)$ for the log-linear model (4.1), under increasing number of additive outliers, a level shift and a transient shift intervention. Results are based on 1000 simulations of length 500. Standard errors are reported in parentheses.

4.1.1 Estimation

The MQLE is defined as a Fisher consistent M-estimator which is given as a solution of the following quasi–score function $S_n(\boldsymbol{\theta}) = 0$ where

$$S_n(\boldsymbol{\theta}) = \sum_{t=1}^{n}\Big(m_t(\boldsymbol{\theta}) - E\Big(m_t(\boldsymbol{\theta}) \parallel \mathcal{F}_{t-1}\Big)\Big) = \sum_{t=1}^{n} s_t(\boldsymbol{\theta}), \tag{4.5}$$

with

$$m_t(\boldsymbol{\theta}) = \psi\Big(r_t(\boldsymbol{\theta})\Big) w_t e^{\nu_t(\boldsymbol{\theta})/2}\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

We will denote the estimator obtained by the solution of the above equations as $S_n(\boldsymbol{\theta}) = 0$ by $\hat{\boldsymbol{\theta}}_{\text{MQLE}}$. Note that $r_t = (Y_t - \lambda_t)/\sqrt{\lambda_t}$ are the so called Pearson residuals and ψ is a suitable weight function that depends on a tuning constant chosen to ensure a desired level of asymptotic efficiency (see also Cantoni and Ronchetti (2001)). The most common choice for the function $\psi(\cdot)$ is the Huber function given by (4.6) but several other choices are available to the literature, like the Tukey biweight function. The vector $\partial\nu_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is a three dimensional vector with components

$$\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial d} = 1 + a\frac{\partial \nu_{t-1}(\boldsymbol{\theta})}{\partial d}, \quad \frac{\partial \nu_t(\boldsymbol{\theta})}{\partial a} = \nu_{t-1} + a\frac{\partial \nu_{t-1}(\boldsymbol{\theta})}{\partial a}, \quad \frac{\partial \nu_t(\boldsymbol{\theta})}{\partial b} = \log(1+Y_{t-1}) + a\frac{\partial \nu_{t-1}(\boldsymbol{\theta})}{\partial b}.$$

The sequence $\{w_t\}$ is an appropriate sequence of weights. Some choices include $w_t = \sqrt{1-h_{tt}}$ where h_{tt} are the diagonal elements of the so called "hat matrix" H. However, as Cantoni and Ronchetti (2001) point out, weights defined on the hat matrix do not have high breakdown properties. Choices of robust weights can be based on the inverse of the robust Mahalanobis distance matrix where location and scatter are robustly estimated to have high breakdown properties using either the minimum volume ellipsoid estimator or the minimum covariance determinant (Rousseeuw and van Zomeren (1990) and Rousseeuw and Driessen (1999)). Regardless of the choice of robust weights, it always holds that $0 < w_t \leq 1$, cf. Seber and Lee (2003, p. 89). The bias correction term appearing in (4.5) implies that the resulting estimator is Fisher–consistent (see also Cantoni and Ronchetti (2001)).

The properties of the MQLE can be studied theoretically by viewing the quasi–score equations $S_n(\boldsymbol{\theta}) = 0$ as a solution to the following maximization problem. More precisely, defining

$$M_t(\boldsymbol{\theta}) = \int_{\tilde{s}}^{\lambda_t(\boldsymbol{\theta})} \Big(\psi\left(\frac{Y_t - z}{\sqrt{z}}\right) - E\Big[\psi\left(\frac{Y_t - z}{\sqrt{z}}\right) \parallel \mathcal{F}_{t-1}\Big]\Big) w_t \frac{1}{\sqrt{z}} dz,$$

with $\tilde{s}$ such that $\Big(\psi\left(\frac{Y_t-\tilde{s}}{\sqrt{\tilde{s}}}\right) - E\Big[\psi\left(\frac{Y_t-\tilde{s}}{\sqrt{\tilde{s}}}\right) \parallel \mathcal{F}_{t-1}\Big]\Big) w_t \frac{1}{\sqrt{\tilde{s}}} = 0$, we obtain that

$$\frac{\partial}{\partial \boldsymbol{\theta}} M_t(\boldsymbol{\theta}) = m_t(\boldsymbol{\theta}) - E\Big(m_t(\boldsymbol{\theta}) \parallel \mathcal{F}_{t-1}\Big).$$

By using Taniguchi and Kakizawa (2000, Thm 3.2.23) which is based on the work by Klimko and Nelson (1978), we can prove existence, consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}_{\text{MQLE}}$. To obtain these results the main assumption imposed on the function $\psi(\cdot)$ is to be bounded and to possess a second derivative which is continuous. Obviously such a condition is not satisfied by the Huber function (4.6). In these cases however, we can develop asymptotic properties of $\hat{\boldsymbol{\theta}}_{\text{MQLE}}$ by appealing to the theory of Z-estimators; in particular van der Vaart (1998, Thm 5.21) lists the necessary conditions for obtaining asymptotic normality provided that $\hat{\boldsymbol{\theta}}_{\text{MQLE}}$ is a consistent root of $S_n(\boldsymbol{\theta}) = 0$. Consistency is established by empirical process theory.

The main problem that we are facing is that we cannot prove the necessary conditions for the unperturbed model to obtain the asymptotic theory (see also Fokianos et al. (2009), Fokianos and Tjøstheim (2011) and Tjøstheim (2012) for detailed discussion about the issues involved). Therefore we prove the corresponding conditions for the perturbed model and then show that the perturbed and unperturbed versions are "close". Towards this goal we define analogously S_n^m to be the MQLE score function for the perturbed model

$$S_n^m(\boldsymbol{\theta}) = \sum_{t=1}^{n} \Big(m_t^m(\boldsymbol{\theta}) - E\Big(m_t^m(\boldsymbol{\theta}) \parallel \mathcal{F}_{t-1}^m\Big)\Big) = \sum_{t=1}^{n} s_t^m(\boldsymbol{\theta}),$$

and m_t^m, r_t^m to have the same form as their counterparts in the non perturbed model but with (Y_t, ν_t) replaced by (Y_t^m, ν_t^m). Then the following theorem follows after proving Lemmas 4.1.1–4.1.5 which verify the conditions of Taniguchi and Kakizawa (2000, Thm 3.2.23).

Theorem 4.1.1 Consider model (4.1). Let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^3$ which is assumed compact and suppose that the true value $\boldsymbol{\theta}_0$ belongs to the interior of $\boldsymbol{\Theta}$. Assume further that ψ is two times continuously differentiable bounded function. Introduce lower and upper values of each component of $\boldsymbol{\theta}_0 = (d_0, a_0, b_0)^T$ such that $d_L < d_0 < d_U$, $-1 < a_L < a_0 < a_U < 1$ and $b_L < b_0 < b_U$ and suppose that at the true value $\boldsymbol{\theta}_0$, $|a_0 + b_0| < 1$ if a_0 and b_0 have the same sign, and $a_0^2 + b_0^2 < 1$ if a_0 and b_0 have different sign. Then, there exists a fixed open neighborhood $O(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$

$$O(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} | d_L < d < d_U, -1 < a_L < a < a_U < 1, b_L < b < b_U\}$$

such that with probability tending to 1 as $n \to \infty$, the equation $S_n(\boldsymbol{\theta}) = 0$ has a unique solution, say $\hat{\boldsymbol{\theta}}_{\mathrm{MQLE}}$. Furthermore, $\hat{\boldsymbol{\theta}}_{\mathrm{MQLE}}$ is strongly consistent and asymptotically normal,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{MQLE}} - \boldsymbol{\theta}_0) \xrightarrow{\mathrm{d}} N(0, V^{-1} W V^{-1})$$

where the matrices W and V are defined in Lemmas 4.1.1 and 4.1.3.

To prove this Theorem we need the following lemmas whose proof is postponed to the Appendix.

Lemma 4.1.1 Define the matrices

$$W^m(\boldsymbol{\theta}) = E\Big(s_t^m(\boldsymbol{\theta}) s_t^m(\boldsymbol{\theta})^T\Big) \quad \text{and} \quad W(\boldsymbol{\theta}) = E\Big(s_t(\boldsymbol{\theta}) s_t(\boldsymbol{\theta})^T\Big).$$

Under the assumptions of Theorem 4.1.1, the above matrices evaluated at the true value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, satisfy $W^m \to W$, as $m \to \infty$.

Lemma 4.1.2 Under the assumptions of Theorem 4.1.1, the score functions for the perturbed (4.2) and unperturbed model (4.1) evaluated at the true value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ satisfy the following:

1. $S_n^m / n \xrightarrow{\mathrm{a.s}} 0$,

2. $S_n^m / \sqrt{n} \xrightarrow{\mathrm{d}} S^m := N(0, W^m)$,

3. $S^m \xrightarrow{\mathrm{d}} N(0, W)$, as $m \to \infty$,

4. $\lim_{m\to\infty} \limsup_{n\to\infty} P(||S_n^m - S_n|| > \epsilon\sqrt{n}) = 0, \quad \forall \epsilon > 0.$

Lemma 4.1.3 Define the matrices

$$V^m(\boldsymbol{\theta}) = -E\left[\frac{\partial}{\partial \boldsymbol{\theta}} s_t^m(\boldsymbol{\theta})\right], \qquad V(\boldsymbol{\theta}) = -E\left[\frac{\partial}{\partial \boldsymbol{\theta}} s_t(\boldsymbol{\theta})\right].$$

Under the assumptions of Theorem 4.1.1, the above matrices evaluated at the true value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, satisfy $V^m \to V$, as $m \to \infty$.

Lemma 4.1.4 Denote by

$$H_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} s_t(\boldsymbol{\theta})\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

where $\ell_t(\boldsymbol{\theta}) = Y_t\nu_t(\boldsymbol{\theta}) - \exp(\nu_t(\boldsymbol{\theta}))$, is the t'th component of the Poisson log-likelihood function. Define analogously $H_n^m(\boldsymbol{\theta})$. Then, under the assumptions of Theorem 4.1.1,

1. $H_n^m \xrightarrow{\text{p}} V^m$ as $n \to \infty$

2. $\lim_{m\to\infty} \limsup_{n\to\infty} P(||H_n^m - H_n|| > \epsilon n) = 0, \quad \forall \epsilon > 0.$

where V^m has been defined in Lemma 4.1.3.

Lemma 4.1.5 Under the assumptions of Theorem 4.1.1,

$$\max_{i,j,k=1,2,3} \sup_{\boldsymbol{\theta}\in O(\boldsymbol{\theta}_0)} \left|\frac{1}{n}\sum_{t=1}^{n} \frac{\partial^2 s_{ti}(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_j}\right| \leq \tilde{M}_n := \frac{1}{n}\sum_{t=1}^{n} \tilde{m}_t$$

where θ_i for $i = 1, 2, 3$ refers to $\theta_i = d, a, b$ respectively and $\{\tilde{m}_t\}$ is defined by (A-8). Define analogously $\tilde{M}_n^m$. Then

1. $\tilde{M}_n^m \xrightarrow{\text{p}} \tilde{M}^m$, as $n \to \infty$ for each $m = 1, 2, ...,$

2. $\tilde{M}^m \to \tilde{M}$, as $m \to \infty$, where $\tilde{M}$ is a finite constant,

3. $\lim_{m\to\infty} \limsup_{n\to\infty} P(|\tilde{M}_n^m - \tilde{M}_n| > \epsilon n) = 0, \quad \forall \epsilon > 0.$

Remark 4.1.1 A related study to ours in that of Elsaied and Fried (2014). The authors are studying robust estimation for a linear count time series model but without feedback. We study a log-linear model which includes feedback and additionally we develop a test statistic for robust testing of existence of the feedback process - see Section (4.2).

4.1.2 A Simulated Example

A brief simulation study is presented to demonstrate the performance of the MQLE estimator. Recall that the observed process in the presence of additive outliers is given by (4.3). The scenario in which the additive outliers appear in consecutive times $\tau_1, \tau_2, \ldots, \tau_k$ is referred to as a patch of outliers. In the following, a time series of size 800 is generated under the assumptions of model (4.1) with a patch of consecutive outliers. The first 300 observations are discarded. We choose $\boldsymbol{\theta} = (0.2, 0.3, 0.5)$ and $\boldsymbol{\theta} = (0.2, 0.3, 0.65)$ and the size of the intervention is set to $\zeta = 20$.

From several simulation studies we have empirically observed that the MQLE estimator is superior when weighted - in fact when robustly weighted. As discussed by Cantoni and Ronchetti (2001), suitable choices for the weights w_t are given by $\sqrt{1 - h_{tt}}$ where h_{tt} are the diagonal elements of the hat matrix $H = X(X^T X)^{-1} X^T$, where X is a general design matrix. Alternatively, we can employ the inverse of the robust Mahalanobis distance where location and scatter are robustly estimated to have a high breakdown point by using either the MVE algorithm or the MCD algorithm (Rousseeuw and van Zomeren (1990) and Rousseeuw and Driessen (1999)). To use robustly weighted methods in this context, it is required to substitute the hidden process $\{\nu_t\}$ by an observable process so that the design matrix can be calculated explicitly.

Two methods are proposed to create the design matrix X for the case of model (4.1) which includes a hidden process.

Method A:

The matrix X is constructed by approximating (4.1) with

$$\hat{\nu}_t = d + a\hat{\nu}_{t-1} + b_1 \log(1 + Y_{t-1}),$$

where $\hat{\nu}_t$ is computed by employing $\hat{\boldsymbol{\theta}}_{MQLE}$ calculated without weights.

Method B:

We approximate (4.1) by

$$\hat{\nu}_t = d^* + \sum_{i=1}^{M} a_i^* \log(1 + Y_{t-i}),$$

for some truncation point M and some regression parameters $\{d^*, a_1^*, \ldots, a_M^*\}$. This choice is motivated by the fact that repeated substitution in (4.1) shows that

$$\nu_t = d\frac{1-a^t}{1-a} + a^t \nu_0 + b\sum_{i=0}^{t-1} a^i \log(1+Y_{t-i-1}).$$

Repeated substitution up to M lags gives

$$\begin{aligned}
\nu_t &= d + a\nu_{t-1} + b\log(1+Y_{t-1}) \\
&= d + a(d + a\nu_{t-2} + b\log(1+Y_{t-2})) + b\log(1+Y_{t-1}) \\
&= d(1+a) + a^2\nu_{t-2} + ab\log(1+Y_{t-2}) + b\log(1+Y_{t-1}) \\
&= d(1+a) + a^2(d + a\nu_{t-3} + b\log(1+Y_{t-3})) + ab\log(1+Y_{t-2}) + b\log(1+Y_{t-1}) \\
&= d(1+a+a^2) + a^3\nu_{t-3} + a^2 b\log(1+Y_{t-3}) + ab\log(1+Y_{t-2}) + b\log(1+Y_{t-1}) \\
&= \cdots \\
&= d(1+a+a^2+\cdots+a^{M-1}) + a^M\nu_{t-M} + \sum_{i=1}^{M} a^{i-1} b\log(1+Y_{t-i}),
\end{aligned}$$

therefore in the approximation of ν_t,

$$d^* = d\frac{1-a^{M-1}}{1-a} + a^M\nu_{t-M} \quad \text{and} \quad a_i^* = a^{i-1}b.$$

In our studies, we use the truncation point $M = 20$. To be more precise, in method B, a total of N observations are generated from which a number of observations are discarded so that n observations remain. We include the last M discarded observations and create the $n \times n$ matrix X as follows

$$X = \begin{bmatrix}
\log(1+Y_{N-(n-1)}) & \log(1+Y_{N-1-(n-1)}) & \log(1+Y_{N-2-(n-1)}) & \dots & \log(1+Y_{N-(M-1)-(n-1)}) \\
\log(1+Y_{N-(n-2)}) & \log(1+Y_{N-1-(n-2)}) & \log(1+Y_{N-2-(n-2)}) & \dots & \log(1+Y_{N-(M-1)-(n-2)}) \\
\log(1+Y_{N-(n-3)}) & \log(1+Y_{N-1-(n-3)}) & \log(1+Y_{N-2-(n-3)}) & \dots & \log(1+Y_{N-(M-1)-(n-3)}) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\log(1+Y_N) & \log(1+Y_{N-1}) & \log(1+Y_{N-2}) & \dots & \log(1+Y_{N-(M-1)})
\end{bmatrix}$$

The behavior of MLE, MQLE without weights, MQLE with ”hat weights” (using the approximation method A or B), MQLE based on the MVE algorithm (using again the methods A and B) and MQLE based on the MCD algorithm (using again methods A and B) is examined in terms of their mean square error (MSE). In all the calculations we use the Huber function

$$\psi_c(x) = \begin{cases} x, & |x| \leq c \\ c\text{sign}(x), & |x| > c \end{cases} \tag{4.6}$$

where sign$(\cdot)$ denotes the sign function. Clearly, the choice of (4.6) does not fulfill the requirements of Theorem 4.1.1, but it is usually employed in robust estimation based methods. Even though (4.6) does not fulfill the requirements of Theorem 4.1.1, $\hat{\boldsymbol{\theta}}_{MQLE}$ is expected to be asymptotically normally distributed, as it was discussed earlier. In fact, by recalling (4.5), we note that for this particular choice of $\psi(\cdot)$ function, the bias term can be calculated upon noticing that

$$E\Big(m_t(\boldsymbol{\theta}) \parallel \mathcal{F}_{t-1}\Big) = E\Big(\psi_c(r_t(\boldsymbol{\theta})) \parallel \mathcal{F}_{t-1}\Big) w_t e^{\nu_t(\boldsymbol{\theta})/2} \frac{\partial \nu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

where

$$\begin{aligned} E\Big(\psi_c\Big(\frac{Y_t - \lambda_t(\boldsymbol{\theta})}{\sqrt{\lambda_t(\boldsymbol{\theta})}}\Big) \parallel \mathcal{F}_{t-1}\Big) &= c\left\{P(Y_t \geq j_2 + 1 \| \mathcal{F}_{t-1}) - P(Y_t \leq j_1 \| \mathcal{F}_{t-1})\right\} \\ &+ \sqrt{\lambda_t(\boldsymbol{\theta})}\left\{P(Y_t = j_1 \| \mathcal{F}_{t-1}) - P(Y_t = j_2 \| \mathcal{F}_{t-1})\right\}, \end{aligned}$$

with j_1 and j_2 defined by $j_1 = \lfloor \lambda_t(\boldsymbol{\theta}) - c\sqrt{\lambda_t(\boldsymbol{\theta})} \rfloor$ and $j_2 = \lfloor \lambda_t(\boldsymbol{\theta}) + c\sqrt{\lambda_t(\boldsymbol{\theta})} \rfloor$, Cantoni and Ronchetti (2001). As the tuning constant c tends to infinity, then the MQLE approximates the ordinary MLE, provided that $w_t \equiv 1$. However, this approximation is no longer true when robust weights are employed for solving (4.5) by using the Huber function (4.6).

Figures 4.1 and 4.4 (respectively Figures 4.2 and 4.5) display the MSE of $\hat{a}_{\text{MQLE}}$ (respectively $\hat{b}_{\text{MQLE}}$) for all estimation methods considered. To solve the score equations (4.5), we initialize $\nu_0 = 1$, $\partial \nu_0 / \partial \boldsymbol{\theta} = 1$ and we employ the above approximation to the bias terms based on (4.6).

For this simulation, we consider the intervention to occur in the first quarter of the series, but simulations where the intervention occurred in the middle of the series provided similar conclusions. It is obvious that in all cases the robustly weighted MQLE outperforms the non-robustly weighted MQLE as well as the MLE. When comparing the two proposed weighting methods we note that method A is superior to method B in all cases considered. It is also noted that for both the estimation of the parameters a and b, as the number of consecutive outliers increases so does the MSE values. Similar results have been obtained for other measures of performance like mean absolute error and bias but they are omitted.

Additionally, Figure 4.3 shows QQ-plots of the simulated values of $\hat{a}$ for all estimating methods considered. For this example, we assume a patch of 15 consecutive outliers of size $\zeta = 20$ and the value of the tuning constant of the Huber function is set to $c = 1.571$. In general, we recommend values of the tuning constant c to be chosen between 1.50 and 2 when the data does not indicate non stationarity. The plot illustrates that the asserted asymptotic normality is achieved quite satisfactorily.

We examine also the way that various estimation procedures perform when the sum of the regression coefficients, say $a + b$, is close to unity in the case of observing a patch of outliers. Figures 4.4 and 4.5 show the MSE of $\hat{a}_{\text{MQLE}}$ and $\hat{b}_{\text{MQLE}}$ respectively, for the choice of parameter $\boldsymbol{\theta} = (0.2, 0.3, 0.65)$. These values are close to the estimated obtained by the data analysis example included in Section 4.3.1. In both cases, the best performing estimating procedure is again the robustly weighted MQLE obtained by method A. Note that when the data are close to the non-stationarity region, the values of MSE remains approximately the same regardless of the number of outliers.

We include an additional simulation example in which a LS or a TS occurs at the first quarter of the series. The size of the intervention is $\zeta = 0.2$ for the case of a LS, and $\zeta = 1$ for the case of a TS. For the case of a TS, the value of δ is chosen to be 0.8; recall (4.4). Figures 4.6 and 4.7 show that the robustly weighted MQLE perform significantly better than all other estimation procedures. Also, when comparing the two proposed methods for obtaining the robustly weighted MQLE, method A yields to considerably better results than method B in terms of MSE for the case of $a+b < 1$. When $a+b$ approach 1, we note that the MSE of $\hat{a}_{\text{MQLE}}$ and $\hat{b}_{\text{MQLE}}$ changes slowly across c. A possible explanation of this behavior is that for $a + b$ close to unity, the process Y_t assumes large values and therefore moderate values of an intervention do not influence the final outcome of estimation.

The choice of the tuning constant c is of high importance in our analysis and depends upon a predefined level of relative efficiency. In the simulation example that follows, we calculate the relative efficiency of the robust estimators for various values of c based on

1000 generated samples of clean data. An efficiency of approximately 95% determines the value of c. The results are displayed in Figure 4.8.

As shown, the estimator that possesses the highest level of efficiency is the MQLE without weights, followed by the non-robustly weighted MQLE with weights based on the hat matrix. On the other hand, the robustly weighted MQLE are considerably less efficient, especially when weighted using the MCD algorithm.

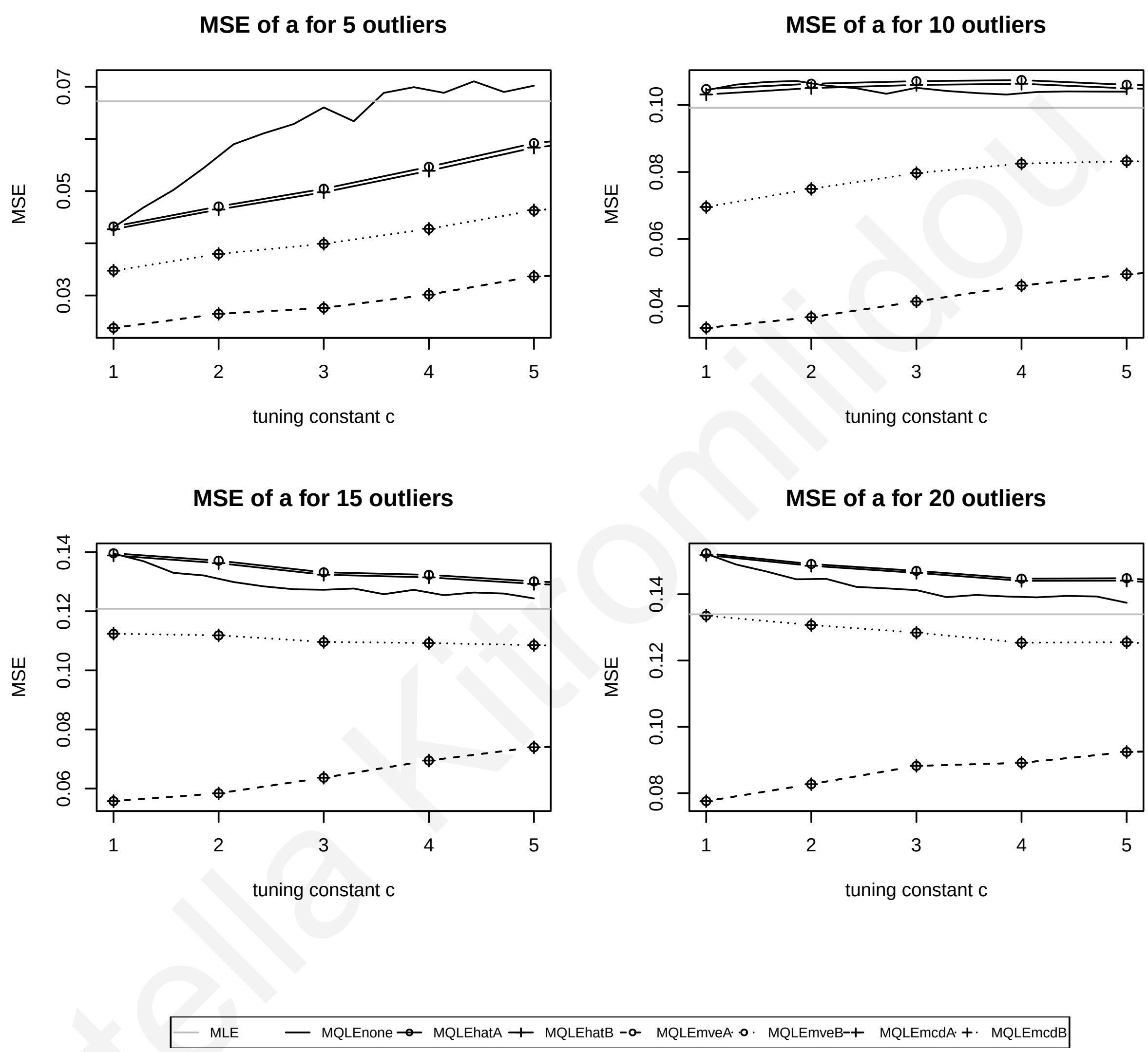

Figure 4.1: MSE of $\hat{a}_{\mathrm{MQLE}}$ as a function of the tuning constant c when using for estimation the Huber function (4.6). Data are generated by model (4.1) with $(d, a, b) = (0.2, 0.3, 0.5)$ and with a patch of outliers–see (4.3)–of size $\zeta = 20$. Results are based on 1000 simulations.

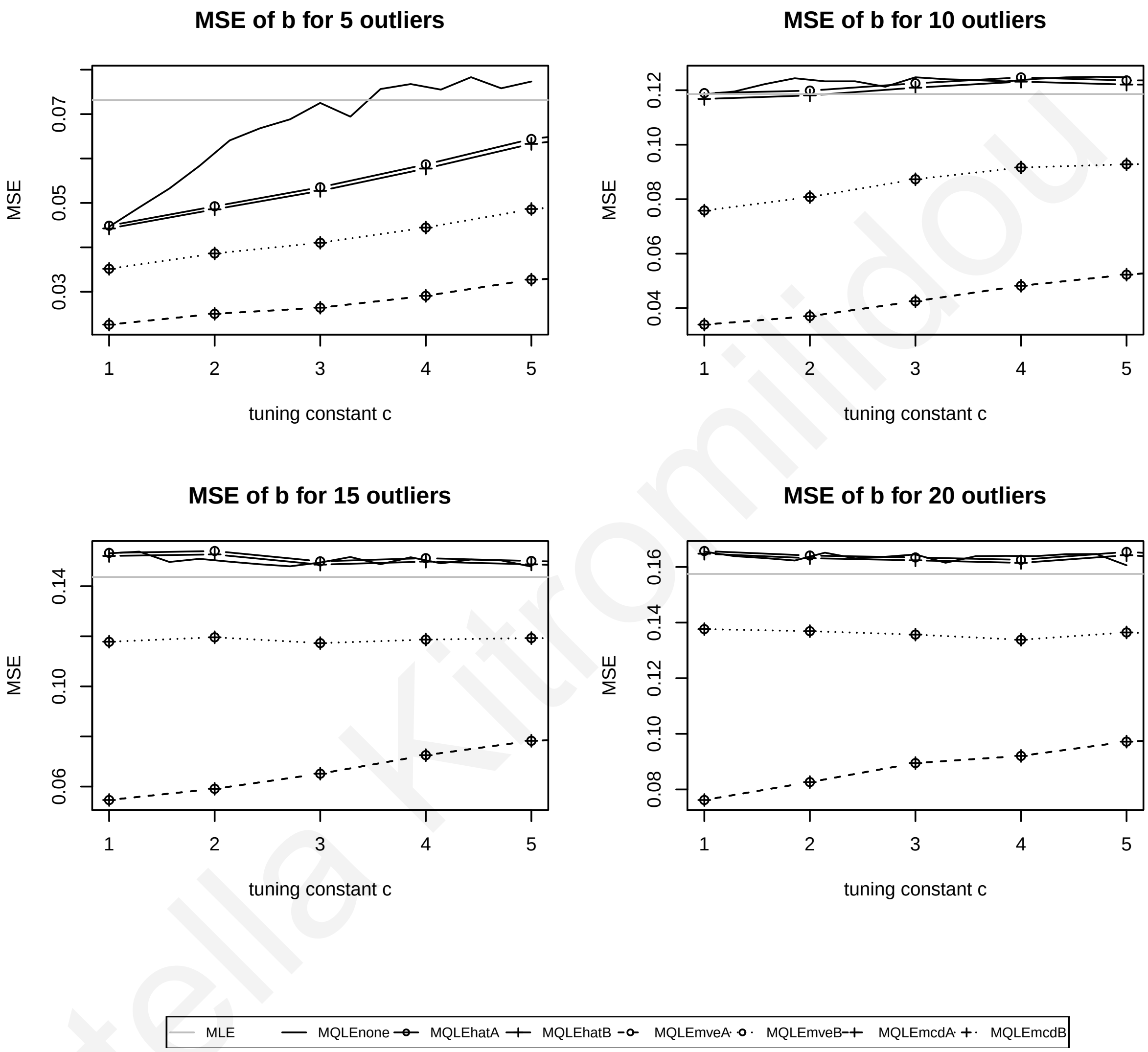

Figure 4.2: MSE values of $\hat{b}_{\mathrm{MQLE}}$ as a function of the tuning constant c when using for estimation the Huber function (4.6). Data are generated by model (4.1) with $(d, a, b) = (0.2, 0.3, 0.5)$ and with a patch of outliers–see (4.3)–of size $\zeta = 20$. Results are based on 1000 simulations.

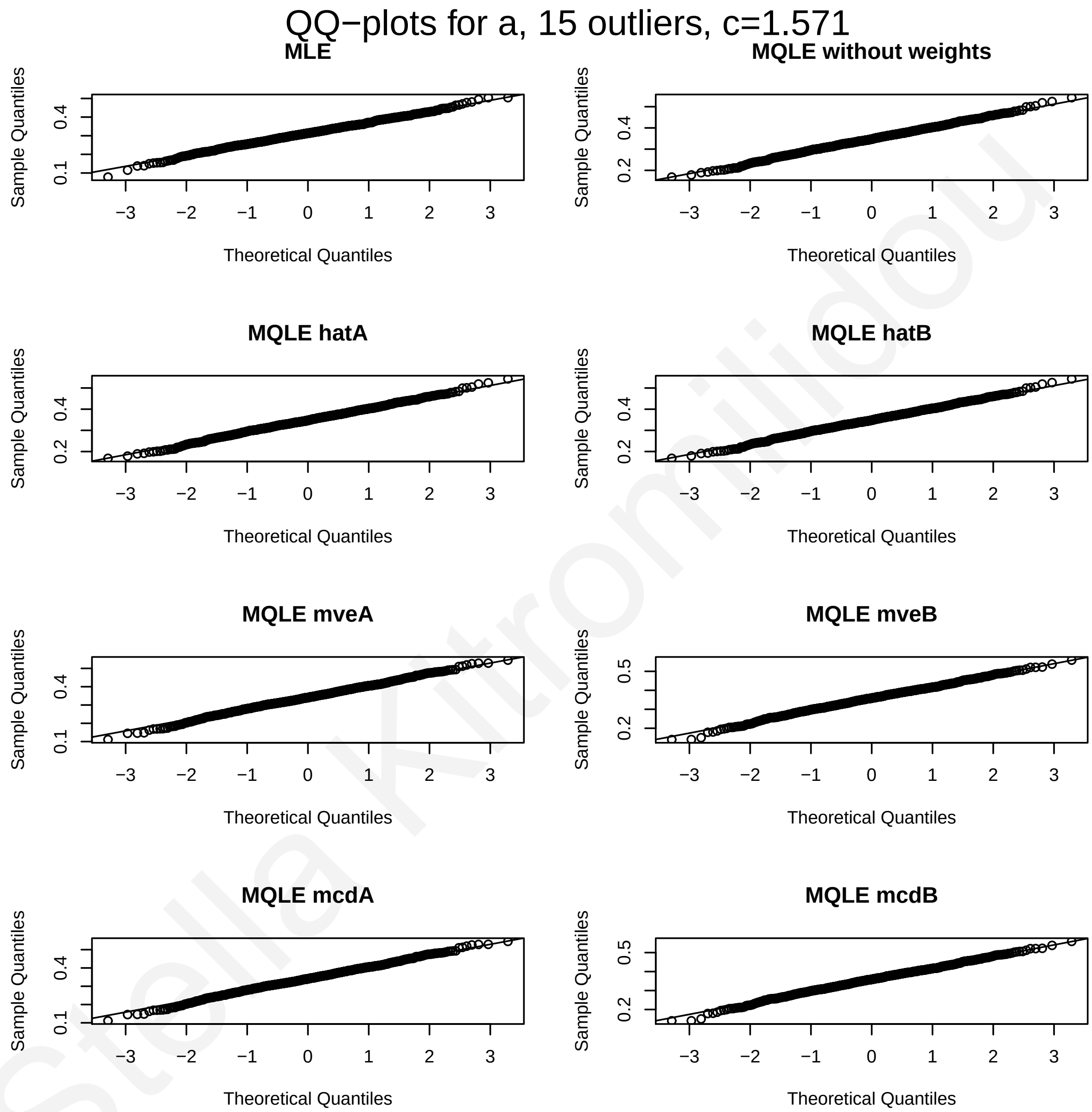

Figure 4.3: QQ-plots for the various estimating methods for $\hat{a}_{\mathrm{MQLE}}$. A patch of 15 consecutive outliers of size $\zeta = 20$–see (4.3)–is considered and the Huber function (4.6) is used for the calculation of MQLE with tuning constant equal to $c = 1.571$. Data are generated by model (4.1) with $(d, a, b) = (0.2, 0.3, 0.5)$. Results are based on 1000 simulations.

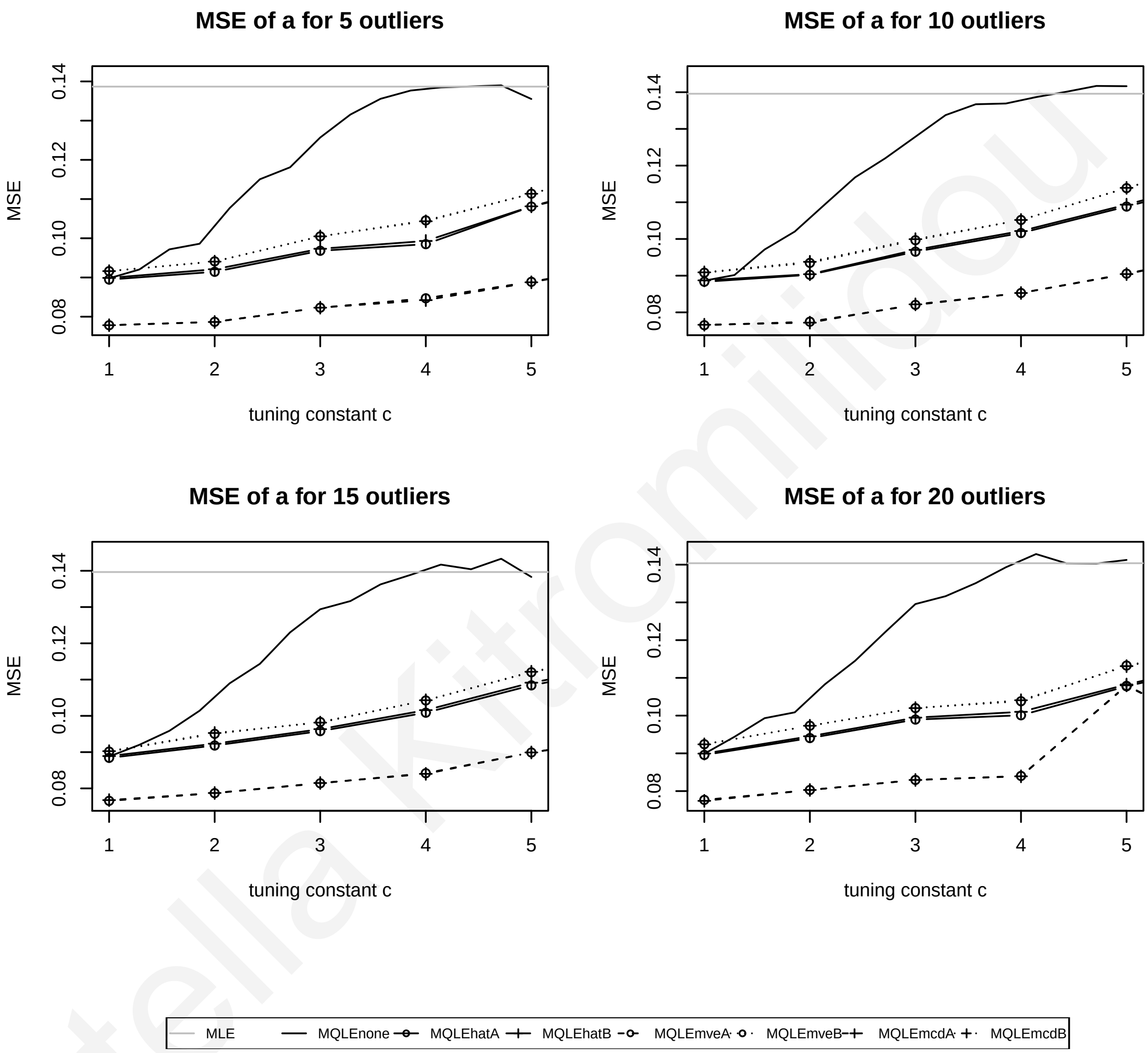

Figure 4.4: MSE of $\hat{a}_{\mathrm{MQLE}}$ as a function of the tuning constant c when using for estimation the Huber function (4.6). Data are generated by model (4.1) with $\boldsymbol{\theta} = (d, a, b) = (0.2, 0.3, 0.65)$ and with a patch of outliers–see (4.3)–of size $\zeta = 20$. Results are based on 1000 simulations.

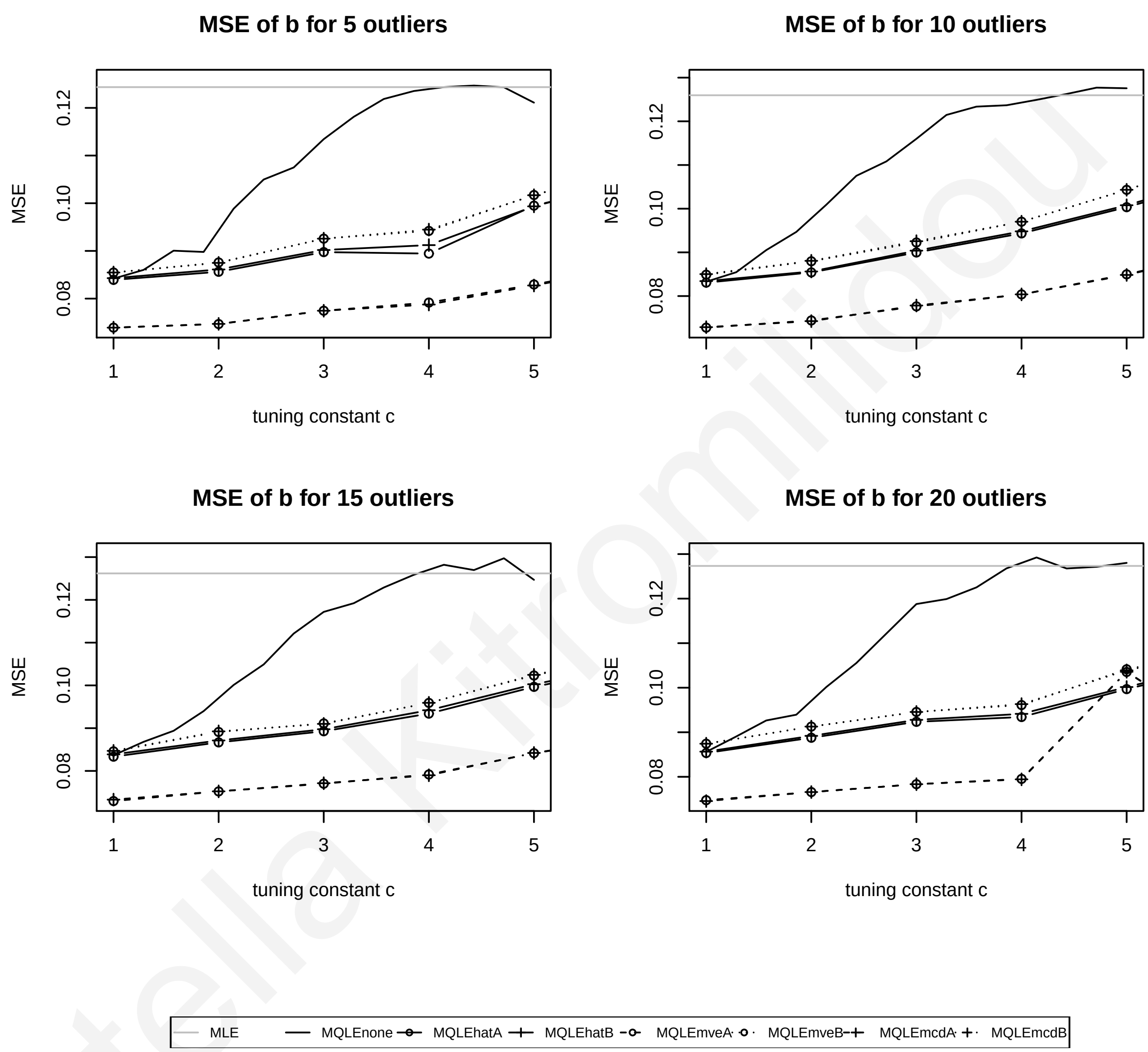

Figure 4.5: MSE of $\hat{b}_{\text{MQLE}}$ as a function of the tuning constant c when using for estimation the Huber function (4.6). Data are generated by model (4.1) with $\boldsymbol{\theta} = (d, a, b) = (0.2, 0.3, 0.65)$ and with a patch of outliers–see (4.3)–of size $\zeta = 20$. Results are based on 1000 simulations.

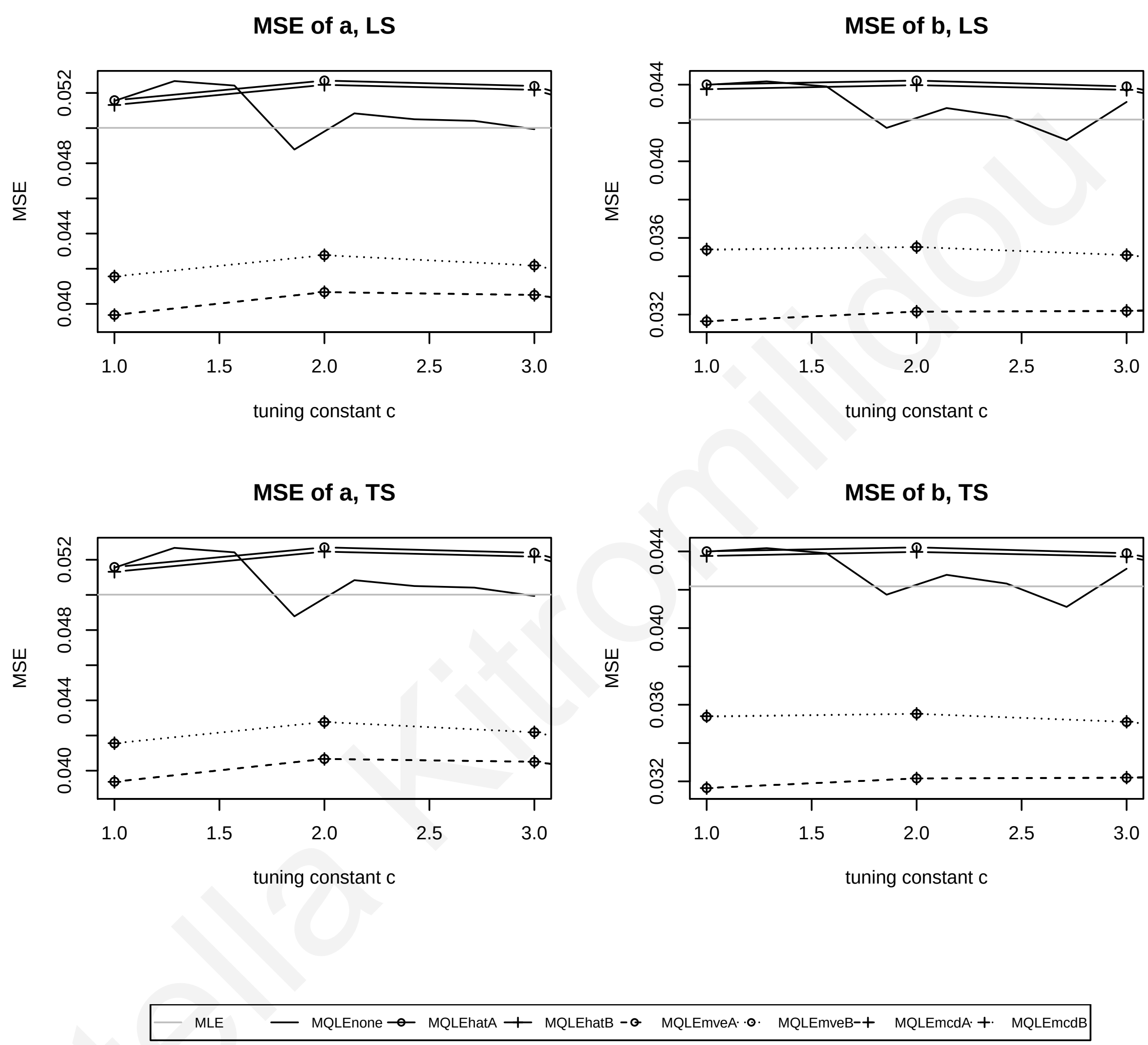

Figure 4.6: MSE of $\hat{a}_{\mathrm{MQLE}}$ (right column) and $\hat{b}_{\mathrm{MQLE}}$ (left column) as a function of the tuning constant c when using for estimation the Huber function (4.6). Data are generated by model (4.4) with $(d, a, b) = (0.2, 0.3, 0.5)$ and with either a LS of size $\zeta = 0.2$ or a TS of size $\zeta = 1$. Results are based on 1000 simulations.

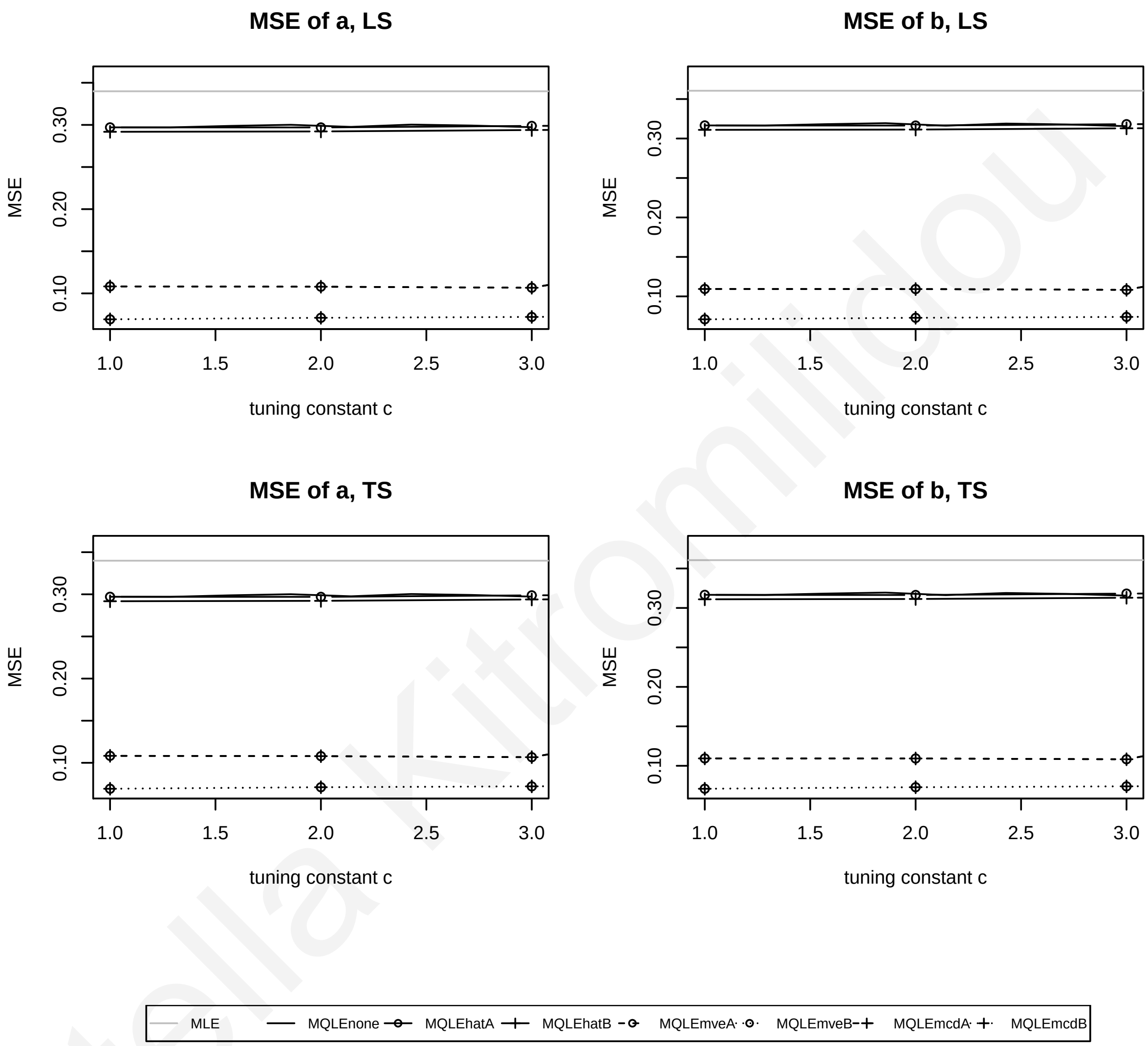

Figure 4.7: MSE of $\hat{a}_{\text{MQLE}}$ (right column) and $\hat{b}_{\text{MQLE}}$ (left column) as a function of the tuning constant c when using for estimation the Huber function (4.6). Data are generated by model (4.4) with $(d, a, b) = (0.2, 0.3, 0.65)$ and with either a LS of size $\zeta = 0.2$ or a TS of size $\zeta = 1$. Results are based on 1000 simulations.

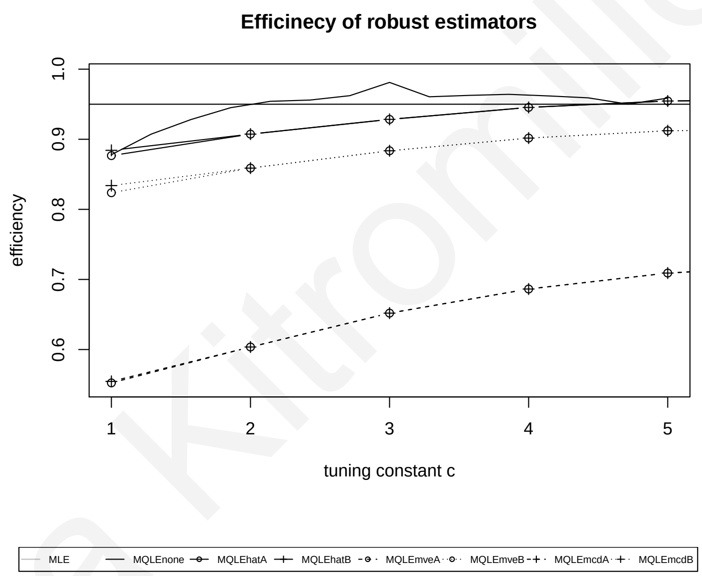

Figure 4.8: Efficiency and choice of c.

4.2 Robust Score Test

In this section we develop a robust procedure for testing the presence of the feedback process $\{\nu_t\}$ when there exist outliers, or unusual observations, in the data. In other words, we will be testing

$$H_0 : a = 0 \quad \text{vs.} \quad H_1 : a \neq 0 \tag{4.7}$$

for model (4.1). Heritier and Ronchetti (1994) consider robust tests based on M-estimators and influence functions. Testing the above hypothesis is performed by employing the score test, because this test statistic is constructed under the null hypothesis and can be easily computed, see Francq and Zakoïan (2010).

4.2.1 Results

For developing the score test, consider the partition $\boldsymbol{\theta} = (\theta^{(1)}, \theta^{(2)})^T$ of the vector of parameters $\boldsymbol{\theta} = (d, a, b)^T$ where $\theta^{(1)} = (d, b)^T$ and $\theta^{(2)} = a$. Then, the score test is defined by

$$ST_n = [S_n^{(2)}(\tilde{\boldsymbol{\theta}}_{\mathrm{MQLE}})]^2 / \tilde{\sigma}^2, \tag{4.8}$$

where $S_n^{(2)}$ is the second component of the partition of the score $S_n = (S_n^{(1)}, S_n^{(2)})^T$ and $\tilde{\boldsymbol{\theta}}_{\mathrm{MQLE}}$ is the constrained MQLE under the null hypothesis (4.7), given by $\tilde{\boldsymbol{\theta}}_n = (\tilde{\boldsymbol{\theta}}_n^{(1)}, 0)^T$ (see Breslow (1990), Harvey (1990) and Christou and Fokianos (2015)). In addition, $\tilde{\sigma}^2$ is a consistent estimator of

$$\sigma^2 = W_{22} - V_{21}{V_{11}}^{-1}W_{12} - W_{21}{V_{11}}^{-1}V_{12} + V_{21}{V_{11}}^{-1}W_{11}{V_{11}}^{-1}V_{12}$$

where V_{ij}, W_{ij}, $i, j = 1, 2$ correspond to partitions of the matrices V and W. The form of the variance term σ^2 is based on the fact that the MQLE is not based on a true log-likelihood function but on a quasi-score function. With this notation, we have the following results:

Theorem 4.2.1 Consider model (4.1) and assume the conditions of Theorem 4.1.1. Then, under the null hypothesis (4.7) we have the following:

1. Define the score test for the perturbed model (4.2) by ST_n^m. Then

$$ST_n^m \xrightarrow{\text{d}} \chi_1^2$$

where χ_d^2 denotes the chi–square distribution with d degrees of freedom.

2. The score statistic for the perturbed model (4.2) and unperturbed model (4.1) satisfy

$$\lim_{m\to\infty} \limsup_{n\to\infty} P(|ST_n^m(\tilde{\boldsymbol{\theta}}_n) - ST_n(\tilde{\boldsymbol{\theta}}_n)| > \epsilon n) = 0, \quad \forall \epsilon > 0.$$

The above results imply that we reject (4.7) whenever $ST_n > \chi^2_{1,\frac{\alpha}{2}}$ or $ST_n < \chi^2_{1,1-\frac{\alpha}{2}}$ where α is a predetermined level of significance.

4.2.2 Simulation Results for the Robust Score Test

Testing is performed in two distinct cases, the case of a patch of outliers and additionally, the case of isolated outliers, in which the outliers are added to the observed series in non-contiguous time positions. Our objective is to obtain a test statistic which achieves the correct size as well as has high power. To study first the MQLE estimator as to the size of the test, we simulate the behavior of the score test statistic given by (4.8). We generate 1000 samples under the null hypothesis and calculate the score statistic (4.8) for the various weighting options we discussed. We use again the Huber function (4.6) for construction of the test statistic. Our study is based on choosing $a = 0$ for the size and $a > 0$ for the power, and b as before (see Section 4.1.2). Note that under (4.7), there is no need to consider an approximation to the hidden process $\{\nu_t\}$ as it was discussed in Section 4.1.2.

We generate 1000 samples of size 800 from model (4.2). The first 300 observations are discarded to ensure stationarity of the process, hence the time series consists of 500 observations. The parameters of the model are set to $d = 0.2$ and $b = 0.5$.

Table 4.2 exhibits the results for the size of the test for selected significance level $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.10$ and for $c = 1.571$. In both scenarios we note that the non-robustly weighted MQLE behave in a similar way. Additionally, the robustly weighted MQLE appear to have a similar behavior as well. We observe that when there do not exist outliers in the time series, then all test statistics achieve the desired size. When there exists a patch of outliers to the time series, we observe that the behavior of the test statistics depends on the percentage of the contamination, as well as on the value of the tuning constant c. If there exist 5 outliers in the series, that is 1% of contamination, then values of the tuning constant over 2.50 make the non-robustly weighted MQLE based test statistics to diverge from the desired size. The robustly weighted MQLE based test statistics do not have this behavior. Increasing further the percentage of contamination to 2% (10 outliers), 3% (15 outliers) and 4% (20 outliers) then, the non-robustly weighted MQLE based test statistics do not achieve the desirable size. The test statistics based on robust weighting depends on the value of the tuning constant. In particular, the smaller the data contamination the larger the value of the tuning constant for which the (4.8) diverges from the desirable size. If there exist isolated outliers in the series, then the non-robustly weighted version of (4.8) does not achieve the desired size regardless of the number of outliers and the value of the tuning constant but the robustly weighted MQLE are not affected by either the number of outliers nor the value of the tuning constant, especially for smaller percentage of contamination.

		Patch of Outliers			Isolated Outliers		
Number of outliers	Weights	Significance level			Significance level		
		$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
no outliers	none	0.003	0.047	0.082	0.009	0.055	0.113
	hat	0.003	0.046	0.084	0.006	0.055	0.114
	mve	0.008	0.049	0.102	0.012	0.056	0.104
	mcd	0.007	0.036	0.090	0.007	0.054	0.102
5 outliers	none	0.009	0.042	0.088	0.028	0.076	0.129
	hat	0.007	0.036	0.084	0.015	0.075	0.132
	mve	0.011	0.046	0.102	0.008	0.048	0.092
	mcd	0.007	0.037	0.091	0.008	0.056	0.113
10 outliers	none	0.281	0.528	0.683	0.018	0.072	0.127
	hat	0.250	0.497	0.649	0.014	0.069	0.133
	mve	0.006	0.048	0.103	0.011	0.049	0.095
	mcd	0.008	0.056	0.108	0.009	0.045	0.084
15 outliers	none	0.693	0.864	0.919	0.019	0.098	0.180
	hat	0.693	0.860	0.916	0.021	0.094	0.181
	mve	0.009	0.048	0.097	0.010	0.051	0.107
	mcd	0.012	0.055	0.110	0.016	0.042	0.091
20 outliers	none	0.825	0.947	0.977	0.021	0.106	0.165
	hat	0.825	0.945	0.978	0.023	0.103	0.168
	mve	0.015	0.061	0.119	0.012	0.054	0.101
	mcd	0.026	0.084	0.155	0.013	0.053	0.106

Table 4.2: Empirical size of the test for the case of a patch of outliers and the case of isolated outliers.

To obtain the power of the test we consider the robustly weighted test statistics since they achieve the nominal size. Figure 4.9 shows the power of the test when the weights implemented are the robust weights based on the MVE and MCD algorithm. The selected value of the tuning constant is $c = 1.571$ and the nominal significance level is $\alpha = 0.05$. The results are generally consistent, showing that the power of the test increases as we move away from the null hypothesis.

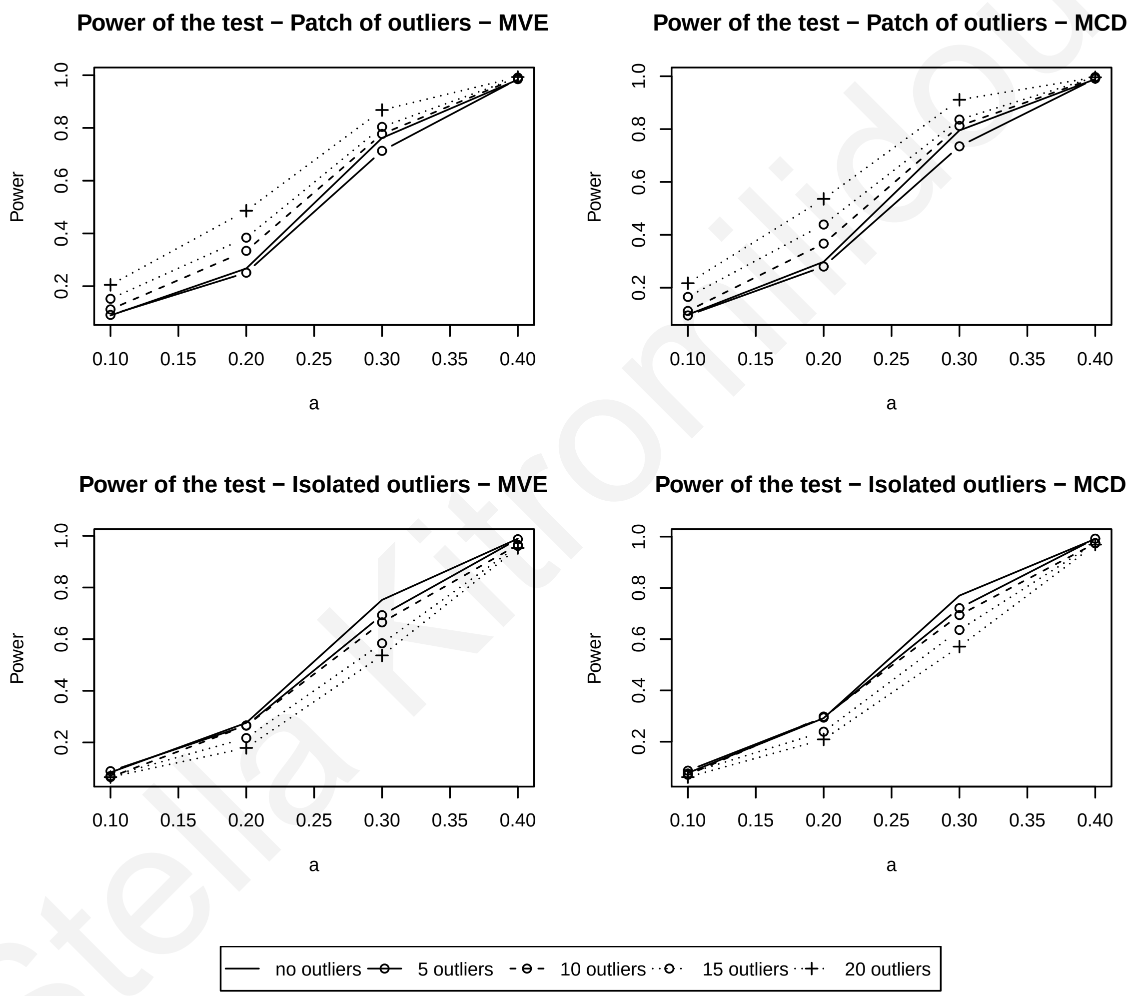

Figure 4.9: Power of the test statistic (4.8) for the cases of a patch of outliers and isolated outliers. Robust weighting is considered (MVE and MCD). The tuning constant is $c = 1.571$ and the nominal significance level is $\alpha = 0.05$.

4.3 Real Data Examples

4.3.1 Measles Data

We apply the methodology to a time series of weekly number of measles infections reported in North Rhine-Westphalia, Germany during the period from January 2001 until November 2003; see Figure 4.10. First, we examine whether the data contain some irregular observations employing the works by Chen and Liu (1993) for the case of real valued data and the method proposed by Fokianos and Fried (2012) for the case of integer valued time series. Note that in the context of count time series, the methodology suggested by Fokianos and Fried (2012) cannot detect additive outliers. The method, however, detects 8 interventions that include three transient shifts at time positions 4, 10 and 48, and five spiky outliers at positions 59, 86, 97, 109 and 122. A spiky outlier is an outlier which is defined by (4.4) for $\delta = 0$. Following the methodology of Chen and Liu (1993), we detect twenty seven interventions, including seven additive outliers, two of which are consecutive. However, we note that some of the interventions detected by the methodology of Chen and Liu (1993) need to be examined carefully because some of its discoveries are not located far apart; for instance we identify a TS followed by an AO. This is a result of the non-stationarity which is pronounced by the data; Note that the sum of the coefficients a and b is close to 1 as Table 4.3 shows.

In fact, Table 4.3 displays the estimated parameters of model (4.1) for the various estimating procedures described in the previous sections, along with the standard deviations of the estimates in parentheses. Estimation is implemented by employing the Huber function (4.6) and $M = 10$ for method B. In all cases, the MQLE provides estimates with equal or smaller standard deviation, regardless of the value of the tuning constant c. Because of the non-stationarity of the data, we choose values of c greater than 3. This choice affects directly the large values observed in the data and makes possible robust fitting of the log-linear model. Smaller values of c might result to convergence issues but for this data examples any choice of $c \geq 3$ yields robust estimators. Even though there are some differences among findings we note that in general method A and B give similar results and the sum of the regression coefficients is close to 1. Furthermore, hypothesis (4.7) is rejected by all test procedures (p-value < 0.001).

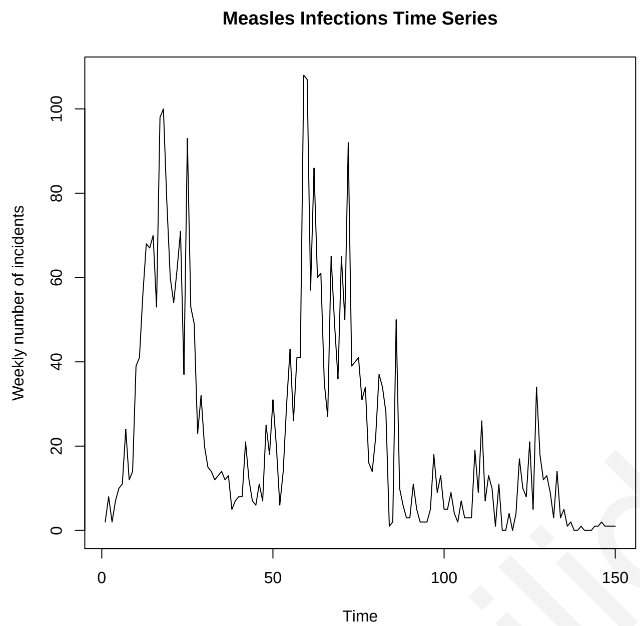

Figure 4.10: Weekly number of measles infections reported.

Estimation procedure	d	a	b
MLE	0.242(0.001)	0.435(0.010)	0.500(0.009)
MQLE no weights	0.077(0.002)	0.379(0.001)	0.587(0.001)
MQLE hat (A)	0.076(0.002)	0.378(0.001)	0.588(0.001)
MQLE hat (B)	0.024(0.002)	0.309(0.001)	0.665(0.001)
MQLE mve (A)	-0.005(0.003)	0.359(0.001)	0.628(0.001)
MQLE mcd (A)	-0.035(0.003)	0.358(0.001)	0.636(0.001)
MQLE mve (B)	0.049(0.002)	0.268(0.001)	0.697(0.001)
MQLE mcd (B)	0.067(0.002)	0.255(0.001)	0.706(0.001)

Table 4.3: Estimates and standard errors (in parentheses) of the parameters of model (4.1) when applied to the measles infection time series. Fitting is done by employing (4.6) with $c = 3$.

It is noted that an analysis of the data without feedback indicates a model of order $q = 13$.

4.3.2 Firearm Homicides Data

The Homicides data consist of the number of weekly firearm homicides recorded at the Salt River state mortuary in Cape Town during the years 1986 to 1991; Zucchini and MacDonald (2009). Figure 4.11 shows a time series plot of the data.

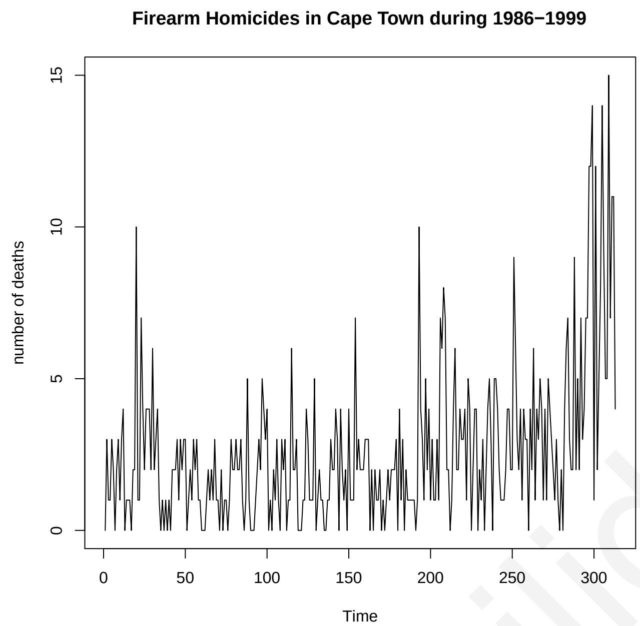

Figure 4.11: Weekly number of firearm homicides reported at the Salt River state mortuary in Cape Town during 1986-1999.

An application of the outlier detection method by Chen and Liu (1993) has detected 5 additive outliers at times 20,193,300,302,309, a transient shift at time 206 and a level shift at time 295. Table 4.4 displays the estimated parameters of model (4.1) and their standard errors. We note that using the first proposed method for the construction of the design matrix, the MVE and MCD algorithms do not work because in this case an error in the singular value decomposition is produced. Analogous problems do not occur when applying the second proposed method for constructing the design matrix. The displayed estimates on Table 4.4 correspond to the value $c = 2.429$ of the Huber function tuning constant. It is observed however that MLE, non weighted MQLE and MQLE with weights based on the hat matrix and the first method do not provide stationary solutions since the stationarity condition is not satisfied. On the contrary, MQLE weighted applying the second method maintains the stationarity condition for all values of the tuning constant considered.

In all cases, the MQLE provides estimates with smaller standard deviations, regardless of the value of the tuning constant c. Here we demonstrate the estimation results for $c = 2.429$-see Table 4.4. Furthermore, hypothesis (4.7) is rejected by all test procedures (p-value < 0.001).

Estimation procedure	d	a	b
MQLE hat (B)	0.060(0.054)	0.396(0.061)	0.303(0.043)
MQLE mve (B)	0.046(0.055)	0.407(0.061)	0.310(0.044)
MQLE mcd (B)	0.026(0.055)	0.425(0.061)	0.322(0.044)

Table 4.4: Estimates of the parameters of model (4.1) for the firearm homicides time series, $c = 2.429$.

4.4 Appendix

In the following, the symbol C denotes a constant which depends upon the context. Define also $d_M = \max(|d_L|, |d_U|)$, $a_M = \max(|a_L|, |a_U|)$ and $b_M = \max(|b_L|, |b_U|)$. In addition, when a quantity is evaluated at the true value of the parameter $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}_0$, then the notation will be simplified by dropping $\boldsymbol{\theta}_0$. For instance, $m_t \equiv m_t(\boldsymbol{\theta}_0)$ and so on. The following two results are taken from Fokianos and Tjøstheim (2011) and are included for completeness.

Lemma A-1 Assume model (4.2) and suppose that $|a| < 1$. In addition, assume that when $b > 0$ then $|a + b| < 1$, and when $b < 0$ then $|a||a + b| < 1$. Then, the following conclusions hold:

1. The process $\{\nu_t^m, t \geq 0\}$ is a geometrically ergodic Markov chain with finite moments of order k, for an arbitrary k.

2. The process $\{(Y_t^m, U_t, \nu_t^m), t \geq 0\}$ is a $V_{(Y,U,\nu)}$-geometrically ergodic Markov chain with $V_{Y,U,\nu}(Y, U, \nu) = 1 + \log^{2k}(1 + Y) + \nu^{2k} + U^{2k}$, k being a positive integer.

Lemma A-2 Suppose that (Y_t, ν_t) and (Y_t^m, ν_t^m) are defined by (4.1) and (4.2) respectively. Assume that $|a + b| < 1$, if a and b have the same sign, and $a^2 + b^2 < 1$ if a and b have different signs. Then the following statements are true:

1. $E|\nu_t^m - \nu_t| \to 0$ and $|\nu_t^m - \nu_t| < \delta_{1,m}$ almost surely for m large.

2. $E(\nu_t^m - \nu_t)^2 \leq \delta_{2,m}$,

3. $E|\lambda_t^m - \lambda_t| \leq \delta_{3,m}$,

4. $E|Y_t^m - Y_t| \leq \delta_{4,m}$,

5. $E(\lambda_t^m - \lambda_t)^2 \leq \delta_{5,m}$,

6. $E(Y_t^m - Y_t)^2 \leq \delta_{6,m}$,

where $\delta_{i,m} \to 0$ as $m \to \infty$ for $i = 1, ..., 6$. Furthermore, almost surely, with m large enough

$$|\lambda_t^m - \lambda_t| \leq \delta \quad \text{and} \quad |Y_t^m - Y_t| \leq \delta, \quad \text{for any} \quad \delta > 0.$$

We will also need the following lemmas whose proof is given below.

Lemma A-3 Define the Pearson residuals for both perturbed and unperturbed models by

$$r_t^m = \frac{Y_t^m - e^{\nu_t^m}}{e^{\nu_t^m/2}}, \quad r_t = \frac{Y_t - e^{\nu_t}}{e^{\nu_t/2}}$$

respectively. Suppose that (Y_t, ν_t) and (Y_t^m, ν_t^m) are defined by (4.1) and (4.2) respectively. Assume that $|a + b| < 1$, if a and b have the same sign, and $a^2 + b^2 < 1$ if a and b have different signs. Then,

1. $E|r_t^m - r_t| \to 0$,

2. $E(r_t^m - r_t)^2 \leq \delta_{7;m}$,

where $\delta_{7,m} \to 0$ as $m \to \infty$. Furthermore, almost surely, with m large enough

$$|r_t^m - r_t| \leq \delta \quad \text{for any} \quad \delta > 0.$$

Proof of Lemma A-3. We have that

$$
\begin{aligned}
|r_t^m - r_t| &= \left| \frac{Y_t^m - e^{\nu_t^m}}{e^{\nu_t^m/2}} - \frac{Y_t - e^{\nu_t}}{e^{\nu_t/2}} \right| = \left| \frac{Y_t^m}{e^{\nu_t^m/2}} - \frac{Y_t}{e^{\nu_t/2}} + \left(e^{\nu_t/2} - e^{\nu_t^m/2}\right) \right| \\
&\leq \left| \frac{Y_t^m e^{\nu_t/2} - Y_t e^{\nu_t^m/2} \pm Y_t^m e^{\nu_t^m/2}}{e^{\nu_t^m/2} e^{\nu_t/2}} \right| + \left| e^{\nu_t/2} - e^{\nu_t^m/2} \right| \\
&= \left| \frac{(Y_t^m - Y_t) e^{\nu_t^m/2} + Y_t^m \left(e^{\nu_t/2} - e^{\nu_t^m/2}\right)}{e^{\nu_t^m/2} e^{\nu_t/2}} \right| + \left| e^{\nu_t^m/2} - e^{\nu_t/2} \right| \\
&\leq \left| \frac{Y_t^m - Y_t}{e^{\nu_t/2}} \right| + \left| \frac{Y_t^m \left(e^{\nu_t/2} - e^{\nu_t^m/2}\right)}{e^{\nu_t^m/2} e^{\nu_t/2}} \right| + \left| e^{\nu_t^m/2} - e^{\nu_t/2} \right| \\
&\leq |Y_t^m - Y_t| + \left| Y_t^m \left(\sqrt{\lambda_t} - \sqrt{\lambda_t^m} \right) \right| + \left| \sqrt{\lambda_t^m} - \sqrt{\lambda_t} \right| \\
&\leq |Y_t^m - Y_t| + |Y_t^m| \sqrt{|\lambda_t^m - \lambda_t|} + \sqrt{|\lambda_t^m - \lambda_t|} \\
&\leq |Y_t^m - Y_t| + \left(|Y_t^m| + 1\right) |\lambda_t^m - \lambda_t|^{1/2} \\
&< \delta,
\end{aligned}
$$

for any $\delta > 0$ almost surely and for m large enough by using the results of Lemma A-2. In addition, we obtain the first and second part of Lemma because of uniform integrability of $|r_t^m - r_t|$. It holds that

$$\sup_t \left\{ \left(|Y_t^m| + 1\right) |\lambda_t^m - \lambda_t|^{1/2} \right\} \leq \sup_t \left\{ |Y_t^m| + 1 \right\}$$

and additionally we have that

$$E \left| |Y_t^m| + 1 \right| < \infty.$$

Therefore, the process

$$\left\{ \left(|Y_t^m| + 1\right) |\lambda_t^m - \lambda_t|^{1/2} \right\}$$

is uniformly integrable, and because it converges almost surely to zero, from the approximation Lemma A-2, we finally obtain that

$$E \left| \left(|Y_t^m| + 1\right) |\lambda_t^m - \lambda_t|^{1/2} \right| \to 0.$$

Finally,

$$E|r_t^m - r_t| \leq E|Y_t^m - Y_t| + E \left| \left(|Y_t^m| + 1\right) |\lambda_t^m - \lambda_t|^{1/2} \right| \to 0.$$

For the second part of the Lemma we have that

$$\begin{aligned}|r_t^m - r_t|^2 &\leq \left\{|Y_t^m - Y_t| + (|Y_t^m| + 1)\,|\lambda_t^m - \lambda_t|^{1/2}\right\}^2 \\ &\leq |Y_t^m - Y_t|^2 + (|Y_t^m| + 1)^2\,|\lambda_t^m - \lambda_t| + 2\,(|Y_t^m| + 1)\,|Y_t^m - Y_t|\,|\lambda_t^m - \lambda_t|^{1/2}.\end{aligned}$$

Then,

$$\begin{aligned}&\sup_t \left\{(|Y_t^m| + 1)^2\,|\lambda_t^m - \lambda_t| + 2\,(|Y_t^m| + 1)\,|Y_t^m - Y_t|\,|\lambda_t^m - \lambda_t|^{1/2}\right\} \\ &\qquad \leq \sup_t \left\{(|Y_t^m| + 1)^2 + 2\,(|Y_t^m| + 1)\right\}\end{aligned}$$

and because

$$E\left\{(|Y_t^m| + 1)^2 + 2\,(|Y_t^m| + 1)\right\} < \infty$$

we conclude that the process

$$\left\{(|Y_t^m| + 1)^2\,|\lambda_t^m - \lambda_t| + 2\,(|Y_t^m| + 1)\,|Y_t^m - Y_t|\,|\lambda_t^m - \lambda_t|^{1/2}\right\}$$

is uniformly integrable. Since it additionally converges almost surely to zero, we finally obtain that

$$E\left\{(|Y_t^m| + 1)^2\,|\lambda_t^m - \lambda_t| + 2\,(|Y_t^m| + 1)\,|Y_t^m - Y_t|\,|\lambda_t^m - \lambda_t|^{1/2}\right\} \to 0.$$

Finally,

$$\begin{aligned}E|r_t^m - r_t|^2 &\leq E|Y_t^m - Y_t|^2 + E\left\{(|Y_t^m| + 1)^2\,|\lambda_t^m - \lambda_t| + 2\,(|Y_t^m| + 1)\,|Y_t^m - Y_t|\,|\lambda_t^m - \lambda_t|^{1/2}\right\} \\ &\to 0\end{aligned}$$

□

Lemma A-4 For any unbiased estimating function $E\left(\psi(x;\theta)\right) = 0$, it holds true that

$$E\left[\psi(x;\theta)\frac{\partial \ell(x;\theta)}{\partial \theta}\right] = -E\left[\frac{\partial \psi(x;\theta)}{\partial \theta}\right]$$

where $\ell(x;\theta)$ denotes the log-likelihood function.

Proof of Lemma A-4.

$$
\begin{aligned}
E\left[\psi(x;\theta)\frac{\partial \ell(x;\theta)}{\partial\theta}\right] &= \int \psi(x;\theta)\frac{\partial \ell(x;\theta)}{\partial\theta} f(x;\theta)dx \\
&= \int \psi(x;\theta)\frac{\partial}{\partial\theta}\log f(x;\theta)\cdot f(x;\theta)dx \\
&= \int \psi(x;\theta)\frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial\theta} f(x;\theta)dx \\
&= \int \left[\psi(x;\theta)\frac{\partial f(x;\theta)}{\partial\theta} \pm \frac{\partial\psi(x;\theta)}{\partial\theta}f(x;\theta)\right]dx \\
&= \int \left[\psi(x;\theta)\frac{\partial f(x;\theta)}{\partial\theta} + \frac{\partial\psi(x;\theta)}{\partial\theta}f(x;\theta)\right]dx - \int \frac{\partial\psi(x;\theta)}{\partial\theta}f(x;\theta)dx \\
&= \int \frac{\partial}{\partial\theta}\left(\psi(x;\theta)f(x;\theta)\right)dx - E\left[\frac{\partial\psi(x;\theta)}{\partial\theta}\right] \\
&= \frac{\partial}{\partial\theta}\int \left(\psi(x;\theta)f(x;\theta)\right)dx - E\left[\frac{\partial\psi(x;\theta)}{\partial\theta}\right] \\
&= \frac{\partial}{\partial\theta}E\left(\psi(x;\theta)\right) - E\left[\frac{\partial\psi(x;\theta)}{\partial\theta}\right] = -E\left[\frac{\partial\psi(x;\theta)}{\partial\theta}\right].
\end{aligned}
$$

□

Proof of Lemma 4.1.1. We will show that

$$E\left(m_t^m(\boldsymbol{\theta})\left(m_t^m(\boldsymbol{\theta})\right)^T\right) - E\left(m_t(\boldsymbol{\theta})\left(m_t(\boldsymbol{\theta})\right)^T\right) \to 0 \qquad \text{(A-1)}$$

and

$$E\left(m_t^m(\boldsymbol{\theta})\right)E^T\left(m_t^m(\boldsymbol{\theta})\right) - E\left(m_t(\boldsymbol{\theta})\right)E^T\left(m_t(\boldsymbol{\theta})\right) \to 0, \qquad \text{(A-2)}$$

as $m \to \infty$. Consider first (A-1). Working along the lines of Fokianos and Tjøstheim (2011), we consider differences of the perturbed and non perturbed matrix along the diagonal individually for $\theta_i = d, a, b$. Then, we need to evaluate

$$E\left|(Z_t^m)^2\left(\frac{\partial \nu_t^m}{\partial\theta_i}\right)^2 - Z_t^2\left(\frac{\partial \nu_t}{\partial\theta_i}\right)^2\right|,$$

with $Z_t = \psi(r_t) w_t e^{\nu_t/2}$ and similarly for Z_t^m. We have,

$$
\begin{aligned}
&E\left|(Z_t^m)^2\left(\frac{\partial \nu_t^m}{\partial \theta_i}\right)^2 - Z_t^2\left(\frac{\partial \nu_t}{\partial \theta_i}\right)^2\right| = E\left|(Z_t^m)^2\left(\frac{\partial \nu_t^m}{\partial \theta_i}\right)^2 - Z_t^2\left(\frac{\partial \nu_t}{\partial \theta_i}\right)^2 \pm (Z_t^m)^2\left(\frac{\partial \nu_t}{\partial \theta_i}\right)^2\right| \\
&E\left|(Z_t^m)^2\left[\left(\frac{\partial \nu_t^m}{\partial \theta_i}\right)^2 - \left(\frac{\partial \nu_t}{\partial \theta_i}\right)^2\right] + ((Z_t^m)^2 - Z_t^2)\left(\frac{\partial \nu_t}{\partial \theta_i}\right)^2\right| \\
&\leq E\left|(Z_t^m)^2\left[\left(\frac{\partial \nu_t^m}{\partial \theta_i}\right)^2 - \left(\frac{\partial \nu_t}{\partial \theta_i}\right)^2\right]\right| + E\left|((Z_t^m)^2 - Z_t^2)\left(\frac{\partial \nu_t}{\partial \theta_i}\right)^2\right| \\
&\leq CE\Big|\exp(\nu_t^m)\Big(\left(\frac{\partial \nu_t^m}{\partial \theta_i}\right) - \left(\frac{\partial \nu_t}{\partial \theta_i}\right)\Big)\Big(\left(\frac{\partial \nu_t^m}{\partial \theta_i}\right) + \left(\frac{\partial \nu_t}{\partial \theta_i}\right)\Big)\Big| + \sqrt{E\Big((Z_t^m)^2 - Z_t^2\Big)^2 E\left(\frac{\partial \nu_t}{\partial \theta_i}\right)^4}
\end{aligned}
$$

The first term can become arbitrarily small because it can be shown (following the proof of Fokianos et al. (2009, Lemma 3.1)) that as $m \to \infty$

$$
\Big|\frac{\partial \nu_t^m}{\partial \theta_i} - \frac{\partial \nu_t}{\partial \theta_i}\Big| < \delta, \quad E\Big|\frac{\partial \nu_t^m}{\partial \theta_i} - \frac{\partial \nu_t}{\partial \theta_i}\Big| < \delta, \quad E\Big(\frac{\partial \nu_t^m}{\partial \theta_i} - \frac{\partial \nu_t}{\partial \theta_i}\Big)^2 < \delta \tag{A-3}
$$

almost surely for any $\delta > 0$. For the second term, note first that $E(\partial\nu_t/\partial\theta_i)^4$ is bounded by a finite constant for $i = 1, 2, 3$ since

$$
\begin{aligned}
\frac{\partial \nu_t}{\partial d} &\leq \frac{1}{(1 - a_M)}, \\
\frac{\partial \nu_t}{\partial a} &\leq \frac{c_0}{(1 - a_M)} + b_M \sum_{i=1}^{t-1} i a_M^{i-1} \log(1 + Y_{t-i-1}), \\
\frac{\partial \nu_t}{\partial b} &\leq \sum_{i=0}^{t-1} a_M^{i-1} \log(1 + Y_{t-i-1}),
\end{aligned} \tag{A-4}
$$

by using Lemma A-1.

Furthermore

$$\begin{aligned} \left|(Z_t^m)^2 - Z_t^2\right| &\le \left|\psi^2(r_t^m)e^{\nu_t^m} - \psi^2(r_t)e^{\nu_t}\right| = \left|\psi^2(r_t^m)e^{\nu_t^m} - \psi^2(r_t)e^{\nu_t} \pm \psi^2(r_t)e^{\nu_t^m}\right| \\ &= \left|(\psi^2(r_t^m) - \psi^2(r_t))e^{\nu_t^m} + \psi^2(r_t)\left(e^{\nu_t^m} - e^{\nu_t}\right)\right| \\ &\le |\psi^2(r_t^m) - \psi^2(r_t)|\lambda_t^m + \psi^2(r_t)|\lambda_t^m - \lambda_t| \\ &= |(\psi(r_t^m) - \psi(r_t))(\psi(r_t^m) + \psi(r_t))|\lambda_t^m + \psi^2(r_t)|\lambda_t^m - \lambda_t| \\ &\le C\Big(|r_t^m - r_t|\lambda_t^m + |\lambda_t^m - \lambda_t|\Big) < \delta \end{aligned}$$

where we have used the boundedness of the function $\psi(\cdot)$, the mean-value theorem and Lemmas A-2 and A-3. Then,

$$E\left|(Z_t^m)^2 - Z_t^2\right|^2 \le C\left\{E|\lambda_t^m - \lambda_t|^2 + E\Big(|r_t^m - r_t|^2(\lambda_t^m)^2 + 2\lambda_t^m|r_t^m - r_t||\lambda_t^m - \lambda_t|\Big)\right\}.$$

The process

$$\left\{|r_t^m - r_t|^2(\lambda_t^m)^2 + 2\lambda_t^m|r_t^m - r_t||\lambda_t^m - \lambda_t|\right\}$$

is uniformly integrable, because

$$\sup_t\left\{|r_t^m - r_t|^2(\lambda_t^m)^2 + 2\lambda_t^m|r_t^m - r_t||\lambda_t^m - \lambda_t|\right\} \le \sup_t\left\{(\lambda_t^m)^2 + 2\lambda_t^m\right\}$$

and

$$E\left|(\lambda_t^m)^2 + 2\lambda_t^m\right| \le E\left|e^{\nu_t^m}\right|^2 + 2E\left|e^{\nu_t^m}\right|$$

is bounded. Also, because the process converges almost surely to zero, we obtain that

$$E\left||r_t^m - r_t|^2(\lambda_t^m)^2 + 2\lambda_t^m|r_t^m - r_t||\lambda_t^m - \lambda_t|\right| \to 0.$$

Therefore,

$$E\left|(Z_t^m)^2 - Z_t^2\right| \le \delta.$$

Hence (A-1) follows.

To prove (A-2), consider

$$\left|E^2\Big(Z_t^m\frac{\partial\nu_t^m}{\partial\theta_i}\Big)-E^2\Big(Z_t\frac{\partial\nu_t}{\partial\theta_i}\Big)\right|\le E\left|Z_t^m\frac{\partial\nu_t^m}{\partial\theta_i}-Z_t\frac{\partial\nu_t}{\partial\theta_i}\right|E\left|Z_t^m\frac{\partial\nu_t^m}{\partial\theta_i}+Z_t\frac{\partial\nu_t}{\partial\theta_i}\right|$$
$$\le CE\left|Z_t^m\frac{\partial\nu_t^m}{\partial\theta_i}-Z_t\frac{\partial\nu_t}{\partial\theta_i}\pm Z_t^m\frac{\partial\nu_t}{\partial\theta_i}\right|E\left|e^{\nu_t^m/2}\frac{\partial\nu_t^m}{\partial\theta_i}+e^{\nu_t/2}\frac{\partial\nu_t}{\partial\theta_i}\pm e^{\nu_t^m/2}\frac{\partial\nu_t}{\partial\theta_i}\right|$$
$$=CE\left|Z_t^m\left(\frac{\partial\nu_t^m}{\partial\theta_i}-\frac{\partial\nu_t}{\partial\theta_i}\right)+(Z_t^m-Z_t)\frac{\partial\nu_t}{\partial\theta_i}\right|E\left|e^{\nu_t^m/2}\left(\frac{\partial\nu_t^m}{\partial\theta_i}+\frac{\partial\nu_t}{\partial\theta_i}\right)+\left(e^{\nu_t^m/2}-e^{\nu_t/2}\right)\frac{\partial\nu_t}{\partial\theta_i}\right|.$$

The above quantity can be made arbitrarily small because of finite moments of $\partial\nu_t/\partial\theta_i$, $\partial\nu_t^m/\partial\theta_i$, $\exp(\nu_t^m)$, (A-3) and the fact that $E\left|(Z_t^m-Z_t)\ \partial\nu_t/\partial\theta_i\right|\to 0$, as $\to\infty$ which is proved following the previous arguments. Indeed, considering each expected value in the above representation separately we have:

for the first expected value,

$$\begin{aligned}E\left|Z_t^m\left(\frac{\partial\nu_t^m}{\partial\theta_i}-\frac{\partial\nu_t}{\partial\theta_i}\right)\right| &\le CE\left|e^{\nu_t^m/2}\left(\frac{\partial\nu_t^m}{\partial\theta_i}-\frac{\partial\nu_t}{\partial\theta_i}\right)\right|\\ &\le C\sqrt{E|e^{\nu_t^m}|E\left(\frac{\partial\nu_t^m}{\partial\theta_i}-\frac{\partial\nu_t}{\partial\theta_i}\right)^2}\\ &< \delta,\end{aligned}$$

for the third expected value,

$$\begin{aligned}E\left|e^{\nu_t^m/2}\left(\frac{\partial\nu_t^m}{\partial\theta_i}+\frac{\partial\nu_t}{\partial\theta_i}\right)\right| &\le E\left|e^{\nu_t^m/2}\left(\frac{\partial\nu_t^m}{\partial\theta_i}\right)\right|+E\left|e^{\nu_t^m/2}\left(\frac{\partial\nu_t}{\partial\theta_i}\right)\right|\\ &\le \sqrt{E|e^{\nu_t^m}|\left(\frac{\partial\nu_t^m}{\partial\theta_i}\right)^2}+\sqrt{E|e^{\nu_t^m}|\left(\frac{\partial\nu_t}{\partial\theta_i}\right)^2}<\infty,\end{aligned}$$

for the fourth expected value

$$\begin{aligned}E\left|\left(e^{\nu_t^m/2}-e^{\nu_t/2}\right)\frac{\partial\nu_t}{\partial\theta_i}\right| &= E\left|(\sqrt{\lambda_t^m}-\sqrt{\lambda_t})\frac{\partial\nu_t}{\partial\theta_i}\right|\\ &\le E\left|\sqrt{|\lambda_t^m-\lambda_t|}\frac{\partial\nu_t}{\partial\theta_i}\right|\\ &\le \sqrt{E|\lambda_t^m-\lambda_t|E\left(\frac{\partial\nu_t}{\partial\theta_i}\right)^2}<\delta\end{aligned}$$

and for the second expected value

$$E\left|(Z_t^m - Z_t)\,\frac{\partial \nu_t}{\partial \theta_i}\right| \leq \sqrt{E\,(Z_t^m - Z_t)^2\, E\left(\frac{\partial \nu_t}{\partial \theta_i}\right)^2}.$$

But,

$$\begin{aligned}
|Z_t^m - Z_t| &\leq |\psi(r_t^m)e^{\nu_t^m/2} - \psi(r_t)e^{\nu_t/2}| = |\psi(r_t^m)e^{\nu_t^m/2} - \psi(r_t)e^{\nu_t/2} \pm \psi(r_t)e^{\nu_t^m/2}| \\
&= \left|(\psi(r_t^m) - \psi(r_t))e^{\nu_t^m/2} + \psi(r_t)\left(e^{\nu_t^m/2} - e^{\nu_t/2}\right)\right| \\
&\leq |\psi(r_t^m) - \psi(r_t)||e^{\nu_t^m/2}| + |\psi(r_t)||\sqrt{\lambda_t^m} - \sqrt{\lambda_t}| \\
&\leq C\{|r_t^m - r_t||e^{\nu_t^m/2}| + \sqrt{|\lambda_t^m - \lambda_t|} < \delta
\end{aligned}$$

and by taking expectation of the square

$$E(Z_t^m - Z_t)^2 \leq C\{E|\lambda_t^m - \lambda_t| + E\left(|r_t^m - r_t|^2|e^{\nu_t^m}| + 2|r_t^m - r_t||e^{\nu_t^m/2}|\sqrt{|\lambda_t^m - \lambda_t|}\right)\}$$

It can be shown that the process

$$\{|r_t^m - r_t|^2|e^{\nu_t^m}| + 2|r_t^m - r_t||e^{\nu_t^m/2}|\sqrt{|\lambda_t^m - \lambda_t|}\}$$

is uniformly integrable because

$$\sup_t\{|r_t^m - r_t|^2|e^{\nu_t^m}| + 2|r_t^m - r_t||e^{\nu_t^m/2}|\sqrt{|\lambda_t^m - \lambda_t|}\} \leq \sup_t\{|e^{\nu_t^m}| + 2|e^{\nu_t^m/2}|\}$$

and

$$E\left(|e^{\nu_t^m}| + 2|e^{\nu_t^m/2}|\right) < \infty.$$

Also, in addition to uniform integrability, the process converges almost surely to zero, therefore its expected value converges to zero. These show that

$$E(Z_t^m - Z_t)^2 < \delta.$$

□

Proof of Lemma 4.1.2. The score function S_n^m for the perturbed model is a martingale sequence, with $E(S_n^m|\mathcal{F}_{t-1}^m) = S_{n-1}^m$ at the true value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\mathcal{F}_{t-1}^m$ denotes the σ-field generated by $\{Y_0^m, ..., Y_{t-1}^m, \mathcal{U}_0, ..., \mathcal{U}_{t-1}\}$. This is because the following statements hold true

i. $\mathcal{F}_n^m \subset \mathcal{F}_{n+1}^m$ since $\mathcal{F}_n^m$ denotes the σ-field generated by $\{Y_0, ..., Y_n, \mathcal{U}_0, ..., \mathcal{U}_n\}$ and $\mathcal{F}_{n+1}^m$ denotes the σ-field generated by $\{Y_0, ..., Y_n, Y_{n+1}, \mathcal{U}_0, ..., \mathcal{U}_n, \mathcal{U}_{n+1}\}$,

ii. S_n^m is $\mathcal{F}_n^m$-measurable,

iii. $E \mid S_n^m \mid < \infty$ since the score is centered and

iv.

$$
\begin{aligned}
E(S_n^m \parallel \mathcal{F}_{n-1}^m) &= E\left(\sum_{t=1}^{n} s_t^m \parallel \mathcal{F}_{n-1}^m)\right) = E\left(\sum_{t=1}^{n-1} s_t^m + s_n^m \parallel \mathcal{F}_{n-1}^m)\right) \\
&= E\left(\sum_{t=1}^{n-1} s_t^m \parallel \mathcal{F}_{n-1}^m)\right) + E\left(s_n^m \parallel \mathcal{F}_{n-1}^m)\right) \\
&= E(S_{n-1}^m \parallel \mathcal{F}_{n-1}^m) + \underbrace{E\left(m_n^m - E(m_n^m \parallel \mathcal{F}_{n-1}^m)\right)}_{=0} \\
&= S_{n-1}^m.
\end{aligned}
$$

We will show that it is square integrable. Proving that $E||s_t^m||^2$ is finite for $\boldsymbol{\theta}_0 = d_0, a_0$ and b_0 guarantees an application of the strong law of large numbers for martingales (Chow (1967)), which gives almost sure convergence to 0 of S_n^m/n as $n \to \infty$. But

$$
E\left\{\left|\psi(r_t^m) w_t e^{\nu_t^m/2} \frac{\partial \nu_t^m}{\partial \theta_i}\right|\right\}^2 \leq C \left(E|e^{\nu_t^m}|^2\right)^{1/2} \left(E\left(\frac{\partial \nu_t^m}{\partial \theta_i}\right)^4\right)^{1/2}
$$

and this is finite because of Lemma A-1 and (A-4). Finite moments of $|e^{\nu_t^m}|^2$ and $(\partial \nu_t^m/\partial \theta_i)^4$ show that the above is finite and conclude the first assertion of the Lemma. To show asymptotic normality of the perturbed score function S_n^m we apply the CLT for martingales (Hall and sHeyde (1980, Cor. 3.1). $(S_n^m)_{n\geq 1}$ is a zero mean, square integrable martingale sequence with $(s_t^m)_{t\geq \mathbb{N}}$ a martingale difference sequence, and so a conditional Lindeberg condition and a condition on the conditional variance hold. To prove the conditional Lindeberg's condition, denote by $\mathbb{1}(.)$ the indicator function and note that

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n} E(\| s_t^m \|^2 \ \mathbb{1}(\| s_t^m \| > \sqrt{n}\delta)|\mathcal{F}_{t-1}^m) &\leq \frac{1}{n}\sum_{t=1}^{n} E\left\{\frac{||s_t^m||^4}{n\delta^2}|\mathcal{F}_{t-1}\right\} \\ &= \frac{1}{n^2\delta^2}\sum_{t=1}^{n} E(||s_t^m||^4|\mathcal{F}_{t-1}) \xrightarrow{\text{p}} 0\end{aligned}$$

because $E||s_t^m||^4 < \infty$ and by applying the Lyapounov condition since the indicator function implies that

$$\begin{aligned}||s_t^m||^2\mathbb{1}(\| s_t^m \| > \sqrt{n}\delta) &\leq ||s_t^m||^2\mathbb{1}(\| s_t^m \| > \sqrt{n}\delta)\frac{||s_t^m||^2}{n\delta^2} \\ &\leq \frac{||s_t^m||^4}{n\delta^2}\mathbb{1}(\| s_t^m \| > \sqrt{n}\delta) \\ &\leq \frac{||s_t^m||^4}{n\delta^2}.\end{aligned}$$

In addition, on the conditional variance

$$\frac{1}{n}\sum_{t=1}^{n} Var(s_t^m|\mathcal{F}_{t-1}^m) \xrightarrow{p} E\Big\{E[(s_t^m)(s_t^m)^T|\mathcal{F}_{t-1}^m]\Big\} = E\left(s_t^m(s_t^m)\right) = W^m.$$

This concludes the second result of the Lemma.

The third result of the Lemma is identical to Lemma 4.1.1 by Prop. 6.3.9. of Brockwell and Davis (1991). Consider now the last result of the Lemma.

$$\begin{aligned}\frac{1}{\sqrt{n}}(S_n^m - S_n) &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{s_t^m - s_t\} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(W_t^m\frac{\partial\nu_t^m}{\partial\boldsymbol{\theta}} - W_t\frac{\partial\nu_t}{\partial\boldsymbol{\theta}}\right) \\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(W_t^m\frac{\partial\nu_t^m}{\partial\boldsymbol{\theta}} - W_t\frac{\partial\nu_t}{\partial\boldsymbol{\theta}} \pm W_t^m\frac{\partial\nu_t}{\partial\boldsymbol{\theta}}\right) \\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[W_t^m\left(\frac{\partial\nu_t^m}{\partial\boldsymbol{\theta}} - \frac{\partial\nu_t}{\partial\boldsymbol{\theta}}\right) + (W_t^m - W_t)\frac{\partial\nu_t}{\partial\boldsymbol{\theta}}\right],\end{aligned}$$

where $W_t = Z_t - E[Z_t \parallel \mathcal{F}_{t-1}]$ and similarly for the perturbed model. For the first summand in the above representation, we obtain that

$$\begin{aligned} P\left(\left\|\sum_{t=1}^{n} W_t^m \left(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}} - \frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)\right\| > \delta\sqrt{n}\right) &\leq P\left(\gamma_m \left|\sum_{t=1}^{n} W_t^m\right| > \delta\sqrt{n}\right) \\ &\leq \frac{\gamma_m^2}{\delta^2 n}\sum_{t=1}^{n} \mathrm{E}\,|W_t^m|^2 \leq C\gamma_m^2 \to 0, \end{aligned}$$

as $m \to \infty$, for some γ_m. For the second summand, note that

$$\begin{aligned} W_t^m &= Z_t^m - E[Z_t^m \parallel \mathcal{F}_{t-1}] = \psi(r_t^m) w_t e^{\nu_t^m/2}\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}} - E\left(\psi(r_t^m) w_t e^{\nu_t^m/2}\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}} \parallel \mathcal{F}_{t-1}\right) \\ &= \psi(r_t^m) w_t e^{\nu_t^m/2}\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}} - E\left(\psi(r_t^m) \parallel \mathcal{F}_{t-1}\right) w_t e^{\nu_t^m/2}\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}} \\ &= \left[\psi(r_t^m) - E\left(\psi(r_t^m) \parallel \mathcal{F}_{t-1}\right)\right] w_t e^{\nu_t^m/2}\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}}. \end{aligned}$$

Consider now the difference

$$\begin{aligned} W_t^m - W_t &= \left[\psi(r_t^m) - E\left(\psi(r_t^m) \parallel \mathcal{F}_{t-1}^m\right)\right] w_t e^{\nu_t^m/2}\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}} - \left[\psi(r_t) - E\left(\psi(r_t) \parallel \mathcal{F}_{t-1}\right)\right] w_t e^{\nu_t/2}\frac{\partial \nu_t}{\partial \boldsymbol{\theta}} \\ &\leq \left[\psi(r_t^m) - E\left(\psi(r_t^m) \parallel \mathcal{F}_{t-1}^m\right)\right] e^{\nu_t^m/2}\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}} - \left[\psi(r_t) - E\left(\psi(r_t) \parallel \mathcal{F}_{t-1}\right)\right] e^{\nu_t/2}\frac{\partial \nu_t}{\partial \boldsymbol{\theta}} \\ &\quad \pm \left[\psi(r_t^m) - E\left(\psi(r_t^m) \parallel \mathcal{F}_{t-1}^m\right)\right] e^{\nu_t/2}\frac{\partial \nu_t}{\partial \boldsymbol{\theta}} \\ &= \Big(\left[\psi(r_t^m) - E\left(\psi(r_t^m) \parallel \mathcal{F}_{t-1}^m\right)\right] - \left[\psi(r_t) - E\left(\psi(r_t) \parallel \mathcal{F}_{t-1}\right)\right]\Big) e^{\nu_t/2}\frac{\partial \nu_t}{\partial \boldsymbol{\theta}} \\ &\quad + \left[\psi(r_t^m) - E\left(\psi(r_t^m) \parallel \mathcal{F}_{t-1}^m\right)\right]\left(e^{\nu_t^m/2}\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}} - e^{\nu_t/2}\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right). \end{aligned}$$

For the first term,

$$\begin{aligned} &\Big|\Big(\psi(r_t^m) - E\left(\psi(r_t^m)||\mathcal{F}_{t-1}^m\right)\Big) - \Big(\psi(r_t) - E\Big(\psi(r_t)||\mathcal{F}_{t-1}\Big)\Big]\Big| \\ &= \big|[\psi(r_t^m) - \psi(r_t)] - \big[E\left(\psi(r_t^m)||\mathcal{F}_{t-1}^m\right) - E\left(\psi(r_t)||\mathcal{F}_{t-1}\right)\big]\big| \\ &= \big|[\psi(r_t^m) - \psi(r_t)] - \big[E\left(\psi(r_t^m) - \psi(r_t)||\mathcal{F}_{t-1}^m\right)\big]\big| \\ &\leq C\Big(|r_t^m - r_t| + E\left(|r_t^m - r_t|||\mathcal{F}_{t-1}^m\right)\Big) \qquad \text{(A-5)} \end{aligned}$$

and therefore its expected value tends to 0 by Lemma A-3.

For the second term,

$$
\begin{aligned}
e^{\nu_t^m/2}\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}} - e^{\nu_t/2}\frac{\partial \nu_t}{\partial \boldsymbol{\theta}} &= e^{\nu_t^m/2}\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}} - e^{\nu_t/2}\frac{\partial \nu_t}{\partial \boldsymbol{\theta}} \pm e^{\nu_t^m/2}\frac{\partial \nu_t}{\partial \boldsymbol{\theta}} \\
&= e^{\nu_t^m/2}\Big(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}} - \frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\Big) + \Big(e^{\nu_t^m/2} - e^{\nu_t/2}\Big)\frac{\partial \nu_t}{\partial \boldsymbol{\theta}} \\
&= e^{\nu_t^m/2}\Big(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}} - \frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\Big) + \Big(\sqrt{\lambda_t^m} - \sqrt{\lambda_t}\Big)\frac{\partial \nu_t}{\partial \boldsymbol{\theta}} \\
&\leq e^{\nu_t^m/2}\Big(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}} - \frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\Big) + \sqrt{|\lambda_t^m - \lambda_t|}\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}
\end{aligned}
$$

whose expected value tends to 0 by (A-4) and Lemma A-2. The fact that $E||\partial\nu_t/\partial\boldsymbol{\theta}||^2 < \infty$ yield the desired conclusion.

$$
\begin{aligned}
P\left(\|\sum_{t=1}^{n}(W_t^m - W_t)\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\| > \delta\sqrt{n}\right) &\leq P\left(\gamma_m\|\sum_{t=1}^{n}\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\| > \delta\sqrt{n}\right) \\
&\leq \frac{\gamma_m^2}{\delta^2 n}\sum_{t=1}^{n}\mathrm{E}\|\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\|^2 \leq C\gamma_m^2 \to 0,
\end{aligned}
$$

as $m \to \infty$, for some γ_m.

□

Proof of Lemma 4.1.3. Because $S_n(\boldsymbol{\theta}) = 0$ is an unbiased estimating function, it holds by Lemma A-4 that

$$
-E\left(\frac{\partial}{\partial \boldsymbol{\theta}}s_t(\boldsymbol{\theta})||\mathcal{F}_{t-1}\right) = E\left(s_t(\boldsymbol{\theta})\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}||\mathcal{F}_{t-1}\right),
$$

where $\ell_t(\boldsymbol{\theta}) = (Y_t\nu_t(\boldsymbol{\theta}) - \exp(\nu_t(\boldsymbol{\theta})))$, is the logarithm of the conditional probability of $Y_t||\mathcal{F}_{t-1}$ under the Poisson assumption. The matrices $V_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{t=1}^{n}E\left(\frac{\partial}{\partial\boldsymbol{\theta}}s_t(\boldsymbol{\theta})||\mathcal{F}_{t-1}\right)$ and $V_n^m(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{t=1}^{n}E\left(\frac{\partial}{\partial\boldsymbol{\theta}}s_t^m(\boldsymbol{\theta})||\mathcal{F}_{t-1}\right)$ are consistent estimators of the matrices V and V^m respectively. Then, the matrix $V_n(\boldsymbol{\theta})$ is rewritten in the form

$$
V_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}E\left(s_t(\boldsymbol{\theta})\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}||\mathcal{F}_{t-1}\right)
$$

and the matrix $V_n^m(\boldsymbol{\theta})$ for the perturbed model is defined analogously. We again examine the difference $s_t^m(\partial \ell_t^m/\partial\theta_i) - s_t(\partial \ell_t/\partial\theta_i)$ for $\theta_i = d, a, b$. Notice that

$$\begin{aligned}
s_t \frac{\partial \ell_t}{\partial \theta_i} &= (m_{ti} - E(m_{ti}||\mathcal{F}_{t-1}))(Y_t - e^{\nu_t})\frac{\partial \nu_t}{\partial \theta_i} \\
&= \left(\psi(r_t) w_t e^{\nu_t/2}\frac{\partial \nu_t}{\partial \theta_i} - E\left(\psi(r_t) w_t e^{\nu_t/2}\frac{\partial \nu_t}{\partial \theta_i}||\mathcal{F}_{t-1}\right)\right)(Y_t - e^{\nu_t})\frac{\partial \nu_t}{\partial \theta_i} \\
&= \left(\psi(r_t) w_t e^{\nu_t/2}\frac{\partial \nu_t}{\partial \theta_i} - E\left(\psi(r_t)||\mathcal{F}_{t-1}\right) w_t e^{\nu_t/2}\frac{\partial \nu_t}{\partial \theta_i}\right)(Y_t - e^{\nu_t})\frac{\partial \nu_t}{\partial \theta_i} \\
&= w_t e^{\nu_t} r_t \Big(\psi(r_t) - E\left(\psi(r_t)||\mathcal{F}_{t-1}\right)\Big)\left(\frac{\partial \nu_t}{\partial \theta_i}\right)^2.
\end{aligned}$$

Then,

$$\begin{aligned}
&E\left|s_t^m \frac{\partial \ell_t^m}{\partial \theta_i} - s_t \frac{\partial \ell_t}{\partial \theta_i}\right| \\
&= E\left|w_t e^{\nu_t^m} r_t^m \Big(\psi(r_t^m) - E\left(\psi(r_t^m)||\mathcal{F}_{t-1}^m\right)\Big)\left(\frac{\partial \nu_t^m}{\partial \theta_i}\right)^2 - w_t e^{\nu_t} r_t \Big(\psi(r_t) - E\left(\psi(r_t)||\mathcal{F}_{t-1}\right)\Big)\left(\frac{\partial \nu_t}{\partial \theta_i}\right)^2\right| \\
&\leq E|e^{\nu_t^m} r_t^m \left(\frac{\partial \nu_t^m}{\partial \theta_i}\right)^2 \Big(\psi(r_t^m) - E\left(\psi(r_t^m)||\mathcal{F}_{t-1}^m\right)\Big) - e^{\nu_t} r_t \left(\frac{\partial \nu_t}{\partial \theta_i}\right)^2 \Big(\psi(r_t) - E\left(\psi(r_t)||\mathcal{F}_{t-1}\right)\Big) \\
&\quad \pm e^{\nu_t^m} r_t^m \left(\frac{\partial \nu_t^m}{\partial \theta_i}\right)^2 \Big(\psi(r_t) - E\left(\psi(r_t)||\mathcal{F}_{t-1}\right)\Big)| \\
&\leq E\left|e^{\nu_t^m} r_t^m \left(\frac{\partial \nu_t^m}{\partial \theta_i}\right)^2 \Big(\left[\psi(r_t^m) - E\left(\psi(r_t^m)||\mathcal{F}_{t-1}^m\right)\right] - \left[\psi(r_t) - E\left(\psi(r_t)||\mathcal{F}_{t-1}\right)\right]\Big)\right| \\
&\quad + E\left|\left(e^{\nu_t^m} r_t^m \left(\frac{\partial \nu_t^m}{\partial \theta_i}\right)^2 - e^{\nu_t} r_t \left(\frac{\partial \nu_t}{\partial \theta_i}\right)^2\right)\left(\psi(r_t) - E\left(\psi(r_t)||\mathcal{F}_{t-1}\right)\right)\right|.
\end{aligned}$$

For the first summand, (A-5) shows that it tends to zero. More specifically, recall that the expected value of (A-5) tends to 0 as $m \to \infty$. This demonstrates that the process

$$\left\{\left[\psi(r_t^m) - E\left(\psi(r_t^m)||\mathcal{F}_{t-1}^m\right)\right] - \left[\psi(r_t) - E\left(\psi(r_t)||\mathcal{F}_{t-1}\right)\right]\right\}$$

is uniformly integrable, and because

$$\left|e^{\nu_t^m} r_t^m \left(\partial \nu_t^m/\partial \theta_i\right)^2\right|$$

is bounded by a finite constant, we obtain uniform integrability of the process

$$\left\{ e^{\nu_t^m} r_t^m \left(\frac{\partial \nu_t^m}{\partial \theta_i} \right)^2 \left[\psi(r_t^m) - E\left(\psi(r_t^m) || \mathcal{F}_{t-1}^m \right) \right] - \left[\psi(r_t) - E\left(\psi(r_t) || \mathcal{F}_{t-1} \right) \right] \right\}.$$

Also, because the above process converges almost surely to 0 as $m \to \infty$ we finally obtain that the first summand tends to zero. We work similarly for the second summand to obtain the desired result. In particular,

$$\begin{aligned}
&\left| e^{\nu_t^m} r_t^m \left(\frac{\partial \nu_t^m}{\partial \theta_i} \right)^2 - e^{\nu_t} r_t \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 \right| = \left| e^{\nu_t^m} r_t^m \left(\frac{\partial \nu_t^m}{\partial \theta_i} \right)^2 - e^{\nu_t} r_t \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 \pm e^{\nu_t^m} r_t^m \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 \right| \\
&\leq \left| e^{\nu_t^m} r_t^m \left[\left(\frac{\partial \nu_t^m}{\partial \theta_i} \right)^2 - \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 \right] \right| + \left| \left(e^{\nu_t^m} r_t^m - e^{\nu_t} r_t \right) \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 \right| \\
&= \left| e^{\nu_t^m} r_t^m \left[\left(\frac{\partial \nu_t^m}{\partial \theta_i} \right)^2 - \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 \right] \right| + \left| e^{\nu_t^m} r_t^m - e^{\nu_t} r_t \pm e^{\nu_t^m} r_t \right| \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 \\
&= \left| e^{\nu_t^m} r_t^m \left[\left(\frac{\partial \nu_t^m}{\partial \theta_i} \right)^2 - \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 \right] \right| + \left| e^{\nu_t^m} (r_t^m - r_t) + \left(e^{\nu_t^m} - e^{\nu_t} \right) r_t \right| \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 \\
&\leq \left| e^{\nu_t^m} r_t^m \right| \left| \left(\frac{\partial \nu_t^m}{\partial \theta_i} \right)^2 - \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 \right| + \left| e^{\nu_t^m} \right| \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 \left| r_t^m - r_t \right| + \left| e^{\nu_t^m} - e^{\nu_t} \right| \left| r_t \right| \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 .
\end{aligned} \tag{A-6}$$

The expected value of (A-6) tends to 0 as $m \to \infty$, by using (A-3) and Lemmas A-2 and A-3, which shows that the process

$$\left\{ e^{\nu_t^m} r_t^m \left(\frac{\partial \nu_t^m}{\partial \theta_i} \right)^2 - e^{\nu_t} r_t \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 \right\}$$

is uniformly integrable. In addition, $|\psi(r_t) - E\left(\psi(r_t) || \mathcal{F}_{t-1} \right)|$ is bounded by a finite constant. Considering also that

$$\left[e^{\nu_t^m} r_t^m \left(\frac{\partial \nu_t^m}{\partial \theta_i} \right)^2 - e^{\nu_t} r_t \left(\frac{\partial \nu_t}{\partial \theta_i} \right)^2 \right] \left[\psi(r_t) - E\left(\psi(r_t) || \mathcal{F}_{t-1} \right) \right]$$

tends to zero as $m \to \infty$, we conclude to the desired result. To show that V_n is positive definite it is sufficient to show that $z' \left(\partial \nu_t / \partial \boldsymbol{\theta} \right) \left(\partial \nu_t / \partial \boldsymbol{\theta} \right)' z > 0$ for any non-zero three dimensional real vector z. If $z' \partial \nu_t / \partial \boldsymbol{\theta} = 0$, then we obtain that $z'(1, \nu_{t-1}, \log(Y_{t-1} + 1))' =$

0. But if the last equation holds, then $z = 0$ because ν_t is expressed as a past function of $\log(Y_t + 1)$ and Y_t is non-zero for some t. The same reasoning holds for V_n^m.

□

Proof of Lemma 4.1.4. The first assertion of the Lemma holds by using a LLN. For the second, the Hessian matrix H_n can be represented as

$$H_n = \frac{1}{n}\sum_{t=1}^{n} s_t \frac{\partial l_t}{\partial \boldsymbol{\theta}} = \frac{1}{n}\sum_{t=1}^{n}\left\{ w_t e^{\nu_t} r_t \left[\psi(r_t) - E\left(\psi(r_t) || \mathcal{F}_{t-1}\right)\right] \left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)^T \right\}.$$

The matrix H_n^m for the perturbed model is defined analogously. Examining the difference $H_n^m - H_n$, we obtain that

$$\begin{aligned}
H_n^m - H_n = \frac{1}{n}\sum_{t=1}^{n} w_t \Big\{ & e^{\nu_t^m} r_t^m \left[\psi(r_t^m) - E\left(\psi(r_t^m) || \mathcal{F}_{t-1}^m\right)\right] \left(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}}\right)^T \\
& - e^{\nu_t} r_t \left[\psi(r_t) - E\left(\psi(r_t) || \mathcal{F}_{t-1}\right)\right] \left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)^T \Big\} \\
= \frac{1}{n}\sum_{t=1}^{n} w_t \Big\{ & e^{\nu_t^m} r_t^m \left[\psi(r_t^m) - E\left(\psi(r_t^m) || \mathcal{F}_{t-1}^m\right)\right] \left[\left(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}}\right)^T - \left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)^T\right] \\
+ \Big(& e^{\nu_t^m} r_t^m \left[\psi(r_t^m) - E\left(\psi(r_t^m) || \mathcal{F}_{t-1}^m\right)\right] - e^{\nu_t} r_t \left[\psi(r_t) - E\left(\psi(r_t) || \mathcal{F}_{t-1}\right)\right]\Big) \left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)^T \Big\}.
\end{aligned}$$

The second term in the above representation tends to zero as $m \to \infty$ because of the previous Lemma and the fact that $E\left\| (\partial \nu_t / \partial \boldsymbol{\theta}) (\partial \nu_t / \partial \boldsymbol{\theta})^T \right\| < \infty$. Indeed,

$$\begin{aligned}
& e^{\nu_t^m} r_t^m \left[\psi(r_t^m) - E\left(\psi(r_t^m) || \mathcal{F}_{t-1}^m\right)\right] - e^{\nu_t} r_t \left[\psi(r_t) - E\left(\psi(r_t) || \mathcal{F}_{t-1}\right)\right] \\
& = e^{\nu_t^m} r_t^m \left(\left[\psi(r_t^m) - E\left(\psi(r_t^m) || \mathcal{F}_{t-1}^m\right)\right] - \left[\psi(r_t) - E\left(\psi(r_t) || \mathcal{F}_{t-1}\right)\right]\right) \\
& \quad + \left(e^{\nu_t^m} - e^{\nu_t}\right)\left[\psi(r_t) - E\left(\psi(r_t) || \mathcal{F}_{t-1}\right)\right] \\
& \le e^{\nu_t^m} |r_t^m| \left| \left[\psi(r_t^m) - E\left(\psi(r_t^m) || \mathcal{F}_{t-1}^m\right)\right] - \left[\psi(r_t) - E\left(\psi(r_t) || \mathcal{F}_{t-1}\right)\right] \right| \\
& \quad + |\lambda_t^m - \lambda_t| \left| \left[\psi(r_t) - E\left(\psi(r_t) || \mathcal{F}_{t-1}\right)\right] \right|
\end{aligned}$$

and both terms tend to zero by (A-5) and Lemma A-2. So,

$$P\left(\left\|\sum_{t=1}^{n}\left\{e^{\nu_t^m} r_t^m\left[\psi(r_t^m)-E\left(\psi(r_t^m)||\mathcal{F}_{t-1}^m\right)\right]-e^{\nu_t} r_t\left[\psi(r_t)-E\left(\psi(r_t)||\mathcal{F}_{t-1}\right)\right]\right\}\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)^T\right\|>\epsilon n\right)$$
$$\leq \frac{\delta}{\epsilon n}\sum_{t=1}^{n} E\left\|\left\{e^{\nu_t^m} r_t^m\left[\psi(r_t^m)-E\left(\psi(r_t^m)||\mathcal{F}_{t-1}^m\right)\right]-e^{\nu_t} r_t\left[\psi(r_t)-E\left(\psi(r_t)||\mathcal{F}_{t-1}\right)\right]\right\}\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)^T\right\|$$
$$\leq C\gamma_m \to 0.$$

For the first term in the representation of $H_n^m - H_n$, we obtain the following

$$P\left(\left\|\sum_{t=1}^{n} e^{\nu_t^m} r_t^m\left[\psi(r_t^m)-E\left(\psi(r_t^m)||\mathcal{F}_{t-1}^m\right)\right]\left[\left(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}}\right)^T-\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)^T\right]\right\|\geq \epsilon n\right)$$
$$\leq \frac{1}{\epsilon n}\sum_{t=1}^{n} E\left\|e^{\nu_t^m} r_t^m\left[\psi(r_t^m)-E\left(\psi(r_t^m)||\mathcal{F}_{t-1}^m\right)\right]\left[\left(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}}\right)^T-\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)^T\right]\right\|$$
$$\to 0.$$

□

Proof of Lemma 4.1.5. Recall that the components of the MQLE score are given by

$$s_{ti}(\boldsymbol{\theta}) = m_{ti}(\boldsymbol{\theta}) - E(m_{ti}(\boldsymbol{\theta})||\mathcal{F}_{t-1}), \quad \text{where} \quad m_{ti}(\boldsymbol{\theta}) = \psi(r_t(\boldsymbol{\theta}))w_t e^{\nu_t(\boldsymbol{\theta})/2}\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial \theta_i}, \quad i = 1, 2, 3.$$

The second derivative of the i-th component of the MQLE score $\partial^2 s_{ti}(\boldsymbol{\theta})/\partial\theta_k\partial\theta_j$ is given by

$$\frac{\partial^2 s_{ti}(\boldsymbol{\theta})}{\partial\theta_k\partial\theta_j} = \frac{\partial^2 m_{ti}(\boldsymbol{\theta})}{\partial\theta_k\partial\theta_j} - E(\frac{\partial^2 m_{ti}(\theta)}{\partial\theta_k\partial\theta_j}||\mathcal{F}_{t-1}),$$

where

$$\begin{aligned}
\frac{\partial^2 m_{ti}(\boldsymbol{\theta})}{\partial\theta_k\partial\theta_j} &= \xi_{1t}(\boldsymbol{\theta})\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial\theta_k}\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial\theta_j}\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial\theta_i} \\
&+ \xi_{2t}(\boldsymbol{\theta})\left\{\frac{\partial^2 \nu_t(\boldsymbol{\theta})}{\partial\theta_k\partial\theta_j}\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial\theta_i} + \frac{\partial^2 \nu_t(\boldsymbol{\theta})}{\partial\theta_k\partial\theta_i}\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial\theta_j} + \frac{\partial^2 \nu_t(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_i}\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial\theta_k}\right\} \\
&+ \xi_{3t}(\boldsymbol{\theta})\frac{\partial^3 \nu_t(\boldsymbol{\theta})}{\partial\theta_k\partial\theta_j\partial\theta_i}.
\end{aligned}$$

where

$$
\begin{aligned}
\xi_{1t}(\boldsymbol{\theta}) &= -\frac{1}{2}\{\psi'(r_t(\boldsymbol{\theta}))e^{\nu_t(\boldsymbol{\theta})} - \frac{1}{2}\psi''(r_t(\boldsymbol{\theta}))(Y_t + e^{\nu_t(\boldsymbol{\theta})})(Y_t e^{-\nu_t(\boldsymbol{\theta})/2} + e^{\nu_t(\boldsymbol{\theta})/2}) \\
&+ \frac{1}{2}\psi'(r_t(\boldsymbol{\theta}))(Y_t + e^{\nu_t(\boldsymbol{\theta})}) - \frac{1}{2}\psi(r_t(\boldsymbol{\theta}))e^{\nu_t(\boldsymbol{\theta})/2}\}w_t \\
\xi_{2t}(\boldsymbol{\theta}) &= -\frac{1}{2}\left\{(Y_t + e^{\nu_t(\boldsymbol{\theta})})\psi'(r_t(\boldsymbol{\theta})) - \psi(r_t(\boldsymbol{\theta}))e^{\nu_t(\boldsymbol{\theta})/2}\right\}w_t \\
\xi_{3t}(\boldsymbol{\theta}) &= \psi(r_t(\boldsymbol{\theta}))w_t e^{\nu_t(\boldsymbol{\theta})/2}.
\end{aligned}
$$

Without loss of generality, we only consider derivatives with respect to a. For the derivatives with respect to d and b we use identical arguments. For the derivatives of ν_t with respect to the parameter a we obtain the following bounds:

$$
\begin{aligned}
&\nu_t \leq \mu_{0t} := b_M \sum_{j=1}^{t-1} a_M^j \log(1 + Y_{t-j-1}) + c_0, \quad \text{where} \quad c_0 = d_M/(1 - a_M) + \nu_0, \\
&\frac{\partial \nu_t}{\partial a} \leq \mu_{1t} := b_M \sum_{j=1}^{t-1} j a_M^{j-1} \log(1 + Y_{t-j-1}) + c_1, \quad \text{where} \quad c_1 = c_0/(1 - a_M), \\
&\frac{\partial^2 \nu_t}{\partial a^2} \leq \mu_{2t} := b_M \sum_{j=1}^{t-2} j(j+1) a_M^{j-1} \log(1 + Y_{t-j-2}) + c_2, \quad \text{where} \quad c_2 = 2c_0/(1 - a_M)^2, \\
&\frac{\partial^3 \nu_t}{\partial a^3} \leq \mu_{3t} := b_M \sum_{j=1}^{t-3} j(j+1)(j+2) a_M^{j-1} \log(1 + Y_{t-j-3}) + c_3, \quad \text{where} \quad c_3 = 6c_0/(1 - a_M)^3.
\end{aligned}
\tag{A-7}
$$

Derivation of the above relations is found in the end of the proof of the Lemma.

With $\theta_i = \theta_j = \theta_k = a$,

$$
\begin{aligned}
\left|\frac{\partial^2 m_{ti}}{\partial\theta_k\partial\theta_j} - E(\frac{\partial^2 m_{ti}}{\partial\theta_k\partial\theta_j}||\mathcal{F}_{t-1})\right| &< C\Big\{|\xi_{1t}(\boldsymbol{\theta}) - E(\xi_{1t}(\boldsymbol{\theta})||\mathcal{F}_{t-1})|\,\mu_{1t}^3 + |\xi_{2t}(\boldsymbol{\theta}) - E(\xi_{2t}(\boldsymbol{\theta})||\mathcal{F}_{t-1})|\,\mu_{1t}\mu_{2t} \\
&+ |\xi_{3t}(\boldsymbol{\theta}) - E(\xi_{3t}(\boldsymbol{\theta})||\mathcal{F}_{t-1})|\,\mu_{3t}\Big\} \\
&\equiv \tilde{m}_t.
\end{aligned}
\tag{A-8}
$$

$\tilde{M}_n^m$ is defined analogously. The first assertion of the Lemma holds by applying the Law of Large Numbers for geometrically ergodic processes; see Jensen and Rahbek (2007). To prove the second assertion of the Lemma, consider the difference $\tilde{m}_t^m - \tilde{m}_t$:

$$\begin{aligned}
|\tilde{m}_t^m - \tilde{m}_t| &< C| \left[\xi_{1t}^m(\boldsymbol{\theta}) - E(\xi_{1t}^m(\boldsymbol{\theta})||\mathcal{F}_{t-1}^m)\right] (\mu_{1t}^m)^3 - \left[\xi_{1t}(\boldsymbol{\theta}) - E(\xi_{1t}(\boldsymbol{\theta})||\mathcal{F}_{t-1})\right] \mu_{1t}^3 \\
&+ \left[\xi_{2t}^m(\boldsymbol{\theta}) - E(\xi_{2t}^m(\boldsymbol{\theta})||\mathcal{F}_{t-1}^m)\right] \mu_{1t}^m \mu_{2t}^m - \left[\xi_{2t}(\boldsymbol{\theta}) - E(\xi_{2t}(\boldsymbol{\theta})||\mathcal{F}_{t-1})\right] \mu_{1t}\mu_{2t} \\
&+ \left[\xi_{3t}^m(\boldsymbol{\theta}) - E(\xi_{3t}^m(\boldsymbol{\theta})||\mathcal{F}_{t-1}^m)\right] \mu_{3t}^m - \left[\xi_{3t}(\boldsymbol{\theta}) - E(\xi_{3t}(\boldsymbol{\theta})||\mathcal{F}_{t-1})\right] \mu_{3t}| \\
&\leq C\{|A_t| + |B_t| + |\Gamma_t|\}.
\end{aligned}$$

We examine separately the expected value of each of the $|A_t|$, $|B_t|$ and $|\Gamma_t|$.

$$\begin{aligned}
E|A_t| &= E| \left[\xi_{1t}^m(\boldsymbol{\theta}) - E(\xi_{1t}^m(\boldsymbol{\theta})||\mathcal{F}_{t-1}^m)\right] (\mu_{1t}^m)^3 - \left[\xi_{1t}(\boldsymbol{\theta}) - E(\xi_{1t}(\boldsymbol{\theta})||\mathcal{F}_{t-1})\right] \mu_{1t}^3| \\
&\leq E| \left[\xi_{1t}^m(\boldsymbol{\theta}) - E(\xi_{1t}^m(\boldsymbol{\theta})||\mathcal{F}_{t-1}^m)\right] ((\mu_{1t}^m)^3 - \mu_{1t}^3)| \\
&+ E| \left[(\xi_{1t}^m(\boldsymbol{\theta}) - \xi_{1t}(\boldsymbol{\theta})) - E(\xi_{1t}^m(\boldsymbol{\theta}) - \xi_{1t}(\boldsymbol{\theta})||\mathcal{F}_{t-1}^m)\right] \mu_{1t}^3|.
\end{aligned}$$

Similarly for $E|B_t|$ and $E|\Gamma_t|$. It follows that

$$E(\mu_{1t}^3) \leq C \sum_{j=1}^{t-1} j^2 a_M^{2(j-1)} < \infty, \quad \text{as} \quad a_M < 1,$$

and analogously $E(\mu_{1t}\mu_{2t}) < \infty$ and $E(\mu_{3t}) < \infty$ by using Lemma A-1. Also,

$$\begin{aligned}
|\mu_{3t}^m - \mu_{3t}| &\leq \sum_{j=1}^{t-3} j(j+1)(j+2) a_M^{j-1} |\log(1 + Y_{t-j-3}^m) - \log(1 + Y_{t-j-3})| \\
&\leq \sum_{j=1}^{t-3} j(j+1)(j+2) a_M^{j-1} |Y_{t-j-3}^m - Y_{t-j-3}|
\end{aligned}$$

tends to zero as $m \to \infty$, by Lemma A-2. The same reasoning holds for the differences $|\mu_{1t}^m - \mu_{1t}|$ and $|\mu_{2t}^m - \mu_{2t}|$. For the difference $|\mu_{1t}^m \mu_{2t}^m - \mu_{1t}\mu_{2t}|$ we have that

$$\begin{aligned}
|\mu_{1t}^m \mu_{2t}^m - \mu_{1t}\mu_{2t}| &= |\mu_{1t}^m \mu_{2t}^m - \mu_{1t}\mu_{2t} \pm \mu_{1t}^m \mu_{2t}| \\
&\leq |(\mu_{1t}^m - \mu_{1t})\mu_{2t}| + |(\mu_{2t}^m - \mu_{2t})\mu_{1t}^m|
\end{aligned}$$

which tends to zero as $m \to \infty$ following previous results. Additionally,

$$\begin{aligned}
|\xi_{3t}^m(\boldsymbol{\theta}) - \xi_{3t}(\boldsymbol{\theta})| &= |\psi(r_t^m(\boldsymbol{\theta}))e^{\nu_t^m(\boldsymbol{\theta})/2} - \psi(r_t(\boldsymbol{\theta}))e^{\nu_t(\boldsymbol{\theta})/2} \pm \psi(r_t^m(\boldsymbol{\theta}))e^{\nu_t(\boldsymbol{\theta})/2}| \\
&\leq |\psi(r_t^m(\boldsymbol{\theta}))(e^{\nu_t^m(\boldsymbol{\theta})/2} - e^{\nu_t(\boldsymbol{\theta})/2})| + |(\psi(r_t^m(\boldsymbol{\theta})) - \psi(r_t(\boldsymbol{\theta})))e^{\nu_t(\boldsymbol{\theta})/2}| \\
&\leq C(\sqrt{|\lambda_t^m(\boldsymbol{\theta}) - \lambda_t(\boldsymbol{\theta})|} + |r_t^m(\boldsymbol{\theta}) - r_t(\boldsymbol{\theta})|e^{\nu_t(\boldsymbol{\theta})/2}),
\end{aligned}$$

therefore the expected value of

$$|(\xi_{3t}^m(\boldsymbol{\theta}) - \xi_{3t}(\boldsymbol{\theta})) - E(\xi_{3t}^m(\boldsymbol{\theta}) - \xi_{3t}(\boldsymbol{\theta}) \parallel \mathcal{F}_{t-1}^m)|$$

tends to zero by Lemmas A-2 and A-3. In the same manner, we obtain that

$$\begin{aligned}
|\xi_{2t}^m(\boldsymbol{\theta}) - \xi_{2t}(\boldsymbol{\theta})| &\leq \frac{1}{2}|\xi_{3t}^m(\boldsymbol{\theta}) - \xi_{3t}(\boldsymbol{\theta})| + \frac{1}{2}|(Y_t^m + e^{\nu_t^m(\boldsymbol{\theta})})(\psi'(r_t^m(\boldsymbol{\theta})) - \psi'(r_t(\boldsymbol{\theta}))) \\
&+ (Y_t^m - Y_t + e^{\nu_t^m(\boldsymbol{\theta})} - e^{\nu_t(\boldsymbol{\theta})})\psi'(r_t(\boldsymbol{\theta}))| \\
&\leq \frac{1}{2}|\xi_{3t}^m(\boldsymbol{\theta}) - \xi_{3t}(\boldsymbol{\theta})| + \frac{1}{2}C\{(Y_t^m + e^{\nu_t^m(\boldsymbol{\theta})})(|r_t^m(\boldsymbol{\theta}) - r_t(\boldsymbol{\theta})| \\
&+ |Y_t^m - Y_t| + |\lambda_t^m(\boldsymbol{\theta}) - \lambda_t(\boldsymbol{\theta})|\}
\end{aligned}$$

and

$$\begin{aligned}
|\xi_{1t}^m(\boldsymbol{\theta}) - \xi_{1t}(\boldsymbol{\theta})| &\leq \frac{1}{4}|\psi''(r_t^m(\boldsymbol{\theta}))(Y_t^m + e^{\nu_t^m(\boldsymbol{\theta})})^2 - \psi''(r_t(\boldsymbol{\theta}))(Y_t + e^{\nu_t(\boldsymbol{\theta})})^2| \\
&+ \frac{1}{2}|\xi_{2t}^m(\boldsymbol{\theta}) - \xi_{2t}(\boldsymbol{\theta})| + \frac{1}{2}(\psi'(r_t^m(\boldsymbol{\theta}))e^{\nu_t^m(\boldsymbol{\theta})} - \psi'(r_t(\boldsymbol{\theta}))e^{\nu_t(\boldsymbol{\theta})}) \\
&= \frac{1}{4}C|(Y_t^m + e^{\nu_t^m(\boldsymbol{\theta})})^2 - (Y_t + e^{\nu_t(\boldsymbol{\theta})})^2| + \frac{1}{2}|\xi_{2t}^m(\boldsymbol{\theta}) - \xi_{2t}(\boldsymbol{\theta})| \\
&+ \frac{1}{2}|(\psi'(r_t^m(\boldsymbol{\theta})) - \psi'(r_t(\boldsymbol{\theta})))e^{\nu_t(\boldsymbol{\theta})} + \psi'(r_t^m(\boldsymbol{\theta}))(e^{\nu_t^m(\boldsymbol{\theta})} - e^{\nu_t(\boldsymbol{\theta})})| \\
&= \frac{1}{4}C|(Y_t^m + e^{\nu_t^m(\boldsymbol{\theta})})^2 - (Y_t + e^{\nu_t(\boldsymbol{\theta})})^2| + \frac{1}{2}|\xi_{2t}^m(\boldsymbol{\theta}) - \xi_{2t}(\boldsymbol{\theta})| \\
&+ \frac{1}{2}C(|r_t^m(\boldsymbol{\theta}) - r_t(\boldsymbol{\theta})|e^{\nu_t(\boldsymbol{\theta})} + |\lambda_t^m(\boldsymbol{\theta}) - \lambda_t(\boldsymbol{\theta})|)
\end{aligned}$$

whose expected value tends to zero as $m \to \infty$. This concludes the second part of the Lemma.

Finally, for the third assertion of the Lemma, we have that

$$P\left(|\tilde{M}_n^m - \tilde{M}_n| > \epsilon n\right) = P\left(\left|\sum_{t=1}^{n}\left[(\tilde{m}_t^m - \tilde{m}_t) - E(\tilde{m}_t^m - \tilde{m}_t \parallel \mathcal{F}_{t-1}^m)\right]\right| \epsilon n\right) \to 0$$

by previous arguments. This concludes the proof of the Lemma.

To show (A-4), recall that after repeated substitution, the process $\{\nu_t\}$ can be expressed as

$$\nu_t = d\frac{1-a^t}{1-a} + a^t\nu_0 + b\sum_{i=0}^{t-1} a^i \log(1 + Y_{t-i-1}).$$

Then,

$$\begin{aligned}
|\nu_t| &\leq \left|d\frac{1-a^t}{1-a}\right| + |a^t||\nu_0| + b_M\sum_{i=0}^{t-1} a_M^i \log(1 + Y_{t-i-1}) \\
&\leq \frac{d_M}{1-a_M} + |\nu_0| + b_M\sum_{i=0}^{t-1} a_M^i \log(1 + Y_{t-i-1}) \\
&= c_0 + b_M\sum_{i=0}^{t-1} a_M^i \log(1 + Y_{t-i-1}), \quad \text{where} \quad c_0 = d_M/(1-a_M).
\end{aligned}$$

On the derivatives of ν_t with respect to d:

$$\frac{\partial \nu_t}{\partial d} = \frac{1-a^t}{1-a} \leq \frac{1}{1-a} \leq \frac{1}{1-a_M}.$$

On the derivative of ν_t with respect to a:

$$\begin{aligned}
\frac{\partial \nu_t}{\partial a} &= \nu_{t-1} + a\frac{\partial \nu_{t-1}}{\partial a} = \nu_{t-1} + a\{\nu_{t-2} + a\frac{\partial \nu_{t-2}}{\partial a}\} = \cdots = \sum_{j=0}^{t-2} a^j \nu_{t-j-1} \\
&\leq \sum_{j=0}^{t-2} a_M^j \left(c_0 + b_M \sum_{i=0}^{(t-j-1)-1} a_M^i \log(1 + Y_{(t-j-1)-i-1})\right) \\
&\leq c_0\sum_{j=0}^{t-2} a_M^j + \sum_{j=0}^{t-2} a_M^j b_M \sum_{i=0}^{(t-j-1)-1} a_M^i \log(1 + Y_{(t-j-1)-i-1}),
\end{aligned}$$

where

$$\sum_{j=0}^{t-2} a_M^j = \frac{1-a^{t-1}}{1-a} \leq \frac{1}{1-a_M}$$

and

$$\sum_{j=0}^{t-2} a_M^j b_M \sum_{i=0}^{(t-j-1)-1} a_M^i \log(1+Y_{(t-j-1)-i-1}) = b_M \sum_{j=0}^{t-2} \sum_{i=0}^{(t-j-1)-1} a_M^{i+j} \log(1+Y_{(t-j-1)-i-1})$$
$$= b_M \sum_{j=0}^{t-1} j a_M^{j-1} \log(1+Y_{t-j-i})$$

which result to

$$\frac{\partial \nu_t}{\partial a} \leq c_1 + b_M \sum_{j=0}^{t-1} j a_M^{j-1} \log(1+Y_{t-j-1}), \quad \text{where} \quad c_1 = c_0/(1-a_M).$$

On the derivative of ν_t with respect to b:

$$\frac{\partial \nu_t}{\partial b} = \sum_{i=0}^{t-1} a^i \log(1+Y_{t-i-1}) \leq \sum_{i=0}^{t-1} a_M^i \log(1+Y_{t-i-1}).$$

To show the bounds on the second and third derivative of ν_t with respect to a, as in (A-7), we work as follows:

On the second derivative of ν_t with respect to a,

$$\begin{aligned}
\frac{\partial^2 \nu_t}{\partial a^2} &= \frac{\partial}{\partial a}\left(\frac{\partial \nu_t}{\partial a}\right) \leq \frac{\partial}{\partial a}\left(\sum_{j=0}^{t-2} a^j \nu_{t-j-1}\right) = \sum_{j=0}^{t-2}\left(j a^{j-1}\nu_{t-j-1} + a^j \frac{\partial \nu_{t-j-1}}{\partial a}\right) \\
&\leq \sum_{j=0}^{t-2}\left(j a_M^{j-1}|\nu_{t-j-1}| + a_M^j \frac{\partial \nu_{t-j-1}}{\partial a}\right) \\
&\leq \sum_{j=0}^{t-2} j a_M^{j-1} c_0 + \sum_{j=0}^{t-2} a_M^j \frac{c_0}{1-a_M} \\
&+ \sum_{j=0}^{t-2}\left(j a_M^{j-1} b_M \sum_{i=0}^{t-j-2} a_M^i \log(1+Y_{t-j-i-2}) + a_M^j b_M \sum_{i=1}^{t-j-2} i a_M^{i-1} \log(1+Y_{t-j-i-2})\right)
\end{aligned}$$

using the bounds on ν_t and $\partial \nu_t / \partial a$. It holds that

$$\sum_{j=0}^{t-2} j a_M^{j-1} c_0 + \sum_{j=0}^{t-2} a_M^j \frac{c_0}{1-a_M} < c_0 \frac{1}{(1-a_M)^2} + \frac{c_0}{1-a_M}\frac{1}{1-a_M} = \frac{2c_0}{(1-a_M)^2}.$$

Also,

$$\sum_{j=0}^{t-2}\left(ja_M^{j-1}b_M\sum_{i=0}^{t-j-2}a_M^i\log(1+Y_{t-j-i-2})+a_M^jb_M\sum_{i=1}^{t-j-2}ia_M^{i-1}\log(1+Y_{t-j-i-2})\right)$$
$$=\sum_{j=0}^{t-2}a_M^{j-1}b_M\sum_{i=0}^{t-j-2}(i+j)a_M^i\log(1+Y_{t-j-i-2})$$
$$=b_M\sum_{j=0}^{t-2}\sum_{i=0}^{t-j-2}(i+j)a_M^{i+j-1}\log(1+Y_{t-j-i-2})$$
$$=b_M\sum_{j=1}^{t-2}j(j+1)a_M^{j-1}\log(1+Y_{t-j-2}).$$

Therefore, we conclude that

$$\frac{\partial^2\nu_t}{\partial a^2}=c_2+b_M\sum_{j=1}^{t-2}j(j+1)a_M^{j-1}\log(1+Y_{t-j-2}),\quad \text{where}\quad c_2=\frac{2c_0}{(1-a_M)^2}.$$

On the third derivative of ν_t with respect to a,

$$
\begin{aligned}
\frac{\partial^3 \nu_t}{\partial a^3} &= \frac{\partial}{\partial a}\left(\frac{\partial^2 \nu_t}{\partial a^2}\right) \leq \frac{\partial}{\partial a}\left\{\sum_{j=0}^{t-2}\left(ja^{j-1}\nu_{t-j-1} + a^j \frac{\partial \nu_{t-j-1}}{\partial a}\right)\right\} \\
&= \sum_{j=0}^{t-2}\left(j(j-1)a^{j-2}\nu_{t-j-1} + 2ja^{j-1}\frac{\partial \nu_{t-j-1}}{\partial a} + a^j \frac{\partial^2 \nu_{t-j-1}}{\partial a^2}\right) \\
&\leq \sum_{j=0}^{t-2}\{j(j-1)a_M^{j-2}\left(c_0 + b_M \sum_{i=0}^{t-j-2} a_M^i \log(1+Y_{t-i-j-2})\right) \\
&\quad +2ja_M^{j-1}\left(c_0/(1-a_M) + b_M \sum_{i=0}^{t-j-2} ia_M^{i-1}\log(1+Y_{t-i-j-2})\right) \\
&\quad +a_M^j\left(\frac{2c_0}{(1-a_M)^2} + b_M \sum_{i=1}^{t-j-3} i(i+1)a_M^{i-1}\log(1+Y_{t-i-j-3})\right)\} \\
&= \sum_{j=0}^{t-2}\{j(j-1)a_M^{j-2}c_0 + 2ja_M^{j-1}\frac{c_0}{1-a_M} + a_M^j\frac{2c_0}{(1-a_M)^2}\} \\
&+ \sum_{j=0}^{t-2}\{j(j-1)a_M^{j-2}b_M \sum_{i=0}^{t-j-2} a_M^i \log(1+Y_{t-i-j-2}) \\
&\quad +2ja_M^{j-1}b_M \sum_{i=0}^{t-j-2} ia_M^{i-1}\log(1+Y_{t-i-j-2}) \\
&\quad +a_M^j b_M \sum_{i=1}^{t-j-3} i(i+1)a_M^{i-1}\log(1+Y_{t-i-j-3})\} \\
&= \left(\frac{2c_0}{(1-a_M)^3} + \frac{2c_0}{1-a_M}\frac{1}{(1-a_M)^2} + \frac{2c_0}{(1-a_M)^2}\frac{1}{1-a_M}\right) \\
&\quad +b_M \sum_{j=1}^{t-3} j(j+1)(j+2)a_M^{j-1}\log(1+Y_{t-j-3}) \\
&= c_3 + b_M \sum_{j=1}^{t-3} j(j+1)(j+2)a_M^{j-1}\log(1+Y_{t-j-3}), \quad \text{where} \quad c_3 = \frac{6c_0}{(1-a_M)^3}.
\end{aligned}
$$

□

Proof of Theorem 4.2.1. The first assertion of the Theorem follows from arguments given in Francq and Zakoïan (2010, Prop. 8.3). For the second assertion of the Theorem, we consider the difference

$$
\begin{aligned}
ST_n^m(\tilde{\boldsymbol{\theta}}_n) - ST_n(\tilde{\boldsymbol{\theta}}_n) &= \frac{[S_n^{(2,m)}(\tilde{\boldsymbol{\theta}}_n)]^2}{(\tilde{\sigma}^m)^2} - \frac{[S_n^{(2)}(\tilde{\boldsymbol{\theta}}_n)]^2}{\tilde{\sigma}^2} \\
&= [S_n^{(2,m)}(\tilde{\boldsymbol{\theta}}_n)]^2 \frac{\tilde{\sigma}^2 - (\tilde{\sigma}^m)^2}{(\tilde{\sigma}^m)^2 \tilde{\sigma}^2} + \frac{[S_n^{(2,m)}(\tilde{\boldsymbol{\theta}}_n)]^2 - [S_n^{(2)}(\tilde{\boldsymbol{\theta}}_n)]^2}{\tilde{\sigma}^2}.
\end{aligned}
$$

The above representation is composed of the following differences, $W_{22}^m - W_{22}$, $W_{12}^m - W_{12}$, $W_{21}^m - W_{21}$, $W_{11}^m - W_{11}$, $V_{11}^{m-1} V_{12}^m - V_{11}{}^{-1} V_{12}$, $V_{21}^m V_{11}^{m-1} - V_{21} V_{11}^{-1}$ and $V_{11}^{m-1} V_{12}^m - V_{11}^{-1} V_{12}$ which all converge to zero as results of Lemmas 4.1.1 and 4.1.3. In particular, for the first summand we have that

$$
\begin{aligned}
|\tilde{\sigma}^2 - (\tilde{\sigma}^m)^2| &= |(W_{22} - V_{21}(V_{11})^{-1} W_{12} - W_{21}(V_{11})^{-1} V_{12} + V_{21}(V_{11})^{-1} W_{11} (V_{11})^{-1} V_{12}) \\
&- \;(W_{22}^m - V_{21}^m V_{11}^{m-1} W_{12}^m - W_{21}^m V_{11}^{m-1} V_{12}^m + V_{21}^m (V_{11}^m)^{-1} W_{11}^m V_{11}^{m-1} V_{12}^m)| \\
&= (W_{22}^m - W_{22}) + (V_{21}^m (V_{11}^m)^{-1} W_{12}^m - V_{21} V_{11}{}^{-1} W_{12}) \\
&+ \;(W_{21}^m (V_{11}^m)^{-1} V_{12}^m - W_{21} V_{11}{}^{-1} V_{12}) \\
&+ \;(V_{21}^m (V_{11}^m)^{-1} W_{11}^m (V_{11}^m)^{-1} V_{12}^m - V_{21} V_{11}{}^{-1} W_{11} V_{11}{}^{-1} V_{12}) \\
&\leq |W_{22}^m - W_{22}| + |V_{21}^m (V_{11}^m)^{-1} W_{12}^m - V_{21} V_{11}{}^{-1} W_{12}| \\
&+ \;|W_{21}^m (V_{11}^m)^{-1} V_{12}^m - W_{21} V_{11}{}^{-1} V_{12}| \\
&+ \;|V_{21}^m (V_{11}^m)^{-1} W_{11}^m (V_{11}^m)^{-1} V_{12}^m - V_{21} V_{11}{}^{-1} W_{11} V_{11}{}^{-1} V_{12}|
\end{aligned}
$$

and the following hold true:

- $W_{22}^m - W_{22}$ converges to zero from Lemma 4.1.1.
- $V_{21}^m (V_{11}^m)^{-1} W_{12}^m - V_{21} V_{11}{}^{-1} W_{12} = V_{21}^m (V_{11}^m)^{-1} (W_{12}^m - W_{12}) + (V_{21}^m (V_{11}^m)^{-1} - V_{21} V_{11}{}^{-1}) W_{12}$ converges to zero from Lemmas 4.1.1 and 4.1.3.
- $W_{21}^m (V_{11}^m)^{-1} V_{12}^m - W_{21} V_{11}^{-1} V_{12} = W_{21}^m ((V_{11}^m)^{-1} V_{12}^m - V_{11}{}^{-1} V_{12}) + (W_{21}^m - W_{21}) V_{11}{}^{-1} V_{12}$ converges to zero from Lemmas 4.1.1 and 4.1.3.
- $V_{21}^m (V_{11}^m)^{-1} W_{11}^m (V_{11}^m)^{-1} V_{12}^m - V_{21} V_{11}{}^{-1} W_{11} V_{11}{}^{-1} V_{12} = V_{21}^m (V_{11}^m)^{-1} (W_{11}^m (V_{11}^m)^{-1} V_{12}^m - W_{11} V_{11}{}^{-1} V_{12}) + (V_{21}^m (V_{11}^m)^{-1} - V_{21} V_{11}{}^{-1}) W_{11} V_{11}{}^{-1} V_{12}$ which converges to zero because

  $$W_{11}^m (V_{11}^m)^{-1} V_{12}^m - W_{11} V_{11}{}^{-1} V_{12} = W_{11}^m ((V_{11}^m)^{-1} V_{12}^m - V_{11}{}^{-1} V_{12}) + (W_{11}^m - W_{11}) V_{11}{}^{-1} V_{12}$$

by Lemmas 4.1.1 and 4.1.3.

For the second summand, we have that

$$\frac{[S_n^{(2,m)}(\tilde{\boldsymbol{\theta}}_n)]^2 - [S_n^{(2)}(\tilde{\boldsymbol{\theta}}_n)]^2}{\tilde{\sigma}^2} = \frac{[S_n^{(2,m)}(\tilde{\boldsymbol{\theta}}_n) - S_n^{(2)}(\tilde{\boldsymbol{\theta}}_n)][[S_n^{(2,m)}(\tilde{\boldsymbol{\theta}}_n) + S_n^{(2)}(\tilde{\boldsymbol{\theta}}_n)]]}{\tilde{\sigma}^2}$$

which converges to zero as a result of Lemma 4.1.2 because $\tilde{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}$.

□

Chapter 5

Orthogonal Polynomial Approximation

In this chapter, our interest centers on the autocovariance and autocorrelation functions of the first order log-linear Poisson model with feedback (4.1) which was studied in the previous chapter.

In the case of the first order linear Poisson model (2.1), the autocovariance function of the observed process $\{Y_t\}$ can be derived explicitly because the model can be represented as an ARMA$(1, 1)$ process; see Section 2.1, Ferland et al. (2006) and Fokianos (2012) for more. However, a similar expression for the autocovariance function of $\{Y_t\}$ in the case of the log-linear Poisson model (4.1) cannot be developed, because of the non-linearity imposed by the log-linear structure.

Our approach to develop an expression for the autocovariance and autocorrelation functions of $\{Y_t\}$ is by using the autocovariance function of a transformation of the process, which we can construct more easily. In particular, we attempt to construct the autocovariance function of $\{Y_t\}$ using the autocovariance function of the one-to-one transformation $\{\log(1+Y_t)\}$ and a connection between the two through orthogonal polynomial expansions.

There exists a close relationship that connects stochastic processes with orthogonal polynomials; see Schoutens (2000) among others. Examples of the connection between stochastic processes and orthogonal polynomials have been examined since the 1950s. For instance, stochastic integration theory with respect to the Brownian motion, the binomial and Poisson processes is connected respectively to the Hermite, Krawtchouk and Charlier polynomials, whilst the so called Karlin-McGregor representation is applied to express the transition probabilities of birth and death processes in terms of orthogonal polynomials. Since then, more relations among stochastic processes and orthogonal polynomials have been instituted. Schoutens (2000) studies this connection and additionally links the wider class of Sheffer polynomials with Lévy processes using a martingale relation. The broad class of Sheffer polynomials includes well known polynomials as are the Hermite, Laguerre and Charlier polynomials and Lévy processes include among others the Gaussian and Poisson processes.

In the following section we explore some of the most important classes of orthogonal polynomials and review their properties. These will be very useful tools in the derivation of the autocovariance function of $\{Y_t\}$ as discussed above.

5.1 Orthogonal Polynomials

Orthogonal polynomials consist of a most classical topic that emerges in many fields of mathematical studies. Many surveys on orthogonal polynomials can be found in the literature, we emphasize however the significant manuscripts of Lebedev (1972), Szegő (1975) and Chihara (1978). The aforementioned books provide the reader an introduction to orthogonal polynomials but at the same time provide additionally an extensive study of the properties and attributes of several particular classes of orthogonal polynomials that are related to specific weight functions.

5.1.1 Hermite Polynomials

The Hermite polynomials $H_n(x)$ constitute one of the most important classes of orthogonal polynomials; see Lebedev (1972) and Szegő (1975). The Hermite polynomials are associated with the standard normal distribution and are defined by the equation

$$H_n(x) = (-1)^n e^{x^2/2} \frac{\partial^n}{\partial x^n} e^{-x^2/2}, \quad n = 0, 1, 2, ..., \quad x \in (-\infty, +\infty).$$

The first few polynomials are

$$H_0(x) = 1, \quad H_1(x) = x, \quad H_2(x) = x^2 - 1, \quad H_3(x) = x^3 - 3x, \quad H_4(x) = x^4 - 6x^2 + 3.$$

An explicit formula to derive the Hermite polynomials is given by

$$H_n(x) = \sum_{k=0}^{[n/2]} \frac{(-1)^k n!}{k!(n-2k)!} \frac{x^{n-2k}}{2^k}$$

where $[\nu]$ denotes the largest integer less than or equal to ν. The generating function of the Hermite polynomials is

$$w(x,t) = e^{xt - t^2/2} = \sum_{n=0}^{\infty} H_n(x) \frac{t^n}{n!}. \tag{5.1}$$

To support the above representation, consider the Taylor expansion of $w(x,t)$ as a function of t:

$$w(x,t) = e^{xt-t^2/2} = \sum_{n=0}^{\infty} \left[\frac{\partial^n w(x,t)}{\partial t^n} \right] \frac{t^n}{n!}, \quad |t| < \infty.$$

This implies that

$$\frac{\partial^n w(x,t)}{\partial t^n} = e^{x^2/2} \left[\frac{\partial^n}{\partial t^n} e^{-(x-t)^2/2} \right]_{t=0} = (-1)^n e^{x^2/2} \frac{\partial^n}{\partial x^n} e^{-x^2/2} = H_n(x)$$

because $e^{xt-t^2} = e^{x^2/2-(x-t)^2/2}$ by completing the square and using the substitution $u = x - t$ we obtain

$$\left[\frac{\partial^n}{\partial t^n} e^{-(x-t)^2/2} \right]_{t=0} = \left[\frac{\partial^n}{\partial u^n} e^{-u^2/2} \left(\frac{\partial u}{\partial t} \right)^n \right]_{u=x} = (-1)^n \left[\frac{\partial^n}{\partial u^n} e^{-u^2/2} \right]_{u=x}.$$

In order to built a three term recurrence relation for the Hermite polynomials, we substitute the generating function (5.1) into the identity

$$\frac{\partial w}{\partial t} - (x-t)w = 0.$$

Then,

$$\sum_{n=1}^{\infty} H_n(x)\frac{nt^{n-1}}{n!} - x\sum_{n=0}^{\infty} H_n(x)\frac{t^n}{n!} + t\sum_{n=0}^{\infty} H_n(x)\frac{t^n}{n!} = 0$$
$$\Rightarrow \sum_{n=1}^{\infty} H_n(x)\frac{t^{n-1}}{(n-1)!} - x\sum_{n=0}^{\infty} H_n(x)\frac{t^n}{n!} + t\sum_{n=0}^{\infty} H_n(x)\frac{t^{n+1}}{n!} = 0$$
$$\Rightarrow \sum_{n=1}^{\infty} H_n(x)\frac{t^{n-1}}{(n-1)!} - x\sum_{n=0}^{\infty} H_n(x)\frac{t^n}{n!} + \sum_{n=-1}^{\infty} H_n(x)(n+1)\frac{t^{n+1}}{(n+1)!} = 0$$
$$\Rightarrow \sum_{n=0}^{\infty} H_{n+1}(x)\frac{t^n}{n!} - x\sum_{n=0}^{\infty} H_n(x)\frac{t^n}{n!} + \sum_{n=0}^{\infty} nH_{n-1}(x)\frac{t^n}{n!} = 0$$

and when the coefficient of t^n is set equal to zero the three term recurrence relation

$$H_{n+1}(x) - xH_n(x) + nH_{n-1}(x) = 0, \quad n = 1, 2, \ldots \tag{5.2}$$

is achieved. The above formula (5.2) can be used to derive the Hermite polynomials starting with $H_0(x)$ and $H_1(x)$.

The Hermite polynomials $H_n(x)$ are orthogonal with respect to the standard normal distribution. In particular, they are orthogonal on the interval $(-\infty, +\infty)$ with respect to the weight function $e^{-x^2/2}$, that is

$$\int_{-\infty}^{\infty} e^{-x^2/2} H_m(x)H_n(x)dx = 0 \quad \text{if} \quad m \neq n.$$

To retrieve an analogous result for $m = n$, we use the recurrence relation (5.2). Replacing the index n by $n-1$ and multiplying the result by $H_n(x)$ gives

$$H_n^2(x) - xH_n(x)H_{n-1}(x) + (n-1)H_n(x)H_{n-2}(x) = 0.$$

Also, multiplying the recurrence relation (5.2) by $H_{n-1}(x)$ gives

$$H_{n+1}(x)H_{n-1}(x) - xH_n(x)H_{n-1}(x) + nH_{n-1}^2(x) = 0.$$

Subtracting the last two equations we obtain that

$$H_n^2(x) + (n-1)H_n(x)H_{n-2}(x) - H_{n+1}(x)H_{n-1}(x) - nH_{n-1}^2(x) = 0, \quad n = 2, 3, \ldots.$$

Finally, multiplying by $e^{-x^2/2}$ and integrating over $(-\infty, +\infty)$ the above equation, application of the property when $m \neq n$ gives

$$\int_{-\infty}^{\infty} e^{-x^2/2} H_n^2(x)dx = n \int_{-\infty}^{\infty} e^{-x^2/2} H_{n-1}^2(x)dx.$$

By repeated application we obtain

$$\begin{aligned}
\int_{-\infty}^{\infty} e^{-x^2/2} H_n^2(x)dx &= n \int_{-\infty}^{\infty} e^{-x^2/2} H_{n-1}^2(x)dx = n(n-1) \int_{-\infty}^{\infty} e^{-x^2/2} H_{n-2}^2(x)dx \\
&= \ldots \\
&= n! \int_{-\infty}^{\infty} e^{-x^2/2} H_1^2(x)dx = n! \int_{-\infty}^{\infty} e^{-x^2/2} dx \\
&= n!\sqrt{2\pi}, \quad n = 0, 1, 2, \ldots
\end{aligned}$$

Combining the two resulting equations for $m \neq n$ and $m = n$ gives the orthogonality relation of the Hermite polynomials

$$\int_{-\infty}^{\infty} e^{-x^2/2} H_m(x) H_n(x)dx = n!\sqrt{2\pi}\delta_{mn} \tag{5.3}$$

where δ_{mn} denotes the Kronecker delta defined by

$$\delta_{mn} = \begin{cases} 0, & m \neq n \\ 1, & m = n. \end{cases}$$

The following result is obtained from Lebedev (1972) and shows that a real function $f(x)$ satisfying certain conditions may be expanded in a series of Hermite polynomials.

Proposition 5.1.1 If the real function $f(x)$ defined on the infinite interval $(-\infty, +\infty)$ is piecewise smooth in every finite interval $[-a, a]$, and if the integral

$$\int_{-\infty}^{\infty} e^{-x^2/2} f^2(x)dx$$

is finite, then the series

$$f(x) = \sum_{n=0}^{\infty} c_n H_n(x), \quad -\infty < x < \infty, \tag{5.4}$$

with coefficients c_n calculated from

$$c_n = \frac{1}{n!\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} f(x) H_n(x) dx, \tag{5.5}$$

converges to $f(x)$ at every continuity point of $f(x)$.

We note that the conditions that are imposed on the function $f(x)$ allow the calculation of the coefficients c_n since they imply the existence of the integral on the right hand side of (5.5).

The coefficients c_n can be determined by multiplying the series (5.4) by $e^{-x^2/2}H_m(x)$, integrating over $(-\infty, +\infty)$ and applying the orthogonality property of the Hermite polynomials (5.3). In particular, we find that

$$\int_{-\infty}^{\infty} e^{-x^2/2} H_m(x) f(x) dx = \sum_{n=0}^{\infty} c_n \int_{-\infty}^{\infty} e^{-x^2/2} H_m(x) H_n(x) dx = c_m m! \sqrt{2\pi}$$

which implies (5.5).

The following is an example of expansion of a function in series of Hermite polynomials.

Example 5.1.1 (Lebedev (1972)) Let $f(x) = e^{ax}$ where a is an arbitrary real number. Setting $t = a$ in the generating function (5.1) we obtain that

$$e^{ax - a^2/2} = \sum_{n=0}^{\infty} H_n(x) \frac{a^n}{n!} \Rightarrow e^{ax} = e^{a^2/2} \sum_{n=0}^{\infty} H_n(x) \frac{a^n}{n!}.$$

Alternatively, we calculate the coefficients c_n in the series expansion of the function $f(x) = e^{ax}$ according to (5.5) as follows:

$$\begin{aligned} c_n &= \frac{1}{n!\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} f(x) H_n(x) dx = \frac{1}{n!\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ax - x^2/2} H_n(x) dx \\ &= \frac{1}{n!\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ax - x^2/2} (-1)^n e^{x^2/2} \frac{\partial^n}{\partial x^n} e^{-x^2/2} dx = \frac{1}{n!\sqrt{2\pi}} \int_{-\infty}^{\infty} (-1)^n e^{ax} \frac{\partial^n}{\partial x^n} e^{-x^2/2} dx \\ &= \ldots \quad \text{(performing integration by parts } n \text{ times)} \\ &= \frac{1}{n!\sqrt{2\pi}} a^n e^{a^2/2} \int_{-\infty}^{\infty} e^{-x^2/2} dx = \frac{1}{n!\sqrt{2\pi}} a^n e^{a^2/2} \sqrt{2\pi} \\ &= a^n e^{a^2/2}. \end{aligned}$$

5.1.2 Laguerre Polynomials

Another important class of orthogonal polynomials is the class of the Laguerre polynomials $L_n^a(x)$ defined by

$$L_n^a(x) = e^x \frac{x^{-a}}{n!} \frac{d^n}{dx^n}(e^{-x}x^{n+a}), \quad n = 0, 1, 2, \ldots$$

for arbitrary real $a > -1$. The first few Laguerre polynomials are

$$\begin{aligned} L_0^a(x) &= 1,\ L_1^a(x) = 1 + a - x,\ L_2^a(x) = \frac{x^2}{2} - (a+2)x + \frac{(a+2)(a+1)}{2}, \\ L_3^a(x) &= -\frac{x^3}{6} + \frac{(a+3)x^2}{2} - \frac{(a+3)(a+2)x}{2} + \frac{(a+3)(a+2)(a+1)}{6}, \\ L_4^a(x) &= \frac{x^4}{24} - \frac{(a+4)x^3}{6} + \frac{(a+4)(a+3)x^2}{4} - \frac{(a+4)(a+3)(a+2)^x}{6} \\ &\quad + \frac{(a+4)(a+3)(a+2)(a+1)}{24}. \end{aligned}$$

An explicit formula gives

$$L_n^a(x) = \sum_{k=0}^{n} \frac{\Gamma(n+a+1)}{\Gamma(k+a+1)} \frac{(-x)^k}{k!(n-k)!}.$$

The generating function of the Laguerre polynomials is

$$w(x,t) = (1-t)^{-a-1} e^{-xt/(1-t)} = \sum_{n=0}^{\infty} L_n^a(x) t^n, \quad |t| < 1. \tag{5.6}$$

To obtain a three term recurrence relation, we substitute the Laguerre generating function (5.6) into the identity

$$(1-t)^2 \frac{\partial w}{\partial t} + [x - (1-t)(1+a)]w = 0.$$

Then,

$$
\begin{aligned}
&(1-t)^2\sum_{n=0}^{\infty} nL_n^a(x)t^{n-1} + [x-(1-t)(1+a)]\sum_{n=0}^{\infty} L_n^a(x)t^n = 0 \\
\Rightarrow\ &(1-2t+t^2)\sum_{n=0}^{\infty} nL_n^a(x)t^{n-1} + [x-(1_a)+(1+a)t]\sum_{n=0}^{\infty} L_n^a(x)t^n = 0 \\
\Rightarrow\ &\sum_{n=0}^{\infty} nL_n^a(x)t^{n-1} - 2\sum_{n=0}^{\infty} nL_n^a(x)t^n + \sum_{n=0}^{\infty} nL_n^a(x)t^{n+1} \\
&+ [x-(1+a)]\sum_{n=0}^{\infty} L_n^a(x)t^n + (1+a)\sum_{n=0}^{\infty} nL_n^a(x)t^{n+1} = 0 \\
\Rightarrow\ &\sum_{n=0}^{\infty}(n+1)L_{n+1}^a(x)t^n - 2\sum_{n=0}^{\infty} nL_n^a(x)t^n + \sum_{n=0}^{\infty}(n-1)L_{n-1}^a(x)t^n \\
&+ [x-(1+a)]\sum_{n=0}^{\infty} L_n^a(x)t^n + (1+a)\sum_{n=0}^{\infty} L_{n-1}^a(x)t^n = 0 \\
\Rightarrow\ &(n+1)L_{n+1}^a(x) + [x-a-2n-1]L_n^a(x) + (n+a)L_{n-1}^a(x) = 0 \qquad (5.7)
\end{aligned}
$$

when the coefficient of t^n is equated to zero. The Laguerre polynomials are orthogonal on the interval $[0,\infty)$ with weight $e^{-x}x^a$, that is

$$\int_0^{\infty} e^{-x}x^a L_m^a(x)L_n^a(x)dx = 0, \quad \text{if} \quad m \neq n,\ a > -1.$$

For $m = n$, working as before, we replace the index n by $n-1$ in the recurrence formula (5.7) multiplied by $L_n^a(x)$ to obtain

$$n[L_n^a(x)]^2 + [x-a-2(n-1)-1]L_{n-1}^a(x)L_n^a(x) + (n-1+a)L_{n-2}^a(x)L_n^a(x) = 0.$$

Multiplying the recurrence relation (5.7) by $L_{n-1}^a(x)$ gives

$$(n+1)L_{n+1}^a(x)L_{n-1}^a(x) + [x-a-2n-1]L_n^a(x)L_{n-1}^a(x) + (n+a)[L_{n-1}^a(x)]^2 = 0.$$

Then, subtracting the two equations we get the equation

$$n[L_n^a(x)]^2-(n+a)[L_{n-1}^a(x)]^2-(n+1)L_{n+1}^a(x)L_{n-1}^a(x)+2L_n^a(x)L_{n-1}^a(x)+(n+a-1)L_{n-2}^a(x)L_n^a(x) = 0.$$

Finally, multiplying the last equation by $e^{-x}x^a$, integrating over the interval $[0,\infty)$ we find that

$$n\int_0^\infty e^{-x}x^a[L_n^a(x)]^2dx=(n+1)\int_0^\infty e^{-x}x^a[L_{n-1}^a(x)]^2dx,\quad n=2,3,\dots$$

Repeated application gives

$$\begin{aligned}
&\int_0^\infty e^{-x}x^a[L_n^a(x)]^2dx=\frac{n+1}{n}\int_0^\infty e^{-x}x^a[L_{n-1}^a(x)]^2dx\\
&=\frac{(n+a)(n+a-1)}{n(n-1)}\int_0^\infty e^{-x}x^a[L_{n-2}^a(x)]^2dx\\
&=\dots\\
&=\frac{(n+a)(n+a-1)\dots(a+2)}{n!}\int_0^\infty e^{-x}x^a[1+a-x]^2dx\\
&=\frac{(n+a)(n+a-1)\dots(a+2)}{n!}\{(1+a)^2\int_0^\infty e^{-x}x^adx-2(1+a)\int_0^\infty e^{-x}x^{a+1}dx\\
&+\int_0^\infty e^{-x}x^{a+2}dx\}\\
&=\frac{(n+a)(n+a-1)\dots(a+2)}{n!}\{(1+a)^2\Gamma(a+1)-2(1+a)\Gamma(a+2)+\Gamma(a+3)\}\\
&=\frac{(n+a)(n+a-1)\dots(a+2)}{n!}(a+1)\Gamma(a+1)\\
&=\frac{(n+a)(n+a-1)\dots(a+2)(a+1)a!}{n!}=\frac{(n+a)!}{n!}\\
&=\frac{\Gamma(n+a+1)}{n!},\quad n=2,3,\dots
\end{aligned}$$

Concluding, the orthogonality relation of the Laguerre polynomials is given by

$$\int_0^\infty e^{-x}x^aL_m^a(x)L_n^a(x)dx=\frac{\Gamma(n+a+1)}{n!}\delta_{mn}. \tag{5.8}$$

A real function $f(x)$ can be expanded in a series of Laguerre polynomials using the following result; see Lebedev (1972).

Proposition 5.1.2 If the real function $f(x)$ which is defined on the interval $(0,\infty)$, is piecewise smooth in every finite subinterval $[x_1,x_2]$ where $0<x_1<x_2<\infty$, $a>-1$, and if the integral $\int_0^\infty e^{-x}x^af^2(x)dx$ is finite, then the series

$$f(x)=\sum_{n=0}^\infty c_nL_n^a(x),\quad 0<x<\infty, \tag{5.9}$$

with coefficients c_n calculated from

$$c_n = \frac{n!}{\Gamma(n+a+1)} \int_0^\infty e^{-x} x^a f(x) L_n^a(x) dx, \tag{5.10}$$

converges to $f(x)$ at every continuity point of $f(x)$.

The coefficients c_n are determined using the orthogonality property (5.8) of the Laguerre polynomials. To be more precise, multiplying the series (5.9) by $e^{-x}x^a L_m^a(x)$ and integrating over $(0, \infty)$ we obtain

$$\begin{aligned} \int_0^\infty e^{-x} x^a f(x) L_m^a(x) dx &= \sum_{n=0}^\infty \int_0^\infty e^{-x} x^a L_m^a(x) c_n L_n^a(x) dx \\ &= c_n \int_0^\infty e^{-x} x^a [L_n^a(x)]^2 dx = c_n \frac{\Gamma(n+a+1)}{n!} \end{aligned}$$

which gives the c_n as in (5.10).

5.1.3 Poisson-Charlier Polynomials

In contrast with the two classes of orthogonal polynomials described in the previous sections, the Poisson-Charlier polynomials constitute a discrete orthogonal polynomial sequence that is orthogonal with respect to the Poisson distribution with rate μ; Chihara (1978).

The monic Charlier polynomials $C_n(\mu; x)$ are defined by the generating function

$$e^{-\mu w}(1+w)^x = \sum_{n=0}^\infty C_n(\mu; x) \frac{w^n}{n!}, \quad \mu \neq 0.$$

An explicit representation is given by

$$C_n(\mu; x) = \sum_{k=0}^n \binom{n}{k} \binom{x}{k} k! (-\mu)^{n-k} \tag{5.11}$$

and the orthogonality relation is given by

$$\sum_{x=0}^\infty C_m(\mu; x) C_n(\mu; x) \frac{e^{-\mu} \mu^x}{x!} = \mu^n n! \delta_{mn}. \tag{5.12}$$

To show the orthogonality relation (5.12) consider the following series expansion of the generating function

$$G(x,w) = e^{-\mu w}(1+w)^x = \sum_{m=0}^{\infty} \frac{(-\mu)^m w^m}{m!} \sum_{n=0}^{\infty} \binom{x}{n} w^n.$$

Forming the Cauchy product of the two series in the above equation we obtain

$$G(x,w) = \sum_{n=0}^{\infty} d_n, \quad \text{where} \quad d_n = \sum_{k=0}^{n} \binom{x}{k} w^k \frac{(\mu)^{n-k} w^{n-k}}{(n-k)!} = \sum_{k=0}^{n} \binom{x}{k} (-\mu)^{n-k} w^n.$$

This gives

$$G(x,w) = \sum_{n=0}^{\infty} P_n(x) w^n \quad \text{where} \quad P_n(x) = \sum_{k=0}^{n} \binom{x}{k} \frac{(-\mu)^{n-k}}{(n-k)!}$$

and $P_n(x)$ is a polynomial of degree n. To show this, consider

$$\mu^x G(x,\nu) G(x,w) = \mu^x e^{-\mu\nu}(1+\nu)^x e^{-\mu w}(1+w)^x = e^{\mu(\nu+w)}[\mu(1+\nu)(1+w)]^x.$$

Then,

$$\begin{aligned}
\sum_{k=0}^{\infty} \frac{\mu^k G(x,\nu) G(x,w)}{k!} &= \sum_{k=0}^{\infty} \frac{e^{-\mu(\nu+w)}[\mu(1+\nu)(1+w)]^k}{k!} \\
&= e^{-\mu(\nu+w)} e^{\mu(1+\nu)(1+w)} = e^{\mu} e^{\mu\nu w} \\
&= \sum_{n=0}^{\infty} e^{\mu} \mu^n \frac{(\nu w)^n}{n!}.
\end{aligned}$$

Also,

$$\sum_{k=0}^{\infty} \frac{\mu^k G(x,\nu) G(x,w)}{k!} = \sum_{k=0}^{\infty} \frac{\mu^k}{k!} \sum_{m,n=0}^{\infty} P_m(k) P_n(k) \nu^m w^n = \sum_{m,n=0}^{\infty} \sum_{k=0}^{\infty} P_m(k) P_n(k) \frac{\mu^k}{k!} \nu^m w^n.$$

By comparing the coefficients of $\nu^m w^n$ in the two resulting series above, we conclude that

$$\sum_{k=0}^{\infty} P_m(k) P_n(k) \frac{\mu^k}{k!} = \begin{cases} 0, & m \neq n \\ \dfrac{e^{\mu}\mu^n}{n!}, & m = n. \end{cases}$$

Finally, $(n!)^{-1} C_n(\mu; x)$ is denoted by $P_n(x)$.

The three term recurrence formula corresponding to the Poisson Charlier polynomials is given by

$$C_{n+1}(\mu;x) = (x-n-\mu)C_n(\mu;x) - \mu n C_{n-1}(\mu;x). \tag{5.13}$$

We note that the Charlier polynomials can also be expressed in terms of the Laguerre polynomials by

$$C_n(\mu;x) = n! L_n^{x-n}(\mu).$$

The first few Charlier polynomials are given by

$$\begin{aligned}
&C_0(\mu;x) = 1, \quad C_1(\mu;x) = x-\mu, \quad C_2(\mu;x) = \mu^2 - 2\mu x + x(x-1),\\
&C_3(\mu;x) = -\mu^3 + 3\mu^2 x - 3\mu x(x-1) + x(x-1)(x-2)\\
&C_4(\mu;x) = \mu^4 - 4\mu^3 x + 6\mu^2 x(x-1) - 4\mu x(x-1)(x-2) + x(x-1)(x-2)(x-3).
\end{aligned}$$

A non-monic form of the Charlier polynomials is given by

$$c_n(\mu;x) = (-1)^n \mu^{-n} C_n(\mu;x) \tag{5.14}$$

and the first few polynomials are

$$\begin{aligned}
&c_0(\mu;x) = 1, \quad c_1(\mu;x) = 1 - \frac{x}{\mu}, \quad c_2(\mu;x) = 1 - \frac{2x}{\mu} + \frac{x(x-1)}{\mu^2}\\
&c_3(\mu;x) = 1 - \frac{3x}{\mu} + \frac{3x(3-1)}{\mu^2} - \frac{x(x-1)(x-2)}{\mu^3}\\
&c_4(\mu;x) = 1 - \frac{4x}{\mu} + \frac{6x(x-1)}{\mu^2} - \frac{4x(x-1)(x-2)}{\mu^3} + \frac{x(x-1)(x-2)(x-3)}{\mu^4}.
\end{aligned}$$

This notation allows the symmetry relation $c_n(\mu;x) = c_x(\mu;n)$ from which follows a "dual orthogonality"

$$\sum_{x=0}^{\infty} c_n(\mu;x)c_m(\mu;x)\frac{e^{-\mu}\mu^x}{x!} = \sum_{x=0}^{\infty} c_x(\mu;n)c_x(\mu;m)\frac{e^{-\mu}\mu^x}{x!} = \mu^{-n} n! \delta_{mn}.$$

An explicit formula for the $c_n(\mu;x)$ is

$$c_n(\mu;x) = \frac{x!}{\mu^x}\nabla^n\left(\frac{\mu^x}{x!}\right) = \frac{x!}{\mu^x}\sum_{k=0}^{n}(-1)^k\binom{n}{k}\frac{\mu^{x-k}}{(x-k)!} \tag{5.15}$$

where ∇ is the backward difference operation defined as

$$\nabla f(x) = f(x) - f(x-1).$$

It holds that

$$\nabla^n f(x) = \nabla(\nabla^{n-1} f(x)) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} f(x-k).$$

Indeed, using equations (5.11) and (5.14) we have that

$$\begin{aligned} c_n(\mu;x) &= (-1)^n \mu^{-n} C_n(\mu;x) = (-1)^n \mu^{-n} \sum_{k=0}^{n} \binom{n}{k}\binom{x}{k} k!(-\mu)^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (-1)^{2n-k} \frac{x!k!}{k!(x-k)!} \mu^{-n+n-k} = \frac{x!}{\mu^x} \sum_{k=0}^{n} (-1)^{-k} \binom{n}{k} \frac{\mu^{x-k}}{(x-k)!} \\ &= \frac{x!}{\mu^x} \sum_{k=0}^{n} (-1)^k \binom{n}{k} \frac{\mu^{x-k}}{(x-k)!} \end{aligned}$$

which verifies the explicit formula (5.15). Normalizing the polynomials we obtain the orthonormal Poisson - Charlier polynomials

$$\mathcal{C}_n(\mu;x) = \sqrt{\frac{\mu^n}{n!}}\, c_n(\mu;x). \tag{5.16}$$

The first few polynomials are

$$\begin{aligned} &\mathcal{C}_0(\mu;x) = 1, \quad \mathcal{C}_1(\mu;x) = \mu\left(1 - \frac{x}{\mu}\right), \quad \mathcal{C}_2(\mu;x) = \frac{\sqrt{2}}{2}\left\{\mu - 2x + \frac{x(x-1)}{\mu}\right\}, \\ &\mathcal{C}_3(\mu;x) = \sqrt{\frac{\mu}{2}}\left\{\frac{\sqrt{3}\mu}{3} - \sqrt{3}x + \frac{\sqrt{3}}{\mu}x(x-1) - \frac{\sqrt{3}}{3\mu^2}x(x-1)(x-2)\right\} \\ &\mathcal{C}_4(\mu;x) = \frac{1}{\sqrt{6}}\left\{\frac{\mu^2}{2} - 2\mu x + 3x(x-1) - \frac{2x(x-1)(x-2)}{\mu} + \frac{x(x-1)(x-2)(x-3)}{2\mu^2}\right\}. \end{aligned}$$

The following proposition exhibits the series expansion to Charlier polynomials of a function, when it possess certain properties.

Proposition 5.1.3 A function $f(x)$ defined on the integers such that

$$\sum_{x=0}^{\infty} f^2(x)\varrho(\mu;x) < \infty$$

for some $\mu > 0$, where

$$\varrho(\mu; x) = e^{-\mu}\frac{\mu^x}{x!}, \quad x = 0, 1, \ldots$$

is the pdf of the Poisson distribution, can be expanded in a series of Poisson-Charlier orthogonal polynomials

$$f(x) = \sum_{n=0}^{\infty} b_n \mathcal{C}_n(\mu; x) \tag{5.17}$$

where

$$b_n = \mathrm{E}[f(x)\mathcal{C}_n(\mu; x)] = \sum_{k=0}^{\infty} f(k)\mathcal{C}_n(\mu; k)\varrho(\mu; k). \tag{5.18}$$

Indeed, multiplying (5.17) by $\mathcal{C}_m(\mu; x)\varrho(\mu; x)$ and taking the sum with respect to x from 0 to infinity we find that

$$\begin{aligned}
\sum_{x=0}^{\infty} f(x)\mathcal{C}_m(\mu; x)\varrho(\mu; x) &= \sum_{x=0}^{\infty}\sum_{n=0}^{\infty} b_n \mathcal{C}_n(\mu; x)\mathcal{C}_m(\mu; x)\varrho(\mu; x) \\
&= \sum_{n=0}^{\infty} b_n \sum_{x=0}^{\infty} \mathcal{C}_n(\mu; x)\mathcal{C}_m(\mu; x)\varrho(\mu; x) \\
&= \sum_{n=0}^{\infty} b_n \sum_{x=0}^{\infty} \sqrt{\frac{\mu^n}{n!}} c_n(\mu; x)\sqrt{\frac{\mu^m}{m!}} c_m(\mu; x)\varrho(\mu; x) \\
&= \sum_{n=0}^{\infty} b_n \sum_{x=0}^{\infty} \sqrt{\frac{\mu^n}{n!}}(-1)^n\mu^{-n}C_n(\mu; x)\sqrt{\frac{\mu^m}{m!}}(-1)^m\mu^{-m}C_m(\mu; x)\varrho(\mu; x) \\
&= b_m \frac{\mu^m}{m!}\mu^{-2m}\mu^m m!\delta_{mn} \\
&= b_m
\end{aligned}$$

using the orthogonality relation (5.12).

Example 5.1.2 Consider the function $f(x) = e^x$. Then, $f(x)$ can be expanded in an orthogonal series as (5.17) with coefficients b_n given as follows:

$$n = 0: \quad b_0 = \sum_{x=0}^{\infty} f(x)\mathcal{C}_0(\mu; x)\varrho(\mu; x) = \sum_{x=0}^{\infty} e^x \frac{e^{-\mu}\mu^x}{x!} = e^{\mu(e-1)}$$

$$n = 1: \quad b_1 = \sum_{x=0}^{\infty} f(x)\mathcal{C}_1(\mu; x)\varrho(\mu; x) = \sum_{x=0}^{\infty} e^x \sqrt{\mu}(1 - \frac{x}{\mu})\frac{e^{-\mu}\mu^x}{x!} = \sqrt{\mu}b_0(1 - e)$$

$$n = 2: \quad b_2 = \sum_{x=0}^{\infty} f(x)\mathcal{C}_2(\mu;x)\varrho(\mu;x) = \sum_{x=0}^{\infty} e^x \frac{1}{\sqrt{2}} \left\{ \mu - 2x + \frac{x(x-1)}{\mu} \right\} \frac{e^{-\mu}\mu^x}{x!}$$
$$= \frac{\mu(1-e)^2}{\sqrt{2}} b_0$$
$$n = 3: \quad b_3 = \sum_{x=0}^{\infty} f(x)\mathcal{C}_3(\mu;x)\varrho(\mu;x)$$
$$= \sum_{x=0}^{\infty} e^x \sqrt{\frac{\mu}{2}} \left\{ \frac{\sqrt{3}\mu}{3} - \sqrt{3}x + \frac{\sqrt{3}}{\mu}x(x-1) - \frac{\sqrt{3}}{3\mu^2}x(x-1)(x-2) \right\} \frac{e^{-\mu}\mu^x}{x!}$$
$$= \frac{\mu\sqrt{\mu}}{\sqrt{3 \cdot 2}}(1-e)^3 b_0$$
$$n = 4: \quad b_4 = \sum_{x=0}^{\infty} f(x)\mathcal{C}_4(\mu;x)\varrho(\mu;x) = \frac{\mu^2}{\sqrt{24}}(1-e)^4 b_0$$

Continuing as above we conclude that

$$b_n = \sqrt{\frac{\mu^n}{n!}}(1-e)^n e^{\mu(e-1)}.$$

5.2 Representation by ARMA

Recall the first order log-linear Poisson model with feedback given by (4.1). The following result is taken from Fokianos and Tjøstheim (2011):

Lemma A-1 $\mathrm{E}(\log(1+Y_t|\nu_t = \nu) - \nu \to 0, \quad \text{as} \quad \nu \to \infty.$

Then, we have the following approximate representation

$$\begin{aligned} \log(1+Y_t) &= \log(1+Y_t) - \nu_t + \nu_t \\ &= \nu_t + \epsilon_t \quad \text{where} \;\; \epsilon_t = \log(1+Y_t) - \nu_t \end{aligned}$$

Hence

$$\begin{aligned} \log(1+Y_t) &= d + a\nu_{t-1} + b\log(1+Y_{t-1}) + \epsilon_t \\ &= d + a(\log(1+Y_{t-1}) - \epsilon_{t-1}) + b\log(1+Y_{t-1}) + \epsilon_t \\ &= d + (a+b)\log(1+Y_{t-1}) + \epsilon_t - a\epsilon_{t-1}. \end{aligned}$$

Therefore,

$$\left\{\log(1+Y_t) - \frac{d}{1-(a+b)}\right\} - (a+b)\left\{\log(1+Y_{t-1}) - \frac{d}{1-(a+b)}\right\} = \epsilon_t - a\epsilon_{t-1}.$$

The above representation shows that the process $\{\log(1+Y_t)\}$ can be represented as an ARMA$(1,1)$ process with mean given by $\frac{d}{1-(a+b)}$. Additionally, the stationarity condition $|a+b|<1$ guarantees the causality of the process. Since the process $\{\log(1+Y_t)\}$ can be represented as an ARMA$(1,1)$ process, then its autocovariance function is given by

$$\gamma(h) = \text{Cov}(\log(1+Y_t), \log(1+Y_{t+h})) = \begin{cases} \frac{(1-a^2-2ab)\sigma_\epsilon^2}{1-(a+b)^2}, & h=0 \\ \frac{b(1-a(a+b))\sigma_\epsilon^2}{1-(a+b)^2}(a+b)^{h-1}, & h\geq 1 \end{cases} \tag{5.19}$$

Set

$$X_t \equiv \log(1+Y_t)$$

and the mean of the process as $\mu_x = \frac{d}{1-(a+b)}$. Then,

$$Y_t = e^{X_t} - 1.$$

Our interest is to derive the autocovariance function of $\{Y_t\}$ using the autocovariance function of $\{X_t\}$.

$$\begin{aligned} \text{Cov}(Y_t, Y_{t+h}) &= \text{Cov}(e^{X_t}-1, e^{X_{t+h}}-1) = \text{Cov}(e^{X_t}, e^{X_{t+h}}) \\ &= \text{Cov}(f(X_t), f(X_{t+h})), \quad \text{where } \; f(x) = e^x \end{aligned} \tag{5.20}$$

Considering the strong connection between orthogonal polynomials and probability theory, we will derive the autocovariance function of the process $\{Y_t\}$ using orthogonal series expansion of the function $f(x)$ as given above. Similar approaches have been attempted mostly in econometrics, for example see Abadir and Talmain (2005) who derive a method to obtain the autocovariance function of transformations of a time series, however not through orthogonal polynomials.

5.2.1 Hermite Expansion

Granger and Newbold (1976) examine the autocovariance of transformed series of Gaussian processes based on Hermite polynomial expansion. Accordingly, assume that the process X_t is stationary Gaussian with mean μ_x, variance σ_x^2 and autocorrelation function $\rho(h) = \mathrm{Corr}(X_t, X_{t+h})$. Additionally, we set

$$Z_t = \frac{X_t - \mu_x}{\sigma_x}$$

and the transformation $f(Z_t)$ can be expanded in a series of Hermite polynomials as

$$f(Z_t) = \sum_{n=0}^{M} c_n H_n(Z_t).$$

Then, the autocovariance function of $f(X_t)$ is given by

$$\mathrm{Cov}(f(X_t), f(X_{t+h})) = \sum_{n=1}^{\infty} c_n^2 n! \mathrm{Corr}^n(X_t, X_{t+h}).$$

In our case, we have the transformation

$$f(X_t) = e^{X_t} = e^{\mu_x + \sigma_x Z_t}$$

and using also Example 5.1.1 we obtain the series expansion

$$f(X_t) = e^{X_t} = \exp(\mu_x + \sigma_x^2/2) \sum_{n=0}^{\infty} \frac{\sigma_x^n}{n!} H_n(Z_t).$$

Hence, the autocovariance function of Y_t (5.20) is given by the following

$$\begin{aligned}
\mathrm{Cov}(Y_t, Y_{t+h}) &= \exp(2\mu_x + \sigma_x^2) \sum_{n=1}^{\infty} \frac{(\sigma_x^2 \rho(h))^n}{n!} \\
&= \exp(2\mu_x + \sigma_x^2) \left\{ \sum_{n=0}^{\infty} \frac{(\sigma_x^2 \rho(h))^n}{n!} - \frac{(\sigma_x^2 \rho(h))^0}{0!} \right\} \\
&= \exp(2\mu_x + \sigma_x^2)\{\exp(\sigma_x^2 \rho(h)) - 1\} \\
&= \exp(2\mu_x + \sigma_x^2)\{\exp(\gamma(h)) - 1\} \qquad (5.21)
\end{aligned}$$

5.2.2 Log-Normal Representation

If a random variable X is normally distributed then the random variable $W = \exp(X)$ follows a Log-Normal distribution. Considering the Gaussian process X_t then the $W_t = \exp(X_t)$ has mean equal to $E(W_t) = \exp(\mu_x + \sigma_x^2/2)$. This is derived directly from the moment generating function of the Normal distribution. Therefore,

$$\mathrm{Cov}(W_t, W_{t+h}) = \mathrm{E}(W_t W_{t+h}) - \mathrm{E}(W_t)\mathrm{E}(W_{t+h}) = \mathrm{E}(W_t W_{t+h}) - \exp(2\mu_x + \sigma_x^2).$$

The mean $\mathrm{E}(W_t W_{t+h})$ is obtain using the bivariate normal distribution. More precisely,

$$(X_t, X_{t+h})$$

has a bivariate normal distribution with mean vector $(\mu, \mu)^T$ and covariance matrix given by

$$\Sigma(h) = \begin{bmatrix} \sigma_x^2 & \gamma(h) \\ \gamma(h) & \sigma_x^2 \end{bmatrix}.$$

Also, the moment generating function of the bivariate normal distribution is given by

$$M_{(X_t, X_{t+h})}(t_1, t_2) = e^{\boldsymbol{\mu}^T \boldsymbol{t} + \frac{1}{2}\boldsymbol{t}^T \Sigma \boldsymbol{t}} = e^{\mu t_1 + \mu t_2 + \frac{1}{2}[t_1^2 \sigma_x^2 + t_2^2 \sigma_x^2 + 2 t_1 t_2 \gamma(h)]}.$$

Using the above moment generating function and $t_1 = t_2 = 1$ we obtain the mean $\mathrm{E}(W_t W_{t+h})$ as

$$\mathrm{E}(W_t W_{t+h}) = e^{2\mu + \sigma_x^2 + \gamma(h)}.$$

The above representation yields the autocovariance function of Y_t, $\mathrm{Cov}(Y_t, Y_{t+h})$ from (5.20) as

$$\mathrm{Cov}(Y_t, Y_{t+h}) = e^{2\mu + \sigma_x^2 + \gamma(h)} - \exp(2\mu_x + \sigma_x^2) = \exp(2\mu_x + \sigma_x^2)\{\exp(\gamma(h)) - 1\}$$

which is directly equivalent to the approximation derived by the Hermite polynomial expansion.

5.2.3 Poisson-Charlier Expansion

Using the explicit formula (5.15) and (5.16) then we obtain the following explicit formula for the Charlier polynomials $\mathcal{C}_n(\mu; x)$:

$$\mathcal{C}_n(\mu; x) = \sqrt{\frac{\mu^n}{n!}} \frac{x!}{\mu^x} \sum_{k=0}^{n} (-1)^k \binom{n}{k} \frac{\mu^{x-k}}{(x-k)!}.$$

We are interested in deriving a formula for the covariance

$$\text{Cov}(f(x), f(y))$$

when the function $f(x)$ is the exponential function, that is $f(x) = e^x$. Using the result of Example 5.1.2 we have that

$$\begin{aligned} \text{Cov}(f(x), f(y)) &= \text{Cov}(\sum_{n=0}^{\infty} b_n \mathcal{C}_n(\mu; x), \sum_{m=0}^{\infty} b_m \mathcal{C}_m(\mu; y)) \\ &= \sum_{n=0}^{\infty} b_n^2 \text{Cov}(\mathcal{C}_n(\mu; x), \mathcal{C}_n(\mu; y)) \end{aligned}$$

where

$$\begin{aligned} &\text{Cov}(\mathcal{C}_n(\mu; x), \mathcal{C}_m(\mu; y)) \\ &\quad = \text{Cov}\left(\sqrt{\frac{\mu^n}{n!}} \frac{x!}{\mu^x} \sum_{k=0}^{n} (-1)^k \binom{n}{k} \frac{\mu^{x-k}}{(x-k)!}, \sqrt{\frac{\mu^n}{n!}} \frac{x!}{\mu^y} \sum_{l=0}^{n} (-1)^l \binom{n}{l} \frac{\mu^{y-l}}{(y-l)!} \right) \\ &\quad = \frac{\mu^n}{n!} \sum_{k=0}^{n} \sum_{l=0}^{n} (-1)^{k+l} \binom{n}{k} \binom{n}{l} \mu^{-(k+l)} \text{Cov}(\frac{x!}{(x-k)!}, \frac{y!}{(y-l)!}) \\ &\quad = \frac{\mu^n}{n!} \sum_{k=0}^{n} \sum_{l=0}^{n} \left(\frac{-1}{\mu}\right)^{k+l} \binom{n}{k} \binom{n}{l} \left\{ \text{E}\left(\frac{x!y!}{(x-k)!(y-l)!}\right) - \text{E}\left(\frac{x!}{(x-k)!}\right) \text{E}\left(\frac{\mu^{y-l}}{(y-l)!}\right) \right\} \\ &\quad = \frac{\mu^n}{n!} \sum_{k=0}^{n} \sum_{l=0}^{n} \left(\frac{-1}{\mu}\right)^{k+l} \binom{n}{k} \binom{n}{l} \left\{ \text{E}\left(\frac{x!y!}{(x-k)!(y-l)!}\right) - \mu^{k+l} \right\}. \qquad (5.22) \end{aligned}$$

For example, we present the covariances for the first few n

$$
\begin{aligned}
n=1: &\quad \mathrm{Cov}(\mathcal{C}_1(\mu;x),\mathcal{C}_1(\mu;y)) = \mu\left\{\frac{\mathrm{E}(xy)}{\mu^2}-1\right\}\\
n=2: &\quad \mathrm{Cov}(\mathcal{C}_2(\mu;x),\mathcal{C}_2(\mu;y)) = 2\mathrm{E}(xy)-\frac{1}{\mu}\{\mathrm{E}(xy(y-1))+\mathrm{E}(x(x-1)y)\}\\
&\qquad +\frac{1}{\mu^2}\mathrm{E}(x(x-1)y(y-1))-\frac{\mu^2}{2}\\
n=3: &\quad \mathrm{Cov}(\mathcal{C}_3(\mu;x),\mathcal{C}_3(\mu;y)) = \frac{3\mu}{2}\mathrm{E}(xy)-\frac{3}{2}\{\mathrm{E}(xy(y-1))+\mathrm{E}(x(x-1)y)\}\\
&\qquad +\frac{1}{2\mu}\{\mathrm{E}(xy(y-1)(y-2))+\mathrm{E}(x(x-1)(x-2)y)\}\\
&\qquad -\frac{1}{2\mu^2}\{\mathrm{E}(x(x-1)y(y-1)(y-2))+\mathrm{E}(x(x-1)(x-2)y(y-1))\}\\
&\qquad +\frac{3}{2\mu}\mathrm{E}(x(x-1)y(y-1))+\frac{1}{6\mu^3}\mathrm{E}(x(x-1)(x-2)y(y-1)(y-2))-\frac{\mu^3}{6}\\
n=4: &\quad \mathrm{Cov}(\mathcal{C}_4(\mu;x),\mathcal{C}_4(\mu;y)) = \frac{2\mu^2}{3}\mathrm{E}(xy)-\mu\{\mathrm{E}(xy(y-1))+\mathrm{E}(x(x-1)y)\}\\
&\qquad +\frac{2}{3}\{\mathrm{E}(xy(y-1)(y-2))+\mathrm{E}(x(x-1)(x-2)y)\}\\
&\qquad -\frac{1}{6\mu}\{\mathrm{E}(xy(y-1)(y-2)(y-3))+\mathrm{E}(x(x-1)(x-2)(x-3)y)\}\\
&\qquad +\frac{3}{2}\mathrm{E}(x(x-1)y(y-1))+\frac{2}{3\mu^2}\mathrm{E}(x(x-1)(x-2)y(y-1)(y-2))\\
&\qquad -\frac{1}{\mu}\{\mathrm{E}(x(x-1)y(y-1)(y-2))+\mathrm{E}(x(x-1)(x-2)y(y-1))\}\\
&\qquad +\frac{1}{4\mu^2}\{\mathrm{E}(x(x-1)y(y-1)(y-2)(y-3))+\mathrm{E}(x(x-1)(x-2)(x-3)y(y-1))\}\\
&\qquad -\frac{1}{6\mu^3}\{\mathrm{E}(x(x-1)(x-2)y(y-1)(y-2)(y-3))\\
&\qquad +\mathrm{E}(x(x-1)(x-2)(x-3)y(y-1)(y-2))\}\\
&\qquad +\frac{1}{24\mu^4}\mathrm{E}(x(x-1)(x-2)(x-3)y(y-1)(y-2)(y-3))-\frac{\mu^4}{24}.
\end{aligned}
$$

5.3 Discussion

We have considered a primary empirical study to examine the behavior of the approximations to the autocovariance function of the observed process through the orthogonal polynomial theory explored in the previous sections.

Notice that the approximation based on orthogonal polynomials depends on the joint distribution of X_t and X_{t+h}. In the case of the Hermite polynomials, this is a bivariate normal distribution. However, in the case of the Poisson-Charlier polynomials this is the joint distribution of two marginally distributed Poisson variables. More specifically, the autocovariance depends on factorial means of the two variables. Unfortunately, the joint distribution of two Poisson variables cannot be explicitly derived. In order to bypass this problem in a simulation, we have to empirically estimate these factorial means.

Our initial empirical results suggest that both the Hermite and the Poisson-Charlier representations fail to adequately approximate the autocovariance function of the response process. In fact, the Hermite approximation produces satisfactory results but for small values of the parameters d,a and b and only for the first few lags, whereas the Poisson-Charlier approximation is overall a very poor approach.

Chapter 6

Discussion and Further Research

6.1 Discussion

This work focuses on robust estimation for count time series models. The motivation for this study arises from the fact that classical statistical methods for inference appear to be immensely sensitive to interventions that result in observations that are distant from the bulk of the data. Initiating from simpler models for the analysis of count data, our goal was to investigate the behavior of several robust estimation procedures under extreme events, and thereafter advance the study to more complicated situations.

In the first part of this thesis, we review two models based on the Poisson distribution, which is the natural assumption on the distribution for the analysis of count data. In particular, we consider a linear and a log-linear Poisson model, both including a feedback component. We propose the log-linear model because it has more advantages and is widely applicable rather than the linear model. Three forms of intervention effects are described: level-shifts, which result in a permanent level change of the mean process, transient-shifts, whose effect on the level of the mean process is not permanent but decays exponentially, and additive outliers, whose effect is directly additive to the observations.

The second part of the thesis concentrates on the log-linear Poisson model that does not include the feedback mechanism (recall (3.3)). We propose two robust estimators, the conditionally unbiased bounded influence estimator (CUBIF) and the Mallows' quasi likelihood estimator (MQLE), and we compare them against the maximum likelihood estimator (MLE), for data contaminated with the aforementioned types of interventions.

Especially, the Mallows' quasi likelihood estimator is studied under four types of weighting schemes. The two choices consist of non robust types of weights and the other two are robust, and more specifically they are based on robust versions of the Mahalanobis distance. Our findings suggest that under level-shifts or transient-shifts, all estimation methods compared behave in a quite similar manner. However, this is not the case when additive outliers are implemented. We consider the cases of a single additive outlier, a patch of outliers and isolated outliers. Generally, the Mallows' quasi likelihood estimator is the best performing estimator in terms of mean square error, mean absolute error and bias, especially when robustly weighted.

Since additive outliers prove to be the most interesting type of intervention effects and the Mallows' quasi likelihood estimator the most prominent estimating procedure, in the third part of the thesis, we examine further the properties of the Mallows' quasi likelihood estimator suitably adjusted to the context of count time series. We convey to the first order log-linear Poisson model with a feedback mechanism (recall model (4.1)). By applying the so called perturbation technique suggested by Fokianos and Tjøstheim (2011), we show that the Mallows' quasi likelihood estimator is asymptotically normally distributed under some conditions on the model parameters. The key to proving asymptotic normality is martingale limit theory. When the data is contaminated with additive outliers, the robustly weighted MQLE exceeds the performance of the maximum likelihood estimator and estimates consistently the regression coefficients.

Additionally, we address the problem of testing whether the first order log-linear model with feedback can be reduced to a model that does not include the feedback mechanism, under the presence of additive outliers. We develop a robust score test and prove that the test statistic follows asymptotically a chi-square distribution. In the case where the data contain additive outliers, the test statistic based on MQLE with robust weights achieve the desirable size of the test and additionally have high power.

In the last part of this study, we approximate the autocovariance function of the time series process $\{Y_t\}$, using orthogonal polynomial expansions. Our approximation is based on the fact that the process $\{\log(1+Y_t)\}$ can be represented as an ARMA process. We therefore question whether the autocovariance function of the series be somehow approximated using the known autocovariance function of the process $\{\log(1+Y_t)\}$. We consider three families of orthogonal polynomials: the Hermite, Laguerre and Poisson-Charlier polynomials. The autocovariance function of the process Y_t is then described by a series that depends on the autocovariance and mean properties of $\{\log(1+Y_t)\}$. The more suitable approximation is suggested by the Hermite polynomials.

6.2 Further Research

A number of suggestions can be made as points for further research. As a first continuation, we plan to construct our R code on the Mallows' quasi likelihood estimation procedure for the log-linear Poisson model with feedback (Appendix B) into an R package. At this point we would like to acknowledge the R package `tscount` which has just been launched and includes functions for likelihood-based estimation analysis of integer-valued time series in the presence of interventions, as studied in the works of Fokianos and Fried (2010) and Fokianos and Fried (2012). For more information see Liboschik et al. (2015). The package provides the first available software for the analysis of dependent structured models following generalized linear model theory, and covers the distributional assumptions of both the Poisson and the Negative Binomial distributions. Considering that the functions available are based on the maximum likelihood estimator, the robust procedure of the Mallows' quasi likelihood estimator that we study, will provide an appealing alternative to MLE and advance the applicability of the package.

Furthermore, keeping in mind that certain conditions on the model parameters are imposed to guarantee stationarity and ergodicity, the problem of constrained optimization deserves more attention.

A notable remark is that even though in Chapter 3 we have studied the (p, q) order log-linear model without feedback, we have only considered the first order model with feedback in Chapter 4. The aim of Chapter 4 is to show how suitably chosen estimating functions can be employed for obtaining robust estimators for the regression coefficients of the log-linear model (4.1), especially in the presence of interventions. This chapter complements the works by Fokianos and Fried (2010, 2012) who studied detection and testing for intervention effects in count time series. Furthermore, we developed a robust test statistic for testing the existence of a feedback mechanism. There are some potential useful extensions of this methodology as we outline below. A first extension is the investigation of the Mallows' quasi likelihood estimator for a higher order log-linear model with feedback. More particularly, consider the model

$$Y_t \parallel \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \qquad \nu_t = d + \sum_{i=1}^{p} a_i \nu_{t-i} + \sum_{j=1}^{q} b_j \log(1 + Y_{t-j})$$

where the log-intensity is linked to the past p values of itself and the past q values of the response. Roughly speaking, we can still employ the quasi-score function (4.5) for estimation and inference with some necessary modifications. For instance, the derivative process $\partial \nu_t(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is a $p + q + 1$-dimensional vector whose elements are given by the following representation

$$\begin{aligned}
\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial d} &= 1 + \sum_{i=1}^{p} a_i \frac{\partial \nu_{t-i}(\boldsymbol{\theta})}{\partial d} \\
\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial a_k} &= \nu_{t-k} + \sum_{i=1}^{p} a_i \frac{\partial \nu_{t-i}(\boldsymbol{\theta})}{\partial a_k}, \quad k = 1, 2, \ldots, p \\
\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial b_l} &= \log(1 + Y_{t-l}) + \sum_{i=1}^{p} a_i \frac{\partial \nu_{t-i}(\boldsymbol{\theta})}{\partial b_l} \qquad l = 1, 2, ..., q.
\end{aligned}$$

However, there are no results reported in the literature about stationarity and ergodicity properties of the above model. Hence, given suitable conditions, we conjecture that a similar robust estimation theory can be developed. Although, deriving asymptotic inference for a higher order model with feedback is not an easy task, particularly due to regressing on a number of past values of the log-intensity process. As we have demonstrated for the first order model, derivation of asymptotic properties depends mostly on derivatives with respect to the feedback coefficient parameter.

A further problem would be to relax the Poisson assumption and to replace it by either a more general discrete distribution (see Christou and Fokianos (2015)) or by a mean specification relation like in Kedem and Fokianos (2002, Ch.4). An alternative is the negative binomial distribution, which has probability mass function given by

$$\mathrm{P}(Y = y) = \binom{y + r - 1}{y} \theta^y (1 - \theta)^r$$

and $\mathrm{E}(Y) = r\theta/(1-\theta)$ and $\mathrm{Var}(Y) = r\theta/(1-\theta)^2$. Linear count models based on the negative binomial distribution have previously been studied by several authors, for example Davis and Wu (2009), Zhu (2011) and Davis and Liu (2015). A possible log-linear negative binomial model is given for instance by

$$Y_t \parallel \mathcal{F}_{t-1} \sim \mathrm{NegBin}(r, \theta_t), \qquad \nu_t = d + \sum_{i=1}^{p} a_i \nu_{t-1} + \sum_{j=1}^{q} b_j \log(1 + Y_{t-j})$$

where $\nu_t = \log\{r\theta_t/(1-\theta_t)\}$. The negative binomial distribution belongs to the exponential family of distributions when r is fixed and therefore an MQLE type estimator may be constructed, as described in Cantoni and Ronchetti (2001). We think that both of these extensions are possible given suitable conditions on the moments of the joint process (Y_t, ν_t). Finally, given an estimating function we can also construct a test statistic analogous to (4.8) for robust testing of feedback presence.

Another interesting point for consideration concerns the ψ-function. Recall that the MQLE score estimating equation depends on the Huber function $\psi_c(x)$ which is used to impose a bound on the influence function of the estimator. However, other functions can be employed instead of the Huber function. For example, consider the Tukey's bisquare ψ function

$$\psi_{Tukey}(x) = \begin{cases} x\left[1 - \left(\frac{x}{c}\right)^2\right]^2, & |x| \le c \\ 0, & |x| > c \end{cases}$$

which is everywhere differentiable but only exists within the interval $[-c, c]$. Other contributions where the Tukey ψ-function is used include Elsaied and Fried (2014) who study robust M-estimation using a modified version of the Huber and Tukey ψ-functions but for a linear Poisson model without feedback. The Tukey's bisquare ψ-function read above may be used as an alternative choice of ψ-function for the MQLE estimator rather than the Huber ψ-function.

We have studied the Mallows' quasi likelihood estimator in the context of the log-linear Poisson model with feedback (2.1). Additionally, the estimator may be surveyed within the context of the linear Poisson model as well. In this context, the lemmas proved in chapter 4 are derived analogously.

Moreover, another important topic for further research is that of binary regression.

That is the response variable follows a binomial distribution. Cantoni and Ronchetti (2001) propose the Mallows' quasi likelihood estimator for binary data, with a logit link implemented in the generalized linear modeling setting. Other than the Mallows' quasi likelihood estimator, one may consider the BLq-estimator suggested byHosseinian (2009) or the Bianco-Yohai estimator suggested by Bianco and Yohai (1996) and studied further by Croux and Haesbroeck (2003), for binary processes.

Lastly, further study on the orthogonal polynomial approximation to the autocovariance and autocorrelation properties of the response process can be considered.

Appendix A

R Code for Chapter 3

Hyde Park Data Analysis

Load the data and plot a time series plot of the data.

```
HydePark=read.table("HydePark.txt",header=T)
ts.plot(HydePark, main="Hyde Park Time Series",ylab="Number of p
        urse snatchings")
```

Use the Chen and Liu (1993) method to detect outliers and interventions.

```
library(tsoutliers)
HydePark.ts=ts(HydePark)
HydePark.outliers=tso(HydePark.ts,maxit.iloop=10)   # detect outliers
HydePark.outliers$outliers
```

Use the method of Fokianos and Fried (2010, 2012) to detect outliers and interventions.

```
library(tscount)
fit=tsglm(HydePark.ts,model=list(past_obs=c(1,2),past_mean=1),
          link="log",distr="poisson")
interventionsHydePark=interv_multiple(fit)  # detect interventions
```

Apply the log linear Poisson model without feedback of order q=4, fit the models and choose the best model using the AIC criterion.

```
response=HydePark[5:71,]
x1=log(1+HydePark[4:70,])
x2=log(1+HydePark[3:69,])
x3=log(1+HydePark[2:68,])
x4=log(1+HydePark[1:67,])
dataout=as.data.frame(cbind(response,x1,x2,x3,x4))
fit1=glm(response~x1,data=dataout[1:57,],family=poisson)
fit2=glm(response~x1+x2,data=dataout[1:57,],family=poisson)
fit3=glm(response~x1+x2+x3,data=dataout[1:57,],family=poisson)
fit4=glm(response~x1+x2+x3+x4,data=dataout[1:57,],family=poisson)
summary(fit1)$aic # similarly we obtain the AIC for all 4 fits
```

The selected model based on the AIC criterion is the second order model (fit2). Fit the selected model. Matrices to save the estimated parameters and their standard errors.

```
library(robust)
cc=seq(1,3.5,length=50)
estCUBIF=estMQLEnone=estMQLEhat=estMQLEmve=
estMQLEmcd=matrix(NA,nrow=length(cc),ncol=3)
std.erCUBIF=std.erMQLEnone=std.erMQLEhat=
std.erMQLEmve=std.erMQLEmcd=matrix(NA,nrow=length(cc),ncol=3)
```

Matrices to save the predicted values of the response for all values of c. every row will be the 10 predicted values for each value of c.

```
CUBIFpr=MQLEnonepr=MQLEhat=prMQLEmvepr=
MQLEmcdpr=matrix(NA,nrow=length(cc),ncol=10)
```

Vectors to save the estimated MSE, MAD and MAE of the predicted values for each c.

```
mseCUBIF=mseMQLEnone=mseMQLEhat=mseMQLEmve=mseMQLEmcd=rep(NA,length(cc))
madCUBIF=madMQLEnone=madMQLEhat=madMQLEmve=madMQLEmcd=rep(NA,length(cc))
maeCUBIF=maeMQLEnone=maeMQLEhat=maeMQLEmve=maeMQLEmcd=rep(NA,length(cc))
```

Matrices to save the residuals from every fit. Each line corresponds to a value of c for CUBIF and MQLE.

```
residCUBIF=residMQLEnone=residMQLEhat=residMQLEmve=
residMQLEmcd=matrix(NA,nrow=length(cc),ncol=57)
```

Fit the chosen model using the estimators of interest and calculate, for all values of c, the estimated parameters of the model with their standard errors and the predicted values.

```
estMLE=fit2$coef    # MLE
std.erMLE=summary(fit2)$coef[,2]
residMLE=fit2$residuals
MLEpr=predict(fit2,dataout[58:67,],type="response")
for (i in 1:length(cc))
{
 fitCUBIF=glmRob(response~x1+x2,
            family=poisson(),data=dataout[1:57,],method="cubif",
             control=glmRob.cubif.control(bpar=cc[i]))
 estCUBIF[i,]=fitCUBIF$coef
 std.erCUBIF[i,]=summary(fitCUBIF)$coef[,2]
 residCUBIF[i,]=fitCUBIF$residuals
 nuhatCUBIF=rep(NA,10)
 lambdaCUBIF=rep(NA,10)
 for (t in 1:10)
 {
  nuhatCUBIF[t]=estCUBIF[i,1]+estCUBIF[i,2]*log(1+yCUBIF[t+57-1])+
               estCUBIF[i,3]*log(1+yCUBIF[t+57-2])
 lambdaCUBIF[t]=exp(nuhatCUBIF[t])
 CUBIFpr[i,t]=exp(nuhatCUBIF[t])
 }
 fitMQLEnone=glmrob(response~x1+x2,family=poisson,data=dataout[1:57,],
                  method="Mqle",control=glmrobMqle.control(tcc=cc[i]))
```

```
 estMQLEnone[i,]=fitMQLEnone$coef
 std.erMQLEnone[i,]=summary(fitMQLEnone)$coef[,2]
 residMQLEnone[i,]=fitMQLEnone$residuals
 MQLEnonepr[i,]=predict(fitMQLEnone,dataout[58:67,],type="response")
 fitMQLEhat=glmrob(response~x1+x2,family=poisson,data=dataout[1:57,],
 method="Mqle",weights.on.x="hat",control=glmrobMqle.control(tcc=cc[i]))
 estMQLEhat[i,]=fitMQLEhat$coef
 std.erMQLEhat[i,]=summary(fitMQLEhat)$coef[,2]
 residMQLEhat[i,]=fitMQLEhat$residuals
 MQLEhatpr[i,]=predict(fitMQLEhat,dataout[58:67,],type="response")
 fitMQLEmve=glmrob(response~x1+x2,family=poisson,data=dataout[1:57,],
 method="Mqle",weights.on.x="robCov",control=glmrobMqle.control(tcc=cc[i]))
 estMQLEmve[i,]=fitMQLEmve$coef
 std.erMQLEmve[i,]=summary(fitMQLEmve)$coef[,2]
 residMQLEmve[i,]=fitMQLEmve$residuals
 MQLEmvepr[i,]=predict(fitMQLEmve,dataout[58:67,],type="response")
 fitMQLEmcd=glmrob(response~x1+x2,family=poisson,data=dataout[1:57,],
 method="Mqle",weights.on.x="covMcd",control=glmrobMqle.control(tcc=cc[i]))
 estMQLEmcd[i,]=fitMQLEmcd$coef
 std.erMQLEmcd[i,]=summary(fitMQLEmcd)$coef[,2]
 residMQLEmcd[i,]=fitMQLEmcd$residuals
 MQLEmcdpr[i,]=predict(fitMQLEmcd,dataout[58:67,],type="response")
}
```

Vector for MLE and matrices for CUBIF and MQLE in which the first 57 observations will be the same as the original ones and the last 10 observations will be the predicted ones. In the cases of CUBIF and MQLE, each row of the matrices corresponds to a value of c.

```
yMLE=rep(NA,67)
yCUBIF=yMQLEnone=yMQLEhat=yMQLEmve=yMQLEmcd=
          matrix(NA,nrow=length(cc),ncol=67)
```

```
for (i in 1:57)
{
  yMLE[i]=HydePark[i+4,]
  for (j in 1:length(cc))
  {
    yCUBIF[j,i]=yMQLEnone[j,i]=yMQLEhat[j,i]=yMQLEmve[j,i]=
                            yMQLEmcd[j,i]=HydePark[i+4,]
  }
}
for (i in 1:10)
{
  yMLE[i+57]=MLEpr[i]
  for (j in 1:length(cc))
  {
    yCUBIF[j,i+57]=CUBIFpr[j,i]
    yMQLEnone[j,i+57]=MQLEnonepr[j,i]
    yMQLEhat[j,i+57]=MQLEhatpr[j,i]
    yMQLEmve[j,i+57]=MQLEmvepr[j,i]
    yMQLEmcd[j,i+57]=MQLEmcdpr[j,i]
  }
}
```

Find the MSE, MAD and MAE of the estimators.

```
mseMLE=(sum((HydePark[58:67,]-yMLE[58:67])^2))/10
madMLE=mad(yMLE[58:67])
maeMLE=mean(abs(yMLE[58:67]-HydePark[58:67,]))
for (i in 1:length(cc))
{
  mseCUBIF[i]=(sum((HydePark[58:67,]-yCUBIF[i,58:67])^2))/10
  mseMQLEnone[i]=(sum((HydePark[58:67,]-yMQLEnone[i,58:67])^2))/10
  mseMQLEhat[i]=(sum((HydePark[58:67,]-yMQLEhat[i,58:67])^2))/10
```

```
  mseMQLEmve[i]=(sum((HydePark[58:67,]-yMQLEmve[i,58:67])^2))/10
  mseMQLEmcd[i]=(sum((HydePark[58:67,]-yMQLEmcd[i,58:67])^2))/10
  madCUBIF[i]=mad(yCUBIF[i,58:67])
  madMQLEnone[i]=mad(yMQLEnone[i,58:67])
  madMQLEhat[i]=mad(yMQLEhat[i,58:67])
  madMQLEmve[i]=mad(yMQLEmve[i,58:67])
  madMQLEmcd[i]=mad(yMQLEmcd[i,58:67])
  maeCUBIF[i]=mean(abs(yCUBIF[i,58:67]-HydePark[58:67,]))
  maeMQLEnone[i]=mean(abs(yMQLEnone[i,58:67]-HydePark[58:67,]))
  maeMQLEhat[i]=mean(abs(yMQLEhat[i,58:67]-HydePark[58:67,]))
  maeMQLEmve[i]=mean(abs(yMQLEmve[i,58:67]-HydePark[58:67,]))
  maeMQLEmcd[i]=mean(abs(yMQLEmcd[i,58:67]-HydePark[58:67,]))
}
```

Appendix B

R Code for Chapter 4

The libraries `robustreg` and `robust` are required to obtain the Huber function and use the `glmrob` function respectively and the packages `foreach` and `doMC` are required for parallel computing.

```
library(robustreg)
library(robust)
library(foreach)
library(doMC)
registerDoMC(cores=12)
```

Program constructed to generate data from the first order log-linear Poisson model with feedback that includes a patch of additive outliers.
`kappa` is the number of outliers of size `w` that occur at time `tau` and `size` is the size of the generated series.

```
loglinearpoisson.ts.patchAO=function(theta,w,tau,kappa,size)
{
  z=rep(NA,size) # contaminated response
  y=rep(NA,size) # clean response
  lambda=nu=rep(NA,size)
  nu[1]=0 # starting value of nu
  lambda[1]=1 # starting value of the mean
```

```
  z[1]=y[1]=rpois(1,lambda[1])
  for (t in 2:size)
  {
    nu[t]=theta[1]+theta[2]*nu[t-1]+theta[3]*log(y[t-1]+1)
    lambda[t]=exp(nu[t])
    y[t]=rpois(1,lambda[t])
    z[t]=y[t]
  }
  end=tau+kappa-1
  for (i in tau:end)
  {
    z[i]=z[i]+w
  }
  return(cbind(z,nu))
}
```

Programs to construct the log-likelihood and score functions, and the Information matrix corresponding to the MLE.

`theta` is the vector of parameters and `data` is the generated data vector using the above data generating function.

```
loglikelihood.poisson=function(theta,data)
{
  loglik=nu=rep(NA,times=length(data))
  nu[1]=1
  loglik[1]=0
  ldata=log(data+1)
  for (t in 2:length(data))
  {
    nu[t]=theta[1]+theta[2]*nu[t-1]+theta[3]*ldata[t-1]
    loglik[t]=-data[t]*nu[t]+exp(nu[t])
  }
```

```
  final=sum(loglik)
}

score.poisson <- function(theta, data)
{
  nu=first=second=third=rep(NA, times=length(data))
  nu[1]=first[1]=second[1]=third[1]=1
  s1=s2=s3=rep(NA, times=length(data))
  ldata=log(data+1)
  for (t in 2:length(data))
  {
    nu[t]=theta[1]+theta[2]*nu[t-1]+theta[3]*ldata[t-1]
    first[t]= (1+theta[2]*first[t-1])
    second[t]=(nu[t-1]+theta[2]*second[t-1])
    third[t]=(ldata[t-1]+theta[2]*third[t-1])
    s1[t]=-( (data[t]-exp(nu[t])))*first[t]
    s2[t]=-( (data[t]-exp(nu[t])))*second[t]
    s3[t]=-( (data[t]-exp(nu[t])))*third[t]
  }
  ss1=sum(s1[-1])
  ss2=sum(s2[-1])
  ss3=sum(s3[-1])
  score=c(ss1,ss2,ss3) # score dianisma
}

mlefromloglinear=function(theta,data)
{
  results=rep(NA,length(theta))
 # Calculate initial values using LSE. We fit an ARMA(1,1) model
  r1=arima(data,order=c(1,0,1),method="CSS")
  phi=r1$coef[1]
```

```
  thetarima=r1$coef[2]
  mu=sigma2=r1$coef[3]
  start=c(mu*(1-phi),sigma2*(1-phi),-thetarima,(phi+thetarima))
  results=optim(loglikelihood.poisson,p=c(start[2],start[3],start[4]),
            data=data,score.poisson,method="BFGS")$par
  return(results)
}

information1.logpoisson <- function(theta, data)
{
theta1=first=second=third=rep(NA, times=length(data))
theta1[1]=first[1]=second[1]=third[1]=1
Information=matrix(0, nrow=3, ncol=3)
s1=s2=s3=rep(NA,times=length(data))
for (t in 2:length(data))
{
theta1[t]=theta[1]+theta[2]*theta1[t-1]+theta[3]*log(data[t-1]+1)
first[t]= (1+theta[2]*first[t-1])
second[t]=(theta1[t-1]+theta[2]*second[t-1])
third[t]=(log(data[t-1]+1)+theta[2]*third[t-1])
s1[t]=first[t]
s2[t]=second[t]
s3[t]=third[t]
var.comp= (exp(theta1[t]/2))*c(s1[t], s2[t], s3[t])
Information=Information+var.comp%*%t(var.comp)
}
return(Information)
}
```

Program to construct the score of the MQLE.

`cc` is the value of the tuning constant c of the Huber function and `wt` is the vector of weights.

```
score.MQLE.poisson.A=function(theta,data,cc,wt)
{
 library(robustreg)
  nu=rep(NA,times=length(data))
  nu[1]=1
  first=second=third=rep(NA,times=length(data))   # derivative of nu
  first[1]=second[2]=third[1]=1
  r.stand=rep(NA,times=length(data)) # Pearson residuals
  r.stand[1]=(data[1]-exp(nu[1]))/sqrt(exp(nu[1]))
  Hub=rep(NA,times=length(data)) # Huber function
  Hub[1]=psiHuber(r.stand[1],cc)
  gs1=gs2=gs3=rep(NA,times=length(data))
  gs1[1]=gs2[1]=gs3[1]=1
  jinf=jsup=rep(NA,times=length(data)) # j1 and j2
  jinf[1]=floor(exp(nu[1])-cc*sqrt(exp(nu[1])))
  jsup[1]=floor(exp(nu[1])+cc*sqrt(exp(nu[1])))
  epsi=rep(NA,times=length(data))
  epsi[1]=-cc*ppois(jinf[1],exp(nu[1])) + cc*(1-ppois(jsup[1],exp(nu[1])))
   + sqrt(exp(nu[1]))*(ppois(jinf[1],exp(nu[1]))-ppois(jinf[1]-1,exp(nu[1]))
    -(ppois(jsup[1],exp(nu[1])) - ppois(jsup[1]-1,exp(nu[1]))))
  alpha=matrix(NA,nrow=length(data),ncol=3)
  alpha[1,1]=epsi[1]*wt[1]*sqrt(exp(nu[1]))*first[1]
  alpha[1,2]=epsi[1]*wt[1]*sqrt(exp(nu[1]))*second[1]
  alpha[1,3]=epsi[1]*wt[1]*sqrt(exp(nu[1]))*third[1]
  alpha.theta=rep(NA,3) # bias correction term alpha(theta)
  s1=s2=s3=rep(NA, times=length(data))
  s1[1]=s2[1]=s3[1]=0
  ldata=log(data+1)
  for (t in 2:length(data))
  {
    nu[t]=theta[1]+theta[2]*nu[t-1]+theta[3]*ldata[t-1]
```

```
    first[t]= (1+theta[2]*first[t-1])
    second[t]=(nu[t-1]+theta[2]*second[t-1])
    third[t]=(ldata[t-1]+theta[2]*third[t-1])
    r.stand[t]=(data[t]-exp(nu[t]))/sqrt(exp(nu[t]))
    Hub[t]=psiHuber(r.stand[t],cc)
    gs1[t]=Hub[t]*wt[t]*sqrt(exp(nu[t]))*first[t]
    gs2[t]=Hub[t]*wt[t]*sqrt(exp(nu[t]))*second[t]
    gs3[t]=Hub[t]*wt[t]*sqrt(exp(nu[t]))*third[t]
    jinf[t]=floor(exp(nu[t])-cc*sqrt(exp(nu[t])))
    jsup[t]=floor(exp(nu[t])+cc*sqrt(exp(nu[t])))
    epsi[t]=-cc*ppois(jinf[t],exp(nu[t])) + cc*(1-ppois(jsup[t],exp(nu[t])))
     + sqrt(exp(nu[t]))*(ppois(jinf[t],exp(nu[t]))-ppois(jinf[t]-1,exp(nu[t]))
      -(ppois(jsup[t],exp(nu[t])) - ppois(jsup[t]-1,exp(nu[t]))))
    alpha[t,1]=epsi[t]*wt[t]*sqrt(exp(nu[t]))*first[t]
    alpha[t,2]=epsi[t]*wt[t]*sqrt(exp(nu[t]))*second[t]
    alpha[t,3]=epsi[t]*wt[t]*sqrt(exp(nu[t]))*third[t]
  }
  alpha.theta=apply(alpha,2,mean)
  s1[-1]=gs1[-1]-alpha.theta[1]  # first score component
  s2[-1]=gs2[-1]-alpha.theta[2]  # second score component
  s3[-1]=gs3[-1]-alpha.theta[3]  # third score component
  ss1=sum(s1)
  ss2=sum(s2)
  ss3=sum(s3)
  score=c(ss1,ss2,ss3) # score vector
}

grad.score.A=function(theta,data,cc,wt)
{
  delta=.Machine$double.eps^.5
  Ident=diag(1,length(theta))
```

```
  thetadelta=theta+delta*Ident
  score.derivative1=(score.MQLE.poisson.A(thetadelta[,1],data,cc,wt)-
                     score.MQLE.poisson.A(theta,data,cc,wt))/delta
  score.derivative2=(score.MQLE.poisson.A(thetadelta[,2],data,cc,wt)-
                     score.MQLE.poisson.A(theta,data,cc,wt))/delta
  score.derivative3=(score.MQLE.poisson.A(thetadelta[,3],data,cc,wt)-
                     score.MQLE.poisson.A(theta,data,cc,wt))/delta
  result=rbind(score.derivative1,score.derivative2,score.derivative3)
  result=as.matrix(result)
  return(result)
}

solve.score.MQLE.A=function(theta,data,cc,wt)
{
  library(MASS) # to use the ginv() command for the inverse of a matrix
  mytol=(.Machine$double.eps)^(1/2)
  max.it=200
  # Calculate initial values using LSE. We fit an ARMA(1,1) model
  r1=arima(data, order=c(1,0,1), method="CSS")
  phi=r1$coef[1]
  thetarima=r1$coef[2]
  mu=sigma2=r1$coef[3]
  start=c(mu*(1-phi), sigma2*(1-phi), -thetarima, (phi+thetarima))
  start1=optim(loglikelihood.poisson, p=c(start[2], start[3], start[4]),
               data=data, score.poisson , method="BFGS")$par
  theta.old=start1
  it=0 # iterations
  test=1
  while(abs(test)>mytol && (it<-it+1)<max.it)
  {
    g.old=score.MQLE.poisson.A(theta.old,data,cc,wt)
```

```
    grad.score.old=grad.score.A(theta.old,data,cc,wt)
    csi=(-1)*(ginv(grad.score.old))%*%(g.old)
    theta.new=as.vector(theta.old+csi)
    test=max((abs(theta.new-theta.old))/(abs(theta.old)))
    theta.old=theta.new
    cat("This is the",it,"iteration","\n")
  }
return(theta.old)
}
```

Program to construct the asymptotic covariance matrix of MQLE. The program contains some commands that are the same as in the program `score.MQLE.poisson.A`. These commands are omitted and we only present the commands that differ.

```
Asym.Variance.Matrix=function(theta,data,cc,wt)
{
  library(robustreg)
⋮
  s1[-1]=gs1[-1]-alpha.theta[1]
  s2[-1]=gs2[-1]-alpha.theta[2]
  s3[-1]=gs3[-1]-alpha.theta[3]
  for (t in 2:length(data))
  {
   matrix.W=matrix.W+(c(s1[t],s2[t],s3[t]))%*%t(c(s1[t],s2[t],s3[t]))
   matrix.H=matrix.H+(Hub[t]*wt[t]*sqrt(exp(nu[t]))*(data[t]-exp(nu[t])))*
      (c(first[t],second[t],third[t])%*%t(c(first[t],second[t],third[t])))
  }
  matrix.W=matrix.W
  H_antistrofo=ginv(matrix.H)
  asym=H_antistrofo%*%matrix.W%*%H_antistrofo
  return(asym)
}
```

Program to estimate the parameter values of the first order log-linear Poisson model with feedback using MQLE.

```
estim_patch=function(theta,w,tau,kappa,size,sim,cc)
{
 simulations.data=foreach(i=1:sim,.combine='rbind',.multicombine=TRUE,
                   .inorder=FALSE,.errorhandling="pass")%dopar%
 {
  cat("\n\n********** Now doing simulation",i,"of", sim, "**********\n\n")
  myalldata=loglinearpoisson.ts.patchAO(theta,w,tau,kappa,size)
  mydata=myalldata[301:size,1]
  mynudata=myalldata[301:size,2]
  # create the design matrix X (method A):
  thetahatA=solve.score.MQLE(theta,mydata,cc)
  ldata=log(mydata+1)
  nuhatA=rep(NA,length(mydata))
  nuhatA[1]=1
  for (t in 2:length(mydata))
  {
   nuhatA[t]=thetahatA[1]+thetahatA[2]*nuhatA[t-1]+thetahatA[3]*ldata[t-1]
  }
  X_A=cbind(nuhatA,ldata)
  Xti_A=(t(X_A))%*%X_A
  Xtitr_A=solve(Xti_A)
  HatMat_A=X_A%*%Xtitr_A%*%(t(X_A)) # hat matrix
  hi_A=diag(HatMat_A) #diagonal elements of the hat matrix
  weights.hii_A=sqrt(1-hi_A) # hat weights
  mve_A=cov.rob(X_A,method="mve")
  mve_center_A=mve_A$center
  mve_cov_A=mve_A$cov
  maha.mve_A=sqrt(mahalanobis(X_A, mve_center_A, mve_cov_A))
  weights.mve_A=1/maha.mve_A    # mve weights
```

```
mcd_A=cov.rob(X_A,method="mcd")
mcd_center_A=mve_A$center
mcd_cov_A=mve_A$cov
maha.mcd_A=sqrt(mahalanobis(X_A, mcd_center_A, mcd_cov_A))
weights.mcd_A=1/maha.mcd_A    # mcd weights
# Create the X matrix (B)
M=20 # truncation constant M
X_B=matrix(NA,nrow=length(mydata),ncol=M)
myally=myalldata[,1]
all_ldata=log(1+myally)
for (i in 1:length(mydata)){ #rows
for (j in 1:M){ #columns
   X_B[i,j]=all_ldata[300+i-(j-1)]
}}
Xti_B=(t(X_B))%*%X_B
Xtitr_B=solve(Xti_B)
HatMat_B=X_B%*%Xtitr_B%*%(t(X_B)) # hat matrix
hi_B=diag(HatMat_B) #diagonal elements of the hat matrix
weights.hii_B=sqrt(1-hi_B) # hat weights
mve_B=cov.rob(X_B,method="mve")
mve_center_B=mve_B$center
mve_cov_B=mve_B$cov
maha.mve_B=sqrt(mahalanobis(X_B, mve_center_B, mve_cov_B))
weights.mve_B=1/maha.mve_B    # mve weights
mcd_B=cov.rob(X_B,method="mcd")
mcd_center_B=mve_B$center
mcd_cov_B=mve_B$cov
maha.mcd_B=sqrt(mahalanobis(X_B, mcd_center_B, mcd_cov_B))
weights.mcd_B=1/maha.mcd_B    # mcd weights
estim=c(mlefromloglinear(theta,mydata),
 solve(information1.logpoisson(mlefromloglinear(theta,mydata),mydata)),
```

```
    solve.score.MQLE(theta,mydata,cc,rep(1,500)),
    Asym.Variance.Matrix(solve.score.MQLE(theta,mydata,cc),mydata,cc,rep(1,500)),
    solve.score.MQLE.A(theta,mydata,cc,weights.hii_A),
    Asym.Variance.Matrix(solve.score.MQLE.A(theta,mydata,cc,weights.hii_A),
    mydata,cc,weights.hii_A),solve.score.MQLE.A(theta,mydata,cc,weights.mve_A),
    Asym.Variance.Matrix(solve.score.MQLE.A(theta,mydata,cc,weights.mve_A),
    mydata,cc,weights.mve_A),solve.score.MQLE.A(theta,mydata,cc,weights.mcd_A),
    Asym.Variance.Matrix(solve.score.MQLE.A(theta,mydata,cc,weights.mcd_A),
    mydata,cc,weights.mcd_A),solve.score.MQLE.A(theta,mydata,cc,weights.hii_B),
    Asym.Variance.Matrix(solve.score.MQLE.A(theta,mydata,cc,weights.hii_B),
    mydata,cc,weights.hii_B),solve.score.MQLE.A(theta,mydata,cc,weights.mve_B),
    Asym.Variance.Matrix(solve.score.MQLE.A(theta,mydata,cc,weights.mve_B),
    mydata,cc,weights.mve_B),solve.score.MQLE.A(theta,mydata,cc,weights.mcd_B),
    Asym.Variance.Matrix(solve.score.MQLE.A(theta,mydata,cc,weights.mcd_B),
    mydata,cc,weights.mcd_B))
  }
 return(simulations.data)
}
```

Testing Program. Again, the commands that are the same as in the `score.MQLE.poisson.A` program are omitted.

```
Asym.Variance.Matrix_testing=function(theta,data,cc,wt)
{
 library(robustreg)
⋮
  s1=s2=s3rep(NA, times=length(data))
  s1[1]=s2[1]=s3[1]=0
  ldata=log(data+1)
  s1MLE=s2MLE=s3MLE=rep(NA, times=length(data))
  matrix.V=matrix.W=matrix.H=matrix(0,nrow=3,ncol=3)
  for (t in 2:length(data))
```

```
{
  nu[t]=theta[1]+theta[2]*nu[t-1]+theta[3]*ldata[t-1]
  first[t]= (1+theta[2]*first[t-1])
  second[t]=(nu[t-1]+theta[2]*second[t-1])
  third[t]=(ldata[t-1]+theta[2]*third[t-1])
  r.stand[t]=(data[t]-exp(nu[t]))/sqrt(exp(nu[t]))
  Hub[t]=psiHuber(r.stand[t],cc)
  gs1[t]=Hub[t]*wt[t]*sqrt(exp(nu[t]))*first[t]
  gs2[t]=Hub[t]*wt[t]*sqrt(exp(nu[t]))*second[t]
  gs3[t]=Hub[t]*wt[t]*sqrt(exp(nu[t]))*third[t]
  jinf[t]=floor(exp(nu[t])-cc*sqrt(exp(nu[t])))
  jsup[t]=floor(exp(nu[t])+cc*sqrt(exp(nu[t])))
  epsi[t]=-cc*ppois(jinf[t],exp(nu[t])) + cc*(1-ppois(jsup[t],exp(nu[t])))
  + sqrt(exp(nu[t]))*(ppois(jinf[t],exp(nu[t]))-ppois(jinf[t]-1,exp(nu[t]))
  -(ppois(jsup[t],exp(nu[t])) - ppois(jsup[t]-1,exp(nu[t]))))
  alpha[t,1]=epsi[t]*wt[t]*sqrt(exp(nu[t]))*first[t]
  alpha[t,2]=epsi[t]*wt[t]*sqrt(exp(nu[t]))*second[t]
  alpha[t,3]=epsi[t]*wt[t]*sqrt(exp(nu[t]))*third[t]
  s1MLE[t]=-( (data[t]-exp(nu[t])))*first[t]
  s2MLE[t]=-( (data[t]-exp(nu[t])))*second[t]
  s3MLE[t]=-( (data[t]-exp(nu[t])))*third[t]
}
alpha.theta=apply(alpha,2,mean)
s1[-1]=gs1[-1]-alpha.theta[1]
s2[-1]=gs2[-1]-alpha.theta[2]
s3[-1]=gs3[-1]-alpha.theta[3]
for (t in 2:length(data))
{
 matrix.V=matrix.V+(c(s1[t],s2[t],s3[t]))%*%t((c(s1MLE[t],s2MLE[t],s3MLE[t])))
 matrix.W=matrix.W+(c(s1[t],s2[t],s3[t]))%*%t(c(s1[t],s2[t],s3[t]))
 matrix.H=matrix.H+(Hub[t]*wt[t]*sqrt(exp(nu[t]))*(data[t]-exp(nu[t])))*
```

```
        (c(first[t],second[t],third[t])%*%t(c(first[t],second[t],third[t])))
    }
    matrix.V=matrix.V
    matrix.W=matrix.W
    V_antistrofo=ginv(matrix.V)
    H_antistrofo=ginv(matrix.H)
    asym=H_antistrofo%*%matrix.W%*%H_antistrofo
    return(list(mat.W=matrix.W,mat.V=matrix.H,asym.mat=asym))
}
testing_program=function(theta,w,tau,kappa,size,sim,cc)
{
 library(robust)
 score.test.none=score.test.hat=score.test.MVE=
 score.test.MCD=matrix(NA,nrow=length(cc),ncol=sim)
 for (k in 1:length(cc))
 {
  for (i in 1:sim)
  {
   # Generate under the null hypothesis, from the model without feedback
   mydata=loglinearpoisson.ts.patchAO(theta,w,tau,kappa,size)[301:size]
   response=mydata[2:(size-300)]
   x=log(1+mydata[1:(size-300-1)])
   dataout=as.data.frame(cbind(response,x))
   # Obtain the MQLE estimates under the null hypothesis
   fit_MQLE_none=glmrob(response~x,family=poisson,data=dataout,method="Mqle",
                control=glmrobMqle.control(tcc=cc[k]))
   fit_MQLE_hat=glmrob(response~x,family=poisson,data=dataout,method="Mqle",
               weights.on.x="hat",control=glmrobMqle.control(tcc=cc[k]))
   fit_MQLE_MVE=glmrob(response~x,family=poisson,data=dataout,method="Mqle",
                weights.on.x="robCov",control=glmrobMqle.control(tcc=cc[k]))
   fit_MQLE_MCD=glmrob(response~x,family=poisson,data=dataout,method="Mqle",
```

```
               weights.on.x="covMcd",control=glmrobMqle.control(tcc=cc[k]))
# MQLE score partition corresponding to a
MQLE.score.none=score.MQLE.poisson.A(c(fit_MQLE_none$coef[1],0,
fit_MQLE_none$coef[2]),mydata,cc[k],wt=rep(1,500))
MQLE.score.hat=score.MQLE.poisson.A(c(fit_MQLE_hat$coef[1],0,
fit_MQLE_hat$coef[2]),mydata,cc[k],wt=c(1,summary(fit_MQLE_hat)$w.x))
MQLE.score.MVE=score.MQLE.poisson.A(c(fit_MQLE_MVE$coef[1],0,
fit_MQLE_MVE$coef[2]),mydata,cc[k],wt=c(1,summary(fit_MQLE_MVE)$w.x))
MQLE.score.MCD=score.MQLE.poisson.A(c(fit_MQLE_MCD$coef[1],0,
fit_MQLE_MCD$coef[2]),mydata,cc[k],wt=c(1,summary(fit_MQLE_MCD)$w.x))
# Partitions of the matrices W and V
#Without weights
Asym.matrices.none=Asym.Variance.Matrix_testing(c(fit_MQLE_none$coef[1],
                   0,fit_MQLE_none$coef[2]),mydata,cc[k],wt=rep(1,500))
Matrix_V.none=Asym.matrices.none$mat.V
Matrix_W.none=Asym.matrices.none$mat.W
W_11.none=rbind(c(Matrix_W.none[1,1],Matrix_W.none[1,3]),
                c(Matrix_W.none[3,1],Matrix_W.none[3,3]))
W_12.none=rbind(Matrix_W.none[1,2],Matrix_W.none[3,2])
W_21.none=cbind(Matrix_W.none[2,1],Matrix_W.none[2,3])
W_22.none=Matrix_W.none[2,2]
V_11.none=rbind(c(Matrix_V.none[1,1],Matrix_V.none[1,3]),
                c(Matrix_V.none[3,1],Matrix_V.none[3,3]))
V_12.none=rbind(Matrix_V.none[1,2],Matrix_V.none[3,2])
V_21.none=cbind(Matrix_V.none[2,1],Matrix_V.none[2,3])
V_22.none=Matrix_V.none[2,2]
# With hat weights
Asym.matrices.hat=Asym.Variance.Matrix_testing(c(fit_MQLE_hat$coef[1],
   0,fit_MQLE_hat$coef[2]),mydata,cc[k],wt=c(1,summary(fit_MQLE_hat)$w.x))
Matrix_V.hat=Asym.matrices.hat$mat.V
Matrix_W.hat=Asym.matrices.hat$mat.W
```

```
W_11.hat=rbind(c(Matrix_W.hat[1,1],Matrix_W.hat[1,3]),
               c(Matrix_W.hat[3,1],Matrix_W.hat[3,3]))
W_12.hat=rbind(Matrix_W.hat[1,2],Matrix_W.hat[3,2])
W_21.hat=cbind(Matrix_W.hat[2,1],Matrix_W.hat[2,3])
W_22.hat=Matrix_W.hat[2,2]
V_11.hat=rbind(c(Matrix_V.hat[1,1],Matrix_V.hat[1,3]),
               c(Matrix_V.hat[3,1],Matrix_V.hat[3,3]))
V_12.hat=rbind(Matrix_V.hat[1,2],Matrix_V.hat[3,2])
V_21.hat=cbind(Matrix_V.hat[2,1],Matrix_V.hat[2,3])
V_22.hat=Matrix_V.hat[2,2]
# With robust Mahalanobis MVE weights
Asym.matrices.MVE=Asym.Variance.Matrix_testing(c(fit_MQLE_MVE$coef[1],
   0,fit_MQLE_MVE$coef[2]),mydata,cc[k],wt=c(1,summary(fit_MQLE_MVE)$w.x))
Matrix_V.MVE=Asym.matrices.MVE$mat.V
Matrix_W.MVE=Asym.matrices.MVE$mat.W
W_11.MVE=rbind(c(Matrix_W.MVE[1,1],Matrix_W.MVE[1,3]),
               c(Matrix_W.MVE[3,1],Matrix_W.MVE[3,3]))
W_12.MVE=rbind(Matrix_W.MVE[1,2],Matrix_W.MVE[3,2])
W_21.MVE=cbind(Matrix_W.MVE[2,1],Matrix_W.MVE[2,3])
W_22.MVE=Matrix_W.MVE[2,2]
V_11.MVE=rbind(c(Matrix_V.MVE[1,1],Matrix_V.MVE[1,3]),
               c(Matrix_V.MVE[3,1],Matrix_V.MVE[3,3]))
V_12.MVE=rbind(Matrix_V.MVE[1,2],Matrix_V.MVE[3,2])
V_21.MVE=cbind(Matrix_V.MVE[2,1],Matrix_V.MVE[2,3])
V_22.MVE=Matrix_V.MVE[2,2]
# With robust Mahalanobis MCD weights
Asym.matrices.MCD=Asym.Variance.Matrix_testing(c(fit_MQLE_MCD$coef[1],
   0,fit_MQLE_MCD$coef[2]),mydata,cc[k],wt=c(1,summary(fit_MQLE_MCD)$w.x))
Matrix_V.MCD=Asym.matrices.MCD$mat.V
Matrix_W.MCD=Asym.matrices.MCD$mat.W
W_11.MCD=rbind(c(Matrix_W.MCD[1,1],Matrix_W.MCD[1,3]),
```

```
                c(Matrix_W.MCD[3,1],Matrix_W.MCD[3,3]))
  W_12.MCD=rbind(Matrix_W.MCD[1,2],Matrix_W.MCD[3,2])
  W_21.MCD=cbind(Matrix_W.MCD[2,1],Matrix_W.MCD[2,3])
  W_22.MCD=Matrix_W.MCD[2,2]
  V_11.MCD=rbind(c(Matrix_V.MCD[1,1],Matrix_V.MCD[1,3]),
                c(Matrix_V.MCD[3,1],Matrix_V.MCD[3,3]))
  V_12.MCD=rbind(Matrix_V.MCD[1,2],Matrix_V.MCD[3,2])
  V_21.MCD=cbind(Matrix_V.MCD[2,1],Matrix_V.MCD[2,3])
  V_22.MCD=Matrix_V.MCD[2,2]
  # calculate Sigma
  Sigma.none=W_22.none-V_21.none%*%solve(V_11.none)%*%W_12.none
       -W_21.none%*%solve(V_11.none)%*%V_12.none
       +V_21.none%*%solve(V_11.none)%*%W_11.none%*%solve(V_11.none)%*%V_12.none
  Sigma.hat=W_22.hat-V_21.hat%*%solve(V_11.hat)%*%W_12.hat
       -W_21.hat%*%solve(V_11.hat)%*%V_12.hat
       +V_21.hat%*%solve(V_11.hat)%*%W_11.hat%*%solve(V_11.hat)%*%V_12.hat
  Sigma.MVE=W_22.MVE-V_21.MVE%*%solve(V_11.MVE)%*%W_12.MVE
       -W_21.MVE%*%solve(V_11.MVE)%*%V_12.MVE
       +V_21.MVE%*%solve(V_11.MVE)%*%W_11.MVE%*%solve(V_11.MVE)%*%V_12.MVE
  Sigma.MCD=W_22.MCD-V_21.MCD%*%solve(V_11.MCD)%*%W_12.MCD
       -W_21.MCD%*%solve(V_11.MCD)%*%V_12.MCD
       +V_21.MCD%*%solve(V_11.MCD)%*%W_11.MCD%*%solve(V_11.MCD)%*%V_12.MCD
  # Calculation of the score test
  score.test.none[k,i]=MQLE.score.none[2]*MQLE.score.none[2]/Sigma.none
  score.test.hat[k,i]=MQLE.score.hat[2]*MQLE.score.hat[2]/Sigma.hat
  score.test.MVE[k,i]=MQLE.score.MVE[2]*MQLE.score.MVE[2]/Sigma.MVE
  score.test.MCD[k,i]=MQLE.score.MCD[2]*MQLE.score.MCD[2]/Sigma.MCD
 }
}
return(list(score.test.no.weights=score.test.none,score.test.hat.weights
  =score.test.hat,score.test.MVE.weights=score.test.MVE,
```

```
  score.test.MCD.weights=score.test.MCD))
}
```

Measles Data Analysis

Load the data from the library `tscount` and plot a time series plot.

```
library(tscount)    # to load the data
measles1=measles[1:150,3]
ts.plot(measles1)   # time series plot of the data
```

Detect outliers using the Chen and Liu (1993) method from the `tsoutliers` package.

```
library(tsoutliers)
measles1.ts=ts(measles1)    # convert the data into a time series format
measles1.outliers=tso(measles1.ts,maxit.iloop=10)
measles1.outliers$outliers
```

Fit an ARIMA(1,0,1)=ARMA(1,1) model to obtain initial estimates of the parameter estimates (theta)

```
r1.measles1=arima(measles1,order=c(1,0,1),method="CSS")
phi.measles1=r1.measles1$coef[1]
thetarima.measles1=r1.measles1$coef[2]
mu.measles1=sigma2.measles1=r1.measles1$coef[3]
theta.initial.measles1=c(sigma2.measles1*(1-phi.measles1),
-thetarima.measles1,(phi.measles1+thetarima.measles1))
theta.initial.measles1
```

Obtain the parameter estimates using MLE and MQLE as described in Chapter 4.

```
# MLE
theta.MLE.measles1=mlefromloglinear(theta.initial.measles1,measles1)
asym.cov.MLE.measles1=solve(information1.logpoisson(theta.initial.measles1,
```

```
                                                            measles1))
# MQLE no weights
cc=seq(1,5,length=15)
theta.MQLEnone.measles1=matrix(NA,nrow=length(cc),
              ncol=length(theta.initial.measles1))
asym.cov.MQLEnone.measles1=array(NA,dim=c(length(theta.initial.measles1),
                           length(theta.initial.measles1),length(cc)))
for (i in 1:length(cc))
{
 theta.MQLEnone.measles1[i,]=solve.score.MQLE.A(theta.initial.measles1,
                             measles1,cc[i],wt=rep(1,length(measles1)))
 asym.cov.MQLEnone.measles1[,,i]=Asym.Variance.Matrix(theta.initial.measles1,
                             measles1,cc[i],wt=rep(1,length(measles1)))
}
# MQLE with weights
# Method A:
thetahatA.measles1=matrix(NA,nrow=length(cc),ncol=length(theta.initial.measles1))
ldata.measles1=log(measles1+1)
nuhatA.measles1=matrix(NA,nrow=length(cc),ncol=length(measles1))
weights.hii_A.measles1=weights.mve_A.measles1=
  weights.mcd_A.measles1=matrix(NA,nrow=length(cc),ncol=length(measles1))
theta.MQLEhatA.measles1=theta.MQLEmveA.measles1=
  theta.MQLEmcdA.measles1=matrix(NA,nrow=length(cc),
  ncol=length(theta.initial.measles1))
matrix(NA,nrow=length(cc),ncol=length(theta.initial.measles1))
asym.cov.MQLEhatA.measles1=asym.cov.MQLEmveA.measles1=asym.cov.MQLEmcdA.measles1=
  array(NA,dim=c(length(theta.initial.measles1),
  length(theta.initial.measles1),length(cc)))
for (i in 1:length(cc))
{
 nuhatA.measles1[i,1]=1
```

```
 thetahatA.measles1[i,]=solve.score.MQLE(theta.initial.measles1,measles1,cc[i])
 for (t in 2:length(measles1))
 {
  nuhatA.measles1[i,t]=thetahatA.measles1[i,1]
                     +thetahatA.measles1[i,2]*nuhatA.measles1[i,(t-1)]
                     +thetahatA.measles1[i,3]*ldata.measles1[t-1]
 }
 X_A.measles1=cbind(nuhatA.measles1[i,],ldata.measles1)
 Xti_A.measles1=(t(X_A.measles1))%*%X_A.measles1
 Xtitr_A.measles1=solve(Xti_A.measles1)
 HatMat_A.measles1=X_A.measles1%*%Xtitr_A.measles1%*%(t(X_A.measles1))
 hi_A.measles1=diag(HatMat_A.measles1) #diagonal elements of the hat matrix
 weights.hii_A.measles1[i,]=sqrt(1-hi_A.measles1)
 mve_A.measles1=cov.rob(X_A.measles1,method="mve")
 mve_center_A.measles1=mve_A.measles1$center
 mve_cov_A.measles1=mve_A.measles1$cov
 maha.mve_A.measles1=sqrt(mahalanobis(X_A.measles1,
        mve_center_A.measles1, mve_cov_A.measles1))
 weights.mve_A.measles1[i,]=pmin(rep(1,length(measles1)),1/maha.mve_A.measles1)
 mcd_A.measles1=cov.rob(X_A.measles1,method="mcd")
 mcd_center_A.measles1=mcd_A.measles1$center
 mcd_cov_A.measles1=mcd_A.measles1$cov
 maha.mcd_A.measles1=sqrt(mahalanobis(X_A.measles1,
        mcd_center_A.measles1, mcd_cov_A.measles1))
 weights.mcd_A.measles1[i,]=pmin(rep(1,length(measles1)),1/maha.mcd_A.measles1)
}
for (i in 1:length(cc))
{
 theta.MQLEhatA.measles1[i,]=solve.score.MQLE.A(theta.initial.measles1,
                         measles1,cc[i],wt=weights.hii_A.measles1[i,])
 asym.cov.MQLEhatA.measles1[,,i]=Asym.Variance.Matrix(theta.initial.measles1,
```

```
                         measles1,cc[i],wt=weights.hii_A.measles1[i,])
 theta.MQLEmveA.measles1[i,]=solve.score.MQLE.A(theta.initial.measles1,
                         measles1,cc[i],wt=weights.mve_A.measles1[i,])
 asym.cov.MQLEmveA.measles1[,,i]=Asym.Variance.Matrix(theta.initial.measles1,
                         measles1,cc[i],wt=weights.mve_A.measles1[i,])
 theta.MQLEmcdA.measles1[i,]=solve.score.MQLE.A(theta.initial.measles1,
                         measles1,cc[i],wt=weights.mcd_A.measles1[i,])
 asym.cov.MQLEmcdA.measles1[,,i]=Asym.Variance.Matrix(theta.initial.measles1,
                         measles1,cc[i],wt=weights.mcd_A.measles1[i,])
}
# Method B:
theta.MQLEhatB.measles1=theta.MQLEmveB.measles1=theta.MQLEmcdB.measles1=
          matrix(NA,nrow=length(cc),ncol=length(theta.initial.measles1))

asym.cov.MQLEhatB.measles1=asym.cov.MQLEmveB.measles1=asym.cov.MQLEmcdB.measles1=
                          array(NA,dim=c(length(theta.initial.measles1),
                          length(theta.initial.measles1),length(cc)))
M=10
X_B.measles1=matrix(NA,nrow=(length(measles1)-M),ncol=M)
ldata=log(1+measles1)
for (i in 1:(length(measles1)-M)){
for (j in 1:M){
 X_B.measles1[i,j]=ldata[M+i-(j-1)]
}}
  Xti_B.measles1=(t(X_B.measles1))%*%X_B.measles1
  Xtitr_B.measles1=solve(Xti_B.measles1)
  HatMat_B.measles1=X_B.measles1%*%Xtitr_B.measles1%*%(t(X_B.measles1))
  hi_B.measles1=diag(HatMat_B.measles1) #diagonal elements of the hat matrix
  weights.hii_B.measles1=sqrt(1-hi_B.measles1) #these are the hat weights
  mve_B.measles1=cov.rob(X_B.measles1,method="mve")
  mve_center_B.measles1=mve_B.measles1$center
```

```
  mve_cov_B.measles1=mve_B.measles1$cov
  maha.mve_B.measles1=sqrt(mahalanobis(X_B.measles1,
         mve_center_B.measles1, mve_cov_B.measles1))
  weights.mve_B.measles1=pmin(rep(1,length(measles1)-M),1/maha.mve_B.measles1)
  mcd_B.measles1=cov.rob(X_B.measles1,method="mcd")
  mcd_center_B.measles1=mcd_B.measles1$center
  mcd_cov_B.measles1=mcd_B.measles1$cov
  maha.mcd_B.measles1=sqrt(mahalanobis(X_B.measles1,
         mcd_center_B.measles1, mcd_cov_B.measles1))
  weights.mcd_B.measles1=pmin(rep(1,length(measles1)-M),1/maha.mcd_B.measles1)
for (i in 1:length(cc))
{
  theta.MQLEhatB.measles1[i,]=solve.score.MQLE.A(theta.initial.measles1,
                     measles1[11:150],cc[i],wt=weights.hii_B.measles1)
  asym.cov.MQLEhatB.measles1[,,i]=Asym.Variance.Matrix(theta.initial.measles1,
                     measles1[11:150],cc[i],wt=weights.hii_B.measles1)
  theta.MQLEmveB.measles1[i,]=solve.score.MQLE.A(theta.initial.measles1,
                     measles1[11:150],cc[i],wt=weights.mve_B.measles1)
  asym.cov.MQLEmveB.measles1[,,i]=Asym.Variance.Matrix(theta.initial.measles1,
                     measles1[11:150],cc[i],wt=weights.mve_B.measles1)
  theta.MQLEmcdB.measles1[i,]=solve.score.MQLE.A(theta.initial.measles1,
                     measles1[11:150],cc[i],wt=weights.mcd_B.measles1)
  asym.cov.MQLEmcdB.measles1[,,i]=Asym.Variance.Matrix(theta.initial.measles1,
                     measles1[11:150],cc[i],wt=weights.mcd_B.measles1)
}
```

Standard deviations of the parameter estimates:

```
sdnone=sdhatA=sdmveA=sdmcdA=sdhatB=sdmveB=sdmcdB=matrix(NA,nrow=15,ncol=3)
for (i in 1:15){
sdnone[i,]=sqrt(diag(asym.cov.MQLEnone.measles1[,,i]))
sdhatA[i,]=sqrt(diag(asym.cov.MQLEhatA.measles1[,,i]))
```

```
sdmveA[i,]=sqrt(diag(asym.cov.MQLEmveA.measles1[,,i]))
sdmcdA[i,]=sqrt(diag(asym.cov.MQLEmcdA.measles1[,,i]))
sdhatB[i,]=sqrt(diag(asym.cov.MQLEhatB.measles1[,,i]))
sdmveB[i,]=sqrt(diag(asym.cov.MQLEmveB.measles1[,,i]))
sdmcdB[i,]=sqrt(diag(asym.cov.MQLEmcdB.measles1[,,i]))
}
```

Bibliography

Abadir, K. M. and G. Talmain (2005). Autocovariance functions of series and of their transforms. *J. Econometrics* ***124***, 227–252.

Abraham, B. and G. E. P. Box (1979). Bayesian analysis of some outlier problems in time series. *Biometrika* ***66***, 229–236.

Barczy, M., M. Ispány, G. Pap, M. Scotto, and M. E. Silva (2012). Additive outliers in INAR(1) models. *Statist. Papers* ***53***, 935–949.

Bianco, A. M. and V. J. Yohai (1996). Robust estimation in the logistic regression model. In *Robust statistics, data analysis, and computer intensive methods (Schloss Thurnau, 1994)*, Volume **109** of *Lecture Notes in Statist.*, pp. 17–34. Springer-Verlag.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* ***31***, 307–327.

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.

Breslow, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *J. Amer. Statist. Assoc.* ***85***, 565–571.

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods* (Second ed.). Springer-Verlag.

Cantoni, E. and E. Ronchetti (2001). Robust inference for generalized linear models. *J. Amer. Statist. Assoc.* ***96***, 1022–1030.

Chan, W. S. (1992). A note on time series model specification in the presence of outliers. *J. Appl. Stat.* ***19***, 117–124.

Chang, I., G. C. Tiao, and C. Chen (1988). Estimation of time series parameters in the presence of outliers. *Technometrics* ***30***, 193–204.

Chen, C. and L. Liu (1993). Joint estimation of model parameters and outlier effects in time-series. *J. Amer. Statist. Assoc.* ***88***, 284–297.

Chihara, T. S. (1978). *An introduction to orthogonal polynomials*. Gordon and Breach Science Publishers, New York-London-Paris. Mathematics and its Applications, Vol. 13.

Chow, Y. S. (1967). On a strong law of large numbers for martingales. *Ann. Math. Statist.* ***38***, 610.

Christou, V. and K. Fokianos (2014). Quasi-likelihood inference for negative binomial time series models. *J. Time Series Anal.* ***35***, 55–78.

Christou, V. and K. Fokianos (2015). Estimation and testing linearity for mixed Poisson autoregerssions. *Electron. J. Stat.* ***9***, 1357–1377.

Cox, D. R. (1981). Statistical analysis of time series: some recent developments. *Scand. J. Statist. Theory and Applications* ***8***, 93–115. (with discussion).

Croux, C. and G. Haesbroeck (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Comput. Statist. Data Anal.* ***44***, 273–295. Special issue in honour of Stan Azen: a birthday celebration.

Davis, R. A., W. T. M. Dunsmuir, and S. B. Streett (2003). Observation-driven models for Poisson counts. *Biometrika* ***90***, 777–790.

Davis, R. A. and H. Liu (2015). Theory and inference for a class of nonlinear models with application to time series of counts. *Statistica Sinica, to appear*.

Davis, R. A. and R. Wu (2009). A negative binomial model for time series of counts. *Biometrika* ***96***, 735–749.

Deutsch, S. J., J. E. Richards, and J. J. Swain (1990). Effects of a single outlier on ARMA identification. *Comm. Statist. Theory Methods* ***19***, 2207–2227.

Douc, R., P. Doukhan, and E. Moulines (2013). Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stochastic Process. Appl.* ***123***, 2620–2647.

Doukhan, P., K. Fokianos, and D. Tjøstheim (2012). On weak dependence conditions for Poisson autoregressions. *Statist. Probab. Lett.* ***82***, 942–948.

Elsaied, H. (2012). *Robust Modelling of Count Time Series: Applications in Medicine.* Ph. D. thesis, TU Dortmund University, Germany.

Elsaied, H. and R. Fried (2014). Robust fitting of INARCH models. *J. Time Ser. Anal.* ***35***, 517–535.

Ferland, R., A. Latour, and D. Oraichi (2006). Integer-valued GARCH process. *J. Time Ser. Anal.* ***27***, 923–942.

Fokianos, K. (2001). Truncated Poisson regression for time series of counts. *Scand. J. Statist.* ***28***, 645–659.

Fokianos, K. (2012). Count time series models. In T. S. Rao, S. S. Rao, and C. R. Rao (Eds.), *Handbook of Statistics: Time Series Analysis–Methods and Applications*, Volume **30**, pp. 315–347. Amsterdam: Elsevier B. V.

Fokianos, K. and R. Fried (2010). Interventions in INGARCH processes. *J. Time Ser. Anal.* ***31***, 210–225.

Fokianos, K. and R. Fried (2012). Interventions in log-linear Poisson autoregression. *Stat. Model.* ***12***, 299–322.

Fokianos, K., A. Rahbek, and D. Tjøstheim (2009). Poisson autoregression. *J. Amer. Statist. Assoc.* ***104***, 1430–1439.

Fokianos, K. and D. Tjøstheim (2011). Log-linear Poisson autoregression. *J. Multivariate Anal.* ***102***, 563–578.

Fokianos, K. and D. Tjøstheim (2012). Nonlinear Poisson autoregression. *Ann. Inst. Statist. Math.* ***64***, 1205–1225.

Fox, A. J. (1972). Outliers in time series. *J. Roy. Statist. Soc. Ser. B* ***34***, 350–363.

Francq, C. and J.-M. Zakoïan (2010). *GARCH models: Structure, Statistical Inference and Financial Applications.* Wiley, Hoboken, NJ.

Fried, R., I. Agueusop, B. Bornkamp, K. Fokianos, J. Fruth, and K. Ickstadt (2015). Retrospective bayesian outlier detection in ingarch series. *Stat. Comput.* ***25***, 365–374.

Fried, R., H. El-Saied, and K. Fokianos (2011). Outliers and interventions in INGARCH time series. *in Conesa, D. et al. (Ed.): Proceedings of IWSM 2011, Valencia*, 234–239.

Fried, R., H. Elsaied, T. Liboschik, K. Fokianos, and S. Kitromilidou (2014). On outliers and interventions in count time series following glms. *Austrian Journal of Statistics* ***43***, 181–193.

Granger, C. W. J. and P. Newbold (1976). Forecasting transformed series. *J. Roy. Statist. Soc. Ser. B* ***38***(2), 189–203.

Hall, P. and C. C. sHeyde (1980). *Martingale Limit Theory and its Applications.* New York: Academic Press.

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

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust statistics*. Wiley, Hoboken, NJ. The approach based on influence functions.

Harvey, A. C. (1990). *The Econometric Analysis of Time Series* (Second ed.). MIT Press, Cambridge, MA.

Heinen, A. (2003). Modelling time series count data : an autoregressive conditional poisson model. Internetausg.: http://www.core.ucl.ac.be/services/psfiles/dp03/dp2003-62.pdf.

Heritier, S. and E. Ronchetti (1994). Robust bounded-influence tests in general parametric models. *J. Amer. Statist. Assoc.* ***89***, 897–904.

Hosseinian, S. (2009). *Robust Inference for Generalized Linear Models: Binary and Poisson Regression.* Ph. D. thesis, Ecole Polytechnique Federale de Lausanne, Switzerland.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* ***35***, 73–101.

Huber, P. J. and E. M. Ronchetti (2009). *Robust statistics* (Second ed.). Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ.

Jensen, S. T. and A. Rahbek (2007). On the law of large numbers for (geometrically) ergodic Markov chains. *Econometric Theory* ***23***, 761–766.

Jung, R. C., M. Kukuk, and R. Liesenfeld (2006). Time series of count data: modeling, estimation and diagnostics. *Comput. Statist. Data Anal.* ***51***, 2350–2364.

Kedem, B. and K. Fokianos (2002). *Regression models for time series analysis.* Wiley, Hoboken, NJ.

Kitromilidou, S. and K. Fokianos (2015). Robust estimation methods for a class of count time series log-linear models. *J. Stat. Comput. Simul.* to appear.

Klimko, L. A. and P. I. Nelson (1978). On conditional least squares estimation for stochastic processes. *Ann. Statist.* ***6***, 629–642.

Krasker, W. S. and R. E. Welsch (1982). Efficient bounded-influence regression estimation. *J. Amer. Statist. Assoc.* ***77***, 595–604.

Künsch, H. R., L. A. Stefanski, and R. J. Carroll (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *J. Amer. Statist. Assoc.* ***84***, 460–466.

Lebedev, N. N. (1972). *Special functions and their applications.* Dover Publications, Inc., New York. Revised edition, translated from the Russian and edited by Richard A. Silverman, Unabridged and corrected republication.

Li, W. K. (1994). Time series models based on generalized linear models: some further results. *Biometrics* ***50***, 506–511.

Liboschik, T., K. Fokianos, and R. Fried (2015). tscount: An r package for analysis of count time series following generalized linear models. *SFB 823 Discussion Paper 6/15, (http://hdl.handle.net/2003/33904).*

Liboschik, T., P. Kerschke, K. Fokianos, and R. Fried (2014). Modelling interventions in INGARCH processes. *Int. J. Comput. Math. , to appear*.

Lô, S. N. and E. Ronchetti (2009). Robust and accurate inference for generalized linear models. *J. Multivariate Anal.* ***100***, 2126–2136.

MacDonald, I. L. and W. Zucchini (1997). *Hidden Markov and Other Models for Discrete–valued Time Series*. London: Chapman & Hall.

Maronna, R. A., R. D. Martin, and V. J. Yohai (2006). *Robust Statistics*. Wiley, Hoboken, NJ.

Martin, R. D. (1983). Robust-resistant spectral analysis. In *Time series in the frequency domain*, Volume **3** of *Handbook of Statist.*, pp. 185–219. North-Holland, Amsterdam.

McCleary, R., R. A. Hay, E. E. Meidinger, and D. McDowall (1980). *Applied time series analysis for the social sciences*. Sage Publications Beverly Hills, CA.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.

Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear models. *Biometrika* ***79***, 747–754.

Mukherjee, K. (2008). *M*-estimation in GARCH models. *Econometric Theory* ***24***, 1530–1553.

Muler, N. and V. J. Yohai (2002). Robust estimates for ARCH processes. *J. Time Ser. Anal.* ***23***, 341–375.

Muler, N. and V. J. Yohai (2008). Robust estimates for GARCH models. *J. Statist. Plann. Inference* ***138***, 2918–2940.

Neumann, M. H. (2011). Absolute regularity and ergodicity of Poisson count processes. *Bernoulli* ***17***, 1268–1284.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robert, C. P. and G. Casella (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag.

Rousseeuw, P. J. and K. V. Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* ***41***, 212–223.

Rousseeuw, P. J. and B. C. van Zomeren (1990). Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.* ***85***, 633–639.

Rydberg, T. H. and N. Shephard (2000). A modeling framework for the prices and times of trades on the New York stock exchange. In W. J. Fitzgerlad, R. L. Smith, A. T. Walden, and P. C. Young (Eds.), *Nonlinear and Nonstationary Signal Processing*, pp. 217–246. Cambridge: Isaac Newton Institute and Cambridge University Press.

Schoutens, W. (2000). *Stochastic processes and orthogonal polynomials*, Volume **146** of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Seber, G. F. and A. J. Lee (2003). *Linear regression analysis* (Second ed.). Wiley, Hoboken, NJ.

Streett, S. B. (2000). *Some observation driven models for time series*. Ph. D. thesis, Colorado State University, USA.

Szegő, G. (1975). *Orthogonal polynomials* (Fourth ed.). American Mathematical Society, Providence, R.I. American Mathematical Society, Colloquium Publications, Vol. XXIII.

Taniguchi, M. and Y. Kakizawa (2000). *Asymptotic theory of statistical inference for time series*. Springer Series in Statistics. New York: Springer-Verlag.

Tjøstheim, D. (2012). Some recent theory for autoregressive count time series. *Test* ***21***, 413–438. (with discussion).

Tsay, R. (1986). Time series model specification in the presence of outliers. *J. Amer. Statist. Assoc.* ***81***, 132–141.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *J. Forecast.* ***7***, 1–20.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to probability and statistics*, pp. 448–485. Stanford Univ. Press, Stanford, Calif.

Valdora, M. and V. J. Yohai (2014). Robust estimators for generalized linear models. *J. Statist. Plann. Inference* ***146***, 31–48.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge: Cambridge University Press.

Weiß, C. H. (2009). Modelling time series of counts with overdispersion. *Stat. Methods Appl.* ***18***, 507–519.

Woodard, D. W., D. S. Matteson, and S. G. Henderson (2011). Stationarity of count-valued and nonlinear time series models. *Electron. J. Stat.* ***5***, 800–828.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika* ***75***, 621–629.

Zeger, S. L. and B. Qaqish (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics.* ***44***, 1019–1031.

Zhu, F. (2011). A negative binomial integer-valued GARCH model. *J. Time Series Anal.* ***32***, 54–67.

Zucchini, W. and I. L. MacDonald (2009). *Hidden Markov models for time series.* CRC Press, Boca Raton, FL.