Some Statistical Models for The Multivariate Analysis of Longitudinal Data

by

Peter Xue-Kun Song

B.Sc., Jilin University, China, 1985
M.Sc., Southwest Jiaotong University, China, 1987

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

in
THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

October 1996

©Peter X.-K. Song, 1996
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Oct. 10, 96
ABSTRACT

This thesis develops some statistical models for the multivariate analysis of longitudinal data on the basis of the dispersion models of Jørgensen (1987a, 1996), consisting of three topics: multivariate dispersion models and their application to regression analysis, stationary time series models with non-normal margins and state space models with Markov latent processes. The goal of the thesis is to develop statistical models which can accommodate features of both trend and dependence for longitudinal data.

This thesis focusses mainly on the following three types of longitudinal data, namely (1) many short time series, (2) a few long stationary time series and (3) a few long non-stationary time series with time-varying covariates.

A class of multivariate dispersion models is proposed to deal with data of type (1), in a spirit similar to multivariate analysis based on the multivariate normal distribution. Under these multivariate parametric models, population-averaged models (Diggle et al., 1994) are revisited, where approximate inferences for regression parameters are presented, including the generalized estimating equation (GEE) of Liang and Zeger (1986) as a special case.

The thesis also presents a class of stationary autoregressive moving-average (ARMA) models with exponential dispersion model margins for data of type (2). The class of ARMA models is defined as a special case of a class of stationary infinite order moving average processes constructed by means of the thinning operation of Joe (1996a).

For analysis of type (3) data, two classes of state space models, including one with stationary latent processes and another with non-stationary latent processes, are proposed. To estimate regression parameters in both classes of models, we develop an algorithm for solving the so-called Kalman estimating equation (KEE), corresponding to a modified EM-algorithm where the E-step is approximated by the Kalman smoother that estimates the latent process via the best linear unbiased predictor (BLUP).
Two simulation studies are conducted in the thesis based on Poisson–gamma models. One is for the comparison of the efficiency of the KEE approach and the Monte Carlo EM (MCEM) algorithm. The other simulation study is for the examination of the utility of the model diagnosis for detecting the misspecification of stationarity and non-stationarity for latent process.

The thesis contains two data analyses. One data set consists of daily counts of emergency room visits for respiratory diseases to the hospital of Prince George, British Columbia, along with covariates of air pollution variables and meteorological variables. These data are analyzed through state space models to investigate the relationship between air pollution and respiratory morbidity. The other data set, consisting of the monthly number of poliomyelitis cases in the USA from 1970 to 1983, is analyzed based on the Poisson stationary-gamma model to study whether or not there is an evidence of a decreasing trend in the rate of polio infections in the USA.
Table of Contents

Abstract ii

List of Tables ix

List of Figures x

Acknowledgements xii

1 Introduction 1
  1.1 General ................................................. 1
  1.2 Literature review ....................................... 3
  1.3 Overview of topics ..................................... 5

2 Dispersion models 10
  2.1 Definition .............................................. 10
  2.2 Exponential dispersion models ......................... 12
  2.3 Tweedie class .......................................... 14
  2.4 Proper dispersion models ............................... 15
  2.5 Thinning operation .................................... 16
  2.6 Multivariate dispersion model densities ............... 19

3 Multivariate dispersion models 24
  3.1 Copulas ................................................. 26
  3.2 Multivariate normal copulas ........................... 28
3.3 Interpretation of dependence parameters .......................... 30
3.4 Multivariate dispersion models ........................................ 35
3.5 Normal scores and their approximations ............................. 36
3.6 Small-dispersion asymptotics for joint density ...................... 40
3.7 Approximate covariance ................................................. 43

4 Likelihood-based inference for population-averaged models 45
4.1 Models ................................................................. 46
4.2 Approximate inference based on the Pearson residual .............. 49
4.3 Approximate inference based on the deviance residual ............. 51
4.4 Approximate inference based on the modified deviance residual . 54
4.5 Implementation of the estimating procedure ....................... 56

5 Stationary models with exponential dispersion model margins 58
5.1 Introduction ............................................................. 58
5.2 Moving average processes of infinite order ......................... 62
  5.2.1 Hidden process .................................................. 63
  5.2.2 Autocorrelation function ...................................... 64
5.3 Autoregressive moving average processes ............................ 65
5.4 Autoregressive processes ............................................. 66
  5.4.1 EDAR(1) process ............................................. 67
  5.4.2 Partial autocorrelation function ................................ 68
5.5 Truncation error ........................................................ 71
5.6 Further questions ..................................................... 72

6 BLUP, Kalman filter and smoother ...................................... 74
6.1 Some general results for linear prediction ............................ 74
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2</td>
<td>General forms of Kalman filter and smoother</td>
<td>80</td>
</tr>
<tr>
<td>7</td>
<td>State space models with stationary latent process</td>
<td>86</td>
</tr>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>86</td>
</tr>
<tr>
<td>7.2</td>
<td>Models</td>
<td>91</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Definition</td>
<td>91</td>
</tr>
<tr>
<td>7.2.2</td>
<td>Moment structure</td>
<td>92</td>
</tr>
<tr>
<td>7.3</td>
<td>Kalman filter and smoother</td>
<td>93</td>
</tr>
<tr>
<td>7.3.1</td>
<td>Derivation</td>
<td>94</td>
</tr>
<tr>
<td>7.3.2</td>
<td>Mean square error matrix</td>
<td>96</td>
</tr>
<tr>
<td>7.4</td>
<td>Estimation</td>
<td>97</td>
</tr>
<tr>
<td>7.4.1</td>
<td>Estimation of regression parameters</td>
<td>98</td>
</tr>
<tr>
<td>7.4.2</td>
<td>Estimating the dispersion and correlation parameters</td>
<td>100</td>
</tr>
<tr>
<td>7.4.3</td>
<td>Calculation of Godambe information matrix</td>
<td>101</td>
</tr>
<tr>
<td>7.5</td>
<td>Model checking</td>
<td>102</td>
</tr>
<tr>
<td>7.6</td>
<td>Data analysis</td>
<td>105</td>
</tr>
<tr>
<td>7.6.1</td>
<td>Poisson stationary-gamma model</td>
<td>107</td>
</tr>
<tr>
<td>7.6.2</td>
<td>Implementation of estimation</td>
<td>108</td>
</tr>
<tr>
<td>7.6.3</td>
<td>Numerical results</td>
<td>109</td>
</tr>
<tr>
<td>7.6.4</td>
<td>Model diagnosis</td>
<td>111</td>
</tr>
<tr>
<td>7.7</td>
<td>Proof of asymptotic normality</td>
<td>115</td>
</tr>
<tr>
<td>8</td>
<td>State space models with non-stationary latent process</td>
<td>121</td>
</tr>
<tr>
<td>8.1</td>
<td>Introduction</td>
<td>121</td>
</tr>
<tr>
<td>8.2</td>
<td>Models</td>
<td>123</td>
</tr>
<tr>
<td>8.3</td>
<td>Kalman filter and smoother</td>
<td>129</td>
</tr>
<tr>
<td>8.4</td>
<td>Parameter estimation</td>
<td>131</td>
</tr>
</tbody>
</table>
# List of Tables

2.1 Summary of Tweedie exponential dispersion models. 14
2.2 Summary of contractions. 18

7.1 Monthly number of U.S. cases of poliomyelitis for 1970 to 1983. 106
7.2 Estimates of coefficients and standard errors 110

9.1 Estimates and standard errors from KEE and MCEM. 162
9.2 Overview of the simulation strategy for model diagnosis. 164
9.3 Estimates of coefficients given by Model I based on the data generated by Model II. 166
9.4 Estimates of coefficients given by Model II based on the data generated by Model I. 167

10.1 Estimates and standard errors for air pollution effects. 180
10.2 Empirical variance-covariance matrix for standardized residuals. 190
10.3 Estimates and standard errors for short-term effects. 191
List of Figures

3.1 Four bivariate normal copulas with different values of $\gamma$. ................. 29
3.2 Monte Carlo simulation plots of three dependence measures ................. 34
3.3 Comparison of three types of residuals from gamma distribution. ........ 41

5.1 Yearly mean sunspot numbers. ......................................................... 59

6.1 The comb structure. ................................................................. 80

7.1 Monthly counts of Polio incidences. ............................................ 106
7.2 Checking correlation structures for both latent and observed processes. .... 112
7.3 Checking distribution assumptions for the latent and observed process. .... 113
7.4 Checking the linearity of the time trend. ........................................ 114
7.5 Checking the latent structure. ....................................................... 115

8.1 Simulation of 100 observations from a process with four categories. ....... 127

9.1 168 simulated counts from the Poisson-gamma model. ........................ 156
9.2 Log-profile-likelihood versus the dispersion parameter given increasingly. 158
9.3 Log-profile-likelihood versus the dispersion parameter given decreasingly. 159
9.4 Patterns for log-likelihood and coefficients estimates. ........................ 161
9.5 Comparison between the true latent process and the estimated latent pro-
cesses. ................................................................................................. 162
9.6 ACF and PACF functions for the residuals of the latent process. ............ 165
9.7 168 simulated counts from a state space model. .................................. 167
<table>
<thead>
<tr>
<th>Section Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.8  ACF and PACF functions for the residuals from the latent process.</td>
<td>168</td>
</tr>
<tr>
<td>10.1 Emergency room visits for four respiratory diseases.</td>
<td>174</td>
</tr>
<tr>
<td>10.2 Meteorological and air pollution variables.</td>
<td>175</td>
</tr>
<tr>
<td>10.3 First-year residuals against second-year residuals for the four categories.</td>
<td>181</td>
</tr>
<tr>
<td>10.4 First-year residuals against second-year residuals for latent process.</td>
<td>182</td>
</tr>
<tr>
<td>10.5 Residuals against log-linear predictions for the four categories.</td>
<td>183</td>
</tr>
<tr>
<td>10.6 Residuals against log-linear predictions for latent process.</td>
<td>184</td>
</tr>
<tr>
<td>10.7 Autocorrelation function of residuals from latent process.</td>
<td>185</td>
</tr>
<tr>
<td>10.8 Autocorrelation functions of residuals for the four categories.</td>
<td>186</td>
</tr>
<tr>
<td>10.9 Residuals against lag 1 residuals for the four categories.</td>
<td>187</td>
</tr>
<tr>
<td>10.10 Residuals against lag 1 residuals for the latent process.</td>
<td>188</td>
</tr>
<tr>
<td>10.11 Residuals for the four categories against the lag 1 smoother.</td>
<td>189</td>
</tr>
<tr>
<td>10.12 The estimated latent process from Kalman smoother.</td>
<td>192</td>
</tr>
<tr>
<td>10.13 Day-of-the-week effects.</td>
<td>192</td>
</tr>
</tbody>
</table>
Acknowledgements

I thank my supervisor, Dr. Bent Jørgensen, who has been a constant source of inspiration, encouragement and guidance for me throughout my graduate career at UBC.

I would like to acknowledge the assistance and valuable comments from other members of my supervising committee, Dr. James Zidek and Dr. Martin Puterman.

Many thanks go to Dr. Harry Joe for introducing me to his research work that was both inspiring and stimulating.

I acknowledge my debt to Dr. Søren Lundbye-Christensen and Dr. Li Sun, without whom this research would not have been possible.

I would also like to thank Dr. Ruben Zamar, Dr. John Petkau, Dr. Jian Liu, Dr. Nancy Heckman and Dr. Paul Gustafson for their sincere support and encouragements during my Ph.D program at UBC.

I thank my wife Ru Yan for her support and patience. My daughter Angela Song came and brought uncountable happiness to me during my stay at UBC.

Finally I would like to thank the Department of Statistics, UBC, as well as my supervisor Dr. Bent Jørgensen for their financial support to me in the past four years.
Chapter 1

Introduction

1.1 General

Statistical studies of longitudinal data are commonly undertaken in a number of practical areas and for a variety of purposes. In the most general sense, "longitudinal data" means any collection of data indexed by time. Following established terminology in biostatistical applications, we shall usually think of longitudinal data as a collection of time-series, one for each of a number of subjects. The main characteristic of longitudinal data is thus the presence of repeated observations for each subject, the observations being correlated over time. Of primary interest in the analysis of longitudinal data are the mechanisms of change, for instance growth, aging, time profiles or the effects of covariates (explanatory variates). See Diggle et al. (1994) or Lindsey (1993) for a general introduction to longitudinal data analysis.

Longitudinal data sets usually consist of either many short time-series or a few long time-series; each series may be either stationary or non-stationary. Classical time-series analysis (in the Box-Jenkins sense, say) corresponds essentially to the special case of one long stationary time-series. See Box and Jenkins (1970) and Brockwell and Davis (1987). Furthermore, the response at a given time may be either univariate or multivariate, and there may be explanatory variates, either constant or time-varying.

Depending on the particular type of longitudinal data at hand, any of several different types of statistical analysis may be appropriate. Some methods for longitudinal
data analysis are adapted from classical time-series analysis, while others such as those based on generalized linear models originate in regression analysis. In particular, state-space models (or dynamical models) have developed in parallel development for both classical time-series analysis and longitudinal data analysis (c.f. Fahrmeir and Tutz, 1994), whereas the generalized estimating equation method of Liang and Zeger (1986) was mainly inspired by ideas from generalized linear models.

A key distinction in longitudinal data analysis is between long and short time-series. For long time-series, stationarity is crucial, and assuming stationarity for non-stationary data, or *vice versa*, may invalidate the analysis. For short time-series, this question is less important, but if the number of subjects is large enough one should still be able to determine the nature of the correlation structure accurately. There is no hard and fast rule for determining when a series is considered long or short, and the choice of method may also depend on other aspects of the data at hand.

In this thesis, we concentrate on three particular types of longitudinal data, which together cover a broad range of data-types and applications. The first type consists of many short time-series, where methods from multivariate data analysis may be employed. The second type involves one or a few long stationary time-series. The third consists of one or a few long non-stationary time-series with multivariate responses. In the first and third cases, we assume that the series are accompanied by time-varying covariates.

For stationary time-series, we can take advantage of the already extensive arsenal of techniques available from classical time-series analysis. Regression analysis for long non-stationary time-series, on the other hand, presents new unsolved problems, the main one being that arising from the distinction between correlation, trend and regression effects for the data.

Understanding the nature of both correlation and trend of the data is the main concern in modelling longitudinal data. However, the marginal model (Diggle et al., 1994),
Currently one of the most popular methods for the analysis of short time series, deals only with the pattern of trend, treating the correlation as a set of nuisance parameters. This model is unsatisfactory in practice when the pattern of correlation is of interest in the analysis. The goal of this thesis is to develop statistical models which can accommodate features of both trend and correlation for longitudinal data.

Corresponding to the three classes of longitudinal data introduced above, this thesis concentrates on three topics: (1) multivariate dispersion models with application to regression analysis under the population-averaged model (Zeger et al., 1988); (2) stationary time series models with exponential dispersion model margins motivated by Wölfer's monthly mean relative sunspot numbers (Yule, 1927); (3) state space models with either stationary or non-stationary Markov latent processes.

1.2 Literature review

Due to the relative paucity of multivariate non-normal distribution families, the analysis of short time series data is in general difficult. For the case of normal data, the theory for longitudinal studies is quite well developed. Lindsey (1993) gives a good summary about such developments and gives further references. Lindsey (1993) also described recent developments in the analysis of longitudinal categorical data, and later Diggle, Liang and Zeger (1994) presented a comprehensive introduction to the theory of longitudinal data analysis based on generalized linear models. The central method in the book of Diggle et al. is what they call the generalized estimating equation (GEE) approach, which is developed by incorporating a so-called "working" correlation matrix into a Fisher scoring equation obtained with independent observations (see also Liang and Zeger, 1986). The correlation matrix is then treated as a set of nuisance parameters in their quasi-likelihood estimation procedure. Fitzmaurice et al. (1993) made an important comment
on this quasi-likelihood approach, arguing that when the experimental design includes
time-varying covariates, it becomes very important to obtain a close approximation to
the correlation structure in order to achieve high efficiency. Zhao and Prentice (1990)
suggest estimating the regression parameters and correlation parameters simultaneously
to improve the efficiency of the GEE approach.

Attempts at likelihood-based estimation for longitudinal data have been seen in the
literature in the past several years (c.f. Laird, 1990 for the general multivariate model
approach to binary outcomes and Joe, 1996b for the copula approach to categorical data).
These attempts give rise to the development of multivariate non-normal distribution fami-
ilies that can be applied in the longitudinal data setting. McCullagh and Nelder (1989,
Chapter 6) and Lipsitz et al. (1990) developed some general multivariate models for
discrete multivariate data for solving certain problems. Some recent models for associ-
ation in discrete data are available (c.f. Carey et al. 1993 and Fitzmaurice et al. 1993).
Other recent work includes Leroux and Puterman (1992), Jørgensen (1992), Jørgensen

The problem of generalizing the classical time series models of Box and Jenkins (1970)
(see also Brockwell and Davis, 1987) to analyze non-normal stationary time series data
has received considerable attention in the past two decades or so, starting with the papers
Lawrance and Lewis (1977) and Jacobs and Lewis (1977) for the case of exponential
margins. Models for other marginal distributions such as gamma, Poisson and negative
binomial followed. A unification and generalization of some of these models were recently
proposed by Joe (1996a), further detailed references are available therein.

To analyze time series data with time-varying covariates, state space models (or dy-
namical models) are very useful. These models offer a convenient conceptual framework
for time series data, where the latent process may be interpreted as an unobserved inten-
sity process, often providing a useful interpretation of the data-generating mechanism.

1.3 Overview of topics

We begin by reviewing (in Chapter 2) the theory of dispersion models, referring to the forthcoming monograph of Jørgensen (1996), including dispersion models, exponential dispersion models, proper dispersion models and Tweedie models. We then study the thinning operation of Joe (1996a) under exponential dispersion models, the key to constructing the stationary time series models introduced in Chapter 5. As a natural multivariate extension of dispersion models, the multivariate dispersion densities of Jørgensen and Lauritzen (1993) are introduced in Section 2.6 where some special cases are exhibited.

Chapters 3 and 4 form the first part of the thesis and deal with short time series data. In Chapter 3, we apply the copula approach (c.f. Joe, 1996b) to form a family of multivariate dispersion models whose dependence structures are inherited from the multivariate normal distribution in a certain way. We introduce a new dependence measure
similar to Kendall's \( \tau \) and Spearman's \( \rho \), the so-called normal scoring \( \nu \), to explain the dependence structure of a copula. The normal copula approach is very useful for dealing with non-normal multivariate data, including integer data, directional data and positive skew data. Compared with existing results, ours provides a more flexible approach to analysis of non-normal longitudinal data.

Chapter 4 presents a likelihood-based approach to the regression analysis of short time series data under the population-averaged model (Zeger et al., 1988) in which the aggregate response for the population is the focus, and each short series forms a response vector assumed to follow a multivariate dispersion model. A theory of approximate inference is proposed for this regression model through saddlepoint approximations and tail area approximations for the joint likelihood function. As a result, we derive several estimating equations for regression parameters, which include the generalized estimating equations of Liang and Zeger (1986) as a special case. Since our approach is based on a fully specified probability model, this sheds new light on the generalized estimating equation approach, and suggests cases where a likelihood-based approach is preferred.

Chapter 5 focusses on the second topic of the thesis: the construction of stationary time series models with exponential dispersion model margins, motivated by Wölfer's monthly mean relative sunspot data (Yule, 1927). The analysis of stationary non-normal long time series data gives rise to the problem of developing stationary time-series models with non-normal marginal distributions. We introduce a class of stationary infinite order moving average processes with margins in the class of infinitely divisible exponential dispersion models. Our main tool for constructing these processes is the thinning operation of Joe (1996a). As a special case we obtain a class of autoregressive moving average processes with exponential dispersion model margins, denoted EDARMA\((p,q)\), which are different from Joe's (1996a) ARMA models.

Our construction is based on first defining an infinite-order moving average model,
and then defining the EDARMA\((p, q)\) model as a special case. Although this approach is less direct than that of Joe (1996a), it is mathematically more convenient, and allows us to transplant many techniques from linear time-series analysis into our framework. In addition, the class of stationary processes could be used to model latent processes in state space models presented in Chapter 7.

The third topic of this thesis is dealt with in Chapters 6 to 10, where we develop state space models for long time series data with time-varying covariates. State space models may have either stationary latent processes or non-stationary latent processes. Most previous work has focussed either on models with stationary latent process, such as Zeger (1988) and Chan and Ledolter (1995), or on linear Gaussian state space models.

Our approach is based on either a stationary latent process or a non-stationary latent process, and uses distributions from the Tweedie class, providing a very flexible approach to the analysis of non-normal time series data.

Following Jørgensen, Labouriau and Lundbye-Christensen (1994), we base our estimation procedure on what we call the Kalman estimating equation (KEE), corresponding to the EM algorithm with the E-step approximated by the Kalman smoother. We also study residual analysis for model checking in detail.

In Chapter 6, we introduce the BLUP and the Kalman filter and smoother, which are crucial for the analysis of state space models in the subsequent chapters.

Chapter 7 presents a class of state space models for long time series data, assuming that the latent processes are given by the class of stationary AR(1) models of Joe (1996a) with non-normal margins and without time-varying covariates. These stationary latent processes are constructed by the thinning operation of Joe (1996a), which leads to a structure different from the linear structure of classical state space models, so standard results for the Kalman filter and smoother do not apply to our models. Instead we use the results of Chapter 6. We also give a complete proof of the asymptotic normality for
Chapter 1. Introduction

the KEE estimates for these models.

Chapter 7 also gives illustrative applications of state space models with a stationary AR(1) gamma process to the analysis of the polio data (Zeger, 1988). The polio data consist of monthly number of poliomyelitis cases in the USA from 1970 to 1993, which were previously analyzed by Zeger (1988) and Chan and Ledolter (1995). We re-analyze the data to study whether or not there is an evidence of a decreasing trend in the rate of polio infection in the USA. Our KEE estimates are closer to Chan and Ledolter's MCEM estimates than Zeger's quasi-likelihood estimates. Both Chan and Ledolter’s method and ours lead to a significantly decreasing trend, whereas Zeger’s method leads to a non-significant trend. We conduct a detailed model checking diagnosis by means of residual analysis to confirm the adequacy of our model assumptions to the data. Neither of the papers cited above by Chan and Ledolter nor Zeger present such a model checking procedure in their analyses.

Chapter 8 proposes a class of state space models for multivariate response vectors with non-stationary latent processes where covariates may enter in both the observed and latent processes. The models open up some new possibilities for detailed investigation of the time-varying structure of long time series data, allowing us to address substantive questions that are not easily dealt with by existing methods. The marginal distributions for the components of the response vector are allowed to be different. Hence our theory provides a very realistic model for many types of multivariate longitudinal data.

Chapter 9 conducts two simulation studies under Poisson-gamma models (Poisson response and gamma latent process): (1) comparison of the efficiency of the Kalman estimating equation approach and the Monte Carlo EM algorithm (MCEM); (2) examination of the utility of residual analysis for detecting of stationarity or non-stationarity for the latent process. The simulation experiments allow us to conclude that the estimates from the KEE method are very similar to those from the MCEM approach. We also
found that the residual analysis is able to effectively distinguish between the stationary and non-stationary cases.

Chapter 10 presents an application of the state space models with non-stationary latent process proposed in Chapter 8 to the analysis of data from Prince George, British Columbia, previously analyzed by Knight et al. (1993). The data consist of daily counts of emergency room visits each of four categories of respiratory diseases, along with covariates of air pollution variables and meteorological variables. These data are analyzed through state space models to investigate the relationship between air pollution and respiratory morbidity. Because our model includes a more realistic correlation structure compared with Knight et al.'s (1993) analysis, we arrive at a much more parsimonious model for the systematic part. The interpretation of the latent process in connection with these data provide new insight into the relation between air pollution and respiratory morbidity.

Finally in Chapter 11 we make some general comments on possible further studies for some of the topics of the thesis.
Chapter 2

Dispersion models

In this chapter, we first give a brief introduction to the theory of univariate dispersion models of Jørgensen (1987a, 1996), which is the basis of the entire theory that we develop in this thesis. We then discuss in detail the thinning operation of Joe (1996a) for the class of exponential dispersion models, and finally discuss a possible multivariate extension of dispersion models.

2.1 Definition

One attempt to generalize the normal distribution, linked to generalized linear models, leads to the class of dispersion models whose statistical properties are very similar to those of the normal distribution. Dispersion models were originally introduced by Sweeting (1981) and Jørgensen (1983) and the models were studied in detail by in Jørgensen (1986, 1987a).

A univariate (regular) dispersion model $\text{DM}(\mu, \sigma^2)$ with position parameter $\mu \in \Omega$ and dispersion parameter $\sigma^2 > 0$ is defined by a probability density function of form

$$p(y; \mu, \sigma^2) = a(y; \sigma^2) \exp \left\{ -\frac{1}{2\sigma^2}d(y; \mu) \right\}, y \in C$$  \hspace{1cm} (2.1.1)

where $a \geq 0$ is a suitable function, $d : C \times \Omega \rightarrow \mathcal{R}$ is a regular unit deviance satisfying

$$d(y; \mu) \geq d(y; y) = 0 \quad \text{and} \quad \frac{\partial^2 d(y; \mu)}{\partial \mu^2} \bigg|_{\mu=y} = \frac{2}{V(y)} > 0,$$

$V$ being called the variance function.
Chapter 2. Dispersion models

Obviously, the normal distribution \( N(\mu, \sigma^2) \) is a dispersion model with unit deviance \((y - \mu)^2\) and variance function \( V(\mu) = 1 \). Other dispersion models have many analogies with the normal distribution, particularly because \( d(y; \mu) \) behaves like a generalized measure of squared distance for small \( \sigma^2 \). Other characteristics of the densities are:

(a) The factor \( a \) in (2.1.1) is independent of the position parameter \( \mu \).

(b) The densities are often unimodal; in general, the deviance \( d \) tends to make the density peak near \( \mu \), and the peak will become higher and sharper the smaller the value of \( \sigma^2 \).

(c) The dispersion models cover a comprehensive range of non-normal distributions such as Poisson, gamma, negative binomial, binomial, inverse Gaussian, von Mises and simplex distributions.

Dispersion models are in general approximately normal \( N \{ \mu, \sigma^2V(\mu) \} \) for \( \sigma^2 \) small, so that \( \mu \) and \( \sigma^2V(\mu) \) represent the asymptotic mean and variance, respectively. According to Jørgensen (1996), as the dispersion parameter \( \sigma^2 \) tends to zero, the density of a regular dispersion model \( DM(\mu, \sigma^2) \) can be approximated as follows,

\[
p(y; \mu, \sigma^2) \simeq \left(2\pi\sigma^2\right)^{-1/2} \frac{1}{V^{-1/2}(y)} \exp \left\{ -\frac{1}{2\sigma^2} d(y; \mu) \right\}, \quad y \in \Omega. \tag{2.1.2}
\]

This asymptotic relation (2.1.2) is usually called the saddlepoint approximation for a regular dispersion model.

We use the convention that throughout the rest of this thesis, the symbol \( \simeq \) in (2.1.2) will always used for asymptotic relationships as the dispersion parameter tends to zero.
2.2 Exponential dispersion models

Exponential dispersion models were originally proposed by Tweedie (1947), and later reintroduced by Nelder and Wedderburn (1972) as the class of error distributions for generalized linear models.

A **reproductive exponential dispersion model** is defined by the following density for a random variable \( Y \), with respect to a suitable measure,

\[
p(y; \theta, \lambda) = c(y; \lambda) \exp \{\lambda \{\theta y - \kappa(\theta)\}\}.
\]  

(2.2.1)

An **additive exponential dispersion model** for a random variable \( Z \) is defined by the probability density function of the following form, with respective to a suitable measure,

\[
p(z; \theta, \lambda) = c^*(z; \lambda) \exp \{\theta z - \lambda \kappa(\theta)\}.
\]  

(2.2.2)

The two types are connected by the duality transformation \( Z = \lambda Y \), so in principle, any exponential dispersion model has both an additive and a reproductive forms. The parameter \( \theta \) is called the **canonical parameter** and \( \lambda \) the **index parameter**. The maximum possible domain for \( \theta \), denoted \( \Theta \subseteq \mathcal{R} \) is called the **canonical parameter domain**. The domain for \( \lambda \), called the **index set**, denoted \( \Lambda \), is a subset of \( \mathcal{R}_+ \). The function \( \kappa \) is called the **cumulant function**.

The reproductive form (2.2.1) is denoted \( Y \sim \text{ED}(\mu, \sigma^2) \), where \( \mu = \tau(\theta) = \kappa'(\theta) \) is the mean and \( \sigma^2 = \lambda^{-1} \) is the dispersion parameter. The variance of \( Y \) is \( \sigma^2 V(\mu) = \sigma^2 \tau'(\tau^{-1}(\mu)) \).

The additive form (2.2.2) is denoted by \( Z \sim \text{ED}^*(\theta, \lambda) \), and the mean and variance of \( Z \) are

\[
E(Z) = \xi = \lambda \tau(\theta) \quad \text{and} \quad \text{Var}(Z) = \lambda V(\xi/\lambda) = \lambda V(\mu),
\]  

(2.2.3)

where \( \tau(\theta) = \kappa'(\theta) \) and \( V(\mu) = \tau'(\tau^{-1}(\mu)) \).
Chapter 2. Dispersion models

We define the unit deviance for the above exponential dispersion models as follows,

\[ d(y; \mu) = 2 \left[ \sup_{\theta \in \Theta} \{y\theta - \kappa(\theta)\} - y\tau^{-1}(\mu) + \kappa \left\{ \tau^{-1}(\mu) \right\} \right]. \]  \hspace{1cm} (2.2.4)

Then the density (2.2.1) can be re-written by the form (2.1.1) in terms of \( d \) of (2.2.4) and the \( a \) function given by

\[ a(y; \sigma^2) = c(y; \sigma^{-2}) \exp \left[ \sup_{\theta \in \Theta} \{y\theta - \kappa(\theta)\} \right]. \]

This proves that the reproductive model (2.2.1) is indeed a dispersion model (2.1.1). However, the additive model (2.2.2) is not, in general, of the form (2.1.1).

Both types of exponential dispersion models satisfy a convolution property given as follows. Let \( Y_1, \ldots, Y_n \) be independent and \( Y_i \sim ED(\mu, \sigma^2/w_i) \) with weights \( w_i \). Then,

\[ \frac{1}{w_+} \sum_{i=1}^{n} w_i Y_i \sim ED(\mu, \sigma^2/w_+) \] \hspace{1cm} (2.2.5)

where \( w_+ = \sum_{i=1}^{n} w_i \).

If \( Z_1, \ldots, Z_n \) are independent and \( Z_i \sim ED^*(\theta_i, \lambda_i) \), then

\[ Z_+ = \sum_{i=1}^{n} Z_i \sim ED^*(\theta, \lambda_1 + \cdots + \lambda_n). \] \hspace{1cm} (2.2.6)

Furthermore, we have \( \Lambda = \mathcal{R}_+ \) if and only if (2.2.2) is infinitely divisible (Jørgensen, 1986).

The saddlepoint approximation to the density function of the additive model (2.2.2) is given by

\[ \{2\pi \lambda V(z/\lambda)\}^{-1/2} \exp \left[ -z\tau^{-1}(z/\lambda) + \lambda\kappa \left\{ \tau^{-1}(z/\lambda) \right\} \right] \exp \{\theta z - \lambda \kappa(\theta)\} \] \hspace{1cm} (2.2.7)

as \( \lambda \) goes to \( \infty \).
Table 2.1: Summary of Tweedie exponential dispersion models.

<table>
<thead>
<tr>
<th>Distributions</th>
<th>$p$</th>
<th>$S$</th>
<th>$\Omega$</th>
<th>$\Theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extreme stable</td>
<td>$p &lt; 0$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{R}_+$</td>
<td>$\mathbb{R}_0$</td>
</tr>
<tr>
<td>Normal</td>
<td>$p = 0$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{R}$</td>
</tr>
<tr>
<td>[Do not exist]</td>
<td>$0 &lt; p &lt; 1$</td>
<td>$\mathbb{R}_0$</td>
<td>$\mathbb{R}_+$</td>
<td>$\mathbb{R}_0$</td>
</tr>
<tr>
<td>Poisson</td>
<td>$p = 1$</td>
<td>$\mathbb{N}_0$</td>
<td>$\mathbb{R}_+$</td>
<td>$\mathbb{R}$</td>
</tr>
<tr>
<td>Compound Poisson</td>
<td>$1 &lt; p &lt; 2$</td>
<td>$\mathbb{R}_0$</td>
<td>$\mathbb{R}_+$</td>
<td>$\mathbb{R}_-$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$p = 2$</td>
<td>$\mathbb{R}_+$</td>
<td>$\mathbb{R}_+$</td>
<td>$\mathbb{R}_-$</td>
</tr>
<tr>
<td>Positive stable</td>
<td>$2 &lt; p &lt; 3$</td>
<td>$\mathbb{R}_+$</td>
<td>$\mathbb{R}_+$</td>
<td>$-\mathbb{R}_0$</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>$p = 3$</td>
<td>$\mathbb{R}_+$</td>
<td>$\mathbb{R}_+$</td>
<td>$-\mathbb{R}_0$</td>
</tr>
<tr>
<td>Positive stable</td>
<td>$p &gt; 3$</td>
<td>$\mathbb{R}_+$</td>
<td>$\mathbb{R}_+$</td>
<td>$-\mathbb{R}_0$</td>
</tr>
<tr>
<td>Extreme stable</td>
<td>$p = \infty$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{R}_-$</td>
</tr>
</tbody>
</table>

Notation: $-\mathbb{R}_0 = (-\infty, 0]$.

2.3 Tweedie class

The Tweedie class is an important class of univariate exponential dispersion models, defined as the special case of (2.2.1) with unit variance functions of the form

$$V(\mu) = \mu^p,$$

for some $p \in (-\infty, 0] \cup [1, \infty)$, denoted $\text{T}_p(\mu, \sigma^2)$, where the variance is $\sigma^2 \mu^p$. Table 2.1 shows a summary of the different types of Tweedie models. The Tweedie models may be characterized as the only exponential dispersion models that satisfy the following scale transformation property,

$$c \: \text{T}_p(\mu, \sigma^2) \sim \text{T}_p(c\mu, c^{2-p} \sigma^2), \quad \forall c > 0. \tag{2.3.1}$$

Hence, each Tweedie model also has an additive form.

The density function of the Tweedie distribution $\text{T}_p(\mu, \sigma^2)$, considered as an additive model, has the following form, with $\beta = (p - 2)/(p - 1)$,

$$f(z; \theta, \lambda, \beta) = c_\beta^*(\lambda; z) \exp \{\theta z - \lambda \kappa_\beta(\theta)\} \tag{2.3.2}$$
where $\kappa_\beta(\theta) = \beta^{-1}(\beta - 1)\{\theta/(\beta - 1)\}^\beta$ and $\tau_\beta(\theta) = \kappa'_\beta(\theta)$. The function $c^*_\beta(\lambda; z)$ is defined as follows,

$$c^*_\beta(\lambda; z) = \begin{cases} 
1 & \text{for } \beta < 0, \ z = 0 \\
\sum_{k=1}^{\infty} \frac{\lambda^k \kappa^k_\beta(-1/z)}{z \Gamma(-k \beta) k!} & \text{for } \beta < 0, \ z > 0 \\
\sum_{k=1}^{\infty} \frac{\Gamma(1 + k \beta)}{\pi z k!} \lambda^k \kappa^k_\beta(-1/z) \sin(-k \pi \beta) & \text{for } 0 < \beta < 1, \ z > 0 \\
\sum_{k=1}^{\infty} \frac{\Gamma(1 + k / \beta)}{\pi z k!} \left\{ \frac{-z}{\lambda \kappa^k_\beta(1)} \right\}^k \sin(-k \pi / \beta) & \text{for } 1 < \beta < 2, \ z \in \mathcal{R}.
\end{cases}$$

(2.3.3)

Many exponential dispersion models have variance functions that are asymptotically of the Tweedie form, leading to a general convergence theorem with the Tweedie models as limiting distributions (Jørgensen, Martínez and Tsao, 1994). For this reason, Tweedie models occupy a central position among exponential dispersion models.

### 2.4 Proper dispersion models

By normalizing the saddlepoint approximation (2.1.2) for a regular dispersion model (2.1.1), we obtain a norming constant $a_0(\mu, \sigma^2)$ that makes the right-hand side of (2.1.2) a density function with respect to Lebesgue measure. Considering the special case where the norming constant $a_0(\mu, \sigma^2) = a(\sigma^2)$ does not depend on $\mu$, we may write (2.1.2) as follows,

$$p(y; \mu, \sigma^2) = a(\sigma^2) V^{-1/2}(y) \exp \left\{ -\frac{1}{2\sigma^2} d(y; \mu) \right\}, \ y \in C.$$

(2.4.1)

This is called a proper dispersion model. Exactly three exponential dispersion models are also proper dispersion models, namely the normal, gamma and inverse Gaussian families. Other examples of proper dispersion models are the von Mises distribution and the simplex distribution of Barndorff-Nielsen and Jørgensen (1991). For more details, refer to Jørgensen (1996).
2.5 Thinning operation

We now describe the thinning operation of Joe (1996a), which is crucial for developing both the stationary ARMA models with exponential dispersion model margins in Chapter 5 and the state space models with stationary Markov latent processes in Chapter 7. The thinning operation may be interpreted as a kind of inverse of convolution. For example, whereas convolution increases the value of $\lambda$ (c.f. (2.2.6)), we show that thinning decreases the value of $\lambda$. For the convenience of convolution expression, we adopt the additive form rather than reproductive exponential dispersion model for this study. Suppose $X_1$ and $X_2$ are independent with

$$X_i \sim \text{ED}^*(\theta_i, \lambda_i), \quad i = 1, 2. \quad (2.5.1)$$

Let $W = X_1 + X_2$ denote their sum. By the property of exponential family, it is easy to show that $W$ is sufficient for the parameter $\theta$. So, the conditional distribution of $X_1$ given $W = x$ does not depend on the parameter $\theta$, and is denoted

$$X_1 | X_1 + X_2 = x \sim G(\lambda_1, \lambda_2, x).$$

The conditional density function corresponding to $G(\lambda_1, \lambda_2, x)$ is given by

$$p(x_1 | x; \lambda_1, \lambda_2) = \frac{c^*(x_1; \lambda_1)c^*(x - x_1; \lambda_2)}{c^*(x; \lambda_1 + \lambda_2)} \quad (2.5.2)$$

where $c^*$ is given in the density (2.2.2). We call the distribution $G(\lambda_1, \lambda_2, x)$ the contraction corresponding to $\text{ED}^*(\theta, \lambda)$ where $\lambda = \lambda_1 + \lambda_2$. This terminology comes from the fact that if the model $\text{ED}^*(\theta, \lambda)$ has support containing only non-negative values (excluding for example the normal density), then $G(\lambda_1, \lambda_2, x)$ is concentrated on the interval from 0 to $x$.

We now define the thinning operation $A(\cdot; \alpha)$, where $\alpha \in [0, 1]$. Let $X$ be a random variable with distribution $\text{ED}^*(\theta, \lambda)$, an infinitely divisible exponential dispersion model,
and let $\bar{\alpha} = 1 - \alpha$. Define the random variable $A(X; \alpha)$ by the conditional distribution

$$A(X; \alpha)|X = x \sim G(\alpha \lambda, \bar{\alpha} \lambda, x).$$

By Proposition 2.1 below, the marginal distribution of $A(X; \alpha)$ is then $ED^*(\theta, \alpha \lambda)$, and we say that $A(X; \alpha)$ is the thinning of $X$ by the proportion $\alpha$. The assumption of infinite divisibility is made to ensure that both $\alpha \lambda$ and $\bar{\alpha} \lambda$ are valid values of the index parameter for any value of $\lambda$ in $(0, 1)$. In the extreme cases $\alpha = 0, 1$, we define $A(X; 1) = X$, and $A(X; 0) = 0$.

The proof of the thinning operation hinges on the next proposition, which is easily proved using density functions when they exist. However, for completeness we include a proof that does not depend on the existence of densities.

**Proposition 2.1** (Joe, 1996a) Let $ED^*(\theta, \lambda)$ be an exponential dispersion model with index set $\Lambda$, and suppose that $X \sim ED^*(\theta, \lambda)$ and $Y|X = x \sim G(\lambda_1, \lambda_2, x)$, where $\lambda, \lambda_1, \lambda_2 \in \Lambda$ and $\lambda_1 + \lambda_2 = \lambda$. Then $Y$ and $X - Y$ are independent, $Y \sim ED^*(\theta, \lambda_1)$ and $X - Y \sim ED^*(\theta, \lambda_2)$.

**Proof** Suppose that $X_1$ and $X_2$ are independent with distributions (2.5.1), and let $W = X_1 + X_2$. Then the conditional distribution of $X_1|W$ and the marginal distribution of $W$ are the same as respectively the conditional distribution of $Y|X$ and the marginal distribution of $X$. Hence the joint distribution of $(X_1, W)$ is the same as the joint distribution of $(Y, X)$, which in turn implies that the joint distribution of $(X_1, X_2)$ is the same as the joint distribution of $(Y, X - Y)$, completing the proof. 

The following moment property, due to Joe (1996a),

$$E \{ A(X; \alpha)|X = x \} = \alpha x,$$

and is important for many calculations. Hence, the thinning operation decreases the value of $\lambda$, one may say that it decreases the expectation, as evidenced by (2.5.3).
Table 2.2: Summary of contractions.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$ED^*(\theta, \lambda)$</th>
<th>$G(\alpha \lambda, \tilde{\lambda}, x)$</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>$N(\theta \lambda, \lambda)$</td>
<td>$N(\alpha x, \alpha \lambda)$</td>
<td>Joe (1996a)</td>
</tr>
<tr>
<td>Poisson</td>
<td>$Po(\lambda e^\theta)$</td>
<td>$Bi(x, \alpha)$</td>
<td>Al-Osh and Alzaid (1987)</td>
</tr>
<tr>
<td>Gamma</td>
<td>$Ga^*(\theta, \lambda)$</td>
<td>$xBeta(\alpha \lambda, \tilde{\lambda})$</td>
<td>Lewis (1983)</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>$IG^*(\theta, \lambda)$</td>
<td>Simplex</td>
<td>Barndorff-Nielsen and Jørgensen (1991)</td>
</tr>
<tr>
<td>Negative binomial</td>
<td>$Nb^*(\theta, \lambda)$</td>
<td>$Beta-Bi(\alpha \lambda, \tilde{\lambda}, x)$</td>
<td>Joe (1996a)</td>
</tr>
<tr>
<td>Generalized Poisson</td>
<td>$GP^*(\theta, \lambda)$</td>
<td>Quasi-binomial</td>
<td>Consul (1989 p. 195)</td>
</tr>
<tr>
<td>Tweedie</td>
<td>$Tw^*_p(\theta, \lambda)$</td>
<td>$GTw_p(\alpha \lambda, \tilde{\lambda}, x)$</td>
<td>Jørgensen and Song (1995)</td>
</tr>
</tbody>
</table>

Table 2.2 presents a summary of the contractions for various exponential dispersion models, including the six examples considered by Joe (1996a). Except for the normal and Poisson distributions, the notation follows that used in equation (2.2.2). Details about the various contractions may be found either in the references mentioned, or in Joe (1996a).

The contraction is particularly simple in the case of a normal or gamma margins. In the normal case with $X \sim N(\theta \lambda, \lambda)$ we may write $A(X; \alpha) = \alpha X + X_0$, where $X_0 \sim N(0, \alpha \tilde{\lambda})$ is independent of $X$. In the gamma case we have $A(X; \alpha) = GX$ where $G \sim Beta(\alpha \lambda, \tilde{\lambda})$ is independent of $X$.

Let

$$g(x; \lambda) = -x\tau^{-1}(x/\lambda) + \lambda \kappa \left\{ \tau^{-1}(x/\lambda) \right\}. \quad (2.5.4)$$

Inserting this in (2.2.7) we obtain an asymptotic approximation for the density of the contraction $G(\alpha \lambda, \tilde{\lambda}, x)$ namely

$$p(w|x; \alpha \lambda, \tilde{\lambda}) \approx \frac{V(x/\lambda)}{2\pi \alpha \tilde{\lambda} V\left\{ w/(\alpha \lambda) \right\} V\left\{ (x-w)/(\tilde{\lambda}) \right\}} e^{g(x;\lambda)-g(w;\alpha \lambda)-g(x-w;\tilde{\lambda})}$$

where $V(\mu)$ is the unit variance function $\tau'\{\tau^{-1}(\mu)\}$. This approximation gives the exact result in two cases, the normal and inverse Gaussian distributions. In several other cases, such as the Poisson and gamma cases, it amounts to replacing all factorials and gamma functions by Stirling's approximation.
2.6 Multivariate dispersion model densities

A simple multivariate extension for the univariate dispersion models might be defined similarly to (2.1.1), except that now $y$ and $\mu$ are $p$-dimensional vectors but $\sigma^2$ is still one-dimensional. Examples of this type of multivariate dispersion models include the multivariate von Mises-Fisher distribution, the multivariate hyperboloid distribution and some special cases of the multivariate simplex distribution of Barndorff-Nielsen and Jørgensen (1991). The inference for this type of multivariate models follows the same pattern as that for univariate models, see Jørgensen (1987a). It is obvious that this multivariate extension oversimplifies the covariance structure because only a single parameter $\sigma^2$ is employed for handling the correlation structure.

Jørgensen and Lauritzen (1993) proposed a different multivariate extension of the dispersion model density, following the spirit of the multivariate normal density.

Consider a $p$-variate multivariate normal distribution, $\mathcal{N}_p(\mu, \Sigma)$ with mean $\mu$ and covariance matrix $\Sigma$. Its density function is of the form

$$p(y; \mu, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu) \right\}, \; y \in \mathbb{R}^p. \quad (2.6.1)$$

Equivalently it may be written in terms of the notations of dispersion models as follows,

$$p(y; \mu, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} \delta(y; \mu) \Sigma^{-1} \delta(y; \mu) \right\} \quad (2.6.2)$$

where $\delta(y; \mu) = (\delta_1, \ldots, \delta_p)^T$ and $\delta_i = \delta(y_i; \mu_i) = d_i^1(y_i; \mu_i) \text{sgn}(y_i - \mu_i) = y_i - \mu_i$.

If we generalize the expression (2.6.2) to allow the $\delta_i$ (the square-root of the unit deviance $d_i$) to be taken from the class of dispersion models, in principal we could define a family of multivariate dispersion model densities as follows,

$$p(y; \mu, \Sigma) = A(y; \Sigma) \exp \left\{ -\frac{1}{2} \delta(y; \mu) \Sigma^{-1} \delta(y; \mu) \right\}, \; y \in C \quad (2.6.3)$$
where $A \geq 0$ is a suitable function, $\Sigma$ is a positive definite matrix, $\delta = (\delta_1, \ldots, \delta_p)^T$ is the vector with components $\pm d^{1/2}(y_i; \mu_i)$ and $C$ is the convex support for $Y$.

The multivariate representation is simplified when $\Sigma$ is a diagonal matrix.

**Proposition 2.2.** Suppose $Y$ has a density of form (2.6.3). If the dispersion matrix $\Sigma$ is diagonal, then the components of $Y$ are independent and $Y_i \sim DM(\mu_i, \sigma_{ii})$ where $\sigma_{ii}$ is the $i$-th diagonal element of the matrix $\Sigma$.

An immediate example of this definition is the multivariate location model defined by the density

$$p(y; \mu, \Sigma) = A(\Sigma) \exp \left\{-\frac{1}{2} \delta^T (y - \mu) \Sigma^{-1} \delta (y - \mu) \right\}$$

where

$$A^{-1}(\Sigma) = \int_{\mathbb{R}^p} \exp \left\{-\frac{1}{2} \delta^T (z) \Sigma^{-1} \delta (z) \right\} dz, \quad \delta(y - \mu) = (\pm \delta(y_1 - \mu_1), \ldots, \pm \delta(y_p - \mu_p))^T,$$

$\pm$ is taken by the sign of $y_i - \mu_i$ and $\delta$ is a non-negative function defined on $\mathcal{R}$ satisfying $\delta(0) = 0$.

If the function $A$ in (2.6.3) factorizes as of form

$$A(y; \Sigma) = C(\Sigma)B(y), \quad \text{for all } y$$

for suitable function $C$ and $B$, we call (2.6.3) a multivariate proper dispersion model.

Obviously, the multivariate normal distribution is a multivariate proper dispersion model. Other examples are given as follows.

**Example 2.1 (Multivariate gamma density)** The univariate gamma distribution, being a proper dispersion model, has unit deviance function

$$d(y; \mu) = 2 \left( \frac{y}{\mu} - \log \frac{y}{\mu} - 1 \right).$$
Now define a multivariate gamma distribution with position parameter vector $\mu$ and dispersion matrix $\Sigma$ by the joint density function

$$p(y; \mu, \Sigma) = A(y; \Sigma) \exp \left\{ -\frac{1}{2} \delta^T (y; \mu) \Sigma^{-1} \delta(y; \mu) \right\}$$

with $\delta(y; \mu) = (\delta_1, \ldots, \delta_p)^T$ where

$$\delta_i = \delta(y_i; \mu_i) = \left\{ 2 \left( \frac{y_i}{\mu_i} - \log \frac{y_i}{\mu_i} - 1 \right) \right\}^{1/2} \text{sign}(y_i - \mu_i).$$

Here the $A$ function is independent of the parameter $\mu$, given by

$$A(y; \Sigma) = \frac{\prod_i^{p} \lambda_i \Gamma(\Sigma)}{\prod_i^{p} \Gamma(\lambda_i)} \exp\left\{ - \sum_i^{p} \lambda_i \right\} \prod_i^{p} y_i^{-1}$$

where $\Sigma^{-1} = \Lambda = (\lambda_{ij})$ and

$$\Gamma^{-1}(\Sigma) = E \left[ \exp \left\{ - \sum_{i \neq j} \lambda_{ij} \delta(X_i) \delta(X_j) \right\} \right]$$

the expectation is taken under the probability measure generated by $p$ independent univariate gamma random variables $X_i \sim Ga\left(1, \lambda_i^{-1}\right)$, $i = 1, \ldots, p$. This expectation can be evaluated by using the method of Monte Carlo simulation. Consequently, we may construct the multivariate exponential distribution and multivariate $\chi^2$-distribution as special cases of multivariate gamma distributions.

**Example 2.2 (Multivariate von Mises density)** The univariate von Mises distribution is a dispersion model, but not an exponential dispersion model, with unit deviance function

$$d(y; \mu) = 2 \left\{ 1 - \cos(y - \mu) \right\}.$$ 

We now define a multivariate von Mises distribution with position parameter $\mu$ and dispersion matrix $\Sigma$ by the joint density function

$$p(y; \mu, \Sigma) = A(\Sigma) \exp \left\{ -\frac{1}{2} \delta^T (y; \mu) \Sigma^{-1} \delta(y; \mu) \right\}$$
with $\delta(y; \mu) = (\delta_1, \ldots, \delta_p)^T$ where

$$
\delta_i = \delta(y_i; \mu_i) = \left[ 2 \{ 1 - \cos(y_i - \mu_i) \} \right]^{1/2} \text{sign}(y_i - \mu_i),
$$

and

$$
A^{-1}(\Sigma) = \int_{[0,2\pi)^p} \exp \left\{ -\frac{1}{2} \delta^T(y; \mu) \Sigma^{-1} \delta(y; \mu) \right\} dy
= \int_{[0,2\pi)^p} \exp \left\{ -\frac{1}{2} \delta_0^T(x) \Sigma^{-1} \delta_0(x) \right\} dx.
$$

where $\delta_0(x) = \delta(y; 0)$. So, the $A$ function is independent of $\Sigma$ since the deviance function is periodic.

**Example 2.3 (Multivariate hyperbola distribution)** As is well-known, the univariate hyperbola distribution (c.f. Barndorff-Nielsen, 1978) is proper dispersion model with unit deviance function

$$
d(y; \mu) = \frac{y}{\mu} + \frac{\mu}{y} - 2, \text{ for } \mu > 0.
$$

Hence we define a multivariate hyperbola distribution with vector of parameter $\mu$ and dispersion matrix $\Sigma$ by the joint density function

$$
p(y; \mu, \Sigma) = C(\Sigma) \left( \prod_{i=1}^p y_i \right)^{-1} \exp \left\{ -\frac{1}{2} \delta^T(y; \mu) \Sigma^{-1} \delta(y; \mu) \right\}, \quad y \in \mathcal{R}_+^p
$$

with $\delta(y; \mu) = (\delta_1, \ldots, \delta_p)^T$ where

$$
\delta_i = \delta(y_i; \mu_i) = \left( \frac{y_i}{\mu_i} + \frac{\mu_i}{y_i} - 2 \right)^{1/2} \text{sign}(y_i - \mu_i).
$$

By the transformation $z_i = y_i/\mu_i$, $i = 1, \ldots, p$, we obtain

$$
C^{-1}(\Sigma) = \int_{\mathcal{R}_+^p} \left( \prod_{i=1}^p y_i \right)^{-1} \exp \left\{ -\frac{1}{2} \delta^T(y; \mu) \Sigma^{-1} \delta(y; \mu) \right\} dy
= \int_{\mathcal{R}_+^p} \left( \prod_{i=1}^p z_i \right)^{-1} \exp \left\{ -\frac{1}{2} \delta_1^T(z) \Sigma^{-1} \delta_1(z) \right\} dz
$$

which is independent of the parameter vector $\mu$ and $\delta_1(z) = \delta(z; 1)$. It is obvious that the multivariate hyperbola distribution is a multivariate proper dispersion model.
Example 2.4 (Multivariate symmetric hyperbolic distribution) The univariate symmetric hyperbolic distribution is a proper dispersion model and has unit deviance function given by
\[ d(y; \mu) = 2 \left\{ \left[ 1 + (y - \mu)^2 \right]^{1/2} - 1 \right\} . \]
Hence the multivariate symmetric hyperbolic distribution is defined by the joint density function
\[ p(y; \mu, \Sigma) = C(\Sigma) \left\{ -\frac{1}{2} \delta^T (y; \mu) \Sigma^{-1} \delta(y; \mu) \right\}, \quad y \in \mathbb{R}^p \]
with \( \delta(y; \mu) = \left( \delta_1, \ldots, \delta_p \right)^T \) where
\[ \delta_i = \delta(y_i; \mu_i) = \sqrt{2} \left\{ \left[ 1 + (y_i - \mu_i)^2 \right]^{1/2} - 1 \right\}^{1/2} \text{sign}(y_i - \mu_i) . \]
By the transformation \( z_i = y_i - \mu_i, \ i = 1, \ldots, p, \) we obtain
\[ C^{-1}(\Sigma) = \int_{\mathbb{R}^p} \exp \left\{ -\frac{1}{2} \delta^T_0 (z) \Sigma^{-1} \delta_0(z) \right\} dz \]
where \( \delta_0(z) = \delta(z; 0) \). Obviously, it is a multivariate proper dispersion model.

In general, although it is difficult to obtain closed forms for the marginal distributions and hence moments, it might be possible, with heavy computing, to develop the multivariate analysis for multivariate non-normal data analogous to the classical theory based on the multivariate dispersion model densities.
Chapter 3

Multivariate dispersion models

In Section 2.6, we have seen that Jørgensen and Lauritzen's (1993) multivariate dispersion model densities do not, in general, have closed forms for their marginals and moments. This deficiency limits their applicability, for example in their capacity to model the pattern of marginal means, which is often of interest for the analysis of short time series. To overcome this, we now study an alternative method for constructing multivariate models with given dispersion model marginals and indices of dependence through the copula approach of Sklar (1959) (see also Joe, 1996b). Our aim is to develop a joint distribution that can provide the given marginals.

Constructing multivariate distributions by mean of copulas has proved popular in recent years (c.f. Joe, 1993, Hutchinson and Lai, 1990, Marshall and Olkin, 1988 and Joe, 1996b). The motivation for the copula approach is probably rooted in the aim of forming multivariate non-normal distributions by combining given non-normal marginal models, in a certain way, with dependence patterns. Being a multivariate joint distribution which contains only information regarding dependence, a copula produces new multivariate distributions whenever new suitable marginals are fed into it. A parametric copula is usually obtained by making certain variable transformations for a multivariate parametric distribution such that the transformed variable's distribution no longer contains any parameters of the marginals. As a consequence, what remains only relates to the parametrization of dependence but nothing to do with that of the marginals.

With regard to dependence indices, it is known that Pearson's correlation coefficient
is mainly adequate for measuring linear dependence in the multivariate normal distribution, so when studying multivariate non-normal distributions, we need new tools to measure dependence between components. Among many possible definitions, Kendall's $\tau$, Spearman's $\rho$, the normal scoring $\nu$ defined in the present chapter, are introduced to describe non-linear dependence.

Amongst several types of copulas available in the literature, we are particularly interested in the multivariate normal copula which is “extracted” from the multivariate normal distribution, say $N_m(\mu, \Gamma)$ where $\Gamma = (\gamma_{ij})$ is a Pearson correlation matrix. The main reason for this option is that the multivariate normal distribution has a nice parameter structure, that is, the marginal parameter vector $\mu$ and the correlation parameter matrix $\Gamma$ are “orthogonal” in the sense that the univariate marginal distributions $N(\mu_i, 1)$ are independent of the correlation parameters $\Gamma$ (this property is not satisfied by, for example, the Dirichlet distribution). This property enables us to get rid of marginal information without any information loss regarding the dependence part. There are a few other reasons that makes the normal copula preferable, notably:

- The normal copula may retain some good properties from the multivariate normal distribution where the dependence is characterized through the correlation matrix; this parametrization has been commonly used in practice and seems easy to comprehend. Additionally, most computational tasks involving the normal distribution are available in statistical software.

- Normal copulas may accommodate both “positive” and “negative” dependence since the matrix $\Gamma$ may have values between $-1$ and $1$.

- For small dispersion parameters, univariate dispersion models may have normal approximations, saddlepoint approximations and Edgeworth expansions to assist asymptotic investigations.
Chapter 3. Multivariate dispersion models

The copula approach to generating multivariate distributions gives rise to an important question of how to interpret the parameters of the matrix $\Gamma$ in new multivariate distributions. A natural approach to this question would be to study the possible relationships between $\Gamma$ and the three measures $\tau$, $\rho$ and $\nu$. We show that $\gamma_{ij}$ is identical to the dependence value of the $(i, j)$-th pair measured by the normal scoring $\nu$ (c.f. (3.3.1) for details), leading to a new interpretation for $\gamma_{ij}$. Also, we obtain two closed forms for the relationships between $\gamma_{ij}$ and $\rho$ and between $\gamma_{ij}$ and $\tau$; both are non-linear. A Monte Carlo stimulation illustrates that the value of $\gamma_{ij}$ is close to both $\tau$ and $\rho$.

A class of multivariate dispersion models is then defined by the multivariate normal copula, which is parameterized by both marginal position parameters, dispersion parameters along with the dependence matrix. It is remarkable that Jørgensen and Lauritzen's (1993) multivariate dispersion model densities in Section 2.6 turn out to be the limiting distributions of our multivariate dispersion models for small dispersion parameters. This result links the two classes of multivariate extensions of dispersion models, and it helps us to understand the multivariate generalization of dispersion models from a different point of view.

3.1 Copulas

We start with the definition of a copula. Let $\mathbf{u}_{-I}$ be a subset of components of $\mathbf{u}$ with those indicated by the set $I$ being omitted, where $I$ is a subset of the indices $\{1, \ldots, m\}$.

**Definition 3.1** A mapping $C : (0,1)^m \rightarrow (0,1)$ is called a copula if

1. it is a continuous distribution function; and
2. each margin is a univariate uniform distribution, namely

$$\lim_{\mathbf{u}_{-i} \rightarrow 1} C(\mathbf{u}) = u_i, u_i \in (0,1)$$
where the limit is taken under $u_j \rightarrow 1, \forall j \neq i$.

It is easy to prove that $\lim_{u \rightarrow 1} C(u)$ is a copula, denoted by $C_I(u_I)$ where $I$ is any subset of $\{1, \ldots, m\}$. Copulas are easy to construct. For example, we may construct a copula from a given multivariate distribution as follows.

Consider a random vector $\mathbf{X} \sim F(x)$ where $F(\cdot)$ is a given continuous multivariate distribution function with margins $F_1(x_1), \ldots, F_m(x_m)$. Then, $U_i = F_i(X_i)$ is a random variable that follows the uniform distribution $U(0,1)$ for $i = 1, \ldots, m$. Now consider the inverse distribution function $x_i = F_i^{-1}(u_i)$ and obtain the copula $C_F(u_1, \ldots, u_m)$ as

$$C_F(u_1, \ldots, u_m) = F\left\{F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m)\right\}.$$ 

By the definition above, a copula might be interpreted as the remainder after transformations making the marginals uniform, where the parameters regarding marginals are hence no longer contained in this copula.

Clearly, the joint distribution $F(x)$ can be fully recovered by complementing the copula with its own marginal distributions,

$$F(x) = C_F\{F_1(x_1), \ldots, F_m(x_m)\}.$$ 

However, it is interesting to observe that the function $C_F(u_1, \ldots, u_m)$ can be used to accommodate other univariate distributions, say $G_1(y_1), \ldots, G_m(y_m)$. Therefore, a new multivariate distribution, $G(y)$, is obtained from the copula $C_F(\cdot)$ by plugging in the margins $G_1(y_1), \ldots, G_m(y_m)$,

$$G(y) = C_F\{G_1(y_1), \ldots, G_m(y_m)\}.$$ 

An important property of the multivariate distribution generated by a copula is that the $i$-th margin of $G(y)$ gives the original $G_i(y_i)$, for $i = 1, \ldots, m$. Note that in general, doing variable transformation for the marginals of a joint distribution $F(x)$ might change
its dependence part. However, the multivariate normal distributions do not suffer from this problem. This is because, as is mentioned above, the multivariate normal distribution is parametrized by two orthogonal sets of parameters with regard to the marginals and the correlations, respectively. This parametrization makes the multivariate normal distribution a good candidate for constructing copulas with easy interpretations.

### 3.2 Multivariate normal copulas

Consider an $m$-variate multivariate normal distribution

$$\mathbf{X} = (X_1, \ldots, X_m)^\top \sim \mathcal{N}_m(\mathbf{0}, \Gamma)$$

where the matrix $\Gamma = (\gamma_{ij})$ is chosen to be a correlation matrix whose diagonal elements are all 1 and in effect, $\gamma_{ij}$ is Pearson’s correlation of $X_i$ and $X_j$. We denote the distribution function of $\mathbf{X}$ by $\Phi$. Clearly, all $m$ marginal components of $\mathbf{X}$ have standard univariate normal distributions $\mathcal{N}(0, 1)$, with distribution and density function denoted by $\phi(\cdot)$ and $\varphi(\cdot)$, respectively. An $m$-dimensional normal copula is defined as

$$C_\Phi(u|\Gamma) = \Phi \left\{ \phi^{-1}(u_1), \ldots, \phi^{-1}(u_m) | \Gamma \right\}, \quad u \in (0, 1)^m$$

and the density of $C_\Phi(u|\Gamma)$ is hence

$$c_\Phi(u|\Gamma) = |\Gamma|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{x}^\top \Gamma^{-1} \mathbf{x} + \frac{1}{2} \mathbf{x}^\top \mathbf{x} \right\} = |\Gamma|^{-1/2} \exp \left\{ \frac{1}{2} \mathbf{x}^\top (I - \Gamma^{-1}) \mathbf{x} \right\}$$

where $\mathbf{x} = (x_1, \ldots, x_m)^\top$ with $x_i = \phi^{-1}(u_i), i = 1, \ldots, m$.

Figure 3.1 shows some plots of bivariate normal copulas (bivariate density functions) with different values of the correlation parameter $\gamma$, namely $-0.9$, $-0.5$, $0.5$ and $0.9$. The normal copulas with negative values of $\gamma$ turn out to be concentrated in an opposite direction in relation to those with positive values of $\gamma$. This figure also indicates a nice curvature pattern which seems similar to that of the normal distribution.
Figure 3.1: Four bivariate normal copulas with different values of $\gamma$. 

- $\gamma = -0.9$
- $\gamma = -0.5$
- $\gamma = 0.5$
- $\gamma = 0.9$
Chapter 3. Multivariate dispersion models

3.3 Interpretation of dependence parameters

An important issue that we discuss in this section is how to interpret the elements of the matrix $\Gamma$ in the normal copula (3.2.2). As is mentioned above, three types of dependence measures are employed to explain $\gamma_{ij}$. First we show that $\gamma_{ij}$ turns out to be identical to the normal scoring $\nu(U_i, U_j)$, and then we proceed to establish its connections with the other two dependence measures: Kendall's $\tau$ and Spearman's $\rho$. Here are the definitions of three dependence measures.

**Definition 3.2 (Kendall's $\tau$)**

Let $F(x_1, x_2)$ be a continuous bivariate distribution function. Then Kendall's dependence measure $\tau$ for any independent pairs $(X_1, X_2)$ and $(X'_1, X'_2)$ with distribution $F$ is defined by,

$$
\tau(X_1, X_2) = P \{(X_1 - X'_1)(X_2 - X'_2) > 0\} - P \{(X_1 - X'_1)(X_2 - X'_2) < 0\} = 4 \iint F(x_1, x_2) dF(x_1, x_2) - 1.
$$

From the definition, it is clear that Kendall's $\tau$ is a bivariate measure of monotone dependence for continuous variables and gauges the difference of the probability of two random concordant pairs and the probability of two random discordant pairs.

**Definition 3.3 (Spearman's $\rho$)**

Let $F(x_1, x_2)$ be a continuous bivariate distribution function with marginal distributions $F_1(x_1)$ and $F_2(x_2)$. Then Spearman's dependence measure $\rho$ is defined as the Pearson correlation of $F_1(X_1)$ and $F_2(X_2)$ for any pair $(X_1, X_2) \sim F$. Therefore,

$$
\rho(X_1, X_2) = \text{Corr}\{F_1(X_1), F_2(X_2)\} = 12 \iint F_1(x_1)F_2(x_2)dF(x_1, x_2) - 3.
$$
Spearman’s $\rho$ is also a bivariate measure of monotone dependence for continuous variables and it reflects the association between two (monotonely increasing) transformed random variables that are non-informative with respect to their original variables.

Since Pearson’s correlation measures the dependence between two normal random variables so satisfactorily, why not to use the Pearson correlation for non-normal variables under normalizing transformations? This leads to another dependence measure defined as follows.

**Definition 3.4 (Normal scoring $v$)**

Let $F(x_1, x_2)$ be a continuous bivariate distribution function with marginal distributions $F_1(x_1)$ and $F_2(x_2)$. Then the normal scoring dependence measure $v$ is defined as the Pearson correlation between $q_1(X_1)$ and $q_2(X_2)$ for any pair $(X_1, X_2) \sim F$, namely

$$v(X_1, X_2) = \text{Corr}\{q_1(X_1), q_2(X_2)\}$$

where $q_i(\cdot)$ are two transformations such that $q_i(X_i)$, $i = 1, 2$, follow normal distributions.

Note that the normality transformation $q(\cdot)$ always exists for a continuous random variable.

**Proposition 3.1** Let $X$ be a continuous random variable distributed by $F(x)$. Define

$$q(x) = \phi^{-1}\{F(x)\}$$

Then $q(X)$ follows a normal distribution. The function $q(x)$ is called the normal scoring function and the values of $q(x)$ are called normal scores.

**Proof** Let $Y = q(X)$ where a one-to-one function $q(\cdot)$ is such that

$$\phi(y) = P(Y \leq y) = F\{q^{-1}(y)\}$$
which leads to \( q(x) = \phi^{-1} \{ F(x) \} \).

Actually, the normal scoring index \( \nu \) is a bivariate measure of dependence for continuous variables and its value is calculated from normal scores. It is not hard to generalize the definition of \( \nu \) to discrete distributions. Details are omitted, but in the following the \( \nu \) is assumed well-defined in both the continuous and discrete cases.

Now we apply the three dependence measures above to interpret the matrix \( \Gamma \) that parametrizes the normal copula given in (3.2.2). Consider the bivariate marginal normal copula with components \((U_i, U_j)\) obtained from (3.2.2),

\[
c_{\Phi_{ij}}(u_i, u_j | \gamma_{ij}) = |\Gamma_{ij}|^{-1/2} \exp \left\{ \frac{1}{2} (x_i, x_j) \left( I_2 - \Gamma_{ij}^{-1} \right) \begin{pmatrix} x_i \\ x_j \end{pmatrix} \right\} \\
= (1 - \gamma_{ij}^2)^{-1/2} \exp \left\{ \frac{1}{2} \frac{x_i^2 + x_j^2 - 2\gamma_{ij}x_i x_j}{1 - \gamma_{ij}^2} + \frac{1}{2} \left( \frac{x_i^2 + x_j^2}{1 - \gamma_{ij}^2} \right) \right\}
\]

where \( \Gamma_{ij} \), as a 2 x 2 symmetric submatrix of \( \Gamma \), has all diagonal elements 1 and all off-diagonal elements \( \gamma_{ij} \) and \( I_2 \) is the bivariate identity matrix.

**Normal scoring \( \nu \)**

The normal scoring \( \nu \) of \((U_i, U_j)\) is seen to be immediately

\[
\nu(U_i, U_j) = \text{Corr} \{ q_i(U_i), q_j(U_j) \} \\
= \text{Corr} \{ \phi^{-1}(U_i), \phi^{-1}(U_j) \}.
\]

Note that

\[
\text{Cov} \left\{ \phi^{-1}(U_i), \phi^{-1}(U_j) \right\} = \int \int_{(0, 1)^2} \phi^{-1}(U_i) \phi^{-1}(U_j) c_{\Phi_{ij}}(u_i, u_j | \gamma_{ij}) du_i du_j \\
= \int \int_{\mathbb{R}^2} x_i x_j \rho_{ij}(x_i, x_j | \gamma_{ij}) dx_i dx_j \\
= \gamma_{ij}
\]
where \( \varphi_{ij}(x_i, x_j | \gamma_{ij}) \) is the joint density function of \( N_2(0, \Gamma_{ij}) \). Similarly, \( \text{Var}(q_i(U_i)) = \text{Var}(q_j(U_j)) = 1 \), and therefore,

\[
\nu(U_i, U_j) = \text{Corr}(q_i(U_i), q_j(U_j)) = \gamma_{ij}.
\] (3.3.1)

**Spearman’s \( \rho \)**

Spearman’s \( \rho \) of \( (U_i, U_j) \) is

\[
\rho_{ij} = \rho(U_i, U_j) = 12 \int_{(0,1)^2} c_{\Phi_{ij}}(u_i, u_j | \gamma_{ij}) du_i du_j - 3.
\]

Note that both \( \phi(X_i) \) and \( \phi(X_j) \) are uniformly distributed as \( U(0,1) \) so it is not hard to prove that

\[
\rho_{ij} = 12E\phi(X_i)\phi(X_j) - 3 = 12\text{Cov}(\phi(X_i), \phi(X_j))
\]

where the expectation is taken under the bivariate distribution of \( (X_i, X_j) \sim N_2(0, \Gamma_{ij}) \).

A Monte Carlo simulation study shows that \( \rho_{ij} \) and \( \gamma_{ij} \) are quite close when the simulation size is large enough, say 5,000. See the two top plots of Figure 3.2.

**Kendall’s \( \tau \)**

Kendall’s \( \tau \) is given by

\[
\tau_{ij} = \tau(U_i, U_j) = 4 \int_{(0,1)^2} C_{\Phi_{ij}}(u_i, u_j | \gamma_{ij}) dC_{\Phi_{ij}}(u_i, u_j | \gamma_{ij}) - 1.
\]

It may be re-written as

\[
\tau_{ij} = 4E\Phi_{ij}(X_i, X_j | \gamma_{ij}) - 1,
\]

where the expectation is taken under the bivariate normal distribution \( (X_i, X_j) \sim N_2(0, \Gamma_{ij}) \).

A Monte Carlo simulation study is shown in the two bottom plots of Figure 3.2. Although Kendall’s \( \tau \) and the values of parameter \( \gamma \) are not identical, they show a close relation to reflect the dependence pattern. In other words, increasing the value of \( \gamma \) results in a similar increase for the value of \( \tau \), and *vice versa.*
Figure 3.2: Monte Carlo simulation study showing plots of Spearman’s $\rho$ versus Pearson’s $\gamma$ (top plots) and Kendall’s $\tau$ versus Pearson’s $\gamma$ (bottom plots), where normal scoring dependence measurements are identical to Pearson correlations given in the normal distribution.
Chapter 3. Multivariate dispersion models

3.4 Multivariate dispersion models

Now we define a family of multivariate dispersion models through multivariate normal copulas. Let $Y_i$, $i = 1, \ldots, m$, be $m$ univariate random variables distributed according to dispersion models $\text{DM} \left( \mu_i, \sigma_i^2 \right)$, respectively. The densities and distributions of $\text{DM} \left( \mu_i, \sigma_i^2 \right)$ are denoted by $g_i(y_i; \mu_i, \sigma_i^2)$ and $G_i(y_i; \mu_i, \sigma_i^2)$, respectively. We may construct a multivariate dispersion model via the normal copula $c_{\Phi}(\cdot|\Gamma)$ by the following definition.

**Definition 3.5** An $m$-variate multivariate dispersion model generated by the normal copula $c_{\Phi}(\cdot|\Gamma)$, denoted by $\text{MDM}_m(\mu, \sigma^2, \Gamma)$, with margins $g_i(y_i; \mu_i, \sigma_i^2)$, $i = 1, \ldots, m$, is defined by the joint distribution of form

$$G(y; \mu, \sigma^2, \Gamma) = C_{\Phi} \left\{ G_1(y_1; \mu_1, \sigma_1^2), \ldots, G_m(y_m; \mu_m, \sigma_m^2) \right\} |\Gamma \right\} (3.4.1)$$

or equivalently by the density given by

$$g(y; \mu, \sigma^2, \Gamma) = c_{\Phi} \left\{ G_1(y_1; \mu_1, \sigma_1^2), \ldots, G_m(y_m; \mu_m, \sigma_m^2) \right\} |\Gamma \right\} g_1(y_1; \mu_1, \sigma_1^2) \cdots g_m(y_m; \mu_m, \sigma_m^2) \right\} (3.4.2)$$

where $\mu = (\mu_1, \ldots, \mu_m)^T$ and $\sigma^2 = (\sigma_1^2, \ldots, \sigma_m^2)^T$.

This definition gives the same marginal distributions as those we started with regardless of the values of the matrix $\Gamma$. Consequently, we may express the logarithm of the density $g(y; \mu, \sigma^2, \Gamma)$ in detail as follows,

$$\log g(y; \mu, \sigma^2, \Gamma) = \log c_{\Phi} \left\{ G_1(y_1; \mu_1, \sigma_1^2), \ldots, G_m(y_m; \mu_m, \sigma_m^2) \right\} |\Gamma \right\} + \sum_{i=1}^{m} \log g_i(y_i; \mu_i, \sigma_i^2)$$

$$= -\frac{1}{2} \log |\Gamma| + \frac{1}{2} \sum_{i=1}^{m} \left[ \phi^{-1}\{G_i(y_i)\} \right]^2 - \frac{1}{2} \sum_{i,j=1}^{m} \gamma^{ij} \phi^{-1}\{G_i(y_i)\} \phi^{-1}\{G_j(y_j)\}$$

$$+ \sum_{i=1}^{m} \log g_i(y_i; \mu_i, \sigma_i^2)$$

$$= -\frac{1}{2} \log |\Gamma| + \frac{1}{2} \sum_{i=1}^{m} (1 - \gamma^{ii}) \left[ \phi^{-1}\{G_i(y_i)\} \right]^2$$
where $G_i(y_i) = G_i(y_i; \mu_i, \sigma_i^2)$, $\Gamma^{-1}$—whose $(i,j)$-th element being denoted by $\gamma^{ij}$—is the inverse matrix of $\Gamma$ and $\phi^{-1}\{G_i(y_i)\}$ are normal scoring functions.

**Proposition 3.2** If $\Gamma$ is the identity matrix $I$, then the components $Y_1, \ldots, Y_m$ are independent.

**Proof** When $\Gamma = I$, the logarithm of the joint density becomes

$$\log g(y; \mu, \sigma^2, I) = \sum_{i=1}^{m} \log g_i(y_i; \mu_i, \sigma_i^2)$$

since $|\Gamma| = 1$ and

$$\sum_{i,j=1}^{m} \gamma^{ij} \phi^{-1}\{G_i(y_i)\} \phi^{-1}\{G_j(y_j)\} = \sum_{i=1}^{m} \left[ \phi^{-1}\{G_i(y_i)\} \right]^2.$$ 

Therefore, $Y_1, \ldots, Y_m$ are independent. \(\square\)

### 3.5 Normal scores and their approximations

For convenience, we denote the normal scores by $q_i = \phi^{-1}\{G_i(y_i)\}, \ i = 1, \ldots, m$

with respect to the $i$-th margin $G_i$. The accuracy of evaluation of normal scores involving the joint density $g(y; \mu, \sigma^2, \Gamma)$ is crucial for its potential applications.

Clearly, when $G_i$ is a univariate normal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$, we obtain

$$q_i = \phi^{-1}\{G_i(y_i)\} = \frac{y_i - \mu_i}{\sigma_i}$$

and therefore,

$$\log g(y; \mu, \sigma^2, \Gamma) = -\frac{1}{2} \log |\Gamma| - \frac{1}{2} \sum_{i,j=1}^{m} \frac{\gamma^{ij}}{\sigma_i \sigma_j} (y_i - \mu_i)(y_j - \mu_j) - \frac{m}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{m} \log(\sigma_i^2)$$
which is exactly the logarithm of the density function of the multivariate normal distribution $\mathcal{N}_m(\mu, \Sigma)$ with $\Sigma = \text{diag}(\sigma_i) \Gamma \text{diag}(\sigma_i)$. However, for non-normal margins $G_i$, the evaluation of $\phi^{-1}\{G_i(y_i)\}$ is not so straightforward, and many approximations have been proposed in recent literature.

According to Jørgensen (1996), (also see Jørgensen, 1987b), the so-called technique of small-dispersion asymptotics plays a key role in approximating normal scores, and this evaluation is basically equivalent to assessing the normal approximation for the distribution of the dispersion model. Suppressing the index variable $i$ for convenience, five related approaches are available to approximate the normal scores, starting with the Pearson residual defined by

$$r_p = \frac{y - \mu}{\sigma V^{1/2}(\mu)},$$

a one-to-one function of $y$ for fixed $\mu$. According to the theory of the tail area approximation (c.f. Jørgensen 1987b and Reid 1995), we have the following small-dispersion asymptotic normality,

$$G(y; \mu, \sigma^2) \approx \phi \left( \frac{r_p}{\sigma} \right)$$

(3.5.1)

where $G(y; \mu, \sigma^2)$ is the distribution function of a dispersion model $\text{DM}(\mu, \sigma^2)$ given in (2.1.1), $\phi(\cdot)$ is the distribution function of $\mathcal{N}(0, 1)$ and $r_p$ is the Pearson residual. Here the notation $\approx$ is introduced in Section 2.1. From (3.5.1), the normal score is therefore approximated as

$$q(y) = \phi^{-1}\{G(y; \mu, \sigma^2)\} \approx \frac{r_p}{\sigma} = \frac{y - \mu}{\sigma V^{1/2}(\mu)}.$$

In addition, there are four other types of standardized residuals similar to the Pearson residuals given as follows:

1. the *Wald residual* defined by

$$r_w = V^{1/2}(y) \{\tau^{-1}(y) - \tau^{-1}(\mu)\}, \ y \in \Omega,$$
(2) the score residual
\[ r_s = -\frac{\partial d}{2\partial \mu} V^{1/2}(\mu), \]

(3) the dual score residual of Reid (1995)
\[ u = \frac{\partial d}{2\partial y} V^{1/2}(y) \]

and

(4) the deviance residual
\[ r = \pm d^{1/2}(y; \mu) \]
where \( \pm \) is taken by the sign of \( y - \mu \).

The approximation equivalence of the five standard residuals gives rise to similar normal asymptotics by

\[ G(y; \mu, \sigma^2) \simeq \phi \left( \frac{r_w}{\sigma} \right) \quad \text{or} \quad q(y) \simeq \frac{r_w}{\sigma} \]
\[ G(y; \mu, \sigma^2) \simeq \phi \left( \frac{r_s}{\sigma} \right) \quad \text{or} \quad q(y) \simeq \frac{r_s}{\sigma}, \]
\[ G(y; \mu, \sigma^2) \simeq \phi \left( \frac{u}{\sigma} \right) \quad \text{or} \quad q(y) \simeq \frac{u}{\sigma} \]

and

\[ G(y; \mu, \sigma^2) \simeq \phi \left( \frac{r}{\sigma} \right) \quad \text{or} \quad q(y) \simeq \frac{r}{\sigma} \]
as the dispersion parameter \( \sigma^2 \) becomes small.

To adjust for a certain bias caused by the above approximations, Barndorff-Nielsen (1990) and Jensen (1992) proposed the modified deviance residual defined by

\[ r^* = \frac{r}{\sigma} + \frac{\sigma}{r} \log \frac{u}{r} \]

and they proved that

\[ G(y; \mu, \sigma^2) = \phi(r^*) \left\{ 1 + O \left( \sigma^3 \right) \right\} \]
where the term $\frac{u}{r}$ is always positive because the two residuals $u$ and $r$ take the same sign. As is shown in Jørgensen (1996), see also Figure 3.3, this approximation tends to be very good, and is often significantly better than the closest contender $\phi(r/\sigma)$ based on the deviance residual. However, this approximation may not be reliable for some extreme cases where $u$ tends to zero while $r$ remains bounded as $y$ tends to the lower end of the domain, making $r^*$ tend to infinity, instead of minus infinity as it ordinarily should.

Now we present a more refined approximation for the normal score based on the modified deviance residual $r^*$. Since $\phi^{-1}$ is continuously differentiable,

\[
q = \phi^{-1}\left\{G(y; \mu, \sigma^2)\right\} \\
= \phi^{-1}\left[\phi(r^*)\{1 + O(\sigma^3)\}\right] \\
= \phi^{-1}\left[\phi(r^*)\right] + \left[\varphi\left\{\phi^{-1}(\xi)\right\}\right]^{-1}\phi(r^*)O(\sigma^3) \\
= r^* + \frac{\phi(r^*)}{\varphi\left\{\phi^{-1}(\xi)\right\}}O(\sigma^3)
\]

where $\xi$ is between $\phi(r^*)O(\sigma^3)$ and $\phi(r^*)$. Due to the monotonicity of $\phi^{-1}$, we have

\[
q = \phi^{-1}\left[\phi(r^*)\{1 + O(\sigma^3)\}\right] \leq r^* + \sqrt{2\pi}\phi(r^*)\exp\left\{\frac{(r^*)^2}{2}\right\}O(\sigma^3).
\]

(3.5.2)

**Proposition 3.3** Let $Y \sim ED(\mu, \sigma^2)$. Then as $\sigma^2 \to 0$,

\[
q(Y) = r^* + O_p(\sigma^3).
\]

**Proof** Consider the limit $\sigma^2 \to 0$. First note that $\log \frac{u}{r} = o_p(1)$, where $u$ is the dual score residual and $r$ the deviance residual. So

\[
r^* = \frac{r}{\sigma} + o_p(\sigma)
\]

and moreover

\[
(r^*)^2 = \frac{d(y; \mu)}{\sigma^2} + o_p(r) + o_p(\sigma^2).
\]
Chapter 3. Multivariate dispersion models

Second, we know that $\text{Var}(Y) = \sigma^2 V(\mu) < \infty$, or

$$\frac{(y - \mu)^2}{\sigma^2 V(\mu)} = O_p(1).$$

By the Taylor expansion, it follows that

$$\frac{d(y; \mu)}{\sigma^2} = \frac{(y - \mu)^2}{\sigma^2 V(\mu)} + o_p(|y - \mu|^2) = O_p(1). \quad (3.5.3)$$

Thus the inequality (3.5.2) can be further written as

$$q = r^* + \exp \left\{ \frac{(r^*)^2}{2} \right\} O(\sigma^3)$$

$$= r^* + \exp \left\{ \frac{d^2}{2\sigma^2} + o_p(r) + o_p(\sigma^2) \right\} O(\sigma^3)$$

$$= r^* + O_p(1) O(\sigma^3).$$

The last equality is obtained by (3.5.3).

Figure 3.3 shows quantile plots for gamma residuals. Clearly, when the dispersion parameter $\sigma^2$ is small, there are no big differences among the three approximations, but as $\sigma^2$ increases, the Pearson residual appears to have a substantial bias away from the diagonal line. Similar phenomena for other dispersion models are illustrated in Jørgensen (1996). In summary, when the dispersion parameter $\sigma^2$ is not small, either $r/\sigma$ or $r^*$ is superior to $r_p/\sigma$ for approximating normal scores. This comparison has a great impact on the approximate inference developed in Chapter 4.

3.6 Small-dispersion asymptotics for joint density

In this section, we focus on the study of approximation for the logarithm of the joint density $g(y)$ given by (3.4.1). According to Jørgensen (1996), the four types of residuals including $r_p$, $r_w$, $r_s$ and $u$ are nearly equivalent in terms of their contributions for the tail area approximation under the class of regular dispersion models. So, the Pearson
Figure 3.3: Comparison of three types of residuals from gamma distribution: \( r^* \) is represented by a solid line (—), \( r/\sigma \) by a dashed line (----), \( r_p/\sigma \) by a broken line (---) and normal scores by a dotted line (......) (mostly covered by the line of \( r^* \)).
residual $r_p$ is opted to be their representative for the following study. The other two types of residuals, the deviance residual $r$ and the modified deviance residual $r^*$, are also used in the following study of approximation.

**Approximation based on Pearson residual**

When the dispersion parameter $\sigma^2$ small,

$$
\log g(y; \mu, \sigma^2, \Gamma) \approx -\frac{1}{2} \log |\Gamma| + \frac{1}{2} \sum_{i=1}^{m} (r_{pi})^2 - \frac{1}{2} \sum_{i,j=1}^{m} \gamma_{ij} r_{pi} r_{pj} + \sum_{i=1}^{m} \log g_i(y_i; \mu_i, \sigma_i^2) \\
\approx -\frac{1}{2} \log |\Gamma| + \sum_{i=1}^{m} \log a(y_i; \sigma_i^2) - \frac{1}{2} (y - \mu)^T (D \Gamma D)^{-1} (y - \mu)
$$

where $D = \text{diag}\{\sigma_i V^{1/2}(\mu_i)\}$.

**Approximation based on deviance residual**

$$
\log g(y; \mu, \sigma^2, \Gamma) \approx -\frac{1}{2} \log |\Gamma| + \frac{1}{2} \sum_{i=1}^{m} \sigma_i^2 - \frac{1}{2} \sum_{i,j=1}^{m} \frac{\gamma_{ij}}{\sigma_i \sigma_j} r_i r_j + \sum_{i=1}^{m} \log a(y_i; \sigma_i^2) \\
= -\frac{1}{2} \log |\Gamma| + \frac{1}{2} \sum_{i=1}^{m} \sigma_i^2 - \frac{1}{2} \sum_{i,j=1}^{m} \frac{\gamma_{ij}}{\sigma_i \sigma_j} r_i r_j + \sum_{i=1}^{m} \log a(y_i; \sigma_i^2) - \frac{1}{2} \sum_{i=1}^{m} d(y_i; \mu_i) \\
= -\frac{1}{2} \log |\Gamma| + \sum_{i=1}^{m} \log a(y_i; \sigma_i^2) - \frac{1}{2} r^T (y; \mu) (G \Gamma G)^{-1} r(y; \mu)
$$

where $r(y; \mu) = (r_1, \ldots, r_m)^T$ and $G = \text{diag}(\sigma_i)$. So, the density might be re-written approximately as

$$
g(y; \mu, \sigma^2, \Gamma) \approx |\Gamma|^{-1/2} \prod_{i=1}^{m} a(y_i; \sigma_i^2) \exp \left\{ -\frac{1}{2} r^T (y; \mu) \Sigma^{-1} r(y; \mu) \right\}
$$

(3.6.1)

where $\Sigma = G \Gamma G$. It is remarkable that the density function obtained by normalizing the right-hand side of the approximation (3.6.1) coincides with the definition for multivariate dispersion density of Jørgensen and Lauritzen (1993) given in Section 2.6.

**Approximation based on modified deviance residual**

$$
\log g(y; \mu, \sigma^2, \Gamma) \approx -\frac{1}{2} \log |\Gamma| + \frac{1}{2} \sum_{i=1}^{m} (r_i^*)^2 - \frac{1}{2} \sum_{i,j=1}^{m} \gamma_{ij} r_i^* r_j^* + \sum_{i=1}^{m} \log g_i(y_i; \mu_i, \sigma_i^2)
$$
and it can be further approximated with the deviance residual $r$ by using $r^* = r/\sigma + o_p(\sigma)$. This approximation provides a more accurate representation to the joint density $g(y; \mu, \sigma^2, \Gamma)$ than the other two; however, because $r^*$ involves the parameter $\sigma^2$, this version of the approximation to inference might result in a complex estimating procedure.

### 3.7 Approximate covariance

It is of interest to investigate how the matrix $\Gamma$ characterizes the correlation structure of $Y_1, \ldots, Y_m$. It follows from (3.3.1) that

$$\nu(Y_i, Y_j) = \nu(U_i, U_j) = \gamma_{ij},$$

so, the $(i, j)$-th element $\gamma_{ij}$ of $\Gamma$ turns out to be the normal scoring $\nu$ measurement for the dependence between $Y_i$ and $Y_j$. In addition, a direct relationship between the Pearson correlation and the parameter $\gamma_{ij}$ could be established approximately as follows. Through some trivial calculations, we find that for any $i, j$,

$$\text{Cov}(Y_i, Y_j) = \text{Cov} \left[ G_i^{-1}\{\phi(X_i)\}, G_j^{-1}\{\phi(X_j)\} \right]$$

where $(X_i, X_j) \sim \mathcal{N}_2(0, \Gamma_{ij})$ and $\Gamma_{ij}$ is a $2 \times 2$ symmetric matrix with diagonal elements 1 and off-diagonal element $\gamma_{ij}$. By (3.5.1), the approximation based on the Pearson residual for $\sigma_k^2$ small, $k = 1, \ldots, m$,

$$G_k^{-1}\{\phi(x_k)\} \simeq \sigma_k V^{1/2}(\mu_k)x_k + \mu_k.$$

Hence, for both $\sigma_i^2$ and $\sigma_j^2$ small,

$$\text{Cov}(Y_i, Y_j) \simeq \text{Cov} \left\{ \sigma_i V^{1/2}(\mu_i)X_i + \mu_i, \sigma_j V^{1/2}(\mu_j)X_j + \mu_j \right\} = \sigma_i V^{1/2}(\mu_i)\sigma_j V^{1/2}(\mu_j)\gamma_{ij}$$

Thus, noting that $\text{Var}(Y_k) = \sigma_k^2 V(\mu_k)$, $k = 1, \ldots, m$, we obtain

$$\text{Cov}(Y_i, Y_j) \simeq \gamma_{ij}.$$
Hence the Pearson correlation of \((Y_i, Y_j)\) is approximately \(\gamma_{ij}\) when both \(\sigma_i^2\) and \(\sigma_j^2\) are small. In other words, \(\gamma_{ij}\) gives a local linear correlation in the neighbourhood of \((\mu_i, \mu_j)\). This property will be used to explain the working correlation matrix of the GEE approach suggested by Liang and Zeger (1986) in Chapter 4.
This chapter studies the regression analysis of short time series data through population-averaged models (Zeger et al., 1988) in which the aggregate response for the population is the focus. A likelihood-based theory for approximate inference about regression parameters is presented, provided that each series as a vector of responses follows a multivariate dispersion model introduced in the previous chapter.

The motivation of such a study is to investigate, from a probability point of view, a possible theoretical explanation of the generalized estimating equation (GEE) proposed by Liang and Zeger (1986). Being a quasi-likelihood approach, the GEE has its limitations. That is, the method incorporates a so-called "working" correlation matrix and then treats the matrix as a set of nuisance parameters in the estimation procedure. Only second moment assumptions are made for the margins, whereas the standard errors of the regression parameter estimates are estimated using the so-called sandwich estimator. This makes the procedure semi-parametric in nature, leaving open the question of the existence of an underlying probability model.

Approximate inference for the regression parameters of the population-averaged models is based on both Pearson and deviance residuals that approximate the normal scores in the multivariate dispersion models. An important finding in this chapter is an asymptotic relation between our likelihood-oriented method and Liang and Zeger's GEE approach as the dispersion parameter becomes small. As a consequence, the GEE method may be justified through a fully specified probability model, thereby reinforcing the theoretical
basis of the GEE approach.

When the dispersion parameter is not small, we propose an alternative approach based on the deviance residual, which provides a much more accurate approximation to the normal score but results in a biased inference function. Following Kuk (1995), a so-called bias-corrected estimate of the regression parameters is suggested, and is shown to be asymptotically unbiased and asymptotically normally distributed.

4.1 Models

Consider a data set consisting of \( n \) independent short time series where each series is observed over the same \( m \) occasions. Assume that the \( i \)-th series (the \( m \)-vector of responses), \( \mathbf{Y}_i \), follows a multivariate exponential dispersion model,

\[
\mathbf{Y}_i \sim \text{MDM}_m(\mu_i, \sigma^2_i, \Gamma), \quad i = 1, \ldots, n
\]

where \( \mu_i = (\mu_{i1}, \ldots, \mu_{im})^T \) and \( \sigma^2_i = (\sigma_{i1}, \ldots, \sigma_{im})^T \) are respectively the vector of mean parameters and the vector of dispersion parameters for the \( i \)-th series. We assume \( \sigma_{it} = \sigma^2/\omega_{it} \) where \( \omega_{it} \) are known positive weights, \( t = 1, \ldots, m \) and \( \sigma^2 \) is an unknown constant dispersion parameter common for all series.

Now assume that each series is associated with a \( p \)-vector of covariates \( \mathbf{x}_{it} \) available at time \( t \). Let \( \mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im})^T \) represent the \( m \times p \) matrix of covariates associated with the \( i \)-th series. Hence, the data for the \( i \)-th subject consist of the pair of matrices \( (\mathbf{Y}_i, \mathbf{X}_i) \).

It follows immediately that for given covariate matrix \( \mathbf{X}_i \), the marginal distribution of the \( t \)-th component, \( Y_{it} \), of \( \mathbf{Y}_i \), \( t = 1, \ldots, m, \quad i = 1, \ldots, n \) is

\[
Y_{it} | \mathbf{X}_i \sim \text{ED} \left( \mu_{it}, \sigma^2/\omega_{it}^2 \right).
\]
We further assume that the marginal means follow a generalized linear model,

\[ h(\mu_{it}) = \eta_{it}, \quad \eta_{it} = x_{it}^T \beta \]

where \( \beta \) is a \( p \)-vector of regression parameters of interest and the link function \( h \) is usually chosen to be the canonical link (for example, the log link for the Poisson distribution and the logistic link for the binomial distribution). The main objective is to estimate the parameters \( \theta = (\beta, \sigma^2, \Gamma) \). We can group the \( \mu_{it}(\beta) \) together to form a \( m \)-vector \( \mu_i(\beta) \) including the marginal means, \( \mu_i(\beta) = (\mu_{i1}, \ldots, \mu_{im})^T \). Note that the dependence among the components of \( Y_i \) is characterized by the joint distribution \( MDM_m(\mu_i, \sigma_i^2, \Gamma) \).

With observations \( y_{it}, t = 1, \ldots, m, \ i = 1, \ldots, n \), the full log-likelihood function for the parameter \( \theta \) is given by

\[
\ell(\beta, \sigma^2, \Gamma; Y) = -\frac{n}{2} \log |\Gamma| + \sum_{i=1}^{n} \sum_{t=1}^{m} \log a(y_{it}; \sigma^2/w_{it}^2) - \frac{1}{2} \sum_{i=1}^{n} q_i^T(y_i; \beta, \sigma^2) (I - \Gamma^{-1}) q_i(y_i; \beta, \sigma^2)
\]

where \( q_i(y_i; \beta, \sigma^2) = (q_{it1}, \ldots, q_{itm})^T \), the vectors of normal scores depending on parameters \( (\beta, \sigma^2) \) and \( y_i = (y_{i1}, \ldots, y_{im})^T \), because indeed \( q_{it} = \phi^{-1} \{ G_{it}(y_{it}; \mu_{it}, \sigma^2/w_{it}^2) \} \)
where \( G_{it} \) are the distribution functions of \( \text{ED} (\mu_{it}, \sigma^2/w_{it}^2) \).

In principle, the maximum likelihood estimate \( \hat{\beta} \) can be obtained, applying the chain rule of differentiation with respect to \( \beta \), as the solution to the estimating equation given by

\[
\Psi(\beta; Y, \sigma^2, \Gamma) = \sum_{i=1}^{n} \left( \frac{\partial q_i}{\partial \beta} \right)^T (I - \Gamma^{-1}) q_i = 0 \quad (4.1.1)
\]

where

\[
\frac{\partial q_{it}}{\partial \beta} = \frac{\partial q_{it}}{\partial G_{it}} \frac{\partial G_{it}}{\partial \mu_{it}} \frac{\partial \mu_{it}}{\partial \beta} = \frac{\mu_{it}}{V(\mu_{it}) \varphi(G_{it})} \left\{ \frac{1}{\mu_{it}} \int_{-\infty}^{y_{it}} u g_{it}(u) du - G_{it} \right\} x_{it}.
\]

Here \( \varphi(\cdot) \) is the density function of the standard normal distribution and \( g_{it} \) is the density function of \( G_{it} \). It is worth noting that since the equation (4.1.1) is derived from an exact
full log-likelihood, the inference function $\Psi$ must be unbiased, that is $E\Psi = 0$. Therefore, under some mild regularity conditions, standard theory for inference functions implies asymptotic normality (Cox and Hinkley, 1974 or Godambe, 1976), as both $\sigma^2$ and $\Gamma$ are fixed

$$\sqrt{n} (\hat{\beta} - \beta) \xrightarrow{d} N_p(0, G), \quad \text{as } n \to \infty$$

with $G = \lim_n nJ^{-1}$ where $J$ is the Godambe information matrix defined by

$$J = S(\beta)V^{-1}(\beta)S^T(\beta)$$

with sensitivity matrix

$$S = E\beta \Psi'(\beta; Y, \sigma^2, \Gamma)$$

and variability matrix

$$V = E\beta \{\Psi(\beta; Y, \sigma^2, \Gamma)\Psi^T(\beta; Y, \sigma^2, \Gamma)\}.$$ 

Here we make several remarks:

(1) In the asymptotic result above the Godambe information matrix is actually the Fisher information matrix, because the inference function 4.1.1 is derived directly from the real likelihood function.

(2) For Poisson count data without overdispersion, $\sigma^2 = 1$, so $\Psi$ depends on only $\beta$ and $Y$ for given $\Gamma$.

(3) Unfortunately, it is difficult to obtain the estimate $\hat{\beta}$ directly from equation (4.1.1) due to the complexity of the normal scores $q_{it}$. Therefore, approximations to the normal scores in (4.1.1) are invoked. Details are given in the following sections.

Similarly, given $\beta$ and $\sigma^2$, differentiating the full log-likelihood function with respect to $\Gamma^{-1}$ and noting that (see for example Rogers, 1980)

$$\frac{d|\Gamma^{-1}|}{d\Gamma^{-1}} = |\Gamma^{-1}|$$
we obtain the maximum likelihood estimate of \( \Gamma \)

\[
\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} q_i(y_i; \beta, \sigma^2) q_i^T(y_i; \beta, \sigma^2). \tag{4.1.2}
\]

Expressing the derivative of the likelihood with respect to the dispersion parameter \( \sigma^2 \) is quite complex because the derivatives for the function \( a \) and the normal scores \( q_i \) are non-linear function in \( \sigma^2 \), so that there is no closed form available, in general, for the estimate of \( \sigma^2 \). In the following sections, various versions of estimate of \( \sigma^2 \) will be given under different choices of the normal scoring approximations. In all cases, to find \( \hat{\sigma}^2 \) we shall use the saddlepoint approximation to the function \( a \),

\[
a(y_i; \sigma^2) \sim \left\{ 2 \pi \sigma^2 V(y_i) \right\}^{-1/2}.
\]

### 4.2 Approximate inference based on the Pearson residual

The first approximation is based on the Pearson residual, which leads to an unbiased inference function shown as follows. Let \( W_i = \text{diag}(w_{it}) \). With the Pearson residual, the log-likelihood function for the parameters \( (\beta, \sigma^2, \Gamma) \) is

\[
\ell(\beta, \sigma^2, \Gamma; \mathbf{Y}) \sim -\frac{n}{2} \log |\Gamma| + \sum_{i=1}^{n} \sum_{t=1}^{p} \log a(y_{it}; \sigma^2/w_{it}^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^T W_i V^{-1} \Gamma^{-1} V^{-1} W_i (y_i - \mu_i)
\]

where \( V = \text{diag}\left\{ V^{1/2}(\mu_{it}) \right\} \).

Taking the derivative of the log-likelihood with respect to the regression parameter \( \beta \), we obtain

\[
\frac{\partial \ell}{\partial \beta} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{\partial \mu_i^T}{\partial \beta} W_i V^{-1} \Gamma^{-1} V^{-1} W_i (y_i - \mu_i)
\]

and hence the estimate of \( \beta \), denoted by \( \hat{\beta}_p \), is given as the solution to the following equation

\[
U_p(\beta; \mathbf{Y}, \Gamma) = \sum_{i=1}^{n} \frac{\partial \mu_i^T}{\partial \beta} W_i V^{-1} \Gamma^{-1} V^{-1} W_i (y_i - \mu_i) = \sum_{i=1}^{n} D_i^T W_i \Sigma_i^{-1} W_i (y_i - \mu_i) = 0
\]
where

\[ D_i = \frac{\partial \mu_i^T}{\partial \beta} \]

and \( \sigma^2 \Sigma_i = \sigma^2 \text{diag} \{ V^{1/2}(\mu_i) \} \) \( \Gamma \text{ diag} \{ V^{1/2}(\mu_i) \} \) is the approximate covariance matrix of \( Y_i \), as shown in Section 3.7 when the dispersion parameter \( \sigma^2 \) becomes small. It is easy to show that for given \( \Gamma \),

\[ E_\beta U_p(\beta; Y, \Gamma) = 0, \]

so the inference function \( U_p(\beta; Y, \Gamma) \) is unbiased. Therefore, we have the asymptotic normality

\[ \sqrt{n} \left( \hat{\beta}_p - \beta \right) \overset{d}{\rightarrow} N_p \left( 0, G_p \right), \quad \text{as } n \rightarrow \infty \]

with \( G_p = \lim_n nJ_p^{-1} \) where \( J_p \) is the Godambe information matrix given by

\[ J_p = S_pV_p^{-1}S_p^T \]

with the sensitivity matrix and the variability matrix are given respectively by

\[ S_p = E_\beta U_p'(\beta; Y, \Gamma), \quad V_p = E_\beta \{ U_p(\beta; Y, \Gamma)U_p'(\beta; Y, \Gamma) \} . \]

Because of the asymptotic normality, we then have

\[ \hat{\beta}_p - \beta = O_p \left( n^{-1/2} \right) . \]

This proves that the estimate \( \hat{\beta}_p \) is consistent and asymptotically unbiased to \( \beta \).

Furthermore, simplifying the above process by letting all \( w_{it} \) be 1, we obtain a simpler version of the estimating equation for \( \beta \) as follows,

\[ \sum_{i=1}^n D_i^T \Sigma_i^{-1}(y_i - \mu_i) = 0. \]  \( (4.2.1) \)

Remarkably, equation (4.2.1) coincides with the form of Liang and Zeger's (1986) generalized estimating equation. It is also remarkable that \( \Sigma_i \) (or \( \Gamma \)), being treated as the
nuisance parameters in their paper, may be interpreted, in our setting, as dependence measures in the sense explained in the previous chapter. The computation of \( \hat{\beta}_p \) may be done by the GEE routine available in S-PLUS.

From the approximate log-likelihood, we could also obtain an estimate of \( \Gamma \)

\[
\hat{\Gamma} = \frac{1}{n\sigma^2} \sum_{i=1}^{n} W_i V_i^{-1} (y_i - \mu_i)(y_i - \mu_i)^T V_i^{-1} W_i
\]

and the following estimate of \( \sigma^2 \)

\[
\hat{\sigma}_p^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \mu_i)^T W_i \Sigma_i^{-1} W_i (y_i - \mu_i).
\]

### 4.3 Approximate inference based on the deviance residual

When the dispersion parameter is not small, we argued in Section 3.5 that the approximation by Pearson residual is likely to cause serious bias and hence results in inaccurate inference. On the other hand, the deviance residual provides a better approximation to the normal scores, leading to a new estimating equation based on the deviance residual as follows.

Let \( r_i = (r_{i1}, \ldots, r_{im})^T \). The full log-likelihood function for the parameter \((\beta, \sigma^2, \Gamma)\) is approximately of the form

\[
\ell(\beta, \sigma, \Gamma; Y) \approx -\frac{n}{2} \log |\Gamma| + \sum_{i=1}^{n} \sum_{t=1}^{m} \log a(y_{it}; \sigma^2/w_{it}^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} r_i^T (y_i; \mu_i) W_i \Gamma^{-1} W_i r_i(y_i; \mu_i).
\]

Let

\[
R_i = \frac{\partial r_i}{\partial \beta}.
\]

Then, the estimate \( \hat{\beta}_r \) is defined as the solution to the equation given as follows,

\[
U_\Gamma(\beta; Y) = \sum_{i=1}^{n} R_i^T W_i \Gamma^{-1} W_i r_i(y_i; \mu_i) = 0 \quad (4.3.1)
\]
where $R_i$ is an $m \times p$ matrix given by

$$R_i(\beta, \sigma^2) = \text{diag} \left( \frac{y_{it} - \mu_{it}}{r_{it} V(\mu_{it})} \right) D_i.$$ 

Clearly, the inference function $U_\Gamma$ is no longer unbiased, i.e., $E_\beta U_\Gamma(\beta; Y) \neq 0$, and therefore the estimate $\hat{\beta}_\Gamma$ is inconsistent (to the true parameter $\beta$).

To compute $\hat{\beta}_\Gamma$, we iterate between approximate Fisher scoring for $\beta$ and updating estimates of $\Gamma$ and $\sigma^2$. Given current estimate $\Gamma$ and $\sigma^2$, we suggest the following modified iterative procedure for $\beta$,

$$\hat{\beta}_\Gamma^{(j+1)} = \hat{\beta}_\Gamma^{(j)} - \left\{ \sum_1^n R_i^{T}(\hat{\beta}_\Gamma^{(j)})W_i\Gamma^{-1}(\hat{\beta}_\Gamma^{(j)})W_iR_i(\hat{\beta}_\Gamma^{(j)}) \right\}^{-1} \times \left\{ \sum_1^n R_i^{T}(\hat{\beta}_\Gamma^{(j)})W_i\Gamma^{-1}(\hat{\beta}_\Gamma^{(j)})W_i \left( y_{it} - \mu_{it}(\hat{\beta}_\Gamma^{(j)}) \right) \right\}.$$ 

Inspired by Kuk (1995), we now consider a bias-corrected estimator of $\beta$. Let $\beta^*$ be a vector of parameters such that

$$E_\beta U_\Gamma(\beta^*; Y) = 0.$$ 

Then, as the sample size $n$ increases, we have the standard asymptotic normality as follows,

$$\sqrt{n} \left( \hat{\beta}_\Gamma - \beta^* \right) \stackrel{d}{\rightarrow} N_p(0, G_\Gamma)$$

with $G_\Gamma = \lim_n nJ_\Gamma^{-1}$ where $J_\Gamma$ is the Godambe information matrix given by

$$J_\Gamma = S_\Gamma V_\Gamma^{-1} S_\Gamma^T$$

with the sensitivity matrix and the variability matrix given by respectively

$$S_\Gamma = E_\beta \left\{ U_\Gamma(\beta^*; Y) \right\} \quad \text{and} \quad V_\Gamma = E_\beta \left\{ U_\Gamma(\beta^*; Y)U_\Gamma^T(\beta^*; Y) \right\}.$$
Suppose that there exists a vector of one-to-one differentiable function \( f \) such that \( \beta^* = f(\beta) \) holds componentwise. Define a new estimator of \( \beta \) by \( \tilde{\beta} = f^{-1}(\hat{\beta}_r) \). Then it follows from Slutsky-Fréchet’s theorem (c.f. Serfling, 1980) that

\[
\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} \mathcal{N}_p(0, \bar{G})
\]

where

\[
\bar{G} = H G_r H^T
\]

and \( H \) is the matrix of the derivatives of the components of \( \beta = f^{-1}(\beta^*) \) with respect to the components of \( \beta^* \). Hence the estimate \( \tilde{\beta} \) is consistent and asymptotically normal.

With a specified function vector \( f \), we derive bias-correction as follows. Define the bias by

\[
b(\beta) = \beta^* - \beta = f(\beta) - \beta.
\]

Start with \( b^{(0)} \) which is an initial estimate for the bias of \( \hat{\beta}_r \) and then update both by

\[
b^{(k+1)} = f(\tilde{\beta}_r - b^{(k)}) - (\tilde{\beta}_r - b^{(k)})
\]

(4.3.2)

and

\[
\tilde{\beta}^{(k+1)} = \tilde{\beta}_r - b^{(k+1)}.
\]

If the limit \( b \) of \( b^{(k)} \) exists, then by letting \( k \to \infty \) in equation (4.3.2), we obtain

\[
b = f(\tilde{\beta}) - (\tilde{\beta}_r - b)
\]

so that \( \tilde{\beta} = f^{-1}(\tilde{\beta}_r) \). On way to initialize the bias is

\[
b^{(0)} = f(\hat{\beta}_p) - \hat{\beta}_p
\]

where \( \hat{\beta}_p \) is the estimate of \( \beta \) obtained from the GEE approach.

In general, it is hard to get a closed form for the function \( f \), but numerical evaluation of \( f \) at given points is feasible. For instance, given the fact that \( \tilde{\beta}_r \) has the asymptotic
mean $\beta^* = f(\beta)$, $f(\beta)$ might be approximated by the average of $\hat{\beta}_r$ over simulated samples from the model with parameter set at $\beta$. For details, we refer to Kuk (1995) and numerical examples therein.

We may consider a further approximation for $r_i$ as follows,

$$r_i(y_i; \mu_i) \simeq V_i^{-1}(y_i - \mu_i)$$

and

$$\frac{\partial r_i^T}{\partial \beta} = D_i^T \text{diag} \left( \frac{\partial r_{it}}{\partial \mu_{it}} \right) = -D_i^T \text{diag} \left( \frac{y_{it} - \mu_{it}}{r_{it}V(\mu_{it})} \right) \simeq -D_i^T V_i^{-1}.$$

Therefore we obtain the same approximate form as equation (4.2.1),

$$\sum_{i=1}^{n} D_i^T W_i \Sigma_i^{-1} W_i (y_i - \mu_i) = 0.$$

Similarly, we could then obtain the following estimate of $\Gamma$,

$$\hat{\Gamma} = \frac{1}{n\sigma^2} \sum_{i=1}^{n} W_i \ r_i(y_i; \mu_i) \ r_i^T(y_i; \mu_i) \ W_i$$

and the following estimate of $\sigma^2$,

$$\hat{\sigma}^2 = \frac{1}{n - p} \sum_{i=1}^{n} r_i^T(y_i; \mu_i) \ W_i \Gamma^{-1} W_i \ r_i(y_i; \mu_i).$$

### 4.4 Approximate inference based on the modified deviance residual

In this case, a closed form for the estimate of $\Gamma$ is available. However, since $r^*$ involves the parameter $\sigma^2$, it is nearly impossible to make inference for $\beta$. For theoretical reasons, we present some calculations in details as follows.

Let $r_i^* = (r_{i1}^*, \ldots, r_{im}^*)^T$ where $r_{it}^*$ denote the modified deviance residuals. The log-likelihood function for parameters $(\beta, \sigma^2, \Gamma)$ may be approximated by

$$\ell(\beta, \sigma, \Gamma; Y) \simeq -\frac{n}{2} \log |\Gamma| + \sum_{i=1}^{n} \sum_{t=1}^{p} \log a(y_{it}; \sigma^2/w_{it}^2) + \frac{1}{2} \sum_{i=1}^{n} \sum_{t=1}^{m} (r_{it}^*)^2 -$$
To derive the likelihood estimating equation, we first study the derivative of the form

\[
\frac{\partial r_i^*}{\partial \mu_i} = \text{diag} \left( \frac{\partial r_i^*}{\partial \mu_{it}} \right)
\]

whose \(t\)-th diagonal element is

\[
\frac{\partial r_{it}^*}{\partial \mu_{it}} = \frac{w_{it}}{\sigma} \frac{\partial e_{it}}{\partial \mu_{it}} + \frac{\sigma}{w_{it}} f_{it},
\]

where \(d_{it} = d(y_{it}; \mu_{it})\) and

\[
e_{it} = 1 - \frac{\sigma^2}{w_{it}^2 d_{it}} - \frac{\sigma^2}{w_{it}^2 d_{it}} \log \frac{u_{it}}{r_{it}}, \quad f_{it} = \frac{1}{u_{it} r_{it}} \frac{V^{1/2}(y_{it})}{V(\mu_{it})}.
\]

In matrix notation, we obtain

\[
\frac{\partial r_i^*}{\partial \mu_i} = \frac{1}{\sigma} W_i E_i \frac{\partial r_i}{\partial \mu_i} + \sigma W_i^{-1} F_i,
\]

where \(E_i = \text{diag}(e_{it})\) and \(F_i = \text{diag}(f_{it})\). So, the estimating equation is finally given by

\[
\sum_{i=1}^n D_i^T \frac{\partial (r_i^*)^T}{\partial \mu_i} (I - \Gamma^{-1}) (r_i^*) = \sum_{i=1}^n D_i^T \frac{\partial r_i}{\partial \mu_i} E_i W_i (I - \Gamma^{-1}) (r_i^*) + \sum_{i=1}^n \sigma^2 F_i W_i^{-1} (I - \Gamma^{-1}) (r_i^*) = 0.
\]

It turns out that the approximation based on modified deviance residual does not reduce the simplicity of the estimation procedure for \(\beta\). However, we obtain the following estimate of \(\Gamma\)

\[
\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^n r_i^* (y_{ii}; \hat{\beta}, \hat{\sigma}^2) r_i^*^T (y_{ii}; \hat{\beta}, \hat{\sigma}^2).
\]
4.5 Implementation of the estimating procedure

To compute the estimate of $\beta$, we choose the estimating equation based on the Pearson residual for small dispersion parameter, and otherwise choose the estimating equation based on the deviance residual. As shown in Section 3.5, along with Figure 3.3, for the accuracy of normal scoring approximation, we implement our estimating procedure as follows.

No matter which approximation residual we use for computing $\hat{\beta}$, we suggest estimating $\Gamma$ always by

$$\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} r_i^T \left( y_i; \hat{\beta}, \hat{\sigma}^2 \right) r_i \right) r_i^T \left( y_i; \hat{\beta}, \hat{\sigma}^2 \right)$$

and $\sigma^2$ by

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} r_i^T \left\{ \frac{y_i}{\mu_i(\hat{\beta})} \right\} W_i \hat{\Gamma}^{-1} W_i r_i \left\{ \frac{y_i}{\mu_i(\hat{\beta})} \right\}.$$  (4.5.2)

The algorithm is as follows.

1. Fit data by a generalized linear model under the assumption of independence, and obtain $\hat{\beta}_0$ and $\hat{\sigma}^2_0$. These estimates should be good starting values since they are a kind of robust estimates against the misspecification of $\Gamma$. Here $\Gamma = I$.

2. Given $\hat{\beta}_0$ and $\hat{\sigma}^2_0$, obtain $\hat{\Gamma}_0$ by equation (4.5.1) above.

3. With $\hat{\sigma}^2_0$ and $\hat{\Gamma}_0$ given, fit a model again to compute $\hat{\beta}_p$ using the GEE method of Liang and Zeger (1986), corresponding to approximate inference based on Pearson residuals, or to compute $\tilde{\beta}$ using the bias-corrected approach of Kuk (1995), corresponding to the approximate inference based on deviance residuals.

4. Use the new $\hat{\beta}$ to update $\hat{\sigma}$ by equation (9.2.11) and then update $\hat{\Gamma}$ by equation (4.5.1).

5. Repeat Steps (3) and (4) until convergence.
Clearly, the standard errors of the estimates could be computed using the Godambe information matrix given in the previous sections. The Jackknife approach is another way to calculate standard errors, see Joe (1996b, Chapter 10) for details.

Finally, to compare the efficiency of the two estimators $\hat{\beta}_p$ and $\tilde{\beta}$, we may calculate the asymptotic relative efficiency (ARE) of $\hat{\beta}_p$ versus $\tilde{\beta}$, which is given by the diagonal elements of

$$\text{ARE} = \text{diag}(J_p^{-1}) \{\text{diag}(G)\} = \text{diag}(G_p) \{\text{diag}(H J_f^{-1} H^T)\}^{-1}.$$
Chapter 5

Stationary models with exponential dispersion model margins

5.1 Introduction

Let us first consider the sunspot data, reported by Andrews and Herzberg (1980), which are monthly mean relative sunspot numbers over the period 1749–1976. These monthly data are the simple monthly means of the daily relative sunspot values that are calculated by Wölfet’s formula (c.f. Andrews and Herzberg, 1980), so they are in fact non-negative continuous data rather than integer data. Due to the large amount of monthly data, we chose to show the Wölfer’s yearly mean relative sunspot numbers instead in Figure 5.1 (the time-series plot based on monthly data appears to be too squeezed and therefore can not exhibit pattern of the data).

Many studies have assumed that the sunspot data (for example, Wölfer’s yearly mean relative sunspot numbers) form a stationary time series (c.f. Yule, 1927). Most such studies assume that the sunspot data approximately follow a normal distribution, allowing one to use classical ARMA models. Noting that the sunspot data are positive numbers with occasional zeroes, we would suggest that the data are likely to follow a compound Poisson marginal distribution, resulting from a sum of gamma random variables (representing Wölfer’s mean relative sunspot numbers) with a Poisson number of terms representing the number of places with the presence of sunspots on the surface of the sun. Thus in the compound Poisson setting, a zero sunspot number is obtained when the Poisson random variable takes the value zero, interpreted as the case where no sunspots occurred for that
Figure 5.1: Yearly mean sunspot numbers.
year (or month). It is natural to ask whether or not the classic ARMA model can really be used to fit such data when we realize the data are not from a normal distribution. More generally, a fundamental question would be whether or not it is possible to generalize the classical ARMA models to fit stationary time series with non-normal margins such as compound Poisson, Poisson, gamma, negative binomial or inverse Gaussian.

In the past two decades or so, this issue has received considerable attention, starting with the papers Lawrance and Lewis (1977) and Jacobs and Lewis (1977) for the case of exponential margins. Models for other marginal distributions such as gamma, Poisson and negative binomial followed. (c.f. Gaver and Lewis 1980, Lawrance 1982, McKenzie 1986, 1988, Lewis, McKenzie and Hugus 1989, McCormick and Park 1992 and Al-Osh and Aly 1992). A unification and generalization of some of these models was recently proposed by Joe (1996a), based on convolution-closed families of infinitely divisible distributions, allowing the construction of autoregressive moving average (ARMA) models of arbitrary orders.

Let us first consider Joe's (1996a) construction of an AR(1) model of the form

\[ X_t = A_t(X_{t-1}; \alpha) + \epsilon_t, \quad t = 1, \ldots, \]  

(5.1.1)

where \( A_t \) is a thinning operation defined in Section 2.5, and the two terms on the right-hand side of the equation are independent. The subscript \( t \) on the operator \( A_t \) emphasizes that different operators are used for each \( t \) and the operator \( A_t \) is assumed to be conditionally independent of \( \epsilon_t \) given \( X_{t-1} \). We want the marginal distribution of \( X_t \) to be an infinitely divisible exponential dispersion model \( ED^*(\theta, \lambda) \). By the definition of the thinning operation, the variable \( X_{t-1} \) with distribution \( ED^*(\theta, \lambda) \) is 'thinned' to a variable \( A_t(X_{t-1}; \alpha) \) with distribution \( ED^*(\theta, \alpha \lambda) \) for some \( \alpha \in (0,1) \). Assuming that the innovations \( \epsilon_t \) are independent and identically distributed \( ED^*(\theta, (1 - \alpha) \lambda) \) and using
the convolution formula

\[ ED^*(\theta, \lambda_1) \ast ED^*(\theta, \lambda_2) = ED^*(\theta, \lambda_1 + \lambda_2), \]  

(5.1.2)

the desired distribution for \( X_t \) is obtained. The thinning operation, defined in Section 2.5, generalizes the notion of binomial thinning of a Poisson variable (c.f. McKenzie, 1988), and the beta 'thinning' of the gamma distribution used by Lewis, McKenzie and Hugus (1989).

The generalization of the thinning operation to the thinning of a vector of correlated variables is not that straightforward. For this reason, the generalization of (5.1.1) to higher-order AR processes is complicated, as evidenced by the technique used by Joe (1996a). In contrast it is easy to construct moving average processes, because only the thinning of independent variables is involved. Thus, following Joe (1996a) we define a \( q \)-th order moving average process \( \{X_t\} \) with margins \( ED^*(\theta, \lambda) \) and coefficients \( \alpha_1, \ldots, \alpha_q \in [0,1] \) by

\[ X_t = \varepsilon_t + \sum_{j=1}^{q} A_{t,j}(\varepsilon_{t-j}; \alpha_j) \]  

(5.1.3)

where \( \{\varepsilon_t\} \) is a sequence of i.i.d. \( ED^*(\theta, \lambda/(1 + \sum_j \alpha_j)) \) innovations, and where the \( q \) terms in the sum on the left-hand side of (5.1.3) are conditionally independent given the innovations.

The definition in equation (5.1.3) is easily extended to a moving average model of infinite order, as made precise in Section 5.2. In the normal case we know that any causal ARMA model may be represented as an infinite order moving average model. Our approach, then, is to use this equivalence to define what we call EDARMA\((p, q)\) processes from the infinite order moving average processes in Section 5.2. Although this approach is less direct than that of Joe (1996a), it is mathematically more convenient, and allows us to transplant many techniques from linear time-series analysis into our framework.
5.2 Moving average processes of infinite order

We now use the thinning operation to construct a class of moving average processes of infinite order with univariate margins of exponential dispersion model form. Moving average processes of infinite order with normal margins have been considered in many practical fields such as geophysics (Dimri, 1992), spectroscopy (Jansson, 1984), and are also known as convolution models. The extension to an infinite order moving average model may be based on the Hilbert space $L^2(\Omega, \mathcal{A}, P)$ whenever second moments exist (c.f. Brockwell and Davis 1987, Chapter 2). Without moment assumptions (at the extremes of the canonical parameter domain), the infinite sums appearing below may be interpreted in terms of convergence in probability.

From now on $ED^*(\theta, \lambda)$ denotes an infinitely divisible exponential dispersion model.

**Definition 5.1** The process $\{X_t\}_{t=-\infty}^{\infty}$ defined by

$$X_t = \sum_{j=0}^{\infty} A_{t,j}(\varepsilon_{t-j}; \alpha_j),\quad (5.2.1)$$

where the coefficients $\alpha_0, \alpha_1, \ldots \in [0, 1]$ satisfy $\alpha_0 = 1$ and $\alpha_+ = \sum_{j=0}^{\infty} \alpha_j < \infty$, is said to be an infinite order moving average process with margins $ED^*(\theta, \lambda)$, denoted $EDMA(\infty)$, provided:

1. the innovations $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ are i.i.d. $ED^*(\theta, \lambda/\alpha_+)$;

2. the thinning operators $\{A_{t,j}(\varepsilon_{t-j}; \alpha_j)\}_{j=0}^{\infty}$ are conditionally independent given the innovations $\{\varepsilon_t\}_{t=-\infty}^{\infty}$.

Since $A_{t,0}(\varepsilon_t; 1) = \varepsilon_t$, we may write the $EDMA(\infty)$ process in the form

$$X_t = \varepsilon_t + \sum_{j=1}^{\infty} A_{t,j}(\varepsilon_{t-j}; \alpha_j).$$
It is useful to remember in the sequel that the terms of this infinite sum are independent with distributions

\[ A_{t,j}(\varepsilon_{t-j}; \alpha_j) \sim ED^*(\theta, \alpha_j \lambda / \alpha_+). \]

The distribution of any partial sum is easily calculated from this result by the convolution formula (5.1.2).

The infinite order EDMA(\infty) processes produce a large class of what appear to be new infinite order moving average models. If we take \( \alpha_j = 0 \) for \( j > q \), the EDMA(\infty) process reduces to a moving average process of finite order \( q \), denoted EDMA(q), as defined in (5.1.3). This model contains several known time-series models as special cases, corresponding to the following marginal distributions: Poisson (Al-Osh and Alzaid, 1987; McKenzie, 1988), gamma (Lewis, 1983; Lewis, McKenzie and Hugus, 1989), inverse Gaussian (Joe, 1996a), negative binomial (Joe, 1996a) and generalized Poisson (Alzaid and Al-Osh, 1993). Note that, in the case of negative binomial margins, the EDMA(q) model is different from that defined by McKenzie (1986) and Al-Osh and Aly (1992).

### 5.2.1 Hidden process

Let us define \( \sigma \)-algebras \( \mathcal{M}_k = \sigma\{\varepsilon_k, \varepsilon_{k-1}, \ldots\} \) and \( \mathcal{M} = \sigma\{\ldots, \varepsilon_1, \varepsilon_0, \varepsilon_1, \ldots\} \). The process \( \{Y_t\} \) defined from \( \{X_t\} \) by

\[ Y_t = E(X_t|\mathcal{M}) = E(X_t|\mathcal{M}_t) \]

is called the hidden process corresponding to \( \{X_t\} \). Noting that

\[ Y_t = \sum_{j=0}^{\infty} \alpha_j \varepsilon_{t-j} \quad (5.2.2) \]

we find that \( Y_t \) is an ordinary strictly stationary infinite order moving average process, except that the innovations \( \varepsilon_t \) follow an exponential dispersion model, rather than Gaussian white noise.
Chapter 5. Stationary models with exponential dispersion model margins

We may then write $X_t$ in the form

$$X_t = \sum_{j=0}^{\infty} (\delta_{jt} + \alpha_j \epsilon_{t-j})$$

$$= \delta_t + Y_t,$$  \hspace{1cm} (5.2.3)

where

$$\delta_{jt} = A_{t,j}(\epsilon_{t-j}; \alpha_j) - \alpha_j \epsilon_{t-j}$$

and

$$\delta_t = \sum_{j=0}^{\infty} \delta_{jt},$$

which both have zero mean, and the $\delta_t$s are independent. When second moments exist, $\delta_{jt}$ and $\epsilon_{t-j}$ are uncorrelated, and so are $\delta_t$ and $Y_t$. We may hence think of $X_t$ as a kind of state space model corresponding to the hidden process $Y_t$.

In the case of a normal marginal distribution, (5.2.2) is a normal moving average process, and $X_t$ is then a normal state space model corresponding to the hidden process $Y_t$. Hence in the case of normal margins, the EDMA(q) processes have a different structure than the classical normal MA(q) processes.

5.2.2 Autocorrelation function

Assume now that $\theta \in \text{int} \Theta$, let $\mu = \lambda \tau(\theta)$ and $\sigma^2 = \lambda \tau'(\theta)$ denote respectively the mean and variance of $X_t$, and let $\mu_\epsilon = \mu/\alpha_+$ and $\sigma^2_\epsilon = \sigma^2/\alpha_+$ denote respectively the mean and variance of $\epsilon_t$. By the properties recorded above, it follows that

$$\text{Var}(\delta_{jt}) = \sigma^2_\epsilon \alpha_j (1 - \alpha_j) \quad \text{and} \quad \text{Var}(\delta_t) = \sigma^2_\epsilon \left( \alpha_+ - \sum_{j=0}^{\infty} \alpha_j^2 \right).$$

The autocovariance function for $Y_t$ of lag $h \geq 0$ is given by

$$\text{Cov}(Y_t, Y_{t+h}) = \sigma^2_\epsilon \sum_{j=0}^{\infty} \alpha_j \alpha_{j+h},$$

and the corresponding autocorrelation function is

$$\rho_Y(h) = \sum_{j=0}^{\infty} \alpha_j \alpha_{j+h} / \sum_{j=0}^{\infty} \alpha_j^2.$$  \hspace{1cm} (5.2.4)
Since $X_t$ and $Y_t$ have the same autocovariances for lags $h > 0$, the autocorrelation function of $X_t$ is, for $h > 0$,

$$\rho_X(h) = \sum_{j=0}^{\infty} \alpha_j \alpha_{j+h} / \alpha_+. \quad (5.2.5)$$

In particular the autocorrelation function of the EDMA($q$) process is $\sum_{j=0}^{q-h} \alpha_j \alpha_{j+h} / \alpha_+ \in [0, (1 + q - h)/(1 + q)]$ for $1 \leq h \leq q$, and 0 otherwise, in agreement with results of Joe (1996a).

We have not investigated the range of the autocorrelation function (5.2.5) in detail, but apparently the range is a subset of that of (5.2.4). Note in particular that $\rho_X(h)$ tends faster to zero as $h$ tends to infinity because of the assumption $\sum \alpha_j < \infty$, compared with (5.2.4) where only $\sum \alpha_j^2 < \infty$ is assumed.

### 5.3 Autoregressive moving average processes

As we know from classical time-series analysis, any causal autoregressive moving average time-series model may be expressed as an infinite order moving average model. This equivalence is now employed to define autoregressive moving average processes with exponential dispersion model margins.

**Definition 5.2** An EDMA($\infty$) process $\{X_t\}$ is said to be an autoregressive moving average process of order $(p, q)$ with margins $ED^*(\theta, \lambda)$, denoted by EDARMA($p, q$), if the coefficients $\alpha_j$ are determined by

$$\sum_{j=0}^{\infty} \alpha_j z^j = \frac{\psi(z)}{\phi(z)}, \quad |z| \leq 1, \quad (5.3.1)$$

where $\phi$ and $\psi$ are polynomials of degree $p$ and $q$, respectively, given by

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p, \quad \psi(z) = 1 + \psi_1 z + \cdots + \psi_q z^q,$$

and where $\phi(z) \neq 0$ for $|z| \leq 1$. 

In the case \( p = 0 \), we have \( \text{EDARMA}(0, q) = \text{EDMA}(q) \), a moving average process of order \( q \) with coefficients \( \alpha_j = \psi_j j = 1, \ldots, q \). When \( q = 0 \) the process is called an autoregressive process of order \( p \), denoted \( \text{EDAR}(p) \), which we study in Section 5.4. The coefficients \( \{\alpha_j\} \) for the \( \text{EDAR}(p) \) process are determined by the relation \( \sum_{j=0}^{\infty} \alpha_j z^j = \phi^{-1}(z) \), and may be found by expanding \( \phi^{-1}(z) \) in a power-series. Similarly, in the general case the coefficients may be obtained by expanding the ratio \( \psi(z)/\phi(z) \) on the right-hand side of (5.3.1) in a power-series, in the usual fashion.

Let us consider the hidden process \( \{Y_t\} \) corresponding to \( \{X_t\} \), assuming that \( X_t \) has finite expectation. From (5.2.2) and (5.3.1), letting \( B \) denote the backward shift operator, we obtain

\[
Y_t = \sum_{j=0}^{\infty} \alpha_j \varepsilon_{t-j} = \left( \sum_{j=0}^{\infty} \alpha_j B^j \right) \varepsilon_t = \frac{\psi(B)}{\phi(B)} \varepsilon_t.
\]

Hence \( Y_t \) is an ordinary ARMA process given by

\[
Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \cdots + \psi_q \varepsilon_{t-q},
\]

in terms of the exponential dispersion model innovations \( \varepsilon_t \). In this case we may hence think of the \( \text{EDARMA}(p, q) \) process \( X_t \) as a state space model corresponding to a hidden ARMA process.

### 5.4 Autoregressive processes

In this section we assume that second moments are finite. When \( X_t \) is an autoregressive \( \text{EDAR}(p) \) process, the hidden process \( Y_t \) is an \( \text{AR}(p) \) process,

\[
Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = \varepsilon_t.
\]

It follows that the autocorrelation function \( \rho_Y(h) \) satisfies the recursion formula, \( \rho_Y(h) = \phi_1 \rho_Y(h-1) + \cdots + \phi_p \rho_Y(h-p) \), just as in the normal case. For the \( \text{EDAR}(p) \) process \( X_t \),
Chapter 5. Stationary models with exponential dispersion model margins 67

a modified recursive relation holds for its autocorrelation function $\rho_X(h)$, as shown in the following theorem. No such recursive relations seem to hold for the autocorrelation function of Joe’s (1996a) autoregressive processes.

**Theorem 5.1** Define $\rho(h) = \rho_X(h)$ for $h \neq 0$ and $\rho(0) = \sum_{j=0}^{\infty} \alpha_j^2 / \alpha_+$. If $X_t$ follows an $EDAR(p)$ process, then $\rho$ satisfies the recursive relation

$$\rho(h) = \phi_1 \rho(h - 1) + \cdots + \phi_p \rho(h - p) \quad \text{for} \quad h > 0.$$  

**Proof** From the equation (5.3.1) with $\psi(z) = 1$, we obtain

$$\sum_{j=0}^{\infty} \alpha_j z^j = 1 + \phi_1 \sum_{j=0}^{\infty} \alpha_j z^{j+1} + \cdots + \phi_p \sum_{j=0}^{\infty} \alpha_j z^{j+p}.$$  

Identifying the coefficients of the term $z^{j+h}$ on both sides, we obtain

$$\alpha_{j+h} = \phi_1 \alpha_{j+h-1} + \cdots + \phi_p \alpha_{j+h-p}.$$  

It follows that

$$\frac{\sum_{j=0}^{\infty} \alpha_j \alpha_{j+h}}{\alpha_+} = \phi_1 \frac{\sum_{j=0}^{\infty} \alpha_j \alpha_{j+h-1}}{\alpha_+} + \cdots + \phi_p \frac{\sum_{j=0}^{\infty} \alpha_j \alpha_{j+h-p}}{\alpha_+},$$  

and so

$$\rho(h) = \phi_1 \rho(h - 1) + \cdots + \phi_p \rho(h - p).$$  

**5.4.1 EDAR(1) process**

We now investigate the EDAR(1) process and compare it with Joe’s (1996a) AR(1) process (5.1.1).

The coefficients of the EDAR(1) process are given by $\alpha_j = \alpha^j$, where $\alpha = \phi_1 \in (0, 1)$. It follows that $\alpha_+ = \sum_{j=0}^{\infty} \alpha^j = 1/(1-\alpha)$, so that $\epsilon_t \sim ED^* \{\theta, (1-\alpha)\lambda\}$. From (5.2.5),
the autocorrelation function is \( \rho_X(h) = \alpha^h/(1 + \alpha) \) for \( h > 0 \). With gamma margins, the process has the form

\[
X_t = \varepsilon_t + A_{t,1}\varepsilon_{t-1} + \cdots + A_{t,j}\varepsilon_{t-j} + \cdots
\]

(5.4.2)

where the \( A_{t,j} \)'s are independent and \( A_{t,j} \sim Beta(\alpha^j\lambda,(1 - \alpha^j)\lambda) \). By comparison, the process (5.1.1) takes the form

\[
(1 - G_tB)X_t = \varepsilon_t
\]

where \( G_t \sim Beta(\alpha\lambda,\alpha\lambda) \). The autocorrelation function of this process has the familiar form \( \alpha^j \). By inversion, we obtain

\[
X_t = \varepsilon_t + G_t\varepsilon_{t-1} + \cdots + G_t^j\varepsilon_{t-j} + \cdots
\]

(5.4.3)

which is different from the expansion (5.4.2). While both models may thus be written as infinite order moving averages with random coefficients, it is clear that there is more than one way to define an autoregressive process.

5.4.2 Partial autocorrelation function

Now we investigate the partial autocorrelation functions for both \( \{X_t\} \) and \( \{Y_t\} \) in the case of an EDAR(\( p \)) process. In order to be able to handle the process \( X_t \) we need the following generalization of the classical definition of the partial autocorrelation function.

**Definition 5.3** The partial autocorrelation function \( \pi_X(\cdot) \) of a stationary time-series \( \{X_t\} \) is defined by

\[
\pi_X(1) = Corr(X_2, X_1) = \rho_X(1),
\]

and for \( h \geq 2 \),

\[
\pi_X(h) = Corr\{X_{h+1} - E(X_{h+1}|X_2,\ldots,X_h), X_1 - E(X_1|X_2,\ldots,X_h)\}.
\]
For the normal distribution the conditional expectation \( E(X_1|X_2,\ldots,X_h) \) is the regression of \( X_1 \) onto the variables \( 1, X_2,\ldots,X_h \), and similarly for \( E(X_{h+1}|X_2,\ldots,X_h) \), so the definition hence reduces to the usual one in this case. It is simple to see that

\[
\pi_X(h) = \text{Corr} \{X_{h+1} - E(X_{h+1}|X_2,\ldots,X_h), X_1\}.
\]

The partial autocorrelation function of \( Y_t \) is the same as for the classical AR\( (p) \) process. In particular, the representation (5.4.1) shows that \( \pi_Y(h) = 0 \) for \( h > p \). While it is difficult to compute the partial autocorrelation function \( \pi_X(h) \) of \( X_t \) for lags \( p \) or less, the following result confirms the autoregressive nature of the EDAR\( (p) \) process by showing that partial autocorrelations for higher lags are zero.

**Theorem 5.2** The partial autocorrelation function \( \pi_X(h) \) of the EDAR\( (p) \) process is zero for lags \( h > p \).

**Proof** By (5.4.4) it is sufficient to prove that for \( h > p \),

\[
\text{Cov} \{X_{h+1} - E(X_{h+1}|X_2,\ldots,X_h), X_1\} = 0.
\]

Combining (5.2.3) and (5.4.1), we obtain

\[
X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = \varepsilon_t + \delta_t - \sum_{j=1}^{p} \phi_j \delta_{t-j}.
\]

It follows that for \( h > p \),

\[
E(X_{h+1}|X_2,\ldots,X_h)
= \sum_{j=1}^{p} \phi_j E(X_{h+1-j} + \mu_e + E(\delta_{h+1}|X_2,\ldots,X_h) - E \left( \sum_{j=1}^{p} \phi_j \delta_{h+1-j}|X_2,\ldots,X_h \right)
= \sum_{j=1}^{p} \phi_j X_{h+1-j} + \mu_e + E(\delta_{h+1}|X_2,\ldots,X_h) - \sum_{j=1}^{p} \phi_j \delta_{h+1-j}
\]
where the last equality is a consequence of \( \sigma(\delta_{h+1-p}, \cdots, \delta_{h}) \subseteq \sigma(X_2, \cdots, X_h) \) because of (5.2.3). Now since
\[
E(\delta_{h+1}|X_2, \cdots, X_h) = E\{E(X_{h+1} - Y_{h+1}|\mathcal{M}, X_2, \cdots, X_h)|X_2, \cdots, X_h\} = E\{E(X_{h+1}|\mathcal{M}) - Y_{h+1}|X_2, \cdots, X_h\} = 0,
\]
it follows that the left hand side of (5.4.5) is
\[
\text{Cov}\{X_{h+1} - E(X_{h+1}|X_2, \cdots, X_h), X_1\} = \text{Cov}\{\delta_{h+1} + \varepsilon_{h+1} - \mu - E(\delta_{h+1}|X_2, \cdots, X_h), X_1\} = \text{Cov}(\delta_{h+1}, \delta_1 + Y_1) - \text{Cov}\{E(\delta_{h+1}|X_2, \cdots, X_h), X_1\} = 0,
\]
completing the proof. \(\Box\)

Finally, we consider the Yule-Walker equations for the EDAR(p) process. First it is easy to show that for \( j = n, \ldots, 1 \) with \( n \leq p \),
\[
\text{Cov}\{X_{n+1} - E(X_{n+1}|X_1, \ldots, X_n), X_j\} = 0.
\]
Taking \( n = p \) and going through arguments similar to the above ones, we obtain
\[
\text{Cov}\{E(X_{p+1}|X_1, \ldots, X_p), X_j\} = \sum_{i=1}^{p} \phi_i \text{Cov}(Y_{p+1-i}, X_j) = \sum_{i=1}^{p} \phi_i \text{Cov}(X_{p+1-i}, X_j).
\]
Therefore, for \( j = p, \ldots, 1 \),
\[
\sum_{i=1}^{p} \phi_i \text{Cov}(X_{p+1-i}, X_j) = \text{Cov}(X_{p+1}, X_j).
\]
Defining $\rho_0 = \sum_j \alpha_j^2 / \alpha_+$, we obtain the following Yule-Walker-like equations,

$$
\begin{pmatrix}
\rho_0 & \rho_X(1) & \rho_X(2) & \cdots & \rho_X(p-1) \\
\rho_X(1) & \rho_0 & \rho_X(1) & \cdots & \rho_X(p-2) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho_X(p-1) & \rho_X(p-2) & \rho_X(p-3) & \cdots & \rho_0
\end{pmatrix}
\begin{pmatrix}
\phi_1 \\
\phi_2 \\
\vdots \\
\phi_p
\end{pmatrix}
= 
\begin{pmatrix}
\rho_X(1) \\
\rho_X(2) \\
\vdots \\
\rho_X(p)
\end{pmatrix}
$$

Unfortunately, it is not completely straightforward to estimate the coefficients $\phi_1, \ldots, \phi_p$ from this equation based on sample moments because of the presence of $\rho_0$. We have

$$
\text{Var} (X_t - Y_t) = \sigma^2 (1 - \rho_0),
$$

allowing us to estimate $\rho_0$ from the variance of $X_t$ and that of $X_t - Y_t$, but of course $Y_t$ is not observable.

### 5.5 Truncation error

We now consider briefly, for the EDMA(∞) process, the error incurred by truncating the infinite sum (5.2.1) to a finite number of terms. Let us denote the truncated series by

$$
X_t^N = \sum_{j=0}^N A_{t,j} (\epsilon_{t-j}; \alpha_j).
$$

Then

$$
X_t - X_t^N \sim ED^* \left( \theta, \lambda \frac{\alpha_+ - \alpha(N)}{\alpha_+} \right),
$$

where $\alpha(N) = \sum_0^N \alpha_j$. When the second moments are finite, the corresponding mean squared error is

$$
E \left( X_t - X_t^N \right)^2 = \sigma^2 \frac{\alpha_+ - \alpha(N)}{\alpha_+} + \mu^2 \left( \frac{\alpha_+ - \alpha(N)}{\alpha_+} \right)^2
\leq (\sigma^2 + \mu^2) \frac{\alpha_+ - \alpha(N)}{\alpha_+}.
$$
Chapter 5. Stationary models with exponential dispersion model margins

To obtain $E(X_t - X_t^N)^2 \leq \eta$, where $\eta$ is a given tolerance level, it is hence enough to choose $N$ so large that

$$\alpha(N) \geq \alpha_+ \left(1 - \frac{\eta}{\sigma^2 + \mu^2}\right).$$

(5.5.2)

In the special case of an EDARMA($p,q$) process we have $\alpha_+ = (1 + \psi_1 + \cdots + \psi_q)/(1 - \phi_1 - \cdots - \phi_p)$, and $N$ may be determined from (5.5.2) by a simple search. For the EDAR(1) process with $\alpha_j = \alpha^j$ we obtain $\alpha_+ = 1/(1 - \alpha)$, and equation (5.5.2) becomes

$$\alpha^N \leq \frac{\eta}{\sigma_\varepsilon^2 + \mu^2}$$

with solution

$$N \geq \frac{\log \eta - \log(\sigma_\varepsilon^2 + \mu^2)}{\log \alpha}.$$

5.6 Further questions

The ideas of this chapter may be extended in several directions, and many issues such as parameter estimation and inference remain to be addressed.

One may construct state space models in various ways from the processes constructed above. To take an example, let $Z_t$ be a given stationary EDMA($\infty$) process with marginal distribution $ED^*(\theta, \lambda/\alpha)$. If $\eta_t$ is independent of the $Z_t$ and has distribution $ED^*(\theta, \delta)$, then the process defined by

$$X_t = A_t(Z_t; \alpha) + \eta_t$$

(5.6.3)

has marginal distribution $ED^*(\theta, \lambda + \delta)$. If $X_t$ is observed and $Z_t$ is latent, we have a simple stationary state-space model. A variation of this idea is obtained by letting $X_t = Z_t + \eta_t$, where $Z_t$ has inverse Gaussian margin and $\eta_t$ follows a reciprocal inverse Gaussian distribution. This gives a reciprocal inverse Gaussian distribution for $Y_t$ (c.f. Seshadri 1992).
A moving average process of infinite order, MA(∞), with normal margins

\[ X_t = \sum_{j=0}^{\infty} \alpha_j \varepsilon_{t-j}, \quad (5.6.4) \]

which, as mentioned earlier, is also known as a convolution process, has a variety of applications in many branches of sciences. An important question is that of deconvolution, whereby the sequence \( \{\varepsilon_t\} \) is recovered from the observed process \( \{X_t\} \) and the parameters \( \{\alpha_j\} \). There is a large literature available on approaches to the deconvolution of (5.6.4) for normal systems (c.f. Dimri 1992, Jansson 1984 and references therein). We may hence ask if it is possible to deconvolve our model

\[ X_t = \sum_{j=0}^{\infty} A_{t,j}(\varepsilon_{t-j}; \alpha_j), \]

that is, to recover \( \{\varepsilon_t\} \) from the observed process \( \{X_t\} \). A possible solution is to estimate the hidden process \( Y_t \) by means of filtering methods. Since \( Y_t \) is of the form (5.6.4), one may then adapt existing deconvolution methods to the filtered version of \( Y_t \).
Chapter 6

BLUP, Kalman filter and smoother

This chapter makes the key technical preparations for results to be developed in the next four chapters. We start with the definition of the best linear unbiased predictor (BLUP) and investigate some basic properties of BLUP. We then develop the Kalman filter and smoother, originally proposed by Kalman (1960) and Kalman and Bucy (1961), under a structure more general than the classic linear model. The reason for studying such a generalization is that the latent process of the state space model considered in Chapter 7 has a non-linear stochastic structure, indeed described by the thinning operation (c.f. Section 2.5), and it requires the Kalman filter and smoother for a broader framework than the classical one presented in, for example, Harvey (1981). In the literature, there are some generalizations of the Kalman filter theory to non-linear models, for instance, that of Kitagawa (1987), Spall (1988) and of Naik-Nimbalkar and Rajarshi (1995). But none has dealt with the theory under a stochastic functional model (c.f. Equation (7.2.2)) yet, as far as we know.

6.1 Some general results for linear prediction

Throughout this section we consider random vectors $X$, $Y$ and $Z$ with finite second moments, and denote their means, variances and covariances by

$$
\begin{align*}
E \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} &= \begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix} \\
\text{Var} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} &= \begin{pmatrix} \Sigma_X & \Sigma_{XY} & \Sigma_{XZ} \\ \Sigma_{YX} & \Sigma_Y & \Sigma_{YZ} \\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_Z \end{pmatrix}
\end{align*}
$$
We define the linear predictor (or the least squares prediction) of \( X \) given \( Y \) by

\[
X|Y \sim [m_{X|Y}; C_{X|Y}],
\]

where \( m_{X|Y} \) and \( C_{X|Y} \), called respectively the predictor or BLUP and the mean square error, are given by

\[
m_{X|Y} = \mu_X + \Sigma_{XY} \Sigma_Y^{-1} (Y - \mu_Y) \tag{6.1.1}
\]

and

\[
C_{X|Y} = \Sigma_X - \Sigma_{XY} \Sigma_Y^{-1} \Sigma_Y X. \tag{6.1.2}
\]

For the true moments we use Greek letters \( \mu \) and \( \Sigma \), while the linear predictor is denoted by the Roman letters \( m \) and \( C \), with appropriate subscripts to denote the variables in question.

When no conditioning is involved, the linear predictor gives the true mean and variance. If \( X \) and \( Y \) are jointly multivariate normally distributed then \( m_{X|Y} \) and \( C_{X|Y} \) coincide with the conditional mean and variance, which is not generally the case otherwise. Furthermore, \( m_{X|Y} \) is that linear combination of the elements of \( Y \) that minimizes the mean square error for prediction of \( X \), i.e. the best linear unbiased predictor (BLUP) (c.f. Brockwell and Davis 1991, p. 64).

In general, the linear predictor behaves much like the conditional mean and variance do in the context of the multivariate normal distribution, as the following list of properties shows.

1. The mean square error may be written in the form

\[
C_{X|Y} = E\{(X - m_{X|Y})(X - m_{X|Y})^\top\} = \text{Var}(X - m_{X|Y}).
\]

Thus, \( C_{X|Y} \) resembles \( \text{Var}\{X - E(X|Y)\} = E\{\text{Var}(X|Y)\} \).
Chapter 6. BLUP, Kalman filter and smoother

2. For suitable matrices or scalars $\alpha$ and $\beta$, we have

$$(\alpha X + \beta|Y \sim [\alpha m_{X|Y} + \beta; \alpha C_{X|Y} \alpha^T].$$

3. For the multivariate normal distribution, the conditional mean is a linear function of $Y$ and the conditional variance is not functionally dependent on the value of $Y$. Similarly for linear predictors, we have that if $E(X|Y)$ is linear in $Y$ then

$$X|Y \sim [E(X|Y); E\{\text{Var}(X|Y)\}]. \quad (6.1.3)$$

4. The prediction error $X - m_{X|Y}$ is uncorrelated with $Y$,

$$\text{Cov}(X - m_{X|Y}, Y) = \text{Cov}(X, Y) - \Sigma_{XY} \Sigma_Y^{-1} \text{Var}(Y) = 0,$$

an important and often used property.

5. If $Y$ and $Z$ are uncorrelated, then $X|(Y, Z) \sim [m_{X|Y, Z}; C_{X|Y, Z}]$, with

$$m_{X|Y, Z} = \mu_X + \Sigma_{XY} \Sigma_Y^{-1} (Y - \mu_Y) + \Sigma_{XZ} \Sigma_Z^{-1} (Z - \mu_Z) \quad (6.1.4)$$

and

$$C_{X|Y, Z} = \Sigma_X - \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX} - \Sigma_{XZ} \Sigma_Z^{-1} \Sigma_{ZX}. \quad (6.1.5)$$

The following results show that the joint, marginal and conditional linear predictors behave much like the joint, marginal and conditional means and variances for the multivariate normal distribution.

**Theorem 6.1** Let $X$, $Y$ and $Z$ be random vectors with finite second moments. Then the joint predictor of $X, Y$ given $Z$ is given by

$$\left( \begin{array}{c} X \\ Y \end{array} \right)|Z \sim \begin{bmatrix} \left( \begin{array}{c} m_{X|Z} \\ m_{Y|Z} \end{array} \right); \left( \begin{array}{cc} C_{X|Z} & C_{XY|Z} \\ C_{YX|Z} & C_{Y|Z} \end{array} \right) \end{bmatrix} \quad (6.1.6)$$
where \( C_{XY|Z} = \Sigma_{XY} - \Sigma_{XZ} \Sigma_{Z}^{-1} \Sigma_{ZY} \). The linear predictor of \( X \) given \( Y \) and \( Z \) (the "conditional" linear predictor) is given by

\[
X \left| \begin{pmatrix} Y \\ Z \end{pmatrix} \right. \sim \left[ m_{X|Z} + C_{XY|Z} C_{Y|Z}^{-1} \left( Y - m_{Y|Z} \right) ; C_{X|Z} - C_{XY|Z} C_{Y|Z}^{-1} C_{XY|Z} \right]. \tag{6.1.7}
\]

**Proof** The joint predictor (6.1.6) is obtained directly from the definition. To show (6.1.7), we note that the vectors \((Y, Z)\) have the same span as \((Y - m_{Y|Z}, Z)\), and that the two components of the latter vector are uncorrelated. Therefore, (6.1.7) follows from (6.1.4) and (6.1.5) by noting that \( \text{Cov}(X, Y - m_{Y|Z}) = C_{XY|Z} \) and \( \text{Var}(Y - m_{Y|Z}) = C_{Y|Z} \). Hence, the linear predictor of \( X \) given \( Y \) and \( Z \) is the same as that given \((Y - m_{Y|Z}), Z\). Using (6.1.4) we find

\[
m_{X|Y^Z} = \mu_X + \left( \text{Cov}(X, Y - m_{Y|Z}), \text{Cov}(X, Z) \right) \begin{pmatrix} C_{Y|Z} & 0 \\ 0 & \Sigma_{Z} \end{pmatrix} \begin{pmatrix} Y - m_{Y|Z} \\ Z - \mu_Z \end{pmatrix}
\]

\[
= \mu_X + \left( C_{XY|Z} C_{Y|Z}^{-1}, \Sigma_{XZ} \Sigma_{Z}^{-1} \right) \begin{pmatrix} Y - m_{Y|Z} \\ Z - \mu_Z \end{pmatrix}
\]

\[
= m_{X|Z} + C_{XY|Z} C_{Y|Z}^{-1} \left( Y - m_{Y|Z} \right)
\]

and by (6.1.5)

\[
C_{X|Y^Z} = \Sigma_X - \left( C_{XY|Z} C_{Y|Z}^{-1}, \Sigma_{XZ} \Sigma_{Z}^{-1} \right) \begin{pmatrix} C_{Y|Z} \\ \Sigma_{ZX} \end{pmatrix}
\]

\[
= \Sigma_X - C_{XY|Z} C_{Y|Z}^{-1} C_{YX|Z} - \Sigma_{XZ} \Sigma_{Z}^{-1} \Sigma_{ZX}
\]

\[
= C_{X|Z} - C_{XY|Z} C_{Y|Z}^{-1} C_{YX|Z}.
\]

This completes the proof. \( \square \)

**Corollary 6.1** If \( X \) and \( Y \) are uncorrelated given \( Z \) and either \( E(X|Z) \) or \( E(Y|Z) \) is
linear in $Z$, then

$$X \left| \begin{pmatrix} Y \\ Z \end{pmatrix} \right. \equiv X|Z \sim [m_{X|Z}; C_{X|Z}].$$  \hspace{1cm} (6.1.8)

**Proof** From (6.1.7), if $C_{XY|Z} = 0$ then (6.1.8) is true. Hence we only need to show $C_{XY|Z} = 0$ under the given conditions. By symmetry, it is sufficient to prove the result in the case $E(X|Z) = \alpha Z + \beta$. Since $\text{Cov}(X, Y|Z) = 0$, we obtain

$$\Sigma_{XY} = \text{Cov}(X, Y)$$
$$= E\{\text{Cov}(X, Y|Z)\} + \text{Cov}\{E(X|Z), E(Y|Z)\}$$
$$= \text{Cov}\{E(X|Z), E(Y|Z)\}$$
$$= \text{Cov}(\alpha Z + \beta, Y)$$
$$= \alpha \Sigma_{ZY}$$

and

$$\Sigma_{XZ} = \text{Cov}(X, Z) = \text{Cov}[E(X|Z), Z]$$
$$= \text{Cov}(\alpha Z + \beta, Z) = \alpha \Sigma_{Z},$$

implying that

$$C_{XY|Z} = \Sigma_{XY} - \Sigma_{XZ} \Sigma_{Z}^{-1} \Sigma_{ZY} = \alpha \Sigma_{ZY} - \alpha \Sigma_{Z} \Sigma_{Z}^{-1} \Sigma_{ZY} = 0.$$  \hspace{1cm} (6.1.9)

This proves the corollary. $\Box$

**Theorem 6.2** Assume $Y|Z \sim [m_{Y|Z}; C_{Y|Z}]$ and $X|Y, Z \sim [\alpha Y + \beta Z + \gamma; C_{X|Y,Z}]$. The joint predictor of $X, Y$ given $Z$ (called the combination) is

$$\begin{pmatrix} X \\ Y \end{pmatrix} \mid Z \sim \begin{pmatrix} \alpha m_{Y|Z} + \beta Z + \gamma \\ m_{Y|Z} \end{pmatrix}; \begin{pmatrix} C_{X|Y,Z} + \alpha C_{Y|Z} \alpha^T & \alpha C_{Y|Z} \\ C_{Y|Z} \alpha^T & C_{Y|Z} \end{pmatrix}. \hspace{1cm} (6.1.9)$$

In particular, $X|Z \sim [\alpha m_{Y|Z} + \beta Z + \gamma; C_{X|Y,Z} + \alpha C_{Y|Z} \alpha^T]$. 
Proof The formula for the combination (6.1.9) may be shown by noting that

\[(\alpha, \beta) = (\Sigma_{XY}, \Sigma_{XZ}) \begin{pmatrix} \Sigma_Y & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_Z \end{pmatrix}^{-1}\]

and that

\[\gamma = \mu_X - \alpha \mu_Y - \beta \mu_Z.\]

Hence

\[\Sigma_{XY} = \alpha \Sigma_Y + \beta \Sigma_{ZY}\]

and

\[\Sigma_{XZ} = \alpha \Sigma_{YZ} + \beta \Sigma_Z.\]

From the definition (6.1.1) we obtain

\[m_{X|Z} = \mu_X + (\alpha \Sigma_{YZ} + \beta \Sigma_Z) \Sigma_Z^{-1} (Z - g_Z)\]

\[= \mu_X + \alpha (m_{Y|Z} - \mu_Y) + \beta (Z - \mu_Z)\]

\[= \alpha m_{Y|Z} + \beta Z + \gamma.\]

Similarly it is seen from (6.1.2) that

\[C_{XY|Z} = \Sigma_{XY} - \Sigma_{XZ} \Sigma_Z^{-1} \Sigma_{ZY}\]

\[= \alpha \Sigma_Y + \beta \Sigma_{ZY} - (\alpha \Sigma_{YZ} + \beta \Sigma_Z) \Sigma_Z^{-1} \Sigma_{ZY}\]

\[= \alpha C_{Y|Z}.\] (6.1.10)

The last part of (6.1.9) is obtained by using \(\beta = (\Sigma_{XZ} - \alpha \Sigma_{YZ}) \Sigma_Z^{-1}\), giving

\[C_{X|YZ} + \alpha C_{Y|Z} \alpha^T = \Sigma_X - \alpha \Sigma_{YX} - \beta \Sigma_{ZX} + \alpha C_{Y|Z} \alpha^T\]

\[= \Sigma_X - \Sigma_{XZ} \Sigma_Z^{-1} \Sigma_{ZX} - \alpha \left(C_{YX|Z} - C_{Y|Z} \alpha^T\right).\]

According to (6.1.10) the last term of this expression vanishes, and hence \(C_{X|Z} = \Sigma_X - \Sigma_{XZ} \Sigma_Z^{-1} \Sigma_{ZX} = C_{X|YZ} + \alpha C_{Y|Z} \alpha^T.\) \(\square\)
6.2 General forms of Kalman filter and smoother

This section presents both Kalman filter and smoother under a general framework where only the mean structure is assumed to be linear, extending the theory given by, for example, Harvey (1981).

Assume that two processes of random vectors \( \{Y_t\} \in \mathcal{R}^p \) and \( \{\theta_t\} \in \mathcal{R}^q \) follow the comb structure defined as follows:

Given \( \theta_t \), \( Y_t \) is uncorrelated with the rest of the \( Y_t \)'s and \( \theta_t \) is first-order Markov.

\[ \text{(6.2.1)} \]

Figure 6.1 illustrates the comb structure at time \( t \).

\[ \text{Figure 6.1: The comb structure.} \]

\[ \begin{align*}
Y_{t-1} & \quad Y_t & \quad Y_{t+1} \\
\bigcirc & \quad \bigcirc & \quad \bigcirc \\
\theta_{t-1} & \quad \theta_t & \quad \theta_{t+1}
\end{align*} \]

In addition, we assume that the processes satisfy the following moment structure:

\[ \text{E} (Y_t | \theta_t) = A_t \theta_t + a_t, \quad \text{and} \quad \text{Var} (Y_t | \theta_t) = W_t(\theta_t) + W^0_t, \quad (6.2.2) \]

and

\[ \text{E} (\theta_t | \theta_{t-1}) = B_t \theta_{t-1} + b_t, \quad \text{and} \quad \text{Var} (\theta_t | \theta_{t-1}) = D_t(\theta_{t-1}) + D^0_t \quad (6.2.3) \]

where the mean structure is assumed to be linear, both \( W_t(\theta_t) \) and \( D_t(\theta_t) \) are matrices of functions in \( \theta_t \) entrywise with constant terms being expressed separately. Such expression
for the conditional variances make the theories developed later in Chapter 7 and 8 easier to implement. Let \( \mathbb{E}\{\mathbf{W}_t(\theta_t)\} = \mathbf{W}_t, \mathbb{E}\{\mathbf{D}_t(\theta_t)\} = \mathbf{D}_t \) and \( \mathbb{E}\theta_t = \tau_t \).

Let \( \mathbf{Y}^t \) be the set of the first \( t \) vectors \( \mathbf{Y}_1, \ldots, \mathbf{Y}_t \). The Kalman filter is the BLUP for \( \theta_t \) given \( \mathbf{Y}^t \), and the Kalman smoother is the BLUP for \( \theta_t \) based on the total of all observations \( \mathbf{Y}^n \), say. Both can be calculated recursively by the following theorems, respectively.

**Theorem 6.3** Under the comb structure and the above moment structure, for given the prediction at time \( t - 1 \), \( \theta_{t-1}|\mathbf{Y}^{t-1} \sim [\mathbf{m}_{t-1}; \mathbf{C}_{t-1}] \), the Kalman filter is given as follows. First compute

\[
\theta_t|\mathbf{Y}^{t-1} \sim [\mathbf{B}_t \mathbf{m}_{t-1} + \mathbf{b}_t; \mathbf{H}_t] \quad \text{and} \quad \mathbf{Y}_t|\mathbf{Y}^{t-1} \sim [\mathbf{f}_t; \mathbf{Q}_t]
\]

where

\[
\mathbf{H}_t = \mathbf{D}_{t-1} + \mathbf{D}_t^0 + \mathbf{B}_t \mathbf{C}_{t-1} \mathbf{B}_t^T
\]

and

\[
\mathbf{f}_t = \mathbf{A}_t (\mathbf{B}_t \mathbf{m}_{t-1} + \mathbf{b}_t) + \mathbf{a}_t, \quad \mathbf{Q}_t = \mathbf{F}_{t-1} + \mathbf{F}_t^0 + \mathbf{A}_t \mathbf{B}_t \mathbf{C}_{t-1} \mathbf{B}_t^T \mathbf{A}_t^T.
\]

Then update the prediction of \( \theta_t \) given \( \mathbf{Y}^t \) by

\[
\theta_t|\mathbf{Y}^t \sim [\mathbf{m}_t; \mathbf{C}_t]
\]

where

\[
\mathbf{m}_t = \mathbf{B}_t \mathbf{m}_{t-1} + \mathbf{b}_t + \mathbf{H}_t^T \mathbf{A}_t^T \mathbf{Q}_t^{-1} (\mathbf{Y}_t - \mathbf{f}_t) \quad \text{and} \quad \mathbf{C}_t = \mathbf{H}_t - \mathbf{H}_t^T \mathbf{A}_t^T \mathbf{Q}_t^{-1} \mathbf{A}_t \mathbf{H}_t.
\]

Note that the Kalman filtering is done by a forward updating procedure.

**Theorem 6.4** Now assume that all Kalman filter \( \mathbf{m}_t \) and \( \mathbf{C}_t \) are given. Under the comb structure and the moment structure given above, the Kalman smoother can be computed
backwards as follows. Given the prediction at time $t+1$, $\theta_{t+1}|Y^n \sim [m^*_{t+1}; C^*_{t+1}]$, the Kalman smoother and the corresponding error are given by

$$m^*_t = m_t + C_tB_{t+1}H^{-1}_{t+1}(m^*_{t+1} - B_{t+1}m_t - b_{t+1})$$

and

$$C^*_t = C_t - C_tB_{t+1}H^{-1}_{t+1}B_{t+1}C_t + C_tB_{t+1}H^{-1}_{t+1}C^*_{t+1}H^{-T}_{t+1}B_{t+1}C_t.$$ 

The recursion starts with $t = n$ by taking $m^*_n = m_n$ and $C^*_n = C_n$.

We now prove the two theorems.

**Proof** of Theorem 6.3: It follows from the comb structure that

$$E(Y_t|\theta_{t-1}) = A_tB_t\theta_{t-1} + A_tb_t + a_t, \quad (6.2.4)$$

$$\text{Var}(Y_t|\theta_{t-1}) = E\{\text{Var}(Y_t|\theta_{t-1})\} + \text{Var}\{E(Y_t|\theta_{t-1})\}$$

$$= F_t(\theta_{t-1}) + F^0_t$$

where

$$F_t(\theta_{t-1}) = E\{W_t(\theta_t)|\theta_{t-1}\} + A_tD_t(\theta_{t-1})A_t^T,$$

and $F^0_t = W^0_t + A_tD^0_tA_t^T$

and

$$\text{Cov}(Y_t, \theta_t|\theta_{t-1}) = \text{Cov}\{E(Y_t|\theta_t), \theta_t|\theta_{t-1}\} = A_tCov(\theta_t, \theta_t|\theta_{t-1})$$

$$= A_tD_t(\theta_{t-1}) + A_tD^0_t.$$ 

Since the both conditional expectations $E(Y_t|\theta_{t-1})$ and $E(\theta_t|\theta_{t-1})$ are linear in $\theta_{t-1}$, by the third property (6.1.3), we immediately obtain

$$\left( \begin{array}{c} Y_t \\ \theta_t \end{array} \right)|\theta_{t-1} \sim \left( \begin{array}{c} A_tB_t\theta_{t-1} + A_tb_t + a_t \\ B_t\theta_{t-1} + b_t \end{array} \right); \left( \begin{array}{cc} \bar{F}_{t-1} + F^0_t & A_tD_{t-1} + A_tD^0_t \\ D^T_tA^T_t + D^0_tA^T_t & D_{t-1} + D^0_t \end{array} \right)$$

$$\left( \begin{array}{c} \bar{F}_{t-1} + F^0_t \\ \bar{D}_{t-1} + D^0_t \end{array} \right) \quad (6.2.5)$$
where

\[
\bar{F}_{t-1} = E \{ F_t(\theta_{t-1}) \} = E \{ E \{ W_t(\theta_t) | \theta_{t-1} \} \} + A_t E \{ D_t(\theta_{t-1}) \} A_t^T = \bar{W}_t + A_t \bar{D}_{t-1} A_t^T.
\]

Now assume that we currently know

\[
\theta_{t-1} | Y^{t-1} \sim [m_{t-1}; C_{t-1}].
\]

Applying Theorem 6.2, we obtain the predictions of \((Y_t, \theta_t)\) given \(Y^{t-1}\) as follows,

\[
\left( \begin{array}{c} Y_t \\ \theta_t \end{array} \right) | Y^{t-1} \sim \left[ \bar{m}_{t|t-1}; \bar{C}_{t|t-1} \right]
\]

(6.2.6)

where the predictor is

\[
\bar{m}_{t|t-1} = \left( \begin{array}{c} A_t \\ I \end{array} \right) B_t m_{t-1} + \left( \begin{array}{c} A_t \\ I \end{array} \right) b_t + \left( \begin{array}{c} a_t \\ 0 \end{array} \right),
\]

and the prediction error is

\[
\bar{C}_{t|t-1} = \left( \begin{array}{cc} \bar{F}_{t-1} + F_t^0 & A_t \bar{D}_{t-1} + A_t D_t^0 \\ \bar{D}_{t-1}^T A_t^T + D_t^0 A_t^T & \bar{D}_{t-1} + D_t^0 \end{array} \right) + \left( \begin{array}{ccc} A_t B_t C_{t-1} B_t^T A_t^T & A_t B_t C_{t-1} B_t^T \\ B_t C_{t-1} B_t^T A_t^T & B_t C_{t-1} B_t^T \end{array} \right)
\]

In particular,

\[
\theta_t | Y^{t-1} \sim [B_t m_{t-1} + b_t; H_t]
\]

where

\[
H_t = \bar{D}_{t-1} + D_t^0 + B_t C_{t-1} B_t^T
\]

and

\[
Y_t | Y^{t-1} \sim [f_t; Q_t]
\]

where

\[
f_t = A_t (B_t m_{t-1} + b_t) + a_t, \quad Q_t = \bar{F}_{t-1} + F_t^0 + A_t B_t C_{t-1} B_t^T A_t^T.
\]
Now we update the prediction of $\theta_t$ given a new vector $Y_t$ and $Y^{t-1}$ from the expression (6.2.6). It follows from Theorem 6.1 that

$$\theta_t|Y^t \sim [m_t; C_t]$$

(6.2.7)

where

$$m_t = B_t m_{t-1} + b_t + \left( \overline{D}_{t-1} + D_t^0 + B_t C_{t-1} B_t^T \right)^T A_t$$

$$\left( \overline{F}_{t-1} + F_t^0 + A_t B_t C_{t-1} B_t^T A_t^T \right)^{-1} (Y_t - f_t)$$

$$= B_t m_{t-1} + b_t + H_t^T A_t^{-1} Q_t^{-1} (Y_t - f_t)$$

(6.2.8)

and

$$C_t = H_t - H_t^T A_t^{-1} Q_t^{-1} A_t H_t,$$

(6.2.9)

so proof is complete. □

**Proof of Theorem 6.4:** Suppose that we know the current Kalman smoother

$$\theta_{t+1}|Y^n \sim [m_{t+1}^*; C_{t+1}^*]$$

(6.2.10)

and the entire sequence of the Kalman filters

$$\theta_t|Y^t \sim [m_t; C_t].$$

(6.2.11)

Note that for given $\theta_t$, $\theta_{t+1}$ is independent of $Y^t$ by the assumption (6.2.1) of the comb structure, then from (6.2.5),

$$\theta_{t+1}|Y^t, \theta_t \overset{d}{=} \theta_{t+1}|\theta_t \sim \left[ B_{t+1} \theta_t + b_{t+1}; \overline{D}_{t} + D_{t+1}^0 \right].$$

(6.2.12)

By Theorem 6.2 and both (6.2.11) and (6.2.12),

$$\begin{pmatrix} \theta_t \\ \theta_{t+1} \end{pmatrix} \left| Y^t \right. \sim \begin{pmatrix} m_t \\ B_{t+1} m_t + b_{t+1} \end{pmatrix}; \begin{pmatrix} C_t & C_t B_{t+1}^T \\ B_{t+1} C_t & H_{t+1} \end{pmatrix}.$$  

(6.2.13)
Furthermore, by Theorem 6.1 and (6.2.13), we have

\[ \theta_t | \theta_{t+1}, Y^t \sim \left[ m_t + C_t B_{t+1}^T H_{t+1}^{-1} (\theta_{t+1} - B_{t+1} m_t - b_{t+1}); C_t - C_t B_{t+1}^T H_{t+1}^{-1} B_{t+1} C_t \right] . \]

(6.2.14)

Similar to the calculation of the equation (6.2.4), it is easy to prove that for each \( s \geq 1 \), \( E(Y_{t+s} | \theta_{t+1}) \) is linear in \( \theta_{t+1} \). From the Corollary 6.1,

\[ \theta_t | \theta_{t+1}, Y^n \overset{d}{=} \theta_t | \theta_{t+1}, Y^t \]

and again from Theorem 6.2 and (6.2.10),

\[
\begin{pmatrix}
\theta_t \\
\theta_{t+1}
\end{pmatrix}
\left| Y^n \right. \sim \left[
\begin{pmatrix}
m_t^* \\
m_{t+1}^*
\end{pmatrix};
\begin{pmatrix}
C_t^* & C_t B_{t+1}^T H_{t+1}^T C_t^* + C_t B_{t+1}^T H_{t+1}^T B_{t+1} C_t
\end{pmatrix}
\right].
\]

(6.2.15)

where \( m_t^* \) and \( C_t^* \) are the Kalman smoother and the corresponding prediction error given by

\[ m_t^* = m_t + C_t B_{t+1}^T H_{t+1}^{-1} (m_{t+1}^* - B_{t+1} m_t - b_{t+1}) \]

(6.2.16)

and

\[ C_t^* = C_t - C_t B_{t+1}^T H_{t+1}^{-1} B_{t+1} C_t + C_t B_{t+1}^T H_{t+1}^{-1} C_{t+1}^* H_{t+1}^{-1} B_{t+1} C_t. \]

(6.2.17)

We complete the proof. \( \square \)
Chapter 7

State space models with stationary latent process

This chapter presents a class of state space models for one-dimensional time series of observations, assuming that the latent processes are the stationary AR(1) models of Joe (1996a) with exponential dispersion model margins and without time-varying covariates. The development of the Kalman filter and smoother under the models is not a straightforward application of the classic theory (c.f. Harvey, 1981) because the latent process is modelled by the thinning operation (c.f. Section 2.5) rather than the linear model in the classic context. Hence we have to derive both predictors, applying the theory of Chapter 6. Estimates of parameters are obtained through the so-called Kalman estimating equation (KEE), corresponding to a modified EM-algorithm with the E-step approximated by the Kalman smoother. The asymptotic normality of the estimates is proven rigorously. Also model checking diagnosis is proposed via residual analysis. We illustrate the KEE method on the analysis of the polio incidence data (Zeger, 1988) that were previously analyzed by both Zeger (1988) and Chan and Ledolter (1995).

7.1 Introduction

For simplicity, in this chapter we concentrate only on one-dimensional time series data; however, most results in this chapter may be generalized to the multivariate setting in a similar way to what we discuss in Chapter 8.

Let \( \{ Y_t, 1 \leq t \leq n \} \) be a time series of observations associated with a set of \( k \)-dimensional time-varying covariates, \( \{ x_t, 1 \leq t \leq n \} \). The objective of the analysis is to
model the observations $Y_t$ as a function of the covariates $x_t$, with a certain correlation structure introduced via an unobserved latent process $\theta_t$. The class of state space models considered in the present chapter is structured as follows: the latent process $\theta_t$ follows the stationary first-order autoregressive model AR(1) of Joe (1996a) with a margin in the family of exponential dispersion models and without time-varying covariates. Given the latent process $\theta_t$, the observations $Y_t$ are conditionally independent with distributions in the Tweedie class (c.f. Section 2.3). Under such a model, the problem of estimation is usually difficult because the latent process results in a complex likelihood function.

Cox (1981) characterized the regression models for time-dependent data into two classes: "parameter-driven models" and "observation-driven models". The class of models defined above may be viewed as of the parameter-driven type. Although observation-driven models have some advantages from a computational point of view, (c.f. Harvey and Fernandes 1989), the parameter-driven model is conceptually more attractive and has therefore attracted a lot of attention in recent years. It is noteworthy that the parameter-driven model in Cox (1981) tends to stress the latent structure feature, which does not fully reflect the characteristic of our models in which the observed part is also (and may be more) important. We prefer and use the term of state space model instead of parameter-driven model, although some authors such as Zeger (1988) and Chan and Ledolter (1995) have used the term of parameter-driven model for their models that are similar to ours.

State space models or dynamical models have been developed extensively since their introduction in the context of Gaussian time-series modelling by Harrison and Stevens (1976) (see also West and Harrison 1989). The models provide a convenient conceptual framework for longitudinal data because the latent process may be interpreted as an unobserved intensity process, often providing a useful interpretation of the data-generating mechanism. An additional important feature of these models is that the latent process

A major obstacle to the development of state space models for non-Gaussian data is the that likelihood-based methods may require computationally intensive methods such as Monte Carlo Markov chain simulation (see Chan and Ledolter, 1995). A computationally simpler alternative to state space models is the quasi-likelihood approach of Liang and Zeger (1986), which accommodates a wide range of data by using only second moment assumptions, and is robust against misspecification of the correlation structure, in the sense of producing asymptotically correct standard errors. In particular, models for serially correlated discrete observations have been studied by Zeger (1988), Kaufmann (1987), Stiratelli, Laird and Ware (1984), Zeger, Liang and Self (1985) and Zeger and Qaqish (1988). Further contributions and references may be found in Zeger and Liang (1986), Diggle, Liang and Zeger (1994) and Dwyer et al. (1991).

The less than full efficiency of the quasi-likelihood estimators is not a major concern, but the lack of full distributional assumptions often makes the method difficult to interpret, and leads to an over-reliance on estimated regression coefficients and their standard errors. Many of the state space models for non-Gaussian data proposed in the literature share with the quasi-likelihood approach the problem that the probability model is not fully specified. Similarly, in the dynamic generalized linear model of West, Harrison and Migon (1985), the latent process is specified in terms of first and second moments only. In the exponential discount model (or the Bayesian steady forecasting model) (c.f. Smith 1979, Key and Godolphin 1981, West 1986 and Smith and Miller 1986), the state process is vaguely described via predictive distributions.

By contrast, the state space models proposed in the present chapter and Chapter 8
are based on a fully specified class of parametric distributions, opening up some new possibilities for detailed modelling of the longitudinal structure of the data. In particular, we allow time-varying covariates to enter either via the latent process or via the observation model. These two types of covariates are called respectively long-term and short-term covariates. The long-term covariates thus feed long-term variations (relatively speaking) via the latent process, whereas the short-term covariates feed variations that are short-term relative to the latent process. As a consequence, we stress the need for checking both the observed and unobserved parts of the model by means of residual analysis, something which may be difficult for partly specified models.

Different structures for latent process gives rise to different interpretations of modelling as well as different implementations for estimation procedures. Therefore, we present our development of the state space models in two separate chapters. In the present chapter, we focus on the state space models with stationary latent process in the absence of long-term covariates. Later in Chapter 8 we consider state space models with more general latent structures including time-varying covariates, leading us to non-stationary latent processes.

With regard to the state space models with stationary latent process, Zeger (1988) as well as Chan and Ledolter (1995) propose two different methods for statistical inference on regression parameters. Zeger (1988) used a quasi-likelihood approach resembling the GEE method developed for population-averaged models (c.f. Liang and Zeger, 1986), while Chan and Ledolter (1995) used a Markov Chain Monte Carlo algorithm, a computationally intensive approach. Both methods were developed for Poisson count data using a log-linear model, but imposed different assumptions regarding the stationary latent structures: Zeger (1988) considered a weakly stationary latent process based on the first two moments while Chan and Ledolter (1995) used a stationary process defined by $\exp\{AR(1)\}$ where the AR(1) is the first-order autoregressive model with a normal
Chapter 7. State space models with stationary latent process

The class of state space models presented in this chapter bases the observed part on the Tweedie family which covers discrete data (including Poisson counts), continuous and mixed data, incorporating a more flexible latent part with the class of AR(1) models having exponential dispersion model margins. To estimate parameters, we use what we call the Kalman estimating equation (KEE), corresponding to a modified version of EM-algorithm with the E-step approximated by the Kalman smoother. We provide a rigorous proof concerning the asymptotic normality of the estimators obtained from the KEE method. Additionally, we develop a systematic approach to model checking diagnosis based on residual analysis, allowing us to avoid any serious misspecification of the models. It is noticeable that both Zeger (1988) and Chan and Ledolter (1995) did not discuss the problem of model diagnosis.

There is a concern about the difficulty of extending the latent process from the stationary AR(1) to higher-order Markov structure. As we have pointed out before, generalizing the AR(1) process of Joe (1996a) to a higher-order AR process is complicated, as evidenced by the technique of Joe (1996a), hampering the further development of the models presented in this chapter to handle high-order correlation structures. A possible path around this problem would model the latent part via the EDAR models discussed in Chapter 5, since all EDAR models have a unified structural presentation regardless of the order of model.

Finally, as an illustration of the parameter-driven model and the KEE estimation approach, we re-analyzed Zeger's polio incidence data, and demonstrated our model checking diagnosis in details. The polio data were previously analyzed by both Zeger (1988) and Chan and Ledolter (1995) who did not discuss any corresponding model diagnoses for the assumptions made in their analyses.
7.2 Models

We focus on one-dimensional time series, although generalization to multivariate serial data could be done, for example, in a similar way to that presented in Chapter 8.

7.2.1 Definition

Conditional on the latent process \( \theta_t \), \( Y_t \) is assumed to be a time series of independent observations following Tweedie exponential dispersion models,

\[
Y_t | \theta_t \sim Tw_p \left( a_t \theta_t, \nu^2 \theta_t^{1-p} \right), \quad t = 1, \ldots, n
\]  

(7.2.1)

with

\[
a_t = \exp(x_t^T \beta)
\]

where \( x_t \) is a \( k \)-vector of time-varying covariates, \( \beta \) is a \( k \)-vector of regression parameters of interest and \( \nu^2 \) is the dispersion parameter. As shown in Section 2.3, the Tweedie class accommodates a considerable range of distributions, including continuous, discrete and mixed types.

For the Poisson case \((p = 1)\), (7.2.1) reduces to \( \nu^2 Po(a_t \theta_t / \nu^2) \), so in this case we take \( \nu^2 = 1 \) in all formulas throughout the chapter.

Assume that latent process \( \{ \theta_t \} \) is first order stationary AR(1) with margins in the class of infinitely divisible exponential dispersion models, defined by Joe (1996a)

\[
\theta_t = A_t(\theta_{t-1}; \alpha) + \varepsilon_t, \quad t = 1, 2, \ldots
\]  

(7.2.2)

where \( A_t(\cdot; \alpha) \) is the thinning operation defined in Section 2.5, \( \alpha \in [0, 1] \) is the autocorrelation coefficient and \( \{ \varepsilon_t \} \) is a sequence of i.i.d. innovations distributed as \( ED^*\{\omega, (1-\alpha)\lambda\} \).

It is worth noting that the values of the latent process must be positive due to the definition of the model (7.2.1), which allows only certain exponential dispersion models.
to be used as marginal distributions for the latent process. This excludes the normal
distribution as a marginal distribution for the latent process. A possible choice for
marginal distribution would be the gamma distribution, exactly the distribution chosen
for the analysis of the polio data (Zeger, 1988) in Section 7.6. The AR(1) model with
gamma margins \( Ga^*(\omega, \lambda) \) was first proposed by Lewis, et al. (1989) as follows,

\[
\theta_t = G_t \theta_{t-1} + \varepsilon_t, \quad t = 1, \ldots
\]

(7.2.3)

where \( \{\varepsilon_t\} \) is a sequence of \( i.i.d \) innovations with distributions \( Ga^*\{\omega, (1 - \alpha)\lambda\} \), and
\( \{G_t\} \) is a sequence of \( i.i.d. \) random variables with beta distribution \( Beta\{\alpha\lambda, (1 - \alpha)\lambda\} \).

### 7.2.2 Moment structure

We now study the moment structure of the model, in order to interpret the correlation
structure of the model, and for use in the Kalman filter and smoother in Section 7.3.
Throughout the rest of this chapter, we let \( \bar{\alpha} = 1 - \alpha \).

**The conditional moment structure**

For the observed process (7.2.1), we have

\[
E(Y_t|\theta_t) = a_t \theta_t, \quad \text{Var}(Y_t|\theta_t) = a_t^2 \nu^2 \theta_t.
\]

(7.2.4)

For the latent process (7.2.2), because of stationarity, let

\[
E(\theta_t) = \mu = \lambda \tau(\omega), \quad \text{Var}(\theta_t) = \sigma^2 = \lambda V(\mu/\lambda)
\]

(7.2.5)

where \( \tau(\omega) = \kappa'(\omega) \) and \( V(\cdot) \) is the variance function of the model \( ED^*(\omega, \lambda) \). Then the
conditional moments of the latent process are given by

\[
E(\theta_t|\theta_{t-1}) = \alpha \theta_{t-1} + \bar{\alpha} \mu, \quad \text{Var}(\theta_t|\theta_{t-1}) = F(\theta_{t-1}) + \bar{\alpha} \sigma^2;
\]

(7.2.6)
where \( F(\theta_{t-1}) = \text{Var}\{A_t(\theta_t; \alpha)|\theta_{t-1}\} \). Note that the conditional variance of the latent model given in (7.2.6) is a non-linear function in \( \theta_{t-1} \) and no longer follow the structure of linear model. By equation (2.5.3), we have \( \text{E}\{A_t(\theta_t; \alpha)|\theta_t\} = \alpha \theta_t \). Furthermore,

\[
\text{E}\{F(\theta_{t-1})\} = \text{Var}\{A_t(\theta_t; \alpha)\} - \text{Var}[\text{E}\{A_t(\theta_t; \alpha)|\theta_{t-1}\}] = \alpha \sigma^2 - \alpha^2 \sigma^2 = \alpha \alpha \sigma^2.
\]

The marginal moment structure

The marginal moments for the observed process are given by

\[
\text{E}(Y_t) = a_t \mu, \quad \text{and} \quad \text{Var}(Y_t) = a_t^2 \nu^2 \mu + a_t^2 \sigma^2,
\]

while (7.2.5) gives the marginal moments for the latent process.

The covariances

We have the following covariances,

\[
\text{Cov}(Y_t, \theta_t) = a_t \sigma^2, \quad \text{Cov}(\theta_t, \theta_{t+s}) = \alpha^s \sigma^2, \quad (7.2.7)
\]

and

\[
\text{Cov}(Y_t, Y_{t+s}) = a_t a_{t+s} \alpha^s \sigma^2, \quad \text{Cov}(Y_t, \theta_{t+s}) = a_t \alpha^s \sigma^2, \quad \text{Cov}(Y_{t+s}, \theta_t) = a_{t+s} \alpha^s \sigma^2. \quad (7.2.8)
\]

By the definition of the latent process (7.2.2), the autocorrelation function is \( \text{Corr}(\theta_t, \theta_{t+s}) = \alpha^s \); moreover, the correlation of the observed process is given by

\[
\text{Corr}(Y_t, Y_{t+s}) = \frac{\sigma^2 \alpha^s}{\sqrt{(a_t^p - 2 \nu^2 \mu + \sigma^2)(a_{t+s}^p - 2 \nu^2 \mu + \sigma^2)}}.
\]

7.3 Kalman filter and smoother

We now develop the Kalman filter and smoother. Because the latent process (7.2.2) is stochastic functional and the variance structure (7.2.6) is non-linear, the classic theory
for the derivation of the Kalman filter and smoother is not applicable to our models. Therefore, we must derive these directly from their definitions by using the tools developed in Chapter 6 under a general setting. Before working out the details, we would like to state a fact that will be useful for the following calculations. Define the innovations of the observed process by

\[ \eta_t = Y_t - E(Y_t|\theta_t) = Y_t - a_t \theta_t. \]

Then \( \{\eta_t\} \) is uncorrelated with any function of \( \{\theta_t\} \), and it is also uncorrelated with \( \{\varepsilon_t\} \).

A sketch of the proof is in fact given as follows. For any measurable function \( f \), by the assumption of conditional independence,

\[ \text{Cov}\{\eta_t, f(\theta_{t+s})\} = E\text{Cov}\{Y_t - a_t \theta_t, f(\theta_{t+s})|\theta_t\} + \text{Cov}[E(Y_t - a_t \theta_t|\theta_t), E\{f(\theta_{t+s})|\theta_t\}] \]

\[ = E\text{Cov}\{Y_t, f(\theta_{t+s})|\theta_t\} - a_t E\text{Cov}\{\theta_t, f(\theta_{t+s})|\theta_t\} + \theta \]

\[ = 0 \quad (7.3.1) \]

and \( \text{Cov}(\eta_t, \varepsilon_{t+s}) = 0 \) by a similar argument.

### 7.3.1 Derivation

The Kalman filter predicts a future value of the observed process and the current value of the latent process based on past observations. Noting that both conditional moments (7.2.4) and (7.2.6) for the observed and latent processes satisfy, respectively, the conditions of (6.2.2) and (6.2.3), by (6.2.6), we obtain the predictions of \( (Y_t, \theta_t)^T \) given \( Y^{t-1} \)

\[ \left( \begin{array}{c} Y_t \\ \theta_t \end{array} \right) \bigg| Y^{t-1} \sim [\tilde{m}_{t|t-1}; \tilde{c}_{t|t-1}] \]

where the predictor is

\[ \tilde{m}_{t|t-1} = \alpha \begin{pmatrix} a_t \\ 1 \end{pmatrix} m_{t-1} + \bar{\alpha} \begin{pmatrix} a_t \\ 1 \end{pmatrix} \mu \]
and the prediction error is
\[
\tilde{\epsilon}_{t|t-1} = \alpha^2 \begin{pmatrix} a_t \\ 1 \end{pmatrix} (a_t, 1) c_{t-1} + (1 - \alpha^2) \sigma^2 \begin{pmatrix} a_t \\ 1 \end{pmatrix} (a_t, 1) + \begin{pmatrix} a_p^p v^2 \mu & 0 \\ 0 & 0 \end{pmatrix}.
\]

In particular,
\[
\theta_t \mid Y^{t-1} \sim [\alpha m_{t-1} + \bar{a} \mu; u_{t-1}].
\]

where \( u_t = \alpha^2 c_t + (1 - \alpha^2) \sigma^2 \). Assume that \( Y_t \mid Y^{t-1} \sim [f_t; Q_t] \). Therefore, by Theorem 6.2 of Chapter 6, we have

\[
f_t = \alpha a_t m_{t-1} + \bar{a} a_t \mu, \quad Q_t = \alpha^2 a_t^2 c_{t-1} + (1 - \alpha^2) \sigma^2 a_t^2 + a_p^p v^2 \mu = a_t^2 u_{t-1} + a_p^p v^2 \mu.
\]

Moreover, applying the equation (6.2.8), we update the prediction of \( \theta_t \) given a new observation \( Y_t \) and \( Y^{t-1} \) as follows,

\[
m_t = \alpha m_{t-1} + \bar{a} \mu + \frac{a_t u_{t-1}}{a_t^2 u_{t-1} + a_p^p v^2 \mu} (Y_t - f_t) \tag{7.3.2}
\]

and

\[
c_t = \alpha^2 c_{t-1} + (1 - \alpha^2) \sigma^2 - \frac{a_t^2 u_{t-1}^2}{a_p^p v^2 \mu + a_t^2 u_{t-1}}
\]

\[
= \frac{\nu^2 \mu a_t^p u_{t-1}}{\nu^2 \mu a_t^p + a_t^2 u_{t-1}}. \tag{7.3.3}
\]

The Kalman smoother predicts the latent process from the total vector of observations, to be used for the estimation procedure presented in Section 8.4. Assume that we know

\[
\theta_t \mid Y^t \sim [m_t; c_t], \quad t = 1, \ldots, n.
\]

Denote the Kalman smoother predictor given all observation \( Y^n \) by

\[
\theta_t \mid Y^n \sim [m_t^*; c_t^*].
\]
Assume that we currently know

\[ \theta_{t+1} | Y^n \sim [m^*_{t+1}; c^*_{t+1}] \]

Then the formulas given in (6.2.16) and (6.2.17) give the smoother,

\[ m^*_t = m_t + \frac{\alpha c_t}{u_t} (m^*_{t+1} - \alpha m_t - \tilde{\alpha} \mu) \tag{7.3.4} \]

and the prediction error

\[ c^*_t = c_t - \frac{\alpha^2 c^2_t}{u_t} + \frac{\alpha^2 c^2_t}{u^2_t} c^*_{t+1} \]

\[ = \frac{(1 - \alpha^2)\sigma^2}{u_t} c_t + \frac{\alpha^2 c^2_t}{u^2_t} c^*_{t+1}. \tag{7.3.5} \]

The recursion starts at \( t = n \) with \( m^*_n = m_n \) and \( c^*_n = c_n \).

### 7.3.2 Mean square error matrix

Let \( m^* = (m^*_0, \ldots, m^*_n)^T \) and \( \theta = (\theta_1, \ldots, \theta_n)^T \). Then

\[ \theta | Y^n \sim [m^*; C^*] \]

where \( C^* \) is the mean square error matrix of form

\[ C^* = \text{E}\{(\theta - m^*)(\theta - m^*)^T\} \]

whose diagonal elements \( c^*_t \) have given by (7.3.5). The off-diagonal elements are given by

\[ c^*_{t, t+s} = c^*_{t+s, t} = \text{E}\{(\theta_t - m^*_t)(\theta_{t+s} - m^*_{t+s})\} = \text{Cov}(\theta_t, \theta_{t+s} - m^*_{t+s}) \]

where the last equation is obtained using \( \text{E}\theta_t = \text{E}m^*_t = \mu \) and

\[ \text{Cov}(m^*_t, \theta_{t+s} - m^*_{t+s}) = \text{E}\text{Cov}\left\{m^*_t, \theta_{t+s} - m^*_{t+s} | Y^n\right\} + \text{Cov}\left\{\text{E}(m^*_t | Y^n), \text{E}(\theta_{t+s} - m^*_{t+s} | Y^n)\right\} \]

\[ = 0. \]
By definition of the model, we have
\[ \theta_t \mid \theta_{t+s}, Y^n \overset{d}{=} \theta_t \mid \theta_{t+s}, Y^{t+s-1} \overset{d}{=} \theta_t \mid \theta_{t+s} - \alpha m_{t+s-1}, Y^{t+s-1}. \]

As discussed by Jørgensen, et al. (1995), we obtain
\[
E \left\{ (\theta_t - m_t^*)(\theta_{t+s} - m_{t+s}^*) \right\} = \frac{\text{Cov}(\theta_t, \theta_{t+s} - \alpha m_{t+s-1})}{\text{Var}(\theta_{t+s} - \alpha m_{t+s-1})} c_{t+s}
\]
where \( \text{Var}(\theta_{t+s} - \alpha m_{t+s-1}) = u_{t+s-1} \) and
\[
\text{Cov}(\theta_t, \theta_{t+s} - \alpha m_{t+s-1}) = \text{Cov}(\theta_t, \theta_{t+s}) - \alpha \{ \text{Cov}(\theta_t, \theta_{t+s-1}) - \text{Cov}(\theta_t, m_{t+s-1}) \}
\]
\[= \alpha \text{Cov}(\theta_t, \theta_{t+s-1}m_{t+s-1}). \]

Following the same procedure, we further obtain
\[
\text{Cov}(\theta_t, \theta_{t+s-1}m_{t+s-1}) = \frac{\text{Cov}(\theta_t, \theta_{t+s-1} - \alpha m_{t+s-2})}{\text{Var}(\theta_{t+s-1} - \alpha m_{t+s-2})} c_{t+s-1}
\]
\[= \frac{\alpha}{u_{t+s-2}} \text{Cov}(\theta_t, \theta_{t+s-2} - \alpha m_{t+s-3}) c_{t+s-1}
\]
and finally,
\[
c_{t, t+s}^* = \alpha^s c_{t+s}^* \prod_{i=0}^{s-1} \frac{c_{t+i}}{u_{t+i}}.
\]

The recursive relation given below is especially useful for calculating the matrix \( C^* \).
\[
c_{t, t+s+1}^* = \alpha^{t+s+1} c_{t+s+1}^* \prod_{i=0}^{s} \frac{c_{t+i}}{u_{t+i}} = \alpha c_{t+s+1}^* c_{t+s}^* \frac{c_{t+s}}{u_{t+s}} c_{t, t+s}^*.
\]

In particular,
\[
c_{t, t+1}^* = \alpha \frac{c_t}{u_t} c_{t+1}^*.
\]

### 7.4 Estimation

This section presents a modified EM-algorithm for the maximum likelihood estimation of the regression parameters \( \beta \), given \( \nu^2 \) and \( \sigma^2 \), by way of the Kalman estimating equation...
proposed first in Jørgensen et al. (1994). Also the asymptotic normal distribution of the resulting estimators is described in this section while the detailed proof of the validity of the distribution appears in Section 7.7. Then we introduce the moment estimators for the dispersion parameters $\nu^2$ and $\sigma^2$ and the correlation parameters $\alpha$, as well as the calculation of the Godambe information matrix.

### 7.4.1 Estimation of regression parameters

Let

$$Y = (Y_1, \ldots, Y_n)^T, \quad X = (x_1, \ldots, x_n), \quad \mu(\beta) = (a_1 \theta_1, \ldots, a_n \theta_n)^T.$$  

Assuming $\theta_i$ fixed, by differentiating the joint likelihood of the model, we obtain the Fisher scoring equation as follows,

$$X A^1 \cdot (Y - \mu(\beta)) = 0 \quad (7.4.1)$$

where $A^c$ denotes a diagonal matrix $\text{diag}(a_1^c, \ldots, a_n^c)$ for any real $c$, and then we obtain the estimate $\hat{\beta}$ as the solution to the equation (7.4.1). By treating the latent process $\theta_i$ as a set of missing values, we would apply the EM-algorithm for estimation as follows.

The M-step is completed by solving the following equation,

$$\psi(\beta) = X A^{1-p} (Y - \tilde{\mu}(\beta)) = \sum_{t=1}^n x_t a_t^{1-p}(Y_t - a_t m_t^*) = 0, \quad (7.4.2)$$

replacing $\mu(\beta)$ in (7.4.1) by $\tilde{\mu}(\beta) = (a_1 m_1^*, \ldots, a_n m_n^*)$ where $m_t^*$ is the Kalman smoother.

Consequently, Equation (7.4.2) is called the Kalman estimating equation (KEE).

On the other hand, the E-step is done by using the Kalman smoother (7.3.4) given both the currently updated estimates $\hat{\beta}$ from the M-step and the moment estimates of $\sigma^2$, $\nu^2$ and $\alpha$.

For most distributions in the Tweedie families, including the cases such as $p = 0$ (normal), $p = 1$ (Poisson) and $p = 2$ (gamma), the M-step may be easily carried out
using standard statistical software for generalized linear models. An alternative algorithm to solve equation (7.4.2) is the so-called Newton scoring algorithm defined by

$$
\beta^{(j+1)} = \beta^{(j)} - S^{-1}\{\beta^{(j)}\} \psi\{\beta^{(j)}\}
$$

where the matrix $S(\beta)$ is given in Theorem 7.1 below.

The asymptotic normality of the distribution for the estimator $\hat{\beta}$ from the KEE holds under some mild regularity conditions, as indicated in the following theorem.

**Theorem 7.1** Suppose $\theta_t$ is a stationary process. Under mild regularity conditions, for $\nu^2, \sigma^2$ and $\alpha$ fixed,

$$
\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N_k\{0, \Sigma(\beta)\}
$$

with $\Sigma(\beta) = \lim_n n J^{-1}(\beta)$ where $J$ is the Godambe information matrix defined by

$$
J(\beta) = S^T(\beta)V^{-1}(\beta)S(\beta),
$$

where $S(\beta)$ is the sensitivity matrix given by

$$
S(\beta) = E_\beta\left\{ \frac{\partial \psi(\beta)}{\partial \beta^T} \right\}
$$

and

$$
V(\beta) = E_\beta\left\{ \psi(\beta)\psi^T(\beta) \right\}
$$

is the variability matrix.

The proof of the theorem is given in Section 7.7.

From the theorem, the asymptotic standard deviation of $\hat{\beta}$ may be calculated from the asymptotic variance $\Sigma(\beta)$. Moreover, an analogue of Wald's test is available for testing the joint significance of a subset of regression parameters. The test statistic is defined by

$$
W = \hat{\beta}_1^T\left\{J^{11}(\hat{\beta}_1)\right\}^{-1}\hat{\beta}_1
$$
where \( \hat{\beta}_1 \) is the subvector of interest to test and \( J^{11}(\hat{\beta}_1) \) is the corresponding block of the asymptotic covariance matrix of \( \hat{\beta}_1 \). Asymptotically, this statistic follows a \( \chi^2(k_1) \)-distribution where \( k_1 \) is the size of \( \beta_1 \).

**7.4.2 Estimating the dispersion and correlation parameters**

The likelihood estimates of dispersion parameters could, in principle, be obtained. However, the numerical computation for obtaining estimates are not easy to carry out in general. For example, in gamma case, dispersion parameters are involved in the gamma function which leads to a digamma function when taking derivatives with respect to the parameters. So, solving the corresponding estimating equation to obtain maximum likelihood estimates of parameters is not straightforward. Alternatively, we use the moment estimates for dispersion parameters.

The parameter \( \mu \) is assumed known in the following discussions. The equation

\[
\text{Var}(Y_t - a_t m_t^*) = a_t^p \mu \nu^2 - a_t^2 c_t^*
\]

leads to the following moment estimator of the dispersion parameter \( \nu^2 \),

\[
\hat{\nu}^2 = \frac{1}{n \mu} \sum_{t=1}^{n} \hat{a}_t^{-p} (Y_t - \hat{a}_t m_t^*)^2 + \frac{1}{n \mu} \sum_{t=1}^{n} \hat{a}_t^{2-p} c_t^*.
\]

(7.4.3)

To estimate the dispersion parameter \( \sigma^2 \) of the latent process, note that

\[
\text{Var}(Y_t) = a_t^p \nu^2 \mu + a_t^2 \sigma^2,
\]

so the moment estimator is of the form

\[
\hat{\sigma}^2 = \sum_{t=1}^{n} \left\{ (Y_t - \hat{a}_t \mu)^2 - \hat{a}_t^p \nu^2 \mu \right\} / \sum_{t=1}^{n} \hat{a}_t^2.
\]

(7.4.4)

We now consider the estimate of the autocorrelation coefficient \( \alpha \). Observing that

\[
\alpha \sigma^2 = \text{Cov}(\theta_t, \theta_{t+1}) = \text{Cov}(m_t^*, m_{t+1}^*) + c_{t,t+1}^* = \text{Cov}(m_t^*, m_{t+1}^*) + \frac{c_t}{u_t} c_{t+1}^*
\]
where \(c_{t+1}^*\) is given in (7.3.6), we obtain

\[
\alpha \left( \sigma^2 - \frac{c_t}{u_t} c_{t+1}^* \right) = \text{Cov} \left( m_t^*, m_{t+1}^* \right).
\]

This suggests the following moment estimate of \(\alpha\),

\[
\hat{\alpha} = \frac{1}{n-1} \sum_{t=1}^{n-1} \frac{(m_t^* - \mu)(m_{t+1}^* - \mu)}{\hat{\sigma}^2 - c_t u_t^{-1} c_{t+1}^*}.
\]

(7.4.5)

It is noted that for small size of data the moment estimate of \(\alpha\), sometimes, might not lead to a positive definite covariance structure.

### 7.4.3 Calculation of Godambe information matrix

We first derive the \((k \times k)\) sensitivity matrix \(S(\beta)\). Define

\[
D_\beta(t) = E \left( \frac{\partial m_t^*}{\partial \beta^T} \right), \quad \text{and} \quad D_\beta^*(t) = E \left( \frac{\partial m_t^*}{\partial \beta^T} \right).
\]

Noting that \(E(Y_t - a_t m_t^*) = 0\), we have

\[
S(\beta) = (1 - p) \sum_{t=1}^{n} a_t^{1-p} x_t x_t^T E(Y_t - a_t m_t^*) - \sum_{t=1}^{n} a_t^{2-p} x_t x_t^T E m_t^* - \sum_{t=1}^{n} a_t^{2-p} x_t E \left( \frac{\partial m_t^*}{\partial \beta^T} \right)
\]

\[
= - \sum_{t=1}^{n} a_t^{2-p} x_t \left( \mu x_t^T + D_\beta^*(t) \right)
\]

The \((k \times k)\) variability matrix is given as follows,

\[
V(\beta) = E \beta \left\{ \psi(\beta) \psi^T(\beta) \right\} = X A^{1-p} V(\mathbf{Y - A}^\dagger) A^{1-p} X^T
\]

(7.4.6)

where \(V(\mathbf{Y - A}^\dagger)\) is a \(k \times k\) matrix whose \((t, s)\)-th element is of form

\[
\text{Cov}(Y_t - a_t m_t^*, Y_s - a_s m_s^*) = \begin{cases} a_t^2 \nu^2 \mu - a_t^2 c_t^*, & t = s \\ -a_t a_s c_{t, s}^*, & t \neq s \end{cases}
\]

A sketch of the derivation of \(V(\mathbf{Y - A}^\dagger)\) is obtained using the following two facts. Under the assumption that two conditional residuals from both processes are uncorrelated, we have the first fact,

\[
\text{Cov}(Y_t - a_t \theta_t, Y_s - a_s \theta_s) = \text{Cov}(Y_t - a_t m_t^*, Y_s - a_s m_s^*) + a_t a_s c_{t, s}^*,
\]
where \( c^*_t \equiv c^*_i \) by the convention in Section 7.3.2. The second fact is that

\[
\text{Cov}(Y_t - a_t \theta_t, Y_s - a_s \theta_s) = \begin{cases} 
\alpha^2 \nu^2 \mu, & t = s \\
0, & t \neq s 
\end{cases}
\]

obtained by the arguments based on the conditional covariance given \( \theta_t \), say.

Finally, we establish recursive relations for both \( D_\beta \) and \( D^*_\beta \). From equation (7.3.2), by taking expectations on derivatives with respect to \( \beta^T \), we have

\[
D_\beta(t) = \alpha D_\beta(t-1) - \frac{a_t u_{t-1}}{\alpha^2 \nu^2 \mu} E \left( \frac{\partial f_t}{\partial \beta^T} \right) 
\]

\[
= \alpha \left( 1 - \frac{a_t^2 u_{t-1}}{\alpha^2 \nu^2 \mu} \right) D_\beta(t-1) - \frac{a_t^2 u_{t-1} \mu}{\alpha^2 \nu^2 \mu} x_t^T,
\]

starting with \( D_\beta(0) = 0 \). The last equality is obtained by noting

\[
\frac{\partial f_t}{\partial \beta^T} = \alpha \mu \frac{\partial a_t}{\partial \beta^T} + \alpha a_t D_\beta(t-1) + \bar{a}_t \mu \frac{\partial a_t}{\partial \beta^T} = \alpha a_t D_\beta(t-1) + \mu a_t x_t^T.
\]

Moreover, noting that \( E(m^*_{t+1} - \alpha m_t - \bar{a} \mu) = 0 \), we obtain

\[
D^*_\beta(t) = D^*_\beta(t) = D_\beta(t) + \alpha \frac{c_t}{u_t} \left\{ D_\beta(t+1) - \alpha D_\beta(t) \right\},
\]

starting at \( t = n \) with \( D^*_\beta(n) = D_\beta(n) \).

### 7.5 Model checking

We now develop methods for model diagnosis for both the observed and unobserved parts of the model. In principle, there may be in all, seven types of residuals classified as either marginal or conditional residuals. Also we could form the residuals based on either the Kalman filter or smoother. In this section, we are interested in only three out of these seven residuals, which are those used for our model checking procedure.

1. For the observed part, we use the conditional residuals based on the Kalman filter, defined by

\[
\hat{\eta}_t = Y_t - f_t
\]
By the fourth basic property and Theorem 6.3 in Chapter 6, for any \( s > 0 \), \( \hat{\eta}_{t+s} \) is uncorrelated with \( Y_{t+s} \), and hence these residuals are uncorrelated and have variances \( Q_t \).

(2) For the latent process, we choose the conditional residuals based on the Kalman filter, given by
\[
\hat{\varepsilon}_t = m_t - (\alpha m_{t-1} + \alpha \mu).
\]
By the similar argument given in part (1), these residuals are uncorrelated and their variances are
\[
\text{Var}(\hat{\varepsilon}_t) = (\nu^2 + \mu^2 a_t^2 c_t^2) Q_t
\]
because of the fact \( \hat{\varepsilon}_t = \nu^2 \mu a_t^{p-1} c_t \hat{\eta}_t \).

(3) Marginal residuals of the latent process based on the Kalman filter are also considered, defined by
\[
r_t = m_t - \mu
\]
with moments given as follows,
\[
E(r_t) = 0, \quad \text{Var}(r_t) = \sigma^2 - c_t,
\]
and
\[
\text{Cov}(r_t, r_{t+s}) = \alpha^2 \sigma^2 - \alpha^2 c_t. \quad (7.5.1)
\]
In the following, we let a bar denote standardized residuals, for instance,
\[
\hat{\varepsilon}_t = \hat{\varepsilon}_t / \sqrt{\text{Var}(\hat{\varepsilon}_t)}.
\]

**Checking correlation structure**

A simple way to check the first-order Markov assumption of the latent process is to plot \( \hat{\varepsilon}_t \) versus \( \hat{\varepsilon}_{t-1} \). Similarly, the plot of \( \hat{\eta}_t \) against \( \hat{\eta}_{t-1} \) provides a check of the
conditional independence for the observed process over time. One may also inspect the
autocorrelation function of \(\hat{\varepsilon}_t\) and that of \(\hat{\eta}_t\) for the same purpose.

**Checking distributional assumptions**

As in generalized linear models, we may check the form of the variance functions (and
hence the distributional forms) by plotting the standardized filter residuals versus the
corresponding fitted values. To check the conditional distribution assumption of the \(\pi_t\)
given \(\pi_{t-1}\), we plot \(\hat{\varepsilon}_t\) against \(\log(\alpha m_{t-1} + \tilde{\alpha}\mu)\). In a similar way, we may plot \(\hat{\eta}_t\) against
\(\log(f_t)\) to check the conditional distribution assumption for the observed part.

**Checking the covariate structure**

Plots of standardized residuals against each covariate are useful for detecting nonlin-
earity of the model. For instance, to check the log link, we could plot \(\log(Y_t)\) against
\(\log(f_t^*)\) where \(f_t^* = a_tm_t^*\). In principle, this plot should show a horizontal linear relationship.

**Checking the latent process**

From (7.5.1), we have

\[
\text{Cov}(r_t, r_{t+s})/\alpha^2 = \sigma^2 - c_t
\]  

(7.5.2)

where the right-hand side of the equation (7.5.2) is constant in lag \(s\). In practice,
\(\text{Cov}(r_t, r_{t+s})\) is indeed replaced by the sample autocovariance function and both \(\alpha\) and
\(\sigma^2\) by their moment estimators. If the values on the left-hand side of the equation (7.5.2)
are about constant as a function of the lag \(s\), the AR(1) structure is confirmed. In prac-
tice, there is certain risk of a misleading conclusion, because equation (7.5.2) contains
estimators, in particular the \(\alpha^2\) are very sensitive to error due to the power function
when \(\hat{\alpha}\) is not precise enough. Nevertheless, we could check the pattern based on the
first several lags, which is more reliable than the pattern with respect to large lags.
For the latent model with gamma margin, we have an alternative way to check the AR(1) structure by inspecting the simulated residuals given by

\[ \hat{e}_t = m_t - G_t m_{t-1} \]  

where \( G_t \) are i.i.d sequence of random variates simulated from the distribution \( \text{Beta} \{ \alpha \lambda, (1 - \alpha) \lambda \} \) where \( \lambda = 1/\sigma^2 \), and they are also independent of the \( Y_t \)’s. For given \( \alpha \) and \( \lambda \), since \( E(G_t) = \alpha \) and \( \text{Var}(G_t) = \alpha(1 - \alpha)/2 \), it is easy to show that

\[ \text{Cov}(\hat{e}_t, \hat{e}_{t+s}) = \text{Cov}(\hat{e}_t, \hat{e}_{t+s}) = 0 \]

and

\[ \text{Var}(\hat{e}_t) = \text{Var}\{ (m_t - \alpha m_{t-1}) + (\alpha - G_t) m_{t-1} \} \]
\[ = \text{Var}(m_t - \alpha m_{t-1}) + \text{Var}\{ (\alpha - G_t) m_{t-1} \} \]
\[ = \text{Var}(\hat{e}_t) + \text{Var}(G_t) \{ \text{Var}(r_t) + \mu^2 \}. \]

Therefore, the standardized \( \hat{e}_t \) are a noise series provided that the structure of AR(1) is correct. Any dependence pattern would indicate inadequacy of the stationary assumption.

### 7.6 Data analysis

We now apply the parameter-driven model to the analysis of the polio incidence data taken from Zeger (1988) as an example. Table 7.1 reports the polio data which are the monthly numbers of poliomyelitis cases in the USA from 1970 to 1983, given in Table 2 of Zeger (1988). Figure 7.1 gives a time series plot of the data. The central question investigated by Zeger (1988) is whether or not the data provide evidence of a long-term decrease in the rate of U.S. polio infections, that is, whether or not the data follow a decreasing time trend.
Chapter 7. State space models with stationary latent process

Table 7.1: Monthly number of U.S. cases of poliomyelitis for 1970 to 1983.

<table>
<thead>
<tr>
<th>Year</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>Aug</th>
<th>Sept</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>1971</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1972</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>1973</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1974</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>1975</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1976</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1977</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>1978</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1979</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>1980</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1981</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1982</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1983</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 7.1: Monthly counts of Polio incidences. Dots show the observations. The solid line represents the estimated latent process by the Kalman smoother. The broken line shows the moving average series of the data with both 6 and 12 lags as the initialization of the latent process for Kalman filter.
Chapter 7. State space models with stationary latent process

These data were previously analyzed by Zeger (1988) and Chan and Ledolter (1995) under a stationary latent structure. They both assumed that the data are conditionally independent given the latent process and follow a Poisson distribution. Here the conditional mean is a function of covariates and the regression parameters. Zeger (1988) used the quasi-likelihood approach similar to the GEE in his model where the latent process is assumed to be stationary without the specification of marginal distribution. Chan and Ledolter (1995) assumed the latent process to be exp\{AR(1)\} where AR(1) is a stationary process with Gaussian margins. They used a Markov chain Monte Carlo approach to deal with the estimation problem.

7.6.1 Poisson stationary-gamma model

In the present section, we tentatively fit the following model to Zeger's (1988) data. The model assumptions will be confirmed through residual analysis later.

Taking \( p = 1 \) in model (7.2.1), we assume that given the latent process \( \theta_t \), the polio counts are conditionally independent over time and follow a Poisson distribution with mean \( a_t \theta_t \) (therefore \( \nu^2 = 1 \)), namely,

\[
Y_t | \theta_t \sim Po(a_t \theta_t), \quad t = 1, \ldots, 168
\]

where \( x_t = (1, t, \cos(2\pi t/12), \sin(2\pi t/12), \cos(2\pi t/6), \sin(2\pi t/6))^T \) which is the same as Zeger (1988), including the constant term, time trend and the seasonal effects.

The latent process is chosen to be an AR(1) model with gamma margins given by the equation (7.2.3),

\[
\theta_t = G_t \theta_{t-1} + \varepsilon_t, \quad t = 1, \ldots,
\]

where the mean \( \mu \) of \( \theta_t \) is set to 1, forcing the unconditional mean of \( E(Y_t) \) in (7.2.4) to depend only on \( x_t^T \beta \) and not on moments of the \( \theta_t \) series, as is desirable in the regression
analysis. So, we have

$$E(Y_t) = a_t, \text{ and } \text{Var}(Y_t) = a_t(1 + a_t\sigma^2).$$

Here the latent process introduces both overdispersion and autocorrelation into $Y_t$. Furthermore, we obtain a log-linear model for the marginal expectation as follows,

$$\log\{E(Y_t)\} = x_t^T \beta. \quad (7.6.1)$$

We must now obtain an estimator of $\beta$ and test whether or not the time trend is significant.

### 7.6.2 Implementation of estimation

Under the Poisson stationary-gamma model, the Kalman filter and smoother take the following form. The filter is given by

$$m_t = \alpha m_{t-1} + \bar{\alpha} + c_t(Y_t - f_t)$$

with prediction error

$$c_t = \frac{a_t u_{t-1}}{a_t + a_0^2 u_{t-1}} = \frac{u_{t-1}}{1 + a_t u_{t-1}},$$

where $u_t = \alpha^2 c_t + (1 - \alpha^2)\sigma^2$, $f_t = \alpha a_t m_{t-1} + \bar{\alpha} a_t$ and $Q_t = a_t^2 u_{t-1} + a_t$.

The smoother is obtained as

$$m_t^* = m_t + \frac{\alpha c_t}{u_t} (m_{t+1}^* - \alpha m_t - \bar{\alpha})$$

with

$$c_t^* = (1 - \alpha^2)\sigma^2 \left( \frac{c_t}{u_t} \right) + \alpha^2 \left( \frac{c_t}{u_t} \right)^2 c_{t+1}.$$

The estimator of $\beta$ is obtained as the solution of the Newton scoring algorithm for the inference function $\psi$ of form

$$\psi(\beta) = \sum_{t=1}^n x_t(Y_t - a_t m_t^*).$$
Chapter 7. State space models with stationary latent process

The sensitivity matrix is given by

\[ S(\beta) = -\sum_{t=1}^{n} a_t x_t \left( x_t^T + D_\beta^*(t) \right). \]

From (7.4.6), the variability matrix is given by

\[ V(\beta) = X \text{Var} \left( Y - Am^* \right) X^T = X \left\{ \text{diag}(a_t) - \text{diag}(a_t)C^* \text{diag}(a_t) \right\} X^T. \]

Also, we could simplify the both forms of \( D_\beta(t) \) and \( D_\beta^*(t) \) as follows,

\[ D_\beta(t) = \alpha(1 - a_t c_t)D_\beta(t-1) - a_t c_t x_t^T \]

and

\[ D_\beta^*(t) = D_\beta(t) + \alpha \frac{c_t}{u_t} \left\{ D_\beta^*(t+1) - \alpha D_\beta(t) \right\}. \]

The estimator of the dispersion parameter \( \sigma^2 \) is of form

\[ \hat{\sigma}^2 = \sum_{t=1}^{n} \frac{(Y_t - \hat{a}_t)^2 - \hat{a}_t}{\sum_{t=1}^{n} \hat{a}_t^2} \]

and the estimator of the correlation parameter is given by

\[ \hat{\alpha} = \frac{1}{n-1} \sum_{t=1}^{n-1} \frac{(m_{t+1}^* - 1)(m_{t+1}^* - 1)}{\hat{\sigma}^2 - c_t u_t^{-1} c_{t+1}^*}. \]

7.6.3 Numerical results

We initialize the Kalman filter by the moving average series of the polio counts obtained based on both 6 and 12 lags, taking the seasonal effects away, shown by the broken line in Figure 7.1. The starting value for \( \alpha \) is chosen to be the autocorrelation of the starting \( \theta \) series, and the starting value for \( \sigma^2 \) is set to 1, the same as in Chan and Ledolter (1995). The convergence criterion is set to \( \max \left( \| \beta^{(i+1)} - \beta^{(i)} \| \right) < 10^{-5} \).

Our results, as well as those reported by Zeger (1988) and Chan and Ledolter (1995), are shown in Table 7.2. In general, the estimates from the three different approaches seem similar; however, a few differences should be pointed out:
Table 7.2: Estimates of coefficients and standard errors: (a) Zeger’s parameter-driven model using GEE; (b) Chan and Ledolter’s parameter-driven model using an MCMC algorithm; (c) our parameter-driven model using the KEE approach.

<table>
<thead>
<tr>
<th></th>
<th>(a) GEE</th>
<th>(b) MCMC</th>
<th>(c) KEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\beta} )</td>
<td>Std. err.</td>
<td>( \hat{\beta} )</td>
</tr>
<tr>
<td>Intercept (log rate in Jan. 1976)</td>
<td>.17</td>
<td>.13</td>
<td>.42</td>
</tr>
<tr>
<td>Trend ( \times 10^{-3} )</td>
<td>-4.35</td>
<td>2.68</td>
<td>-4.62</td>
</tr>
<tr>
<td>( \cos(2\pi t/12) )</td>
<td>-.11</td>
<td>.16</td>
<td>-.15</td>
</tr>
<tr>
<td>( \sin(2\pi t/12) )</td>
<td>-.48</td>
<td>.17</td>
<td>-.50</td>
</tr>
<tr>
<td>( \cos(2\pi t/6) )</td>
<td>.20</td>
<td>.14</td>
<td>.44</td>
</tr>
<tr>
<td>( \sin(2\pi t/6) )</td>
<td>-.41</td>
<td>.14</td>
<td>-.04</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>.77</td>
<td>.54</td>
<td>.81</td>
</tr>
<tr>
<td>( \hat{\rho}(1) )</td>
<td>.77</td>
<td>.88</td>
<td>.36</td>
</tr>
</tbody>
</table>

Note: \( \hat{\rho}(1) \) represents the estimator of the first lag autocorrelation for the latent process.

1. Our KEE estimates are closer to Chan and Ledolter’s maximum likelihood (MCMC) estimates than Zeger’s estimates, noting that the estimates corresponding to the four seasonal effects are almost the same in (b) and (c).

2. With regard to the time trend, both our method and Chan and Ledolter’s method found the slope to be significantly different from zero, implying a significant decreasing time trend for the polio data. Zeger reported insignificance for the slope of the time-trend at the .05 level, having a \( t \) ratio of 1.62, while both Chan and Ledolter’s and our \( t \) ratios exceed 2.3.

3. Standard errors for the KEE estimates corresponding non-constant covariates are uniformly smaller than the those given by Zeger, being nearly the same as the those obtained by Chan and Ledolter.

4. Our estimate for the lag 1 autocorrelation, \( \hat{\rho} \), is almost twice as small as the one
Chapter 7. State space models with stationary latent process

reported by Zeger and two-and-a half times smaller than that obtained by Chan and Ledolter. The lag 1 autocorrelation for the raw data of polio counts is close to .3, so the value of .36 is reasonable for the lag 1 autocorrelation of latent process, under the assumption that the observations are conditionally independent given the latent part.

The solid line in Figure 7.1 shows the estimated latent process from the Kalman smoother. Although the computation for the KEE involves the updating of predictions of the latent process, the most time-consuming part of the method, it is worthwhile, because the availability of prediction for the latent process makes model checking diagnosis feasible, as discussed in the following section.

With regard to the computational aspects for the analysis of the polio data, our KEE program took about 16 minutes on a Sun Sparc-10 Workstation to complete a total of 166 iterations until convergence. The program was written in S-PLUS code rather than C code, so the KEE approach is generally fairly competitive in terms of computational time. Also the KEE approach is very easy to implement, compared with the MCMC algorithm which usually needs programming in the C language due to the tremendous demand for computing memory on the simulations. More detailed comparisons between the MCEM and KEE methods on the computational aspects are found in Chapter 9.

7.6.4 Model diagnosis

When reading this section, the reader should bear one thing in mind: all the time series plots such as the autocorrelation function (acf) and partial autocorrelation function (pacf) plots are just approximate demonstrations, since the distributions considered here are non-normal.

Figure 7.2 provides a check of the correlation structures of both the latent part and
Figure 7.2: Checking correlation structures for both latent and observed processes.
Chapter 7. State space models with stationary latent process

observed part. The two top displays are respectively the plot of $\hat{\epsilon}_t$ versus $\hat{\epsilon}_{t-1}$ and the acf plot of $\hat{\epsilon}_t$. The two bottom displays are respectively the plot of $\hat{\eta}_t$ versus $\hat{\eta}_{t-1}$ and the acf plot of $\hat{\eta}_t$. All four plots show absence of serial correlation for both residuals respectively from the latent process and observed process. This confirms the assumptions of a first-order Markov latent process and the conditional independence for the observed process.

Figure 7.3: Checking distribution assumptions for the latent and observed process.

Figure 7.3 shows plots of filter residuals $\hat{\epsilon}_t$ and $\hat{\eta}_t$ against respectively the log linear predictors $\log(\alpha m_{t-1} + \bar{\alpha})$ and $\log(f_t)$, for the purpose of checking the distributional assumptions. The plot for checking the assumption of gamma distribution (latent part) looks fine, while the one for checking the assumption of Poisson distribution (observed part) shows clear discreteness. Taking into account the small expected counts, the second
plot seems to be in agreement with the proposed model. Figure 7.4 shows the plot of \( \hat{\eta}_t \) versus time \( t \) to check the adequacy of linearity for the time trend covariate. The plot looks fairly random, confirming the proposed model structure for the time trend covariate which is the main focus in the analysis of the data.

Figure 7.4: Checking the linearity of the time trend.

Figure 7.5 provides a check of the structure of the latent process by plotting the autocorrelation and partial autocorrelation functions of the standardized residuals \( \hat{e}_t \), which are supposed to be a series of noise values, where \( G_t \) is a sequence of beta numbers randomly generated from the distribution \( Beta(\hat{\alpha}/\hat{\sigma}^2, (1 - \hat{\alpha})/\hat{\sigma}^2) \). Both the acf and pacf plots show no significant peaks, confirming the adequacy of the assumption for the latent structure.
7.7 Proof of asymptotic normality

**Lemma 7.1** If \( \{\theta_t\} \) is a stationary Markov process of order 1, then \( \{\theta_t\} \) is a \( \phi \)-mixing process.

For the proof, refer to Billingsley (Page 167, 1968).

**Lemma 7.2** Assume \( \{\theta_t\} \) is a \( \phi \)-mixing process and given \( \theta_t \), \( \{Y_t\} \) is conditionally independent. Then the \( \{Y_t\} \) is a \( \phi \)-mixing process, and therefore under some mild regularity conditions

\[
\frac{1}{n} \sum_{t=1}^{n} (Y_t - EY_t) \to 0, \quad \text{as } n \to \infty, \tag{7.7.1}
\]

and \( \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (Y_t - EY_t) \) is asymptotically normal with mean 0 and variance

\[
\tau_{\theta}(0) + \lim_{n \to \infty} \frac{1}{n} \sum_{j=-n+1}^{n-1} \tau_{\theta}(j).
\]
where \( \tau_y(j) = \text{Cov}(Y_{t+j}, Y_t) \).

**Proof** First we prove that \( \{Y_t\} \) is a \( \phi \)-mixing process. Let

\[ B_{i,j} = \sigma(Y_t, i \leq t \leq j) \]

for all \( 1 \leq i \leq j \leq \infty \), and let \( A \in B_{1, n} \) and \( B \in B_{n+s, \infty} \). Since given \( \theta_t, Y_t \) is conditionally independent,

\[
P(A \cap B) = \int_R P(A \cap B | \theta_{n+s} = x) p_{n+s}(x) dx \]
\[
= \int_R P(A | \theta_{n+s} = x) P(B | \theta_{n+s} = x) p_{n+s}(x) dx
\]

here \( p_t(x) \) denotes the density function of \( \theta_t \). Since \( \{\theta_t\} \) is a \( \phi \)-mixing process, there exists a function \( \phi(s) \) with \( \phi(s) \rightarrow 0 \) as \( s \rightarrow \infty \) such that

\[
|p_{n,n+s}(z, x) - p_n(z)p_{n+s}(x)| \leq \phi(s)
\]

for all \( 1 \leq n, s \leq \infty \). Hence,

\[
\frac{P(A | \theta_{n+s})}{P(A)} = \int_R P(A | \theta_{n+s} = x, \theta_n = z) p_{n+s \mid n}(x | z) p_n(z) dz
\]
\[
= \int_R p_n(z | A) P_{n+s}(z, x) \frac{p_{n+s}(z, x)}{p_n(z)p_{n+s}(x)} dz
\]
\[
\leq \int_R p_n(z | A) \left( 1 + \frac{\phi(s)}{p_n(z)p_{n+s}(x)} \right) dz
\]
\[
= 1 + \phi(s) \int_R \frac{p_n(z | A)}{p_n(z)p_{n+s}(x)} dz
\]

where \( p_{t,t'}(x, z) \) denotes the joint density of \( \theta_t \) for the state \( z \) at time \( t \) and the state \( x \) at time \( t' \), \( p_{t \mid t'}(x \mid z) \) denotes the conditional density of \( \theta_t, \theta_{t'} \), and \( p_t(z | A) \) denotes the conditional density of \( \theta_t \) given a Borel set \( A \).

It follows that

\[
P(A \cap B) \leq P(A)P(B) + \phi(s)P(A) \int_R \int_R p_n(z | A) \frac{p_{n+s}(z, x)}{p_n(z)} P(B | \theta_{n+s} = x) dz dx
\]
\[
= P(A)P(B) + \phi(s) \left[ \int_R P(A | \theta_n = z) dz \right] \left[ \int_R P(B | \theta_{n+s} = x) dx \right]
\]
Proceeding the similar procedure above, we may obtain
\[
P(A \cap B) \geq P(A)P(B) - \phi(s) \left[ \int_R P(A|\theta_n = z)dz \right] \left[ \int_R P(B|\theta_{n+s} = x)dx \right].
\]
So,
\[
|P(A \cap B) - P(A)P(B)| \leq \phi^*(s)
\]
where
\[
\phi^*(s) = \phi(s) \left[ \int_R P(A|\theta_n = z)dz \right] \left[ \int_R P(B|\theta_{n+s} = x)dx \right] \to 0 \text{ as } s \to 0.
\]
The proof for the rest parts of this theorem is just an application of theorems appeared in Chapter 3 of Stout (1974) and Chapter 4 of Billingsley (1968), so the details are omitted.

Now we prove Theorem 7.1. Under some regularity conditions, \( \sqrt{n}(\hat{\beta} - \beta) \) can be approximated by
\[
\frac{1}{n} \left[ - \left( \frac{\partial \psi(\beta)}{\partial \beta^T} \right) / n^2 \right]^{-1} \left[ \frac{\psi(\beta)}{\sqrt{n}} \right]
\]
where
\[
- \left( \frac{\partial \psi(\beta)}{\partial \beta^T} \right) = - \sum_{t=1}^n x_t a_t^{2-p} x_t^T m_t^* - \sum_{t=1}^n x_t a_t^{2-p} \left( \frac{\partial m_t^*}{\partial \beta^T} \right) + (1 - p) \sum_{t=1}^n x_t a_t^{1-p} x_t^T (Y_t - a_t m_t^*).
\]
From the definition of the Kalman smoother, denoting
\[
H_t = (h_{t1}, \ldots, h_{tn}) = \text{Cov}(\theta_t, Y) \text{Var}^{-1}(Y),
\]
we obtain
\[
m_t^* - E(m_t^*) = m_t^* - \mu = H_t(Y - EY).
\]
The first term on the right-hand side of the equation (7.7.2) may be written as follows,
\[
\frac{1}{n^2} \sum_{t=1}^n x_t a_t^{2-p} x_t^T m_t^* = \frac{1}{n^2} \sum_{t=1}^n x_t a_t^{2-p} x_t^T E m_t^* + \frac{1}{n^2} \sum_{t=1}^n x_t a_t^{2-p} x_t^T (m_t^* - E m_t^*)
\]
\[
= \frac{\mu}{n^2} \sum_{t=1}^n x_t a_t^{2-p} x_t^T + o_p(1)
\]
since
\[ \frac{1}{n^2} \sum_{t=1}^{n} x_t a_t^{2-p} x_t^T (m_t^* - E m_t^*) = \frac{1}{n^2} \sum_{t=1}^{n} x_t a_t^{2-p} x_t^T \sum_{j=1}^{n} h_{tj} (Y_j - E Y_j) \]
\[ = \frac{1}{n} \sum_{j=1}^{n} \left( \frac{1}{n} \sum_{t=1}^{n} h_{tj} x_t a_t^{2-p} x_t^T \right) (Y_j - E Y_j) \]
\[ = o_p(1), \]

provided that
\[ \frac{1}{n} \sum_{t=1}^{n} h_{tj} x_t a_t^{2-p} x_t^T = O(1). \]  

Similarly, the second term on the right-hand side of the equation (7.7.2) can be evaluated as follows,
\[ \frac{1}{n^2} \sum_{t=1}^{n} x_t a_t^{2-p} \left( \frac{\partial m_t^*}{\partial \beta^T} \right) = \frac{1}{n^2} \sum_{t=1}^{n} x_t a_t^{2-p} \left( E \frac{\partial m_t^*}{\partial \beta^T} \right) + \frac{1}{n^2} \sum_{t=1}^{n} x_t a_t^{2-p} x_t^T \left( \frac{\partial m_t^*}{\partial \beta^T} - E \frac{\partial m_t^*}{\partial \beta^T} \right) \]
\[ = \frac{1}{n^2} \sum_{t=1}^{n} x_t a_t^{2-p} D_{\beta}^T + o_p(1) \]
where the last equality is obtained by the following arguments. First we note that
\[ \frac{\partial m_t^*}{\partial \beta^T} - E \frac{\partial m_t^*}{\partial \beta^T} = \frac{\partial H_t}{\partial \beta^T} (Y - E Y). \]

It follows from Lemma 7.2 that
\[ \frac{1}{n^2} \sum_{t=1}^{n} a_t^{2-p} x_t \frac{\partial H_t}{\partial \beta^T} (Y - E Y) = \frac{1}{n} \sum_{j=1}^{n} \left\{ \frac{1}{n} \sum_{t=1}^{n} a_t^{2-p} x_t \left( \frac{\partial h_{tj}}{\partial \beta^T} \right) \right\} (Y_j - E Y_j) \]
\[ = o_p(1), \]
assuming
\[ \frac{1}{n} \sum_{t=1}^{n} a_t^{2-p} x_t \left( \frac{\partial h_{tj}}{\partial \beta^T} \right) = O(1). \]  

Because \( \frac{1}{\sqrt{n}} \sum_{t=1}^{n} a_t^{1-p} x_t x_t^T (Y_t - a_t m_t^*) = o_p(1) \) due to the asymptotic normality (which will be proven below), the third term in the right-hand side of equation (7.7.2) is \( O_p(1/\sqrt{n}). \)
Chapter 7. State space models with stationary latent process

By now we prove that the equation (7.7.2) has
\[
-\frac{1}{n^2} \left( \frac{\partial \psi(\beta)}{\partial \beta^T} \right) = -\frac{\mu}{n^2} \sum_{t=1}^{n} a_t^{2-p} x_t x_t^T - \frac{1}{n^2} \sum_{t=1}^{n} a_t^{2-p} x_t D_\beta^*(t) + o_p(1)
\]
\[
= -\frac{1}{n^2} \sum_{t=1}^{n} a_t^{2-p} x_t \left( \mu x_t + D_\beta^*(t) \right)^T + o_p(1)
\]
where \( D_\beta^*(t) \) is given in (7.4.7).

On the other hand, noting that
\[
Y_t - a_t m_t^* = Y_t - a_t \mu - a_t H_t(Y - EY)
\]
\[
= (1 - a_t h_{tt})(Y_t - EY_t) - \sum_{j \neq t} a_t h_{tj}(Y_j - EY_j)
\]
\[
= \sum_{j=1}^{n} \tilde{h}_{tj}(Y_j - EY_j)
\]
where
\[
\tilde{h}_{tj} = \begin{cases} 
1 - a_t h_{tt} & \text{if } j = t \\
-a_t h_{tj} & \text{if } j \neq t
\end{cases}
\]
we have
\[
\frac{1}{\sqrt{n}} \psi(\beta) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} a_t^{1-p} x_t \sum_{j=1}^{n} \tilde{h}_{tj}(Y_j - EY_j)
\]
\[
= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left( \sum_{t=1}^{n} a_t^{1-p} x_t \tilde{h}_{tj} \right) (Y_j - EY_j).
\]
By Lemma 7.2, \( \frac{1}{\sqrt{n}} \psi(\beta) \) is asymptotically normal with mean 0 and variance
\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i,j}^{n} X A^{1-p} \tilde{h}_i \tau_p (i - j) \tilde{h}_j^T A^{1-p} X^T,
\]
where \( \tilde{h}_i = (\tilde{h}_{ii}, \ldots, \tilde{h}_{ni})^T \). By (7.4.6), the asymptotic variance may also be written as
\[
\lim_{n \to \infty} \frac{1}{n} X A^{1-p} \text{Var}(Y - A \bar{m}^*) A^{1-p} X^T
\]
where \( Y = (Y_1, \ldots, Y_n)^T \) and \( \bar{m}^* = (m_1^*, \ldots, m_n^*)^T \). Summarizing above arguments, we finally prove that \( \sqrt{n}(\tilde{\beta} - \beta) \) is asymptotically normal with mean 0 and variance matrix
given by

\[ \lim_{n} n \left[ \sum_{t=1}^{n} a_t^{2-\rho} x_t \left\{ \mu x_t + \mathcal{D}_\beta(t) \right\}^\top \right]^{-1} \times \]

\[ \left\{ X A^{1-\rho} \text{Var}(Y - A m^*) A^{1-\rho} X^\top \right\} \left[ \sum_{t=1}^{n} a_t^{2-\rho} \left\{ \mu x_t + \mathcal{D}_\beta(t) \right\} x_t^\top \right]^{-1} \]

which may also be written as

\[ \lim_{n} n S^{-1}(\beta) V(\beta) S^{-\top}(\beta). \]

The proof of the asymptotic normality is completed.
State space models with non-stationary latent process

In this chapter, we present a class of state space models structured as follows: the latent process is assumed to be a Markov process including time-varying covariates and the observations are conditionally independent given the latent process, over time as well as over the components of the response vector.

The class of models generalizes the state space models discussed in the previous chapter in two aspects: (1) the models allow time-varying covariates to enter latent process, which results in a non-stationary latent process and (2) the models deal with multivariate time series data where the correlations over both time and components are introduced by latent process.

The models are inspired by the analysis of the Prince George data regarding the number of emergency room visits for four categories of respiratory diseases. Refer to Chapter 10 for details regarding these data. This provides a fully parametric alternative to the quasi-likelihood approach of Liang and Zeger (1986). Following the KEE method presented in Chapter 7, we establish the estimation for the regression parameters in the present chapter, and the analysis of residuals from both the observation and the latent process is also proposed for model diagnosis.

8.1 Introduction

Let $Y_t$ be a $d$-dimensional response vector such that the components of $Y_t$ (called categories) all reflect the same underlying tendency $\theta_t$, say, but where the categories may not
all have the same distribution. We assume that the latent process \( \theta_t \) is a Markov process including time-varying covariates and that the observations are conditionally independent given the latent process, both across categories and over time, giving a realistic correlation structure for many types of data. We consider a log-linear regression model for the marginal means \( \mathbb{E}(Y_t) \).

These state space models generalize the class of state space models with stationary latent process proposed in Chapter 7 in two aspects: (1) they allow the latent process to include the time-varying covariates and (2) they deal with the multivariate time series data. The two types of covariates that enter either via the latent process or via the observation model are called respectively long-term and short-term covariates, a terminology that derives from the fact that, for the long-term covariates, the Markov structure of the latent process creates a carry-over effect, which to some extent obviates the need for lagged covariates. Another important feature of these models is that the latent process naturally introduces both the correlation over time within each time series and the correlation across categories among all time series, giving a more realistic model setting to multivariate longitudinal data.

Additionally, similar to the previous chapter, the models are also based on distributions in the class of Tweedie exponential dispersion models (see Section 2.3). This accommodates a wide variety of data, including discrete, continuous and mixed data. The model provides a fully parametric alternative to the quasi-likelihood models of Liang and Zeger (1986) (c.f. Diggle, Liang and Zeger, 1994). However, the Tweedie models all have unbounded support, excluding in particular binary data from our method.

The class of state space models proposed in the present chapter is inspired by the special case of a Poisson response vector and a gamma latent process which is used for the analysis of the Prince George data in Chapter 10, where the relation between ambient air pollution and emergency room visits for respiratory diseases is discussed.
In this analysis the air pollution variables are taken as long-term covariates, the latent process represents the air pollution's current respiratory morbidity potential, and the short-term variables such as temperature and humidity affect the rate at which this potential morbidity produces respiratory disease. Note that in this case the categories all have the same marginal distribution.

As the estimation method used in the previous chapter for the models with stationary latent process, we also center our estimation approach for the state space models with non-stationary latent process at the Kalman estimating equation (KEE), corresponding to the EM algorithm with the E-step approximated by the Kalman smoother. We also use a Newton scoring algorithm in the M-step, which is an adaptation of Fisher’s scoring method to estimating equations. The dispersion parameters are estimated by a bias corrected version of the Pearson estimator. We also consider analysis of residuals from both the latent and observed parts of the model.

8.2 Models

Consider a $d$-dimensional vector of observations $Y_t$ recorded at equally spaced times $t = 1, \ldots, n$, and denote the full data vector by

$$Y = (Y_1^T, \ldots, Y_n^T)^T.$$ 

Let $\theta_t$ be a univariate latent process and denote the vector of $\theta$s by

$$\theta = (\theta_0, \theta_1, \ldots, \theta_n)^T,$$

where $\theta_0$ is an initializing variable. We assume that the $nd$ components of $Y$ are conditionally independent given $\theta$, and that the conditional distribution of $Y_t$ given $\theta$ depends only on $\theta_t$. Let $x_t \in \mathbb{R}^k$ and $z_t \in \mathbb{R}^l$ be two separate vectors of time-varying covariates,
containing respectively short-term and long-term covariates, none of which contains the constant covariate.

The conditional distribution of the *i*th category of $Y_t$ given $\theta_t$ is assumed to follow a Tweedie exponential dispersion model,

$$Y_{it} | \theta_t \sim Tw_{pi} \left( a_{it} \theta_t, \frac{\nu_i^2}{\theta_t^{p_i-1}} \right),$$  \hspace{1cm} \text{(8.2.1)}

for $i = 1, \ldots, d$, $t = 1, \ldots, n$, where

$$a_{it} = \exp \left( \mathbf{x}_t^{T} \alpha_i \right).$$  \hspace{1cm} \text{(8.2.2)}

The short-term covariates represent modulating factors that have an immediate effect on $Y_{it}$ relative to the value of $\theta_t$. Note that the corresponding regression parameters $\alpha_i \in \mathbb{R}^k$ may vary among the $d$ categories.

The shape parameters $p_i$ of the Tweedie model, which must satisfy $p_i \geq 1$ or $p_i \leq 0$, are known, but may vary from category to category, allowing each category to have a different distribution. The parameters $\nu_i^2$ in (8.2.1) are dispersion parameters, so that even in the case where all $p_i$s are equal, the $d$ categories may have different dispersions.

Note that $p_i \geq 1$ corresponds to non-negative distributions while $p_i \leq 0$ corresponds to distributions with support $\mathcal{R}$ (c.f. Section 2.3). In particular, $p_i = 0$ gives the normal distribution, $p_i = 1$ gives the Poisson distribution, $1 < p_i < 2$ are compound Poisson distributions with a positive probability in zero, $p_i = 2$ is the gamma distribution and $p_i = 3$ is the inverse Gaussian distribution. The Tweedie class hence accommodates a considerable range of distributions, including continuous, discrete and mixed ones.

In the Poisson case ($p_i = 1$), (8.2.1) reduces to

$$Y_{it} | \theta_t \sim Tw_1 \left( a_{it} \theta_t, \nu_i^2 \right) = \nu_i^2 Po(a_{it} \theta_t / \nu_i^2),$$

so in this case we take $\nu_i^2 = 1$ in all formulas in the following.
In (8.2.1) the conditional expectation is the first argument $a_i t \theta_t$ and the conditional variance is
\[
\text{Var}(Y_{it}|\theta_t) = \nu_i^2 a^2_{it} \theta_t.
\]
Hence $\theta_t$ may be interpreted as the "intensity" of the observation $Y_t$, being modified for the $i$th category by the covariates $x_t$ via $a_{it}$.

The latent process of the state space model is assumed to be an exponential dispersion model Markov process, defined by the transition distribution
\[
\theta_t|\theta_{t-1} \sim \text{T}_{q}(b_t \theta_{t-1}, \frac{\sigma^2}{\theta_{t-1}}) \tag{8.2.3}
\]
for $t = 1, \ldots, n$, where the parameter $q$ is known and satisfies $q \geq 2$, making the latent process positive. In particular, this excludes the possibility of a normal distribution for the latent process. The conditional expectation is $b_t \theta_{t-1}$ and the conditional variance is
\[
\text{Var}(\theta_t|\theta_{t-1}) = \sigma^2 b_t^2 \theta_{t-1}.
\]

The latent process may be interpreted as a random walk with drift, and is hence non-stationary.

The parameter $\sigma^2$ in (8.2.3) is a dispersion parameter, and $b_t$ depends on the long-term covariates $z_t$ via their increments $\Delta z_t = z_t - z_{t-1}$ as follows,
\[
b_t = \exp \left( \Delta z_t \beta \right), \tag{8.2.4}
\]
$\beta \in \mathbb{R}^l$ being a regression parameter. For convenience, we assume from now on that $z_0 = 0$, which may be obtained by subtracting $z_0$ from all $z_t$.

Note that the Markov structure of the latent process produces a carry-over effect, such that an increase in the level of $z$ at a given time $t$ may have an effect on subsequent time-points. An important characteristic of long-term covariates is that they have the
same effect on all \( d \) categories of \( Y_t \), in contrast to short-term covariates whose effects may vary across categories.

For a single series of observations \( Y \), we take \( \theta_0 = g = \exp(\gamma) \) to be degenerate and equal to the exponential of the constant covariate. The means of the latent and observed processes are log-linear in the covariates, as shown by

\[
E(\theta_t) = \tau_t = gb_1 \cdots b_t = \exp(\mathbf{z}_t^\top \beta + \gamma)
\]

and

\[
E(Y_{it}) = \exp(\mathbf{x}_t^\top \alpha_i + \mathbf{z}_t^\top \beta + \gamma).
\]

The marginal variance of the observed vector \( Y_t \) consists of two terms,

\[
\text{Var}(Y_t) = \Lambda_t \tau_t + \mathbf{a}_t \mathbf{a}_t^\top \text{Var}(\theta_t),
\]

where \( \Lambda_t = \text{diag}(\nu_1^2 a_{11}^{p_1}, \ldots, \nu_d^2 a_{dt}^{p_d}) \), \( \mathbf{a}_t = (a_{1t}, \ldots, a_{dt})^\top \) and \( \text{Var}(\theta_t) \) is given in (8.2.7) below. The second term shows overdispersion relative to the model (8.2.1) and shows that the correlation between categories is positive. The covariance between two observation vectors separated by a lag of \( s > 0 \) is

\[
\text{Cov}(Y_t, Y_{t+s}) = \mathbf{a}_t \mathbf{a}_{t+s}^\top b_{t+1} \cdots b_{t+s} \text{Var}(\theta_t).
\]

Figure 8.1 shows simulations of the model for \( q = 2 \) (a gamma latent process) and four categories with \( p_1 = 0, p_2 = 1, p_3 = 1.5 \) and \( p_4 = 2 \). The four observation processes represent the following characteristics, respectively: symmetric continuous (normal observation process), discrete (Poisson observation process), positive skew with positive probability at zero (compound Poisson observation process) and positive skew (gamma observation process).

Now consider the case where \( h \) independent series of observations are available indexed by \( j = 1, \ldots, h \), the \( j \)th series having length \( n_j \). We let \( Y_{ijt} \) denote the observation for
Figure 8.1: Simulation of 100 observations from a process with four categories with $q = 2$, $p_i = 0, 1, 1.5, 2$, $\nu_i^2 = 1$, $\sigma^2 = 0.05$ and $\theta_0 = 2$ without any covariate effects. Observations are shown with solid lines, and the latent processes with dashed lines.
category \( i \), series \( j \) and time \( t \). The vector of observations and the latent value at time \( t \) for series \( j \) are denoted \( Y_{jt} \) and \( \theta_{jt} \), respectively, and the corresponding values of the time-varying covariates are denoted \( z_{jt} \) and \( x_{jt} \). We assume that series \( j \) has an associated vector of covariates \( w_j \) that are constant over time, and which by convention contains the constant covariate. The covariates \( w_j \) represent characteristics specific to the \( j \)th observation series.

Let the conditional distribution of the data vector for series \( j \) given \( \theta_{j0} \) be given by the model (8.2.1)–(8.2.3) with parameters \( \alpha_{ij} \in \mathbb{R}^k, \beta_j \in \mathbb{R}^l \) depending on \( j \) but \( \sigma^2 \) being independent of \( j \). Let the marginal distribution of \( \theta_{j0} \) be

\[
\theta_{j0} \sim Tw_r(g_j, \omega^2),
\]

with

\[
g_j = \exp \left( w_j^T \gamma \right),
\]

where \( \gamma \in \mathbb{R}^m \) is a third regression parameter. The variance of \( \theta_{j0} \) is

\[
\text{Var}(\theta_{j0}) = \omega^2 g_j^2.
\]

The parameter \( \omega^2 \) is a dispersion parameter, reflecting random variation between series, whereas \( w_j \) reflects systematic variation. The variance of \( \theta_{jt} \) is

\[
\text{Var}(\theta_{jt}) = \sigma^2 \phi_{jt} r_{jt} + \omega^2 r_{jt}^2
\]

where

\[
\phi_{jt} = b_{jt}^{q-1} + b_{jt} b_{jt-1}^{q-1} + \cdots + b_{jt} b_{jt-1} \cdots b_{j1}^{q-1}.
\]

In this random effects model the marginal expectation of the observations are

\[
E(Y_{ijt}) = \exp \left( x_{jt}^T \alpha_{ij} + z_{jt}^T \beta_j + w_j^T \gamma \right).
\]

Note that if we interpret \( Tw_r(g, 0) \) as the degenerate distribution at \( g \), taking \( \omega^2 = 0 \) in (8.2.5) gives the special case of the general model (8.2.1)–(8.2.6) in which \( \theta_{j0} \) is constant and equal to \( g_j \).
8.3 Kalman filter and smoother

We now consider the Kalman filter and smoother for the model (8.2.1)-(8.2.6). For notational convenience we consider a single series of observations, suppressing the index \( j \) from the notation. The Kalman filter predicts a future observation \( Y_{t+1} \) and the current value of the latent process \( \theta_t \) based on past and current observations \( Y_1, \ldots, Y_t \). The smoother predicts the latent process from the total vector of observations \( Y_1, \ldots, Y_n \).

The innovations for the latent process are given by

\[
\xi_t = \theta_t - E(\theta_t|\theta_{t-1}) = \theta_t - b_t \theta_{t-1} \tag{8.3.1}
\]

and for the observed process

\[
\varphi_t = Y_t - E(Y_t|\theta^t) = Y_t - a_t \theta_t, \tag{8.3.2}
\]

where \( \theta^t = (\theta_0, \ldots, \theta_t)^T \). Using the innovations, the model may be put on additive form as follows,

\[
Y_t = a_t \theta_t + \varphi_t, \quad \theta_t = b_t \theta_{t-1} + \xi_t,
\]

where \( \varphi_t \) and \( \theta_t \) are uncorrelated, and so are \( \xi_t \) and \( \theta_{t-1} \). All \( 2n \) innovations are uncorrelated by the argument similar to those given in 7.3.1. Hence the model is structured as that of an ordinary Gaussian state space model. The variances of the innovations are

\[
\text{Var}(\xi_t) = E(\xi_t^2) = E\{\text{Var}(\theta_t|\theta^{t-1})\} = \sigma^2 b_t^q \tau_t
\]

and

\[
\text{Var}(\varphi_t) = E(\varphi_t \varphi_t^T) = \Lambda_t \tau_t. \tag{8.3.3}
\]

With the notations introduced in Chapter 6, we may either use the standard results under the ordinary Gaussian state space models (see e.g. Harvey, 1981) or apply the
results developed in Chapter 6 under a general setup to obtain the recursive formulas for prediction, Kalman filter and smoother. Let

\[ \theta_{t-1} | Y^{t-1} \sim [m_{t-1}; C_{t-1}] \]

and

\[ D_t = b_t C_{t-1} + \sigma^2 b_t^2 \tau_{t-1}, \]

where \( Y^{t-1} = (Y_1, \ldots, Y_{t-1}) \). The prediction is then given by

\[ Y_t | Y^{t-1} \sim [f_t; Q_t], \tag{8.3.4} \]

where

\[ f_t = a_t b_t m_{t-1}; \quad Q_t = b_t D_t a_t^T + \Lambda_t \tau_t. \]

The filtered values of the latent process are, for \( t = 1, \ldots, n \)

\[ \theta_t | Y^t \sim [m_t; C_t], \tag{8.3.5} \]

starting with \( m_0 = g \) and \( C_0 = \omega^2 g^* \), where

\[ m_t = b_t \{m_{t-1} + D_t a_t^T Q_t^{-1} (Y_t - f_t)\} \tag{8.3.6} \]

and

\[ C_t = b_t D_t \left(1 - b_t D_t a_t^T Q_t^{-1} a_t\right). \]

Given all \( n \) observations, the smoothed version of the latent process is given by the following backward recursion for \( t = n - 1, \ldots, 0 \),

\[ \theta_t | Y^n \sim [m^*_t; C^*_t], \tag{8.3.7} \]

starting with \( m^*_n = m_n \) and \( C^*_n = C_n \), where

\[ m^*_t = m_t + \frac{C_t}{D_{t+1}} (m^*_{t+1} - b_{t+1} m_t) \]

and

\[ C^*_t = \sigma^2 \frac{C_t}{D_{t+1}} b_t^2 \tau_{t+1} + \frac{C_t^2}{D_{t+1}^2} C^*_{t+1}. \]
8.4 Parameter estimation

We now consider parameter estimation for the model defined in Section 8.2 based on \( h \) independent series of observations, corresponding to \( N = n_1 + \cdots + n_h \) vectors of observations \( Y_{jt} \). The unknown parameters of this model are \( \alpha_{ij} \), \( \beta_j \), \( \gamma \), \( \nu_i^2 \), \( \sigma^2 \) and \( \omega^2 \).

Using a slightly different notation from Sections 8.1 and 8.3, we now denote the \( dN \times 1 \) data vector by 
\[
Y = \left( Y_{11}^T, \ldots, Y_{1n_1}^T, \ldots, Y_{h1}^T, \ldots, Y_{hn_h}^T \right)^T.
\]
and we define the vector of smoothed values by 
\[
m^* = (m_{10}, \ldots, m_{1n_1}, \ldots, m_{h0}, \ldots, m_{hn_h})^T.
\]
We also define the two parameter vectors
\[
\alpha = \left( \alpha_{11}^T, \ldots, \alpha_{d1}^T, \ldots, \alpha_{1h}^T, \ldots, \alpha_{dh}^T \right)^T, \quad \beta = \left( \beta_1^T, \ldots, \beta_h^T \right)^T.
\]

Using the similar arguments in Section 7.3.2, the mean squared error matrix \( C^* \), defined by
\[
C^* = E \left\{ \left( \theta - m^* \right) \left( \theta - m^* \right)^T \right\},
\]
has diagonal elements \( C_t^* \) and off-diagonal elements given by
\[
C_{t+s}^* = C_{t+s}^* \prod_{i=1}^{s} \frac{C_{t+i-1}}{D_{t+i}}.
\]

8.4.1 Kalman estimating equation

Following ideas presented in Chapter 7, we estimate the regression parameters based on the Kalman estimating equation. This estimation method corresponds to replacing the E-step of the EM algorithm by the Kalman smoother, but instead of the EM algorithm we use the Newton scoring algorithm.
Let us consider estimation of $\alpha$, $\beta$ and $\gamma$ for known values of the dispersion parameters $v^2$, $\sigma^2$ and $\omega^2$. We define the estimate $\hat{\eta}$ as the solution to the Kalman estimating equation $\psi(\eta) = 0$. Here $\eta = (\alpha^T, \beta^T, \gamma^T)^T$, and $\psi$ is the vector with components

$$
\psi_1(\eta) = X^T K_a^{-1} (Y - A m^*) ,
$$
(8.4.1)

$$
\psi_2(\eta) = \Delta Z^T K_b^{-1} B m^* ,
$$
(8.4.2)

and

$$
\psi_3(\eta) = W^T K_g^{-1} (m_0^* - g) ,
$$
(8.4.3)

where $A$ and $B$ are suitably defined matrices such that the elements of $Y - A m^*$ are $Y_t - a_t m_t^*$ and those of $B m^*$ are $m_t^* - b_t m_{t-1}^*$. The matrices $X$, $\Delta Z$ and $W$ are the design matrices corresponding to the generalized linear models (8.2.2), (8.2.4) and (8.2.6), and $K_a$, $K_b$ and $K_g$ and the vectors $g$ and $m_0^*$ are defined by

$$
K_a = \text{diag} \left\{ \nu_1^2 a_{111}^{-1}, \ldots, \nu_d^2 a_{dh}^{-1} \right\},
$$

$$
K_b = \text{diag} \left\{ b_1^{q_{-1}}, \ldots, b_{m_1}^{q_{-1}}, \ldots, b_h^{q_{-1}} \right\},
$$

$$
K_g = \text{diag} \left\{ g_1^{r_{-1}}, \ldots, g_h^{r_{-1}} \right\},
$$

and

$$
g = (g_1, \ldots, g_h)^T , \quad m_0^* = (m_{10}^*, \ldots, m_{h0}^*)^T .
$$

The estimating equation $\psi(\eta) = 0$ is unbiased, and the asymptotic standard errors of the estimator $\hat{\eta}$ may be calculated from the inverse of the Godambe information matrix,

$$
J(\eta) = S^T(\eta) V^{-1}(\eta) S(\eta) ,
$$

where $S(\eta)$ is the sensitivity matrix,

$$
S(\eta) = \text{E}_\eta \left( \frac{\partial \psi}{\partial \eta^T} \right) .
$$
and where

\[ V(\eta) = E_\eta \{ \psi(\eta)\psi^T(\eta) \} \]

is the \textit{variability matrix}. Both these matrices are calculated in Section 8.5. If the dispersion parameters are unknown, this gives approximately correct asymptotic standard errors if \( \psi \) depends little on \( \nu_i^2, \sigma^2 \) and \( \omega^2 \).

An analogue of Wald's test is available for testing the joint significance of a subset of regression parameters. The test statistic is defined as follows,

\[ W = \tilde{\eta}_1^T \{ J^{11}(\tilde{\eta}) \}^{-1} \tilde{\eta}_1, \]

where \( \eta_1 \) is the subvector whose significance we want to test, and \( J^{11}(\tilde{\eta}) \) is the corresponding block of the asymptotic covariance matrix of \( \tilde{\eta} \). Asymptotically, this statistic follows a \( \chi^2(u) \)-distribution, where \( u \) is the size of the subvector \( \eta_1 \).

The \textit{Newton scoring algorithm}, a parallel to Fisher's scoring method, is defined as the Newton algorithm applied to the equation \( \psi(\eta) = 0 \), but with the derivative of \( \psi \) replaced by its expectation \( S(\eta) \). The resulting algorithm gives the following updated value for \( \eta \),

\[ \eta^* = \eta - S^{-1}(\eta)\psi(\eta). \]  

An advantage of this algorithm is that the calculation of \( S(\eta) \) may be done recursively in parallel with the calculation of the Kalman smoother given in Section 8.5.

To start the algorithm, we first estimate the \( \theta \)-process for a given series as the average of the \( d \) components of \( Y_{jt} \). If \( d = 1 \) some smoothing is required, using for example a moving average of the series. Initial parameter estimates may then be obtained by fitting the three generalized linear models mentioned above.
8.4.2 Estimation of dispersion parameters

Based on the augmented data vector \((Y, \theta)\), simple estimates of the parameters \(\nu^2\), \(\sigma^2\) and \(\omega^2\) are *Pearson estimates*, based on weighted sums of the squared innovations, estimated from the Kalman smoother. These estimates would be downwards biased due to the substitution of the smoother. However, the bias may be expressed in terms of the mean square error matrix \(C^*\) of the smoother, as explained in the following, leading to the following adjusted *Pearson estimates* for the dispersion parameters.

Note that \(E(\xi^2_{jt}) = \sigma^2 b_{jt}^{-1} \tau_{jt}\), and, since \(\xi_{jt}\) and \(\tilde{\xi}_{jt}^* = m_{jt}^* - b_{jt} m_{jt-1}^*\) are uncorrelated, that

\[
\sigma^2 b_{jt}^{-1} \tau_{jt} = E(\tilde{\xi}_{jt}^2) + E\left(\xi_{jt} - \tilde{\xi}_{jt}^*\right)^2.
\]

Straightforward calculations (c.f. Jørgensen et al., 1995) yield

\[
E\left(\xi_{jt} - \tilde{\xi}_{jt}^*\right)^2 = C_{jt}^* + b_{jt}^* C_{jt-1}^* - 2 b_{jt} C_{jt-1}^* D_{jt}^*/D_{jt},
\]

Hence we obtain

\[
\sigma^2 = \frac{E(\tilde{\xi}_{jt}^2)}{b_{jt}^{-1} \tau_{jt}} + \frac{C_{jt}^* + b_{jt}^* C_{jt-1}^* - 2 b_{jt} C_{jt-1}^* D_{jt}^*/D_{jt}}{b_{jt}^{-1} \tau_{jt}},
\]

and an unbiased estimate for \(\sigma^2\) is

\[
\hat{\sigma}^2 = \frac{1}{N} \sum_{j=1}^{h} \sum_{t=1}^{n_j} \frac{\tilde{\xi}_{jt}^2}{b_{jt}^{-1} \tau_{jt}} + \frac{1}{N} \sum_{j=1}^{h} \sum_{t=1}^{n_j} \frac{C_{jt}^* + b_{jt}^* C_{jt-1}^* - 2 b_{jt} C_{jt-1}^* D_{jt}^*/D_{jt}}{b_{jt}^{-1} \tau_{jt}}.
\]  

(8.4.5)

where the last term corrects for the bias introduced by using the smoother for the latent process.

Similarly, estimates for \(\nu_i^2\) may be obtained from the following,

\[
\nu_i^2 a_{ijt}^2 \tau_{jt} = E\left(\tilde{\epsilon}_{ijt}^2\right) = E\left(\tilde{\xi}_{ijt}^2\right) + a_{ijt}^2 C_{jt}^* + \tilde{\varphi}_{jt}^* = Y_{jt} - a_{ijt} m_{jt}^*.
\]

This leads to the corresponding adjusted estimate for \(\nu_i^2\)

\[
\hat{\nu}_i^2 = \frac{1}{N} \sum_{j=1}^{h} \sum_{t=1}^{n_j} \frac{\tilde{\epsilon}_{ijt}^2}{\tau_{jt} a_{ijt}^2} + \frac{1}{N} \sum_{j=1}^{h} \sum_{t=1}^{n_j} \frac{a_{ijt}^2 C_{jt}^*}{\tau_{jt} a_{ijt}^2}.
\]

(8.4.6)
Based on the same arguments, the adjusted estimator for the random effect variance is

$$\hat{\omega}^2 = \frac{1}{h} \sum_{j=1}^{h} \frac{m_{ij}^2}{g_j^r} + \frac{1}{h} \sum_{j=1}^{h} \frac{C_{ijo}}{g_j^r}. \quad (8.4.7)$$

When the dispersion parameters are unknown, the Newton scoring algorithm swaps between updating $\eta$ via (8.4.4) and updating $\nu_t^2, \sigma^2$ and $\omega^2$ via (8.4.5), (8.4.6) and (8.4.7).

### 8.5 Derivation of Godambe information matrix

This section presents the calculations of the sensitivity matrix $S$ and the variability matrix $V$ entering in the Godambe information matrix $J = S^T V^{-1} S$ and in the Newton scoring algorithm.

#### 8.5.1 Some moment calculations

Since $\widetilde{\epsilon}_i = \epsilon_i - a_t(m_i^* - \theta_i)$, where $a_t\theta_i$ is the conditional expectation of $Y_t$ given $\theta_i$, we have for all $s, t > 0$

$$\text{Cov}(\epsilon_t, \epsilon_s) = \text{Cov}(\widetilde{\epsilon}_t^*, \widetilde{\epsilon}_s^*) + a_t a_s^T C_{ts}^*.$$

According to results in Section 8.3, the $\widetilde{\epsilon}_t^*$s are mutually uncorrelated, and we hence obtain

$$E(\widetilde{\epsilon}_t^* \widetilde{\epsilon}_s^{*T}) = \begin{cases} A_t \tau_t - a_t a_t^T C_{ts}^* & \text{for } t = s \\ -a_t a_s^T C_{ts}^* & \text{for } t \neq s. \end{cases} \quad (8.5.1)$$

A special case of this result is

$$\text{Cov} (\widetilde{\epsilon}_t^*, m_0^*) = a_t C_{t0}^*.$$

From Section 8.3 we know that $\text{Cov}(\epsilon, \theta) = 0$ and using again that $\theta_t = m_t^* - (m_t^* - \theta_t)$, we obtain

$$\text{Cov} (\widetilde{\epsilon}_t^*, \widetilde{\epsilon}_s^*) = a_t \text{Cov} (m_t^* - \theta_t, m_s^* - \theta_s - b_s(m_{s-1}^* - \theta_{s-1}))$$

$$= a_t C_{ts}^* - a_t b_s C_{t,s-1}^*. \quad (8.5.2)$$
To obtain the covariance matrix of $\xi^* = (\hat{\xi}_1^*, \ldots, \hat{\xi}_n^*)$ we note that the innovation $\xi_t$ may be written as a sum of two uncorrelated terms

$$\xi_t = \hat{\xi}_t^* + \{(\theta_t - m_t^*) - b_t(\theta_{t-1} - m_{t-1}^*)\}.$$

The innovations are mutually uncorrelated with variance $\sigma^2 b_t^2 \tau_t$, and since $\text{Var}(\theta - m^*) = C^*$, we find

$$E \left( \frac{\hat{\xi}_t^* \hat{\xi}_{t+s}^*}{\xi_t} \right) =
\begin{cases}
\sigma^2 b_t^2 \tau_t - C_t^* + 2b_t C_{t-1}^* - b_t^2 C_{t-1}^* & \text{for } s = 0 \\
-C_{t+s}^* + b_t C_{t+s}^* - b_t^2 C_{t+s}^* - b_t b_t C_{t+s}^* - b_t^2 C_{t-1}^* & \text{for } s > 0
\end{cases} \quad (8.5.3)$$

A special case of this formula is

$$E \left\{ (m_0^* - \tau_0) \hat{\xi}_t^* \right\} = -C_{0t}^* + b_t C_{0t-1}^*.$$

### 8.5.2 Derivatives of Kalman filter and smoother

In the calculation of the sensitivity matrix we need the expected values of the partial derivatives of $m_t$ and $m_t^*$ with respect to the parameters $\alpha$, $\beta$ and $\gamma$, denoted by

$$D_\alpha(t) = E \left( \frac{\partial m_t}{\partial \alpha^T} \right), \quad D_\alpha^*(t) = E \left( \frac{\partial m_t^*}{\partial \alpha^T} \right)$$

e etc., which we now calculate.

For the Kalman filter, we have from (8.2.5) and (8.3.6) that

$$m_0 = g = \exp(w^T \gamma) \quad \text{and} \quad m_t = m_t = b_t \{m_{t-1} + D_t a_t^T Q_t^{-1} (Y_t - f_t)\},$$

where $a_t$ is a function of $\alpha$, $b_t$ is a function of $\beta$ and $D_t$ and $Q_t$ are functions of $\alpha$ and $\beta$. Since $E(Y_t - f_t) = 0$, we obtain the recursion

$$D_\alpha(t) = b_t D_\alpha(t-1) - b_t^2 D_t a_t^T Q_t^{-1} E \left( \frac{\partial}{\partial \alpha^T a_t m_{t-1}} \right)$$

$$= b_t \left( 1 - b_t D_t a_t^T Q_t^{-1} a_t \right) D_\alpha(t-1) - b_t D_t \tau_t a_t^T Q_t^{-1} \frac{\partial}{\partial \alpha^T a_t}$$

$$= C_t D_\alpha(t-1) - b_t D_t \tau_t a_t^T Q_t^{-1} \frac{\partial}{\partial \alpha^T a_t}.$$
starting with $\mathcal{D}_\alpha(0) = 0$. Note that
\[
\frac{\partial}{\partial \alpha^T} a_t = \text{blockdiag} \left( x_t^T a_{1t}, \ldots, x_t^T a_{dt} \right).
\]

Analogously, we obtain
\[
\mathcal{D}_\beta(t) = \frac{C_t}{b_t D_t} \left\{ z_t^T r_t + b_t \mathcal{D}_\beta(t-1) \right\},
\]
starting with $\mathcal{D}_\beta(0) = 0$, and
\[
\mathcal{D}_\gamma(t) = \frac{C_t}{D_t} \mathcal{D}_\gamma(t-1),
\]
starting with $\mathcal{D}_\gamma(0) = w^T g$.

For the smoother we immediately obtain the recursions
\[
\mathcal{D}_\alpha^*(t) = \mathcal{D}_\alpha(t) + \frac{C_t}{D_{t+1}} \left\{ \mathcal{D}_\alpha^*(t+1) - b_{t+1} \mathcal{D}_\alpha(t) \right\},
\]
\[
\mathcal{D}_\beta^*(t) = \mathcal{D}_\beta(t) + \frac{C_t}{D_{t+1}} \left\{ \mathcal{D}_\beta^*(t+1) - b_{t+1} \mathcal{D}_\beta(t) - z_{t+1}^T r_{t+1} \right\},
\]
and
\[
\mathcal{D}_\gamma^*(t) = \mathcal{D}_\gamma(t) + \frac{C_t}{D_{t+1}} \left\{ \mathcal{D}_\gamma^*(t+1) - b_{t+1} \mathcal{D}_\gamma(t) \right\},
\]
all starting for $t = n$ with $\mathcal{D}_\alpha(n) = \mathcal{D}_\alpha(n)$, $\mathcal{D}_\beta^*(n) = \mathcal{D}_\beta(n)$ and $\mathcal{D}_\gamma^*(n) = \mathcal{D}_\gamma(n)$.

### 8.5.3 Variability matrix

We now switch to the notation of Section 8.4 again. By definition, the variability matrix for the estimating function $\psi$ is
\[
V = \begin{pmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_{31} & V_{32} & V_{33} \end{pmatrix}
\]
where $V_{ij} = E\{\psi_i \psi_j^T\}$. From Section 8.5.1 we have the variance-covariance structure of $Y - Am^*$, $Bm^*$ and $m_0^*$. Straightforward calculations give

$$V_{11} = X^T K_a^{-1} \text{Var}(Y - Am^*) K_a^{-1} X,$$

$$V_{12} = X^T K_a^{-1} \text{Cov}(Y - Am^*, Bm^*) K_b^{-1} \Delta Z,$$

$$V_{13} = X^T K_a^{-1} \text{Cov}(Y - Am^*, m_0^*) K_g^{-1} W,$$

$$V_{22} = \Delta Z^T K_b^{-1} \text{Var}(Bm^*) K_b^{-1} \Delta Z,$$

$$V_{23} = \Delta Z^T K_b^{-1} \text{Cov}(Bm^*, m_0^*) K_g^{-1} W$$

and

$$V_{33} = W^T K_g^{-1} \text{Var}(m_0^*) K_g^{-1} W.$$

The variances and covariances involved in these expressions may be calculated from the results of Section 8.5.1 by noting that $Y - Am^*$ is the vector of predicted innovations $\hat{e}_j^*$ and $Bm^*$ is the vector of predicted innovations $\hat{\xi}_j^*$.  

### 8.5.4 Sensitivity matrix

As in the previous section, we may write the sensitivity matrix for $\psi$ on the form

$$S = \begin{pmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{pmatrix}$$

where $S_{ij} = E\{\partial \psi_i / \partial \eta_j^T\}$, with $j = 1, 2, 3$ referring to $\alpha$, $\beta$ and $\gamma$, respectively. Using the derivatives of the Kalman smoother from Section 8.5.2 we obtain

$$S_{11} = -X^T K_a^{-1} \left( \frac{\partial A}{\partial \alpha} \tau + AD^* \right),$$
where $D^*_{\alpha}$ denotes the vectors $D_{\alpha}(t)$ stacked in the appropriate order, and where

$$\left( \frac{\partial A}{\partial \alpha^T} \right)_{ij} = \sum_{i=1}^{n} \frac{\partial a_{it}}{\partial \alpha_j}. $$

The remaining blocks of $S$ are, using a similar notation,

$$S_{12} = -X^T K_a^{-1} AD_{\beta},$$

$$S_{13} = -X^T K_a^{-1} AD_{\gamma},$$

$$S_{21} = \Delta Z^T K_b^{-1} BD_{\alpha},$$

$$S_{22} = \Delta Z^T K_b^{-1} \left( \frac{\partial B}{\partial \beta^T} + BD_{\beta} \right),$$

$$S_{23} = \Delta Z^T K_b^{-1} BD_{\gamma},$$

$$S_{31} = W^T K_g^{-1} D_0^* \alpha,$$

$$S_{32} = W^T K_g^{-1} D_0^* \beta$$

and

$$S_{33} = W^T K_g^{-1} \left( D_0^* \gamma - \frac{\partial g}{\partial \gamma^T} \right).$$

Here a subscript 0 refers to differentiation of $m_0^*.$

8.6 Residual analysis

Residual analysis for both the observed and unobserved parts of the model is an important ingredient of our approach. The basic idea is to use plots of standardized residuals (i.e. residuals divided by their standard deviations) in much the same way as in generalized linear models, in order to check the distributional assumptions and the regression part of the model. We may also check the correlation structure of the model by means of methods from time-series analysis. We may consider each series separately, or we may combine residuals from several series.
Chapter 8. State space models with non-stationary latent process

The main type of residuals are the conditional residuals, defined as the predicted values of the innovations $\varphi$ and $\xi$, based on either the Kalman filter or smoother. All residuals have means 0. The properties derived in the following do not take into account the effect of substituting parameter estimates.

The predicted values of the innovations based on the Kalman filter are

$$\hat{\varphi}_{jt} = Y_{jt} - f_{jt} \quad \text{and} \quad \hat{\xi}_{jt} = m_{jt} - a_{jt}m_{jt}.$$  

The prediction errors $\hat{\varphi}_{jt}$ are mutually uncorrelated over time, and since $\hat{\xi}_{jt} = b_{jt}D_{jt}a_{jt}Q_{jt}^{-1}\hat{\varphi}_{jt}$, then so are the predicted innovations of the latent process. This property of the filter residuals $\hat{\varphi}_{jt}$ and $\hat{\xi}_{jt}$ makes them especially useful for residual plots. Moreover, $\hat{\xi}_{jt}$ is uncorrelated with $f_{jt}$, and $\hat{\xi}_{jt}$ is uncorrelated with $b_{jt}m_{jt-1}$. The variances of the two sets of residuals are, respectively,

$$\text{Var} \left( \hat{\varphi}_{jt} \right) = Q_{jt} \quad \text{and} \quad \text{Var} \left( \hat{\xi}_{jt} \right) = b_{jt}D_{jt} - C_{jt}.$$

A disadvantage of the filter residuals is that their variances may be large for the first few observations of each series. This disadvantage is not shared by residuals based on the Kalman smoother, which in fact have smaller variances than the corresponding filter residuals. These residuals are

$$\hat{\varphi}^*_{jt} = Y_{jt} - a_{jt}m^*_j \quad \text{and} \quad \hat{\xi}^*_{jt} = m^*_j - b_{jt}m^*_{j,t-1}.$$  

However, the smoother residuals do not have the same uncorrelatedness properties as the filter residuals, making the smoother residuals less useful for some purposes. Their variances are

$$\text{Var} \left( \hat{\varphi}^*_{jt} \right) = \Lambda_{jt}r_{jt} - a_{jt}a_{jt}^T C^*_j$$

and

$$\text{Var} \left( \hat{\xi}^*_{jt} \right) = \sigma^2b_{jt}^2r_{jt}^2 - C^*_j - b_{jt}^2C^*_{j,t-1} + 2b_{jt}C_{jt-1}C^*_j.$$
Standardized residuals (having unit variance) are denoted $\hat{\xi}_{jt}$ etc. The residuals $\varphi_{jt}$ are standardized componentwise.

The correlation structure of the model is easy to check by means of the autocorrelation functions of the standardized filter residuals (the $\hat{\varphi}_{ijt}$s are considered separately for each $i$). We may also plot $\hat{\xi}_{jt}$ against $\hat{\xi}_{j,t-1}$, to check the Markov assumption for the latent process, and $\hat{\varphi}_{ijt}$ against $\hat{\varphi}_{ij,t-1}$, to check the conditional independence of the counts over time.

To check the assumption that the categories are conditionally independent given the latent process, we consider the empirical variance-covariance matrices of the vector of standardized prediction errors,

$$\frac{1}{n_j} \sum_{t=1}^{n_j} Q_{jt}^{-1/2} \varphi_{jt} \varphi_{jt}^T Q_{jt}^{-T/2},$$

whose expectations are all the $d \times d$ identity matrix. The off-diagonal elements have asymptotic standard errors $n_j^{-1/2}$. Note that the standardization depends on the version of the square-root matrix $Q_{jt}^{1/2}$ chosen.

As in generalized linear models, we may check the form of the variance functions (and hence distributional forms) by plotting the standardized residuals against the corresponding log fitted values. A "megaphone" shape would indicate that the chosen shape parameter in the corresponding variance function is incorrect. In the case of several series we may use the standardized smoother residuals $m_{j0}^* - g_j$ for checking the form of the random effects distribution. It is crucial to use the smoother residuals for this purpose, because, by definition, $m_{j0} = g_j$, which is not random.

Plots of standardized residuals against each covariate are useful for detecting non-linearity of the model. The smoother residuals are the best choice of residuals for this purpose. To check the log link assumption, we plot $\log Y_{ijt}$ against the log fitted values $\log(a_{jt} m_{jt}^*)$. This plot should show a horizontal linear relationship; a curvature or other
unusual shape would indicate inadequacy of the log link assumption.

Residual analysis may also help to determine whether a covariate is long-term or short-term. The basic idea is that a short-term covariate would show an association with the observation residuals ($\hat{\phi}$ or $\hat{\phi}^*$) if it had been incorrectly fitted as a long-term covariate, and similarly, a long-term covariate would show association with the latent process residuals ($\hat{\xi}$ or $\hat{\xi}^*$) if it had been incorrectly fitted as a short-term covariate.
Chapter 9

Simulation studies under Poisson-gamma models

In this chapter, Poisson-gamma models are used for our simulation studies to help address the two issues: (1) the comparative efficiency between the KEE approach and the Monte Carlo EM approach and (2) the utility of the residual analysis for detecting misspecifications between stationarity and non-stationarity for latent process.

For issue (1), our study is based on a one-dimensional observed Poisson process (with one category) and non-stationary gamma latent process including time trend covariate. With regard to the model with stationary gamma latent process, a comparison has been made on the basis of the analysis of the polio data (see Table 7.1). As shown in Chapter 7, the KEE approach gives competitive estimates of regression parameters compared with the MCEM approach. So, in this chapter we decided to conduct our comparison on the model with non-stationary latent process. As the time-varying covariates enter the latent process, the feature of the latent process is substantially changed to non-stationarity, which increases the complexity of the model and complicates the statistical inference. As seen in Chapter 8, different covariates may play different roles in statistical models, and hence arranging them on different levels in models according to their nature is reasonable approach. The basic goal for the comparison is to understand the differences of the corresponding estimates of the coefficients as well as the standard errors given by the MCEM and KEE approaches. As a matter of fact, Chapter 8 suggests using the Godambe information matrix as the asymptotic variance matrix from which the standard errors may be computed. But such a suggestion is rooted in the asymptotic result which

143
was only proven when observations are independent (c.f. Godambe, 1976). Therefore, the comparison is expected to provide a kind of guidance for the further application of KEE approach, in particular as regards calculation of standard errors. In addition, the derivation of the MCEM method presented in this chapter may be applicable to a state space model whose latent process includes many covariates rather than one covariate (time trend) considered in our simulation study.

The simulation study for question (2) mainly concerns misspecification of the latent process. Apparently, incorporating a stationary or non-stationary structure in latent process is crucial for modelling and hence for analysis of data. It seems rather important to stress the utility of the residual analyses developed in respectively Chapter 7 and 8 for detecting the misspecification of stationarity and non-stationarity for the latent process. We use the strategy of a case-and-control study; that is, we generate a data set from the model with a stationary latent process but fit the generated data by the model with a non-stationary latent structure, and then apply the diagnostic tools of model checking to detect the inadequacy for such a switch. As shown in the simulation study, the residual analyses proposed in the previous chapters effectively detect the mistakes of the modelling structures and hence assure appropriate model assumptions in the analysis of data.

9.1 Models

We use two types of Poisson-gamma models for our simulation studies in this chapter. They are:

**Model I:** The same model as studied in Section 7.6.1 with stationary latent process,

\[ Y_t | \theta_t \sim Po(A_t \theta_t), \quad t = 1, \ldots, n \]

where \( A_t = \exp \{ X_t^T \eta \} \), \( X_t = (t, \cos(2\pi t/12), \sin(2\pi t/12), \cos(2\pi t/6), \sin(2\pi t/6))^T \) includes time-trend and seasonal effects covariates, and \( \eta \in \mathcal{R}^5 \) is the parameter
Chapter 9. Simulation studies under Poisson-gamma models

of interest. Note that in the model we use slightly different notation as before in order to distinguish it from the Model II given below.

The stationary latent process is assumed to be

$$\theta_t = G_t \theta_{t-1} + \epsilon_t, \quad t = 1, \ldots, n$$  \hspace{1cm} (9.1.1)

where $G_t$ are i.i.d random variables with beta distribution $\text{Beta}\{\rho \lambda, (1-\rho) \lambda\}$ where $\rho$ is the autocorrelation.

**Model II:** The following one-dimensional Poisson-gamma model is a special case of the models proposed in Chapter 8,

$$Y_t|\theta_t \sim P_0(a_t \theta_t), \quad t = 1, \ldots, n$$

where $a_t = \exp \left\{ x_t^\top \alpha \right\}$, 

$$x_t = (\cos(2\pi t/12), \sin(2\pi t/12), \cos(2\pi t/6), \sin(2\pi t/6))^\top$$

includes only the seasonal effects covariates and $\alpha \in \mathcal{R}^4$. The latent process is the first-order Markov gamma process

$$\theta_t|\theta_{t-1} \sim Ga \left( b \theta_{t-1}, \sigma^2/\theta_{t-1} \right), \quad t = 1, \ldots, n,$$  \hspace{1cm} (9.1.2)

where $\theta_0$ is set to 1 and $b = e^\beta \in \mathcal{R}$. This allows the time-trend (assumed to be a long-term covariate) to enter the latent process and hence results in a non-stationary latent process. Both $\beta$ and $\alpha$ are parameters of interest in the model, and $\sigma^2$ is the dispersion parameter.

Note that in both models, we rule out the intercept covariate due to the issue of model identification, which is different from the way we did for the analysis of the polio data in Section 7.6.1 where we imposed the condition $E(\theta_t) = 1$ instead. Also note that only Model II is used for the study of question (1).
9.2 Comparison between MCEM and KEE

It follows from Model II that the conditional moment structures are given by

\[ E(Y_t | \theta_t) = a_t \theta_t, \quad \text{Var}(Y_t | \theta_t) = a_t \theta_t, \]

and

\[ E(\theta_t | \theta_{t-1}) = b \theta_{t-1}, \quad \text{Var}(\theta_t | \theta_{t-1}) = b^2 \theta_{t-1} \sigma^2. \]

Up to the condition on \( \theta_0 \), the marginal mean for the observed process is

\[ \log E(Y_t) = x_t^T \alpha + \beta t \]

which is the same marginal model as in the stationary case, although the interpretations of the parameters are different.

Let \( \eta = (\alpha^T, \beta)^T \) be the vector of parameters to be estimated. Also let \( Y = (Y_1, \ldots, Y_n)^T \) and \( \theta = (\theta_0, \ldots, \theta_n)^T \). Conditionally on the initial latent variable \( \theta_0 \), the full log-likelihood of the augmented data \((Y, \theta)\) is

\[ \ell(\eta) = \ell_{Y | \theta}(\alpha) + \ell_{\theta | \theta_0}(\beta) \quad (9.2.1) \]

which shows that the parameter vector \( \alpha \) and the parameter \( \beta \) are \( \lambda \)-independent. Apart from constant terms, the likelihood for \( \alpha \) is

\[
\ell(\alpha) = \ell_{Y | \theta}(\alpha) = \sum_{t=1}^{n} \left\{ Y_t \log(\theta_t) + Y_t \log(a_t) - a_t \theta_t \right\} = \sum_{t=1}^{n} \left\{ -\theta_t \exp(x_t^T \alpha) + Y_t \log(\theta_t) + Y_t x_t^T \alpha \right\}. \quad (9.2.2)
\]

By letting \( \lambda = 1/\sigma^2 \), the likelihood for \( \beta \) is

\[ \ell(\beta) = \ell_{\theta | \theta_0}(\beta) = \sum_{t=1}^{n} \left\{ \theta_{t-1} \lambda \log \theta_t - \beta \lambda \theta_{t-1} + \lambda \theta_{t-1} \log \theta_t - \log \theta_t - \lambda \theta_t e^{-\beta} - \log \Gamma(\lambda \theta_{t-1}) \right\} \]
Chapter 9. Simulation studies under Poisson-gamma models

\[ \lambda (\log \lambda - \beta) \sum_{t=1}^{n} \theta_{t-1} + \lambda \sum_{t=1}^{n} \theta_{t-1} \log(\theta_{t}) - \sum_{t=1}^{n} \log(\theta_{t}) - \]
\[ \lambda e^{-\beta} \sum_{t=1}^{n} \theta_{t} - \sum_{t=1}^{n} \log(\lambda \theta_{t-1}). \]

The EM algorithm is natural to use for estimating the parameter vector \( \eta \) by treating the latent process \( \{\theta_{t}\} \) as a set of missing values. The real difficulty for the implementation of EM-algorithm is that the E-step cannot be computed directly and easily. An approximation to the E-step may be given by using a sample average and the Monte Carlo method, which is discussed in detail in the following section.

9.2.1 MCEM algorithm

The EM algorithm consists of two steps:

- **E step**: given the current \( \eta^{(i)} = (\alpha^{(i)} \beta^{(i)})^{T} \) and \( Y \), compute

\[ Q(\eta, \eta^{(i)}) = E \log f(\eta|\theta, Y) \]  \hspace{1cm} (9.2.3)

where the expectation is taken under the probability density \( f(\theta|\eta^{(i)}, Y) \).

- **M step**: find the maximizer \( \eta^{(i+1)} \) of \( Q(\eta, \eta^{(i)}) \).

The algorithm is iterated until \( \| \eta^{(i+1)} - \eta^{(i)} \| \) or \( |Q(\eta^{(i+1)}, \eta^{(i)}) - Q(\eta^{(i)}, \eta^{(i)})| \) is sufficiently small. When the E step cannot be implemented explicitly, we use the Monte Carlo method to approximate the \( Q \) function. In particular, the Monte Carlo E-step is given as

(a) Draw \( \theta_{1}, \ldots, \theta_{m} \stackrel{iid}{\sim} f(\theta|Y, \eta^{(i)}) \)

(b) Let \( \hat{Q}_{i+1}(\eta, \eta^{(i)}) = \frac{1}{m} \sum_{j=1}^{m} \log f(\eta|\theta_{j}, Y) \)
Accordingly, the M-step maximizes the $\hat{Q}_{i+1}$ to obtain $\eta^{(i+1)}$. It is worth noting that the step (a) in the Monte Carlo E-step is usually conducted by the Gibbs sampler approach that generates an ergodic Markov chain $\theta_1, \ldots, \theta_m$ whose invariant distribution is the conditional distribution $f(\theta|Y, \eta^{(i)}).$ The ergodicity implies that as $m \to \infty$, $\hat{Q}_{i+1}(\eta, \eta^{(i)}) \overset{a.s.}{\to} Q(\eta, \eta^{(i)}).$ Gelfand and Smith (1990) advocated using every $h$-th ($h \approx 50$) value in the sequence to obtain more independent contributions, but some authors such as Zeger and Karim (1991) opposed such an implementation and insisted on using the whole sequence for the computation. In our simulation studies, we have tried both; since there is no big difference between the two implementations in our computation of the E-step, we use the whole sequence for our computations.

Within the context of Monte Carlo EM, the observed Fisher information matrix of the estimator $\hat{\eta} = (\hat{\alpha}^T, \hat{\beta})^T$ may be approximately estimated via the complete information minus the missing information, given as respectively

$$\text{complete info.} = \begin{pmatrix} \frac{1}{m} \sum_{j=1}^{m} I_{\alpha \alpha, j} & 0 \\ 0 & \frac{1}{m} \sum_{j=1}^{m} I_{\beta \beta, j} \end{pmatrix}$$

and

$$\text{missing info.} = \begin{pmatrix} \frac{1}{m} \sum_{j=1}^{m} \ell_{\alpha, j}(\alpha) \ell_{\alpha, j}^T(\alpha) - \bar{\ell}_\alpha(\alpha) \bar{\ell}_\alpha^T(\alpha) & \frac{1}{m} \sum_{j=1}^{m} \ell_{\alpha, j}(\alpha) \ell_{\beta, j}(\beta) - \bar{\ell}_\alpha(\alpha) \bar{\ell}_\beta(\beta) \\ \frac{1}{m} \sum_{j=1}^{m} \ell_{\beta, j}(\beta) \ell_{\alpha, j}^T(\alpha) - \bar{\ell}_\beta(\beta) \bar{\ell}_\alpha^T(\alpha) & \frac{1}{m} \sum_{j=1}^{m} \ell_{\beta, j}(\beta) \ell_{\beta, j}(\beta) - \bar{\ell}_\beta(\beta) \bar{\ell}_\beta^T(\beta) \end{pmatrix}$$

where

$$I_{\alpha \alpha, j} = \frac{-\partial^2 \log f(\eta|Y, \theta_j)}{\partial \alpha \partial \alpha^T} |\hat{\alpha}, \quad I_{\beta \beta, j} = \frac{-\partial^2 \log f(\eta|Y, \theta_j)}{\partial \beta \partial \beta} |\hat{\beta},$$

$$\bar{\ell}_\alpha(\alpha) = \frac{1}{m} \sum_{j=1}^{m} \ell_{\alpha, j}(\alpha) = \frac{1}{m} \sum_{j=1}^{m} \frac{\partial \log f(\eta|Y, \theta_j)}{\partial \alpha} |\hat{\alpha}$$

and

$$\bar{\ell}_\beta(\beta) = \frac{1}{m} \sum_{j=1}^{m} \ell_{\beta, j}(\beta) = \frac{1}{m} \sum_{j=1}^{m} \frac{\partial \log f(\eta|Y, \theta_j)}{\partial \beta} |\hat{\beta}.$$
Chapter 9. Simulation studies under Poisson-gamma models

Letting $B = \sum_{t=1}^{n} \theta_{t-1}$, $C = \sum_{t=1}^{n} (Y_{t} + \lambda \theta_{t-1}) \log(\theta_{t})$, $D = \sum_{t=1}^{n} \log(\theta_{t})$, $E = \sum_{t=1}^{n} \theta_{t}$ and $F = \sum_{t=1}^{n} \log(\lambda \theta_{t-1})$, we may re-write the full log-likelihood of $\eta$ by

$$
\ell(\eta) = \sum_{t=1}^{n} \left\{ -\theta_{t} \exp\left( x_{t}^\top \alpha \right) + Y_{t} x_{t}^\top \alpha \right\} + \lambda (\log \lambda - \beta) B + C - D - \lambda e^{-\beta} E - F. \tag{9.2.6}
$$

where $E - B = \theta_{n} - \theta_{0}$. Clearly, the log-likelihood $\ell(\eta)$ is linear in $B, C, D, E$ and $F$.

So, for given $\theta_{1}, \ldots, \theta_{m} \sim f(\theta|Y, \eta^{(i)})$, the Monte Carlo E-step is given by

$$
\hat{Q}_{i+1}(\eta, \eta^{(i)}) = \sum_{t=1}^{n} \left\{ -\bar{\theta}_{t} \exp\left( x_{t}^\top \alpha \right) + Y_{t} x_{t}^\top \alpha \right\} + \lambda (\log \lambda - \beta) B + C - D - \lambda e^{-\beta} E - F \tag{9.2.7}
$$

where we use bars to denote the sample averages for corresponding variables; for instance, $\bar{\theta}_{t} = \frac{1}{m} \sum_{j=1}^{m} \theta_{i,j}$ and $\theta_{i,j}$ is the $i$-th component of $\theta_{j}$.

For the M-step, by taking the derivatives of the function $\hat{Q}_{i+1}$ with respect to the parameter $\alpha$, we obtain

$$
\sum_{t=1}^{n} x_{t} (Y_{t} - a_{t} \bar{\theta}_{t}) = 0 \tag{9.2.8}
$$

and the estimator $\hat{\alpha}^{(i+1)}$ is the solution to the equation (9.2.8). To solve the equation, we may either use the GLIM software for Poisson regression or use the Newton-Raphson algorithm directly. Similarly, by differentiating $\hat{Q}_{i+1}$ with respect to the parameter $\beta$, we obtain

$$
\hat{\beta}^{(i+1)} = \log(\bar{E}) - \log(\bar{B}). \tag{9.2.9}
$$

The standard errors of the estimators $\hat{\alpha}$ and $\hat{\beta}$ may be given as the square roots of the diagonal elements of the inverse of the observed Fisher information matrix approximated as follows,

$$
\text{Fisher info.} \approx \text{complete info.} - \text{missing info.} \tag{9.2.10}
$$

where both the complete info. and missing info. are approximated on the basis of the estimated likelihood $\hat{Q}_{i}$ for such a large $i$ that convergence is obtained.
Chapter 9. Simulation studies under Poisson-gamma models

One may also update the estimate of the dispersion parameter $\sigma^2$ in the MCEM procedure, although sometimes it goes out of control in simulated samples for some unclear reason. Here we propose two types of estimators for $\sigma^2$. One is based on the fact that

$$\text{Var}(Y_i) = a_t b^t + a_t^2 b^{t+1} \frac{(1 - b^t)}{1 - b} \sigma^2.$$

We hence propose a moment estimator of $\sigma^2$ as

$$\hat{\sigma}^2 = \frac{(1 - b) \sum_{t=1}^{n} \left\{ (Y_i - a_t b^t)^2 - a_t b^t \right\}}{b \sum_{t=1}^{n} a_t^2 b^t (1 - b^t)}. \quad (9.2.11)$$

Another estimator is based on the likelihood function using Stirling’s approximation for the $\Gamma$ function. The estimate of $\sigma^2$ or $\lambda$ is obtained by taking derivative with respect to the parameter $\lambda$ in equation (9.2.7),

$$\hat{\lambda} = \frac{2}{n} \left( \beta \bar{B} + b^{-1} \bar{E} - \bar{G} - \bar{B} \right), \quad (9.2.12)$$

where $\bar{G}$ is defined as the average of the term $\sum_t \log(\theta_t)$ over the samples $\theta_1, \ldots, \theta_m$.

We now derive the formula for computing the approximate observed Fisher information matrix as follows. Suppose $\hat{\eta} = \left( \hat{\alpha}^T, \hat{\beta} \right)^T$ is the parameter estimate obtained from the MCEM algorithm and $\bar{t}_i$ are the smoothed means output at the last iteration. For the complete information,

$$\frac{1}{m} \sum_{j=1}^{m} I_{\alpha,j} = \frac{1}{m} \sum_{j=1}^{m} \sum_{t=1}^{n} \theta_{t,j} a_t x_t x_t^T = \sum_{t=1}^{n} a_t x_t x_t^T \bar{t}_t. \quad (9.2.13)$$

and

$$\frac{1}{m} \sum_{j=1}^{m} I_{\beta,j} = \frac{1}{m} \sum_{j=1}^{m} \lambda e^{-\beta} \sum_{t=1}^{n} \theta_{t,j} = \lambda e^{-\beta} \bar{E}. \quad (9.2.14)$$

For the missing information,

$$\ell_{\alpha,j}(\alpha) = \sum_{t=1}^{n} x_t (Y_t - a_t \theta_{t,j}), \quad \ell_{\beta,j}(\beta) = \left( -\lambda \sum_{t=1}^{n} \theta_{t-1,j} + \lambda e^{-\beta} \sum_{t=1}^{n} \theta_{t,j} \right). \quad (9.2.15)$$
Also
\[ \ell_\alpha(\alpha) = \frac{1}{m} \sum_{j=1}^{m} \sum_{t=1}^{n} x_t(Y_t - a_t \theta_{t,j}) = \sum_{t=1}^{n} x_t(Y_t - a_t \bar{\theta}_t) \] 
(9.2.16)
and
\[ \ell_\beta(\beta) = \frac{1}{m} \sum_{j=1}^{m} \sum_{t=1}^{n} \left( -\lambda \sum_{i=1}^{n} \theta_{i-1,j} + \lambda e^{-\beta} \sum_{i=1}^{n} \theta_{i,j} \right) = -\lambda \bar{B} + \lambda e^{-\beta} \bar{E}. \] 
(9.2.17)

Notice that both \( \ell_\alpha(\hat{\alpha}) \approx 0 \) and \( \ell_\beta(\hat{\beta}) \approx 0 \) since \( \hat{\alpha} \) and \( \hat{\beta} \) are maximum likelihood estimators making the first-order derivatives zero. Therefore, Tanner (1991) suggested approximating the observed Fisher information matrix by excluding the terms of \( \ell_\alpha(\alpha) \) and \( \ell_\beta(\beta) \) in the missing information matrix given by equation (9.2.5).

9.2.2 Gibbs sampler

In order to implement the MCEM algorithm, we must sample from the conditional distribution \( f(\theta|Y, \eta) \) where \( \eta = (\alpha^T, \beta)^T \) is given by the previous M-step. We use the Gibbs sampler to generate the required Markov chain. Let us first review the method by considering only the case of three variables \( U, V \) and \( W \). Suppose that the conditional distribution of each, given the remaining two has a simple form while the joint distribution is complicated. For simplicity, denote the conditional distributions by \([U|V, W], [V|U, W] \) and \([W|U, V] \), respectively. Then the Gibbs sampler generates a random variate from the joint distribution \([U, V, W] \) as follows. Given arbitrary starting values, \( U_0, V_0, W_0 \), draw \( U_1 \) from \([U|V_0, W_0] \), then draw \( V_1 \) from \([V|U_1, W_0] \), and finally, complete the first iteration by drawing \( W_1 \) from \([W|U_1, V_1] \). After a large number, \( m \), of iterations, we obtain \( U_m, V_m, W_m \). Geman and Geman (1984) showed that under mild conditions, the joint distribution of \((U_m, V_m, W_m) \) converges at an exponential rate to the joint distribution \([U, V, W] \) as \( m \to \infty \).

With regard to our implementation, each iteration needs to generate an \( n \)-vector from \( n \) conditional densities, respectively. In effect, we incorporate the so-called modified
rejection sampling in the Gibbs sampling algorithm to generate each component \( \theta_t \) of \( \theta \) for \( t \neq n \), and this modified version of rejection sampling is given as follows.

The rejection sampling (c.f. Ripley, 1987) is a generating method for univariate distributions. Assume that we want to sample from a density \( f(x) \), say, which is rather difficult to do directly. Assume that a constant \( c \geq 1 \) is available such that \( c \cdot g(x) \geq f(x) \) where \( g(x) \) is a density from which sampling is easy to perform. Then the following steps result in a random variate \( x \) with density \( f(x) \):

1. Generate a random variate \( x^* \) from \( g(x) \)
2. Generate a uniform variate \( u \) from \( U(0,1) \)
3. If \( f(x^*) / \{c \cdot g(x^*)\} < u \), assign \( x = x^* \); otherwise return to step (1).

The modified version of rejection sampling is described as follows. Suppose now that the density \( f(x) \) is only known up to a multiplicative constant \( c_f \), say, \( f(x) = c_f f^*(x) \). And assume similarly that another density is given by \( g(x) = c_g g^*(x) \) where \( c_g \) might be also unknown. If, by the known constant \( K \), we have \( f^*(x) < Kg^*(x) \), a random variate \( x \) can be generated from density \( f(x) \) by the following modified rejection sampling:

1. Generate a random variate \( x^* \) from \( g(x) \)
2. Generate a uniform variate \( u \) from \( U(0,1) \)
3. If \( f^*(x^*) / \{K \cdot g^*(x^*)\} < u \), assign \( x = x^* \); otherwise return to step (1)

To find the constant \( K \) for our Poisson-gamma model, we need the following lemma.

**Lemma 9.1** The function \( f(z) = z^{1/z} \), \( z > 0 \) is bounded and has a (unique) maximum \( e^{1/e} \) at \( z = e \).
**Proof** Clearly, the function $f(z)$ is continuous on $(0, \infty)$. For $z \in (0, 1]$, $f(z) \leq 1$. Since

$$
\frac{\log(z)}{z} \to 0, \quad z \to \infty,
$$

we know that

$$
f(z) = \exp\left(\frac{\log(z)}{z}\right) \to 1, \quad z \to \infty.
$$

By taking the first derivative of $f(z)$, we get the maximum $e^{1/e}$ of $f(z)$ at $z = e$.  

Let $\theta_{-t}$ be $\theta$ with $\theta_t$ being omitted. Under our Poisson-gamma setting, we attempt to sample the $t$-th component $\theta_t$ from $f_{\eta}(\theta_t|\theta_{-t}, Y)$. Note that for $t \neq n$,

$$
f_{\eta}(\theta_t|\theta_{-t}, Y) \propto f_{\eta}(\theta, Y) \propto f_\alpha(Y|\theta)f_\beta(\theta) \propto f_\alpha(y_t|\theta_t)f_\beta(\theta_t|\theta_{t-1})f_\beta(\theta_{t+1}|\theta_t)
$$

$$
\propto \theta_t^{\alpha_t} \exp(-a_t\theta_t) \theta_t^{\lambda_t-1} \exp\left\{-\frac{\lambda \theta_t}{b}\right\} \frac{1}{\Gamma(\lambda \theta_t)} \left(\frac{\lambda \theta_{t+1}}{b}\right)^{\lambda \theta_t}
$$

$$
= \frac{(\lambda \theta_{t+1})^{\lambda \theta_t}}{\Gamma(\lambda \theta_t)} \theta_t^{(\alpha_t+\lambda \theta_{t-1})-1} \exp\left\{-\theta_t \left(a_t + \lambda b^{-1} + \lambda \log b\right)\right\}.
$$

Using the notations in the modified rejection sampling, we thus have

$$
f^*(\theta_t) = \frac{(\lambda \theta_{t+1})^{\lambda \theta_t}}{\Gamma(\lambda \theta_t)} \theta_t^{(\alpha_t+\lambda \theta_{t-1})-1} \exp\left\{-\theta_t \left(a_t + \lambda b^{-1} + \lambda \log b\right)\right\} \quad (9.2.18)
$$

and the function $g^*(\cdot)$ is obtained by applying Stirling’s formula

$$
\Gamma(p) \sim \sqrt{2\pi} e^{-p} p^{p-1/2}, \quad \text{as} \quad p \to \infty.
$$

or the fact that

$$
\Gamma(p) > \sqrt{2\pi} e^{-p} p^{p-1/2}.
$$

It follows that

$$
\Gamma(\lambda \theta_t) > \sqrt{2\pi} (\lambda \theta_t)^{\lambda \theta_t-1/2} \exp(-\lambda \theta_t).
$$

And furthermore,

$$
\frac{(\lambda \theta_{t+1})^{\lambda \theta_t}}{\Gamma(\lambda \theta_t)} < \sqrt{\frac{\lambda}{2\pi}} \theta_t^{1/2} \exp(\lambda \theta_t) \left(\frac{\theta_{t+1}}{\theta_t}\right)^{\lambda \theta_t} \leq \sqrt{\frac{\lambda}{2\pi}} \exp\left(\frac{\lambda \theta_{t+1}}{e}\right) \theta_t^{1/2} \exp(\lambda \theta_t)
$$
where the last inequality is obtained by Lemma 9.1 and the fact that
\[
\left( \frac{\theta_{t+1}}{\theta_t} \right)^{\lambda \theta_t} = \left\{ \left( \frac{\theta_{t+1}}{\theta_t} \right)^{\theta_t/\theta_{t+1}} \right\}^{\lambda \theta_{t+1}} \leq \exp \left( \frac{\lambda \theta_{t+1}}{e} \right). \tag{9.2.19}
\]
Thus, we obtain
\[
g^*(\theta_t) = \theta_t^{(y_t + \lambda \theta_{t-1} + 1/2) - 1} \exp \left\{ -\theta_t \left( a_t + \lambda b^{-1} - \lambda + \lambda \log b \right) \right\}
\]
and \( f^*(\theta_t) < K g^*(\theta_t) \) where
\[
K = \sqrt{\frac{\lambda}{2\pi}} \exp \left( \frac{\lambda \theta_{t+1}}{e} \right)
\]
and
\[
g^*(\theta_t) \propto \text{gamma} \left( y_t + \lambda \theta_{t-1} + \frac{1}{2}, a_t + \lambda b^{-1} - \lambda + \lambda \log b \right).
\]
In summary, we generate the \( \theta_t \) by executing the following routine:

(1) Generate
\[
\theta_t^* \sim \text{gamma} \left( y_t + \lambda \theta_{t-1} + \frac{1}{2}, a_t + \lambda b^{-1} - \lambda + \lambda \log b \right).
\]

(2) Generate \( u \sim \mathcal{U}(0,1) \).

(3) If
\[
\sqrt{\frac{\lambda}{2\pi}} \Gamma(\lambda \theta_t^*) (\theta_t^*)^{1/2} \exp \left( \lambda \theta_t^* + \lambda e^{-1} \theta_{t+1} \right) (\lambda \theta_{t+1})^{-\lambda \theta_t^*} > u^{-1}, \tag{9.2.20}
\]
let \( \theta_t = \theta_t^* \), otherwise return to step (1).

For \( t = n \), a similar derivation leads to the conditional density
\[
f_{\theta}^*(\theta_n | \theta_{n-1}, Y) \propto f(y_n | \theta_n) f(\theta_n | \theta_{n-1}) \propto \text{gamma} \left( y_n + \lambda \theta_{n-1}, a_n + \lambda b^{-1} \right).
\]
In some cases, the bound given in equation (9.2.19) is too crude to execute the rejection sampling effectively. An alternative bound we actually used for our programming is given
as follows. For a small $\epsilon < 1$,
\[
\left( \frac{\theta_{t+1}}{\theta_t} \right)^{\lambda \theta_t} < \left( \epsilon + \frac{\theta_{t+1}}{\theta_t} \right)^{\lambda \theta_t}
\]
\[
= e^{\lambda \theta_t} \left\{ \left( 1 + \frac{\theta_{t+1}}{\epsilon \theta_t} \right)^{\frac{\epsilon \theta_t}{\theta_t + 1}} \right\} \leq e^{\lambda \theta_t} \exp \left( \frac{\lambda \theta_{t+1}}{\epsilon} \right).
\]  
(9.2.21)

Consequently, the rejection sampling could be modified as follows.

(1') Generate
\[
\theta_t^* \sim \text{gamma} \left( y_t + \lambda \theta_{t-1} + \frac{1}{2}, a_t + \lambda b^{-1} - \lambda - \lambda \log(\epsilon) + \lambda \log b \right).
\]

(2') Generate $u \sim U(0,1)$.

(3') If
\[
\sqrt{\frac{\lambda}{2\pi}} \Gamma(\lambda \theta_t^*) (\theta_t^*)^{1/2} \exp \left( \lambda \theta_t^* + \lambda e^{-1} \theta_{t+1} \right) (\lambda \theta_{t+1})^{-\lambda \theta_t^*} > u^{-1},
\]
let $\theta_t = \theta_t^*$, otherwise return to step (1).

The method gives some flexibility for the sampling procedure through adjustment of $\epsilon$.

For some cases, it is necessary to monitor outcomes of rejection sampling to assure that the samples are somehow balanced in order to make a reasonable inference.

9.2.3 Numerical results

The true parameters for Model II were chosen to be $\alpha = (0.10, -0.20, 0.20, -0.05)^T$ with covariates $x = (\cos(2\pi t/12), \sin(2\pi t/12), \cos(2\pi t/6), \sin(2\pi t/6))^T$, $t = 1, \ldots, 168$, and $\beta = 0.005$ (increasing trend). The dispersion parameter $\sigma^2$ is set to 0.1 and $\theta_0$ is set to 1.
Chapter 9. Simulation studies under Poisson-gamma models

Figure 9.1 shows the 168 simulated observations from the Poisson-gamma model with these parameters, where the solid line represents the latent process and the dots are the Poisson counts. There is an apparent increasing trend in the latent part, so the underlying process appears to be non-stationary.

Figure 9.1: 168 simulated counts from the Poisson-gamma model where the latent process is non-stationary. Observations are indicated by dots, and the latent process with solid line.

To perform the MCEM algorithm, we set initial values as follows. The starting values of $\alpha$ were obtained by fitting a log-linear model to the time series of counts with an offset specified by the moving average of the count data. In order to remove the seasonal
effects, the moving average was computed based on both lag 6 and lag 12 (first average

the raw series by lag 6 and then average averaged series by lag 12), and the resulting

series characterized mainly the trend. It was also used as the initial latent process for the

Gibbs sampler. Consequently, we obtained the starting value of $\beta$ based on the initial

latent process, by differentiating the joint log-likelihood with respect to $\beta$. The starting

value for $\sigma^2$ was chosen as 0.001. The same initializations was also used for the KEE

approach. The Markov chain sample size ($m$) was initially chosen to be 2,000.

As mentioned above, the Gibbs sampler starts with the $\theta$ process which was set to

the moving average of the count data. A run of 4,000 $\theta$'s was generated by the Gibbs

sampler, and the first 2,000 $\theta$'s were discarded as transient values. The last 2,000 $\theta$'s

were retained to compute $Q^{(i+1)}$. The final value of $\theta$ from the previous run was then

used as the starting value of $\theta$ for the Gibbs sampler in the next M-step. This procedure

was tried to run repeatedly 400 times where the estimate of the dispersion parameter $\sigma^2$

was updated at each iteration by equation (9.2.12).

Unfortunately, the above tentative scheme for computing the MCEM algorithm did

not succeed in the estimation of the dispersion parameter $\sigma^2$, because Stirling's approx­

imation to the gamma function was applied to obtain the estimate. In fact, whenever

the estimate of $\sigma^2$ became negative, the MCEM algorithm stopped. This relates to the

rejection sampling that can not always produce the “robust” samples against the approx­

imation for the estimate of $\sigma^2$. Actually, we have tried a few alternative ways to estimate

$\sigma^2$, but none was successful. So, we eventually came to use the plotting profile likelihood

approach to overcoming the difficulty. This method requires much more computing time

and space because now each iteration needs much more runs to obtain the maximizer

of profile likelihood. For a fixed value of $\sigma^2$, the profile likelihood with respect to the

parameter $\eta$ was given by equation (9.2.6) where the parameter $\lambda$ was substituted by

the given value $1/\sigma^2$. Therefore the whole implementation for the MCEM algorithm was
then based on likelihood with known \( \sigma^2 \), rather than the joint likelihood with unknown \( \sigma^2 \). Accordingly, the M-step, consisting of solving equations (9.2.8) and (9.2.9), was run until the convergence criterion \( \max || \eta^{(i+1)} - \eta^{(i)} || < 10^{-4} \) attains. So, the resulting maximizers of profile likelihood, denoted by \( \hat{\alpha}(\sigma^2) \) and \( \hat{\beta}(\sigma^2) \) are the so-called maximum profile-likelihood estimates of \( \alpha \) and \( \beta \) for given \( \sigma^2 \).

Figure 9.2: Log-profile-likelihood versus the dispersion parameter given increasingly.
To implement the plotting profile likelihood method numerically, referring to the pattern of the adjusted Pearson estimate of $\sigma^2$ produced by the KEE approach, we chose the range of the dispersion parameter $\sigma^2$ as the interval [0.001, 2.15]. We implemented the procedure based on both cases where the dispersion parameter was updated, respectively, increasingly and decreasingly with a difference of 0.001 between two consecutive iterations.

Figure 9.3: Log-profile-likelihood versus the dispersion parameter given decreasingly.
Figure 9.2 shows the plot of the estimated log-profile-likelihood of the observed data versus the values of the dispersion parameter $\sigma^2$ which was fed increasingly starting at 0.001. Conversely, given the values of the dispersion parameter $\sigma^2$ decreasingly starting at 0.3, the estimated log-profile-likelihood of the observed data versus $\sigma^2$ is plotted in Figure 9.3.

The top display of Figure 9.2 indicates that the log-profile-likelihood first goes up rapidly and then goes down slowly when $\sigma^2$ exceeds 0.03 or so. The bottom display shows a detailed pattern of the log-profile-likelihood for the $\sigma^2$ given on the interval $[0.01, 0.1]$. The two displays in Figure 9.3 show similar pattern to Figure 9.2, where the log-profile-likelihood goes up slowly and then goes down quickly. Both figures indicate that the log-profile-likelihood functions attain maximum around $\sigma^2 = 0.025$.

Then we took the parameter values at $\sigma^2 = 0.028$, which gives the highest profile likelihood in Figure 9.2, as the starting but fixed value of dispersion parameter for a new MCEM algorithm. This time the starting latent process and the starting values of regression parameters are set to the values of the profile likelihood maximizers corresponding to $\sigma^2 = 0.028$. Now the size of the Markov chain sample is increased to 4,000. Figure 9.4 shows the patterns for the log-likelihood and the coefficients over 100 iterations. All displays are stable, and the values of the parameters change only slightly; the first two significant digits are virtually unchanged.

### 9.2.4 Conclusion

Table 9.1 reports the approximate maximum likelihood estimates from the MCEM approach, those based on Figure 9.2 corresponding to $\sigma^2 = 0.028$, and from the KEE approach.

Except that the estimate of $\beta$ (trend covariate) is a little overestimated, the rest of estimates are relatively close to the corresponding true values. The standard errors of the
Figure 9.4: Patterns for log-likelihood and coefficients estimates.
Table 9.1: Estimates and standard errors from KEE and MCEM.

<table>
<thead>
<tr>
<th></th>
<th>MCEM Estimates</th>
<th>MCEM Std err.</th>
<th>KEE Estimates</th>
<th>KEE Std err.</th>
<th>True values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trend</td>
<td>0.0125</td>
<td>0.0300</td>
<td>0.0120</td>
<td>0.0149</td>
<td>0.0050</td>
</tr>
<tr>
<td>cos(2\pi t/12)</td>
<td>0.1348</td>
<td>0.0494</td>
<td>0.1014</td>
<td>0.0774</td>
<td>0.1000</td>
</tr>
<tr>
<td>sin(2\pi t/12)</td>
<td>-0.3300</td>
<td>0.0546</td>
<td>-0.2281</td>
<td>0.0835</td>
<td>-0.2000</td>
</tr>
<tr>
<td>cos(2\pi t/6)</td>
<td>0.2264</td>
<td>0.0513</td>
<td>0.1592</td>
<td>0.0698</td>
<td>0.2000</td>
</tr>
<tr>
<td>sin(2\pi t/6)</td>
<td>-0.0731</td>
<td>0.0522</td>
<td>-0.0571</td>
<td>0.0702</td>
<td>-0.0500</td>
</tr>
<tr>
<td>(\sigma^2)</td>
<td>0.0280</td>
<td>0.1310</td>
<td>0.1000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

KEE method are uniformly slightly higher than those of MCEM method. However, it is unfair to turn down the KEE in favour of MCEM only through the simple comparison between standard errors which are approximately computed. For example, one may compare the predicted latent processes obtained from two methods. Figure 9.5 presents the true latent process, the Kalman smoother \(m^*_t\) related to the KEE approach and the smoothed means \(\bar{\theta}_t\) from the MCEM approach.

Figure 9.5: Comparison between the true latent process (solid line) and the estimated latent processes from the Gibbs sampler (dashed line) and from Kalman smoother (broken line).
It is easy to compute the mean squared error for the two estimated processes: the \( \bar{\theta}_t \) produces \( n^{-1} \sum_1^n (\theta_i - \bar{\theta}_i)^2 = 7.9140 \) and the Kalman smoother \( m^*_t \) gives \( n^{-1} \sum_1^n (\theta_i - m^*_i)^2 = 0.7410 \). The BLUP gives an overall better prediction for the latent process than the smoothed means of the Gibbs sampler. This might be the reason why the KEE produces a stable sequence for estimation of the dispersion parameters converging towards the true value.

Another advantage of KEE method is in the aspect of computational efficiency in terms of the simplicity of programming and far less time-consuming computation. For the analysis of the simulated data, we coded the KEE method in the S-PLUS language, whereas we were forced to program the MCEM in the C language due to the huge dynamic space demand for the Gibbs sampler and plotting profile likelihood approach, which are impossible to run based on S-PLUS code. Even though C code is much more efficient than S-PLUS in terms of computational speed, our program for the plotting profile likelihood estimation via MCEM took about 30 hours to complete, whereas the program for KEE took only about 16 minutes to converge. As is well-known, the rejection sampling involved in the Gibbs sampler is the part which takes considerable time. In addition, the plotting profile likelihood estimation is very time-consuming because the MCEM has to be carried out for a sequence of \( \sigma^2 \) values. As a result, we would like to make a recommendation concerning the MCEM approach: if there is a failure to get simple starting values for the MCEM algorithm such as through GLIM or if a faster computing procedure is needed, the KEE definitely provides high quality starting values because those values given by the KEE are already near the true values. For example, the Kalman smoother provides very competitive starting values for the latent process as shown in Figure 9.5. We have tentatively tried to run the MCEM algorithm by using the KEE output as starting values, and the log-likelihood fluctuated without any substantial changes occurring. A future careful investigation of this issue would be interesting.
9.3 Detecting non-stationarity

This section is mainly devoted to examining the utility of the residual analyses developed respectively in Chapters 7 and 8. It seems that the assumption of stationarity or non-stationarity for the latent process substantially affects the interpretation of the modelling and hence the data analysis. So, developing tools for checking whether the latent process is stationary or not is important in order to assure the appropriateness of the data analysis. The reliability of the model checking diagnosis based on analysis of residuals will be the main concern in the present section.

Our studies are organized through the case-and-control strategy as indicated in the following table.

<table>
<thead>
<tr>
<th>Model that fits data</th>
<th>Model that generates data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I (stationarity)</td>
<td>Model I (stationarity)</td>
</tr>
<tr>
<td>Model II (non-stationarity)</td>
<td>Model II (non-stationarity)</td>
</tr>
<tr>
<td>see Chapter 7</td>
<td>see Section 9.3.1</td>
</tr>
<tr>
<td>see Section 9.3.2</td>
<td>see Chapter 8</td>
</tr>
</tbody>
</table>

In Section 9.3.2 the data that are generated from Model I are fitted by Model II, and in Section 9.3.1 the data that are generated from Model II are fitted by Model I. Then we study the residuals from both fittings.

9.3.1 When non-stationarity is present

Actually, the data set given in the previous section (also shown in Figure 9.1) is used for our study here. Now we fit the data by Model I where the latent process is stationary AR(1) with gamma margins. The starting values for the dispersion parameter $\sigma^2$ and autocorrelation parameter $\rho$ are set to 0.05 and 0.1, respectively. As usual, the starting
latent process is set to the moving average of the data. And 500 iterations were run; however, according to our experience, 150 iterations are normally enough to get convergence. Unfortunately, for the current case the estimate of $\alpha$ does not seem convergent where the maximum difference of two consecutive estimates of $\alpha$ remains about as high as 1,500, and the last 100 iterations only improve the value of the maximum difference from 1,507 to 1,498.

Figure 9.6: ACF and PACF functions for the residuals of the latent process.

It is clear that all estimates reported in Table 9.3 are fairly poor compared with the true values. Nevertheless, let us try the residual plot to see if there is any indication of violation against the assumption of stationarity. If the stationary AR(1) structure is correct, the sequence of residuals $\tilde{e}_t = m_t - G_t m_{t-1}$ should be a noise series where $G_t$ are independent beta random variates from $\text{Beta}(\rho \lambda, (1 - \rho) \lambda)$. Figure 9.6 shows the
Chapter 9. Simulation studies under Poisson-gamma models

Table 9.3: Estimates of coefficients given by Model I based on the data generated by Model II.

<table>
<thead>
<tr>
<th></th>
<th>Estimates</th>
<th>Std. err.</th>
<th>True values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trend</td>
<td>0.0237</td>
<td>0.0002</td>
<td>0.0050</td>
</tr>
<tr>
<td>cos(2πt/12)</td>
<td>0.2718</td>
<td>0.0179</td>
<td>0.1000</td>
</tr>
<tr>
<td>sin(2πt/12)</td>
<td>-0.8187</td>
<td>0.0186</td>
<td>-0.2000</td>
</tr>
<tr>
<td>cos(2πt/6)</td>
<td>0.0910</td>
<td>0.0153</td>
<td>0.2000</td>
</tr>
<tr>
<td>sin(2πt/6)</td>
<td>-0.3079</td>
<td>0.0154</td>
<td>-0.0500</td>
</tr>
<tr>
<td>σ²</td>
<td>0.6512</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ρ</td>
<td>0.7175</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

autocorrelation and partial autocorrelation functions for $\hat{e}_t$. Both displays clearly suggest that the latent process is neither an autoregressive process nor a moving-average process.

9.3.2 When stationarity is present

First we generated a data set from Model I with true values of the parameters given as follows. The covariates were the same as those we used to generate the data in Section 9.3.1, and so are the values of the parameters associated with the corresponding covariates. The correlation coefficient $\rho$ is set to 0.5 and the dispersion parameter $\sigma^2 = 0.5$ or $\lambda = 2$. The 168 simulated observations and latent process are shown in Figure 9.7.

These counts were fitted by Model II, starting with $\sigma^2 = 0.05$ and the moving-average of the data used as the initial latent process. After 200 iterations, the program stopped according to the criterion of convergence – the maximum difference between two consecutive estimates was less than $10^{-5}$. Table 9.4 reports the values of estimates.

The estimate for covariate $\cos(2\pi t/12)$ had the opposite sign compared with the true value 0.1. The rest of estimates coincide in sign with the true values, and the covariates of trend, $\cos(2\pi t/12)$ and $\sin(2\pi t/6)$ are not significant at the significance level of 95%. If the non-stationary latent process were correct, the residuals $\xi_t = (m_t - bm_{t-1})/\sqrt{bD_t - C_t}$
Figure 9.7: 168 simulated counts (represented by dots) from a state space model with a stationary gamma latent process (represented by a solid line).

Table 9.4: Estimates of coefficients given by Model II based on the data generated by Model I.

<table>
<thead>
<tr>
<th></th>
<th>Estimates</th>
<th>Std. err.</th>
<th>True values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trend</td>
<td>0.0013</td>
<td>0.0321</td>
<td>0.0050</td>
</tr>
<tr>
<td>$\cos(2\pi t/12)$</td>
<td>-0.0612</td>
<td>0.0730</td>
<td>0.1000</td>
</tr>
<tr>
<td>$\sin(2\pi t/12)$</td>
<td>-0.2700</td>
<td>0.0813</td>
<td>-0.2000</td>
</tr>
<tr>
<td>$\cos(2\pi t/6)$</td>
<td>0.3101</td>
<td>0.0671</td>
<td>0.2000</td>
</tr>
<tr>
<td>$\sin(2\pi t/6)$</td>
<td>-0.1053</td>
<td>0.0667</td>
<td>-0.0500</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1.9748</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
should be uncorrelated. Figure 9.8 shows the autocorrelation function and partial autocorrelation function of these residuals. It is clear that the $\xi_t$ are correlated, which suggests that there is something wrong with the assumptions for the latent process.

Figure 9.8: ACF and PACF functions for the residuals from the latent process.

9.3.3 Conclusion

As discussed above, the residual analyses developed in the previous chapters are very effective for detecting the misspecification of the assumptions concerning the structure of the latent processes. Such tools are valuable in practice for checking model assumptions for data analysis. This use of residual analysis for model checking provides an objective
statistical inference. Evidently, the ability to detect the misspecification of the structure of the latent process gives some credibility to the KEE approach and increases the confidence to its users in applied areas. It is worth noting that the model checking diagnosis via residual analyses for both MCEM and GEE methods are not currently available. That is because that even though the residuals may be computed in both approaches, their properties that might be used for model checking are not easy to be drawn in general. This hampers the development of model checking diagnosis on the basis of both methods.
Chapter 10

Data analysis

In this chapter we present an application of state space models to a regression analysis of data from Prince George, British Columbia, previously analyzed by Knight, et al. (1993). The data consist of daily recordings of the number of emergency room visits for each of four categories of respiratory diseases, along with measurements of meteorological variables and air pollution. We used a state-space model assuming conditionally independent Poisson counts for the four categories given a latent morbidity process, the latent process being a gamma Markov process. The main objective of the investigation was to examine the relationship between air pollution and respiratory morbidity, taking into account seasonality and meteorological conditions. We found that Total Reduced Sulphur significantly influences the expected number of emergency room visits for the four disease categories, in agreement with the conclusion by Knight et al. (1993). However, our final model is simpler than theirs; in particular we found no evidence of seasonal variation beyond that explained by the meteorological variables.

10.1 Introduction

We consider data from Prince George, British Columbia, consisting of daily counts of emergency room visits for respiratory diseases, classified into four categories (Asthma, Bronchitis, Ear infections and Others) for the period 1984 to 1986, along with daily measurements of air pollution and meteorological variables. The air pollution variables were
Sulphur (total reduced sulphur compounds) and Particulates (total suspended particulates). The main objective of the investigation was to examine the relationship between air pollution and respiratory morbidity.

The monitoring of air pollution in Prince George (c.f. Lambert et al., 1987) shows that there are frequent excursions above the provincial air quality standards, and there has long been public concern in Prince George that the air quality in the city may be adversely affecting the health of the residents.

Recent studies of the association between air pollution and human health that have used emergency room visits or hospital admissions as an indicator of human health include Levy et al. (1977), Pope (1989), Samet et al. (1981) and Bates and Sizto (1983). Other indicators such as mortality (Dockery et al., 1993) and elementary school absences (Ransom and Pope, 1992) have also been used. We are particularly concerned with the modelling of the dynamics of the disease-generating process (at an aggregate level), and our model is well suited to deal with count indicators such as emergency room visits and school absences that can be presumed to be closely related to daily levels of ambient air pollution.

The Prince George data were previously analyzed by Knight et al. (1993) who used a log-linear Poisson model, in which the counts were assumed independent over time, as well as between the four categories. They found evidence of over-dispersion, and hence used a quasi-likelihood analysis.

A state-space model based on Poisson counts, proposed in Chapter 8, is used for the analysis, assuming that the counts are conditionally independent given a latent Markov process. This implies serial correlation for the counts as well as correlations between the four disease categories, giving a more realistic modelling of the correlation structure for this type of data. The ability of our model to account for correlations among the four categories turns out to be crucial in the data analysis.
A new feature of our model is that the air pollution covariates are treated as long-term covariates and enter the model via the latent Markov process. We may hence interpret the latent process as a latent morbidity process representing the air pollution's potential for causing respiratory diseases. In particular, the Markov assumption implies a carry-over effect from day to day for the effect of air pollution.

The meteorological and seasonal variables are assumed to have an immediate effect on the counts, and hence enter as covariates in the conditional Poisson distribution for the counts. So, such covariates are short-term covariates, because they have an immediate effect on the counts.

This approach gives a realistic modelling of the longitudinal structure of the data not easily achieved with the quasi-likelihood method developed by Liang and Zeger (c.f. Diggle, Liang and Zeger, 1994). In particular, the new model has both a marginal and a conditional interpretation, and a detailed residual analysis allows all model assumptions to be checked from the data.

10.2 Data

The available data consist of daily counts of emergency room visits for respiratory diseases and the corresponding daily values of various meteorological variables and air pollution parameters from April 1, 1984 to March 31, 1986, a total of 730 days. For a more detailed description of the data, we refer to Knight et al. (1993).

The counts refer to emergency room visits by residents of Prince George to the single hospital in Prince George. All visits were classified according to discharge diagnosis. The counts were originally classified into 33 diagnostic categories. For this analysis these categories were collapsed into four categories, referred to as Asthma, Bronchitis, Ear infections and Other respiratory diseases. Figure 10.1 shows time-series plots of the daily
counts for each of the four categories.

The meteorological data were all obtained at the Prince George airport, and consist of daily summaries of temperature and humidity. The following three measurements were recorded: temperature, maximum relative humidity and minimum relative humidity. The temperature is the average of 24 hourly readings in degrees Celsius, and the maximum/minimum relative humidity are the largest/smallest of 24 hourly readings. There were no missing values for temperature, but a missing value of minimum humidity for September 29, 1985 was substituted by 38, the average of the remaining records of minimum humidity.

Air quality parameters have been measured at six monitoring stations in Prince George since 1980, and the daily averages of the readings were used. The two summaries used in this study were Sulphur (total reduced sulphur compounds) and Particulates (total suspended particulates).

Sulphur (recorded in \( \text{ppb} \)) was measured by a continuous monitor, recording daily averages. The original data had many individual missing values, but these were "predicted" using available measurements from the other stations (for details, see Knight et al., 1993).

Particulates (recorded in \( \mu g/m^3 \)) was measured by a 24-hour vacuum sampler once every six days, on the same day for all six monitoring stations. For the purpose of the analysis, we have filled in the missing five-day blocks by linear interpolation. Figure 10.2 shows time-series plots of the meteorological and air pollution variables.

### 10.3 Methodology

Referring to Chapter 8, we summarize the main features of the model and estimation technique as follows.
Figure 10.1: Emergency room visits for four respiratory diseases.

**Asthma Counts**

**Bronchitis Counts**

**Ear Infection Counts**

**Others**
Chapter 10. Data analysis

Figure 10.2: Meteorological and air pollution variables.

Temperature

Humidity (Max, Min)

Sulphur (TRS)

Particulates (TSP)
10.3.1 The model

The response variable $Y_t$ is a four-dimensional vector, consisting of daily counts of emergency room visits for the four respiratory disease categories. The main idea in the model is to consider the number of daily emergency room visits to be a Poisson process driven by a latent Markov process, denoted $\{\theta_t\}$. The components of $Y_t$ are assumed to be conditionally independent given $\theta_t$, and the conditional distribution for the $i$th component of $Y_t$ follows a Poisson distribution,

$$Y_{it}|\theta_t \sim Po(a_{it}\theta_t), \ i = 1, 2, 3, 4,$$

where $\log a_{it}$ is a linear predictor depending on the short-term covariates $x_t$, i.e.,

$$a_{it} = \exp(x_t^T\alpha_i),$$

$\alpha_i$ denoting the parameter vector for the $i$th category.

The latent morbidity process $\theta_t$ represents the "overall potential" for air pollution to cause emergency room visits with respiratory symptoms. It is assumed to follow a gamma Markov process, defined by

$$\theta_t|\theta_{t-1} \sim Ga\left(b_t\theta_{t-1}, \frac{\sigma^2}{\theta_{t-1}}\right),$$

where $Ga(\mu, \delta^2)$ denotes the gamma distribution with mean $\mu$ and coefficient of variation $\delta$. The parameter $\sigma^2$ is a dispersion parameter expressing the smoothness of the process, and $\log b_t$ is a linear predictor depending on the long-term covariates via their increments. Hence,

$$b_t = \exp(\Delta z_t^T\beta), \ \Delta z_t = z_t - z_{t-1},$$

where $z_t$ is the vector process of air pollution measurements and $\beta$ is the parameter vector. The latent process is standardized by taking $\theta_0 = 1$, and the long-terms covariates are centralized by subtracting $z_1$ from all values and taking $z_0 = 0$. 
Chapter 10. Data analysis

The marginal moments of $Y_t$ are

$$E(Y_t) = a_t \tau_t, \quad \text{Var}(Y_t) = \Lambda_t \tau_t + a_t a_t^T \phi_t \sigma^2 \tau_t$$

where $\Lambda_t = \text{diag}(a_t)$, $\tau_t = E(\theta_t) = \theta_0 b_1 \cdots b_t$ and $\phi_t = b_t + b_t b_{t-1} + \cdots + b_t b_{t-1} \cdots b_1$. The log-expectation of $Y_{it}$ is

$$\log E(Y_{it}) = x_t^T \alpha_i + z_t^T \beta,$$

showing that the model is log-linear in the covariates $x_t$ and $z_t$.

The two sets of covariates $x_t$ and $z_t$ are called respectively short-term and long-term covariates, a crucial distinction, although the terminology short-term/long-term should not be taken too literally. We now explain the differences between the two types of covariates.

The long-term covariates $z_t$ represent measurements of the two air pollutants, expected to have an effect on respiratory diseases. Because the long-term covariates enter the model via the latent process, this process represents the potential for air pollution to cause emergency room visits with respiratory symptoms. The terminology “long-term” is used because the Markov structure of the latent process produces a carry-over effect from day to day, such that an increase in the air pollution level on a given day may have an effect on subsequent days. An important characteristic of long-term covariates is that they have the same effect on all four categories of $Y_t$.

The short-term covariates $x_t$ enter via the conditional distribution of the observed emergency room visit counts given the latent process $\theta_t$, and hence represent modulating factors that have an immediate (acute) effect on visit counts relative to the value of $\theta_t$. Such covariates may include local weather conditions and seasonal effects, and also include separate constants for the four categories, corresponding to the different incidence rates for the symptoms.
10.3.2 Parameter estimation

Following Chapter 8, we use the so-called Kalman estimating equation for estimation. For a sample of \( n \) observations, equations (8.4.1) and (8.4.2) become

\[
\psi_1(\eta, \theta) = \sum_{i=1}^{n} \left( \sum_{t=1}^{4} Y_{it} - \sum_{t=1}^{4} a_{it} \theta_t \right) x_t = 0
\]

and

\[
\psi_2(\eta, \theta) = \sum_{i=1}^{n} b_t^{-1} (\theta_t - b_t \theta_{t-1}) \Delta z_t = 0,
\]

where \( \eta = (\alpha^T, \beta^T)^T \) and \( \theta = (\theta_0, \theta_1, \ldots, \theta_n) \) denotes the latent process. Through the Newton scoring algorithm given by (8.4.4), the M-step of the EM algorithm devotes to solve the Kalman estimating equations

\[
\psi_1(\eta, m^*(\eta)) = 0, \quad \psi_2(\eta, m^*(\eta)) = 0,
\]

where \( m^* = m^*(\eta) \) is the Kalman smoother (c.f. Section 8.3). And the standard errors of the estimates may be computed via the inverse of Godambe information matrix.

10.3.3 Residual analysis

The Kalman filter and smoother allow us to define residuals for both the observed and latent processes, to be used for model checking. The most useful type of residuals for our purpose are the conditional residuals, which may be characterized as observed values minus predicted values.

There are two types of conditional residuals for the \( ith \) observed process, defined in terms of the Kalman filter and smoother by

\[
R_{it} = Y_{it} - f_{it}, \quad \text{and} \quad R^*_{it} = Y_{it} - a_{it} m^*_t,
\]

respectively. For the latent process, we define the conditional residuals as
\[ r_t = m_t - b_t m_{t-1}, \]

based on the Kalman filter. The residuals are then standardized to have zero mean and unit variance. All references to residuals from now on are to these standardized residuals with parameter estimates inserted.

The standardized residuals may be used for checking the correlation structure, the distributional assumptions and the covariate structure of the model, as done in Section 10.4 for the Prince George data. The residuals \( r_t \) and \( R_t \) are particularly useful for checking the correlation structure of the data, because they are uncorrelated over time.

10.4 Data analysis

10.4.1 Model identification

We now describe the main considerations that went into finding the appropriate model for the Prince George data.

Following Knight et al. (1993), we transformed Sulphur and Particulates to respectively \( \log(0.5 + \text{Sulphur}) \) and \( \log(\text{Particulates}) \). For convenience, we shall refer to \( \log(0.5 + \text{Sulphur}) \) as simply \( \log\)-Sulphur. We also made a logarithmic transformation of the minimum and maximum humidity (note that Knight et al. (1993) did not transform humidity). For ease of interpretation, the difference and the sum of high and low \( \log\)-humidities were used as short-term covariates.

The three meteorological covariates were centralized by subtracting their respective means. In this way, the intercept parameter represents the log expected number of emergency room visits for a day with average temperature and humidity, and with air pollution as on April 1, 1984. The intercept is allowed to depend on the day of the week.
We now investigate the need for lagging the log-Sulphur covariate. Following Knight et al. (1993), we consider up to lag 2 for log-Sulphur. It does not seem reasonable to include lags for log-Particulates, because of the linear interpolation used for this covariate (c.f. Section 10.2).

Our preliminary model for the Prince George data is hence as follows.

Short-term covariates:

- log-temperature,
- difference of log-humidities,
- sum of log-humidities,
- 7 day-of-week factors.

Long-term covariates:

- lag 0, 1 and 2 of log-Sulphur,
- log-Particulates.

Table 10.1: Estimates and standard errors for air pollution effects.

<table>
<thead>
<tr>
<th></th>
<th>log-Sulphur</th>
<th>log-Particulates</th>
</tr>
</thead>
<tbody>
<tr>
<td>lag 0</td>
<td>0.029 (0.016)</td>
<td>-0.120(0.058)</td>
</tr>
<tr>
<td>lag 1</td>
<td>-0.023 (0.017)</td>
<td>—</td>
</tr>
<tr>
<td>lag 2</td>
<td>0.048 (0.016)</td>
<td>—</td>
</tr>
</tbody>
</table>

The estimates of the coefficients for the air pollution covariates are reported in Table 10.1. To verify the need for lag 1 and lag 2 for log-Sulphur, we applied the Wald test described in Section 10.3. The result was $W = 10.251$, which, compared with a $\chi^2(2)$-distribution, gave a $p$-value of 0.006. This confirms the need for including the two lags in the model. Only the lag 2 coefficient is significant individually, but we have no
explanation for this finding. Knight et al. (1993) found similar results, except that they found both the lag 0 and lag 2 coefficients to be significant, but not the lag 1 coefficient.

Figure 10.3: First-year residuals against second-year residuals for the four categories.

To investigate the possibility of a seasonality effect, we plotted the second year's residuals against the first year's, for both \( r_t \) and \( R_{tt} \) shown in Figure 10.3 and Figure 10.4, respectively.

In the presence of seasonality not accounted for by the model, we would expect to see a correlation between the first and second year's residuals. None of these five plots
Figure 10.4: First-year residuals against second-year residuals for latent process.

indicated any such correlation, but it should be kept in mind that a two-year period may be too short to detect seasonality. However, it seems that the meteorological variables have accounted for all seasonality variation in the present data.

10.4.2 Model checking

In order to verify the main model assumptions, we now analyse the residuals from the above model.

Figure 10.5 shows plots of $R_{it}$ against $\log f_{it}$ for each $i$, where the $f_{it}$'s are the predicted values of the observed processes based on the Kalman filter. The discreteness of the data combined with the low expected counts, especially for the Asthma process, account for the curved set of bands with negative slope found in these plots. The lowest band in each plot corresponds to the observation $y = 0$, the next lowest to $y = 1$ and so on. In particular, there can be no large negative residuals for small fitted values. This feature, together with the negative slope of the curves on which the residuals lie, make
Figure 10.5: Residuals against log-linear predictions for the four categories.
Chapter 10. Data analysis

The residual plots show an apparent downward trend. These traits dominate the plots for log expected counts less than 1 (and are hence prominent in Figure 10.5), but disappear above this value, where the Poisson residuals become approximately normal. Taking this into account, the plots are compatible with the Poisson assumption. In particular the plots confirm the appropriateness of the Poisson variance function. There appears to be an outlier in the category Others, but otherwise the number of larger residuals seems to be compatible with the number of observations.

The downward trend in Figure 10.5 will be apparent in any plot of residuals against fitted values, making it difficult to interpret some plots. For example, Figure 10.6 plots the residuals $r_t$ against the log fitted values $\log(b_t * m_{t-1})$, which is useful for checking the gamma assumption for the latent process, showed such a slight downward trend. Otherwise this plot showed no distinctive features, thereby confirming the appropriateness of the gamma assumption.

Figure 10.6: Residuals against log-linear predictions for latent process.
Chapter 10. Data analysis

To check the correlation structure for the latent process, we plotted the autocorrelation function (ACF) for $r_t$ (Figure 10.7). The plot shows that the autocorrelations for lags 1, 16 and 27 fall slightly outside the asymptotic 95% confidence bands. However, the confidence bands are approximate, even for Gaussian time-series, and it seems that the significance would be slight in any case, confirming the Markov assumption for the latent process.

Figure 10.7: Autocorrelation function of residuals from latent process.

![Autocorrelation function](image)

Similar ACF plots for $R_{it}$ for each of the four categories shown in Figure 10.8, did not show anything unexpected. This confirms that the $R_{it}$ are uncorrelated over time for each of the four categories and moreover confirms the adequacy of assuming conditional independence of the Poisson counts over time, given the latent process.

Further plots of $R_{it}$ against $R_{it-1}$ for each $i$, and of $r_t$ against $r_{t-1}$ in respectively Figure 10.9 and Figure 10.10 confirmed these conclusions.

An additional check of the correlation structure for the model was obtained by plotting $R_{it}^*$ against $m_{t-1}^*$ for each category, see Figure 10.11, of the latent process, and not on
Figure 10.8: Autocorrelation functions of residuals for the four categories.

Asthma

Bronchitis

Ear infections

Others
Figure 10.9: Residuals against lag 1 residuals for the four categories.
yesterday's value. These plots exhibited a slight downward trend, especially for asthma and bronchitis, but since this effect is to be expected for plots of residuals against fitted values, as noted above, we interpreted the plot to be in agreement with the model.

In order to check the assumption that the four categories share the same latent process, we plotted the residuals $R_{it}$ against the long-term covariates, in order to confirm that the dependence of the counts on the long-term covariates is \textit{via} the latent process. These eight plots (not shown) did not exhibit any unexpected patterns. Further plots of the residuals $r_t$ against the long-term covariates, and of $R_{it}$ against the short-term covariates, also confirmed the adequacy of the form of the regression model.

Finally, to check the correlations between the four categories, we standardized the vector consisting of $R_{it}$, $i = 1, 2, 3, 4$ by the square-root of its variance-covariance matrix, and calculated the corresponding empirical variance-covariance matrix, which is expected to be the identity matrix (for details, see Chapter 8). The result is given in Table 10.2.

The interpretation of the results in Table 10.2 is complicated by the fact that they
Figure 10.11: Residuals for the four categories against the lag 1 smoother.
depend on the choice of the square root-matrix used for the standardization. However, the four $R_t$s are not far from being uncorrelated, so the matrix in Table 10.2 is not far from being the empirical correlation matrix for the four $R_t$'s. If we take the asymptotic standard errors of the elements of the matrix to be $1/\sqrt{730} = 0.037$, some of the elements of the fourth row are somewhat larger than predicted by the model, suggesting a problem with the category Others. It is not surprising that this category should show correlation with the categories bronchitis and ear infections, for example due to practical problems with the classification of certain cases. Otherwise, this does not seem to indicate any serious shortcoming of our model.

10.5 Conclusions

10.5.1 Final model

We now comment on the interpretation of the model, whose parameter estimates are summarized in Tables 10.1 and 10.3, with asymptotic standard errors in brackets.

The Wald test for the joint significance lag 0, 1 and 2 of log-Sulphur had the value $W = 11.706$, which gave a $p$-value of 0.0085 compared with a $\chi^2(3)$-distribution. Hence, the effect of log-Sulphur on emergency room visits is highly significant.

The coefficient for log-Particulates is negative, but only slightly significant. Moreover, due to the use of linear interpolation (described in Section 10.2), this result should be
Table 10.3: Estimates and standard errors for short-term effects.

<table>
<thead>
<tr>
<th></th>
<th>Asthma</th>
<th>Bronchitis</th>
<th>Ear</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>0.0106 (0.0060)</td>
<td>-0.0147 (0.0045)</td>
<td>0.0002 (0.0039)</td>
<td>-0.0018 (0.0035)</td>
</tr>
<tr>
<td>Diff. Humidity</td>
<td>0.0682 (0.2336)</td>
<td>0.2964 (0.1763)</td>
<td>0.0903 (0.1168)</td>
<td>-0.0569 (0.0906)</td>
</tr>
<tr>
<td>Sum Humidity</td>
<td>-0.0388 (0.1590)</td>
<td>0.1780 (0.1207)</td>
<td>-0.0099 (0.0816)</td>
<td>-0.1638 (0.0646)</td>
</tr>
<tr>
<td>Sunday</td>
<td>0.0796 (0.1760)</td>
<td>0.9099 (0.1620)</td>
<td>1.8208 (0.1552)</td>
<td>2.4185 (0.1531)</td>
</tr>
<tr>
<td>Monday</td>
<td>-0.1050 (0.1807)</td>
<td>0.3878 (0.1695)</td>
<td>1.1237 (0.1601)</td>
<td>1.8075 (0.1554)</td>
</tr>
<tr>
<td>Tuesday</td>
<td>-0.3667 (0.1893)</td>
<td>0.1634 (0.1745)</td>
<td>1.0236 (0.1614)</td>
<td>1.5843 (0.1571)</td>
</tr>
<tr>
<td>Wednesday</td>
<td>-0.6121 (0.1986)</td>
<td>0.4878 (0.1685)</td>
<td>1.1242 (0.1606)</td>
<td>1.7100 (0.1564)</td>
</tr>
<tr>
<td>Thursday</td>
<td>-0.5163 (0.1951)</td>
<td>0.3026 (0.1722)</td>
<td>1.0648 (0.1613)</td>
<td>1.6419 (0.1570)</td>
</tr>
<tr>
<td>Friday</td>
<td>-0.1725 (0.1834)</td>
<td>0.1881 (0.1741)</td>
<td>1.1538 (0.1604)</td>
<td>1.8060 (0.1561)</td>
</tr>
<tr>
<td>Saturday</td>
<td>0.1798 (0.1747)</td>
<td>1.0689 (0.1612)</td>
<td>1.9462 (0.1554)</td>
<td>2.5418 (0.1535)</td>
</tr>
</tbody>
</table>

\[ \hat{\sigma}^2 = 0.019 \]

interpreted with care. We have fitted the above model without log-Particulates, in order to see if correlation between log-Particulates and log-Sulphur could have influenced the coefficient of log-Sulphur. The resulting estimates for the effect of log-Sulphur (not shown) were close to those of Table 10.1. The actual correlation between log-Particulates and log-Sulphur was 0.36, a modest value, confirming that the effect of log-Particulates on the coefficient for log-Sulphur is likely to be of minor importance.

Figure 10.12 shows the Kalman smoother estimate of the latent process, which estimates the potential morbidity due to air pollution without regard to meteorological conditions. The estimated process is highly variable, and the estimate of the dispersion parameter (\( \hat{\sigma} = 0.13 \), from Table 10.3) shows that the stochastic variation of the latent morbidity process \( \theta_t \) itself is high.

There is evidence of major episodes in December 1985 and late March 86 of elevated risk of respiratory morbidity due to air pollution, and minor episodes in November–December 1984 and April 85. These episodes are to some extent evident in the counts for three of the categories (Bronchitis, Ear and Others), see Figure 10.1.
Figure 10.12: The estimated latent process from Kalman smoother.

Figure 10.13: Day-of-the-week effects.
Figure 10.13 shows a plot of the day-of-the-week effects from Table 10.3 for the four categories, which are very similar to the plot of daily averages shown by Knight et al. (1993). This plot shows a fairly clear weekly cycle, with more emergency room visits during the weekends than on workdays.

10.5.2 Comparisons with the previous analysis

While our main findings agree with those of Knight et al. (1993), some of the significant effects that they found were not confirmed by our analysis, which produced a more parsimonious model. We shall argue that there are objective reasons for these discrepancies.

It is crucial in our analysis that we did not find any seasonal variation other than what can be explained by the meteorological variables. As a rule, one does not expect to find an interaction when the corresponding main effects are not significant, ruling out the possibility of finding interactions involving seasonality based on our model.

Thus, while Knight et al. (1993) found interactions between month and category, between temperature and season, and between humidity and season, no such effects were found in our analysis. We also found no interactions between season and the air pollution covariates, contrary to Knight et al. (1993), who found significantly different effects of the air pollution covariates for the four seasons.

Due to the fact that we consider air pollution measurements as being long-term covariates, it is an assumption in our model that the air pollution effects are the same for all four categories. However, this assumption was confirmed via the residual analysis, in agreement with the findings by Knight et al. (1993).

At a general level, we expect that an analysis in which the correlation structure of the data is taken into account should produce larger standard errors of estimates, and hence fewer significant results, compared with an analysis based on assuming independent counts. However, while Knight et al. (1993) worked mostly with the assumption of
Chapter 10. Data analysis

no serial correlation for the counts, they did perform a check of this assumption by incorporating lagged counts as covariates, creating a kind of autoregressive model. This check showed that their conclusions are not much affected by the assumption of no serial correlation.

The principal difference between the two approaches is that, while Knight et al. (1993) assumed that the four daily counts are conditionally independent given the covariates, we assume conditional independence given the covariates and the latent process, implying that, in our model, the daily counts are correlated given the covariates. The estimated correlations from our analysis were in the range 0.235 to 0.888, indicating that an analysis based on assuming independence of the four categories might create spurious significances. By taking these correlations into account, our model is hence more realistic. As mentioned in Section 10.4, there may in fact be more correlation present than explained by our model, but a more detailed analysis of the correlation structure is unlikely to alter the conclusions by much.

Finally, it should be pointed out that the $\chi^2$-tests used by Knight et al. (1993) in their quasi-likelihood Poisson analysis have not yet been studied in detail in the literature. They use a $\chi^2$-approximation to the scaled deviance difference, but this approximation ignores the randomness of the scale estimator. If anything, this would tend to overemphasize parameter effects, although the effect would disappear for large sample sizes. Moreover, this aspect is likely to be less important than the question of the correlation structure discussed above.
Chapter 11

Further research

This chapter gives a brief discussion of further research for some topics of the thesis, including the following three parts: (1) comparison of the approximate inferences for population-averaged models in Chapter 4; (2) development of an estimation procedure for the stationary time-series models proposed in Chapter 5; and (3) some theoretical questions concerning state space models proposed in Chapters 7 and 8. We will mainly list questions and, for some problems, give specific suggestions.

11.1 Comparison of approximate inferences

The main features of the two likelihood-based approximate inference methods proposed in Chapter 4 may be summarized as follows:

Method based on Pearson residuals: This method relates to an unbiased inference function derived from Pearson residuals. These residuals give bad approximation to normal scores when the dispersion parameter is large. The GEE approach is a limiting case of the method when the dispersion parameter becomes small.

Method based on deviance residuals: This method is associated with a biased inference function derived from deviance residuals. These residuals give good approximations to normal scores even when the dispersion parameter is not small. A bias-correction is usually needed to obtain reasonable estimates.
It is not clear to us whether the bias-correction approach is useful to improve the estimation procedure of the inference based on deviance residuals and how. Before proceeding with any theoretical work, a stimulation study for numerical comparison should be made first. In general, the following questions should be addressed through the simulation.

(a) With regard to the estimates from the approach based on deviance residual approximations, does a bias-correction improve the estimates substantially? Is there any relationship between the roles of bias-correction and the values of dispersion parameter and sample size?

(b) Compare the estimates from the two approximate methods, taking both values of dispersion parameter and sample size into account.

In addition, the application of the multivariate dispersion models in the survival analysis (see Lawless, 1982) would be interesting.

11.2 Estimation for stationary time-series models

For the stationary time-series models, we need to develop the estimation theory of the model and fit the model to the sunspot data that motivated our study.

Due to the similarity of our model and the classic ARMA models, an analogous estimating approach such as least-squares estimation could be used for our models. Maximum likelihood estimation might be difficult because this method requires evaluation of high-dimensional integrals.

Another interesting question is to develop a numerical tool for computing the partial autocorrelation function defined in Chapter 5. It is very difficult to compute such function directly, and therefore approximations seem inevitable. But we do not know by now how to approximate this function.
11.3 Extensions of state space models

For the state space models proposed in Chapters 7 and 8, here are some questions concerning further studies.

(a) Develop models that are able to accommodate high-order Markov structure to replace the need for lagged covariates which appeared in the model for the Prince George data.

(b) Give a theoretical proof of the asymptotic normality for the estimators from the Kalman estimating equation in Chapter 8.

(c) Give a theoretical proof for the convergence of the modified EM-algorithm where the E-step is replaced by the Kalman smoother.

(d) Study the orthogonality of the regression parameters and the dispersion parameters. If this is true, the asymptotic variance matrix of the estimates is block diagonal.

(e) In Chapter 8, extend the state space models to deal with the spatial correlation likely to be present in the case where data from several air monitoring stations and hospitals are available, a problem we avoided by taking daily averages over the six monitoring stations in Prince George. One might address this problem by working with a multi-dimensional latent process, having for example a component for each hospital, while allowing all monitoring stations to influence each component.

(f) Develop a test for the hypothesis that the dispersion parameter of the latent process is zero. This test would provide information about the relevance of the latent process.
Bibliography


