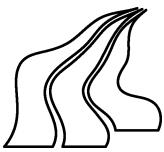
Gradually Changing Seasonal Variation of Cardiovascular Diseases

- A Danish Nationwide Cohort Study

Master of Science Thesis

Spring 2009

Anette Luther Christensen



Department of Mathematical Sciences, Aalborg University
Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital

Department of Mathematical Sciences, Aalborg University
Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital

TITLE:
Gradually Changing Seasonal Variation of Cardiovascular Diseases
- A Danish Nationwide Cohort Study

SEMESTER:
Master of Science Thesis
MAT6, tenth semester

PROJECT PERIOD:
1/2/2009 − 27/5/2009

WRITTEN BY:
Anette Luther Christensen

SUPERVISOR:
Claus Dethlefsen

Number of copies: 7

Number of pages: 144

Finished: D. 27/5/2009

SYNOPSIS:
In this Master of Science Thesis the basic theory of Gaussian and non-Gaussian state space models is outlined, with an application in modelling seasonal variation of incident cardiovascular diseases in the Danish population from 1980 to 2008, identified using the Danish National Registry of Patients.
The thesis consists of three parts. The first part comprises crude analyses of daily incidence rates of incident cardiovascular diseases, i.e. acute coronary syndrome, stroke and venous thromboembolism, modelling seasonal variation characterised by a single cycle during a year and the secular trend as a cubic spline. Furthermore, stratified analyses according to gender and age groups (20-49, 50+) are performed, modelling seasonal variation characterised by four cycles during a year, the secular trend as a cubic spline and the effect of the day of week as unstructured seasonality. All analyses are modelled by a non-Gaussian state space model and residual analyses are performed. Results indicate that incident cardiovascular diseases exhibit gradually changing seasonal variation during the study period.
The second part consists of an article based on a simulation study comparing geometrical models and Poisson regression when modelling seasonal variation. Results of the simulation study indicate that Poisson regression is superior in modelling seasonal variation, when data sets are small.
The third part consists of the basic theory of Gaussian and non-Gaussian state space models. Furthermore, implementation of Kalman forecasting and the EM algorithm is performed, and formulation of state space models modelling seasonal variation in `sspir` is outlined. Several residuals are proposed as model diagnostics.

Preface

This Master of Science Thesis is written by Anette Luther Christensen in the period from February 2009 to May 2009. Parts of Chapter 2 were developed during the ninth semester from September 2008 to December 2008. The thesis is developed at Department of Mathematical Sciences, Aalborg University, in cooperation with Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital. It is assumed, that, as a minimum, the reader possesses the mathematical qualifications corresponding to completion of the bachelor education of Mathematical Sciences at Aalborg University.
The aim of the thesis is to model seasonal variation exhibited by incident cardiovascular diseases in the Danish population from 1980 to 2008 featuring state space models. In addition, the thesis provides a methodological development of modelling seasonal variation, taking as a starting point the results obtained during the ninth semester. The thesis consists of three parts. The first part provides an analysis of seasonal variation of daily frequencies of cardiovascular diseases in the Danish population from 1980 to 2008, applying state space models. During the ninth semester a simulation study comparing geometrical models and Poisson regression when modelling seasonal variation was performed, resulting in an article, which was submitted to *Computer Methods and Programs in Biomedicine*, in April, 2009, and which composes the second part of this thesis. Finally, the third part comprises the basic theory of Gaussian and non-Gaussian state space models, including implementation of Kalman forecasting and the EM algorithm. The code is available by request to lutherstatistiker@hotmail.com.
Citation is given by, e.g. (Joensen et al., 2009) or Joensen et al. (2009). The bibliography is found at the end of the thesis. Scalars are interpreted as 1×1 matrices and vectors of dimension p as $p \times 1$ matrices. Hence no notational differentiation between scalars, vectors or matrices is applied. A list of the notation employed in the thesis is found on page 123.
The author wishes to thank Center for Cardiovascular Research for providing data as well as computer equipment, and her supervisor, Claus Dethlefsen, for his supervision. Furthermore, the author wants to thank all personnel at Center for Cardiovascular Research, especially the team of statisticians, comprised of Claus Dethlefsen, Søren Lundbye-Christensen, Tina Obel and Martin Bøgsted.

Dansk Resumé

I dette speciale skrevet ved Institut for Matematiske Fag, Aalborg Universitet i samarbejde med Kardiovaskulært Forskningscenter, Aalborg Sygehus, Aarhus Universitets Hospital, beskrives den basale teori for Gaussiske og ikke-Gaussiske state space modeller. Desuden indeholder specialet en applikation af denne teori, omhandlende modellering af sæsonvariation udvist af incidente kardiovaskulære sygdomme i den danske befolkning fra 1980 til 2008, identificeret ved hjælp af det Danske Landspatientregister.

Specialet består er tre dele. Den første del omhandler anvendelsen af state space modeller. Materialet, der ligger til grund for anvendelsen, inklusiv datakilder, samt kardiovaskulære sygdomme, beskrives. Endvidere skitseres analysestrategien, der opdeles i to dele. Initielt foretages overordnede analyser af sæsonvariationen af daglige incidens rater af kardiovaskulære sygdomme, nærmere bestemt ved akut koronar syndrom, slagtilfælde og venetromboser, beskrevet ved en enkelt svingning per år, samt en sekulær trend i form af en kubisk spline. Dernæst foretages analyser stratificerede på køn og aldersgrupper (20-49, 50+), for hvilke sæsonvariationen er beskrevet ved fire svingninger per år, en sekulær trend i form af en kubisk spline, samt en ugedagseffekt karakteriseret som ustrukturet sæsonvariation. Alle analyser modelleres som ikke-Gaussiske state space modeller, og endvidere verificeres modellerne ved hjælp af residual analyse. Analyserne viser, at incidente kardiovaskulære sygdomme udviser en sæsonvariation der gradvist ændre sig i løbet af studieperioden.

Den anden del af specialet består af et artikeludkast omhandlende et simulationsstudie, i hvilket der udføres en sammenligning af geometriske modeller og Poisson regression til at modellere sæsonvariation. Resultaterne indikerer, at Poisson regression er at foretrække til at modellere sæsonvariation, når datamaterialet er begrænset. Artiklen er submittet til tidsskriftet *Computer Methods and Programs in Biomedicine* i April, 2009, og er per dags dato under review.

Den trejde del består af den basale teori for Gaussiske og ikke-Gaussiske state space modeller. Implementering af Kalman prediktion og EM algoritmen er udført og beskrevet. Endvidere skitseres formuleringen af state space modeller i programpakken `sspir` i `R` med fokus på modellering af sæsonvariation. Residualer udledes og foreslås anvendt i forbindelse med modelkontrol. Afslutningsvis

eksemplificeres formuleringen af state space modeller med fokus på modellering af sæsonvariation ved hjælp af simulerede data. Desuden foretages et simulationsstudie af EM algoritmen på de simulerede data, efterfulgt af modelkontrol i form af residual analyse.
Appendiks, hørende til specialet, består af tre dele, hvor første del indeholder diverse teoremer og definitioner anvendt i teoridelen. Anden del består af henholdsvis figurer vedrørende det afsluttende eksempel i teoridelen af specialet, samt illustrationer af udvalgte resultater fra analyserne i anvendelsesdelen af specialet. Tredje og sidste del består af en nomenklaturliste. Specialet afsluttes med en litteraturliste.

Contents

Introduction

In Denmark cardiovascular diseases, i.e. diseases related to the heart and blood vessels, are some of the most serious and resource demanding diseases. In fact, cardiovascular diseases are the most frequent cause of death. During the last 20-30 years the relative number of subjects dying from cardiovascular diseases have decreased remarkably for both genders, due to reduction of exposure to risk factors and improvements of treatment.
In 2005, the total expenses for hospitalisations of cardiovascular diseases were nearly five billions Danish kroner. On average, the expenses were higher for men. The expenses of cardiovascular medicine increased from approximately one billion Danish kroner in 2004 to approximately two billions in 2005.
Risk factors of cardiovascular diseases are manifold, and are represented by both biological and genetical factors, as well as the social background. Fatty diet, smoking, alcohol consumption, physical inactivity and working conditions are known risk factors. Exposure to risk factors are dependent on gender, age, level of education and area of living (Nissen and Rasmussen, 2008).

Multiple studies report that the occurrence of incident cardiovascular diseases, i.e. the first time occurrence, varies within the year. This association is reported in both Danish and foreign studies (Frost et al., 2006; Ornato et al., 1996). The majority of reported results indicate, that the frequency of occurrence is higher during the winter than the summer, however some studies report the opposite. The conflicting results may indicate, that the association varies between climatic areas (Ku et al., 1998), and may arise from the fact, that there is a disagreement of which statistical models to be used, along with few cases in each study. Furthermore, there have been reported higher occurrences of incident cardiovascular diseases and higher mortality of cardiovascular diseases during holidays (Phillips et al., 2004; Zubaid et al., 2006).
This reported association of the occurrence of incident cardiovascular diseases and the time of year is commonly called **seasonal variation**. In epidemiology (Manfredini et al., 2004; Fischer et al., 2004; Dowell and Ho, 2004; Altizer et al., 2006; Fisman, 2007; Eilers et al., 2008; Wallis et al., 2008) and economics (Findley et al., 1998) the hypotheses may concern seasonal variation

of events. The **period of seasonal variation** may vary dependent on the context. Studies, investigating seasonal variation during twenty-four hours, a week or a year, are found (Spielberg et al., 1996; Sharma et al., 2001; Spengos et al., 2003; Stein et al., 2004). Commonly, the seasonal variation regardless of the period is described by a single maximum during the period (Frost et al., 2006; Fischer et al., 2005), however studies describing the seasonal variation with multiple locally maxima and a single globally maxima exist (Fischer et al., 2004). Furthermore, seasonal variation is commonly assumed being constant during consecutive periods. However, it is plausible that the seasonal variation changes over time concurrently with progression in treatments and awareness of risk factors (Lundbye-Christensen et al., 2009). Globally, seasonal variation of cardiovascular diseases is an acknowledged feature (Gerber et al., 2006), however studies reporting no seasonal variation of cardiovascular diseases, are also found (Bounameaux et al., 1996).
In Denmark multiple registries exist, which facilitate historically epidemiological cohort studies. The usage of this resource is manifold, and is ideal when investigating diseases for seasonal variation, e.g. cardiovascular diseases. The Danish National Registry of Patients holds information on 99.4% of all non-psychiatric hospitalisations in Denmark from 1977 till 2009, which makes it a valuable resource in epidemiological studies (Andersen et al., 1999). Studies show that the positive predictive values of specific cardiovascular diseases within the Danish National Registry of Patients are estimated to be in the range 58%-80% (Johnsen et al., 2002; Joensen et al., 2009; Severinsen et al., 2008). Preliminary results indicate, that specific incident cardiovascular diseases exhibit seasonal variation described by a single cycle during the year and with highest frequency in January (Christensen, 2008).

Statistically Modelling Seasonal Variation

There exists several statistical models for modelling seasonal variation, including parametrical and non-parametrical models. However in some studies such statistical models are not applied to investigate data, instead subjective conclusions are made based on tabulations and graphical representations. Furthermore, when statistical models are in fact applied, the conclusions are based on simple χ^2 tests (Elwood and Little, 1992).
In 1961, Edwards published his work regarding a statistical model for modelling seasonal variation, commonly referred to as a **geometrical model** (Edwards, 1961; Frangakis and Varadhan, 2002; Brookhart and Rothman, 2008). The basic idea of the geometrical model is to visualise the period of seasonal variation as a circle divided into an appropriate number of time intervals, dependent on frequency by which data are observed, e.g. when the period of seasonal variation is a single year and we have monthly data, the circle may be divided into twelve time intervals. A given weight is assigned to each month according

to the number of events in that specific month and the center of gravity of these twelve weights is determined and exploited to fit an appropriate single cycle sinusoidal curve to data. This model has been considered as standard (Roger, 1977), and seems intuitive and the interpretation is simple. However, the model has several limitations and consequently conclusions based on this model may be incorrect (Wehrung and Hay, 1970; Hewitt et al., 1971; Pocock, 1974; Walter and Elwood, 1975; Roger, 1977).
The assumption of the model is, the time intervals being of equal sizes. This assumption is not fulfilled, when the time intervals represent e.g. months or quarters. An additional assumption is, the population at risk, i.e. the collection of subjects in risk of developing a specific disease, being constant during the period of seasonal variation (Edwards, 1961). Improvements of the geometrical model regarding the assumption of constant time intervals and constant population at risk were introduced in 1975 (Walter and Elwood, 1975). In common for both geometrical models is, that adjustment for an overall trend in data, called the secular trend, is not possible. Furthermore, the models rely on a Gaussian approximation, which becomes poor in case of few events in each time interval, hence conclusions may be incorrect (Gao et al., 2006).
Adjusting for explanatory variables is not explicitly possible, when applying geometrical models, however, by stratification on given variables and performing analysis on each strata, adjusting for explanatory variables is performed implicitly. As a consequence the number of observations in each strata may be too small, hence the Gaussian approximation becomes poor and conclusions may be unreliable.
The **generalised linear models** introduced by Nelder and Wedderburn (1972) enable modelling of events as being Poisson distributed, and additionally, it is possible to adjust for the secular trend and perform regression on explanatory variables (Nelder and Wedderburn, 1972). Results show, that Poisson regression provides reliable conclusions, when the number of observations are small, as opposed to the geometrical models (Christensen et al., 2009a). Seasonal variation modelled by Poisson regression is seen in some epidemiological studies (Fischer et al., 2004, 2005; Thorpe et al., 2004; Eilers et al., 2008).
Both geometrical models and generalised linear models do not allow parameters of the model to vary over time, hence the models are static models. However, it is plausible that the seasonal variation may vary over time, i.e. the amount of fluctuation and time for extrema may change (Lundbye-Christensen et al., 2009).
Harrison and Stevens (1976) introduced a new class of statistical models called **state space models**, which are dynamic models, hence allowing parameters of the model to change over time (Harrison and Stevens, 1976). The model consists of two processes, a latent process and an observation process, the latter being considered as indirect observations of the latent process. Assessment of the latent process is performed by filtering the observations.

The concept of filtering was first introduced by Thiele (1880), however, it was not until Kalman published his work in 1960 and 1963 the application of the filter became clear, consequently the filter is named after Kalman (Thiele, 1880; Kalman, 1960, 1963). Generalised linear models are special cases of state space models, since it is possible to parameterise a generalised linear model as a state space model, hence all distributions belonging to an exponential family may be modelled by a state space model. Additionally, distributions specified only by the first and second moments may be modelled by a state space model (West et al., 1985).
Modelling seasonal variation using a state space model, the seasonal variation is allowed to evolve over time, and we may include the secular trend and regression on explanatory variables also with evolving parameters. Hence, applying state space models in epidemiological studies concerning seasonal variation of events, may provide a more clarifying description of the seasonal variation.

In this thesis we investigate data of incident cardiovascular diseases among the Danish population, since 1980 for gradually changing seasonal variation. Preliminary studies in which data are analysed using Poisson regression indicate that incident cardiovascular diseases exhibit seasonal variation (Christensen, 2008). By analysing the same data, featuring state space models, we may clarify the seasonal variation by allowing the parameters to evolve over time.

Structure of the Thesis

The thesis is organised in three parts, starting with an application of state space models consisting of an analysis of seasonal variation exhibited by incident cardiovascular diseases. The second part consists of a manuscript of an article based on a simulation study performed by the author during autumn 2008 as a part of the ninth semesters project, and the manuscript was submitted in April, 2009 (Christensen, 2008; Christensen et al., 2009a). The third part provides the basic theory of state space models. The chapters are outlined as follows

Chapter 1 Description of data on incident cardiovascular diseases in Denmark from 1980 until 2008 to be analysed, including a description of data sources and data preparation. The study design and analysis strategy are outlined.

Chapter 2 Results of analyses regarding seasonal variation of incident cardiovascular diseases are presented, including residual analyses.

Chapter 3 Manuscript of article regarding simulation study of geometrical models and Poisson regression in modelling seasonal variation.

Chapter 4 Basic theory concerning Gaussian state space models. Kalman filtering, forecasting and smoothing are derived, along with disturbance fil-

tering and smoothing. Estimation of variance matrices, based on the EM algorithm, is derived.

Chapter 5 Basic theory concerning non-Gaussian state space models. Filtering, conjugate filtering, forecasting and smoothing are derived. Iterated extended Kalman smoothing is outlined, as well as estimation of variance matrices, based on the adjusted Pearson algorithm.

Chapter 6 Model diagnostics for Gaussian state space models are derived.

Chapter 7 Formulation of state space models in `R` using the package `sspir` including formulation of secular trend as a cubic spline, harmonic and unstructured seasonal variation and regression on explanatory variable. Ending with an example illustrating the formulation of a Gaussian state space model, estimation of variance matrices, and residual analysis of simulated data.

Part I

Analysis of Gradually Changing Seasonal Variation of Incident Cardiovascular Diseases

Chapter 1

Materials and Methods

This chapter concerns the materials and methods of the analyses. The employed data sources are described, which include the Danish National Registry of Patients and the Central Person Registry. A description of chosen cardiovascular diseases is given and specific cardiovascular diseases are selected to be investigated for seasonal variation. An outline of data preparation is given, followed by a specification of the study design along with an analysis strategy.

1.1 The Danish National Registry of Patients

The Danish National Registry of Patients was established January, 1977, and holds 99.4% of all hospital non-psychiatric records in Denmark (Andersen et al., 1999). The registry includes among other information, the civil registry number, dates of admission and discharge, one or several diagnoses classified according to the Danish version of the International Classification of Diseases, 8th Revision (ICD8) until the beginning of 1994 and afterwards according to the Danish version of the 10th Revision (ICD10) and surgical procedures performed. Furthermore, the registry specifies the hospital and ward of discharge diagnosis, as well as the type of patient and diagnosis. Since 1995 not only hospital admissions are recorded also emergency room- and outpatient contacts are recorded.
The general health and hospital systems in Denmark are non-profit and non-charging systems, that are financed through taxes. Further information about the Danish National Registry of Patients is available at `http://www.sst.dk`.

1.2 The Central Person Registry

The Central Person Registry includes vital status for every resident in Denmark, since April, 1968. Changes in vital status are recorded in the registry. Furthermore, the registry includes, among other information, the civil registry

number and possible change of this, information about civil status, date of possible change, residence and emigration, date of possible change and date of birth. More information about the Central Person Registry is available at `http://www.cpr.dk`.

1.3 Selection of Cardiovascular Diagnoses

Cardiovascular diseases are related to the heart and blood vessels. Among others, these diseases include aneurysm, angina, atherosclerosis, strokes, heart failure, coronary artery diseases, acute myocardial infarction and thromboembolism.
An aneurysm is a pathological dilation of a blood vessel. Angina is characterised by severe chest pain caused by lack of blood supply to the heart and is an ischemic disease, i.e. diseases caused by a reduced blood flow, hence reduced oxygen supply. Atherosclerosis is a slowly progressed hardening of a blood vessel caused by storing of macrophages due to a chronic inflammation in the blood vessel. If a rupture of the atherosclerosis occurs it may lead to narrowing, called stenosis, of the blood vessel or an aneurysm. Strokes are rapidly developing loss of brain functions due to either lack of blood supply, i.e. ischemia, or a haemorrhage, i.e. a bleeding, in the brain.
The coronary arteries, the left and the right, provides the heart muscle with oxygen-rich blood, this is also called the coronary circulation. Coronary artery diseases are the final result of atherosclerosis within the coronary arteries. This may lead to acute myocardial infarction. Myocardium is the heart muscle tissue and infarction means death, hence acute myocardial infarction is characterised by acute damage or death of a part of the heart muscle. Damage of the heart muscle tissue is caused by lack of oxygen supply, hence acute myocardial infarction is an ischemic disease.
Thrombos is the Greek word for coagulation of the blood, hence thrombosis is the formation of a blood clot on the blood vessel and occurs when a blood vessel is injured to prevent loss of blood. When the blood clot dislodge from the blood vessel it is called a thrombus, whereas an embolus is a foreign object, e.g. a thrombus, within a blood vessel, lead through the circulation, embodying risk of blockage of a blood vessel. Hence, thromboembolism is a formation of a blood clot, that is lead through the circulation causing a blockage in a blood vessel in another part of the body (Andersen et al., 2002).
In this study we focus on **acute coronary syndrome** (ACS), i.e. symptoms related to the heart, **stroke**, i.e. symptoms related to the brain and **venous thromboembolism** (VTE), i.e. thromboembolism in the veins. Each of the diagnoses can be divided into several subdiagnoses. Acute coronary syndrome is divided into acute myocardial infarction (AMI), unstable angina pectoris and cardiac arrest as proposed by Joensen et al. (2009). The latter is characterised by an abrupt stop of heart beat leading to arrest of the circulation, which leads

to oxygen deficiency in the whole body. This is different from an AMI, since in that case the blood flow, to a still beating heart, is reduced.
Stroke is divided into four subdiagnoses as proposed by Johnsen et al. (2002), subarachnoid haemorrhage (SAH), intracerebral haemorrhage (ICH), ischemic stroke and unspecified stroke. A subarachnoid haemorrhage is located right outside the brain in the so-called subarachnoid space separated from the cerebral cortex by a membrane called pia matter, whereas an intracerebral haemorrhage is located inside the brain, hence the terms subarachnoid and intracerebral refer to the location of the haemorrhage (Andersen et al., 2002).
Venous thromboembolism diagnoses are divided into deep vein thrombosis (DVT) and pulmonary embolism (PE) as proposed by Severinsen et al. (2008). The first representing the formation of a blood clot, thrombus, in a deep-lying vein often in the legs, which may lead to a pulmonary embolism, i.e. a thrombus dislodge from a vein and lead through the veins and lungs causing a blockage of the artery that leads blood from the heart to the lungs, called the pulmonary artery (Andersen et al., 2002).

1.4 Validation of Selected Diagnoses

The three studies Joensen et al. (2009), Johnsen et al. (2002) and Severinsen et al. (2008) reports estimated positive predictive values (PPV) of discharge diagnoses in the Danish National Registry of Patients of ACS, stroke and VTE, respectively. All studies are based on the Danish prospective cohort called *Diet, Cancer and Health*, see Tjønneland et al. (2007) for a detailed description. Subjects born in Denmark, living in the urban areas of Aarhus and Copenhagen, aged 50-64 years and not registered with a diagnosis of cancer were invited during December, 1993, until May, 1997, to participate in the cohort. A total of 80,996 males and 79,729 females were invited, whereas 27,179 males and 29,876 females accepted. Based on the three studies certain criteria were selected for each disease, ACS, stroke and VTE, in order to obtain high reliability of diagnoses. In Table 1.1, the selected diagnoses to be included in the study are listed with corresponding ICD8 and ICD10 codes along with the corresponding estimated PPV.

1.4.1 Validation of Acute Coronary Syndrome

In the study Joensen et al. (2009) medical records, retrieved from 54 different hospitals, were reviewed by one of three reviewers. In total, 1,654 patients were identified with an incident discharge diagnosis of ACS in the Danish National Registry of Patients among the participants in the cohort, *Diet, Cancer and Health*. Of these, 96 were not characterised either because the medical records were not retrievable ($n = 77$), or because the medical records included insufficient data to classify the patients ($n = 19$).

Disease	ICD8, ICD10	PPV
Acute coronary syndrome (ACS)	410, 427.27, I21, I46	65.5%
Stroke	430-434, 436, I60, I61, I62, I63, I64	79.3%
Venous thromboembolism (VTE)	I26, I80	58.5%

Table 1.1: *Selected cardiovascular diseases with corresponding diagnoses codes (ICD8, ICD10) based on the studies Joensen et al. (2009), Johnsen et al. (2002), Frost et al. (2006) and Severinsen et al. (2008).*

After exclusion of missing medical records the PPV of ACS was 65.5% (95% CI [63.1, 67.9]). When stratifying on subdiagnoses, the PPV of AMI was 81.9% (95% CI [79.47, 8.2]), unstable angina pectoris was 27.5% (95% CI [23.4, 31.9]) and cardiac arrest was 50.0% (95% CI [34.2, 65.8]). Stratifying on type of department of discharge, i.e. ward, emergency room or outpatient, the PPV for a discharge diagnosis on a ward was 80.1% (95% CI [77.7, 82.3]) and the PPV for discharge diagnoses from emergency room or as outpatient was 16.1% (95% CI [12.4, 20.4]). Furthermore, it is reported that stratifying on the type of discharge diagnosis, i.e. primary or secondary, the PPVs are 67.1% (95% CI [64.6, 69.5]) and 47.0% (95% CI [37.6, 56.5]), respectively (Joensen et al., 2009).
Hence, the PPVs differ substantially for the specific subdiagnoses, discharge department and type of diagnosis, furthermore, the study reports a gender specific PPV, males having a significantly higher value for all diagnoses than women. Based on these reports we chose in this study to only include diagnoses of AMI and cardiac arrest discharged from a ward, preferring a high PPVs rather than a large sample size.

1.4.2 Validation of Stroke

In the study Johnsen et al. (2002) medical records were retrieved and validated by one reviewer. In total, 389 patients were identified with an incident discharge diagnosis of stroke in the Danish National Registry of Patients among the participants in the cohort, *Diet, Cancer and Health*. Of these, 377 (96.9%) medical records were retrievable and validated. The PPV of stroke was 79.3% (95% CI [74.9, 83.3]). Stratifying on subdiagnoses the PPV of SAH was 48.3% (95% CI [29.5, 67.5]), ICH 65.7% (95% CI [47.8, 80.9]), ischemic stroke 87.7% (95% CI [80.1, 93.1]) and unspecified stroke 76.0% (95% CI [69.5, 81.7]). Stratifying on department of discharge diagnosis the PPV of stroke from emergency room was 48.8% (95% CI [39.9, 57.8]), whereas from non-speciality departments the PPV was 68.8% (95% CI [61.3, 75.5]) and speciality departments the PPV was 77.9% (95% CI [72.3, 82.7]). It is reported that this characteristic is the same within all subdiagnoses. The study does not find differences in the PPVs, when stratifying on gender or age (Johnsen et al., 2002).
As noted in the study the PPVs are based on a relatively small sample size, hence, making them rather imprecise. However, based on these reports, we chose

to include discharge diagnoses of SAH, ICH, ischemic stroke and unspecified strokes, and addition, as proposed by Frost et al. (2006) also discharge diagnoses identified by I62 (ICD10) are included in this study. Only discharge diagnoses from a ward or outpatient are included.

1.4.3 Validation of Venous Thromboembolism

In the study Severinsen et al. (2008) medical records were retrieved and reviewed by one reviewer. In total, 1,135 patients were identified with an incident discharge diagnosis of VTE in the Danish National Registry of Patients among the participants in the cohort, *Diet, Cancer and Health*. Of these, 1,100 (96.9%) medical records were retrieved and validated. The PPV of VTE was 58.5% (95% CI [55.5, 61.4]). Stratifying on subdiagnoses the PPV of PE was 66.5% (95% CI [62.3, 72.3]) and DVT was 54.6% (95% CI [50.9, 58.2]). Stratifying on department of discharge diagnoses the PPV of VTE diagnosed on a ward was 75.0% (95% CI [71.9, 77.9]) and VTE discharge diagnoses from an emergency room was 31.3% (95% CI [27.0, 35.8]). Stratifying on types of diagnoses, i.e. primary or secondary, the PPVs for VTE are 77.0% (95% CI [73.7, 80.1]) and 66.5% (95% CI [58.4, 73.8]), respectively. The study does not report a significantly difference of PPV stratifying on gender or age for VTE.
Hence, for this study we chose to include both DVT and PE discharge diagnoses from a ward or outpatient and both primary and secondary diagnoses.

1.5 Data Preparation

The data were received as four separated SAS files. Using StatTransfer version 9 all data were transferred to Stata files (StataCorp, 2007). Data preparation was performed in Stata version 10. The statistical analyses were performed in R version 2.9.0 (R Development Core Team, 2008).
Two of the received data sets contained data from the Danish National Registry of Patients, one with primary diagnoses and appurtenant information, containing more than 1.3 million subjects, and the other with the additional diagnoses corresponding to each primary diagnoses. The two data sets were merged using a record number identifying one primary diagnosis with corresponding additional diagnoses. Using a self-written function in Stata subjects having an incident discharge diagnosis of ACS, stroke or VTE were identified, as well as date of diagnosis. Three data sets were created containing subjects identified with ACS, stroke or VTE, respectively. Notice that a subject may be represented in more than one data set. By merging each data set with the Central Person Registry information of gender and birthdate were linked to each subject, as well as date of possible death. Age at time of diagnosis was extracted using birthdate, and grouped into two age groups, 20-49, and 50+.
The Danish population size stratified on gender and age determined January first

every year was extracted as Excel files from `http://www.statistikbanken.dk` and transferred to Stata using StatTransfer. In Stata the data were adequately reshaped and the population size was linearly interpolated in order to obtain data for each day, since January first, 1977. Information of the Danish population size is merged with each data set containing subjects with either ACS, stroke or VTE.

1.6 Study Design

The study is a **historical cohort study**, i.e. a group of selected individuals observed over a given period of time (Rothman, 2002). Furthermore, the cohort study is a **prospective observational study**, i.e. no exposure is assigned to any individual, merely observed through registries. The selected individuals to be observed, i.e. the **study population** or **cohort**, are all inhabitants in Denmark, and the period of time in which the study population is observed, i.e. the **study period**, is January, 1980, until August, 2008. To be included in the study, the subjects must fulfill predefined criteria, which are dependent on the hypothesis of the study, i.e. the **inclusion criteria**. In this study the first criterion to be fulfilled, is subjects being aged 20 or more. This criterion is stated to avoid interference of possible different pathology in children compared with adults.
During the study period the total Danish population increased from approximately 5.12 million to 5.48 million inhabitants, whereas the population restricted to be aged 20 or more, increased from approximately 3.65 million to 4.13 million. In 1980, the population in Denmark consisted of approximately 1.87 million female aged 20 or more. This number increased to approximately 2.11 million in 2008. The number of males aged 20 or more increased from 1.78 million in 1980 to 2.02 in 2008 (Statistik, 2008). An additional characteristic of the cohort is, that the study population may change during the study period, which is referred to as being an **open cohort**.
The **primary endpoint** of the study is the daily frequency of incident diagnoses of cardiovascular diseases, i.e. the daily frequency of first time occurrences of a cardiovascular disease identified using the Danish National Registry of Patients. Consequently subjects with previous diagnosis of a cardiovascular disease are excluded from the study, which leads to a second inclusion criterion, which is, subjects being required to have no previous diagnosis of a cardiovascular disease. Since the Danish National Registry of Patients includes only hospitalisations from January, 1977, information of diagnoses of cardiovascular diseases prior to January, 1977 are not obtainable, hence identified subjects may have a prevalent diagnosis, i.e. at least one previous occurrence of cardiovascular disease. To overcome this circumstance, we provide an **exclusion criterion**, which is to exclude subjects with a diagnosis of a cardiovascular disease from January, 1977 until December, 1979 (Frost et al., 2006).

The daily frequency of incident cardiovascular diseases is determined and the daily **incidence rate** (IR) is calculated according to

$$IR = \frac{\text{number of cases}}{\text{total time of risk}}.$$

The numerator represents the cumulated time at risk of developing an incident cardiovascular disease for all subjects and ensures that the issue of **competing risk** is addressed (Rothman, 2002). Say, we observe 10 occurrences of incident cardiovascular diseases in a population of 100 during a single year. The **risk** of developing an incident cardiovascular diseases within a year is the number of cases divided by the total population during the observation time, hence the risk is 0.1. However, it may happen that some subjects, not developing an incident cardiovascular disease, die from other diseases during the observation time, hence the risk is underestimated. This is the concept of competing risk. The influence of competing risk becomes more pronounced as the study period becomes longer. As for the incidence rate we compare the daily frequency of incident cardiovascular diseases with the total **population at risk**, hence the number of subjects alive except from subjects with previous diagnosis of cardiovascular diseases.
As an **effect measure** of the seasonal variation we introduce the **incidence rate difference** (IRD). By calculating the difference between the highest incidence rate and the smallest, i.e. the incidence rate difference, we obtain an estimate of the peak-to-trough measure, which is independent of the underlying frequency level. In comparison, the **incidence rate ratio** (IRR) determined by the ratio of the highest and smallest incidence rates is dependent on the underlying frequency level.

1.7 Analysis Strategy

For each of the endpoints, ACS, stroke and VTE, the following analyses are performed. Initially, a crude analysis on the daily incidence rates of the disease per $100,000$ is performed by modelling the secular trend as a cubic spline, see Section 6.1, and harmonic seasonal variation described by a single cycle during the year, see Section 6.2.
This analysis is followed by an analysis stratified according to gender and age groups. Hence we perform four analyses for each group of disease, hence allowing the seasonal variation having different levels and shapes for each gender and age group. Furthermore, the harmonic seasonal variation is altered to be described by four cycles during the year, and the effect of the day of week modelled as unstructured seasonality, see Section 6.3, is included in the model (Spielberg et al., 1996; Fischer et al., 2004; Lundbye-Christensen et al., 2009).
Additionally, we analyse a fourth endpoint defined by the daily incidence rates of the first occurrence of an incident diagnosis of either ACS, stroke or VTE,

hence all cardiovascular diseases, this endpoint is denoted cardio. Preliminary results suggest that the overall level of the incidence rates of incident ACS is decreasing during the study period, whereas the level of both stroke and VTE are increasing, in particular the latter (Christensen, 2008). This may suggest that instead of developing an incident ACS, subjects develop an incident VTE, hence the slope of the secular trend, estimated for cardio, may be approximately zero.

Since the daily incidence rates, Y_t, are based on daily observed frequencies, i.e. possibly serially correlated count data, we assume that the incidence rates are Poisson distributed with intensity parameter $(\mu_t/m_t)100,000$, where m_t denotes the total time at risk. The subscript, t, emphasises the dynamic evolvement of the intensity parameter during the study period. Furthermore, we assume the latent process, $\{\theta_t\}$, being Gaussian with constant evolution variance matrix, W. This holds for all analyses.

Due to the dynamic nature of the parameters of the model, the effect measure inherits this characteristic. The estimate of the incidence rate difference at time t may be interpreted as the peak-to-trough measure of a window of the size equaling one year, beginning at time t. In the simple case of a single cycle, the incidence rate difference is directly determined by

$$IRD_t = \exp(A_t) - \exp(-A_t),$$

where $A_t = \sqrt{\alpha_t^2 + \beta_t^2}$ and α_t and β_t are the coefficients of the harmonic seasonality. When the harmonic seasonal variation is characterised by four cycles during a year, the incidence rate difference may be determined by evaluating the seasonal component at an appropriate grid of time points, e.g. daily, in a window of size equaling one year starting at time t in order to identify the maximum and minimum incidence rates, denoted $IR_{t,\max}$ and $IR_{t,\min}$, respectively, and determine

$$IRD_t = \exp(IR_{t,\max}) - \exp(IR_{t,\min})$$

(Lundbye-Christensen et al., 2009).

1.7.1 Model Structure

Each of the analyses, the crude and stratified analyses, are described in the following and the model formulations are outlined. The formulated model for the crude analysis is denoted Model 1, and the formulated model of the stratified analysis is denoted Model 2. Formally mathematical descriptions of model structure and formulation of state space models are given in Chapter 6.

Model 1 - Crude Analyses

Parallel analyses are performed for each of the four endpoints, ACS, stroke, VTE and cardio. The secular trend is modelled by a cubic spline and the seasonal

variation is modelled as a harmonic seasonal variation with a single cycle. Hence we obtain the non-Gaussian state space model

$$\begin{aligned} p(Y_t|\mu_t) &= \exp\left(Y_t \log(\mu_t) - \mu_t + Y_t!\right) \\ \log(\mu_t) &= \lambda_t = F_t^\top \theta_t \\ \theta_t &= G_t\theta_{t-1} + \omega_t, \quad \omega_t \sim \mathcal{N}_4\left(0\,,\, W(\phi)\right) \\ \theta_0 &\sim \mathcal{N}_4\left(m_0\,,\, C_0\right), \end{aligned}$$

where we have

$$F_t = \begin{bmatrix} 1 \\ 0 \\ \cos\left(\dfrac{2\pi t}{365}\right) \\ \sin\left(\dfrac{2\pi t}{365}\right) \end{bmatrix},\ \theta_t = \begin{bmatrix} q(t) \\ q'(t) \\ \alpha_t \\ \beta_t \end{bmatrix},\ G_t = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

and

$$W(\phi) = \begin{bmatrix} \dfrac{\phi_1}{3} & \dfrac{\phi_1}{2} & 0 & 0 \\ \dfrac{\phi_1}{2} & \phi_1 & 0 & 0 \\ 0 & 0 & \phi_2 & 0 \\ 0 & 0 & 0 & \phi_2 \end{bmatrix}.$$

The hyper parameters, $\phi^\top = \begin{bmatrix}\phi_1 & \phi_2\end{bmatrix}$, each represents a feature of the model, where the first hyper parameter, ϕ_1, represents the variance of the secular trend, and the second, ϕ_2, represents the variance of the harmonic seasonality.
This model is comparable with traditional models modelling seasonal variation, since we model the seasonal variation by a harmonic seasonality with a single cycle (Christensen et al., 2009a). Each analysis is subsequently comparable with results reported in Christensen (2008).

Model 2 - Stratified Analyses

We have for each endpoint, ACS, stroke, VTE and cardio, four strata. The first strata is characterised by subjects being females and aged 20-49, the second by subjects being females and aged 50+. The third strata is characterised by subjects being males and aged 20-49, whereas the fourth and final strata is characterised by subjects being males and aged 50+. Hence we perform sixteen parallel analyses, one for each endpoint and each strata.
For each strata the secular trend is still modelled as a cubic spline, whereas the harmonic seasonality is described by four cycles during the year. In addition, we model the effect of the day of week as an unstructured seasonality with a

period of seven. We obtain the non-Gaussian state space model

$$\begin{aligned} p(Y_t|\mu_t) &= \exp\left(Y_t \log(\mu_t) - \mu_t + Y_t!\right) \\ \log(\mu_t) &= \lambda_t = F_t^\top \theta_t \\ \theta_t &= G_t \theta_{t-1} + \omega_t, \quad \omega_t \sim \mathcal{N}_{16}\left(0\,,\, W(\phi)\right) \\ \theta_0 &\sim \mathcal{N}_{16}\left(m_0\,,\, C_0\right), \end{aligned}$$

where

$$F_t^\top = \left[1\ 0\ \cos\left(\tfrac{2\pi t}{365}\right)\ \sin\left(\tfrac{2\pi t}{365}\right)\ \cdots\ \cos\left(4\tfrac{2\pi t}{365}\right)\ \sin\left(4\tfrac{2\pi t}{365}\right)\ 1\ 0\ \cdots\ 0\right]_{1\times 16},$$

$$\theta_t^\top = \left[q(t) \quad q'(t) \quad \alpha_{1,t} \quad \beta_{1,t} \quad \cdots \quad \alpha_{4,t} \quad \beta_{4,t} \quad \gamma_t \quad \cdots \quad \gamma_{t-6}\right]_{1\times 16},$$

and the evolution transfer matrix, G_t, is block diagonal consisting of the three block matrices, G_1, G_2, and G_3, given by

$$G_1 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad G_2 = I_8, \quad G_3 = \begin{bmatrix} -1 & -1 & \cdots & -1 \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}_{6\times 6},$$

each representing a feature of the model, i.e. the secular trend, the harmonic seasonality and the unstructured seasonality, respectively.
The evolution variance matrix is block diagonal consisting of three block matrices, denoted W_1, W_2, and W_3. The first block matrix, W_1, represents the secular trend, the second, W_2, represents the harmonic seasonality, and finally, the third, W_3, represents the unstructured seasonality. Having three hyper parameters $\phi^\top = \left[\phi_1 \quad \phi_2 \quad \phi_3\right]$, the first, ϕ_1, represents the variance of the secular trend, the second, ϕ_2, represents the variance of the harmonic seasonality, assuming equal variances for all four cycles, and finally, the third, ϕ_3, represents the variance of the effect of the day of week. Hence, we have

$$W_1 = \begin{bmatrix} \frac{\phi_1}{3} & \frac{\phi_1}{2} \\ \frac{\phi_1}{2} & \phi_1 \end{bmatrix}, \quad W_2 = \phi_2 I_8, \quad W_3 = \operatorname{diag}(\phi_3, 0, \ldots, 0)_{6\times 6}.$$

The secular trend is parameterised by the level, denoted q, and the slope, denoted q'. The harmonic seasonality is parameterised by four pairs of coefficients, denoted α_i and β_i, where the subscript i denotes which of the four cycles the coefficients parameterise, hence $i = 1, \ldots, 4$. Finally, the unstructured seasonality, i.e. the effect of the day of week, is parameterised by a single coefficient, denoted γ. From this point on, the dependency of the hyper parameters, ϕ, is suppressed, hence the evolution variance matrix is merely denoted W.

1.7.2 Estimation of the Evolution Variance Matrix

In order to obtain a reasonable initial value of the evolution variance matrix, W, we perform a square root transformation of the observed frequencies. The daily incidence rates based on the transformed data, denoted $\breve{Y}_t$, are approximately Gaussian, hence we may estimate the corresponding variance matrices, $\breve{V}$ and $\breve{W}$, using the EM algorithm assuming that, as well as the evolution variance matrix, the observation variance matrix is constant. The EM algorithm provides the variance estimates of the approximated Gaussian state space model, specified by

$$\{F_t, G_t, \breve{V}, \breve{W}\}, \tag{1.1}$$

denoted $\hat{V}^{(0)}$ and $\hat{W}^{(0)}$, respectively. Hence, the EM algorithm maximises the likelihood function, $L_{true}(\phi|\breve{Y}_t)$, of (1.1).
Defining the non-Gaussian state space model given by

$$\{F_t, G_t, \tilde{V}^{(0)}, \hat{W}^{(0)}\}, \tag{1.2}$$

where $\tilde{V}^{(0)}$ denotes the initial value of the observation variance matrix. Notice, that we do not use $\hat{V}^{(0)}$ as an initial value, since this matrix is the estimate of the observation variance for model (1.1) of the daily incidence rates based on the transformed observations, however, we use the estimate of the evolution variance matrix, $\hat{W}^{(0)}$, provided by the EM algorithm.
Applying the iterated extended Kalman smoother to model (1.2), we obtain an approximated Gaussian state space model specified by $\tilde{Y}_t^{(1)}$ and $\tilde{V}_t^{(1)}$ with the property that upon convergence the mode of the likelihood function, $L_{true}(\theta|Y)$, of model (1.2) equals the mode of the likelihood function, $L_{approx}(\theta|\tilde{Y}^{(1)})$, of the approximated Gaussian model. Specifying this approximated Gaussian state space model by

$$\{F_t, G_t, \tilde{V}_t^{(1)}, \hat{W}^{(0)}\}, \tag{1.3}$$

we apply the EM algorithm to model (1.3), hence maximising the likelihood function, $L_{approx}(\phi|\tilde{Y}^{(1)})$, to obtain a new estimate of the evolution variance matrix, W, denoted $\hat{W}^{(1)}$ assuming the observation variance matrix, $\tilde{V}_t^{(1)}$, being known, hence not to be estimated. This may be performed iteratively by initialising a non-Gaussian state space model by (1.2) and apply the iterative extended Kalman smoother. In each iteration we perform the following two steps.

IEKS-step : Apply the iterative extended Kalman smoother on the non-Gaussian state space model

$$\{F_t, G_t, \tilde{V}^{(m-1)}, \hat{W}^{(m-1)}\},$$

which provides the approximated Gaussian model specified by $\tilde{Y}_t^{(m)}$ and $\tilde{V}_t^{(m)}$, denoted

$$\{F_t, G_t, \tilde{V}_t^{(m)}, \hat{W}^{(m-1)}\}.$$

EM-step : Apply the EM algorithm on the approximated Gaussian model from the IEKS-step, assuming $\tilde{V}_t^{(m)}$ being known, which provides $\hat{W}^{(m)}$.

Iterations are performed until convergence is reached, which is defined as

$$\left(\hat{W}^{(m-1)}\right)^{-1} \left|\hat{W}^{(m)} - \hat{W}^{(m-1)}\right| < \epsilon, \tag{1.4}$$

where ϵ is chosen to equal 10^{-3}.
As for Model 2, the stratified analyses, we only estimate the evolution variance matrix based on the transformed observed frequencies for the specific strata consisting of subjects being males and aged 50+ with incident ACS, according to the outlined strategy. This estimated evolution variance matrix, $\hat{W}^{(0)}$, is applied as the initial variance matrix in the first IEKS-step in the remaining fifteen strata.

Chapter 2

Results

This chapter contains the results of the analyses of daily incidence rates per $100,000$ of incident discharge diagnoses of ACS, stroke, VTE and cardio, respectively. Only the results of the analyses regarding incident ACS are thoroughly outlined, including illustrations and residual analysis, whereas only the comprised results of the analysis regarding stroke, VTE and cardio are given, and corresponding illustration are to be found in the Appendix.

We identified in total $274,965$ primary discharge diagnoses of incident ACS from a ward, and $349,099$ primary or secondary discharge diagnoses of incident stroke from a ward or as outpatient contacts, and finally, $130,929$ primary or secondary discharge diagnoses of incident VTE from a ward or as outpatient contacts, using the Danish National Registry of Patients.
Of these identified subjects, $37,869$ were having an incident discharge diagnosis of both ACS and stroke, $10,526$ were having an incident discharge diagnosis of both ACS and VTE, whereas $15,387$ were identified having an incident discharge diagnosis of both stroke and VTE. A total of $2,538$ subjects were identified having an incident discharge diagnosis of both ACS, stroke and VTE.
Well over every third of ACS are females, whereas approximately every second of stroke and VTE are female. In general, females are older, when developing an incident cardiovascular disease, than males. In addition, we identified $686,135$ cases of first occurrences of either ACS, stroke or VTE. See Table 2.1 for a demographic description of identified subjects.

2.1 Acute Coronary Syndrome

Daily incidence rates of incident ACS per $100,000$ were fitted by the non-Gaussian state space model specified by Model 1, which include a secular trend, modelled as a cubic spline and a harmonic seasonal variation with a single cycle

	Total	Female	Age*	
			Female	Male
ACS	274,965	100,009 (36.37%)	75 (66-82)	67 (58-76)
Stroke	349,099	173,742 (49.77%)	76 (67-83)	71 (61-79)
VTE	130,929	69,495 (53.08%)	70 (55-80)	66 (54-75)
Cardio	686,135	313,886 (45.75%)	74 (64-87)	68 (58-77)

Table 2.1: *Demographic description of identified subjects using the Danish National Registry of Patients. Cardio represents the first occurrences of either ACS, stroke or VTE.* * *Median along with first and third quartile in brackets.*

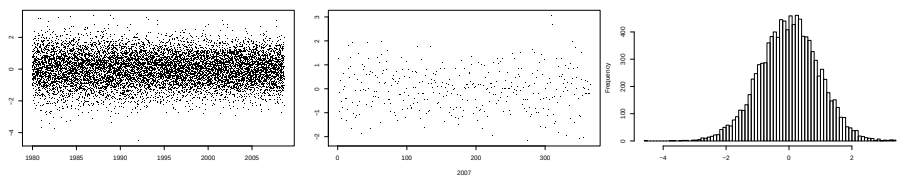


Figure 2.1: *Residual analysis of the approximated Gaussian state space model of daily incidence rates based on the square root transformed observed frequencies. Time plots and histogram of filter residual of origin Y. The latter time plot shows the residuals for the year* 2007.

during the year. The initial evolution variance matrix, W, of the first IEKS-step was estimated by fitting a Gaussian state space model to the daily incidence rates of incident ACS per 100,000 based on the square root transformed daily frequencies.

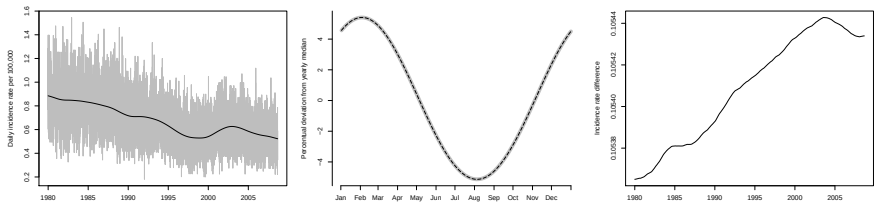
2.1.1 Square Root Transformed Daily Frequencies

Applying the EM algorithm on the Gaussian state space model of the transformed daily frequencies with initial values

$$V^{(0)} = \frac{1}{n}\sum_{t=1}^{n} 100,000\sqrt{Y_t}/m_t = 0.13, \quad \phi^{(0)} = \begin{bmatrix} 10^{-9} \\ 10^{-9} \end{bmatrix},$$

and initial distribution specified by

$$m_0 = \begin{bmatrix} 0.05 \\ 0.05 \\ 0.03 \\ 0.03 \end{bmatrix}, \quad C_0 = 10I_4,$$



(a) (b) (c)

Figure 2.2: *Estimated components of Model* 1. *(a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*

we obtain the estimates of the observation variance matrix and the hyper parameters given by

$$\hat{V}^{(0)} = 2.13 \cdot 10^{-4}, \quad \hat{\phi}^{(0)} = \begin{bmatrix} 3.921399 \cdot 10^{-10} \\ 9.787605 \cdot 10^{-10} \end{bmatrix}.$$

The algorithm converged after 260 iterations with `epsilon` equaling 10^{-3}. Residual analysis is performed for the approximated Gaussian state space model specified by

$$\{F_t, G_t, \hat{V}^{(0)}, \hat{W}^{(0)}\},$$

in order to verify the assumption that the daily incidence rates based on the transformed frequencies are Gaussian. In Figure 2.1, two time plots and a histogram of the filter residual with origin Y are shown. Neither of the time plots indicate misspecification of the observation model, and the histogram may indicate that the daily incidence rates based on the transformed frequencies may reasonably be modelled by a Gaussian state space model.

2.1.2 Crude Analysis

Applying recursively the iterated extended Kalman smoother, with default maximum number of iterations to run equaling 50 and `epsilon` equaling 10^{-6}, followed by the EM algorithm, with maximum number of iterations to run equaling $1,000$ and `epsilon` equaling 10^{-3}, until convergence given by (1.4) on page 20, we obtain estimates of the hyper parameters given by

$$\hat{\phi} = \begin{bmatrix} 3.920857 \cdot 10^{-10} \\ 9.787546 \cdot 10^{-10} \end{bmatrix}.$$

Notice, that the estimates have not altered notably from the initial values, hence, indicating that the iterated extended Kalman smoothing may be redundant.

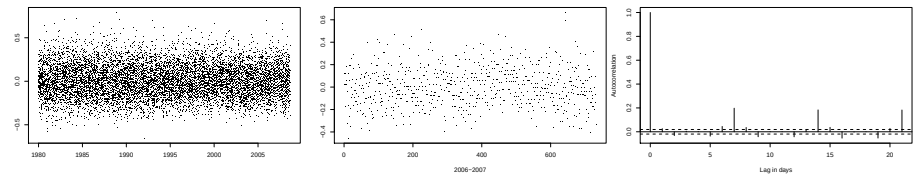


Figure 2.3: *Residual analysis of Model* 1. *Time- and autocorrelation plots of the filter residual with origin* Y.

The estimated secular trend of Model 1 is shown in Figure 2.2(a), which is superimposed on the observed daily incidence rates per $100,000$. In Figure 2.2(b) the seasonal component is shown for 1980 and 2008. Due to the dynamic nature of state space models, the seasonal component alters every day, hence the seasonal component is determined by the estimated coefficients at January first for each year. From this point and on, all estimated measures, e.g. incidence rates or incidence rate differences, are determined by the estimated coefficients at January first the corresponding year. The amplitudes of the seasonal component for each year do not seem to differ. According to Figure 2.2(c), the estimated incidence rate differences vary between 10.537% and 10.544%. In addition, we see that the amplitude of the seasonal variation increases from 1980 until approximately 2004 and decreases afterwards, however, the differences from 1980 to 2004 may not be of any clinically relevance, and too small to be visualised in Figure 2.2(b).
In 1980, the average daily incidence rate of incident ACS is 0.89 per $100,000$. The estimated incidence rate difference is approximately 10.537% with a peak in February, hence the daily incidence rate in February is approximately 0.93 per $100,000$, whereas in August it is only 0.84 per $100,000$. In 2008, the average daily incidence rate is approximately 0.53 per $100,000$. During January, the daily incidence rate is 0.56 per $100,000$ and in July it is 0.51 per $100,000$, since the estimated incidence rate difference is approximately 10.543%. Hence, although the estimated incidence rate difference increases during the study period, the fluctuation in absolute values of the daily incidence rates becomes smaller.
Residual analysis is performed on the resulting non-Gaussian state space model after applying recursively the iterated extended Kalman smoother and the EM algorithm. Time plots of the filter residual for the entire study period and from the years 2006 and 2007 with origin Y are given in Figure 2.3, along with the autocorrelation plot. As seen in the time plots, especially the second, the residuals lie on parallel bands, which is to be expected due to the discreteness of data. The time plot of the residuals from 2006-2007 may indicate a misspecification of the observation model, since the residuals seem to lie on curved bands with a shape, that is repeated in both years. Furthermore, the autocorrelation plot indicates the existence of an effect of the day of week, since every seventh lag

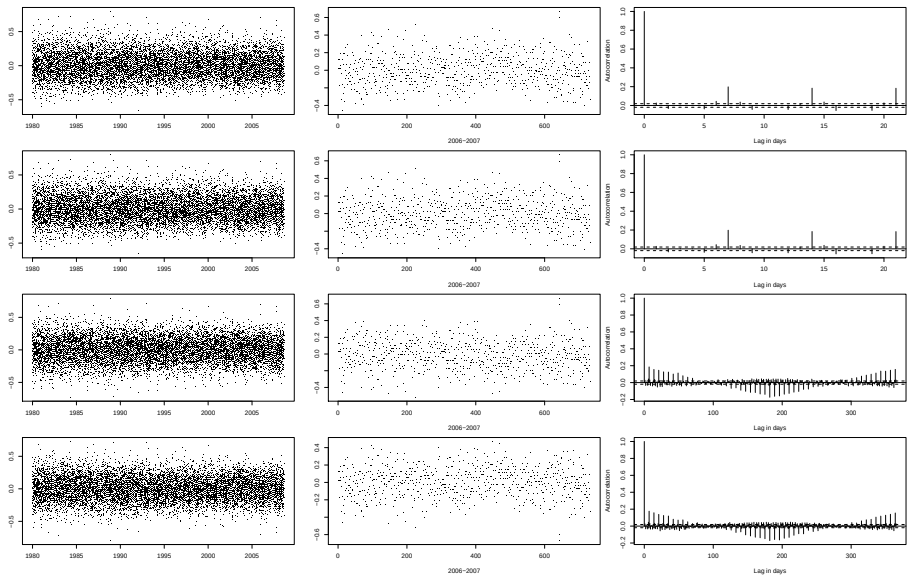


Figure 2.4: *Residual analysis of Model* 1. *Time plots of the components, q, q', α and β, of the filter residuals with origin θ.*

peaks.
In Figure 2.4, time plots of the components of the filter residuals with origin θ are shown. All four plots indicates that the assumption of constant evolution variance matrix may be verified. The time plots of residuals from 2006-2007 of the components q and q' also exhibit the repeated curved bands as the filter residual with origin Y, whereas for the components α and β this pattern is not as pronounced.
The autocorrelation plots of the components q and q' also indicate that an effect of the day of week exists, and in addition, the autocorrelation plots of the components, α and β, both indicate, that the seasonal variation of the observed daily incidence rates may not be explained exclusively be a harmonic seasonality with a single cycle.

2.1.3 Stratified Analyses

Daily incidence rates of incident ACS per $100,000$ stratified according to gender and age groups were fitted by the non-Gaussian state space model specified by Model 2, which include a secular trend, modelled as a cubic spline, a harmonic seasonal variation with four cycles during the year and adjustments of the effect of the day of week modelled as unstructured seasonality of period seven.
The evolution variance matrix is estimated by fitting a Gaussian state space

model to daily incidence rates based on the square root transformed observed frequencies of males aged 50+ with incident ACS and applying the EM algorithm. Initialising the hyper parameters of the Gaussian state space model as

$$\phi^{(0)} = \begin{bmatrix} 10^{-9} \\ 10^{-5} \\ 10^{-7} \end{bmatrix},$$

and applying the EM algorithm, we obtain the estimates of the hyper parameters given by

$$\hat{\phi}^{(0)} = \begin{bmatrix} 1.75946 \cdot 10^{-10} \\ 7.577474 \cdot 10^{-7} \\ 9.176332 \cdot 10^{-8} \end{bmatrix}.$$

These estimates of the hyper parameters, ϕ, are applied as initial values in the iterated extended Kalman smoother for each strata. The final estimates of the hyper parameters for females aged 20-49, and 50+, along with males aged 20-49, and 50+ are

$$\hat{\phi}_{\text{females, 20-49}} = \begin{bmatrix} 1.758898 \cdot 10^{-10} \\ 7.577311 \cdot 10^{-7} \\ 9.176317 \cdot 10^{-8} \end{bmatrix}, \quad \hat{\phi}_{\text{females, 50+}} = \begin{bmatrix} 1.760734 \cdot 10^{-10} \\ 7.575716 \cdot 10^{-7} \\ 9.176022 \cdot 10^{-8} \end{bmatrix},$$

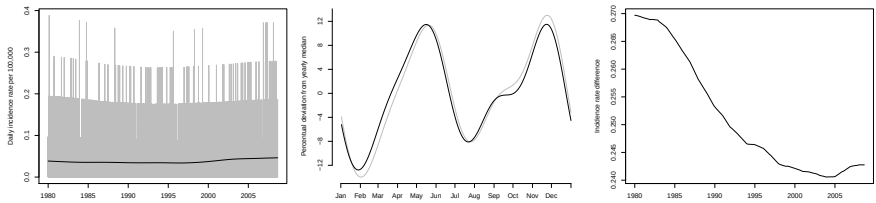
$$\hat{\phi}_{\text{males, 20-49}} = \begin{bmatrix} 1.758747 \cdot 10^{-10} \\ 7.576994 \cdot 10^{-7} \\ 9.176276 \cdot 10^{-8} \end{bmatrix}, \quad \hat{\phi}_{\text{males, 50+}} = \begin{bmatrix} 1.757109 \cdot 10^{-10} \\ 7.572185 \cdot 10^{-7} \\ 9.175041 \cdot 10^{-8} \end{bmatrix},$$

respectively. We see, that the estimates are not notably different between the four strata, and in fact, the estimates have not altered notably from the initial values, which may indicate the iterated extended Kalman smoothing may be redundant, as in the crude analysis.

Females, aged 20-49

The estimated secular trend of Model 2 is shown in Figure 2.5(a) superimposed on the observed daily incidence rates. In general, the secular trend does not change notably during the study period. Figure 2.5(b) shows the seasonal component for 1980 and 2008. The estimated seasonal component indicates that in 1980 two distinct peaks exist, which occur in May and November, whereas the global peak occur in November. The amplitudes of the two peaks seem to be nearly equal in 2008. A small peak occurs during September and is followed immediately by a trough in both years. In addition, two distinct troughs occur during February and July, in both 1980 and 2008. In general, the shape of the seasonal component does not alter notably during the study period. Only three peaks and troughs are immediately visible, indicating that the harmonic seasonality may appropriately be modelled by only three cycles during the year.

The estimated incidence difference rate decreases from 1980 to 2005, afterwards is increases until 2007, where it seems to stagnate, see Figure 2.5(c).



(a) (b) (c)

Figure 2.5: *Estimated components of Model* 2 *for females aged* 20-49. *(a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*

In 1980, the daily incidence rate is approximately 0.038 per 100, 000 and the estimated incidence rate difference is 26.97%. The global peak occurs during November with a maximum daily incidence rate of 0.043 per 100, 000, whereas the global trough occurs in February with a minimum incidence rate of 0.033 per 100, 000. In 2008, the estimated incidence rate difference decreases to 24.28% and the average daily incidence rate is 0.046 per 100, 000, hence, during November, in which the global peak occurs, the daily incidence rate is approximately 0.051 per 100, 000 and during February the lowest daily incidence rate is estimated and equals 0.040 per 100, 000. Due to the relatively low observed daily incidence rates in this strata, the seasonal component may be influenced by noise, hence the actual seasonality may be concealed by noise.

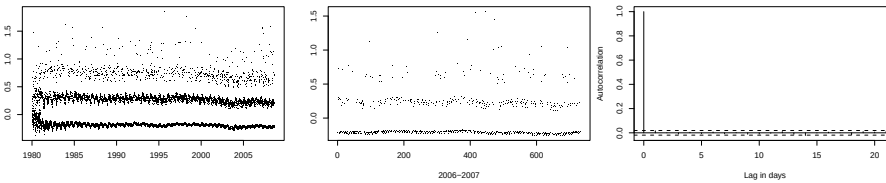


Figure 2.6: *Residual analysis of Model* 2 *for females aged* 20-49. *Time- and autocorrelation plots of the filter residual with origin* Y *of the non-Gaussian state space model. The latter time plot shows the residuals for the year* 2007.

Time plots of the filter residual with origin Y clearly show the discreteness of data, and in addition, the residuals seem to lie on curved bands throughout the entire study period, see Figure 2.6. However, the autocorrelation plot does not indicate any systematic time dependence. Notice, that the peaks in every seventh lag in the autocorrelation plot from the crude analysis, see Figure 2.3,

have been eliminated, possibly due to the modelled effect of the day of week.

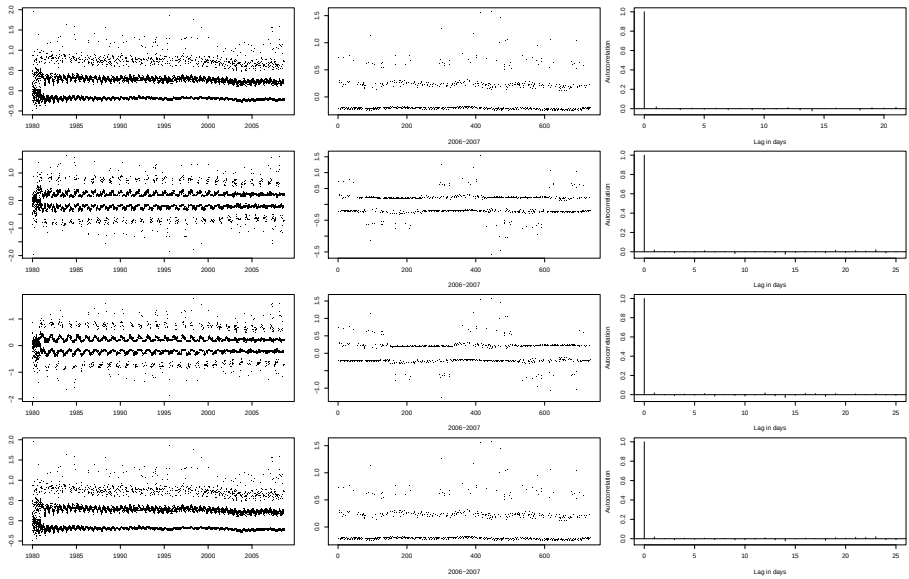


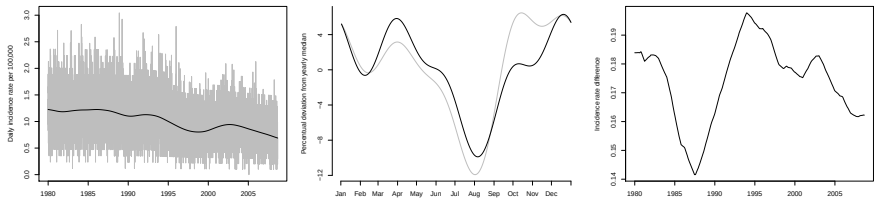
Figure 2.7: *Residual analysis of Model* 2 *for females aged* 20-49. *Time- and autocorrelation plots of the components, q, α_1, β_1 and γ, of the filter residual with origin θ of the non-Gaussian state space model. The time plots in the second column shows the residuals for the years* 2006 *and* 2007.

In Figure 2.7, time- and autocorrelation plots of the components, q, α_1, β_1 and γ, of the filter residuals with origin θ are given. We see, that the plots of the component q are similar to the plots of the filter residual with origin Y, hence the residuals lie on curved bands, whereas the autocorrelation plots do not indicate any time dependence. Again notice that the distinct peaks in every seventh lag in the autocorrelation plot in Figure 2.4 are now eliminated. The time plots of the components α_1 and β_1 exhibit a characteristic behavior, since the residuals lie on curved bands, this behavior becomes less distinct at the end of the study period. The residuals from 2006 and 2007 clearly show a curved behavior, however, the autocorrelation plots do not indicate any systematic time dependence. Neither residual plots of the component γ indicate any misspecification of the effect of the day of week, hence the effect may reasonably be modelled by an unstructured seasonality with period seven.

Females, aged 50+

The estimated secular trend of Model 2 is given in Figure 2.8(a) superimposed on the observed incidence rates and is, in general, decreasing during the study

period, however, with three humps in 1986, 1992 and 2002. The shape of the secular trend is similar with the estimated secular trend of the crude analysis, see Figure 2.2(a). The shape of the seasonal component seems to alter notably as time goes, see Figure 2.8(b). In 1980, the seasonality is described by three distinct peaks during March, October and December and a fourth less distinct peak during June. Similarly, three distinct troughs occur during February, August and November, and a fourth less distinct trough during May. The global peak occurs in October and the global trough occurs in August. As time goes the less distinct peak which occur in June and trough in May becomes more pronounced and in addition, in 2008 the global peak occurs in December. The estimated incidence rate difference decreases from 1980 to 1987, afterwards it increases until 1993, whereas it decreases until 2008 with a peak in 2002, see Figure 2.8(c).



(a) (b) (c)

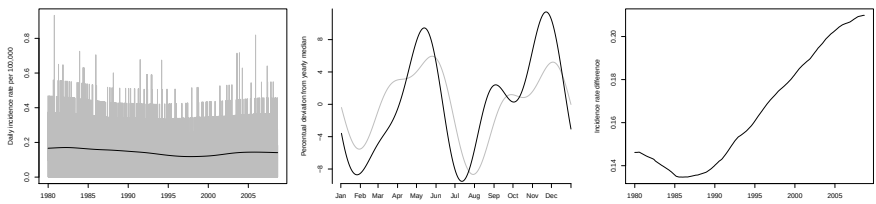
Figure 2.8: *Estimated component of Model* 2 *for females aged* 50+*. (a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*

The average daily incidence rate of 1980 is approximately 1.23 per 100,000. The estimated incidence rate difference is 18.39%, hence during October the daily incidence rate is 1.31 per 100,000 and during August the daily incidence rate is 1.08 per 100,000. This changes during the study period and in 2008 the average daily incidence rate is only 0.72 per 100,000. During December the daily incidence rate is 0.76 per 100,000 and during August it is 0.65 per 100,000, and the estimated incidence rate difference is approximately 16.19%. Residual analysis indicates no obvious misspecification of the model. Time plots of the filter residual with origin from both Y and θ do not indicate any misspecification and the autocorrelation plots do not exhibit any time dependence. The plots have been inspected, however, are not provided in the thesis.

Males, aged 20-49

The estimated secular trend of Model 2 is given in Figure 2.9(a) superimposed on the observed incidence rates, and decreases from 1980 to 2000, and afterwards it increases until 2004, where it seems to stagnate. The seasonal component is

given in Figure 2.9(b) for 1980 and 2008. In 1980, the global peak occurs in May followed immediately by the global trough in August. An additional peak occurs in December. All four peaks and troughs are visible. As time goes the peaks and troughs shift back in time, and in 2008, the global peak occurs in November. The global trough still occurs during summer, however it has shifted back to June. Now only three peaks and troughs are immediately visible, with two distinct peaks and troughs. The estimated incidence rate difference decreases from 1980 until 1985, and afterwards, it increases for the rest of the study period, see Figure 2.9(c). As with females aged 20-49 the daily incidence rates are relatively small, hence the actual seasonality may be concealed by noise.



(a) (b) (c)

Figure 2.9: *Estimated components of Model* 2 *for males aged* 20-49. *(a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*

In general, the observed daily incidence rates of male aged 20-49 with incident ACS are higher than females of same age. In 1980, the average daily incidence rate is 0.17 per 100,000, whereas during May the daily incidence rate is 0.18 per 100,000 and during August it is 0.15 per 100,000. The estimated incidence rate difference is 14.62%. In 2008, the average daily incidence rate is 0.14 per 100,000 and the incidence rate difference is approximately 20.94%. The highest daily incidence rate is 0.16 per 100,000 and the lowest is 0.13 per 100,000.
Residual analysis does not indicate any immediately misspecifications of the model. The time- and autocorrelation plots of the filter residuals are similar with the corresponding plots in Figure 2.6 and Figure 2.7, hence not provided in the thesis.

Males, aged 50+

The observed frequencies of male aged 50+ with incident ACS, are square root transformed and the daily incidence rates based on the transformed frequencies are fitted by a Gaussian state space model, in order to estimate the evolution variance matrix using the EM algorithm. This estimate is applied as the initial evolution variance matrix for the first IEKS step, in order to estimate the corresponding variance matrix of the daily incidence rates based on the observed

frequencies fitted by a non-Gaussian state space model.
In Figure 2.10, two time plots and a histogram of the filter residual with origin Y of the Gaussian state space model are given. All three plots may verify that the daily incidence rates based on the transformed observed frequencies may reasonably be modelled by a Gaussian state space model.

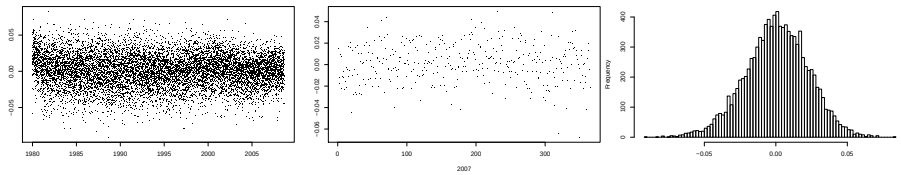
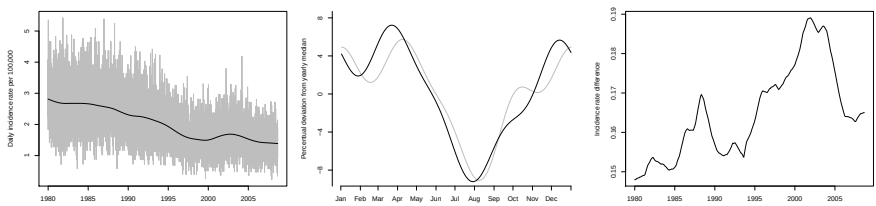


Figure 2.10: *Residual analysis of the approximated Gaussian state space model of the daily incidence rates based on the square root transformed observed frequencies. Time plot and histogram of filter residual with origin Y of the Gaussian state space model. The latter time plot shows the residuals for the year* 2007*.*

The estimated secular trend of Model 2 is given in Figure 2.11(a) superimposed on the daily observed incidence rates. The shape of secular trend is similar with the estimated secular trend of both Model 1 and Model 2 for females aged 50+, see Figure 2.2(a) and Figure 2.8(a). Hence, in general, the secular trend is decreasing throughout the entire study period. The seasonal component is given in Figure 2.11(b) for 1980 and 2008. In general, the seasonal component is characterised by only three immediately visible peaks and troughs. In 1980, the global peak occurs in April, and the global trough occurs in August. In 2008, the global peak occurs in March and the global trough in July, hence the peaks and troughs seem to have been shifted back in time during the study period. The estimated incidence rate difference increases during the entire study period with three notably peaks in 1982, 1988 and 2001.



(a) (b) (c)

Figure 2.11: *Estimated components of Model* 2 *for males aged* 50+*. (a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*

In 1980, the daily incidence rate is approximately 2.81 per 100, 000, and the

estimated incidence rate difference is 14.80%. The global peak occurs during April with a maximum daily incidence rate of 2.98 per 100,000, whereas the global trough occurs in August with a minimum daily incidence rate of 2.56 per 100,000. In 2008, the estimated incidence rate difference increases to 16.44% and the average daily incidence rate is 1.40 per 100,000, hence during March the daily incidence rate is approximately 1.50 per 100,000 and during July the lowest daily incidence rate is estimated and equals 1.27 per 100,000.
Residual analysis does not indicate any misspecifications of the model. Time- and autocorrelation plots of the filter residual with origin from both Y and θ have been inspected, however are not provided in the thesis.

2.2 Stroke

The estimated secular trend of Model 1 is increasing from 1980 to 2003, and decreasing from 2003 to 2008. The seasonal component is characterised by a peak in January, and a trough in July. The estimated incidence rate difference is decreasing in the entire study period, from approximately 9.403% to 9.387%. In 1980, the average daily incidence rate is approximately 0.74 per 100,000, which increases to 0.89 per 100,000 in 2008. The estimated incidence rate differences in 1980 and 2008 are approximately 9.40% and 9.39%, respectively. During January, in 1980, the daily incidence rate is approximately 0.77 per 100,000, whereas during July it is 0.70 per 100,000. This changes in 2008, since the daily incidence rate is approximately 0.94 per 100,000 during January, and only 0.85 per 100,000 in July.
Residual plots indicate that the daily incidence rates based on the transformed observed frequencies may reasonably be modelled by a Gaussian state space model. Furthermore, time plots of the filter residual with origin θ of Model 1 indicate that the evolution variance matrix is constant. The autocorrelation plots indicate an effect of the day of week exists along with indication that not all seasonality is modelled. These plots are, however, not provided in the thesis.

The estimated secular trend of Model 2 for females aged 20-49 increases during the entire study period. In 1980, the average daily incidence rate is 0.07 per 100,000 and the estimated incidence rate difference is 26.85%. The highest daily incidence rate occurs in November and equals 0.08 per 100,000, whereas the lowest occurs in July and equals 0.06 per 100,000. In 2008, the average daily incidence rate increases to 0.14 per 100,000, whereas the incidence rate difference decreases to 22.73%. The highest daily incidence rate still occurs in November and equals 0.16 per 100,000, and the lowest occurs in July and equals 0.13 per 100,000. The estimated incidence rate difference decreases during the study period.
Considering females aged 50+ the estimated secular trend of Model 2 is similar with the estimated secular trend of Model 1. In average, the daily incidence rate,

in 1980, is 1.46 per 100,000 and the incidence rate difference is 18.68%, hence during February the daily incidence rate is 1.58 per 100,000 and during July the daily incidence rate is 1.30 per 100,000. In 2008, the average daily incidence rate is 1.61 per 100,000. The estimated incidence rate difference decreases to 13.76%. During December the daily incidence rate is 1.68 per 100,000, which is the highest, whereas during July it is 1.46 per 100,000 as the lowest. The incidence rate difference is in general decreasing during the study period with peaks in 1990, 1998 and 2008.
The estimated secular trend of Model 2 for males aged 20-49 seems to increase during the study period. In 1980, the average daily incidence rate is 0.071 per 100,000 and the estimated incidence rate difference is 17.32%. During November the daily incidence rate is 0.078 per 100,000 and during July it is 0.066 per 100,000. In 2008, the average daily incidence rate increases to 0.15 and the incidence rate difference increases to 20.79%. The global peak now occurs in December with a daily incidence rate of 0.16 per 100,000, and the global trough is still in July with a daily incidence rate of 0.13 per 100,000. The incidence rate difference is increasing during the study period with a notably peak starting in 1997.
For males aged 50+ the estimated secular trend of Model 2 seems to be constant with three peak occurring in 1983, 1991 and 2008. In 1980, the average daily incidence rate is 1.78 per 100,000, and in 2008, it is 1.87 per 100,000. The incidence rate differences were 13.15% in 1980 and 13.27% in 2008. The global peak occurs in December, in 1980, and the daily incidence rate is 1.89 per 100,000, whereas the global trough occurs in July, in which the daily incidence rate is 1.66 per 100,000. In 2008, the global peak still occurs in December with a daily incidence rate of 1.95 per 100,000, and the global trough occurs in July, with a daily incidence rate of 1.70 per 100,000. The estimated incidence rate difference alters notably during the entire study period.
The estimated components of Model 1 and Model 2 are illustrated in figures provided in Section B.2. Furthermore, by inspection of time- and autocorrelation plots of the filter residuals with origin from both Y and θ, misspecification of the model is not suspected. The residual plots are not provided in the thesis.

2.3 Venous Thromboembolism

The estimated secular trend or Model 1 is decreasing from 1980 to 1992, whereas it is increasing from 1992 to 2008. The seasonal component is characterised by a peak in January followed by a trough in July. The estimated incidence rate difference decreases from approximately 14.352% in 1980 to 14.329% in 2008. See figures in Section B.3.
In average, the daily incidence rate is 0.37 per 100,000 in 1980, whereas in 2008, the average daily incidence rate is 0.45 per 100,000. During January, the daily incidence rate is approximately 0.39 per 100,000 and during July it is 0.34 per

100,000, and the estimated incidence rate difference is 14.35%. In 2008, the peak still occurs during January in which the daily incidence rate is approximately 0.48 per 100,000 and only 0.41 per 100,000 in July with an estimated incidence rate difference equaling 14.33%.
Indicated by a histogram, the distribution of the daily incidence rates based on the transformed observed frequencies seems skewed compared with a Gaussian distribution and in fact the histogram may indicate that the actual distribution is bimodal. Nonetheless the variance matrices are initially estimated by the EM algorithm, since treatment of such distribution is beyond the scope of this thesis. Residual plots of Model 2 indicate a misspecification of the state model, since the residuals are more concentrated around zero in the period from 1990 to 1993 than before and after, hence the evolution variance matrix may not be constant. This is, in fact, indicated by the time plot of the estimated secular trend, by inspecting the observed daily incidence rates, which seem to vary more pronounced in the beginning and end of the study period. Autocorrelation plots indicate that an effect of the day of week and additional seasonality are present. These residual plots are provided in Section B.3.

The estimated secular trend of Model 2 for females aged 20-49 is increasing during the study period. In 1980, the average daily incidence rate is 0.082 per 100,000 and the incidence rate difference is 12.26%. The highest daily incidence rate occurs in November and equals 0.087 per 100,000, whereas the lowest occurs in April and equals 0.077 per 100,000. In 2008, the average daily incidence rate increases to 0.20 per 100,000, and the incidence rate difference increases to 13.76%. The highest daily incidence rate occurs in October and equals 0.21 per 100,000, and the lowest occurs in April and equals 0.19 per 100,000. The estimated incidence rate difference increases during the entire study period.
For females aged 50+ the estimated secular trend of Model 2 is decreasing from 1980 to 1992 and increasing from 1992 to 2008. In average, the daily incidence rate, in 1980, is 0.65 per 100,000 and the incidence rate difference is 27.69%, hence during January the daily incidence rate is 0.74 per 100,000 and during July the daily incidence rate is 0.56 per 100,000. In 2008, the average daily incidence rate is 0.70 per 100,000. The estimated incidence rate difference decreases to 17.96%. During February, the daily incidence rate is 0.76 per 100,000, whereas during August it is 0.63 per 100,000. The incidence rate difference is in general decreasing during the study period with a peak in 1994.
The estimated secular trend of Model 2 for males aged 20-49 increases during the entire study period. In 1980, the average daily incidence rate is 0.073 per 100,000 and the estimated incidence rate difference is 18.50%. During September, the daily incidence rate is 0.079 per 100,000, and during December, it is 0.066 per 100,000. In 2008, the average daily incidence rate increases to 0.15 and the incidence rate difference decreases to 17.29%. The global peak still occurs in September with a daily incidence rate of 0.14 per 100,000, and the global

trough is still in December with a daily incidence rate of 0.12 per 100,000. Notice, that the behavior of the seasonal component seems opposite of the seasonal component for all other strata and from ACS and stroke, since the trough occurs during winter. The incidence rate difference is decreasing during the study period.
For males aged 50+ the estimated secular trend of Model 2 is similar with the estimated secular trend of Model 2 for females aged 50+. In 1980, the average daily incidence rate is 0.73 per 100,000 with a global peak during February of 0.85 per 100,000 and a global trough during July of 0.63 per 100,000. The estimated incidence rate difference is 30.27%. In 2008, the incidence rate difference is only 19.66% and the average daily incidence rate is 0.68 per 100,000. During January, the highest daily incidence rate occurs and equals 0.75 per 100,000, whereas the lowest occurs during May and equals 0.62 per 100,000. The estimated incidence rate difference decreases during the entire study period. The estimate components of Model 1 and Model 2 are illustrated in figures provided in Section B.3. Furthermore, time- and autocorrelation plots of the filter residuals with origin from both Y and θ, are inspected and indicate that assumption of constant evolution variance matrix may not be verified as recognised in the crude analysis. The residual plots are not provided in the thesis.

2.4 Cardio

In general, the estimated secular trend of Model 1 is decreasing during the entire study period, however with humps in 1994 and 2004. Hence, the suggested behavior, that subjects in stead of developing an incident ACS, develop an incident stroke or VTE, does not immediately show. The seasonal component is characterised by a peak in January. The estimated incidence rate difference is decreasing from approximately 10.670%, in 1980, to 10.636%, in 2008.
In 1980, the average daily incidence rate is 1.92 per 100,000 and the estimated incidence rate difference is 10.67%, hence, during January, the daily incidence rate is approximately 2.02 per 100,000, whereas in July it is 1.81 per 100,000. The average daily incidence rate decreases to 1.70 per 100,000, in 2008 and the incidence rate difference decreases to 10.64%. In 2008, the peak occurs in January and the daily incidence rate is approximately 1.79 per 100,000, and, in July, it is 1.61 per 100,000.
The daily incidence rates based on the transformed observed frequencies may reasonably be modelled by a Gaussian state space model. Residual plots of Model 1 indicate that the assumption of constant evolution variance matrix may be verified. As seen in ACS, stroke and VTE autocorrelation plots indicates that the effect of day of week and additional seasonality is present. The plots are not provided on the thesis.

The estimated secular trend of Model 2 for females aged 20-49 is increasing

during the entire study period. In 1980, the average daily incidence rate is 0.19 per 100,000 and the estimated incidence rate difference is 20.46%. The highest daily incidence rate occurs in November and equals 0.21 per 100,000, whereas the lowest occurs in July and equals 0.17 per 100,000. In 2008, the average daily incidence rate increases to 0.39 per 100,000, whereas the incidence rate difference decreases to 13.00%. The highest daily incidence rate occurs in November and equals 0.42 per 100,000, and the lowest occurs in July and equals 0.37 per 100,000. In general, the estimated incidence rate difference decreases during the study period.

Considering females aged 50+ the estimated secular trend of Model 2 is decreasing during the study period. In average, the daily incidence rate, in 1980, is 3.26 per 100,000 and the incidence rate difference is 19.78%, hence during December the daily incidence rate is 3.50 per 100,000 and during July the daily incidence rate is 2.86 per 100,000. In 2008, the average daily incidence rate is 2.79 per 100,000. The estimated incidence rate difference decreases to 13.76%. During December the daily incidence rate is 2.92 per 100,000, whereas during July it is 2.54 per 100,000. The incidence rate difference is in general decreasing during the entire study period with a peak in 1990.

The estimated secular trend of Model 2 for males aged 20-49 seems to increase during the study period. In 1980, the average daily incidence rate is 0.30 per 100,000 and the estimated incidence rate difference is 8.67%. During November the daily incidence rate is 0.31 per 100,000 and during July it is 0.28 per 100,000. In 2008, the average daily incidence rate increases to 0.41 and the incidence rate difference has increased to 14.68%. The global peak still occurs in November with a daily incidence rate of 0.44 per 100,000, and the global trough is still in July with a daily incidence rate of 0.38 per 100,000. The incidence rate difference is increasing during the study period.

For males aged 50+ the estimated secular trend of Model 2 is decreasing during the study period. In 1980, the average daily incidence rate is 5.22 per 100,000, and in 2008, 3.65 per 100,000. The incidence rate differences were 14.90% in 1980 and 14.79% in 2008. The global peak occurs during December, in 1980, and the daily incidence rate is 5.53 per 100,000, whereas the global trough occurs in August in which the daily incidence rate is 4.76 per 100,000. In 2008, the global peak occurs in March with a daily incidence rate of 3.88 per 100,000, and the global trough occurs in July, with a daily incidence rate of 3.34 per 100,000. The estimated incidence rate difference alters notably during the entire study period.

The estimated components of Model 1 and Model 2 are illustrated in figures provided in Section B.4. Furthermore, time- and autocorrelation plots of the filter residuals with origin from both Y and θ, of Model 2 for females and males aged 20-49 may indicate non-constant evolution variance matrix, whereas for females and males aged 50+ no indication of misspecification is revealed. See figures in Section B.4.

2.5 Summary of Results

To summarise the results, we notice that when not stratifying on gender and age groups, the daily highest incidence rates occur during the winter for all cardiovascular diseases, hence the results are in accordance with other reported results (Christensen, 2008). In addition, we notice that the seasonal component does not alter notably during the study period, hence the evolution of the incidence rate difference from the beginning of the study period to the end may not be of any clinically relevance. Commonly, residual analysis indicates that the seasonality may not be exclusively modelled by a harmonic seasonality with a single cycle, furthermore, autocorrelation plots indicate that an effect of the day of week exists.

When stratifying on both gender and age groups, we notice, that subjects aged 50+ more frequently develop an incident cardiovascular disease in comparison with subjects aged 20-49. In general, the global trough occurs during summer, whereas several peaks occur during winter. The seasonal component alters notably during the study period, hence the evolution of the incidence rate difference becomes more pronounced compared with the crude analyses. Most commonly, the estimated incidence rate difference decreases during the study period for subjects aged 50+, except from males with incident ACS or stroke. This behavior also holds for females of any aged and disease, except from females aged 20-49 with incident VTE.

Residual analyses indicate that the seasonal variation of cardiovascular diseases may reasonably be modelled by a harmonic seasonality with four cycles during the year, the secular trend as a cubic spline and the effect of the day of week, indicated by the crude analyses, as unstructured seasonality with period seven.

Discussion

In order to assess the seasonal variation during a year, we have analysed daily incidence rates per $100,000$ of incident cardiovascular diseases. Analyses were performed on four endpoints, i.e. ACS, stroke and VTE, as well as the first occurrence of these three diseases. The incidences were identified using the Danish National Registry of Patients. For each endpoint a crude analysis was performed, i.e. a non-Gaussian state space model was fitted to data, modelling the secular trend as a cubic spline and the seasonal variation as a harmonic seasonality with a single cycle during the year. In addition, each of the four endpoints were analysed stratified on gender and age groups. For each strata a stratified analysis was performed, i.e. a non-Gaussian state space model was fitted to data, modelling the secular trend as a cubic spline and the seasonal variation as a harmonic seasonality with four cycles during the year, and the effect of the day of week as unstructured seasonality.
Estimation of the latent process, θ, and the evolution variance matrix, W, was performed by iteratively applying the iterated extended Kalman smoother and the EM algorithm. At first, the data was square root transformed, hence obtaining an approximated Gaussian state space model, in order to estimate the evolution variance matrix using the EM algorithm. This estimate was applied as the initial evolution variance matrix in the iterated extended Kalman smoother. After convergence of the iterated extended Kalman smoother, we obtained an approximated Gaussian state space model, having likelihood function with same mode as the likelihood function of the non-Gaussian state space model. Iteratively, the iterated extended Kalman smoother and the EM algorithm were applied until convergence was reached.

The present study is essential, since it contributes in clarifying the etiology of cardiovascular diseases, which may improve treatment and preventive strategies. By clarifying the changes of the seasonal component over time, we may evaluate the efficiency of preventive strategies, e.g. the new legislation on smoking, which became effective in 2007, as well as changes in definitions of specific diseases, e.g. a new definition of acute myocardial infarction was introduced in 2000 and implemented during the following years (Nissen and Rasmussen, 2008).

In Denmark, several administrative registries exist and are linked through the civil registration number, which make the present study possible. However, when analysing data obtained from such registries, several sources of errors occur. In the Danish National Registry of Patients hospitalisations from the entire country are registered, hence numerous individuals take part in the updating procedure. Consequently, registrations may be performed in numerous ways, whereas some variables in such registries rely on subjective considerations, e.g. the actual diagnoses. Although, definitions of diseases exist, registrations of the diagnoses are not consistently correct, and are highly dependent on the ward from which the registrations originate (Johnsen et al., 2002; Severinsen et al., 2008; Joensen et al., 2009).

The seasonal variation exhibited by incident cardiovascular diseases is characterised by a notably trough during the summer, as seen in both the crude and stratified analyses. This may indicate a possible association between the development of an incident cardiovascular disease and weather. However, in general, the distinct trough occur during July, in which the regular personnel might take their annual leave. The consequence may be that registrations are performed differently in comparison with the rest of the year. In addition, residual analyses of the crude analyses indicated that an effect of the day of week exist. The present study indicate that Monday is the day of week with the highest frequency of hospitalisations, which is recognised by other studies (Spielberg et al., 1996; Christensen, 2008). This characteristic may be an administrative consequence, rather than a pathological consequence.

In addition, residual analyses of the crude analyses indicated that the seasonal variation exhibited by incident cardiovascular diseases may not exclusively be explained by a single harmonic cycle during the year. By modelling the harmonic seasonality with four cycles during the year, and by including the effect of the day of week, residual analyses of the stratified analyses, show a better fit of the model to the observed daily incidence rates. When applying the geometrical model, which have been considered standard (Roger, 1977), this characteristic was not revealed. Hence state space models may be superior, when modelling seasonal variation, since such model are adaptable to simple seasonality with a single cycle during the period of seasonality, as well as more complex patterns, including several cycles, secular trend, effect of the day of week, and regression on several explanatory variables.

When performing analyses on stratified data according to age groups, we assume that the seasonal variation may be different in each strata. This means that, when subjects relocate to a higher age group, the seasonal variation changes abrupt. However, this strong assumption may seem implausible. When stratifying only on gender, and merely adjust for age as a continuous explanatory variable, the issue still holds, and in addition the shape of the seasonal variation is assumed being identical for all ages, unless interaction between time and age are modelled.

We have restricted ourselves to only focus on the explanatory variables, gender and age, whereas other available adjustments are straightforward, e.g. adjusting for co-morbid diseases, such as diabetes or cancer (Charlson et al., 1987). In fact, when determining the incidence rates, we indirectly adjust for co-morbidity, by determining the total time at risk, hence subjects dying from other diseases, than cardiovascular diseases, are censored. However, it is plausible that subjects having a co-morbid disease, may have higher risk in developing an incident cardiovascular disease. Hence, having a co-morbid disease may have a confounding influence on the seasonality exhibited by incident cardiovascular diseases. Preliminary results indicate that this, in fact, is plausible, since the amplitude of the seasonal variation exhibited by incident unprovoked VTE, i.e. subjects with incident VTE having no previous diagnosis of cancer, or any diagnosis within three months before the diagnosis of incident VTE, with none co-morbid diseases was notably higher, than corresponding subjects with a least one co-morbid disease, for which the amplitude was nearly zero (Christensen et al., 2009b).
In the present study we define a fourth endpoint, as the occurrence of either ACS, stroke or VTE, on which we perform the crude analysis as well as the stratified analyses. It must be noted, that this endpoint consists of an accumulation of two diseases occurring in the arteries, ACS and stroke, and one disease occurring in the veins, VTE. This may rise a question wether these diseases are comparable, however, since the majority of risk factors coincide it may be in order to perform the analyses. The daily incidence rates of ACS seems to be decreasing during the study period, whereas for stroke and VTE the daily incidence rates are increasing, indicating that subjects develop incident stroke or VTE rather than ACS, which should be exhibited by the fourth endpoint. However, results of the analyses of the fourth endpoint do not exhibit this behaviour.

In order to estimate the latent process and the evolution variance matrix, we iteratively apply the iterated extended Kalman smoother, which maximises the likelihood function, $L_{approx}(\theta|\tilde{Y})$, of the approximated Gaussian state space model, followed by applying the EM algorithm, which maximises the likelihood function, $L_{approx}(\phi|\tilde{Y})$, of the same approximated Gaussian state space model, where ϕ denotes the hyper parameters. Hence, we have an ad hoc estimation procedure, since we have not derived, that the maxima of the likelihood functions of the two algorithms are identical. This may be derived analytically or verified by an appropriately designed simulation study. In fact, the estimates obtained upon convergence, do not differ notably from the initial values provided by the EM algorithm applied on the daily incidence rates based on the square root transformed observed frequencies. This may indicate that applying the iterated extended Kalman smoother is redundant, however, further investigations of this issue must be performed, since the actual values of the hyper parameters are small. Simulations may clarify the consequences of applying the iterated extended Kalman smoother, when the values of the evolution variance matrix

are higher. In addition, estimation of the observation variance matrix was not performed, when applying the EM algorithm on the approximated Gaussian state space model provided by the iterated extended Kalman smoother. The sufficiency of estimation of the observation variance matrix, when applying the iterated extended Kalman smoother, may as well by clarified by simulations.
No general derivation of confidence intervals of the peak-to-trough measure is available. Attempts of providing confidence intervals of the peak-to-trough ratio in a static setting, assuming the seasonal variation may be modelled a single cycle sinusoidal curve, e.g. geometrical models, have been published (Frangakis and Varadhan, 2002; Brookhart and Rothman, 2008). Commonly, the estimators of the peak-to-trough ratio rely on the assumption of the square root transformed data being Gaussian, hence the estimators may be influenced by bias in case of small data sets (Christensen et al., 2009a). In constructing confidence intervals for the peak-to-trough measure, the lower limit is restricted to be non-negative, hence complicating the derivations, when maintaining the coverage of the confidence interval to equal e.g. 95% (Frangakis and Varadhan, 2002). No derivations of confidence intervals of the peak-to-trough measure in the dynamic setting are available.
Furthermore, no formal test exists to determine wether an explanatory variable is time varying or static during a given study period. Hence, we can not determine if the seasonal variation exhibited by incident cardiovascular diseases changes significantly during the study period. In case, an explanatory variable is static, the corresponding variance component of the evolution variance matrix equals zero (Dethlefsen and Lundbye-Christensen, 2006), hence the variance component is not to be estimated. As a consequence the computational calculations are reduced and estimation of the evolution variance matrix may be performed faster, which is preferred. However, when no formal tests exist to determine wether an explanatory variable is time varying or static, such decisions rely on subjective considerations, which may require some sort of expert knowledge.

Future work may include analyses of the association between development of incident cardiovascular diseases and the weather conditions, e.g. humidity, precipitation and temperature. Also we want to be able to model gradually changing seasonal variation as subjects becomes older, as we have modelled gradually changing seasonal variation over time. Derivations of a verified estimation procedure, as well as confidence intervals of peak-to-trough measures and formal tests of explanatory variables being time varying or static, are desirable.

Part II

Manuscript of Article

Comparison of geometrical models and Poisson regression modelling seasonal variation - A simulation study

A. L. Christensen[*,a,b], C. Dethlefsen[b], S. Lundbye-Christensen[b]

[a]*Department of Mathematical Sciences, Aalborg University, Fredrik Bayers vej 7G, DK-9210 Aalborg Øst*
[b]*Department of Cardiology, Center for Cardiovascular Research, Aalborg Hospital, Aarhus University, Sdr. Skovvej 15, DK-9000 Aalborg*

Abstract

Seasonal variation is when part of the variation in a time series is described by a repeated temporal cyclic pattern. This is recognized in several epidemiological and economical studies. Geometrical models and Poisson regression are applicable in modelling seasonal variation. In this study we compare two geometrical models and Poisson regression using stochastic simulations. Each model is fitted to simulated data sets consisting of 12 counts of events representing the months of the year, with a total number of events, referred to as the sample size, ranging from 25 to 100,000. The probability of type I error is simulated for each sample size for significance levels ranging from 1% to 20%. The power is simulated for each sample size for amplitudes of seasonal variation ranging from 1% to 20%. Results show that geometrical models too often detect false seasonal variation for sample sizes less than 500, whereas Poisson regression detects false seasonal variation at a frequency equaling the significance level for all sample sizes. The simulated power equals for all three models for sample sizes larger than 500. Based on this simulation study Poisson regression is preferable, when modelling seasonal variation, for small sample sizes in comparison with geometrical models.

Key words: Seasonal variation, Poisson regression, geometrical model, simulation, epidemiology

1. Introduction

In several contexts events are more frequent observed at specific times during a given period of time, e.g. the year or week. This repeated temporal cyclic pattern is commonly referred to as seasonal variation, and is often recognized in epidemiology [1, 2, 3, 4, 5, 6, 7] and economics [8]. Elucidation of seasonal variation of events provides essential knowledge in understanding the nature of the underlying system generating the events.

Edwards (1961) derived a geometrical model to detect seasonal variation of events [9, 10, 11]. This model was considered as the standard in epidemiological studies [12]. However notably critique regarding the model has been published, such as inaccuracies due to small number of events, and neglecting the difference between the absolute count of events and the count relative to both the size of population at risk and length of time interval, e.g. months or quarters [12, 13, 14, 15, 16, 17]. Walter and Elwood proposed a refinement of Edwards' model by handling a varying population at risk and true length of month in 1975. The refined model, however was still not able to handle small number of events [15].

The models are lacking the ability to adjust for covariates. The behaviour of a given system of events may differ according to specific conditions. In epidemiological studies such conditions may be gender, age, time and interventions whereas in economics the state of the market may change behaviour of the system. Stratification on conditions is a solution, however, when several conditions are present, the number of events in each strata may becomes to small. Present in some systems is the secular trend, the overall trend describing the average behaviour of the system.

The generalized linear models of Nelder and Wedderburn (1972) [18] provides models which are applicable to Poisson distributed observations, furthermore seasonal variation, secular trend and regression on covariates may be modelled, hence the models overcome the limitations of the existing geometrical models. Seasonal variation studies using Poisson regression are published by [2], [6], [19], and [20].

In this study we investigate by stochastic simulations the performance of three models in order to estimate seasonal variation quantified by the peak-to-trough ratio. The three models are, first, the model derived by Edwards, second, the refined model derived by Walter and Elwood, and third, Poisson regression. For each model the probability of type I error and power are simulated for sample sizes ranging from 25 to 100,000. Probability of type I errors are simulated for significance levels, 1%, 2%, 5%, 10% and 20%, whereas powers are simulated at a significance level of 5% with seasonal variation of amplitudes 1%, 2%, 5%, 10% and 20%.

2. Modelling Seasonal Variation

Geometrical models are based on the assumption of seasonal variation being characterised by a single cycle sinusoidal

[*]Corresponding author
Email addresses: anluc@rn.dk (A. L. Christensen), cld@rn.dk (C. Dethlefsen), solc@rn.dk (S. Lundbye-Christensen)

curve. The basic idea of the model derived by Edwards is to represent the period of time under investigating for seasonal variation by a unit circle [9]. The circle is divided into k equally sized sectors, representing k time intervals. Each sector is weighted by the square root of the number of events happening in the corresponding time interval.

The number of events in sector i is denoted n_i, and the distribution is assumed being proportional with

$$P_1(N_i = n_i) \propto 1 + \alpha_1 \cos\left((\theta_i - \theta^*) - \frac{\pi}{k}\right), \tag{1}$$

where $\theta_i = \frac{2\pi i}{k}$, $i = 1, \ldots, k$, and θ^* corresponds to the time interval with highest number of events and α_1 is the amplitude of the curve, hence the relative risk, RR_1, is estimated by the peak-to-through ratio

$$\hat{RR}_1 = \frac{1 + \alpha_1}{1 - \alpha_1}. \tag{2}$$

The center of gravity of the observed counts, denoted (x_s, y_s), is given by

$$(x_s, y_s) = \left(\frac{\sum_{i=1}^{k} \sqrt{n_i} \cos\left(\theta_i - \frac{\pi}{k}\right)}{\sum_{i=1}^{k} \sqrt{n_i}}, \frac{\sum_{i=1}^{k} \sqrt{n_i} \sin\left(\theta_i - \frac{\pi}{k}\right)}{\sum_{i=1}^{k} \sqrt{n_i}} \right), \tag{3}$$

where n_i is the observed number of events in the i'th time interval. Denote by d, the distance from origo to (x_s, y_s), then $d = \sqrt{x_s^2 + y_s^2}$. Define a test statistic, T_e, as

$$T_e = 8Nd \tag{4}$$

where N is the sample size, i.e. the summation of all events in the k time intervals. Assuming that x_s and y_s are independent and Gaussian, the test statistic, T_e, is approximately χ^2 distributed with two degrees of freedom. An estimate of the amplitude, α_1, is $\hat{\alpha}_1 = 4d$.

The refined model proposed by Walter and Elwood [15] is based on the same idea as Edwards, however instead of dividing the circle into k equally sized sectors they model the possible difference in length of time intervals. Additionally, the model incorporates a varying population at risk. Denote by m_i the size of population at risk in the ith time interval and let $M = \sum_{i=1}^{k} m_i$. The distribution of the number of events, n_i, are assumed being proportional with

$$P_2(N_i = n_i) \propto m_i \left\{ 1 + \alpha_2 \cos\left[\left(\tilde{\theta}_i - \theta^*\right) - \frac{\pi}{k} \right] \right\}, \tag{5}$$

where $i = 1, \ldots, k$, and $\tilde{\theta}_i$ represents the true angles of the endpoints of the k time intervals, hence the estimated peak-to-trough ratio is

$$R_2 = \frac{1 + \alpha_2}{1 - \alpha_2}. \tag{6}$$

Define a test statistic, T_w, as

$$T_w = \left(\frac{x_s - \mu_{x_s}}{\sigma_{x_s}} \right)^2 + \left(\frac{y_s - \mu_{y_s}}{\sigma_{y_s}} \right)^2,$$

where

$$\mu_{x_s} = \frac{\sum_{i=1}^{k} \sqrt{m_i} \cos\left(\tilde{\theta}_i - \frac{\pi}{k}\right)}{\sum_{i=1}^{k} \sqrt{m_i}}, \tag{7}$$

and

$$\sigma^2_{x_s} = \frac{\sum_{i=1}^{k} \frac{1}{4} \cos^2\left(\tilde{\theta}_i - \frac{\pi}{k}\right)}{\left(\sum_{i=1}^{k} \sqrt{\frac{Nm_i}{M}} \right)^2}. \tag{8}$$

Assuming that x_s and y_s are approximately Gaussian distributed, the test statistic, T_e, is approximately χ^2 distributed with two degrees of freedom. An estimate of the amplitude, α_2, is

$$\hat{\alpha}_2 = \frac{2\left\{ d\sqrt{kM} - \sum_{i=1}^{k} \sqrt{m_i} \cos\left[\left(\tilde{\theta}_i - \theta^*\right) - \frac{\pi}{k} \right] \right\}}{\sum_{i=1}^{k} \sqrt{m_i} \cos^2\left[\left(\tilde{\theta}_i - \theta^*\right) - \frac{\pi}{k} \right]}, \tag{9}$$

where $d = \sqrt{(x_s - \mu_{x_s})^2 + (y_s - \mu_{y_s})^2}$.

Notice, when $\tilde{\theta}_i = \theta_i$ and m_i is constant for all time intervals, the model (1) is a special case of the refined model (5).

2.1. Poisson Regression

Assuming, the number of events for each time interval is Poisson distributed and mutually independent, the number of events may be described by a Poisson regression with intensity parameter λ_i, denoted $n_i \sim \text{Poisson}(\frac{\lambda_i}{m_i})$. Using log link function the linear predictor, η_i, may be

$$\log(\lambda_i) = \eta_i = \Psi_{season}. \tag{10}$$

The distribution of the number of events, n_i, is assumed to be proportional to

$$P_3(N_i = n_i) \propto m_i \exp\left\{ \alpha_3 \cos\left[\left(\tilde{\theta}_i - \theta^*\right) - \frac{\pi}{k} \right] \right\}, \tag{11}$$

hence the seasonal variation is specified by

$$\Psi_{season} = \gamma_1 \sin(\tilde{\theta}_i - \theta^*) + \gamma_2 \cos(\tilde{\theta}_i - \theta^*), \tag{12}$$

where $\alpha_3 = \sqrt{\gamma_1^2 + \gamma_2^2}$. Notice that in general the seasonal variation is not restricted to be characterised by a single cycle sinusoidal curve. The relative risk, estimated by the peak-to-trough ratio, is

$$\hat{RR}_3 = \exp(2\alpha_3). \tag{13}$$

Furthermore, the linear predictor is given by

$$\eta_i = \log(m_i) + \Psi_{season}, \tag{14}$$

where the coefficient of m_i is restricted to be unity, hence the size of population at risk at time interval i is treated as a covariate with coefficient restricted to equal unity. A simultaneous test for $\gamma_1 = 0$ and $\gamma_2 = 0$, corresponding to no seasonal variation, may be performed by a deviance test, which is $\chi^2(2)$.

Model	Deficits	Features
(1)	Small sample sizes Varying time intervals Varying population at risk	
(5)	Small sample sizes	Varying time intervals Varying population at risk
(11)		Small sample sizes Varying time intervals Varying population at risk

Table 1: Deficits and features of the three models.

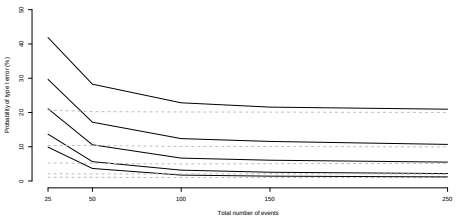


Figure 1: Simulated probability of type I error of the model (5) (solid line) compared with the model (11) (dashed line) for sample sizes in the range 25-250. Results are based on the number of events simulated with true length of month and varying population at risk.

3. Design of Simulation Study

The deficits and features of the three models are summarised in Table 1. On the basis of these deficits and features, the design of the simulation study is created.

For each of the models (1), (5) and (11), the probability of type I error and power are simulated using test for seasonality. The probability of type I error are simulated for thirteen sample sizes ranging from 25-100, 000 and five values of significance levels, 1%, 2%, 5%, 10% and 20%, altogether 25 situations. The powers are likewise simulated for thirteen sample sizes ranging from 25-100, 000 and five values of seasonal variation amplitude, 1%, 2%, 5%, 10% and 20%, all 25 situations at a significance level of 5%. Each situation is based on 100, 000 simulated data sets, where the test statistics, T_e and T_w, for each of the models (1) and (5) are compared with a $\chi^2(2)$ distribution. For the model (11), the test is performed with the $\chi^2(2)$ distributed deviance test.

For the model (1) results are based on two simulated data sets, one appropriate to the model assumptions in (1) and one with events proportional with (5). Results based on the second simulated data set describe the inaccuracies of the model (1). Results for the model (5) is based only on a data set simulated according to (5). Results for the model (11) is based on data sets simulated according to (11) for constant length of month, hence $\tilde{\theta}_i = \theta_i$ and constant population at risk, as well as true length of month and varying population at risk.

To summarise we have the two following simulated data sets,

- True length of months and constant population at risk
- Varying length of months and varying population at risk

where each of the two types of data set is simulated for all thirteen sample sizes.

All simulations are performed in R version 2.8.1 [21] using an Intel Pentium 4 3.2GHz processor with Microsoft Windows XP, version 2002, Service Pack 2.

4. Results

Simulations with constant length of month and constant population at risk show, that the simulated probability of type I error of the model (1) equals the significance level for sample sizes larger than 500. For smaller sample sizes the simulated probability of type I error is notably higher than the significance level, hence the model too often detects false seasonal variation. The simulated type I error of the model (11) equals the significance level for all sample sizes.

Simulations with true length of month and varying population at risk show, that the probabilities of type I error of the model (11) equal the significance level for all sample sizes and significance levels, whereas for the model (5) the probability of type I error is notably higher than the significance level for sample sizes smaller than 250, see Figure 1. For larger sample sizes, the probability of type I error equals the significance level, hence equals the probability of type I error of the model (11).

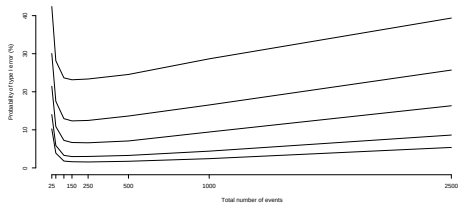
The simulated powers of the model (11) equal the powers of both the models (1) and (5) for all amplitudes of seasonal variation and for sample sizes larger than 500. The power for sample sizes smaller than 250 of the model (5) is higher, than the model (11) as a consequence of the high probability of type I error of the model (5).

As for the model (1) the probability of type I errors are notably larger than the significance level and the profile of the probability of type I error for increasing sample size is described by a 'u'-shaped curve for small sample sizes and a monotonically increasing curve for larger sample sizes, which is illustrated in Figure 2.

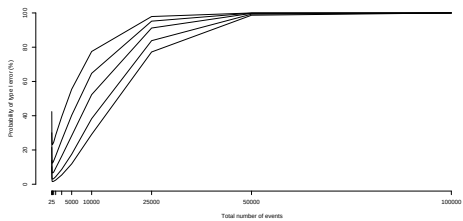
The simulated type I error for each model for selected sample sizes, 25, 1, 000 and 100, 000, and significance levels, 0.01, 0.05 and 0.10, are shown in Table 2 as well as the simulated powers for selected sample sizes, 25, 1, 000 and 100, 000, and amplitudes, 0.01, 0.05 and 0.20.

5. Discussion

In this study three models to model seasonal variation were evaluated according to the simulated probability of type I error and simulated power for several sample sizes and significance levels, two geometrical models and a regression model. The two geometrical models performed poorly for small sample



(a) Simulated probability of type I error for sample sizes ranging from 25-2, 500 illustrating the 'u'-shape of the curve for small sample sizes.



(b) Simulated probability of type I error for sample sizes ranging from 25-100, 000.

Figure 2: Simulated probability of type I error of the model (1). Results are based on the number of events proportional with (5). Notice, that the simulated probability of type I error increases as the sample sizes increases.

	Simulated probability of type I error			Simulated power		
	Significance level			Amplitude		
N	0.01	0.05	0.10	0.01	0.05	0.20
	10.25	21.45	30.05	25.71	25.95	31.78
25	9.88	21.10	29.66	25.69	25.97	31.31
	1.10	5.26	10.34	5.31	5.54	9.38
	2.43	9.43	16.53	9.81	20.83	97.79
1, 000	1.08	5.11	10.28	5.62	15.55	98.62
	0.98	4.99	10.22	5.36	15.52	98.50
	99.99	100.00	100.00	99.96	99.71	100.00
100, 000	0.98	5.00	9.82	50.48	100.00	100.00
	1.00	5.13	10.06	50.35	100.00	100.00

Table 2: Simulated probability of type I error and power the models (1), (5) and (11), respectively. N represents the sample size.

sizes, whereas the regression model performed well for all sample sizes. For larger sample sizes all three models performed well, when data were simulated accordingly to the given model.

Statistical models may only be viewed as an attempt to model the reality, since no models are correct in describing the reality. One model may be appropriate in a given context, whereas to fail in another. Hence a model is neither correct or wrong, however a model may be more appropriate to model a given association than another. It is crucial to pay attention to the reality, which a given model attempts to model, since violations of the assumptions of the model may result in incorrect conclusions.

Using the model derived by Edwards to model seasonal variation it is assumed that the time intervals in which the number of events is observed are constant, further it is assumed that in each time interval the maximum number of events possible is constant, hence the population at risk is constant. The latter assumption may not be fulfilled, and the assumption of constant time intervals depends on the given seasonal variation of interest. Seasonal variation during a week provides constant time intervals as well as during a single day, whereas monthly seasonal variation provides time intervals of length 28 to 31. It is shown in this study, that attempting to model seasonal variation with the model derived by Edwards, when the assumptions are not fulfilled results in a high probability of type I error, hence too often a false seasonal variation is detected.

The refined model compensate these assumption by allowing varying length of time intervals as well as varying population at risk. In fact the probability of type I error equals the significance level for large sample sizes. However for small sample sizes the probability of type I error is higher than the significance level.

Neither the model derived by Edwards or the refined model are appropriate for small sample sizes, due to high probability of type I error, however results show that Poisson regression performs acceptable for small sample sizes as well as large sample sizes, based on the probability of type I error, which equals the significance level for all significance levels. The Poisson regression allows varying length of time intervals and varying population at risk.

An additional assumption of the two geometrical models is that the possible seasonal variation must be described be a single cycle sinusoidal curve during the period of seasonal variation. It is plausible that in some context this assumption can not be fulfilled, hence the models are not appropriate even though all other assumptions are accommodated.

Using a regression model the seasonal variation may be specified in several ways, including multiple cycles sinusoidal curves and polynomials. Additionally, secular trend and regression on covariates may be modelled, hence providing a more complex model that may explain data more appropriately.

The model derived by Edwards is easy to interpret and compute. The output from the model is an estimate of the relative risk and the corresponding test statistic, hence the model performs the test of significantly seasonal variation. In the statistical software Stata [22] the model is implemented and further an

implementation on the internet is available. However as shown in this study, when the assumptions of the model are violated, the conclusions may be wrong.

An advantage of the Poisson regression model is that in the majority of statistical software an implementation of generalized linear models, of which Poisson regression is a special case, exists. Hence this model is also available to the researcher using statistical software, like Stata or R [21]. However the interpretation of this model may be more advanced to a researcher, since the model provides estimates of the regression coefficient, and do not perform a test of significantly seasonal variation, which must manually performed by a deviance test.

Often a researcher wants to compare his result with other studies. The model derived by Edwards provides a comparable output, in the sense that the relative risk is based on only the data representing the number of events for each time interval, no other information is used. As for the Poisson regression, the secular trend and regression on several covariates may be modelled to estimate the relative risk, hence making a comparison fragile, in the sense that all covariates may not be obtainable in all studies. Hence comparisons of results provided by Poisson regression must be performed with caution.

However when only the linear predictor in the Poisson regression consists of the seasonal variation term and the offset, the result is not only comparable with other such specified Poisson regression, but also with the result provided by the model derived by Edwards.

This study shows that when the assumptions of the model derived by Edwards are not accommodated, the model tends to detect false seasonal variation for small sample sizes, hence for rare events and small studies. The model is highly available to the researcher and it provides comparable results. Poisson regression provides a model of seasonal variation that performs acceptable for all sample sizes. This model is also highly available, whereas comparisons of results must be performed with caution. In conclusion, based on the simulations in this study Poisson regression is preferable to model seasonal variation in comparison with the geometrical models.

References

[1] E. Manfredini, M. Gallerani, B. Boari, R. Salmi, R. H. Mehta, Seasonal Variation in Onset of Pulmonary Embolism is Independent of Patients' Underlying Risk Comorbid Conditions, Clinical and Applied Thrombosis/Hemostasis 10 (1) (2004) 39–43.

[2] T. Fischer, S. Lundbye-Christensen, S. P. Johnsen, H. C. Schønheyder, H. T. Sørensen, Secular Trends and Seasonality in First-Time Hospitalization for Acute Myocardial Infarction — a Danish Population-Based Study, International Journal of Cardiology 97 (3) (2004) 425–431.

[3] S. F. Dowell, M. S. Ho, Seasonality of Infectious Diseases and Severe Acute Respiratory Syndrome — What WeDdon't Know Can Hurt Us, Lancet Infect Dis 4 (2004) 704–708.

[4] S. Altizer, A. Dobson, P. Hosseini, P. Hudson, M. Pascual, P. Rohani, Seasonality and the Dynamics of Infectious Diseases, Ecology Letters 9 (2006) 467–484.

[5] D. N. Fisman, Seasonality of Infectious Diseases, Annu. Rev. Public Health 28 (2007) 127–143.

[6] P. H. C. Eilers, J. Gampe, B. D. Marx, R. Rau, Modulation Models for Seasonal Time Series and Incidence Tables, Statistics in Medicine 27 (2008) 3430–3441.

[7] D. E. Wallis, S. Penckofer, G. W. Sizemore, The "Sunshine Deficit" and Cardiovascular Disease, Circulation 118 (2008) 1476–1485.

[8] D. F. Findley, B. C. Monsell, W. R. Bell, M. C. Otto, B. C. Chen, New Capabilities and Methods of the X-12-ARIMA Seasonal Adjustment Program, Journal of Business and Economic Statistics 16 (2) (1998) 127–176.

[9] J. H. Edwards, The Recognition and Estimation of Cyclic Trends, Ann. Hum. Genet. Lond. 25 (1961) 83–86.

[10] M. A. Brookhart, K. J. Rothman, Simple Estimators of the Intensity of Seasonal Occurrence, BMC Medical Research Methodology 8 (2008) 67–75.

[11] C. E. Frangakis, R. Varadhan, Confidence Intervals for Seasonal Relative Risk with Null Boundary Values, Epidemiology 13 (2002) 734–737.

[12] J. H. Roger, A Significance Test for Cyclic Trends in Incidence Data, Biometrika 64 (1) (1977) 152–155.

[13] D. A. Wehrung, S. Hay, A Study of Seasonal Incidence of Congenital Malformations in the United States, British Journal of Preventive and Social Medicine 24 (1970) 24–32.

[14] D. Hewitt, J. Milner, A. Csima, A. Pakula, On Edwards' Crition of Seasonality and a Non-parametric Alternative, British Journal of Preventive and Social Medicine 25 (1971) 174–176.

[15] S. D. Walter, J. M. Elwood, A Test for Seasonality of Events with a Variable Population at Risk, British Journal of Preventive and Social Medicine 29 (1975) 18–21.

[16] S. J. Pocock, Harmonic Analysis Applied to Seasonal Variations in Sickness Absence, Applied Statistic 23 (2) (1974) 103–120.

[17] F. Gao, K. Chia, I. Krantz, P. Nordin, D. Machin, On the Application of the von Mises Distribution and Angular Regression Methods to Investigate the Seasonality of Disease Onset, Statistics in Medicine 25 (2006) 1593–1618.

[18] J. A. Nelder, R. W. M. Wedderburn, Generalized Linear Models, Journal of Royal Statistical Society 135 (3) (1972) 370–384.

[19] L. E. Thorpe, T. R. Frieden, K. F. Laserson, C. Wells, G. R. Khatri, Seasonality of Tuberculosis in India: Is It Real and What Does It Tell Us?, Lancet 364 (9445) (2004) 1613–1614.

[20] T. Fischer, S. P. Johnsen, L. Pedersen, D. Gaist, H. T. Sørensen, K. J. Rothman, Seasonal Variation in Hospitalization and Case Fatality of Subarachnoid Heamoorhage - A Nationwide Danish Study on 9, 367 Patients, Neuroepidemiology 24 (2005) 32–37.

[21] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3900051070, `http://www.R-project.org` (2008).

[22] StataCorp, Stata Statistical Software: Release 10, StataCorp, College Station, TX: StataCorp LP, `http://www.stata.com/` (2007).

Part III

Basic Theory of State Space Models

Chapter 3

Gaussian State Space Models

This chapter provides the basic theory of Gaussian state space models, which forms the basis for the non-Gaussian state models in the consecutive chapters. The definition of Gaussian state space models is stated and the Kalman filter, -forecaster and -smoother are derived. Furthermore, the EM algorithm is derived for Gaussian state space models to estimate the variance matrices.

Establishing some introductive terminology we define a **time series** as a collection of random variables $\{Y_{k_t} \mid t = 1, \ldots, n\}$, measuring a single response of a single subject at times k_t. In case, several responses for a single subject are measured, data is referred to as a **multivariate time series**. **Longitudinal data** consists of a collection of univariate or multivariate time series measuring the same response or responses, respectively, over time, for several subjects. The subscript of k indicates that observations of the random variables Y_{k_t} are not necessarily observed at constant time intervals, hence observations are not necessarily **equidistant**. However, to avoid heavy notation, observations of the random variable Y_{k_t} over time are denoted merely Y_t in case of both equidistant and non-equidistant observations.
State space models are **dynamic models**. Thus state space models are adaptable to changes in parameters, and, in fact, to a reduction or expansion of the dimension of parameters. They apply to multivariate time series and longitudinal data.
The observations of the random variables $\{Y_t \mid t = 1, \ldots, n\}$ are considered as indirect measurements of the **latent process**, $\{\theta_t \mid t = 1, \ldots, n\}$. The latent process is assumed being a Markov process with parameters varying over time. A state space model is specified by the distribution of the observations conditional on the latent process. As observations arrive the distributional parameters are

updated.
Regarding inference, concerning state space models, we distinguish between three terms, **assessment**, **prediction** and **forecasting**. The first term concerns inference of the latent process, whereas the term prediction concerns inference of future states of the latent process, and finally, forecasting referrers to inference concerning future observations.
At each time t, all information available is defined formally as follows.

Definition 3.1 (Information sets)
*Let Y_t be a $(d \times 1)$ vector observed at times $t = 1, \ldots, n$. At time t, the **information set**, D_t, is recursively defined by*

$$D_t = \{D_{t-1}, Y_t\}.$$

*Especially, at time $t = 0$, the information set D_0 is called the **initial information set** and contains all available and relevant prior information at time $t = 0$.* ◇

As a consequence of Definition 3.1 all past information is contained in the information set, D_t, at time t.

3.1 Definition

Gaussian state space models are defined formally as follows.

Definition 3.2 (Gaussian state space model)
*Let Y_t be a $(d \times 1)$ vector observed at times $t = 1, \ldots, n$. Y_t is described by a **Gaussian state space model** if, given a set of quadruples*

$$\{F_t, G_t, V_t, W_t\} = \{F, G, V, W\}_t$$

for each t, where

$$\begin{aligned} F_t &\text{ is a known } (p \times d) \text{ matrix} \\ G_t &\text{ is a known } (p \times p) \text{ matrix} \\ V_t &\text{ is a known } (d \times d) \text{ matrix} \\ W_t &\text{ is a known } (p \times p) \text{ matrix,} \end{aligned}$$

the relations between Y_t and a $(p \times 1)$ parameter vector, θ_t, at time t, as well as the relation between θ_t and the sequence of θ_t through time are determined by the conditional distributions

$$\text{Observation model} : Y_t|\theta_t, D_{t-1} \sim \mathcal{N}_d\left(F_t^\top \theta_t\,,\, V_t\right) \tag{3.1}$$

$$\text{State model} : \theta_t|\theta_{t-1}, D_{t-1} \sim \mathcal{N}_p\left(G_t\theta_{t-1}\,,\, W_t\right) \tag{3.2}$$

$$\text{Initial distribution} : \theta_0|D_0 \sim \mathcal{N}_p\left(m_0\,,\, C_0\right), \tag{3.3}$$

hence

$$\begin{aligned} \textit{Observation equation} &: Y_t = F_t^\top \theta_t + \nu_t, \quad \nu_t \sim \mathcal{N}_d\left(0\,,\, V_t\right) && (3.4)\\ \textit{State equation} &: \theta_t = G_t\theta_{t-1} + \omega_t, \quad \omega_t \sim \mathcal{N}_p\left(0\,,\, W_t\right), && (3.5) \end{aligned}$$

where the sequences $\{\nu_t\}$ *and* $\{\omega_t\}$ *are internally and mutually independent, and independent of* θ_0 *conditional on* D_0. ◇

In Figure 3.1, the conditional independence structure of a state space model is illustrated.

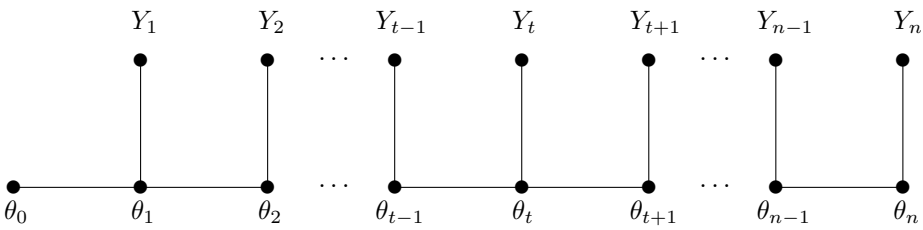


Figure 3.1: *Illustration of the conditional independence structure of the latent process and the observations in a state space model.*

It follows from Definition 3.2, and is illustrated in Figure 3.1, that

$$\begin{aligned} Y_t \perp\!\!\!\perp D_{t-1} \mid \theta_t, \quad \forall t \\ Y_{t+k} \perp\!\!\!\perp D_{t-1} \mid D_t, \quad \forall k > 0 \\ \theta_t \perp\!\!\!\perp D_{t-1} \mid \theta_{t-1}, \quad \forall t. \end{aligned}$$

The matrix F_t is called the **design matrix** and contains all the observed explanatory variables, the vector θ_t is called the **state vector**, and the quantity $\lambda_t = F_t^\top \theta_t$ is called the **signal**. For Gaussian state space models the signal equals the mean of the observations conditional on the latent process, $\mu_t = \mathbb{E}[Y_t \mid \theta_t]$, however, this is not a general result. The matrix G_t is the design matrix of the state model, and is called the **evolution transfer matrix**. Usually, the evolution transfer matrix is block diagonal, each block representing a feature of the model, e.g. secular trend and seasonal variation. Occasionally, the state model may be referred to as the evolution model, emphasising the dynamics of the latent process.
The variance matrices V_t and W_t determines the uncertainty associated with the observation- and state models, respectively, and are called the **observation-** and **evolution variance matrix**, respectively. The vectors ν_t and ω_t are called the **observation-** and **evolution errors**, respectively. All the matrices, F_t, G_t, V_t and W_t, may depend on an unknown parameter vector $\phi^\top = \begin{bmatrix}\phi_1 & \cdots & \phi_l\end{bmatrix}$,

called **hyper parameters**, denoted e.g. $W_t(\phi)$. The dependency of the hyper parameters is suppressed in the notation.
The vector of all observations is denoted $Y^\top = \begin{bmatrix} Y_1^\top & \cdots & Y_n^\top \end{bmatrix}$, hence Y contains all information except for the initial information, and, in fact, Y is an ordered version of the information set, D_n. Similar the vector of all state vectors is denoted $\theta^\top = \begin{bmatrix} \theta_1^\top & \cdots & \theta_n^\top \end{bmatrix}$.

3.2 Kalman Filtering

We are often interested in assessing the latent process based on past observations. Using the Kalman filter, knowledge of the latent process is updated, whenever a new observation becomes available. This estimation procedure was first derived by Thiele (1880), however, it was not until the works by Kalman, (Kalman, 1960) and (Kalman, 1963), were published the applications became clear, therefore the estimation procedure is named after Kalman. The Kalman filter is a recursive updating scheme and calculates the best linear unbiased predictor of the mean, $\mathbb{E}[\theta_t \mid D_t]$, and the corresponding variance matrix, $\mathbb{V}\mathrm{ar}[\theta_t \mid D_t]$, of the posterior distribution of θ_t conditional on D_t (Klein, 2003), and is stated as follows.

THEOREM 3.1 (KALMAN FILTER)
Let $Y_1, \ldots, Y_n$ be described by a Gaussian state space model. For each t, the updating of the state vector is performed according to the following conditional distributions,

$$\begin{aligned}
&\textit{Prior}: && \theta_t | D_{t-1} \sim \mathcal{N}_p \left(\underbrace{G_t m_{t-1}}_{a_t}, \underbrace{G_t C_{t-1} G_t^\top + W_t}_{R_t} \right) \\
&\textit{One-step forecast}: && Y_t | D_{t-1} \sim \mathcal{N}_d \left(\underbrace{F_t^\top G_t m_{t-1}}_{f_t}, \underbrace{F_t^\top R_t F_t + V_t}_{Q_t} \right) \\
&\textit{Posterior}: && \theta_t | D_t \sim \mathcal{N}_p \left(m_t, C_t \right),
\end{aligned}$$

with $m_t = a_t + R_t F_t Q_t^{-1}(Y_t - f_t)$ and $C_t = R_t - R_t F_t Q_t^{-1} F_t^\top R_t$.

PROOF The theorem is proofed by induction on t. The basis step follows from (3.3) in Definition 3.2. Now assume it holds that,

$$\theta_{t-1} | D_{t-1} \sim \mathcal{N}_p \left(m_{t-1}, C_{t-1} \right).$$

Since all variables are Gaussian as will their sums. From the state equation, (3.5), we have, conditional on D_{t-1}, that

$$\theta_t = G_t \theta_{t-1} + \omega_t,$$

hence θ_t is Gaussian with mean $a_t = G_t m_{t-1}$, due to linearity of the mean operator, and variance matrix $R_t = G_t C_{t-1} G_t^\top + W_t$, due to independence between θ_{t-1} and ω_t conditional on D_{t-1}. Hence

$$\theta_t | D_{t-1} \sim \mathcal{N}_p\left(a_t\,,\, R_t\right). \tag{3.6}$$

Using (3.6) and the observation equation, (3.4), we have, conditional on D_{t-1}, that

$$Y_t = F_t^\top \theta_t + \nu_t,$$

hence, Y_t is Gaussian with mean $f_t = F_t^\top a_t$, and variance matrix $Q_t = F_t^\top R_t F_t + V_t$, due to independence between θ_t and ν_t conditional on D_{t-1}. Hence,

$$Y_t | D_{t-1} \sim \mathcal{N}_p\left(f_t\,,\, Q_t\right). \tag{3.7}$$

The covariance of θ_t and Y_t is

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left[\theta_t\,,\, Y_t \mid D_{t-1}\right] &= \mathbb{C}\mathrm{ov}\left[\theta_t\,,\, F_t^\top \theta_t + \nu_t \mid D_{t-1}\right] \\
&= \mathbb{C}\mathrm{ov}\left[\theta_t\,,\, F_t^\top \theta_t \mid D_{t-1}\right] + \mathbb{C}\mathrm{ov}\left[\theta_t\,,\, \nu_t \mid D_{t-1}\right] \\
&= \mathbb{V}\mathrm{ar}[\theta_t \mid D_{t-1}] F_t = R_t F_t.
\end{aligned}$$

From multivariate Gaussian theory, we have that

$$\begin{bmatrix}\theta_t \\ Y_t\end{bmatrix} | D_{t-1} \sim \mathcal{N}_{n+r}\left(\begin{bmatrix}a_t \\ f_t\end{bmatrix}, \begin{bmatrix}R_t & R_t F_t \\ F_t^\top R_t & Q_t\end{bmatrix}\right),$$

and conditional on Y_t, it holds, from Definition 3.1, that

$$\theta_t | D_t \sim \mathcal{N}_p\left(m_t\,,\, C_t\right), \tag{3.8}$$

hence, θ_t is Gaussian with mean $m_t = a_t + R_t F_t Q_t^{-1}(Y_t - f_t)$ and variance matrix $C_t = R_t - R_t F_t Q_t^{-1} F_t^\top R_t$, completing the proof. □

When using the Kalman filter, we obtain the mean vector and variance matrix based on the observations, which are updated for each new observation. The mean vector m_t is called the **filtered mean** and the variance matrix C_t is called the **filtered variance**. The mean vector f_t is called the **one-step forecast mean** and the variance matrix Q_t is the corresponding **one-step forecast variance**. Letting $e_t = Y_t - f_t$ denote the **one-step forecasting error**, the posterior mean of θ_t is a linear combination of the prior mean, a_t, and the one-step forecasting error, hence $m_t = a_t + A_t e_t$, where $A_t = R_t F_t Q_t^{-1}$ is called the **adaptive matrix**.

3.3 Disturbance Filtering

The disturbance filter is a mathematical equivalent to the Kalman filter and the outputs of the disturbance filter are the one-step forecast errors, e_t, the inverses

of the one-step forecast variances, Q_t^{-1}, and the so-called **scaled adaptive coefficient matrices**, K_t.
The disturbance filter is initialised by

$$a_1 = G_1 m_0, \quad \text{and} \quad R_1 = G_1 C_0 G_1^\top + W_1.$$

For $t = 1, \ldots, n$ the outputs are updated recursively by the equations

$$\begin{aligned}
e_t &= Y_t - F_t^\top a_t \\
Q_t &= F_t^\top R_t F_t + V_t \\
K_t &= G_{t+1} A_t \\
a_{t+1} &= G_{t+1} m_t \\
&= G_{t+1}(a_t + A_t e_t) \\
&= G_{t+1} a_t + K_t e_t \\
R_{t+1} &= G_{t+1} C_t G_{t+1}^\top W_{t+1} \\
&= G_{t+1}(R_t - A_t Q_t A_t^\top) G_{t+1}^\top + W_{t+1} \\
&= G_{t+1} R_t G_{t+1}^\top - K_t Q_t K_t^\top + W_{t+1} \\
&= G_{t+1} R_t (G_{t+1} - K_t F_t^\top)^\top + W_{t+1}.
\end{aligned}$$

Upon filtering, only e_t, Q_t^{-1} and K_t are stored, hence, dependent on the sizes of p and d, the disturbance filter uses less computer storage and may be faster than the Kalman filter (Dethlefsen, 2001).

3.4 Kalman Forecasting

Prediction of the latent process and forecasting of observations after observing Y_n, may be of interest. To simplify the notation we denote the time $n+k$ merely as k. Hence the distributions $\theta_k|D_n$ and $Y_k|D_n$, $k \in \mathbb{N}$, respectively, must be determined. This requires knowledge of G_k, F_k, V_k and W_k for all k of interest.

Theorem 3.2 (Kalman forecaster)
Let $Y_1, \ldots, Y_n$ be described by a Gaussian state space model. For each $k \in \mathbb{Z}_+$ the k-step forecasting distributions are

$$\begin{aligned}
\theta_k|D_n &\sim \mathcal{N}_p\left(\vec{m}_k\,,\, \vec{C}_k\right) \\
Y_k|D_n &\sim \mathcal{N}_d\left(\vec{f}_k\,,\, \vec{Q}_k\right),
\end{aligned}$$

where

$$\begin{aligned}\vec{m}_k &= G_k \vec{m}_{k-1} \\ \vec{C}_k &= G_k \vec{C}_{k-1} G_k^\top + W_k \\ \vec{f}_k &= F_k^\top \vec{m}_k \\ \vec{Q}_k &= F_k^\top \vec{C}_k F_k + V_k,\end{aligned}$$

with starting values $\vec{m}_0 = m_n$ *and* $\vec{C}_0 = C_n$.

PROOF The theorem is proofed by induction on k. The basis steps for both distributions follow from Theorem 3.1. Now assume that

$$\theta_{k-1}|D_n \sim \mathcal{N}_n\left(\vec{m}_{k-1}\,,\,\vec{C}_{k-1}\right),$$

and

$$Y_{k-1}|D_n \sim \mathcal{N}_d\left(\vec{f}_{k-1}\,,\,\vec{Q}_{k-1}\right).$$

From the state equation, we have that $\theta_k = G_k\theta_{k-1} + \omega_k$, hence, θ_k is Gaussian with mean $\vec{m}_k = G_k\vec{m}_{k-1}$, and variance matrix $\vec{C}_k = G_k\vec{C}_{k-1}G_k^\top + W_k$, since both θ_{k-1} and ω_k are Gaussian and independent conditional on D_n. Hence

$$\theta_k \sim \mathcal{N}_p\left(\vec{m}_k\,,\,\vec{C}_k\right).$$

From the observation equation, we have that $Y_k = F_k^\top\theta_k + \nu_k$. Since θ_k and ν_k are Gaussian, it follows, that Y_k is Gaussian with mean $\vec{f}_k = F_k^\top\vec{m}_k$ and variance matrix $\vec{Q}_k = F_k^\top\vec{C}_kF_k + V_k$, due to the independence between θ_k and ν_k conditional on D_n. Hence,

$$Y_k \sim \mathcal{N}_d\left(\vec{f}_k\,,\,\vec{Q}_k\right),$$

completing the proof. □

The mean vector, $\vec{m}_k$, is the k**-step forecasted mean** of θ_k and the variance matrix, $\vec{C}_k$, is the k**-step forecasted variance** of θ_k. Furthermore, the mean vector, $\vec{f}_k$, is the k**-step forecast mean** of Y_k and the variance matrix, $\vec{Q}_k$, is the k**-step forecast variance** of Y_k.

3.4.1 Implementation

The Kalman forecaster is implemented as a function in the package `sspir` in `R` (Dethlefsen and Lundbye-Christensen, 2006; R Development Core Team, 2008).

Description

Forecasted distributions of observations and predicted distributions of latent states in a state space model.

Usage

```
forecast(ss,k=10)
```

Arguments

`ss` an object of class `SS` or `ssm`.

`k` a positive integer giving the time for forecasting and prediction.

Details

Forecasting of observations and prediction of the latent process are performed according to the input state space model `ss` by estimating the distributions of the observation and latent process, respectively. The integer `k` defines the number of future forecasts and predictions to be estimated, hence having observed n observations, the forecasted observations and predicted latent states at times $n+1, \ldots, n+10$ (as default) is estimated.
The distribution of the predicted latent states are given by

$$\theta_k | D_n \sim N(\vec{m}_k, \vec{C}_k),$$

where $\vec{m}_k = G_k \vec{m}_{k-1}$ and $\vec{C}_k = G_k \vec{C}_{k-1} G_k^\top + W_k$ and the distribution of the forecasted observations are given by

$$Y_k | D_n \sim N(\vec{f}_k, \vec{Q}_k),$$

where $\vec{f}_k = F_k^\top \vec{m}_k$ and $\vec{Q}_k = F_k^\top \vec{C}_k F_k + V_k$.

Value

The returned value from either `smoother` (Gaussian case) or `extended`. An object containing the forecasted distributions of the observations specified by `forecast$f` and `forecast$Q` and the predicted distributions of the latent states specified by `forecast$m` and `forecast$C`.

3.5 Kalman Smoothing

The Kalman filter assess the latent process based on past observations. Suppose that all n observations are available, assessment of the latent process based on the entire information at time n, D_n, is the objective of smoothing. The Kalman smoother provides the mean, denoted $\tilde{m}_t$, and variance, denoted $\tilde{C}_t$, of θ_t conditional on D_n, $t = 1, \ldots, n$, which are referred to as the **smoothed mean** and the **smoothed variance**, respectively.

THEOREM 3.3 (KALMAN SMOOTHER)
Let $Y_1, \ldots, Y_n$ be described by a Gaussian state space model. For each $t = 1, \ldots, n$, the smoothed conditional distribution is

$$\theta_t | D_n \sim \mathcal{N}_p \left(\tilde{m}_t \, , \, \tilde{C}_t \right),$$

where

$$\begin{aligned} \tilde{m}_t &= m_t + B_t[\tilde{m}_{t+1} - a_{t+1}] \\ \tilde{C}_t &= C_t + B_t[\tilde{C}_{t+1} - R_{t+1}]B_t^\top \\ B_t &= C_t G_{t+1}^\top R_{t+1}^{-1}, \end{aligned}$$

with starting values $\tilde{m}_n = m_n$ and $\tilde{C}_n = C_n$.

PROOF The theorem if proofed by backwards induction on t, starting with $t = n$. The basis step follows from Theorem 3.1. Assume that

$$\theta_{t+1} | D_n \sim \mathcal{N}_p \left(\tilde{m}_{t+1} \, , \, \tilde{C}_{t+1} \right).$$

From the state- and observation equations, we have that

$$\begin{aligned} \theta_{t+1} | D_t &\sim \mathcal{N}_p \left(a_{t+1} \, , \, R_{t+1} \right) \\ \theta_t | D_t &\sim \mathcal{N}_p \left(m_t \, , \, C_t \right). \end{aligned}$$

The covariance matrix of θ_{t+1} and θ_t conditional on D_t is

$$\begin{aligned} \mathbb{C}\text{ov} \left[\theta_t \, , \, \theta_{t+1} \mid D_t \right] &= \mathbb{C}\text{ov} \left[\theta_t \, , \, G_{t+1}\theta_t + \omega_{t+1} \mid D_t \right] \\ &= \mathbb{V}\text{ar}[\theta_t \mid D_t] G_{t+1}^\top + \mathbb{C}\text{ov} \left[\theta_t \, , \, \omega_{t+1} \mid D_t \right] \\ &= C_t G_{t+1}^\top, \end{aligned}$$

due to independence between θ_t and ω_{t+1} conditional on D_t. Hence, we have that

$$\begin{bmatrix} \theta_t \\ \theta_{t+1} \end{bmatrix} | D_t \sim \mathcal{N}_{2p} \left(\begin{bmatrix} a_{t+1} \\ m_t \end{bmatrix}, \begin{bmatrix} R_{t+1} & C_t G_{t+1}^\top \\ G_{t+1} C_t & C_t \end{bmatrix} \right).$$

Conditional on θ_{t+1}, we obtain from multivariate Gaussian theory, that

$$\mathbb{E}[\theta_t \mid D_t, \theta_{t+1}] = m_t + C_t G_{t+1}^\top R_{t+1}^{-1} [\theta_{t+1} - a_{t+1}] = m_t + B_t[\theta_{t+1} - a_{t+1}],$$

and

$$\mathbb{V}\text{ar}[\theta_t \mid D_t, \theta_{t+1}] = C_t - C_t G_{t+1}^\top R_{t+1}^{-1} G_{t+1} C_t^\top = C_t - B_t R_{t+1} B_t^\top.$$

Hence,

$$\theta_t|D_t, \theta_{t+1} \sim \mathcal{N}_p\left(m_t + B_t[\theta_{t+1} - a_{t+1}]\,,\, C_t - B_t R_{t+1} B_t^\top\right). \tag{3.9}$$

Since $\theta_t \perp\!\!\!\perp D_n \setminus D_t | \theta_{t+1}$, the distribution of $\theta_t|D_t, \theta_{t+1}$ equals the distribution of $\theta_t|D_n, \theta_{t+1}$. We now obtain, that

$$\begin{aligned}
\mathbb{E}[\theta_t \mid D_n] &= \mathbb{E}[\mathbb{E}[\theta_t \mid D_n, \theta_{t+1}] \mid D_n] \\
&= \mathbb{E}[m_t + B_t[\theta_{t+1} - a_{t+1}] \mid D_n] \\
&= m_t + B_t[\tilde{m}_{t-1} - a_{t+1}],
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\theta_t \mid D_n] &= \mathbb{V}\mathrm{ar}[\mathbb{E}[\theta_t \mid D_n, \theta_{t+1}] \mid D_n] + \mathbb{E}[\mathbb{V}\mathrm{ar}[\theta_t \mid D_n, \theta_{t+1}] \mid D_n] \\
&= \mathbb{V}\mathrm{ar}[m_t + B_t[\theta_{t+1} - a_{t+1}] \mid D_n] \\
&\quad + \mathbb{E}\Big[C_t - B_t[\tilde{C}_{t+1} - R_{t+1}]B_t^\top \mid D_n\Big] \\
&= B_t[C_{t+1} - a_{t+1}]B_t^\top + C_t - B_t[\tilde{C}_{t+1} - R_{t+1}]B_t^\top \\
&= C_t - B_t[\tilde{C}_{t+1} - R_{t+1}]B_1^\top.
\end{aligned}$$

Hence,

$$\theta_t|D_n \sim \mathcal{N}_p\left(\tilde{m}_t\,,\, \tilde{C}_t\right),$$

completing the proof. □

The conditional means obtained from the Kalman smoother, denoted

$$\tilde{m}^\top = \begin{bmatrix}\tilde{m}_1^\top & \cdots & \tilde{m}_n^\top\end{bmatrix},$$

maximise the posterior, $p(\theta|Y)$, since mean equals mode when observations are Gaussian. From the definition of conditional densities, we have

$$p(\theta|Y) \propto p(\theta, Y),$$

and it follows that $\tilde{m}$ also maximises $p(\theta, Y)$. Consequently, $\tilde{m}$ also maximises

$$\begin{aligned}
\log(p(\theta, Y)) &= \log(p(Y|\theta)) + \log(p(\theta)) \\
&= \sum_{t=1}^{n} \log(p(Y_t|\theta_t)) + \sum_{t=1}^{n} \log(p(\theta_t|\theta_{t-1})) + \log(p(\theta_0)).
\end{aligned}$$

Differentiating with respect to θ_t and equating to zero, yields

$$\begin{aligned}
&\frac{\partial \log(p(Y_t, \theta_t))}{\partial \theta_t} \\
&= \frac{\partial \log(p(Y_t|\theta_t))}{\partial \theta_t} + \frac{\partial \log(p(\theta_t|\theta_{t-1}))}{\partial \theta_t} + \frac{\partial \log(p(\theta_{t+1}|\theta_t))}{\partial \theta_t} \mathbb{1}[t \neq n] = 0,
\end{aligned} \tag{3.10}$$

which are solved by $\tilde{m}$. Hence, we may interpret the Kalman smoother as an algorithm to solve equations like (3.10) efficiently. According to (3.4) and (3.5), the equations (3.10) gives

$$F_t V_t^{-1}(Y_t - \mu_t) - W_t^{-1}(\theta_t - G_t\theta_{t-1}) + G_{t+1}^\top W_{t+1}^{-1}(\theta_{t+1} - G_{t+1}\theta_t)\mathbb{1}[t \neq n] = 0. \tag{3.11}$$

3.6 Disturbance Smoothing

Using the outputs of the disturbance filter, e_t, Q_t^{-1}, and K_t, for $t = 1, \ldots, n$, we are able to estimate the disturbances, ω_t and ν_t, instead of focusing on the states (Koopman, 1993). Hence, we estimate $\tilde{\omega}_t = \mathbb{E}[\omega_t \mid D_n]$ and $\tilde{\nu}_t = \mathbb{E}[\nu_t \mid D_n]$ for $t = 1, \ldots, n$. Letting r_t be a $(p \times 1)$ vector and ϵ_t a $(d \times 1)$ vector, the smoothed disturbances are backwards recursively determined by the equations

$$\begin{aligned} r_n &= 0 \\ \epsilon_t &= Q_t^{-1} e_t - K_t^\top r_t \\ \tilde{\omega}_t &= W_t r_t \\ \tilde{\nu}_t &= V_t \epsilon_t \\ r_{t-1} &= F_t \epsilon_t + G_{t+1}^\top r_t. \end{aligned}$$

Notice that in comparison with the Kalman smoother no matrix inversion is needed, since the disturbance filter provides the inverted one-step forecast variances, Q_t. The smoothed means of the states are recursively determined by

$$\tilde{m}_t = G_t \tilde{m}_{t-1} + \tilde{\omega}_t,$$

initialised by $\tilde{m}_1 = G_1 m_0 + \tilde{\omega}_1$.

3.7 Estimation of Parameters

Until now we have assumed the matrices, F_t, G_t, V_t and W_t, being known for all t, and the dependency of the hyper parameters, ϕ, has been suppressed. This section describes two methods to obtain the maximum likelihood estimate of the hyper parameters.

3.7.1 Direct Maximum Likelihood Estimation

The likelihood function of the hyper parameters conditional on the observations is $L(\phi|Y) = p(Y|\phi)$, hence, we have the decomposition

$$\log(L(\phi|Y)) = \sum_{t=1}^{n} \log(p(Y_t|\phi, D_{t-1}))$$

of the log likelihood function due to the independence structure of the Gaussian state space model. From the one-step forecast distribution provided by the Kalman filter we have

$$Y_t|\phi, D_{t-1} \sim \mathcal{N}_d\left(f_t\,,\, Q_t\right).$$

Hence the log likelihood function is

$$\log(L(\phi|Y)) = -\frac{nd}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}(\log|Q_t| + (Y_t - f_t)^\top Q_t^{-1}(Y_t - f_t)). \tag{3.12}$$

For a given ϕ the log likelihood function can be obtained from the Kalman filter and (3.12) can be maximised numerically with respect to ϕ, providing a maximum likelihood estimate, denoted $\hat{\phi}$.

3.7.2 EM Algorithm

The Expectation-Maximisation (EM) algorithm is a two-step, iterative estimation algorithm used to estimate unknown parameters by maximum likelihood estimation (Dempster et al., 1977). Let Y be the observed Gaussian data with likelihood function $L(\psi|Y) = p(Y|\psi)$, which is dependent on the parameter vector $\psi \in \Psi \subseteq \mathbb{R}^p$ to be estimated. Furthermore, we extent data by some latent Gaussian data, denoted $Z \in \mathbb{R}^k$, so that the augmented data (Y, Z) have the joint likelihood function $L(\psi|Y, Z) = p(Y, Z|\psi)$. The likelihood function of ψ can be expressed as

$$L(\psi|Y) = \frac{p(Y, Z|\psi)}{p(Z|Y, \psi)}.$$

Maximisation of $L(\psi|Y)$ is equivalent to maximisation of the log likelihood function, i.e.

$$\log(L(\psi|Y)) = \log(p(Y, Z|\psi)) - \log(p(Z|Y, \psi)).$$

Letting $\psi^* \in \Psi$ be a temporary value of ψ, we have

$$\begin{aligned}
&\log(L(\psi|Y)) \\
&\quad = \int \log(p(Y, Z|\psi))p(Z|Y, \psi^*)\mathrm{d}Z - \int \log(p(Z|Y, \psi))p(Z|Y, \psi^*)\mathrm{d}Z \\
&\quad = Q(\psi, \psi^*) - H(\psi, \psi^*),
\end{aligned}$$

where

$$Q(\psi, \psi^*) = \int \log(p(Y, Z|\psi))p(Z|Y, \psi^*)\mathrm{d}Z = \mathbb{E}[\log(p(Y, Z|\psi)) \mid Y, \psi^*],$$

$$H(\psi, \psi^*) = \int \log(p(Z|Y, \psi))p(Z|Y, \psi^*)\mathrm{d}Z = \mathbb{E}[\log(p(Z|Y, \psi)) \mid Y, \psi^*].$$

Letting $\psi^{(m-1)}$ be the $(m-1)$th estimate of ψ, the two steps in the EM algorithm are

E-step : Calculate the conditional expectation

$$Q\left(\psi, \psi^{(m-1)}\right) = \mathbb{E}\Big[\log(p(Y, Z|\psi)) \mid Y, \psi^{(m-1)}\Big] \tag{3.13}$$

as a function of ψ.

M-step : Determine $\psi^{(m)}$ by maximise (3.13) with respect to ψ so

$$Q\left(\psi^{(m)}, \psi^{(m-1)}\right) \geq Q\left(\psi, \psi^{(m-1)}\right). \tag{3.14}$$

The algorithm alternates between these two steps until a predefined convergence criterion is reached. Notice, the algorithm only focuses on $Q(\cdot, \cdot)$ and not $H(\cdot, \cdot)$, since $H(\cdot, \cdot)$ is non-decreasing for each iteration, which is stated in the following theorem.

THEOREM 3.4
Let $\psi^{(m)}$ be the current estimate of ψ and choose $\psi^{(m+1)}$ according to the EM algorithm so that (3.14) *is fulfilled. Then*

$$\log\left(L\left(\psi^{(m+1)}|Y\right)\right) \geq \log\left(L\left(\psi^{(m)}|Y\right)\right),$$

with equality if and only if

$$Q\left(\psi^{(m+1)}, \psi^{(m)}\right) = Q\left(\psi^{(m)}, \psi^{(m)}\right)$$

and

$$H\left(\psi^{(m+1)}, \psi^{(m)}\right) = H\left(\psi^{(m)}, \psi^{(m)}\right).$$

PROOF Consider

$$\begin{aligned}
&\log\left(L\left(\psi^{(m+1)}|Y\right)\right) - \log\left(L\left(\psi^{(m)}|Y\right)\right) \\
&= \underbrace{Q\left(\psi^{(m+1)}, \psi^{(m)}\right) - Q\left(\psi^{(m)}, \psi^{(m)}\right)}_{(*)} + \underbrace{H\left(\psi^{(m+1)}, \psi^{(m)}\right) - H\left(\psi^{(m)}, \psi^{(m)}\right)}_{(**)}.
\end{aligned}$$

The choice of $\psi^{(m+1)}$ implies that $(*)$ is non-negative, hence $Q(\cdot, \cdot)$ is non-decreasing for each iteration. Since the logarithmic function is concave, we may

use Jensen's inequality, see A.5, and we have

$$\begin{aligned} H\left(\psi^{(m+1)}, \psi^{(m)}\right) - H\left(\psi^{(m)}, \psi^{(m)}\right) & \\ &= -\mathbb{E}\left[\log\left(\frac{p\left(Z|Y, \psi^{(m+1)}\right)}{p\left(Z|Y, \psi^{(m)}\right)}\right) \mid Y, \psi^{(m)}\right] \\ &\geq -\log\left(\mathbb{E}\left[\frac{p\left(Z|Y, \psi^{(m+1)}\right)}{p\left(Z|Y, \psi^{(m)}\right)} \mid Y, \psi^{(m)}\right]\right) \\ &= -\log\left(\int \frac{p\left(Z|Y, \psi^{(m+1)}\right)}{p\left(Z|Y, \psi^{(m)}\right)} p\left(Z|Y, \psi^{(m)}\right) \mathrm{d}Z\right) \\ &= 0. \end{aligned}$$

Hence, $(**)$ is non-negative, and thereby $H(\cdot, \cdot)$ is non-decreasing for each iteration, which completes the proof. □

In the following, the EM algorithm is applied to estimate the variance matrices $V_t = V$ and $W_t = W$, which are assumed being constant in a Gaussian state space model.
Assuming Y has been observed and θ is the latent data, the joint likelihood function of the augmented data (Y, θ) is $p(Y, \theta|V, W)$, which is Gaussian and the joint log likelihood function is

$$\log\left(p\left(Y, \theta|V, W\right)\right) \propto \sum_{t=1}^{n} \log\left(p\left(Y_t|\theta_t, V\right)\right) \tag{3.15}$$

$$+ \sum_{t=1}^{n} \log\left(p\left(\theta_t|\theta_{t-1}, W\right)\right), \tag{3.16}$$

since the initial values m_0 and C_0 are assumed being known.
The two steps in the EM algorithm may be formulated as maximising the expectation of the log likelihood function of the augmented data conditional of the observed data and the previous estimates of V and W, yielding new estimates for the next iteration. Letting $V^{(0)}$ and $W^{(0)}$ denote the initial estimates, the two steps in the mth iteration are

E-step : Calculate the conditional expectation

$$Q\left(V, W, V^{(m-1)}, W^{(m-1)}\right) = \mathbb{E}\Big[\log\left(p\left(Y, \theta|V, W\right)\right) \mid Y, V^{(m-1)}, W^{(m-1)}\Big] \tag{3.17}$$

as a function of V and W.

M-step : Maximise (3.17) with respect to V and W so

$$Q\left(V^{(m)}, W^{(m)}, V^{(m-1)}, W^{(m-1)}\right) \geq Q\left(V, W, V^{(m-1)}, W^{(m-1)}\right).$$

Hence, $V^{(m)}$ and $W^{(m)}$ are the mth estimates to be used in the E-step in the $(m+1)$th iteration.
Due to the decomposition of the joint likelihood function the estimation of V and W may be performed separately by using the EM algorithm for each of (3.15) and (3.16). Assume in the following that the smoothed mean, $\tilde{m}_t$, and variance matrix, $\tilde{C}_t$, are obtained from the Kalman smoother by replacing V and W with their current estimates, $V^{(m)}$ and $W^{(m)}$, respectively.

Estimation of V

Maximising the expectation of $\log\left(p\left(Y_t|\theta_t, V\right)\right)$ conditional on D_n is equivalent to minimising the expectation of

$$\log|V| + \frac{1}{n}\sum_{t=1}^{n}\|Y_t - \mu_t\|^2_{V^{-1}}$$

conditional on D_n. Using Theorem A.2, where $Z = \mu_t$ and $\epsilon = Y_t$, we have

$$\mathbb{E}\big[\|Y_t - \mu_t\|^2_{V^{-1}} \mid D_n\big] = \operatorname{trace}\left(V^{-1}F_t^\top \tilde{C}_t F_t\right) + \|Y_t - F_t^\top \tilde{m}_t\|^2_{V^{-1}}.$$

This yields

$$\begin{aligned}
&\mathbb{E}\left[\log|V| + \frac{1}{n}\sum_{t=1}^{n}\|Y_t - \mu_t\|^2_{V^{-1}} \mid D_n\right] \\
&= \log|V| + \frac{1}{n}\sum_{t=1}^{n}\left(\operatorname{trace}(V^{-1}F_t^\top \tilde{C}_t F_t) + \|Y_t - F_t^\top \tilde{m}_t\|^2_{V^{-1}}\right) \\
&= \log|V| + \operatorname{trace}\left[V^{-1}\frac{1}{n}\sum_{t=1}^{n}\left(F_t^\top \tilde{C}_t F_t + \left(Y_t - F_t^\top \tilde{m}_t\right)\left(Y_t - F_t^\top \tilde{m}_t\right)^\top\right)\right].
\end{aligned}$$

This is minimised for $V = V^{(m+1)}$ according to Theorem A.3, where

$$V^{(m+1)} = \frac{1}{n}\sum_{t=1}^{n}\left(F_t^\top \tilde{C}_t F_t + \left(Y_t - F_t^\top \tilde{m}_t\right)\left(Y_t - F_t^\top \tilde{m}_t\right)^\top\right). \tag{3.18}$$

Estimation of W

Similarly maximising the expectation of $\log(p(\theta_t|\theta_{t-1}, W))$ conditional on D_n is equivalent to minimising the expectation of

$$\log|W| + \frac{1}{n}\|\theta_t - G_t\theta_{t-1}\|^2_{W^{-1}}$$

conditional on D_n. First notice that

$$\begin{aligned}
&\mathbb{C}\mathrm{ov}\left[\theta_t\,,\,\theta_{t-1}\mid D_n\right]\\
&=\mathbb{E}[\mathbb{C}\mathrm{ov}\left[\theta_t\,,\,\theta_{t-1}\mid\theta_t,D_n\right]\mid D_n]+\mathbb{C}\mathrm{ov}\left[\mathbb{E}[\theta_t\mid\theta_t,D_n]\,,\,\mathbb{E}[\theta_{t-1}\mid\theta_t,D_n]\mid D_n\right]\\
&=\mathbb{C}\mathrm{ov}\left[\theta_t\,,\,\mathbb{E}[\theta_{t-1}\mid\theta_t,D_n]\mid D_n\right]\\
&=\mathbb{C}\mathrm{ov}\left[\theta_t\,,\,m_{t-1}+B_{t-1}(\theta_t-a_t)\mid D_n\right]\\
&=\mathbb{C}\mathrm{ov}\left[\theta_t\,,\,B_{t-1}\theta_t\mid D_n\right]\\
&=\tilde{C}_tB_{t-1}^\top,
\end{aligned}$$

where the third equality follows from (3.9) on page 62. Furthermore, we have

$$\mathbb{E}[\theta_t-G_t\theta_{t-1}\mid D_n]=\tilde{m}_t-G_t\tilde{m}_{t-1}$$

and

$$\mathbb{V}\mathrm{ar}[\theta_t-G_t\theta_{t-1}\mid D_n]=\tilde{C}_t+G_t\tilde{C}_{t-1}G_t^\top-\tilde{C}_tB_{t-1}^\top G_t^\top-G_tB_{t-1}\tilde{C}_t=L_t. \tag{3.19}$$

Using Theorem A.2, where $Z=\theta_t-G_t\theta_{t-1}$ and $\epsilon=0$, we have

$$\begin{aligned}
&\mathbb{E}\left[\log|W|+\frac{1}{n}\|\theta_t-G_t\theta_{t-1}\|^2_{W^{-1}}\mid D_n\right]\\
&=\log|W|+\frac{1}{n}\sum_{t=1}^n\left(\mathrm{trace}(W^{-1}L_t)+\|\tilde{m}_t-G_t\tilde{m}_{t-1}\|^2_{W^{-1}}\right)\\
&=\log|W|+\mathrm{trace}\left[W^{-1}\frac{1}{n}\sum_{t=1}^n\left(L_t+[\tilde{m}_t-G_t\tilde{m}_{t-1}]\,[\tilde{m}_t-G_t\tilde{m}_{t-1}]^\top\right)\right],
\end{aligned} \tag{3.20}$$

which is minimised for $W=W^{(m+1)}$, where

$$W^{(m+1)}=\frac{1}{n}\sum_{t=1}^n\left(L_t+[\tilde{m}_t-G_t\tilde{m}_{t-1}]\,[\tilde{m}_t-G_t\tilde{m}_{t-1}]^\top\right), \tag{3.21}$$

according to Theorem A.3.
Hence, with initial values $V^{(0)}$ and $W^{(0)}$, we apply the Kalman filter and smoother with current estimates $V^{(m)}$ and $W^{(m)}$ and calculate new estimates $V^{(m+1)}$ and $W^{(m+1)}$ according to (3.18) and (3.21) with the components provided by the Kalman smoother. This is repeated until convergence is reached.

Structure of Variance Matrices

Occasionally, the structure of the variance matrices, V and W, is known. This knowledge may be employed in the EM algorithm by parameterising the variances as functions of the hyper parameters, ϕ, hence $V=V(\phi)$ and $W=W(\phi)$.

Dependent on the structure of the variance matrices, the formulaes (3.18) and (3.21) are altered.
The following applies for both the observation variance matrix and the evolution variance matrix, letting

$$K_t = F_t^\top \tilde{C}_t F_t + \left(Y_t - F_t^\top \tilde{m}_t\right)\left(Y_t - F_t^\top \tilde{m}_t\right)^\top,$$

or

$$K_t = L_t + \left(\tilde{m}_t - G_t \tilde{m}_{t-1}\right)\left(\tilde{m}_t - G_t \tilde{m}_{t-1}\right)^\top,$$

respectively.
Assume that the components of the state vector, θ_t, are mutually independent, i.e. the evolution variance matrix becomes

$$W(\phi) = \phi I_p.$$

By substituting W with ϕI_p in (3.20), we have that the E-step is

$$\log|\phi I_p| + \frac{1}{n}\sum_{t=1}^{n} \operatorname{trace}((\phi I_p)^{-1} K_t) = d\log(\phi) + \frac{1}{\phi n}\sum_{t=1}^{n} \operatorname{trace}(K_t),$$

which is minimised for

$$\hat{\phi} = \frac{1}{dn}\sum_{t=1}^{n} \operatorname{trace}(K_t).$$

Hence, the M-step is

$$\hat{\phi}^{(m)} = \frac{1}{dn}\sum_{t=1}^{n} \operatorname{trace}(L_t + \left(\tilde{m}_t - G_t \tilde{m}_{t-1}\right)\left(\tilde{m}_t - G_t \tilde{m}_{t-1}\right)^\top).$$

This result implies that estimation of the evolution variance matrix may be performed as if no knowledge of the structure is available, and after the M-step, the average of the diagonal elements of the current estimate, is determined, which equals the estimate of $\hat{\phi}^{(m)}$, and the mth estimate of the evolution variance matrix is then $\hat{W}^{(m)} = \hat{\phi}^{(m)} I_p$.
Now assume we have a two dimensional state vector and the evolution variance matrix has the structure

$$W(\phi) = \phi W^* = \phi \begin{bmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}.$$

The E-step is

$$\log|\phi W^*| + \frac{1}{n}\sum_{t=1}^{n} \operatorname{trace}((\phi W^*)^{-1} K_t) = 2\log\left(\frac{\phi}{12}\right) + \frac{1}{\phi n}\sum_{t=1}^{n} \operatorname{trace}((W^*)^{-1} K_t),$$

which is maximised for

$$\hat{\phi} = \frac{1}{2n} \sum_{t=1}^{n} \text{trace}((W^*)^{-1} K_t).$$

Hence the M-step is

$$\hat{\phi}^{(m)} = \frac{1}{2n} \sum_{t=1}^{n} \text{trace} \left[(W^*)^{-1} \left(L_t + (\tilde{m}_t - G_t \tilde{m}_{t-1}) (\tilde{m}_t - G_t \tilde{m}_{t-1})^\top \right) \right].$$

Commonly, the evolution variance matrix is block diagonal, since the features of the model, e.g. secular trend and seasonal variation, are mutually independent. Assume the evolution transfer matrix, G_t, is block diagonal consisting of two independent features of the model, consequently the state vector and evolution variance matrix may be denoted

$$\theta_t = \begin{bmatrix} \theta_t^{(1)} \\ \theta_t^{(2)} \end{bmatrix}, \quad W = \begin{bmatrix} W^{(1)} & 0 \\ 0 & W^{(2)} \end{bmatrix},$$

respectively. The log likelihood function becomes

$$\log \left[p \left(\theta_t | \theta_{t-1}, W \right) \right] = \log \left[p \left(\theta_t^{(1)} | \theta_{t-1}^{(1)}, W^{(1)} \right) \right] + \log \left[p \left(\theta_t^{(2)} | \theta_{t-1}^{(2)}, W^{(2)} \right) \right],$$

hence the two block variance matrices may be estimated independently, and knowledge of the structure of each block may be exploited.

3.7.3 Implementation

The EM algorithm is implemented as a function in the package `sspir` in `R` and uses the functions `kfilter` and `smoother`, which implement Theorem 3.1 and 3.3, respectively (R Development Core Team, 2008).

Description

Estimates variance matrices of the observation- and latent process in a Gaussian state space model given as input, using the EM algorithm.

Usage

```
EMalgo(ss, maxiter = 50, epsilon = 1e-06, Vstruc=NA, Wstruc=NA)
```

Arguments

`ss` an object of class `SS`.

`maxiter` a positive integer giving the maximum number of iterations to run.

`epsilon` a (small) positive numeric giving the tolerance of the maximum relative differences of `Vmat` and `Wmat` between iterations.

`Vstruc` a function specifying the structure of the variance matrix of the observation model if such structure is known.

`Wstruc` a function specifying the structure of the variance matrix of the state model if such structure is known.

Details

The initial variance matrices are to be specified in the model specification and structures of the variance matrices may be specified by the user by the functions `Vstruc` and `Wstruc`. As default these are assigned `NA`, and if not specified, the variance matrices are not estimated, hence it is possible to estimate only the observation variance matrix or the evolution variance matrix assuming the other being known. The EM algorithm requires that the variance matrices to be estimated are constant, however if a variance matrix is not be estimated it may be non-constant.
The output provided by the function is the smoothed Gaussian model along with the estimated variance matrices, maximum values of the log likelihood function for each iteration and the number of iterations upon convergence.

Value

`ss` the value from `smoother`.

`Vmat.est` the estimate of the observation variance matrix, which is provided if the input of `Vstruc` is of class `function`, otherwise as input of `Vmat`.

`Wmat.est` the estimate of the observation variance matrix, which is provided if the input of `Wstruc` is of class `function`, otherwise as input of `Wmat`.

`loglik` maximum value of log likelihood function for each iteration.

`iteration` number of iterations upon convergence.

Chapter 4

Non-Gaussian State Space Models

This chapter provides the basic theory of non-Gaussian state space models, including partially specified-, exponential family- and general non-Gaussian state space models. The adjusted Pearson estimation algorithm is derived, and furthermore, assessment of the latent process using conjugate filtering and iterated extended Kalman smoothing are outlined. Finally, the chapter provides an example of assessing the latent process of a Poisson time series.
Initially, we note that, according to Bayes' Theorem a step in the Kalman filter may be described by

$$p(\theta_t|D_t) \propto p(\theta_t|D_{t-1})p(Y_t|\theta_t),$$

hence the Kalman filter may be considered as a Bayesian updating scheme (Lee, 2004). The prior distribution, $p(\theta_t|D_{t-1})$, contains all the prior information upon observing Y_t and is determined by the state equation, whereas the likelihood function, $p(Y_t|\theta_t)$, is determined by the observation equation. The posterior distribution, $p(\theta_t|D_t)$, is obtained by the prior distribution and the likelihood function.
In the following non-Gaussian state space models are considered, hence the observation model and the latent process may both be non-Gaussian. Several special cases are handled each with different relaxed distributional assumptions.

4.1 Partially Specified non-Gaussian State Space Models

Distributions may be specified only by the first and second order moments. This may be the case for both the observation- and the state model, which leads to

the following definition.

Definition 4.1 (Partially specified non-Gaussian state space model) *Let Y_t be a $(d \times 1)$ vector observed at times $t = 1, \ldots, n$. Y_t is described by a* ***partially specified non-Gaussian state space model*** *if given $\{F, G, W, V\}_t$ for each t, the relations between Y_t and a $(p \times 1)$ state vector, θ_t, at time t, as well as the relation between θ_t and the sequence of θ_t through time are determined by the conditional distributions specified by the first and second moments, given by*

$$\begin{aligned} \textit{Observation model} &: Y_t|\theta_t, D_{t-1} \sim [F_t^\top \theta_t, V_t] && (4.1)\\ \textit{State model} &: \theta_t|\theta_{t-1}, D_{t-1} \sim [G_t\theta_{t-1}, W_t] && (4.2)\\ \textit{Initial distribution} &: \theta_0|D_0 \sim [m_0, C_0], && (4.3) \end{aligned}$$

hence,

$$\begin{aligned} \textit{Observation equation} &: Y_t = F_t^\top \theta_t + \nu_t, \quad \nu_t \sim [0, V_t] && (4.4)\\ \textit{State equation} &: \theta_t = G_t\theta_{t-1} + \omega_t, \quad \omega_t \sim [0, W_t], && (4.5) \end{aligned}$$

where the sequences $\{\nu_t\}$ and $\{\omega_t\}$ are internally and mutually uncorrelated, and uncorrelated of θ_0 conditional on D_0. ◇

4.1.1 Linear Bayes' Estimator

In order to assess the latent process of a partially specified non-Gaussian state space model we introduce the **linear Bayes' estimator**. Suppose $\hat{\theta}$ is an estimate of the stochastic variable θ with support Ω for which $Y^\top = \begin{bmatrix} Y_1 & \cdots & Y_n \end{bmatrix}$, $Y_t \sim f(Y_t|\theta)$ are observed. Given a function L, for which the expectation,

$$\mathbb{E}\Big[L(\theta, \hat{\theta})\Big] = \int_\Omega L(\theta, \hat{\theta}) p(\theta) \mathrm{d}\theta$$

is to minimised, the function, L, is called a **loss function**. A quadratic loss function may be given by

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^\top (\theta - \hat{\theta}) = \operatorname{trace}\Big[(\theta - \hat{\theta})(\theta - \hat{\theta})^\top\Big].$$

The **overall risk function** is given by

$$r(\hat{\theta}) = \mathbb{E}\Big[L(\theta, \hat{\theta})\Big] = \operatorname{trace}\Big(\mathbb{E}\Big[(\theta - \hat{\theta})(\theta - \hat{\theta})^\top\Big]\Big).$$

An estimate, $\hat{\theta}^*(Y)$, of the linear form $\hat{\theta}^*(Y) = h + HY$, where h is an appropriate vector and H is an appropriate matrix, which fulfils

$$\mathbb{E}\Big[L(\theta, \hat{\theta}^*(Y))\Big] = \min_{\hat{\theta}} \mathbb{E}\Big[L(\theta, \hat{\theta})\Big],$$

is called a linear Bayes' estimate.

LEMMA 4.1
Let the distribution of $Y^\top = \begin{bmatrix} Y_1^\top & Y_2^\top \end{bmatrix}$ *be specified by its first and second moments,*

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \left[\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right].$$

Then the linear Bayes' estimate of Y_1 *is*

$$\hat{\theta}^*(Y_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2),$$

using a quadratic loss function. The covariance matrix of the estimate, called the ***risk matrix****, is*

$$RM = \mathbb{E}\Big[\Big[Y_1 - \hat{\theta}^*(Y_2)\Big]\Big[Y_1 - \hat{\theta}^*(Y_2)\Big]^\top\Big] = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

PROOF For $\hat{\theta} = h + HY_2$, define $R(\hat{\theta}) = \mathbb{E}\Big[(Y_1 - \hat{\theta})(Y_1 - \hat{\theta})^\top\Big]$. It follows that

$$\begin{aligned} R(\hat{\theta}) &= \mathbb{E}\big[Y_1 Y_1^\top\big] + \mathbb{E}\big[(h + HY_2)(h + HY_2)^\top\big] - \mathbb{E}\big[(h + HY_2)Y_1^\top\big] \\ &\quad - \mathbb{E}\big[Y_1(h + HY_2)^\top\big] \\ &= \Sigma_{11} + \mu_1\mu_1^\top + H\Sigma_{22}H^\top + (h + H\mu_2)(h + H\mu_2)^\top - H\Sigma_{21} \\ &\quad - \mu_1(h + H\mu_2)^\top - \Sigma_{12}H^\top - (h + H\mu_2)\mu_1^\top \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} + (H - \Sigma_{12}\Sigma_{22}^{-1})\Sigma_{22}(H - \Sigma_{12}\Sigma_{22}^{-1})^\top \\ &\quad + (\mu_1 - h - HY_2)(\mu_1 - h - HY_2)^\top. \end{aligned}$$

Hence, the overall risk function is a sum of three terms, the first term,

$$\operatorname{trace}\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

being independent of Y_1, the second term,

$$\operatorname{trace}\left[(H - \Sigma_{12}\Sigma_{22}^{-1})\Sigma_{22}(H - \Sigma_{12}\Sigma_{22}^{-1})^\top\right]$$

having a minimum value of zero, when $H = \Sigma_{12}\Sigma_{22}^{-1}$, and the third term,

$$\operatorname{trace}\left[(\mu_1 - h - HY_2)(\mu_1 - h - HY_2)^\top\right]$$

having a minimum value of zero, when $\mu_1 = h + H\mu_2$. Consequently, the risk function is minimised, when $H = \Sigma_{12}\Sigma_{22}^{-1}$ and $h = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2$, hence the linear Bayes' estimate is $\hat{\theta}^* = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2)$.
The risk matrix is

$$\begin{aligned} RM &= \mathbb{E}\Big[\big\{Y_1 - [\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2)]\big\}\big\{Y_1 - [\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2)]\big\}^\top\Big] \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\mathbb{C}\mathrm{ov}\left[Y_2\,,\,Y_1\right] - \mathbb{C}\mathrm{ov}\left[Y_1\,,\,Y_2\right]\Sigma_{22}^{-\top}\Sigma_{21} \\ &\quad + \Sigma_{12}\Sigma_{22}^{-1}\mathbb{V}\mathrm{ar}[Y_2](\Sigma_{22}^{-1})^\top\Sigma_{21} \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \end{aligned}$$

completing the proof. □

Using Lemma 4.1 we can derive the Kalman filter for a partially specified non-Gaussian state space model. When distributions are determined using Lemma 4.1, we denote it by $\hat{\sim}$ to emphasise the approximative nature of the distributions.

THEOREM 4.2
Let Y_t be described by a partially specified non-Gaussian state space model. For each t, the updating of the state vector is performed according to the following conditional distributions specified by the first and second moments,

$$\begin{aligned}
&\textit{Prior}: && \theta_t|D_{t-1} \hat{\sim} [\underbrace{G_t m_{t-1}}_{a_t}, \underbrace{G_t C_{t-1} G_t^\top + W_t}_{R_t}] \\
&\textit{One-step forecast}: && Y_t|D_{t-1} \hat{\sim} [\underbrace{F_t^\top G_t m_{t-1}}_{f_t}, \underbrace{F_t^\top (G_t C_{t-1} G_t^\top + W_t) F_t + V_t}_{Q_t}] \\
&\textit{Posterior}: && \theta_t|D_t \hat{\sim} [m_t, C_t],
\end{aligned}$$

with $m_t = a_t + R_t F_t Q_t^{-1}(Y_t - f_t)$ and $C_t = R_t - R_t F_t Q_t F_t^\top R_t^\top$.

PROOF The proof follows the proof of Theorem 3.1, using Lemma 4.1 as argument instead of results from multivariate Gaussian theory. □

The smoothing equations equal the corresponding equations of Gaussian state space models and are derived by using Lemma 4.1, where the components of the Kalman smoother are given by the Kalman filter.

4.2 Adjusted Pearson Algorithm

Assuming the variance matrices of the observation- and state models, V and W, are constant over time, however unknown, in model (4.1)-(4.3). As proposed by Jørgensen et al. (1996) estimation of the variance matrices may be performed by the **adjusted Pearson algorithm**, which is an iterative ad-hoc estimation algorithm. The vector $\theta_t - \tilde{m}_t$ is independent of the vector $Y_t - F_t^\top \tilde{m}_t$, hence by definition we have

$$\begin{aligned}
V &= \mathbb{V}\mathrm{ar}\big[Y_t - F_t^\top \theta_t\big] \\
&= \mathbb{V}\mathrm{ar}\big[Y_t - F_t^\top \tilde{m}_t - F_t^\top(\theta_t - \tilde{m}_t)\big] \\
&= \mathbb{V}\mathrm{ar}\big[Y_t - F_t^\top \tilde{m}_t\big] + F_t^\top \mathbb{V}\mathrm{ar}[(\theta_t - \tilde{m}_t)] F_t \\
&= \mathbb{V}\mathrm{ar}\big[Y_t - F_t^\top \tilde{m}_t\big] + F_t^\top \tilde{C}_t F_t \\
&= \mathbb{E}\big[(Y_t - F_t^\top \tilde{m}_t)(Y_t - F_t^\top \tilde{m}_t)^\top\big] + F_t^\top \tilde{C}_t F_t,
\end{aligned}$$

provided the model is correct, hence $\mathbb{E}\big[Y_t - F_t^\top \tilde{m}_t\big] = 0$. This derivation is only valid if the Kalman filter and smoother have been applied to the model

with correct variance matrices. In that case we may estimate the observation variance matrix by

$$\hat{V} = \underbrace{\frac{1}{n}\sum_{t=1}^{n}(Y_t - F_t^\top \tilde{m}_t)(Y_t - F_t^\top \tilde{m}_t)^\top}_{\text{Pearson estimate}} + \underbrace{\frac{1}{n}\sum_{t=1}^{n} F_t^\top \tilde{C}_t F_t}_{\text{Adjustment}}. \tag{4.6}$$

According to Jørgensen et al. (1996) the Pearson estimates are downwards biased due to substitution of the smoothed components, whereas the bias may be corrected for by the adjustment in (4.6).
Similarly, we have

$$\begin{aligned} W &= \mathbb{V}\mathrm{ar}[\theta_t - G_t\theta_{t-1}] \\ &= \mathbb{V}\mathrm{ar}[\tilde{m}_t - G_t\tilde{m}_{t-1} + \theta_t - \tilde{m}_t - G_t(\theta_{t-1} - \tilde{m}_{t-1})] \\ &= \mathbb{V}\mathrm{ar}[\tilde{m}_t - G_t\tilde{m}_{t-1}] + \mathbb{V}\mathrm{ar}[\theta_t - \tilde{m}_t - G_t(\theta_{t-1} - \tilde{m}_{t-1})] \\ &= \mathbb{E}\big[(\tilde{m}_t - G_t\tilde{m}_{t-1})(\tilde{m}_t - G_t\tilde{m}_{t-1})^\top\big] + L_t, \end{aligned}$$

where the matrix L_t is given by (3.19) on page 68. This derivation is only valid if the model is correct, hence, $\mathbb{E}[\tilde{m}_t - G_t\tilde{m}_{t-1}] = 0$. Assuming that the model is correctly specified, we may estimate the evolution variance matrix by

$$\hat{W} = \underbrace{\frac{1}{n}\sum_{t=1}^{n}(\tilde{m}_t - G_t\tilde{m}_{t-1})(\tilde{m}_t - G_t\tilde{m}_{t-1})^\top}_{\text{Pearson estimate}} + \underbrace{\frac{1}{n}\sum_{t=1}^{n} L_t}_{\text{Adjustment}}. \tag{4.7}$$

Hence, with initial values $V^{(0)}$ and $W^{(0)}$, we apply the Kalman filter and smoother with current estimates $V^{(m)}$ and $W^{(m)}$ and calculate new estimates $V^{(m+1)}$ and $W^{(m+1)}$ according to (4.6) and (4.7) with the components provided by the Kalman smoother. This is repeated until convergence is reached.
As noted by Dethlefsen et al. (1997), the adjusted Pearson algorithm is equivalent to the EM algorithm, provided the distributions of both the observation and state models are Gaussian, hence convergence result for the EM algorithm applies to the adjusted Pearson algorithm in this case. Otherwise no convergence results are shown for the adjusted Pearson algorithm (Klein, 2003).

4.3 Exponential Family State Space Models

Establishing the notation, we define the natural exponential families to be all distributions with log density functions of the form

$$\log(p(Y|\eta)) = Y^\top\eta - b(\eta) + c(Y),$$

where η is called the **natural parameter**, the function b is convex and twice differentiable and c is a suitable function only dependent on the observations.

From the theory of exponential families we have that

$$\mathbb{E}[Y \mid \eta] = \mu = \frac{\partial b(\eta)}{\partial \eta} = b'(\eta) = \tau(\eta)$$
$$\mathbb{V}\mathrm{ar}[Y \mid \eta] = \Sigma = \frac{\partial^2 b(\eta)}{\partial \eta \partial \eta^\top} = b''(\eta),$$

where the function τ is called the **mean value mapping**.
The natural exponential families are special cases of the generalised linear models, which are specified by three steps, i.e. the random- and systematic components and the link between these two components. Given a sample of n independent observations Y_t with corresponding explanatory variables, x_t, the **random component** is that each Y_t belonging to the same exponential family with possibly differing natural parameters. The **systematic component** is the **linear predictor**, which specifies the relation between the explanatory variables and the observations through

$$\lambda_t = F_t^\top \theta_t,$$

where $F_t^\top$ is the $(d \times p)$ **design matrix** consisting of known function of the explanatory variables and θ is the $(p \times 1)$ state vector. As usual we refer to λ_t as the signal. The link between the random- and systematic component is determined by the relation between the mean, $\mu_t = \tau(\eta_t)$, and the linear predictor, λ_t, and is specified by a **response function**

$$\mu_t = b'(\eta_t) = h(\lambda_t).$$

The inverse of the response function is the **link function**

$$g(\mu_t) = \lambda_t.$$

The relation between the linear predictor and the natural parameter is determined by a function

$$v(\lambda_t) = \eta_t = \tau^{-1}(h(\lambda_t)).$$

If $g = \tau^{-1}$, we say that g is a canonical link, since then $\eta_t = \lambda_t$. A canonical link implies that $h = \tau$.
Now suppose the observation model is a natural exponential family and let the latent process be specified only by its first and second moments. The natural **exponential family state space model** of a one dimensional observation process $\{Y_t\}$ with partially specified state vector may be described by

$$p(Y_t|\eta_t) = \exp\left(Y_t\eta_t - b(\eta_t) + c(Y_t)\right) \tag{4.8}$$
$$\eta_t = v(\lambda_t) = v(F_t^\top \theta_t) \tag{4.9}$$
$$\theta_t = G_t\theta_{t-1} + \omega_t, \quad \omega_t \sim [0, W_t] \tag{4.10}$$
$$\theta_0 \sim [m_0, C_0]. \tag{4.11}$$

The same conditional independence structure from Gaussian state space models is assumed and the process $\{\omega_t\}$ is assumed being serially uncorrelated. Assessment of the latent process is performed using the following theorem.

THEOREM 4.3 (CONJUGATE FILTERING)
Let Y_t be described by the model in (4.8)-(4.11)*. For each t, the updating of the state vector is performed according to the following approximated conditional distributions specified by the first and second moments,*

$$\begin{aligned} &\textit{Prior}: && \theta_t|D_{t-1} \sim [\underbrace{G_t m_{t-1}}_{a_t}, \underbrace{G_t C_{t-1} G_t^\top + W_t}_{R_t}] \\ &\textit{Posterior}: && \theta_t|D_t \tilde{\sim} [m_t, C_t] \end{aligned}$$

with $m_t = a_t + R_t F_t (f_t^* - f_t)/q_t$ *and* $C_t = R_t - \left(1 - \frac{q_t^*}{q_t}\right) R_t F_t F_t^\top R_r / q_t$.

PROOF The theorem is proofed by induction on t. The basis step follows from (4.11). Assume that

$$\theta_{t-1}|D_{t-1} \tilde{\sim} [m_{t-1}, C_{t-1}]. \tag{4.12}$$

We have from (4.10), that the prior distribution of θ_t conditional on D_{t-1} is

$$\theta_t|D_{t-1} \sim [\underbrace{G_t m_{t-1}}_{a_t}, \underbrace{G_t C_{t-1} G_t^\top + W_t}_{R_t}].$$

This leads to the prior distribution of λ_t conditional on D_{t-1} using (4.9),

$$\lambda_t|D_{t-1} \sim [\underbrace{F_t^\top a_t}_{f_t}, \underbrace{F_t^\top R_t F_t}_{q_t}].$$

The covariance matrix of θ_t and λ_t conditional on D_{t-1} is

$$\begin{aligned} \mathbb{C}\text{ov}\left[\theta_t\,,\,\lambda_t \mid D_{t-1}\right] &= \mathbb{C}\text{ov}\left[\theta_t\,,\,F_t^\top \theta_t \mid D_{t-1}\right] \\ &= \mathbb{V}\text{ar}[\theta_t \mid D_{t-1}] F_t \\ &= R_t F_t. \end{aligned}$$

Hence, the joint distribution of θ_t and λ_t conditional on D_{t-1} is

$$\begin{bmatrix} \theta_t \\ \lambda_t \end{bmatrix} |D_{t-1} \sim \left[\begin{bmatrix} a_t \\ f_t \end{bmatrix}, \begin{bmatrix} R_t & R_t F_t \\ F_t^\top R_t & q_t \end{bmatrix} \right].$$

From Lemma 4.1 the approximated distribution of θ_t conditional on λ_t and D_{t-1} is specified as

$$\theta_t|\lambda_t, D_{t-1} \tilde{\sim} \left[\frac{a_t + R_t F_t (\lambda_t - f_t)}{q_t}, \frac{R_t - R_t F_t F_t^\top R_t}{q_t} \right].$$

We choose a conjugate prior on the natural parameter η_t, hence

$$p(\eta_t|D_{t-1}) = c(r_t, s_t) \exp\left(r_t\eta_t - s_t b(\eta_t)\right), \tag{4.13}$$

where r_t and s_t are given so $\mathbb{E}[g(\tau(\eta_t)) \mid D_{t-1}] = f_t$ and $\mathbb{V}\mathrm{ar}[g(\tau(\eta_t)) \mid D_{t-1}] = q_t$ and $c(r_t, s_t)$ is the normalising constant ensuring $p(\eta_t|D_{t-1})$ being a density (Lee, 2004). It follows that the one-step forecast of Y_t is

$$\begin{aligned}
p(Y_t|D_{t-1}) &= \int p(Y_t, \eta_t|D_{t-1})\mathrm{d}\eta_t \\
&= \int p(Y_t|\eta_t, D_{t-1})p(\eta_t|D_{t-1})\mathrm{d}\eta_t \\
&= c(r_t, s_t) \int \exp\left[(\underbrace{Y_t + r_t}_{r_t^*})\eta_t - (\underbrace{1 + s_t}_{s_t^*})b(\eta_t)\right]\mathrm{d}\eta_t \\
&= \frac{c(r_t, s_t)}{c(r_t^*, s_t^*)}.
\end{aligned}$$

This leads to the posterior of η_t conditional on D_t by

$$\begin{aligned}
p(\eta_t|D_t) &= p(\eta_t|D_{t-1})p(Y_t|\eta_t, D_{t-1})\frac{1}{p(Y_t|D_{t-1})} \\
&= p(\eta_t|D_{t-1})\frac{\mathrm{d}p(Y_t|D_{t-1})}{\mathrm{d}\eta_t}\frac{1}{p(\eta_t|D_{t-1})p(Y_t|D_{t-1})} \\
&= c(r_t^*, s_t^*)\exp(r_t^*\eta_t - s_t^* b(\eta_t)).
\end{aligned} \tag{4.14}$$

We see that the updating scheme of the natural parameter, η_t, is a conjugate updating scheme, since the prior and posterior distribution belong to the same family.
Hence, the posterior distribution of λ_t conditional on D_t is

$$\lambda_t|D_t \sim \left[\underbrace{\mathbb{E}[g(\tau(\eta_t)) \mid D_t]}_{f_t^*}, \underbrace{\mathbb{V}\mathrm{ar}[g(\tau(\eta_t)) \mid D_t]}_{q_t^*}\right].$$

The posterior distribution of θ_t conditional on D_t is obtained by the joint distribution of λ_t and θ_t conditional on D_t, which is, using Bayes' Theorem,

$$\begin{aligned}
p(\lambda_t, \theta_t|D_t) &= p(\lambda_t, \theta_t|D_{t-1})p(Y_t|\lambda_t, D_{t-1})p(Y_t|D_{t-1}) \\
&\propto p(\lambda_t, \theta_t|D_{t-1})p(Y_t|\lambda_t) \\
&= p(\theta_t|\lambda_t, D_{t-1})p(\lambda_t|D_{t-1})p(Y_t|\lambda_t) \\
&\propto p(\theta_t|\lambda_t, D_{t-1})p(\lambda_t|D_t).
\end{aligned}$$

Hence,

$$p(\theta_t|D_t) = \int p(\theta_t|\lambda_t, D_{t-1})p(\lambda_t|D_{t-1})\mathrm{d}\eta_t. \tag{4.15}$$

The first and second moments of θ_t conditional on D_t are

$$m_t = \mathbb{E}[\theta_t \mid D_t] = \mathbb{E}[\mathbb{E}[\theta_t \mid \lambda_t, D_{t-1}] \mid D_t] = a_t + R_t F_t (f_t^* - f_t)/q_t \tag{4.16}$$

$$\begin{aligned} C_t &= \mathbb{V}\mathrm{ar}[\theta_t \mid D_t] = \mathbb{V}\mathrm{ar}[\mathbb{E}[\theta_t \mid \lambda_t, D_{t-1}] \mid D_t] + \mathbb{E}[\mathbb{V}\mathrm{ar}[\theta_t \mid \lambda_t, D_{t-1}] \mid D_t] \\ &= R_t - \left(1 - \frac{q_t^*}{q_t}\right) R_t F_t F_t^\top R_t / q_t. \end{aligned} \tag{4.17}$$

Hence

$$\theta_t | D_t \tilde{\sim} [m_t, C_t],$$

completing the proof. □

The smoothing equations equal the corresponding equations of Gaussian state space models and are derived by using Lemma 4.1, where the components of the Kalman smoother are given by the conjugate filter.
Notice, letting $x_t = r_t/s_t$ the prior of the natural parameter, η_t, can be written as

$$p(\eta_t | D_{t-1}) = c(r_t, s_t) \exp[s_t(x_t \eta_t - b(\eta_t))].$$

Since b is convex, the prior is unimodal with mode $x_t = b'(\eta_t)$, hence x_t is a location parameter and s_t is the precision. The posterior can be rewritten as

$$p(\eta_t | D_t) = c(r_t^*, s_t^*) \exp[s_t^*(x_t^* \eta_t - b(\eta_t))],$$

where

$$r_t^* = r_t + Y_t \tag{4.18}$$

$$s_t^* = s_t + 1 \tag{4.19}$$

$$x_t^* = \frac{r_t + Y_t}{s_t + 1} = \frac{s_t x_t + Y_t}{s_t + 1} = \frac{s_t}{s_t + 1} x_t + \left(1 - \frac{s_t}{s_t + 1}\right) Y_t.$$

Hence, upon observing Y_t, we obtain a gain in the posterior precision and we see that the posterior location is a weighted average between the prior location parameter and the observation. Given a high prior precision, the observation is weighted low in the posterior precision and vice versa. Using a conjugate updating scheme simplifies the updating, since only the posterior parameters, (4.18) and (4.19) are to be updated and the prior determines the posterior distribution. Notice, that the conjugate prior is specific according to the sampling distribution.
As proposed by West et al. (1985) the parameters of the conjugate prior, r_t and s_t, may be chosen such that f_t and q_t equal other quantities than the first and second moments of the natural parameter, which may provide more convenient values. They propose f_t to equal the mode and q_t^{-1} to equal the curvature at the mode. By this West et al. emphasise that the link between $\tau(\eta_t)$ and λ_t is merely a guide to form the prior of η_t (West et al., 1985).

4.3.1 Conjugate Filtering for Poisson Time Series

To illustrate the conjugate filtering, assume we have a time series of serially correlated count data. Choosing the canonical link for the Poisson distribution, the log link, we have

$$\lambda_t = \eta_t = \log(\mu_t) = F_t^\top \theta_t.$$

The conjugate prior for Poisson distributed observations, Y_t, with intensities, μ_t, is the gamma distribution,

$$p(\mu_t|\alpha_t, \beta_t) = \frac{\beta_t^{\alpha_t} \mu_t^{\alpha_t - 1} \exp(-\beta_t \mu_t)}{\Gamma(\alpha_t)}$$

(DeGroot, 1989). Hence, the conjugate prior of the natural parameter, $\eta_t = \log(\mu_t)$, is

$$p(\eta_t|\alpha_t, \beta_t) = \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \exp(\alpha_t \eta_t - \beta_t \exp(\eta_t)).$$

According to (4.13) we have $r_t = \alpha_t$ and $s_t = \beta_t$. Notice, that the moment generating function of η_t, denoted $M(s)$, is

$$\begin{aligned} M(s) &= \int \exp(s\eta_t) p(\eta_t|\alpha_t, \beta_t) \mathrm{d}\eta_t \\ &= \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \int \exp[(\alpha_t + s)\eta_t - \beta_t \exp(\eta_t)] \mathrm{d}\eta_t \\ &= \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \frac{\Gamma(\alpha_t + s)}{\beta_t^{\alpha_t + s}} = \beta_t^{-s} \frac{\Gamma(\alpha_t + s)}{\Gamma(\alpha_t)}, \end{aligned}$$

hence, the mean of the prior is

$$f_t = M'(s)|_{s=0} = \frac{\Gamma'(\alpha_t)}{\Gamma(\alpha_t)} - \log(\beta_t),$$

and the variance is

$$q_t = M''(s)|_{s=0} - (M'(s)|_{s=0})^2 = \left(\frac{\Gamma'(\alpha_t)}{\Gamma(\alpha_t)} \right)',$$

which both may be solved numerically to obtain r_t and s_t (West et al., 1985; DeGroot, 1989).

Upon observing Y_t, it follows that the posterior of the natural parameter is of the form (4.14) with parameters $r^* = r_t + Y_t$ and $s^* = s_t + 1$, hence we have

$$f_t^* = \frac{\Gamma'(\alpha_t + Y_t)}{\Gamma(\alpha_t + Y_t)} - \log(\beta_t + 1),$$

and

$$q_t^* = \left(\frac{\Gamma'(\alpha_t + Y_t)}{\Gamma(\alpha_t + Y_t)} \right)'.$$

By insertion in (4.16) and (4.17), the filtered moments, m_t and C_t, are obtained, and smoothing may be performed.

4.4 General non-Gaussian State Space Models

This section generalises the exponential family state space models with partially specified state vector, by not restricting the observation model to be a natural exponential family. The state model is modelled as a first order Markov process and is in general non-Gaussian, and furthermore, the process $\{\omega_t\}$ is serially uncorrelated and uncorrelated of the observations. Formally, the general non-Gaussian state space model is denoted

$$\text{Observation model : } p(Y_t|\lambda_t) \tag{4.20}$$

$$\text{State model : } \theta_t = G_t\theta_{t-1} + \omega_t, \quad \omega_t \sim p(\omega_t), \tag{4.21}$$

where $\lambda_t = F_t^\top \theta_t$ is the signal. We aim to maximise the posterior, $p(\theta|Y)$, with respect to θ, which is equivalent to maximise the joint density, $p(Y, \theta)$, hence, we must solve the equations

$$\frac{\partial \log(p(Y,\theta))}{\partial \theta_t} = \frac{\partial \log(p(Y_t|\theta_t))}{\partial \theta_t} + \frac{\partial \log(p(\theta_t|\theta_{t-1}))}{\partial \theta_t} + \frac{\partial \log(p(\theta_{t+1}|\theta_t))}{\partial \theta_t}\mathbb{1}[t \neq n] = 0.$$

The first term is determined by

$$\frac{\partial \log(p(Y_t|\theta_t))}{\partial \theta_t} = \left(\frac{\partial \log(p(Y_t|\lambda_t))}{\partial \lambda_t}\right)^\top \frac{\partial \lambda_t}{\partial \theta_t} = F_t \frac{\partial \log(p(Y_t|\lambda_t))}{\partial \lambda_t}. \tag{4.22}$$

From the definition of the distribution function we have

$$\begin{aligned} F_{\theta_t|\theta_{t-1}}(x) &= P(\theta_t < x|\theta_{t-1}) \\ &= P(\theta_t - G_t\theta_{t-1} < x - G_t\theta_{t-1}) \\ &= P(\omega_t < x - G_t\theta_{t-1}) = F_{\omega_t}(x - G_t\theta_{t-1}), \end{aligned}$$

hence $p_{\theta_t|\theta_{t-1}}(\theta_t) = p_{\omega_t}(\theta_t - G_t\theta_{t-1}) = p(\omega_t)$. This yields

$$\frac{\partial \log(p(\theta_t|\theta_{t-1}))}{\partial \theta_t} = \left(\frac{\partial \log(p(\omega_t))}{\partial \omega_t}\right)^\top \frac{\partial \omega_t}{\partial \theta_t} = \frac{\partial \log(p(\omega_t))}{\partial \omega_t}. \tag{4.23}$$

The equations to solve become

$$\frac{\partial \log(p(Y,\theta))}{\partial \theta_t} = F_t \frac{\partial \log(p(Y_t|\lambda_t))}{\partial \lambda_t} + \frac{\partial \log(p(\omega_t))}{\partial \omega_t} - G_{t+1}^\top \frac{\partial \log(p(\omega_{t+1}))}{\partial \omega_{t+1}}\mathbb{1}[t \neq n] = 0. \tag{4.24}$$

In order to solve the equations, the strategy is to obtain an approximation of the state space model by linearising the non-linear terms and identifying $\tilde{Y}_t$, $\tilde{V}_t$ and $\tilde{W}_t$ by comparing (4.24) with (3.11) on page 63. We then obtain a linearised approximated Gaussian state space model specified by $\tilde{Y}_t$, $\tilde{V}_t$ and $\tilde{W}_t$ with the property that the posterior, $p(\theta|\tilde{Y})$, of this model has the same mode as the posterior, $p(\theta|Y)$, of the state space model, (4.20)-(4.21).

4.5 Iterated Extended Kalman Smoothing

Assuming the second derivative of $\log(p(Y_t|\lambda_t))$ is positive definite, we may linearise the first term of (4.24), the observation term, with a first order Taylor expansion about an initial value of θ_t, denoted $\tilde{\theta}_t$, hence, $\tilde{\lambda}_t = F_t^\top \tilde{\theta}_t$. The first order Taylor expansion around $\tilde{\lambda_t}$ is given by

$$\frac{\partial \log(p(Y_t|\lambda_t))}{\partial \lambda_t} \approx \frac{\partial \log(p(Y_t|\lambda_t))}{\partial \lambda_t}\bigg|_{\lambda_t=\tilde{\lambda}_t} + \frac{\partial^2 \log(p(Y_t|\lambda_t))}{\partial \lambda_t \partial \lambda_t^\top}\bigg|_{\lambda_t=\tilde{\lambda}_t} (\lambda_t - \tilde{\lambda}_t). \quad (4.25)$$

Letting,

$$\tilde{V}_t^{-1} = -\frac{\partial^2 \log(p(Y_t|\lambda_t))}{\partial \lambda_t \partial \lambda_t^\top}\bigg|_{\lambda_t=\tilde{\lambda}_t},$$

and

$$\tilde{Y}_t = \tilde{\lambda}_t + \tilde{V}_t \frac{\partial \log(p(Y_t|\lambda_t))}{\partial \lambda_t}\bigg|_{\lambda_t=\tilde{\lambda}_t},$$

we have

$$\frac{\partial \log(p(Y_t|\lambda_t))}{\partial \lambda_t} \approx \tilde{V}_t^{-1}(\tilde{Y}_t - \lambda_t), \quad (4.26)$$

which is of the same form as the first term of (3.11) on page 63. However, if the second derivative of $\log(p(Y_t|\lambda_t))$ is not positive definite, the following method may be used.

Assume, that the densities $p(Y_t|\lambda_t)$ and $p(\omega_t)$ both are functions of a quadratic form of the arguments, denoted $(Y_t - \lambda_t)^2$ and $(\omega_t)^2$, respectively. Notice that

$$\frac{\partial \log(p(Y_t|\lambda_t))}{\partial \lambda_t} = \frac{\partial \log(p(Y_t|\lambda_t))}{\partial (Y_t - \lambda_t)^2} \frac{\partial (Y_t - \lambda_t)^2}{\partial \lambda_t} = -2\frac{\partial \log(p(Y_t|\lambda_t))}{\partial (Y_t - \lambda_t)^2}(Y_t - \lambda_t), \quad (4.27)$$

and

$$\frac{\partial \log(p(\omega_t))}{\partial \omega_t} = \frac{\partial \log(p(\omega_t))}{\partial (\omega_t)^2} \frac{\partial (\omega_t)^2}{\partial \omega_t} = 2\frac{\partial \log(p(\omega_t))}{\partial (\omega_t)^2}(\theta_t - G_t\theta_{t-1}). \quad (4.28)$$

Hence, (4.24) is approximated by

$$\begin{aligned}\frac{\partial \log(p(Y,\theta))}{\partial \theta_t} \approx & -2F_t \frac{\partial \log(p(Y_t|\lambda_t))}{\partial (Y_t - \lambda_t)^2}(Y_t - \lambda_t) + 2\frac{\partial \log(p(\omega_t))}{\partial (\omega_t)^2}(\theta_t - G_t\theta_{t-1}) \\ & - 2G_{t+1}^\top \frac{\partial \log(p(\omega_{t+1}))}{\partial (\omega_{t+1})^2}(\theta_{t+1} - G_{t+1}\theta_t)\mathbb{1}[t \neq n].\end{aligned}$$

By comparison with (3.11) we have

$$\tilde{V}_t^{-1} = -2\frac{\partial \log(p(Y_t|\lambda_t))}{\partial (Y_t - \lambda_t)^2}\bigg|_{\lambda_t=\tilde{\lambda}_t},$$

and

$$\tilde{W}_t^{-1} = -2\frac{\partial \log(p(\omega_t))}{\partial (\omega_t)^2}\bigg|_{\omega_t=\tilde{\omega}_t},$$

where $\tilde{\omega}_t = \tilde{\theta}_t - G_t\tilde{\theta}_{t-1}$.
Letting $\tilde{\lambda}_t^{(0)} = F_t^\top a_t$ and apply the Kalman filter and smoother on the approximated model specified by $\tilde{Y}_t$, $\tilde{V}_t$ and $\tilde{W}_t$ to obtain $\tilde{m}_t^{(0)}$, we have recursively $\tilde{\lambda}_t^{(m)} = F_t^\top \tilde{m}_t^{(m-1)}$. Hence, the approximated Gaussian state space model is iteratively improved and upon convergence this has the same mode as the non-Gaussian state space model, (4.20)-(4.21). A convergence criterion, proposed by Dethlefsen (2001), is, when all values fulfill

$$\left(\tilde{Y}_t^{(m-1)}\right)^{-1}\left|\left(\tilde{Y}_t^{(m)} - \tilde{Y}_t^{(m-1)}\right)\right| < \epsilon,$$

$$\left(\tilde{V}_t^{(m-1)}\right)^{-1}\left|\left(\tilde{V}_t^{(m)} - \tilde{V}_t^{(m-1)}\right)\right| < \epsilon,$$

and

$$\left(\tilde{W}_t^{(m-1)}\right)^{-1}\left|\left(\tilde{W}_t^{(m)} - \tilde{W}_t^{(m-1)}\right)\right| < \epsilon,$$

for $t = 1, \ldots, n$, where ϵ is a small positive number. Furthermore, if the number of iterations exceeds a predefined number, before the convergence criterion is reached, the algorithm may be stopped and the final estimates of $\tilde{Y}_t$, $\tilde{V}_t$ and $\tilde{W}_t$ specify the approximating model. This procedure is called **iterated extended Kalman smoothing**.

4.5.1 Example

Assume we have a time series of serially correlated count data. In case of canonical link function we have

$$p(Y_t|\lambda_t) = \exp(Y_t^\top \lambda_t - b(\lambda_t) + c(Y_t))$$
$$\lambda_t = F_t^\top \theta_t = \log(\mu_t).$$

Furthermore, assume the state process is first order Markov with t_r-distributed evolution error, i.e.

$$\theta_t = G_t\theta_{t-1} + \omega_t$$
$$p(\omega_t) = \frac{\Gamma(\frac{r+1}{2})}{\sqrt{r\pi}\Gamma(\frac{r}{2})}\left(1 + \frac{\omega_t^2}{r}\right)^{-\frac{r+1}{2}}.$$

We linearise the observation model according to (4.25) and we obtain

$$\begin{aligned}\tilde{V}_t^{-1} &= b''(\tilde{\lambda}_t) = \exp(F_t^\top \tilde{\theta}_t)\\ \tilde{Y}_t &= \tilde{\lambda}_t - b''(\tilde{\lambda}_t)\left[Y_t - b'(\tilde{\lambda}_t)\right]\\ &= F_t^\top \tilde{\theta}_t - \exp(F_t^\top \tilde{\theta}_t)\left[Y_t - \exp(F_t^\top \tilde{\theta}_t)\right].\end{aligned}$$

The state model is linearised according to (4.28) and we obtain

$$\tilde{W}_t = \frac{r + \tilde{\omega}_t^2}{r+1}.$$

Let $\tilde{\lambda}_t^{(0)} = F_t^\top a_t$ and $\tilde{\omega}_t^{(0)} = a_t - G_t a_{t-1}$. Apply the Kalman filter stated in Theorem 4.2 and Kalman smoother on the approximated Gaussian state space model

$$\begin{aligned}\text{Observation model } &: Y_t|\theta_t \sim \left[\exp\left(\tilde{\lambda}_t^{(0)}\right), \exp\left(\tilde{\lambda}_t^{(0)}\right)\right],\\ \text{State model } &: \theta_t|\theta_{t-1} \sim \left[G_t a_{t-1}, \frac{r + \left(\tilde{\omega}_t^{(0)}\right)^2}{r+1}\right],\end{aligned}$$

to obtain $\tilde{m}_t^{(0)}$. Now $\tilde{\lambda}_t^{(0)}$ and $\tilde{\omega}_t^{(0)}$ are updated, hence $\tilde{\lambda}_t^{(1)} = F_t^\top \tilde{m}_t^{(0)}$ and $\tilde{\omega}_t^{(1)} = \tilde{m}_t^{(0)} - G_t \tilde{m}_{t-1}^{(0)}$. Applying the Kalman filter and the Kalman smoother on the approximated Gaussian state space model

$$\begin{aligned}\text{Observation model } &: Y_t|\theta_t \sim \left[\exp\left(\tilde{\lambda}_t^{(m)}\right), \exp\left(\tilde{\lambda}_t^{(m)}\right)\right]\\ \text{State model } &: \theta_t|\theta_{t-1} \sim \left[G_t \tilde{m}_{t-1}^{(m-1)}, \frac{r + \left(\tilde{\omega}_t^{(m)}\right)^2}{r+1}\right],\end{aligned}$$

we obtain the $(m+1)$th estimates of $\tilde{\lambda}_t^{(m+1)}$ and $\tilde{\omega}_t^{(m+1)}$. Iterations are performed until convergence is reached.

Chapter 5

Model Diagnostics

Having formulated and applied a given model to data, it is important to evaluate the adequacy of the model, that is investigate if the model assumptions are fulfilled. This may be performed by residual analysis, as proposed by Jørgensen et al. (1999), where diagnostics, in form of residuals for a non-Gaussian state space model for longitudinal Poisson data driven by a gamma distributed latent process, are outlined. This chapter is mainly inspired by Dethlefsen (2001).

5.1 Residuals for Gaussian State Space Models

The residuals are based on the Kalman filter, the Kalman smoother and the one-step forecasts. Each residual origins from either the observation model or the state model, denoted Y and θ, respectively, giving rise to six types of residuals listed in Table 5.1. The residuals are all Gaussian distributed with zero mean, assuming both the observation- and the state model are Gaussian.
From the Kalman filter, Theorem 3.1, we have $a_t = G_t(a_{t-1} + A_{t-1}e_{t-1})$, which

Type	Origin	Definition	Variance matrix
Filter	Y	$v_t = Y_t - F_t^\top m_t$	$(I - F_t^\top A_t)Q_t(I - F_t^\top A_t)^\top$
	θ	$w_t = m_t - G_t m_{t-1}$	$A_t Q_t A_t^\top$
Smoother	Y	$\tilde{v}_t = Y_t - F_t^\top \tilde{m}_t$	$V_t - F_t^\top \tilde{C}_t F_t$
	θ	$\tilde{w}_t = \tilde{m}_t - G_t \tilde{m}_{t-1}$	$W_t - L_t$
Forecast	Y	$\vec{v}_t = Y_t - f_t$	Q_t
	θ	$\vec{w}_t = a_t - G_t a_{t-1}$	$G_t A_{t-1} Q_{t-1} (G_t A_{t-1})^\top$

Table 5.1: *Residuals based on Kalman filter, the Kalman smoother and the one-step forecasts, which origin from either the observation model, denoted Y, or the state model, denoted θ.*

yields $\vec{w}_t = G_t A_{t-1} \vec{v}_t$. Now we can rewrite the filter residuals as

$$v_t = Y_t - F_t^\top m_t = Y_t - F_t^\top a_t - F_t^\top A_t \vec{v}_t = (I - F_t^\top A_t)\vec{v}_t,$$

and

$$w_t = m_t - G_t m_{t-1} = a_t + A_t \vec{v}_t - G_t a_{t-1} - G_t A_{t-1} \vec{v}_{t-1} = A_t \vec{v}_t.$$

Derivation of the variance matrices for the filter- and forecast residuals is trivial. The variance matrices for the smoother residuals are derived by noticing from Theorem A.4, that $\theta_t - \tilde{m}_t$ is uncorrelated, hence independent, of Y. Then,

$$\begin{aligned} V_t &= \mathbb{V}\mathrm{ar}\big[Y_t - F_t^\top \theta_t\big] \\ &= \mathbb{V}\mathrm{ar}\big[Y_t - F_t^\top \tilde{m}_t - F_t^\top(\theta_t - \tilde{m}_t)\big] \\ &= \mathbb{V}\mathrm{ar}\big[Y_t - F_t^\top \tilde{m}_t\big] + F_t^\top(\mathbb{V}\mathrm{ar}[\theta_t - \tilde{m}_t])F_t \\ &= \mathbb{V}\mathrm{ar}\big[Y_t - F_t^\top \tilde{m}_t\big] + F_t^\top \tilde{C}_t F_t, \end{aligned}$$

which yields

$$\mathbb{V}\mathrm{ar}[\tilde{v}_t] = \mathbb{V}\mathrm{ar}\big[Y_t - F_t^\top \tilde{m}_t\big] = V_t - F_t^\top \tilde{C}_t F_t.$$

Furthermore,

$$\begin{aligned} W_t &= \mathbb{V}\mathrm{ar}[\theta_t - G_t \theta_{t-1}] \\ &= \mathbb{V}\mathrm{ar}[\tilde{m}_t - G_t \tilde{m}_{t-1} + \theta_t - \tilde{m}_t - G_t(\theta_{t-1} - \tilde{m}_{t-1})] \\ &= \mathbb{V}\mathrm{ar}[\tilde{m}_t - G_t \tilde{m}_{t-1}] + \mathbb{V}\mathrm{ar}[\theta_t - \tilde{m}_t] + G_t \mathbb{V}\mathrm{ar}[\theta_{t-1} - \tilde{m}_{t-1}] G_t^\top \\ &\quad - \mathbb{C}\mathrm{ov}\left[\theta_t - \tilde{m}_t\,,\, \theta_{t-1} - \tilde{m}_{t-1}\right] G_t^\top - G_t \mathbb{C}\mathrm{ov}\left[\theta_{t-1} - \tilde{m}_{t-1}\,,\, \theta_t - \tilde{m}_t\right] \\ &= \mathbb{V}\mathrm{ar}[\tilde{m}_t - G_t \tilde{m}_{t-1}] + L_t, \end{aligned}$$

where L_t is as given in (3.19) on page 68. This yields that

$$\mathbb{V}\mathrm{ar}[\tilde{w}_t] = \mathbb{V}\mathrm{ar}[\tilde{m}_t - G_t \tilde{m}_{t-1}] = W_t - L_t.$$

Notice, that in the univariate case the residuals $\vec{w}_t$, v_t and w_t is proportional to $\vec{v}_t$, hence, no additional information about the adequacy of the model is revealed from these residuals. However, in the multivariate case they are all linear transformations of $\vec{v}_t$, hence, inadequacies not revealed by $\vec{v}_t$ may be revealed by $\vec{w}_t$, v_t and w_t.
Letting $Y^t = \begin{bmatrix} Y_1^\top & \dots & Y_t^\top \end{bmatrix}^\top$, we notice, that

$$\mathbb{C}\mathrm{ov}\left[Y_t - f_t\,,\, Y^{t-1}\right] = \mathbb{C}\mathrm{ov}\left[\mathbb{E}\big[Y_t - f_t \mid Y^{t-1}\big]\,,\, Y^{t-1}\right] = 0,$$

since $\mathbb{E}\big[Y_t - f_t \mid Y^{t-1}\big] = 0$. This means, that the one-step forecast residual, $\vec{v}_t$, at time t is independent of the previous observations. Since $\vec{v}_{t-s}$, $s \geq 1$ is a linear function of Y^{t-s}, we have that the one-step forecast residuals are

mutually independent over time, i.e. $\vec{v}_t \perp\!\!\!\perp \vec{v}_{t-s}$, $s \geq 1$. It follows, since $\vec{w}_t$, v_t and w_t are all linear transformations of $\vec{v}_t$, that the arguments of independence hold for these residuals as well. Using this property we may investigate the independence structure of the model by inspection of these residuals.
The smoother residuals are not independent over time, hence, a large residual due to an outlier, causes succeeding residuals to be large as well, due to the serially correlation. Hence, these residuals must be used with caution and consequently can not be used to investigate the independence structure of the model. However, the residuals may be used to reveal misspecifications of the observation model, as well as the state model.
To obtain standardised residuals, we multiply the residuals in Table 5.1 with the appropriate inverse variance matrix, e.g.

$$r_t^{\vec{v}_t} = Q^{-\frac{1}{2}}\vec{v}_t,$$

introducing the notation $r_t^{\bullet}$ for the standardised residuals.

5.1.1 Independence Structure

To investigate the independence structure of the model we consider the **autocovariance function** defined by

$$\gamma(t,s) = \mathbb{E}\big[(Y_t - \mu_t)(Y_s - \mu_s)^\top\big] = \mathbb{C}\mathrm{ov}\left[Y_t\,,\, Y_s\right].$$

The error sequences, $\{\nu_t\}$ and $\{\omega_t\}$, are both serially uncorrelated, and in addition, assuming the sequences being Gaussian with zero mean and variance σ^2, they are serially independent. Such sequences are commonly referred to as white noise processes. The autocovariance function of a white noise process is

$$\gamma(t,s) = \gamma(|\Delta t|) = \mathbb{E}[\nu_t\nu_s] = \begin{cases} \sigma^2, & t = s \\ 0, & t \neq s, \end{cases}$$

hence, only dependent on the time difference $|\Delta t| = |t-s|$. White noise processes are characterised as being **second order** or **weak stationary**, for which the autocovariance function only dependents on the time difference, $|\Delta t|$ (Diggle, 1990). On the contrary, a random walk process, e.g. the process $\{Z_t\}$ recursively defined by $Z_t = Z_{t-1} + \nu_t$, is not second order stationary, since, firstly, the mean value of the process is not constant for all t, unless assuming $Z_0 = 0$. Secondly, even in case of constant zero mean, the variance of the process is

$$\mathbb{V}\mathrm{ar}[Z_t] = \mathbb{E}\big[Z_t^2\big] = \mathbb{V}\mathrm{ar}[Z_{t-1}] + \sigma^2 = t\sigma^2.$$

In fact, the forecast residuals interpreted as a process, $\{\vec{v}_t\}$, is a white noise process, as well as the processes $\{\vec{w}_t\}$, $\{v_t\}$ and $\{w_t\}$.
The **autocorrelation function** is defined as

$$\rho(|\Delta t|) = \frac{\gamma(|\Delta t|)}{\gamma(0)}.$$

This function may be estimated from the observations $Y_1, \ldots, Y_n$ be the k**th sample autocovariance coefficient** defined by

$$g_k = \frac{1}{n} \sum_{t=k+1}^{n} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y}), \quad k = 0, 1, \ldots, n-1,$$

where $\bar{Y}$ is the sample mean. The k**th autocorrelation coefficient** is now defined as

$$r_k = \frac{g_k}{g_0}.$$

The plot of r_k against k is called an autocorrelation plot and may be exploited to justify the independence structure of the state space model (Diggle, 1990). Autocorrelation plots of the components of the residuals based on the filter and the one-step forecast with origin from the observation- or the latent process may be used to investigate the independence structure of both the observation- and state models. Autocorrelation plots of the residuals v_t and $\vec{v}_t$ reveal the dependence structure of the observation model, whereas the residuals w_t and $\vec{w}_t$ reveal the dependence structure of the state model.

5.1.2 Observation Model

Plots of the components of the residuals with origin from the observation model against time are investigated for any systematic time dependence, which may reveal a possible misspecification of the observation model. Plots of the components of the residuals with origin from the observation model against the corresponding components of a_t, m_t and $\tilde{m}_t$ are investigated to reveal misspecifications of the explanatory variables in the observation equation.

5.1.3 State Model

Plots of the components of the residuals with origin from the state model against time are investigated for any systematic time dependence, which may reveal a possible misspecification of the state model. Furthermore, plots of the components of the residuals with origin from the state model against the corresponding components of a_t, m_t and $\tilde{m}_t$ may also reveal a possible misspecification of the state model.

Chapter 6

Formulating State Space Models In `sspir`

In this chapter we derive a formulation of state space models for a time series consisting of serially correlated data exhibiting seasonal variation, which applies to the package `sspir` implemented i `R` (Dethlefsen and Lundbye-Christensen, 2006; R Development Core Team, 2008). An example illustrating formulation, estimation of variance matrices and residual analysis follows. We assume that data are equidistant and the distribution of data is a natural exponential family. Let $\{Y_t\}$ denote the time series, the model is described by

$$p(Y_t|\eta_t) = \exp\left(Y_t\eta_t - b(\eta_t) + c(Y_t)\right) \tag{6.1}$$

$$g(b'(\eta_t)) = \lambda_t = F_t^\top \theta_t \tag{6.2}$$

$$\theta_t = G_t\theta_{t-1} + \omega_t, \quad \omega_t \sim [0, W_t] \tag{6.3}$$

$$\theta_0 \sim [m_0, C_0]. \tag{6.4}$$

The linear predictor may be decomposed into four components, T_t, H_t, S_t and R_t, describing the secular trend, harmonic seasonality, unstructured seasonality and regression with possibly time varying explanatory variables, respectively, i.e.

$$\lambda_t = T_t^\top \theta_t^{(1)} + H_t^\top \theta_t^{(2)} + S_t^\top \theta_t^{(3)} + R_t^\top \theta_t^{(4)}.$$

The block diagonal evolution transfer matrix, G_t, has the form

$$G = \begin{bmatrix} G_t^{(1)} & 0 & 0 & 0 \\ 0 & G_t^{(2)} & 0 & 0 \\ 0 & 0 & G_t^{(3)} & 0 \\ 0 & 0 & 0 & G_t^{(4)} \end{bmatrix},$$

and the design matrix, F_t, is of the form

$$F_t^\top = \begin{bmatrix} T_t^\top & H_t^\top & S_t^\top & R_t^\top \end{bmatrix}.$$

Each component of the evolution transfer matrix, G_t, is derived in the following.

6.1 Secular Trend

The secular trend of the observations, Y_t, may be modelled by a sufficiently smooth function. In a dynamic setting a low degree polynomial with time varying coefficients may suffice, whereas in a static setting a sufficiently smooth function is provided by e.g. a high degree polynomial or splines (Dethlefsen and Lundbye-Christensen, 2006). A **local polynomial growth** secular trend is when the latent process evolves smoothly in time according to a polynomial of order p, i.e.

$$\theta_t^{(1)} = q(t) = \epsilon_{t,0} + \epsilon_{t,1}t + \ldots + \epsilon_{t,p}t^p. \tag{6.5}$$

The coefficients of the polynomial, $\epsilon_{t,i}$, are allowed to vary in time as emphasised by the subscript. Assuming equidistant observations we may approximate θ_t by a Taylor expansion of order p around the previous state at time $t-1$, hence,

$$q(t) \approx q(t-1) + q'(t-1) + \frac{1}{2}q''(t-1) + \ldots + \frac{1}{p!}q^{(p)}(t-1). \tag{6.6}$$

The evolution transfer matrix becomes

$$G_t^{(1)} = \begin{bmatrix} 1 & 1 & \cdots & \frac{1}{p!} \\ 0 & 1 & \cdots & \frac{1}{(p-1)!} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{(p+1)\times(p+1)},$$

and the $(p+1)$ state vector is $\theta_t^{(1)} = \begin{bmatrix} q(t) & q'(t) & \cdots & q^{(p)}(t) \end{bmatrix}^\top$, hence

$$\theta_t^{(1)} = G_t^{(1)}\theta_{t-1}^{(1)} + \omega_t^{(1)}, \quad \omega_t^{(1)} \sim [0, W_t^{(1)}].$$

Since the secular trend component is the first element in the state vector, we have

$$T_t^\top \theta_t^{(1)} = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix} \theta_t^{(1)}.$$

Formulaes for non-equidistant observations are derived trivially.
In case $p = 1$ and the evolution variance matrix is of the form

$$W^{(1)} = \omega^{(1)} \begin{bmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix} \tag{6.7}$$

the secular trend is modelled by a cubic spline (Carter and Kohn, 1994).

6.2 Harmonic Seasonality

Harmonic seasonal variation is characterised by a cyclic behavior during a given period, hence the same behavior is repeated through consecutive periods. Formally, this can be described by a real-valued function $g(t)$, $t \in \mathbb{Z}_+$, where t is a time index. The function $g(t)$ is called **cyclical**, if, for some integer $\Delta > 1$ and for all integers $t, d \geq 0$, it holds that

$$g(t + d\Delta) = g(t). \tag{6.8}$$

The smallest integer Δ such that (6.8) holds is called the **period**. Hence, the function g exhibit a full **cycle** in any time interval of length $[t, t + \Delta - 1]$ for all t. The **seasonal factors** are the Δ values taken at any full cycle. Decompose the seasonal factors into one deseasonalised level, and Δ deviations from this level, these Δ seasonal deviations are called the **seasonal effects**.
Such seasonal effects pattern may be modelled by a dth degree trigonometrical polynomial

$$\begin{aligned} H_t^\top \theta_t^{(2)} &= \sum_{i=1}^{d} \left\{ \alpha_{t,i} \cos\left(i\frac{2\pi}{\Delta} t \right) + \beta_{t,i} \sin\left(i\frac{2\pi}{\Delta} t \right) \right\} \\ &= \begin{bmatrix} c_{t,1} & \cdots & c_{t,d} & s_{t,1} & \cdots & s_{t,d} \end{bmatrix} \begin{bmatrix} \alpha_{t,1} \\ \vdots \\ \alpha_{t,d} \\ \beta_{t,1} \\ \vdots \\ \beta_{t,d} \end{bmatrix}, \end{aligned}$$

where $c_{t,i} = \cos(i2\pi t/\Delta)$ and $s_{t,i} = \sin(i2\pi t/\Delta)$ (Dethlefsen and Lundbye-Christensen, 2006).
The dynamic structure of this component is modelled in the state model by letting the seasonal effects follow a first order random walk, hence

$$\theta_t^{(2)} = \theta_{t-1}^{(2)} + \omega_t^{(2)} = \begin{bmatrix} \alpha_{t-1,1} \\ \vdots \\ \alpha_{t-1,d} \\ \beta_{t-1,1} \\ \vdots \\ \beta_{t-1,d} \end{bmatrix} + \omega_t^{(2)}, \quad \omega_t^{(2)} \sim [0, W_t^{(2)}],$$

and we see that $G_t^{(2)} = I_{2d}$.

6.3 Unstructured Seasonality

Seasonal patterns not explained by a harmonic seasonality may be modelled by an **unstructured seasonality**. This seasonality may be parameterised by letting the seasonal effects sum to an uncorrelated zero-mean error sequence in a dynamic setting, whereas the seasonal effects sum to zero in a static setting. Let Δ denote the period, then the constraint in the dynamic setting can be expressed as

$$\sum_{i=0}^{\Delta-1} \gamma_{t-i} = \omega_t^{*(3)}, \quad \omega_t^{*(3)} \sim [0, W_t^{*(3)}],$$

yielding

$$\gamma_t = -\gamma_{t-1} - \gamma_{t-2} - \ldots - \gamma_{t-\Delta+1} + \omega_t^{*(3)}.$$

In matrix notation we have

$$\begin{aligned}\theta_t^{(3)} &= G_t^{(3)}\theta_{t-1}^{(3)} + \omega_t^{(3)} \\ &= \begin{bmatrix} -1 & -1 & \cdots & -1 \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \gamma_{t-1} \\ \gamma_{t-2} \\ \vdots \\ \gamma_{t-\Delta+1} \end{bmatrix} + \begin{bmatrix} \omega_t^{*(3)} \\ 0 \\ \vdots \\ 0 \end{bmatrix},\end{aligned}$$

where $G_t^{(3)}$ is a $(\Delta-1)\times(\Delta-1)$ matrix, $\theta_t^{(3)}$ is a $(\Delta-1)$ vector and $W_t^{(3)} = \text{diag}(W_t^{*(3)}, 0, \ldots, 0)$ (Kitagawa and Gersch, 1984; Dethlefsen and Lundbye-Christensen, 2006). Since the first term in $\theta_t^{(3)}$ is γ_t, we have

$$S_t^\top \theta_t^{(3)} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \theta_t^{(3)}.$$

6.4 Regression on Explanatory Variables

Observed, possibly time varying, explanatory variables, represented by the matrix R_t, are modelled by the usual regression term

$$R_t^\top \theta_t^{(4)},$$

where the regression coefficients are allowed to evolve over time according to a random walk, hence $\theta_t^{(4)} = \theta_{t-1}^{(4)} + \omega_t^{(4)}$, where $\omega_t^{(4)} \sim [0, W_t^{(4)}]$. Furthermore, the regression component is specified by the usual Wilkinson-Rogers formula notation in `R` (Dethlefsen and Lundbye-Christensen, 2006).

The variance structures of all the components are specified by the modeller and reflect the dependence internally in each component. The variance structure may be known to the modeller, however, the exact variance may be unknown. In estimation of unknown variances, these structures may be exploited as described in Section 3.7.3.

6.5 Example - Simulation Study

To illustrate the formulation of a state space model in `sspir` we provide a hypothetical example of Gaussian observations exhibiting seasonal variation. Estimation of variance matrices using the EM algorithm is performed, and model verification using residual analysis is outlined.

6.5.1 Formulation of Model

Assume the secular trend is described by a local polynomial growth curve with $p = 1$ and variance structure giving by (6.7), hence a cubic spline and the seasonality exhibited by data is described by harmonic seasonality of degree one, hence a single cycle. The period of seasonality, Δ, is dependent on the frequency of which data are observed, say we have daily observations and we want to asses the seasonal variation during a year, the period may be $\Delta = 365$. In matrix notation the model is

$$Y_t = \begin{bmatrix} 1 & 0 & \cos\left(\frac{2\pi t}{365}\right) & \sin\left(\frac{2\pi t}{365}\right) \end{bmatrix} \begin{bmatrix} q(t) \\ q'(t) \\ \alpha_t \\ \beta_t \end{bmatrix} + \nu_t, \quad \nu_t \sim \mathcal{N}(0\,,\, V) \tag{6.9}$$

$$\theta_t = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} q(t-1) \\ q'(t-1) \\ \alpha_{t-1} \\ \beta_{t-1} \end{bmatrix} + \omega_t, \quad \omega_t \sim \mathcal{N}_4\left(0\,,\, W(\phi)\right), \tag{6.10}$$

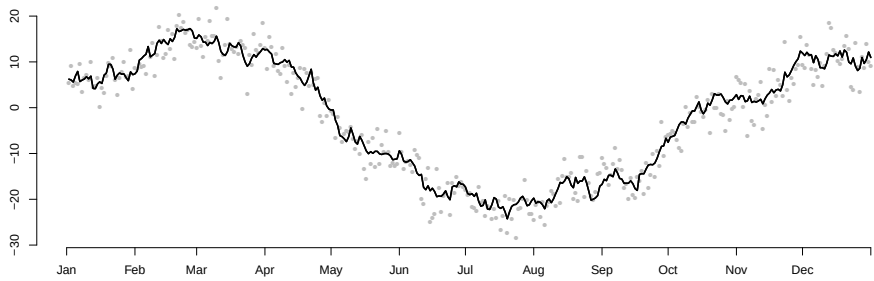
assuming the variance matrices V and $W(\phi)$ are constant.
Using the function `recursion` implemented in `sspir`, we may simulate observations according to the model (6.9)-(6.10) with prespecified variance matrices and initial distribution of the latent process. Let the observation variance matrix, V, and the evolution variance matrix, $W(\phi)$, be given by

$$V = 10, \quad W(\phi) = \begin{bmatrix} \frac{\phi_1}{3} & \frac{\phi_1}{2} & 0 & 0 \\ \frac{\phi_1}{2} & \phi_1 & 0 & 0 \\ 0 & 0 & \phi_2 & 0 \\ 0 & 0 & 0 & \phi_2 \end{bmatrix}, \quad \phi = \begin{bmatrix} 10^{-6} \\ 1 \end{bmatrix}, \tag{6.11}$$

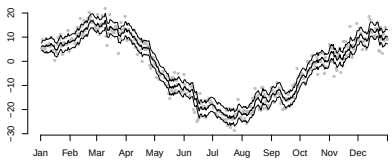
and the initial distribution of the latent process be specified by

$$m_0 = \begin{bmatrix} 1 \\ 10^{-4} \\ 5 \\ 5 \end{bmatrix} \quad \text{and} \quad C_0 = \begin{bmatrix} 3 \cdot 10^{-4} & 0 & 0 & 0 \\ 0 & 3 \cdot 10^{-6} & 0 & 0 \\ 0 & 0 & 10^{-5} & 0 \\ 0 & 0 & 0 & 10^{-5} \end{bmatrix}. \tag{6.12}$$

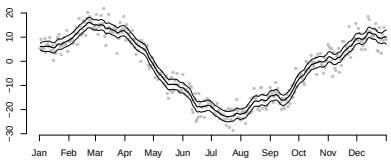
The model is defined in `sspir` as



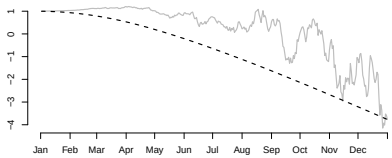
(a) Simulated observations and true latent process.



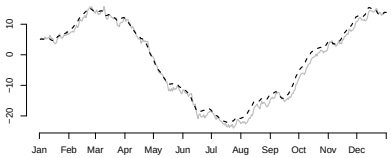
(b) Kalman filter applied to simulated observations.



(c) Kalman smoother applied to simulated observations.

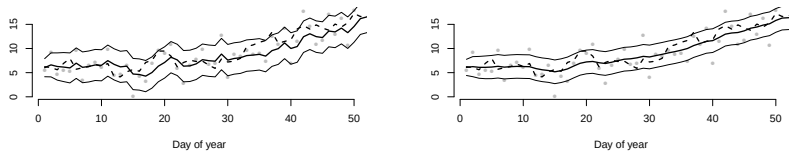


(d) Secular trend.



(e) Harmonic seasonal variation.

Figure 6.1: *The first plot shows simulated observations from model* (6.9)-(6.10) *with constant variance matrices given as* (6.11)*, represented by dots, and the solid line represents the true latent process. The second plot shows the filtered estimates of the latent process,* $\{m_t\}$*, with* 95% *confidence limits, whereas the third plot shows the smoothed estimates of the latent process,* $\{\widetilde{m}_t\}$*, with* 95% *confidence limits. The fourth plot illustrates the filtered and smoothed secular trend component by the solid and dashed lines, respectively. The last plot illustrates the filtered and smoothed harmonic seasonal variation component by the solid and dashed lines, respectively.*



(a) Filtered estimates, $\{m_t\}$. (b) Smoothed estimates, $\{\tilde{m}_t\}$.

Figure 6.2: *The first plot illustrates the filtered latent process for the first* 50 *observations (solid line) with* 95% *confidence limits, and the true latent process (dashed line). The second plot illustrates the smoothed latent process for the first* 50 *observations (solid line) with* 95% *confidence limits, and the true latent process (dashed line).*

```
m1 <- SS(
    Fmat = function(tt,x,phi) {
            Fmat      <- matrix(0,4,1)
            Fmat[1,1] <- 1
            Fmat[3,1] <- cos(2*pi*tt/365)
            Fmat[4,1] <- sin(2*pi*tt/365)
            return(Fmat) },
    Gmat = function(tt,x,phi) {
            return(matrix(c(1,0,0,0,1,1,0,0,0,0,1,0,0,0,0,1),nrow=4)) },
    Wmat = function(tt,x,phi) {
            return( matrix(c(phi[1]/3,phi[1]/2,0,0,
            phi[1]/2,phi[1],0,0,
            0,0,phi[2],0,
            0,0,0,phi[2]),nrow=4)) },
    Vmat = matrix(10),
    phi  = c(1e-6,1),
    m0   = matrix(c(1,1e-3,5,5),nrow=1),
    C0   = matrix(c(3e-3,0,0,0,0,3e-5,0,0,0,0,1e-4,0,0,0,0,1e-4),
            nrow=4,ncol=4)
)
```

Daily observations are simulated by the function `recursion(model,n)`. This function provides observations simulated according to the defined model `m1` and returns the entire model including the simulated observations. Furthermore, the function also provides the values of the true latent process. Letting $n = 365$, we simulate observations from a single year, setting a seed provides reproducible results, hence,

```
set.seed(51885273)
m1 <- recursion(m1,365)
```

The simulated observations are plotted in Figure 6.1(a) represented by grey

dots, whereas the true latent process is represented by the solid line. When the observation variance matrix, V, has small values the observations are closer to the latent process than in case of large values of V. Similarly, the latent process fluctuates less, when the evolution variance matrix, $W(\phi)$, has small values, than for large values.

6.5.2 Estimation of Variance Matrices

In order to estimate the variance matrices, V and $W(\phi)$, we apply the EM algorithm. Choosing different initial values and convergence criterion, `epsilon`, we obtain different estimates. Letting m_0, C_0, V and ϕ be given as (6.11) and (6.12), in Table 6.1 the different initial values are listed together with the corresponding estimates and the number of iterations upon convergence, where the maximum number of iterations to run is $10,000$, along with the maximum value of the log likelihood function determined by the model with the estimated variance matrices. Furthermore, the structure of the evolution variance matrix is ensured being block diagonal with $W(\phi)^{(1)}$, given in (6.7), as the first block, and $W(\phi)^{(2)} = I_2$ as the second, see Section 3.7.3.

When using the true values as initial values the estimates of V and $W(\phi)$ are slightly underestimated in comparison with true values, especially V and ϕ_2. We see that estimation initialised by a misspecified initial mean, $0.05m_0$, together with a relatively weak initial variance matrix, $10C_0$, results in estimates of V and $W(\phi)$ relatively close to the true values, whereas a misspecified initial mean, $10m_0$, together with a relatively strong initial variance, $0.05C_0$, results in estimates far from the true values, except from the estimate of ϕ_1.

Furthermore, initialising the observation variance, V, as relatively weak, $10V$, or strong, $0.05V$, results in estimates close to the true value, whereas a weak initial variance of the evolution variance matrix, 10ϕ, results in a relatively large overestimate of the variance of the secular trend, $W^{(1)}(\phi)$, and a relatively large underestimate, when initialising with a strong variance, 0.05ϕ. The estimate of the harmonic seasonality, $W^{(2)}(\phi)$, is relatively close to the true value regardless of the initialising.

In this example, estimation initialised by a relatively strong initial observation variance matrix along with a relatively weak evolution variance matrix produces the highest maximum value of the log likelihood function, whereas the lowest are produced, when the initial mean is misspecified by $10m_0$, along with a strong initial variance, $0.05C_0$. Furthermore, we see that, in general, as `epsilon` decreases, the maximum value of the log likelihood increases along with the number of iterations upon convergence.

In Figure 6.1(b), the Kalman filter is applied to the simulated observations and the estimated variance matrices $\hat{V}$ and $W(\hat{\phi})$ provided by the EM algorithm initialised by m_0, C_0, $0.05V$ and 10ϕ along with `epsilon`$= 10^{-3}$, obtaining the estimates $\{m_t\}$ of the latent process with 95% confidence limits, whereas in

epsilon	$m_{(0)}$	$C_{(0)}$	$V_{(0)}$	$\phi_{(0)}$	$\hat{V}$	$\hat{\phi}$	**Iterations**	**Max**
	m_0	C_0	V	ϕ	9.2106	$(1.0089 \cdot 10^{-6}\ ,\ 0.6434)^\top$	98	-973.3704
	$0.05m_0$	$10C_0$	V	ϕ	9.0974	$[0.9958 \cdot 10^{-6}\ \ 0.8841]^\top$	51	-979.7101
	$10m_0$	$0.05C_0$	V	ϕ	4.7158	$[1.0096 \cdot 10^{-6}\ \ 17.7258]^\top$	74	-1117.3420
10^{-3}	m_0	C_0	$0.05V$	10ϕ	9.2253	$[9.0685 \cdot 10^{-6}\ \ 0.6313]^\top$	192	-973.4573
	m_0	C_0	$10V$	0.05ϕ	9.2534	$[0.0510 \cdot 10^{-6}\ \ 0.6124]^\top$	167	-973.4588
	m_0	C_0	$0.05V$	0.05ϕ	9.2533	$[0.0506 \cdot 10^{-6}\ \ 0.6124]^\top$	152	-973.4588
	m_0	C_0	$10V$	10ϕ	9.2256	$[9.1232 \cdot 10^{-6}\ \ 0.6310]^\top$	159	-973.4585
	m_0	C_0	V	ϕ	9.2498	$[1.8262 \cdot 10^{-6}\ \ 0.6137]^\top$	4566	-973.3343
	$0.05m_0$	$10C_0$	V	ϕ	9.1251	$[0.9899 \cdot 10^{-6}\ \ 0.8600]^\top$	124	-979.7034
	$10m_0$	$0.05C_0$	V	ϕ	4.6446	$[1.0099 \cdot 10^{-6}\ \ 17.8392]^\top$	116	-1117.3400
10^{-4}	m_0	C_0	$0.05V$	10ϕ	9.2560	$[3.5006 \cdot 10^{-6}\ \ 0.6088]^\top$	3740	-973.3310
	m_0	C_0	$10V$	0.05ϕ	9.2258	$[0.0511 \cdot 10^{-6}\ \ 0.6328]^\top$	253	-973.4520
	m_0	C_0	$0.05V$	0.05ϕ	9.2258	$[0.0507 \cdot 10^{-6}\ \ 0.6328]^\top$	238	-973.4521
	m_0	C_0	$10V$	10ϕ	9.2560	$[3.5009 \cdot 10^{-6}\ \ 0.6088]^\top$	3716	-973.3310
	m_0	C_0	V	ϕ	9.2531	$[2.4896 \cdot 10^{-6}\ \ 0.6110]^\top$	10000	-973.3269
	$0.05m_0$	$10C_0$	V	ϕ	9.1276	$[0.5538 \cdot 10^{-6}\ \ 0.8579]^\top$	10000	-979.6914
	$10m_0$	$0.05C_0$	V	ϕ	4.6373	$[1.0102 \cdot 10^{-6}\ \ 17.8510]^\top$	159	-1117.3400
10^{-5}	m_0	C_0	$0.05V$	10ϕ	9.2542	$[2.7744 \cdot 10^{-6}\ \ 0.6102]^\top$	9935	-973.3265
	m_0	C_0	$10V$	0.05ϕ	9.2231	$[0.0632 \cdot 10^{-6}\ \ 0.6348]^\top$	10000	-973.4503
	m_0	C_0	$0.05V$	0.05ϕ	9.2231	$[0.0627 \cdot 10^{-6}\ \ 0.6348]^\top$	10000	-973.3265
	m_0	C_0	$10V$	10ϕ	9.2545	$[3.0700 \cdot 10^{-6}\ \ 0.6100]^\top$	8550	-973.3228

Table 6.1: *Estimation of variance matrices, V and $W(\phi)$, with different initial values and the number of iterations upon convergence along with corresponding maximum value of the log likelihood function. The initial values of V, ϕ, m_0 and C_0 correspond to the values of* (6.11) *and* (6.12)*, respectively.*

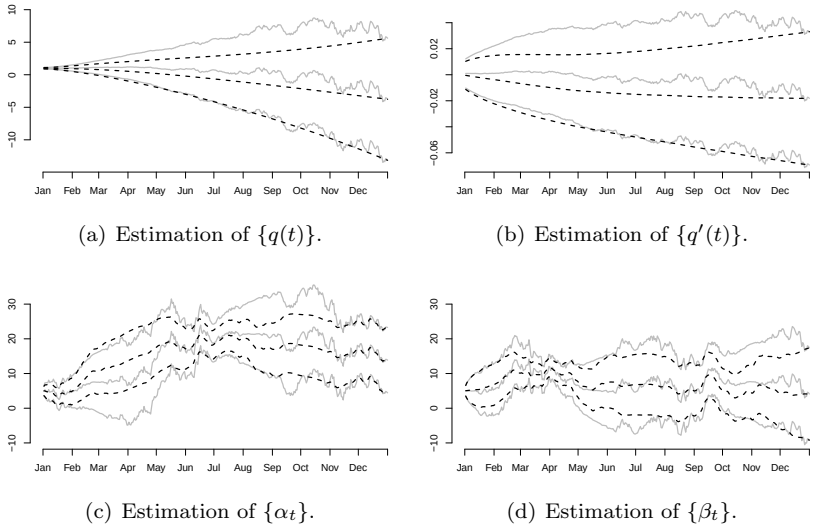
Figure 6.1(c), the Kalman smoother is applied to obtain $\{\tilde{m}_t\}$ and corresponding 95% confidence limits. We see that the estimates m_t fluctuates more than $\tilde{m}_t$, hence the estimates of the latent process becomes more smooth, when applying the Kalman smoother. Furthermore, the elements of the filtered variance matrix, C_t, are larger than the elements of the smoothed variance matrix, $\tilde{C}_t$. This is illustrated in Figure 6.2.
In Figure 6.2(a), the filtered estimates of the latent process for the first fifty observations are plotted with 95% confidence limits, solid lines, and the true latent process superimposed, dashed line and similarly in Figure 6.2(b) with the smoothed estimates. Although the confidence intervals for the smoothed estimates are more narrow, than for the filtered estimates, the true latent process is surrounded by the limits, hence the smoothed values do fit the latent process rather well.

6.5.3 Residual Analysis

The following residual analysis is performed using the estimates of V and $W(\phi)$ obtained with the EM algorithm initialised by m_0, C_0, $0.05V$ and 10ϕ along with `epsilon`$= 10^{-5}$, hence

$$\hat{V} = 9.2542, \quad \hat{\phi} = \begin{bmatrix} 2.7744 \cdot 10^{-6} & 0.6102 \end{bmatrix}^\top .$$

The Kalman filter and smoother are applied to the simulated observations and the estimated variance matrices to obtain the standardised residuals as specified in Section 5.1. Due to computational problems we were not able to standardise the residuals with origin from θ, since the corresponding variance matrices were computationally singular, however each component of the residuals were divided by the square root of the corresponding diagonal entry of the variance matrix.
In Figure 6.3, plots of estimated components of the latent process, q, q', α and β are given. For each component the filtered and smoothed estimates are plotted, with solid and dashed lines, respectively. Furthermore, the approximated 95% confidence limits are provided for q and q', due to serially correlation of q and q', whereas the 95% confidence limits for α and β are exact, since the two components are mutually independent and independent of q and q'. The components, q and q', indicate that the secular trend seems to be decreasing during the year of simulation with a decreasing slope during the first half of year, whereas it becomes approximately constant, see Figure 6.3(a) and 6.3(b).
The harmonic seasonal variation component is seen in Figure 6.1(e). The coefficients, α_t and β_t are both modelled by random walks, which is recognised in Figures 6.3(c) and 6.3(d). It seems that the coefficient α_t fluctuates more than β_t and take values in a broader range.
Autocorrelation plots of each component of the standardised forecast and filter residuals with origin from both Y and θ are given in Figure 6.4. These plots may be exploited to verify the independence structure of the model. None of



(a) Estimation of $\{q(t)\}$.

(b) Estimation of $\{q'(t)\}$.

(c) Estimation of $\{\alpha_t\}$.

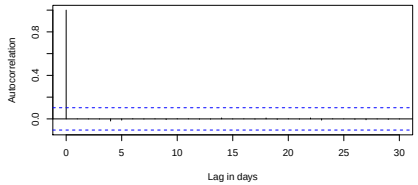
(d) Estimation of $\{\beta_t\}$.

Figure 6.3: *The filtered (solid lines) and smoothed (dashed lines) estimates of the components of the latent process with approximated* 95% *confidence limits for* q *and* q' *and exact* 95% *confidence limits for* α *and* β.

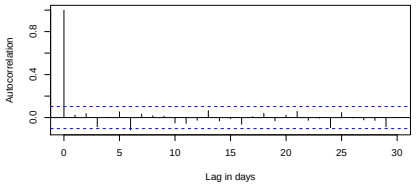
the plots indicates systematic serially correlation of the components, hence the independence structure may be verified.
Time plots of the standardised filter, smoother and forecast residuals with origin from Y are given in Figure 6.5. Misspecification of the observation model may be revealed by these plots. Both the filter and smoother residuals do not reveal any misspecification, however, the forecast residuals include few extreme outliers making the time plot different from the two others. Time plot of the forecast residual without the most extreme outliers shows no indication of misspecification.
Furthermore, time plots of the components of the standardised filter, smoother and forecast residuals with origin from θ are given in Figure 6.6. Misspecification of the state model may be revealed by these plots. No components of the filter and forecast residuals give rise to concern about misspecification of the state model, whereas the components of the smoother residuals clearly show the serially correlation of the components of the residuals, especially the components q and q'.
In order to illustrate the consequence of applying estimates provided by misspecified initial mean m_0 and a strong initial variance, C_0, residual analysis is performed for the model specified by the estimates

$$\hat{V} = 4.6373, \quad \hat{\phi} = \begin{bmatrix} 1.0102 \cdot 10^{-6} & 17.8510 \end{bmatrix}^\top.$$

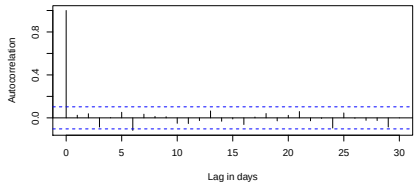
As a consequence of the relatively small observation variance, the filtered and smoothed latent process follow the observations extensively compared to the true latent process, see Figure 6.1(a) and Figure B.1. The estimated secular trend seems to be increasing and the estimated harmonic seasonal variation characterised by a single cycle, seems to be drowning in noise, as a consequence of the relatively high estimated ϕ_2, see Figure B.2. Autocorrelation plots indicate misspecification of both the observation- and the state models, in this case the variance matrices are misspecified, see Figure B.4. Additional plots are shown in Appendix B.



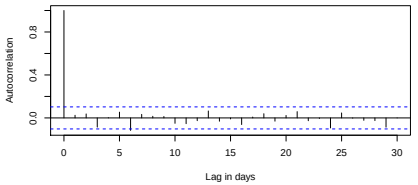
(a) Forecast residual, $r^{\vec{v}}$.



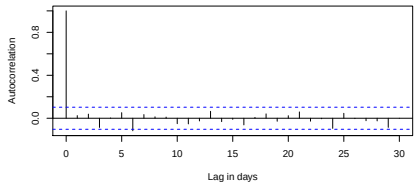
(b) Filter residual, r^{v}.



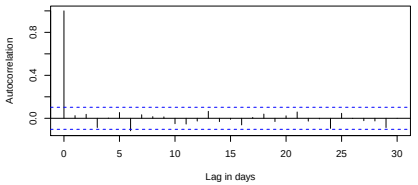
(c) Forecast residual, $r^{\vec{w}}$, q.



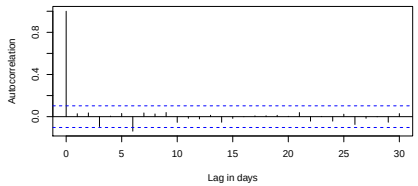
(d) Filter residual, r^{w}, q.



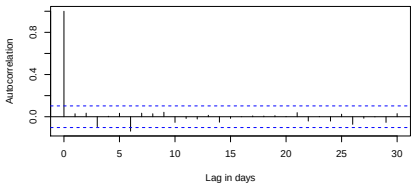
(e) Forecast residual, $r^{\vec{w}}$, q'.



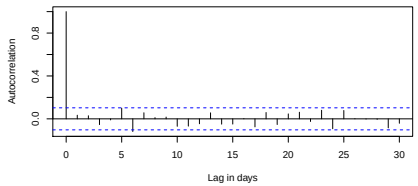
(f) Filter residual, r^{w}, q'.



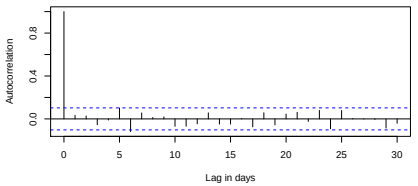
(g) Forecast residual, $r^{\vec{w}}$, α.



(h) Filter residual, r^{w}, α.

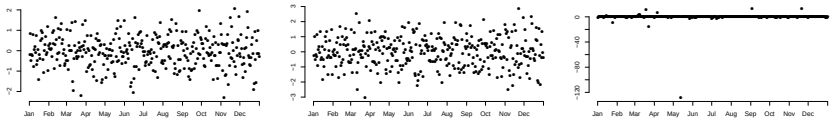


(i) Forecast residual, $r^{\vec{w}}$, β.



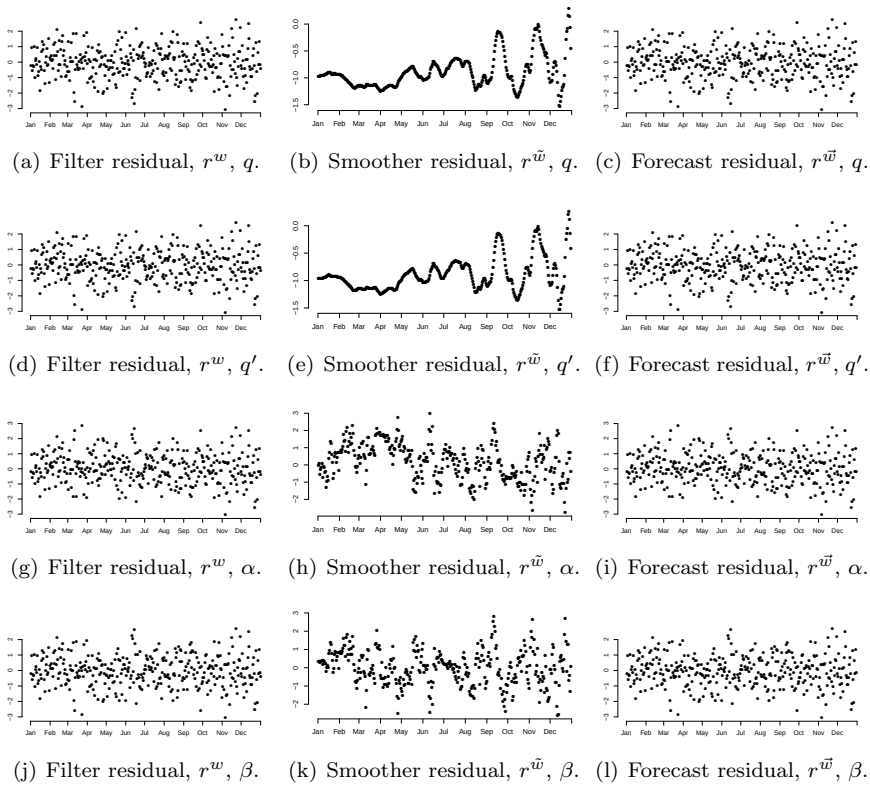
(j) Filter residual, r^{w}, β.

Figure 6.4: *Autocorrelation plots of each component of the standardised residuals with origin from Y and θ.*



(a) Filter residual, r^{v}. (b) Smoother residual, $r^{\tilde{v}}$. (c) Forecast residual, $r^{\vec{v}}$.

Figure 6.5: *Time plots of the standardised residuals with origin from Y.*



(a) Filter residual, r^{w}, q. (b) Smoother residual, $r^{\tilde{w}}$, q. (c) Forecast residual, $r^{\vec{w}}$, q.

(d) Filter residual, r^{w}, q'. (e) Smoother residual, $r^{\tilde{w}}$, q'. (f) Forecast residual, $r^{\vec{w}}$, q'.

(g) Filter residual, r^{w}, α. (h) Smoother residual, $r^{\tilde{w}}$, α. (i) Forecast residual, $r^{\vec{w}}$, α.

(j) Filter residual, r^{w}, β. (k) Smoother residual, $r^{\tilde{w}}$, β. (l) Forecast residual, $r^{\vec{w}}$, β.

Figure 6.6: *Time plots of the components of the standardised residuals with origin θ.*

Appendices

Appendix A

Miscellaneous Results

This appendix contains several results to be used in the thesis.

$$\begin{aligned}
\mathbb{E}[g(X)] &= \mathbb{E}[\mathbb{E}[g(X) \mid Z]] \\
\mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X \mid Z]] \\
\mathbb{V}\mathrm{ar}[X] &= \mathbb{V}\mathrm{ar}[\mathbb{E}[X \mid Z]] + \mathbb{E}[\mathbb{V}\mathrm{ar}[X \mid Z]] \\
\mathbb{C}\mathrm{ov}\left[X\,,\,Z\right] &= \mathbb{E}[\mathbb{C}\mathrm{ov}\left[X\,,\,Z \mid X\right]] + \mathbb{C}\mathrm{ov}\left[\mathbb{E}[X \mid X]\,,\,\mathbb{E}[Z \mid X]\right]
\end{aligned}$$

Definition A.1 (Multivariate Normal Distribution)
Let $Z^\top = \begin{bmatrix} Z_1 & \dots & Z_d \end{bmatrix}$ *be a vector of random variables. The vector* Z *is said to have a multivariate normal distribution with mean vector* μ *and a positive definite variance matrix* Σ *if its density function is given by*

$$p(Z) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(Z-\mu\right)^\top \Sigma^{-1}\left(Z-\mu\right)\right].$$

This is denoted

$$Z \sim \mathcal{N}_d\left(\mu\,,\,\Sigma\right).$$

◇

Theorem A.1 (Azzalini (1996))
Suppose $Z \sim \mathcal{N}_d\left(\mu\,,\,\Sigma\right)$ *and let*

$$Z = \begin{bmatrix} Z^{(1)} \\ Z^{(2)} \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where, for $i = 1, 2$, $Z^{(i)}$ *and* $\mu^{(i)}$ *are* $d_i \times 1$ *vectors,* Σ_{ii} *are* $d_i \times d_i$ *matrices and* $d_1 + d_2 = d$, *then*

$$Z_1 | Z^{(2)} \sim \mathcal{N}_{d_1}\left(\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(Z^{(2)} - \mu^{(2)})\,,\,\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

THEOREM A.2
Suppose $Z \sim \mathcal{N}_d(\mu, \Sigma)$, where Σ is a positive definite $(p \times p)$ matrix, A is a positive definite $(p \times p)$ matrix and ϵ is a $(p \times 1)$ constant vector, then

$$\mathbb{E}\left[\|Z - \epsilon\|^2_{A^{-1}}\right] = \mathit{trace}(A^{-1}\Sigma) + \|\mu - \epsilon\|^2_{A^{-1}}.$$

PROOF

$$\begin{aligned}
\mathbb{E}\left[\|Z - \epsilon\|^2_{A^{-1}}\right] &= \mathbb{E}\left[\text{trace}\left(\|Z - \epsilon\|^2_{A^{-1}}\right)\right] \\
&= \mathbb{E}\left[\text{trace}\left(A^{-1}(Z-\epsilon)(Z-\epsilon)^\top\right)\right] \\
&= \text{trace}\left(A^{-1}\mathbb{E}\left[(Z-\epsilon)(Z-\epsilon)^\top\right]\right) \\
&= \text{trace}\left(A^{-1}(\Sigma + \mu\mu^\top - \mu\epsilon^\top - \epsilon\mu^\top + \epsilon\epsilon^\top)\right) \\
&= \text{trace}\left(A^{-1}\Sigma\right) + \text{trace}\left(A^{-1}(Z-\epsilon)(Z-\epsilon)^\top\right) \\
&= \text{trace}\left(A^{-1}\Sigma\right) + \|\mu - \epsilon\|^2_{A^{-1}}.
\end{aligned}$$

□

THEOREM A.3
Consider the matrix function

$$f(\Sigma) = \log|\Sigma| + \mathit{trace}\left(\Sigma^{-1}A\right).$$

If A and Σ both are positive definite, then, with respect to Σ, $f(\Sigma)$ is minimised uniquely at $\Sigma = A$.

PROOF Note that $\Sigma^{-1}A$ has the same eigenvalues, denoted $\lambda_1, \ldots, \lambda_d$, as $\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}}$. The latter matrix is positive definite, since both A and Σ are positive definite, and hence all eigenvalues are positive. Instead of minimising $f(\Sigma)$ directly, we minimise $f(\Sigma) - f(A)$. Hence

$$\begin{aligned}
f(\Sigma) - f(A) &= \log|\Sigma| + \text{trace}\left(\Sigma^{-1}A\right) - \log|A| - \text{trace}\left(A^{-1}A\right) \\
&= \log|\Sigma A^{-1}| + \text{trace}\left(\Sigma^{-1}A\right) - d \\
&= -\log|\Sigma^{-1}A| + \text{trace}\left(\Sigma^{-1}A\right) - d \\
&= -\log\left(\prod_{i=1}^{d}\lambda_i\right) + \sum_{i=1}^{d}\lambda_i - d \\
&= \sum_{i=1}^{d}\left(-\log(\lambda_i) + \lambda_i - 1\right) \\
&\geq 0,
\end{aligned}$$

since $\log(\lambda_i) \leq \lambda_i - 1$ for positive eigenvalues, equality holds when $\lambda_i = 1$ for all i, thus $f(\Sigma) - f(A)$ and thereby $f(\Sigma)$ is minimised, when $\Sigma = A$. □

THEOREM A.4
Let X and Z be random vectors. It holds that $X - \mathbb{E}[X \mid Z]$ is uncorrelated with Z.

THEOREM A.5 (JENSEN'S INEQUALITY, (AZZALINI, 1996))
Let $f(X)$ be a convex function, i.e. $\rho f(X) + (1-\rho) f(Z) \geq f(\rho X + (1-\rho)Z)$ for all $\rho \in (0,1)$. It holds that

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

Appendix B

Figures

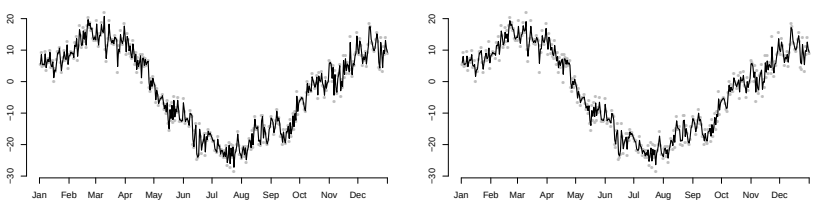
This chapter provides additional residual plots supporting statements in Section 6.5.3, and figures of results supporting statements of the analyses in Chapter 2.

B.1 Example - Simulation Study

Residual plots for the model specified by

$$\hat{\phi} = \begin{bmatrix} 1.0102 \cdot 10^{-6} \\ 17.8510 \end{bmatrix}$$

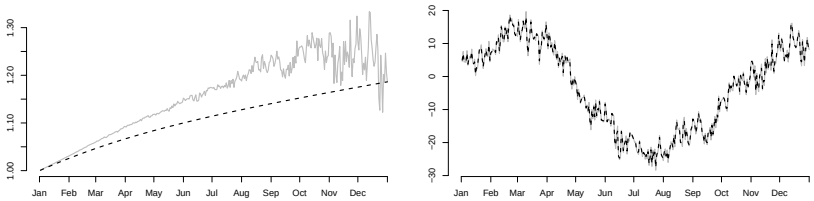
and $\hat{V} = 4.6373$ provided by the EM algorithm with initial values $10m_0$, $0.05C_0$, V and ϕ, see Table 6.1 on page 99.



(a) Kalman filter applied to simulated observations.

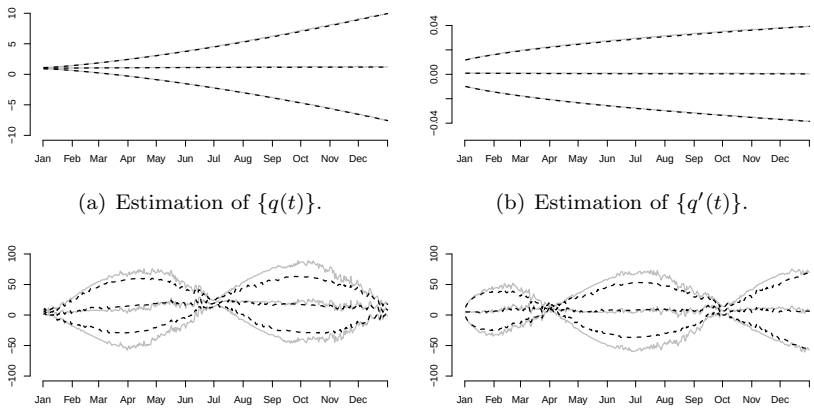
(b) Kalman smoother applied to simulated observations.

Figure B.1: *The first plot shows the filtered estimates of the latent process, $\{m_t\}$, whereas the second plot shows the smoothed estimates of the latent process, $\{\widetilde{m}_t\}$.*



(a) Secular trend. (b) Harmonic seasonal variation.

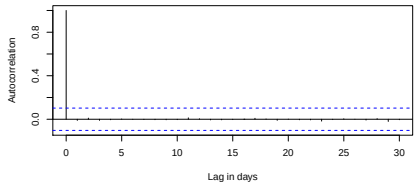
Figure B.2: *The first plot illustrates the filtered and smoothed secular trend component by the solid and dashed lines, respectively. The second plot illustrates the filtered and smoothed harmonic seasonal variation component by the solid and dashed lines, respectively.*



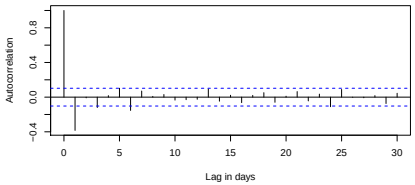
(a) Estimation of $\{q(t)\}$. (b) Estimation of $\{q'(t)\}$.

(c) Estimation of $\{\alpha_t\}$. (d) Estimation of $\{\beta_t\}$.

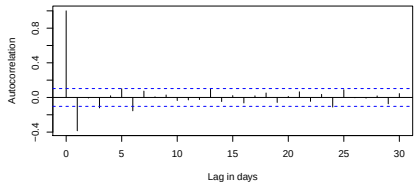
Figure B.3: *The filtered (solid lines) and smoothed (dashed lines) estimates of the components of the latent process with approximated* 95% *confidence limits for* q *and* q' *and exact* 95% *confidence limits for* α *and* β.



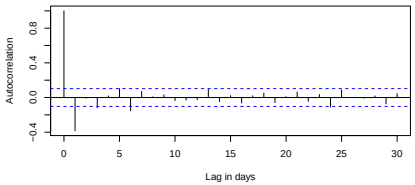
(a) Forecast residual, $r^{\vec{v}}$.



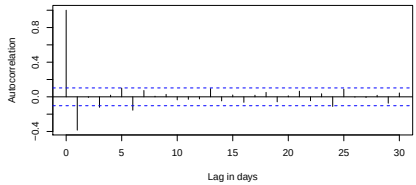
(b) Filter residual, r^{v}.



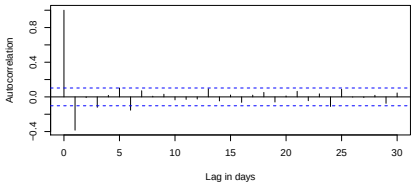
(c) Forecast residual, $r^{\vec{w}}$, q.



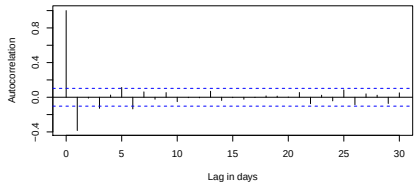
(d) Filter residual, r^{w}, q.



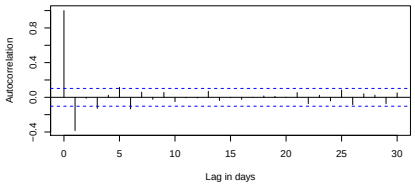
(e) Forecast residual, $r^{\vec{w}}$, q'.



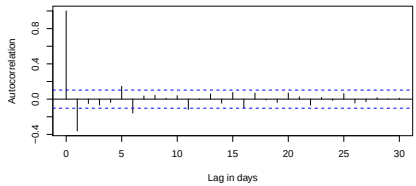
(f) Filter residual, r^{w}, q'.



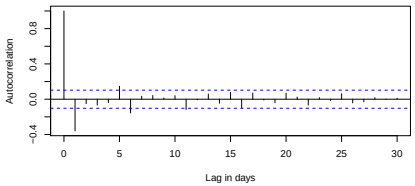
(g) Forecast residual, $r^{\vec{w}}$, α.



(h) Filter residual, r^{w}, α.

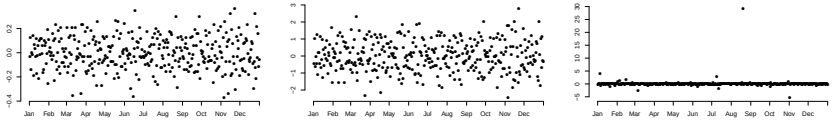


(i) Forecast residual, $r^{\vec{w}}$, β.



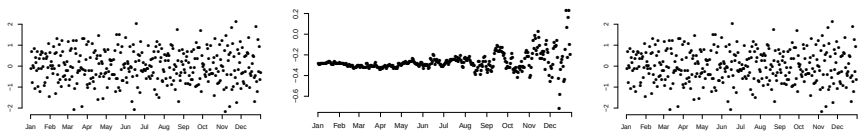
(j) Filter residual, r^{w}, β.

Figure B.4: *Autocorrelation plots of each component of the standardised residuals with origin from Y and θ.*

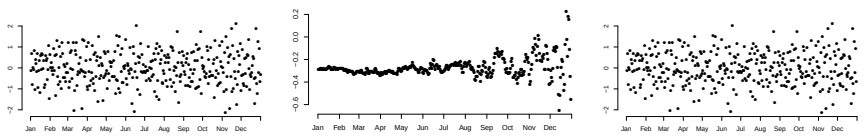


(a) Filter residual, r^{v}. (b) Smoother residual, $r^{\tilde{v}}$. (c) Forecast residual, $r^{\vec{v}}$.

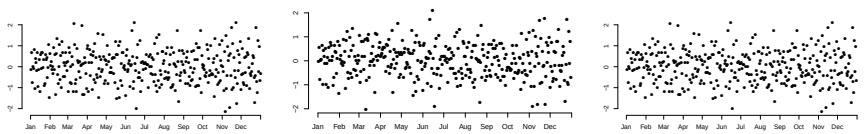
Figure B.5: *Time plots of the standardised residuals with origin from Y.*



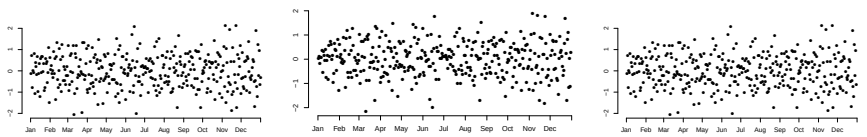
(a) Filter residual, r^{w}, q. (b) Smoother residual, $r^{\tilde{w}}$, q. (c) Forecast residual, $r^{\vec{w}}$, q.



(d) Filter residual, r^{w}, q'. (e) Smoother residual, $r^{\tilde{w}}$, q'. (f) Forecast residual, $r^{\vec{w}}$, q'.



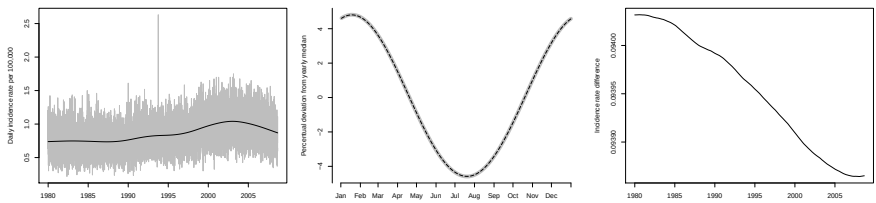
(g) Filter residual, r^{w}, α. (h) Smoother residual, $r^{\tilde{w}}$, α. (i) Forecast residual, $r^{\vec{w}}$, α.



(j) Filter residual, r^{w}, β. (k) Smoother residual, $r^{\tilde{w}}$, β. (l) Forecast residual, $r^{\vec{w}}$, β.

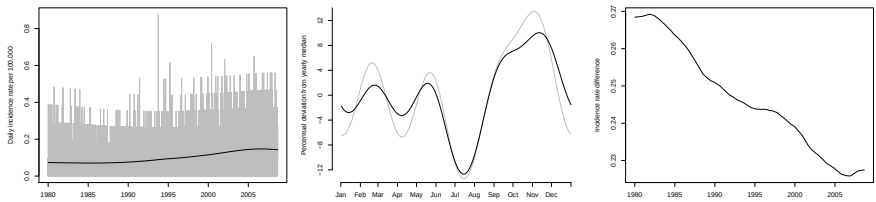
Figure B.6: *Time plots of each component of the standardised residuals with origin θ.*

B.2 Stroke



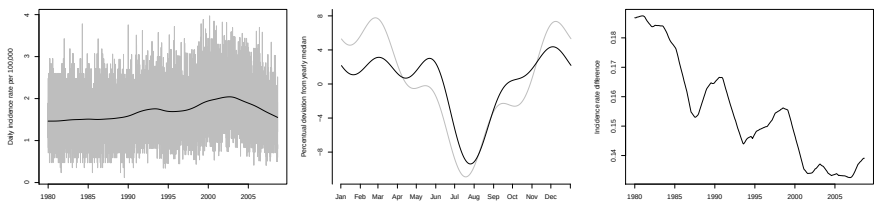
(a) (b) (c)

Figure B.7: *Estimated components of Model* 1*. (a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*



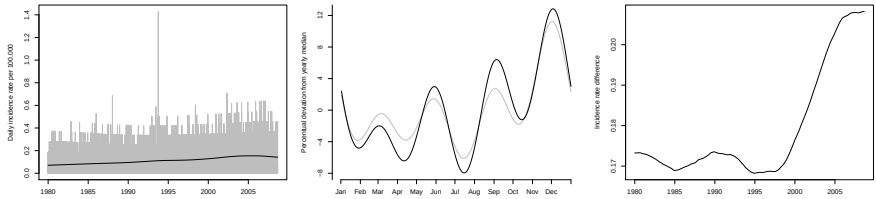
(a) (b) (c)

Figure B.8: *Estimated components of Model* 2 *for females aged* 20-49*. (a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*



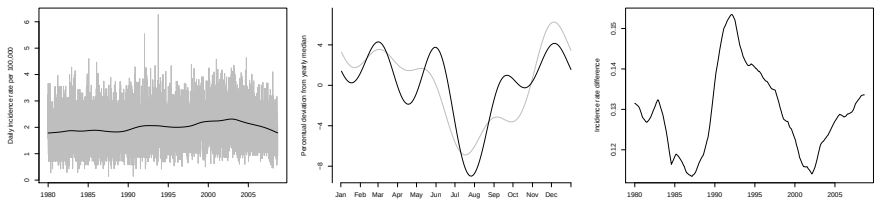
(a) (b) (c)

Figure B.9: *Estimated components of Model* 2 *for females aged* 50+*. (a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*



(a) (b) (c)

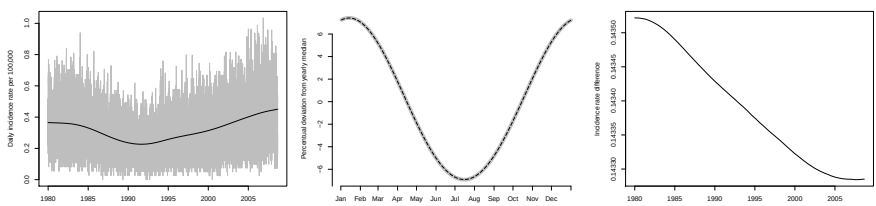
Figure B.10: *Estimated components of Model* 2 *for males aged* 20-49*. (a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*



(a) (b) (c)

Figure B.11: *Estimated components of Model* 2 *for males aged* 50+*. (a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*

B.3 VTE



(a) (b) (c)

Figure B.12: *Estimated components of Model* 1*. (a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*

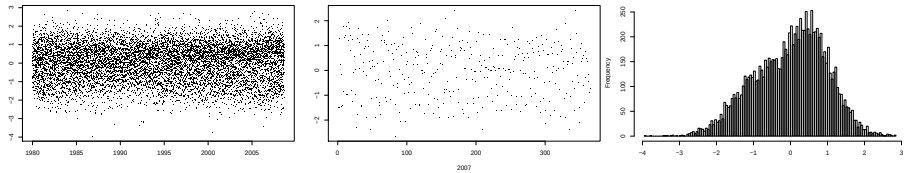


Figure B.13: *Residual analysis of the approximated Gaussian state space model of daily incidence rates based on the square root transformed observed frequencies. Time plot and histogram of the filter residual of origin Y. The latter time plot shows the residuals for the year* 2007*.*

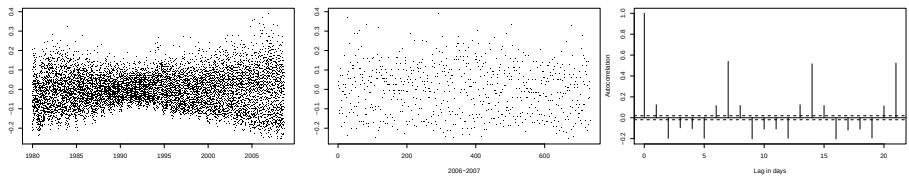


Figure B.14: *Residual analysis of Model* 1*. Time- and autocorrelation plots of the filter residual with origin Y.*

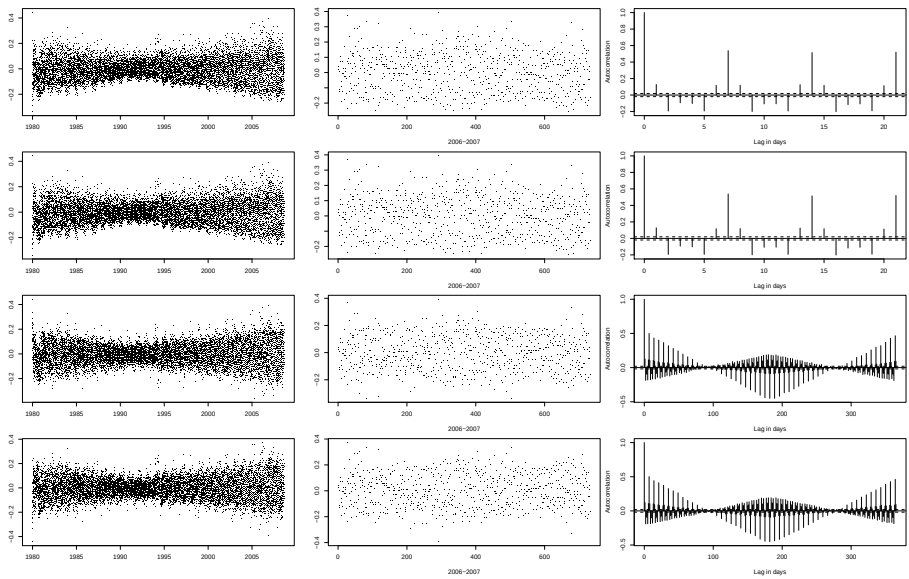


Figure B.15: *Residual analysis of Model* 1*. Time plots of the components, q, q', α and β, of the filter residual with origin θ.*

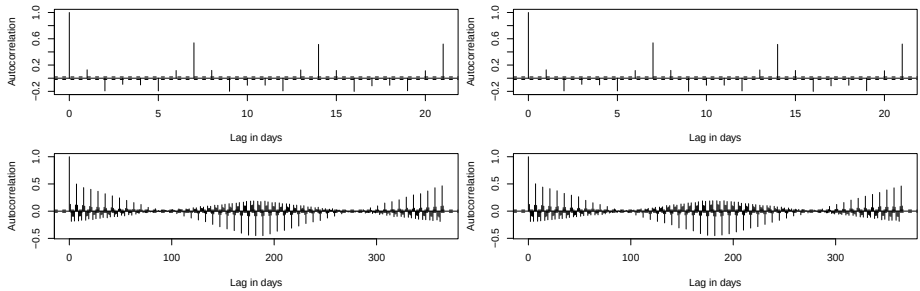
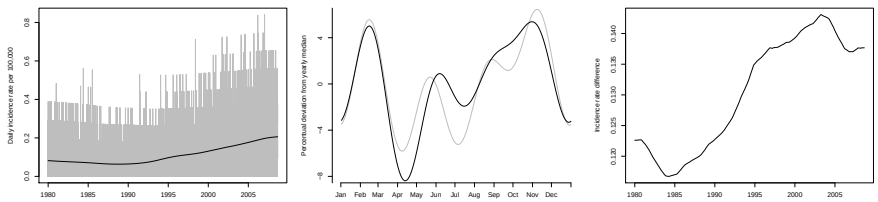
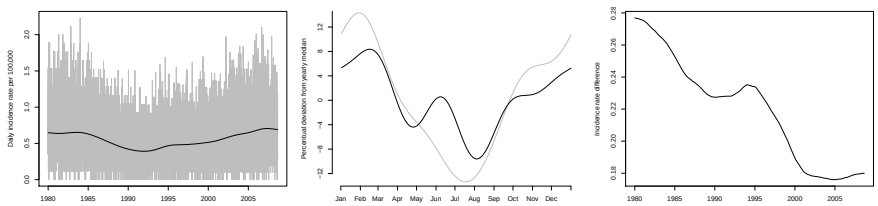


Figure B.16: *Residual analysis of Model* 1*. Autocorrelation plots of the components, q, q', α and β, of the filter residual with origin θ.*



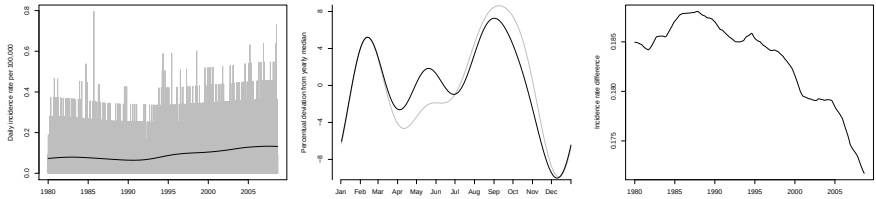
(a) (b) (c)

Figure B.17: *Estimated components of Model* 2 *for female aged* 20-49*. (a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*



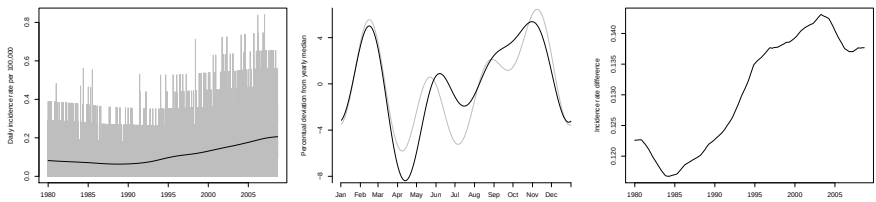
(a) (b) (c)

Figure B.18: *Estimated components of Model* 2 *for females aged* 50+*. (a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*



(a) (b) (c)

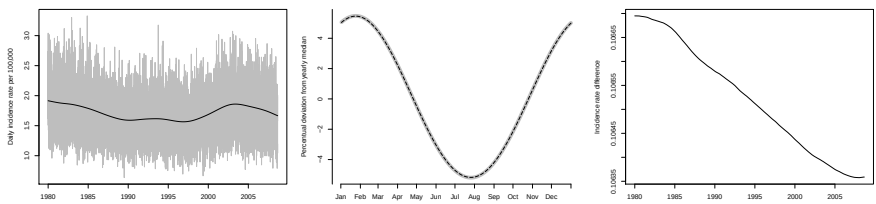
Figure B.19: *Estimated components of Model* 2 *for males aged* 20-49. *(a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*



(a) (b) (c)

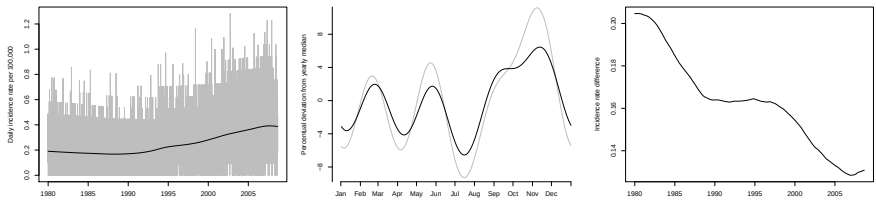
Figure B.20: *Estimated components of Model* 2 *for males aged* 50+. *(a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*

B.4 Cardio



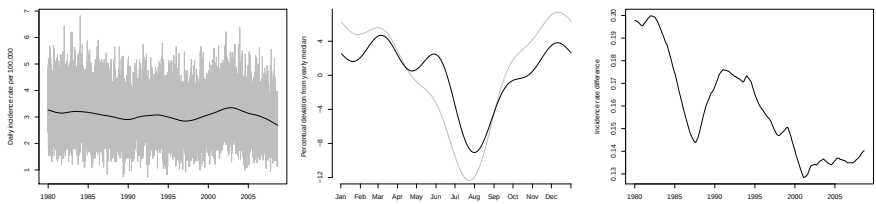
(a) (b) (c)

Figure B.21: *Estimated components of Model* 1. *(a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*



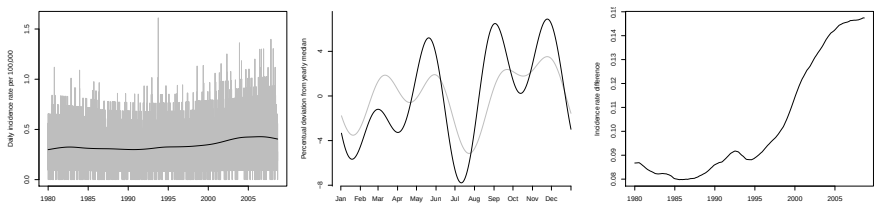
(a) (b) (c)

Figure B.22: *Estimated components of Model* 2 *for female aged* 20-49. *(a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*



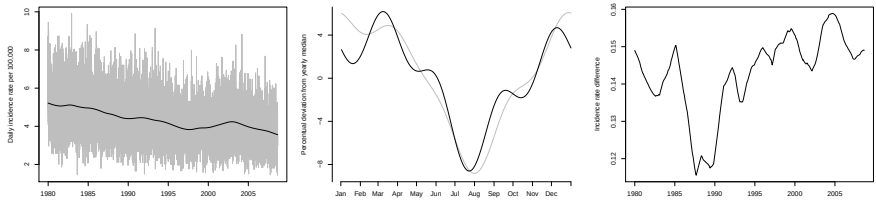
(a) (b) (c)

Figure B.23: *Estimated components of Model* 2 *for females aged* 50+. *(a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*



(a) (b) (c)

Figure B.24: *Estimated components of Model* 2 *for males aged* 20-49. *(a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*



(a) (b) (c)

Figure B.25: *Estimated components of Model* 2 *for males aged* 50+*. (a) Secular trend. (b) Seasonal variation component determined by January first for* 1980 *(grey curve) and* 2008 *(black curve). (c) Daily incidence rate differences.*

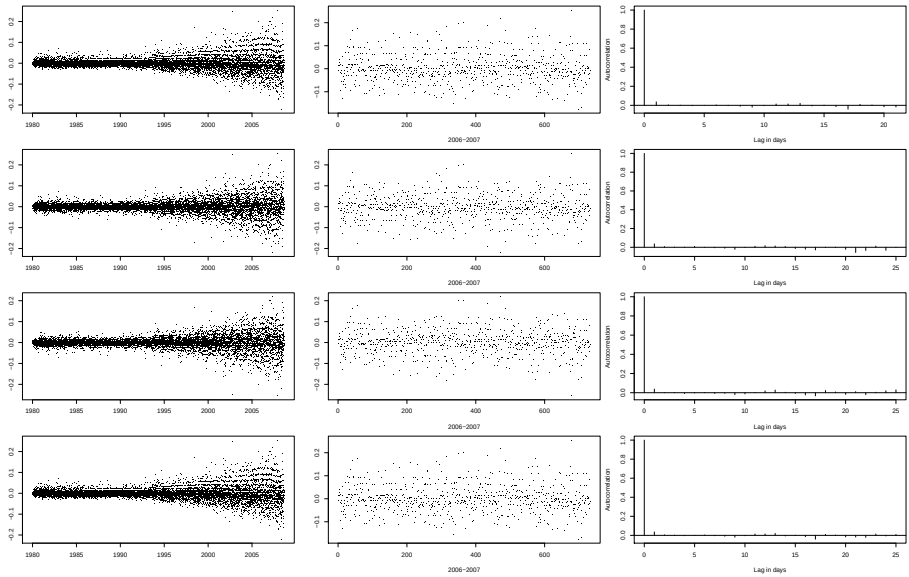


Figure B.26: *Residual analysis of Model* 2 *for females aged* 20-49*. Time- and autocorrelation plots of the components,* q*,* α_1 *and* β_1*, and* γ*, of the filter residual with origin* θ*.*

Appendix C

Nomenclature

$\mathbb{1}[t = n]$	Indicator function taking value 1 when expression is true and 0 when false
0	Matrix of zeros
1	Matrix of ones
I_p	$(p \times p)$ identity matrix
$\mathbb{R}$	Real numbers
$\mathbb{R}^n$	Real vector space of n-dimensional real vectors
$\mathbb{Z}$	Integers
$\mathbb{Z}_+$	Positive integers including zero
$\mathbb{N}$	Natural numbers, $1, 2, \ldots$
$A^\top$	Transpose of a real matrix A
A^{-1}	Inverse of a real matrix A
$\text{trace}(A)$	Trace of a real matrix A
$\text{diag}(a)_{p\times d}$	$(p \times d)$ matrix with the elements of a at the diagonal and zeros at the off-diagonal
$\lvert A \rvert$	Determinant of a real matrix A
$f'(x)$	Differentiation of f with respect to x
$\bar{Z}$	Sample mean
$\gamma(t, s)$	Autocovariance function
$\rho(\lvert \Delta t \rvert)$	Autocorrelation function
g_k	kth sample autocovariance coefficient
r_k	kth autocorrelation coefficient

State Space Models

D_t	Information set at time t
θ_t	Latent state at time t
Y_t	Observation at time t
ν_t	Observation error at time t
ω_t	Evolution error at time t
F_t	Design matrix at time t
V_t	Observation variance matrix at time t
G_t	Evolution transfer matrix at time t
W_t	Evolution variance matrix at time t
f_t	One-step forecast mean at time t
Q_t	One-step forecast variance matrix at time t
e_t	One-step forecast error at time t
A_t	Adaptive matrix at time t
K_t	Scaled adaptive coefficient matrix at time t
m_t	Filtered mean at time t
C_t	Filtered variance matrix at time t
$\{F, G, V, W\}_t$	Quardruple defining a state space model at time t
$\vec{m}_k$	k-step forecast mean of θ_{n+k} at time $n+k$
$\vec{C}_k$	k-step forecast variance matrix of θ_{n+k} at time $n+k$
$\vec{f}_k$	k-step forecast mean of Y_{n+k} at time $n+k$
$\vec{Q}_k$	k-step forecast variance matrix of Y_{n+k} at time $n+k$
$\tilde{m}_t$	Smoothed mean of θ_t at time t
$\tilde{C}_t$	Smoothed variance matrix of θ_t at time t
$\tilde{\omega}_t$	Smoothed evolution disturbance at time t
$\tilde{\nu}_t$	Smoothed observation disturbance at time t
v_t	Filter residual with origin Y at time t
w_t	Filter residual with origin θ at time t
$\tilde{v}_t$	Smoother residual with origin Y at time t
$\tilde{w}_t$	Smoother residual with origin θ at time t
$\vec{v}_t$	Forecast residual with origin Y at time t
$\vec{w}_t$	Forecast residual with origin θ at time t
$r_t^\bullet$	Standardised residual at time t

Distributions

$p(\cdot)$	Generic notation for the density- or probability function of arguments
$p(\cdot\|\cdot)$	Generic notation for conditional density function
$[\mu, V]$	Partially specified distribution with mean vector μ and variance matrix V
$\mathcal{N}_n(\mu, \Sigma)$	Multivariate normal distribution of dimension n with mean vector μ and variance matrix Σ
$\chi^2(n)$	Chi square distribution with n degrees of freedom
$L(\cdot)$	Generic notation for likelihood function
$l(\cdot)$	Generic notation for log likelihood function
$\sim$	Distributed as
$\dot{\sim}$	Approximately distributed as
$\hat{\sim}$	Partial distributed derived by using Linear Bayes' estimate with respect to a quadratic loss function
$A \perp\!\!\!\perp B$	A is independent of B
$A \perp\!\!\!\perp B\|C$	Conditional on C, A is independent of B
$\mathbb{E}[\cdot]$	Expected value of argument
$\mathbb{E}[\cdot \mid \cdot]$	Conditional expected value of argument
$\mathbb{V}\mathrm{ar}[\cdot]$	Variance of argument
$\mathbb{V}\mathrm{ar}[\cdot \mid \cdot]$	Conditional variance of argument
$\mathbb{C}\mathrm{ov}[\cdot, \cdot]$	Covariance of arguments
$\mathbb{C}\mathrm{ov}[\cdot, \cdot \mid \cdot]$	Conditional covariance of arguments
$\forall$	For all

Bibliography

S. Altizer, A. Dobson, P. Hosseini, P. Hudson, M. Pascual, and P. Rohani. Seasonality and the Dynamics of Infectious Diseases. *Ecology Letters*, 9:467–484, 2006.

E. Andersen, S. Bertelsen, E. Bindseil, O. Bonnevie, P. Bretlau, P. Christoffersen, E. Dickmeis, F. Gjerris, S. Haunsø, T. Heckscher, P. Hertoft, L. Heslet, N. Holm-Nielsen, P. Halm-Nielsen, H. Høyer, H. Iversen, H. G. Jespersen, J. P. Kampmann, H. Kirk, H. E. Larsen, J. F. Larsen, B. Lund M. Mellergård, N. Michelsen, J. Monrad, L. Mosekilde, S. Nørby, O. B. Paulson, E. B. Pedersen, K. D. Pedersen, H. Permin, K. Rasmussen, I. Sewerin, A. K. Sjølie, P. Skinhøj, F. Stadil, K. Stengaard-Pedersen, and K. Thestrup-Pedersen. *Klinisk ordbog*. Munksgaard Danmark, 15. edition, 2002. ISBN 8762800582.

T. F. Andersen, M. Madsen, J. Jørgensen, L. Mellemkjær, and J. H. Olsen. The Danish National Hospital Register - A valuable Source of Data for Modern Health Sciences. *Danish Medical Bulletin*, 46(3):263–268, 1999.

A. Azzalini. *Statistical Inference Based on the Likelihood*. Chapman & Hall/CRC, 1. edition, 1996. ISBN 041260650X.

H. Bounameaux, L. Hicklin, and S. Desmarais. Seasonal Variation in Deep Vein Thrombosis. *British Medical Journal*, 312:284–285, 1996.

M. A. Brookhart and K. J. Rothman. Simple Estimators of the Intensity of Seasonal Occurrence. *BMC Medical Research Methodology*, 8:67–75, 2008.

C. K. Carter and R. Kohn. On Gibbs Sampling for State Space Models. *Biometrika*, 81(3):541–553, 1994.

M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie. A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation. *Journal of Chronic Diseases*, 40(5):373–383, 1987.

A. L. Christensen. *Seasonal Variation of Cardiovascular Diseases - a Danish Nationwide Cohort Study*. 2008. Project written at Department of Mathe-

matical Sciences, Aalborg University in cooperation with Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital. Available at `http://people.math.aau.dk/~luther/Mat5.pdf`.

A. L. Christensen, C. Dethlefsen, and S. Lundbye-Christensen. Comparison of Geometrical Models and Poisson Regression Modelling Seasonal Variation - a Simulation Study. *Computer Methods and Programs in Biomedicine*, 2009a. Submitted in April.

A. L. Christensen, M. T. Severinsen, C. Dethlefsen, S. Lundbye-Christensen, and S. R. Kristensen. Occurrence of Seasonal Variation in Incident Unprovoked Venous Thromboembolism in the Danish Population. Poster from "Forskningens Dag", attained at Aalborg Hospital, Aarhus University Hospital, April 2009b. Available at `http://people.math.aau.dk/~luther/poster.pdf`.

M. H. DeGroot. *Probability and Statistics*. Addison Wesley, 2. edition, 1989. ISBN 020111366X.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood From Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

C. Dethlefsen. *Space Time Problems and Applications*. PhD thesis, Aalborg University, 2001.

C. Dethlefsen and S. Lundbye-Christensen. Formulating State Space Models in R with Focus on Longitudinal Regression Models. *Journal of Statistical Software*, 16(1), 2006.

C. Dethlefsen, B. Klein, and H. Thomsen. *State Space Models and Kalman Filtering*. 1997. Master Thesis written at Department of Mathemetics, Institute for Electronic Systems, Aalborg University.

P. J. Diggle. *Time Series. A Biostatistical Introduction*. Oxford Science Publications, 1. edition, 1990. ISBN 0198522266.

S. F. Dowell and M. S. Ho. Seasonality of Infectious Diseases and Severe Acute Respiratory Syndrome — What We Don't Know Can Hurt Us. *The Lancet Infectious Diseases*, 4:704–708, 2004.

J. H. Edwards. The Recognition and Estimation of Cyclic Trends. *Ann. Hum. Genet. Lond.*, 25:83–86, 1961.

P. H. C. Eilers, J. Gampe, B. D. Marx, and R. Rau. Modulation Models for Seasonal Time Series and Incidence Tables. *Statistics in Medicine*, 27:3430–3441, 2008.

J. M. Elwood and J. Little. *Seasonal Variation in Epidemiology and Control of Neural Tube Defects*. Oxford Medical Publications, 1. edition, 1992. ISBN 0192618849.

D. F. Findley, B. C. Monsell, W. R. Bell, M. C. Otto, and B. C. Chen. New Capabilities and Methods of the X-12-ARIMA Seasonal Adjustment Program. *Journal of Business and Economic Statistics*, 16(2):127–176, 1998.

T. Fischer, S. Lundbye-Christensen, S. P. Johnsen, H. C. Schønheyder, and H. T. Sørensen. Secular Trends and Seasonality in First-Time Hospitalization for Acute Myocardial Infarction — a Danish Population-Based Study. *International Journal of Cardiology*, 97(3):425–431, 2004.

T. Fischer, S. P. Johnsen, L. Pedersen, D. Gaist, H. T. Sørensen, and K. J. Rothman. Seasonal Variation in Hospitalization and Case Fatality of Subarachnoid Heamorrhage - A Nationwide Danish Study on 9,367 Patients. *Neuroepidemiology*, 24:32–37, 2005.

D. N. Fisman. Seasonality of Infectious Diseases. *Annual Review Public Health*, 28:127–143, 2007.

C. E. Frangakis and R. Varadhan. Confidence Intervals for Seasonal Relative Risk with Null Boundary Values. *Epidemiology*, 13:734–737, 2002.

L. Frost, L. V. Andersen, L. S. Mortensen, and C. Dethlefsen. Seasonal Variation in Stroke and Stroke-Associated Mortality in Patients with a Hospital Diagnosis of Nonvalvular Atrial Fibrillation or Flutter. *Neuroepidemiology*, 26:220–225, 2006.

F. Gao, K. Chia, I. Krantz, P. Nordin, and D. Machin. On the Application of the von Mises Distribution and Angular Regression Methods to Investigate the Seasonality of Disease Onset. *Statistics in Medicine*, 25:1593–1618, 2006.

Y. Gerber, S. J. Jacobsen, J. M. Kilian, S. A. Weston, and V. L. Roger. Seasonality and Daily Weather Conditions in Relation to Myocardial Infarction and Sudden Cardiac Death in Olmsted County, Minnesota, 1979 to 2002. *Journal of the American College of Cardiology*, 48(2):287–292, 2006.

P. J. Harrison and C. F. Stevens. Bayesian Forecasting. *Journal of the Royal Statistical Society. Series B*, 38(3):205–247, 1976.

D. Hewitt, J. Milner, A. Csima, and A. Pakula. On Edwards' Crition of Seasonality and a Non-parametric Alternative. *British Journal of Preventive and Social Medicine*, 25:174–176, 1971.

A. M. Joensen, M. K. Jensen, K. Overvad, C. Dethlefsen, E. B. Schimdt, L. Rasmussen, A. Tjønneland, and S. P. Johnsen. Predictive Values of Acute Coronary Syndrome Disharge Diagnoses Differed in the Danish National Patient Registry. *Journal of Clinical Epidemiology*, 62(2):188–194, 2009.

S. P. Johnsen, K. Overvad, H. Toft-Sørensen, A. Tjønneland, and S. E. Husted. Predictive Value of Stroke and Transient Ischemic Attack Discharge Diagnoses in the Danish National Registry of Patients. *Journal of Clinical Epidemiology*, 55:602–607, 2002.

B. Jørgensen, S. Lundbye-Christsensen, P. X. Song, and L. Sun. State-Space Models for Multivariate Longitudinal Data of Mixed Types. *The Canadian Journal of Statistics*, 24(3):385–402, 1996.

B. Jørgensen, S. Lundbye-Christsensen, P. X. Song, and L. Sun. A State Space Model for Multivariate Longitudinal Count Data. *Biometrika*, 86(1):169–181, 1999.

R. E. Kalman. *In Proceedings of the First Symposium on Engineering Application of Random Function Theory and Probability.* John Wiley and Sons, 1. edition, 1963.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82:35–45, 1960.

G. Kitagawa and W. Gersch. A Smoothness Priors - State Space Modeling of Time Series with Trend and Seasonality. *Journal of American Statistical Association*, 79(386):378–389, 1984.

B. M. Klein. *State Space Models for Exponential Family Data.* PhD thesis, Department of Statistics, University of Southern Denmark, 2003.

S. J. Koopman. Disturbance Smoother for State Space Models. *Biometrika*, 80 (1):117–126, 1993.

C. Ku, C. Yang, W. Lee, H. Chiang, C. Liu, and S. Lin. Absence of a Seasonal Variation in Myocardial Infarction Onset in a Region Without Temperature Extremes. *Cardiology*, 89:277–282, 1998.

P. M. Lee. *Bayesian Statistics an Introduction.* Hodder Arnold, 3. edition, 2004. ISBN 0340814055.

S. Lundbye-Christensen, C. Dethlefsen, A. Gorst-Rasmussen, T. Fischer, H. C. Schønheyder, K. J. Rothman, and H. T. Sørensen. Examing Secular Trends and Seasonality in Count Data Using Dynamic Generalized Linear Models: a New Methodological Approach Illustrated with Hospital Discharge Data on Myocardial Infarction. *European Journal of Epidemiology*, 24(5):225–230, 2009.

E. Manfredini, M. Gallerani, B. Boari, R. Salmi, and R. H. Mehta. Seasonal Variation in Onset of Pulmonary Embolism is Independent of Patients' Underlying Risk Comorbid Conditions. *Clinical and Applied Thrombosis/Hemostasis*, 10 (1):39–43, 2004.

J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of Royal Statistical Society*, 135(3):370–384, 1972.

N. K. Nissen and S. Rasmussen. *HjerteStatistik2008 - fokus på køn og sociale forskelle*. Hjerteforeningen, Statens Institut for Folkesundhed, Syddansk Universitet, 1. edition, 2008. ISBN 9788778991317.

J. P. Ornato, M. A. Peberdy, N. C. Chandra, and D. E. Bush. Seasonal Pattern of Acute Myocardial Infarction in the National Registry of Myocardial Infarction. *Journal of the American College of Cardiology*, 28:1684–1688, 1996.

D. P. Phillips, J. R. Jarvinen, I. S. Abramson, and R. R. Phillips. Cardiac Mortality is Higher around Christmas and New Year's than at any other time - The Holidays as a Risk Factor for Death. *Circulation*, 110:3781–3788, 2004.

S. J. Pocock. Harmonic Analysis Applied to Seasonal Variations in Sickness Absence. *Applied Statistic*, 23(2):103–120, 1974.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3900051070, `http://www.R-project.org`.

J. H. Roger. A Significance Test for Cyclic Trends in Incidence Data. *Biometrika*, 64(1):152–155, 1977.

K. Rothman. *Epidemiology: An Introduction*. Oxford University Press, USA, 1. edition, 2002. ISBN 9780195135541.

M. T. Severinsen, S. R. Kristensen, K. Overvad, C. Dethlefsen, A. Tjønneland, and S. P. Johnsen. Positive Predictive Value of Venous Thromboembolism Disharge Diagnoses in the Danish National Patient Registry. 2008. In preparation.

G. V. R. K. Sharma, J. H. Frisbie, D. E. Tow, S. V. Yalla, and S. F. Khuri. Circadian and Circannual Rhythm of Nonfatal Pulmonary Embolism. *American Journal of Cardiology*, 87:922–924, 2001.

K. Spengos, K. Vemmos, G. Tsivgoulis, E. Manios, N. Zakopoulos, M. Mavrikakis, and D. Vassilopoulos. Diurnal and Seasonal Variation of Stroke Incidence in Patients with Cardioembolic Stroke due to Atrial Fibrillation. *Neuroepidemiology*, 22:204–210, 2003.

C. Spielberg, D. Falkenhahn, S. N. Willich, K. Wegscheider, and H. Völler. Circadian, Day-of-Week, and Seasonal Variability in Myocardial Infarction: Comparison Between Working and Retired Patients. *American Heart Journal*, 132(7):579–585, 1996.

StataCorp. *Stata Statistical Software: Release 10.* StataCorp, College Station, TX: StataCorp LP, 2007. `http://www.stata.com/`.

Danmarks Statistik. Statistikbanken, 2008. `http://www.statistikbanken.dk/statbank5a/default.asp?w=1280`.

P. D. Stein, F. Kayali, and R. E. Olson. Analysis of Occurrence of Venous Thromboembolic Disease in the Four Seasons. *American Journal of Cardiology*, 93:511–513, 2004.

T. Thiele. Sur la Compensation de Quelques Erreurs Quasi-Systematiques par la Methode des Moindres Carrees. *Copenhagen: Reitzel*, 43, 1880.

L. E. Thorpe, T. R. Frieden, K. F. Laserson, C. Wells, and G. R Khatri. Seasonality of Tuberculosis in India: Is It Real and What Does It Tell Us? *Lancet*, 364(9445):1613–1614, 2004.

A. Tjønneland, A. Olsen, K. Boll, C. Stripp, J. Christensen, G. Engholm, and K. Overvad. Study Design, Exposure Variables, and Socioeconomic Determinants of Participation in Diet, Cancer and Health: A Population-based Prospective Cohort Study of 57,053 Men and Women in Denmark. *Scandivian Journal of Public Health*, 35:432–441, 2007.

D. E. Wallis, S. Penckofer, and G. W. Sizemore. The "Sunshine Deficit" and Cardiovascular Disease. *Circulation*, 118:1476–1485, 2008.

S. D. Walter and J. M. Elwood. A Test for Seasonality of Events with a Variable Population at Risk. *British Journal of Preventive and Social Medicine*, 29: 18–21, 1975.

D. A. Wehrung and S. Hay. A Study of Seasonal Incidence of Congenital Malformations in the United States. *British Journal of Preventive and Social Medicine*, 24:24–32, 1970.

M. West, P. J. Harrison, and H. S. Migon. Dynamic Generalized Linear Models and Bayesian Forecasting. *Journal of the American Statistical Association*, 80(389):73–83, 1985.

M. Zubaid, L. Thalib, and C. G. Suresh. Incidence of Acute Myocardial Infarction During Islamic Holiday Season. *European Journal of Epidemiology*, 21: 191–195, 2006.