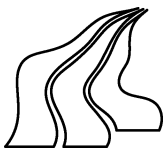
Detection of Disease Outbreaks
using State Space Models

Master of Science Thesis
Autumn 2011/Spring 2012
Tina Graungaard



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:
Detection of Disease Outbreaks
using State Space Models

SEMESTER:
Master of Science Thesis
Ninth and tenth semester

PROJECT PERIOD:
2/9/2011 – 1/6/2012

WRITTEN BY:

Tina Graungaard

SUPERVISOR:
P. Svante Eriksen

Number of copies: 6

Number of pages: 103

Finished: D. 1/6/2012

SYNOPSIS:
Disease surveillance is the systematic collection, analysis, interpretation and distribution of health data for preventing health related problems. The primary purpose of disease surveillance is early detection of disease outbreaks for prevention of further morbidity and mortality. In Denmark disease surveillance is carried out by Statens Serum Institut, which defines disease outbreaks as an unusual high number of incidences of a disease. The objective of this thesis is to compare different statistical models for prospective detection of possible outbreaks. Adjustments for seasonal variations, secular trends and past outbreaks should be incorporated into the model. Three different models are used: Farringtons algorithm, a dynamic linear model and a multi-process dynamic linear model. Comparison of the models is presented applying data from Statens Serum Institut consisting of all samples tested positive for Mycoplasma pneumoniae infections from July 1994 to July 2005. The analysis indicate that the dynamic linear model and the multi-process dynamic linear model are superior to Farringtons algorithm. The threshold value in Farringtons algorithm is highly affected by the baseline values used in the calculations, where the dynamic linear model and the multi-process dynamic linear model are better at adapting to the seasonal variations and past outbreaks. The multi-process dynamic linear model has the advantage that it can identify outliers.

Preface

This master of science thesis is written by Tina Graungaard in the period from September 2011 to June 2012. The thesis is composed at the Department of Mathematical Sciences, Aalborg University, in cooperation with Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital. The author would like to thank Center for Cardiovascular Research for making data available and the statisticians for their inputs and comments.
A basic knowledge equivalent to a bachelor degree in Mathematical Sciences at Aalborg University is required.

Reading instructions

References throughout the report will be presented according to the number method, and when a reference is placed before a period, it refers to the previous paragraph.
Figures, tables, mathematical definitions, etc. are enumerated in reference to the chapter i.e. the first figure in chapter 4 has number 4.1, the second has number 4.2 etc.

The project is divided into two parts: Analysis and theory. In the first part the problem of detection of disease outbreaks is outlined, and the analysis of the samples tested positive for Mycoplasma pneumoniae using three different methods, Farringtons algorithm, the dynamic linear model and the multi-process dynamic linear model, is presented. The first part is completed by a discussion and a conclusion of the results. In the second part the basic theory for the three methods is presented. First generalized linear models and in particular Poisson regression are presented. Then state space models and dynamic linear model are introduced, and the part is finished with a presentation of the multi-process model.

Mathematical notation and symbols

Throughout the rapport the following mathematical notation and symbols are applied.

$\mathbb{R}$	The set of all real numbers
$\mathbb{R}^n$	Real vector space of n-dimensional real vector
A^T	Transpose of a real matrix A
A^{-1}	Inverse of a real matrix A
$f(\cdot)$, $\pi(\cdot)$	The density- or probability function of arguments
$f(\cdot\|\cdot)$, $\pi(\cdot\|\cdot)$	Conditional density function of arguments
f'	The derivative of f
$N(\mu, V)$	Normal distribution with mean μ and variance V
$N_p(\mu, \Sigma)$	Multivariate normal distribution of dimension p with mean vector μ and variance matrix Σ
$Po(\mu)$	Poisson distribution with mean and variance μ
$\chi^2(p)$	Chi square distribution with p degrees of freedom
$\sim$	Distributed as
$L(\cdot)$	Likelihood function
$\ell(\cdot)$	Log-likelihood function
$U(\cdot)$	Score statistic
$\mathfrak{I}$	The information matrix
$\lambda(\cdot)$,$W(\cdot)$	The likelihood ratio
r_i	Residuals
D	The deviance
E[$\cdot$]	Expected value
Var[$\cdot$]	Variance
Var[$\cdot\|\cdot$]	Conditional variance
Cov[$\cdot, \cdot$]	Covariance
$(Y_t)_{t\geq 1}$	Time series
$y_{1:t}$	$y_1, y_2, \ldots, y_t$
Y_t	Observation at time t
θ_t	State at time t
v_t	Observation error at time t
w_t	Evolution error at time t
F_t	Design matrix at time t
G_t	Evolution matrix at time t
V_t	Observation variance matrix at time t
W_t	Evolution variance matrix at time t
e_t	Forecast error
$\tilde{e}_t$	Standard innovation
r	The signal-to-noise ratio

Dansk resumé

Overvågning af sygdomme involverer systematisk indsamling, dynamisk modellering, analyse og fortolkning af sundhedsrelateret data for at forebygge og kontrollere sygdomme, skader og andre sundhedsrelaterede problemer. Det primære formål med overvågning af sygdomme er at detektere udbrud og epidemier tidligt for at forebygge yderligere morbiditet og mortalitet. Overvågning af sygdomme udføres i Danmark af Statens Serum Institut, der definerer et udbrud som et unaturligt højt antal af inficerede af en bestemt sygdom.

Formålet med dette speciale er at sammenligne tre forskellige metoder til detektering af mulige udbrud. Metoder til automatisk detektering af mulige sygdomsudbrud skal tage højde for forskellige ting. Systematiske variationer som sæsonvariation og trend skal inkorporeres i modellen, og der skal tages højde for tidligere udbrud.

Den første metode blev præsenteret af Farrington et al. i 1996 og bruges af Statens Serum Institut til detektering af mulige sygdomsudbrud. Farringtons algoritme er baseret på en log-lineær regressionsmodel, som justeres for overspredning, sæsonvariation, trends og tidligere udbrud. En tærskelværdi udregnes baseret på baseline værdier fra tidligere år, og hvis det observerede antal er højere end denne tærskelværdi, markeres observationen som et muligt udbrud.

Den anden metode, der bruges, er en dynamisk lineær model, hvor det antages, at data er normalfordelt. Denne model bruger data fra tidligere uger til at forudsige et interval en tidsenhed frem, hvor det forventes at næste observation ligger indenfor. Hvis det observerede antal ligger over dette interval, markeres observationen som et muligt udbrud.

Den sidste metode er en multi-proces dynamisk lineær model, hvor det antages, at en enkelt dynamisk lineær model ikke kan beskrive data. I stedet for bruges tre dynamisk lineære modeller til at beskrive tre forskellige tilstande, som det antages, at data kan være i. De tre tilstande er stabil tilstand, hvor der ikke er udbrud, outlier, hvor en observation afviger uden der er udbrud og den tredje

mulighed er muligt sygdomsudbrud. Der bruges en første ordens Markov struktur til at beskrive overgangen mellem de tre forskellige modeller.

De tre metoder sammenlignes ved at analysere data fra Statens Serum Institut, som består af alle prøver, som er testet positiv for Mycoplasma pneumoniae infektioner fra juli 1994 til juli 2005. I denne periode er der to udbrud, som er identificeret af Statens Serum Institut. Analysen viser, at Farringtons algoritme påvirkes meget af hvilke baseline værdier, som bruges til udregning af tærskelværdien. Hvis baseline værdier falder sammen med tidligere udbrud, afspejles dette i tærskelværdien, og hvis der kun haves en begrænset mængde data, bliver sæsonvariation detekteret som mulige udbrud. Dette giver stor usikkerhed i, hvornår et muligt sygdomsudbrud detekteres. Den dynamiske lineære model og den multi-proces dynamisk lineære model er derimod bedre til at tilpasse sig sæsonvariationen og bliver ikke påvirket af tidligere udbrud. Begge modeller giver færre falsk positive alarmer end Farringtons algoritme. Det ser ud til, at den dynamiske lineære model detekterer udbruddene før multi-proces dynamisk lineære modellen. Multi-proces dynamiske lineære modellen har den fordel, at posterior sandsynligheder for de tre mulige tilstande fås, og derfor er muligt at identificere outliers.

Contents

Chapter 1

Introduction

Disease surveillance is the systematic collection, analysis, interpretation and distribution of health data for preventing and controlling disease, injury and other health related problems. In public health services disease surveillance has several purposes, where the primary purpose is the detection of disease outbreaks and epidemics. Early detection is important, because more effective disease intervention can prevent further morbidity and mortality [1]. Interventions could be removal of contaminated food, vaccination or preventively treating individuals at risk [2]. Disease surveillance can also give information about the natural development of diseases, for example how long the incubation period is, or it can be used to determine the size and range of an outbreak or epidemic. Finally disease surveillance can be used for evaluating and monitoring how interventions affect the public health [1],[3].

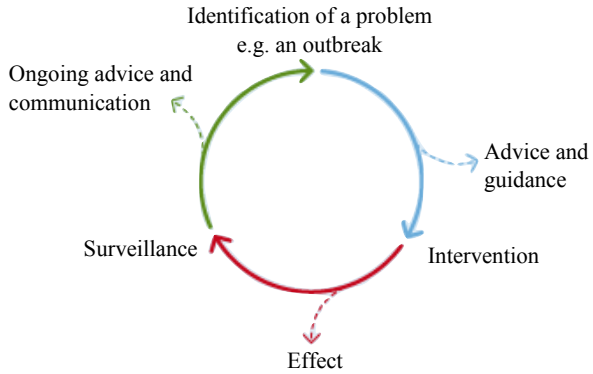


Figure 1.1: *Illustration of surveillance system [3].*

Figure 1.1 illustrates a surveillance system consisting of several components, that are part of a cycle. During the surveillance, data is collected and registered, then the data is analysed, and possible problems are identified, for example an outbreak. Afterwards, if necessary, interventions can be made, and new data can be collected for analysis to study the effects of the interventions. Not all diseases are under surveillance, only diseases of serious character, and diseases that can be prevented e.g. through vaccinations [3].

In disease surveillance one critical problem is the definition of an outbreak. The Centers for Disease Control and Prevention, which carry out disease surveillance throughout the United States, define an outbreak as two or more cases of infection, that are epidemiologically connected. This definition can be useful in retrospective analysis, when detailed epidemiological information is available, but in prospective analysis this definition is of little help [2]. In Denmark the surveillance is carried out by Statens Serum Institut under the Danish Ministry of Health [4]. Statens Serum Institut defines disease outbreaks as an unusual high number of incidences of a specific disease [5]. This definition, however, can be used for prospective detection of outbreaks, where the aim is to detect unusual high number of incidences as they occur. Thus an outbreak is first identified, when the number of reported infected is higher than an expected level. When the number of reported infected exceeds the expected level, it is called an aberration. Further epidemiological investigations have to be carried out to determine if it is an actual outbreak, or it is clusters caused by the reporting system [2].

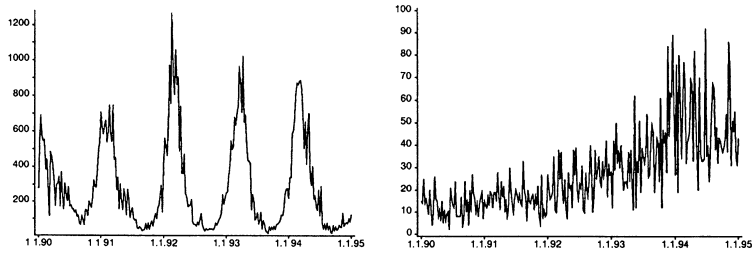
In prospective analysis the data must be analysed, as it is collected, and it is therefore not possible to accurately account for potential reporting delays as in retrospective analysis, where complete data often are available. Because it is difficult to determine the exact time of infection, the date of report is often used as a reference date. Alternatively the delay can be estimated and incorporated into the model, thereby introducing additional uncertainty. Thus it is important to reduce the variability of the reporting delay for example by standardizing the procedure for identification of infected individuals throughout the area under surveillance. Another problem in prospective analysis is validation of data because of the importance of early intervention for prevention of further morbidity or mortality. Therefore the uncertainty is higher in prospective analysis. The need for early interventions is also further complicated by the reporting delay [2].

During the last century the development in public health surveillance has been monumental, and the demand from diseases under surveillance exceeds the capabilities for manual scanning [1],[6]. Therefore a computer-aided detection system for detection of potential outbreaks is desired [6].

A computer-aided detection system has to fulfill several requirements for it to be reliable. One requirement is that the detection system has high sensitivity and a low false detection rate, where the sensitivity is given by the proportion of true outbreaks detected by the system, and the false detection rate is the proportion of false positives i.e. the proportion of observations marked as abnormal but not associated with outbreaks. If the detection system had low sensitivity or a high number of false positive the confidence in the system would be compromised. Reporting delays also affect the detection sensitivity and specificity of the system as well as the timeliness [2].

A robust algorithm is needed to handle a wide variety of organisms with differing epidemiologies and frequencies. In other words it must be capable of handling rare organisms with low frequency counts, and common organisms with high frequency count [2].

Disease surveillance can be applied to various diseases, but the most common area is infectious diseases, which also will be the focus of this project. Infectious diseases can go through systematic variations without it being an actual outbreak. The systematic variations can for example follow seasonal cycles or secular trends. The climate might influence the number of incidences, and some infectious diseases might peak in early winter or late winter/early spring [2]. The weekly counts reported to the Communicable Disease Surveillance Centre from 1990 to 1995 are shown in figure 1.2. In figure 1.2(a) the weekly counts for rotavirus are shown, which is a virus that can cause diarrhea in infants and young children. Rotavirus shows seasonal fluctuations that peak in the beginning of the year. Figure 1.2(b) shows the Clostridium difficile, which is an infection in the intestinal system. Clostridium difficile shows a slight trend during the period [6],[7].



(a) Rotavirus (b) Clostridium difficile

Figure 1.2: *Weekly count reported to the Communicable Disease Surveillance Centre from 1990 to 1995 [6].*

Systematic variations should be incorporated into the system when calculating baselines and thresholds to reduce the false detection rate. Past aberrations or outbreaks also needs to be included in the model to reduce the false detection rate. The easiest way to do this is by omitting the data corresponding to past aberrations or outbreaks from baselines and thresholds calculations. Alternatively a weight function can be used, where data relating to past aberrations or outbreaks are down-weighted [2].

In 1996 Farrington et al. presented an algorithm for early detection of outbreaks of infectious diseases based on a log-linear regression model. The model accounts for overdispersion, seasonality, trends and past outbreaks. Historical data is used to calculate a threshold value, and if the observed count is above this threshold it is declared an aberration [6]. Farringtons algorithm is a widely used algorithm for detection of disease outbreaks, and is currently being used by Statens Serum Institut in Denmark to monitor the gastrointestinal pathogens Salmonella, Campylobacter, Yersinia enterocolitica, Shigella and E. coli [8],[9]. In England and Wales Farringtons algorithm is used by the Health Protection Agency to detect outbreaks in laboratory-based surveillance data [10].

In 2006 Cowling et al. compared three different methods for monitoring influenza surveillance data. The focus was to find a valid and reliable way to detect the onset of a peak season, which did not require more than 9 weeks of baseline data. The first method was a dynamic linear model, which is a special case of a state space models. This model uses the previous information to calculate a forecast interval, and if the observed count falls outside this interval, then the count is identified as an aberration. The second method is a regression model, where a forecast interval is calculated based on the normal distribution from the preceding weeks. The third method is a cumulative sum method, CUSUM, where the prediction error from the past d weeks is summed up, and if it exceeds a predefined threshold, it is defined an aberration. The comparison is made using data from Hong Kong and the United States, where the dynamic linear model was superior to the other models in the data from Hong Kong, and in the data from the United States the dynamic linear model and the CUSUM method performed similarly but better than the regression model. Thus the dynamic linear model is the preferred method of the three [11].

In 1983 Smith et al. used a multi-process dynamic linear model for monitoring renal transplants. The interest was in developing an on-line statistical procedure for monitoring the kidney function of patients who had received kidney transplants, specially changes that indicated rejection of the transplant. They assumed that the system could be in different states: Steady state, changes in the system or outlier [12]. Similarly to the model presented by Smith et al. Whittaker et al. presented a dynamic change-point model for detecting the

onset of growth in bacteriological infections. They used an on-line decision procedure to determine whether bacteriological infections were present in feedstuff [13].

Aim of the thesis

The aim of this thesis is to compare different methods for detection of outbreaks. The first method is the algorithm presented by Farrington et al. and is currently being used by Statens Serum Institut for detection of aberrations [6],[8]. The second method is a dynamic linear model equivalent to that presented by Cowling et al. [11]. The last method is a multi-process dynamic linear model, which is similar to the model presented by Smith et al. and the change-point model by Whittaker et al. for detection of abrupt changes in patterns [12],[13].

The comparison is presented applying data from Statens Serum Institut in Denmark consisting of all samples tested positive for Mycoplasma pneumoniae infections from July 1994 to July 2005.
Mycoplasma pneumoniae is the cause of a broad spectrum of respiratory infections. Incidences occur all year but is most frequent in the fall and the winter. In Denmark outbreaks occur every four to six years, where they typically begin slowly during the late summer and have a duration of 3 to 4 months. Mycoplasma pneumoniae is diagnosed by detection of Mycoplasma pneumoniae DNA in respiratory secretion. It is not possible to prevent Mycoplasma pneumoniae infections, but they can be limited by treatment and isolation of infected individuals [14].

Part I

Analysis: Detection of Disease Outbreaks

Chapter 2

Materials and Methods

In this chapter the materials and methods of analysis are described. First the data preprocessing is presented, then the method described by Farrington et al., Farringtons algorithm, which is currently being used by Statens Serum Institut, is introduced [6],[8]. Furthermore a dynamic linear model and a multi-process dynamic linear model are presented.

2.1 Mycoplasma pneumoniae

Mycoplasma pneumoniae is a microorganism that causes of a broad spectrum of respiratory infections e.g. pneumonia, bronchitis and infections in the upper respiratory system. The transmission of the microorganism occur in areas with many people, and the incubation time is about 2 to 3 weeks. The symptoms are dry cough, fever, headache, sore throat, rash and ear complications. No effective vaccine exist, but the duration of the disease can be shortened by treatment e.g. antibiotics [15]. Mycoplasma pneumoniae infections occur all year, but is most common in the fall and early winter. Epidemics occur every 4 to 6 years and the extent varies with average about 3 to 4 months [14].
There were two outbreaks of Mycoplasma pneumoniae in the period Juli 1st 1994 to Juli 29th 2005. The first outbreak was in 1998/1999, and the second outbreak was in 2004/2005 [16],[17].

2.2 Data preprocessing

The data was received as a csv file containing 4047 observations obtained daily from Juli 1st 1994 to Juli 29th 2005. Each observation consist of the observation

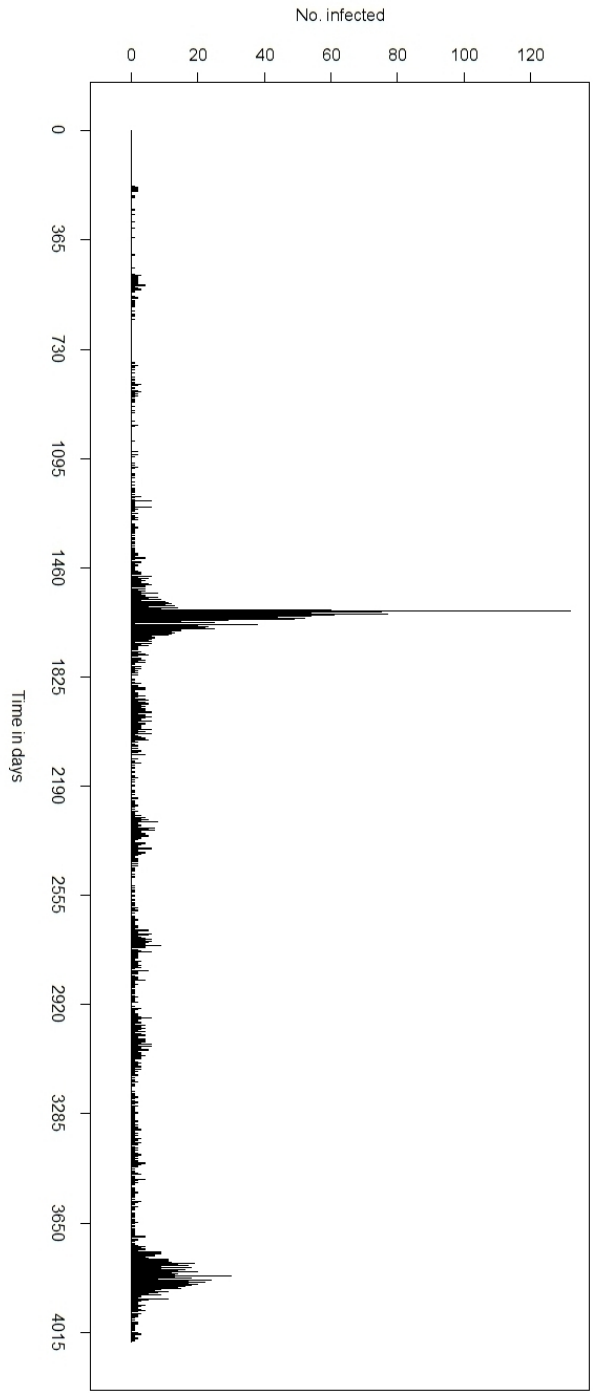


Figure 2.1: *Number of infected in days*

number, the date, and the number of samples tested positive for Mycoplasma pneumoniae the current day. The data analysis using Farringtons algorithm was preformed using R 2.10.0 and the R-package *surveillance* [18], which includes Farringtons algorithm. The analysis using the dynamic linear model and the multi-process dynamic linear model were carried out using R 2.15.0 and the R-package *dlm* [19].

The number of infected each day is shown in figure 2.1. The days are numbered from 1 to 4047 i.e. number 1 correspond to Juli 1st 1994 and so on. The total number of infected is 5295. The distribution of the number of infected over the week is shown in figure 2.2, which shows that during the weekend the number of infected drops compared to Monday to Friday.

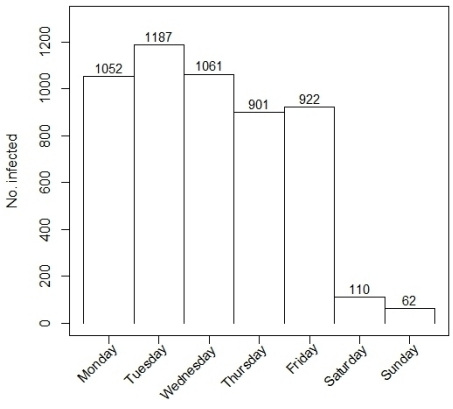


Figure 2.2: *Number of samples recorded on each day.*

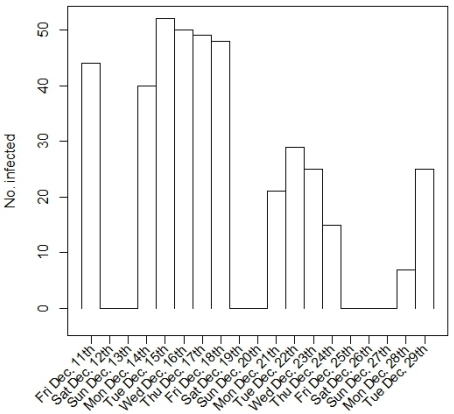


Figure 2.3: *Number of infected.*

In figure 2.3 the number of infected from Friday December 11th 1998 to Tuesday December 29th 1998 is shown. The figure illustrates that there are no observations in the weekend. Therefore the data is aggregated from daily counts to weekly counts, where each week is from Thursday to Wednesday. In Farringtons algorithm and the dynamic linear model it is assumed, that there are 52 weeks in a year, i.e. it does not account for leap years and years with 53 weeks. In the multi-process dynamic linear model it is assumed that there is $\frac{7}{365.25}$ weeks in a year. The number of infected each week is shown in figure 2.4. The first year is used for baseline calculations for all three methods, and it is therefore omitted through the rest of the report, which means that week 1 through the rest of the report correspond to the week beginning June 29th 1995.

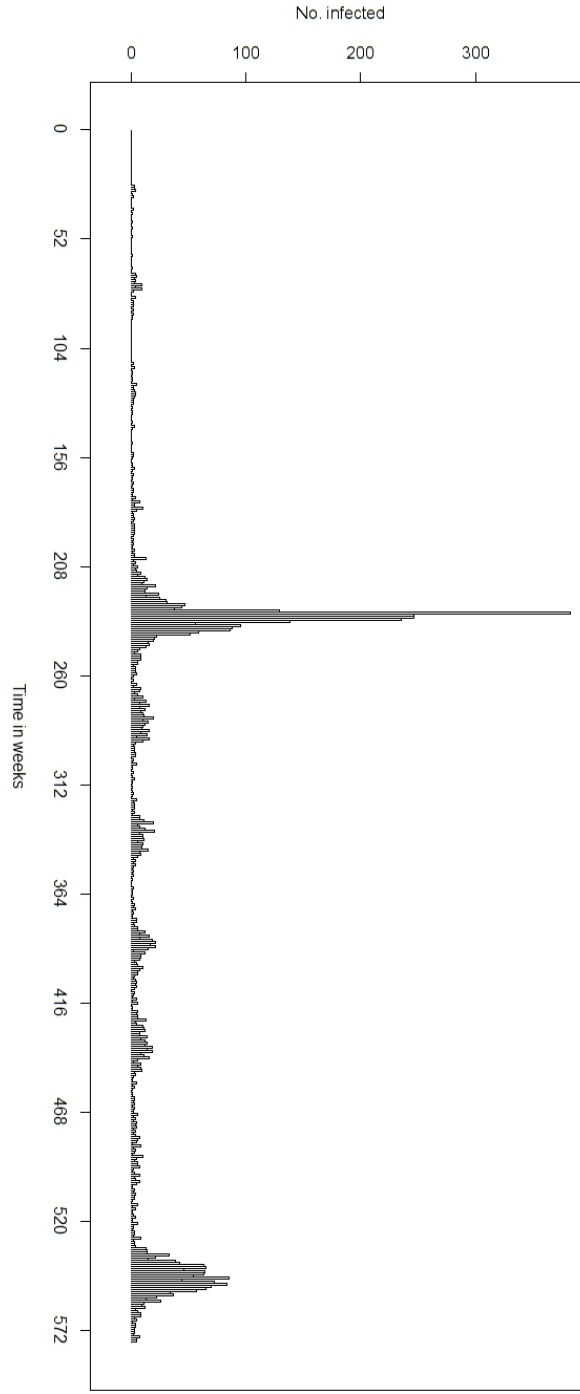


Figure 2.4: *Number of infected in weeks*

Figure 2.5 is a frequency plot of the observations, which shown that the distribution is skewed. Four observations, where the number of observed is higher than 140, are omitted from the figure. The weeks omitted are week 230 to 233, where the number of infected is 383, 246, 246, and 235, respectively.

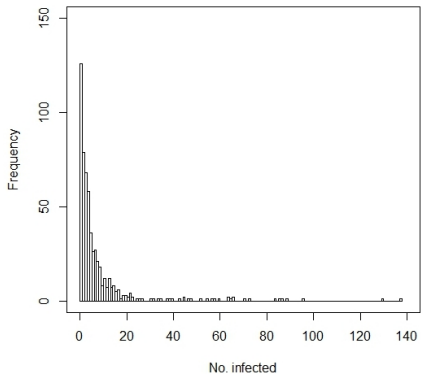


Figure 2.5: *Frequency of observations*

2.3 Farringtons algorithm

In this section the first method for analysis is presented. The described method is a generalized linear model or more specifically a log-linear regression model presented by Farrington el al. in 1996 [6].

Farringtons algorithm is an algorithm for epidemiological surveillance and was developed to assist in early detection of outbreaks of infectious diseases [6]. This algorithm is being used by Statens Serum Institut in Denmark for monitoring of the gastrointestinal pathogens pathogens Salmonella, Campylobacter, Yersinia enterocolitica, Shigella, and E. coli [8]. Farringtons algorithm is also being used by the Health Protection Agency in England and Wales to detect outbreaks in laboratory-based surveillance data [10].

The primary purpose is to detect outbreaks early enough to have time for intervention. The algorithm must take seasonal cycles, secular trends and past outbreaks into consideration, and it must be sufficiently robust to handle a wide range of different diseases. Data collected for surveillance systems are often subject to bias and delays in reporting. This makes the use of such data problematic for early detection of outbreaks [6].

Seasonal variations affect many diseases, and the number of affected individuals may peak at different times of the year or show long-term trends. The primary interest of the surveillance system is to detect increases greater than the seasonal

variability and the trends. Some variation is not of primary interest such as abnormally low counts, or if the count is unusual high but does not constitute an outbreak [6].

A routine scanning system has to fulfill different requirements such as timeliness, sensitivity, specificity, and the output has to be easy to interpret. The statistical features of the system is determined by these requirements, and one algorithm that can analyse all diseases is developed [6].

2.3.1 Model structure

A flexible algorithm that takes seasonal patterns, underlying trends and noise in the data into account is designed. This is done by developing a log-linear regression model which is adjusted for overdispersion, seasonality, secular trend and past outbreaks. The model is used to calculate a threshold value, where it is expected that the next observed count is below. If the observed count for the next observation is above the threshold value, the count is considered unusual [6].

There will occur delays between the time of infection and when it is reported because it takes time to diagnose diseases. Since it is difficult to determine the exact time of infection the date of report is used as reference date. Trends are taken into account in the model by fitting a linear time variable in the regression model, and seasonality is considered by calculating the threshold value based on comparable baseline periods from previous years [6].

Baseline

The baseline periods in weeks are calculated by letting b be the number of years back in time and w be half of the width of a chosen window. The present week is denoted x of year y, and data for weeks $x - w$ to $x + w$ of years from $y - b$ to $y - 1$ is used, which gives $n = b(2w + 1)$ baseline weeks. The value of n affect the precision, the need for a high n value for high precision, must be considered in relation to the width of the window and the seasonal variations [6].

Regression model

Let y_i denote the baseline count connected to baseline week t_i, which is assumed to be distributed with mean μ_i and variance $\phi\mu_i$, and the baseline values are assumed to be independent of each other. If the frequency of the disease is low the assumption of independence between the baseline counts is expected, but for diseases of high frequency correlation between baseline counts are expected. Farrington et al. examined the correlation between baseline values for organisms

with high frequency count and found that is has a small effect on the threshold values. Correlation between baseline counts is therefore not included in the model.

The systematic component of the model is given by

$$\log(\mu_i) = \alpha + \beta t_i,$$

where trend is the only effect included, and the estimates $\widehat{\alpha}$ and $\widehat{\beta}$ are fitted using Poisson regression. If historical data is available for minimum 3 years, if β is significant at the 5% level and if

$$\widehat{\mu}_0 \leq \max\{y_i | i = 1, \ldots, n\},$$

then the linear time trend is included in the model.

The dispersion parameter ϕ is estimated by a quasi-likelihood method

$$\widehat{\phi} = \max\left\{\frac{1}{n-p}\sum_{i=1}^{n}\omega_i\frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}, 1\right\},$$

where ω_i is a weight, which will be explained later, and $p = 1$ if no time trend is fitted, or $p = 2$ if a time trend is fitted [6]. A quasi-likelihood method is a method that allows for overdispersion when using the Poisson distribution by specifying a variance function depending on the mean value. The introduction of the dispersion parameter allows the variability to be larger than the expected variability [20, pp. 258-260]. The expected count is estimated by

$$\widehat{\mu}_0 = \exp\left(\widehat{\alpha} + \widehat{\beta} t_0\right) s,$$

where t_0 is the current week and y_0 is the current count of infected individuals. It is assumed that the frequency count of infected individuals is Poisson distributed if the frequency count is small. For more frequent types of infections the data is assumed to be normal distributed.

Farrington et al. [6] tested the model using simulated data, where overdispersion is generated by assuming that the Poisson mean varies according to a gamma distribution with mean μ and variance $\mu(\phi - 1)$. This is the same as assuming that the data is negative binomial distributed with mean μ and variance $\phi\mu$ [6].

Threshold

The distribution is highly skewed for diseases with low frequency counts, which have to be taken into account in the calculation of the threshold value. To correct for the skewness a $\frac{2}{3}$-power transformation of the data is used, which gives a more symmetric distribution if the data is Poisson distribution, and the threshold

become more accurate. High frequency data remains almost unaffected by the transformation.

Given that the data is Poisson distributed and using the delta method, then

$$f(y_0) = y_0^{2/3}$$

can be approximated by

$$f(y_0) \approx f(\mu_0) + f'(\mu_0)(y_0 - \mu_0),$$

and the mean value of $f(y_0)$ is given by

$$\mathrm{E}\left[y_0^{2/3}\right] = \mu_0^{2/3}.$$

The variance of $f(y_0)$ is

$$\begin{aligned} \mathrm{Var}\left[y_0^{2/3}\right] &= f'(\mu_0)^2 \mathrm{Var}\left[y_0\right] \\ &= \left(\frac{2}{3}\mu_0^{-1/3}\right)^2 \cdot \phi\mu_0 \\ &= \frac{4}{9}\phi\mu_0^{1/3}, \end{aligned}$$

and

$$\mathrm{Var}\left[\widehat{\mu}_0^{2/3}\right] = \frac{4}{9}\mu_0^{-2/3}\mathrm{Var}[\widehat{\mu}_0],$$

where $\mathrm{Var}[\widehat{\mu}_0]$ is given as the variance of the fitted Poisson regression. On the $\frac{2}{3}$-power scale the prediction error variance is given by

$$\mathrm{Var}\left[y_0^{2/3} - \widehat{\mu}_0^{2/3}\right] = \frac{4}{9}\tau\mu_0^{1/3},$$

where

$$\tau = \phi + \frac{\mathrm{Var}[\widehat{\mu}_0]}{\mu_0}.$$

Then an approximate $100(1-2\alpha)\%$ prediction interval (L, U) for y_0 is defined as

$$\begin{aligned} U &= \widehat{\mu}_0 \left\{1 + \frac{2}{3} z_\alpha \left(\frac{\widehat{\tau}}{\widehat{\mu}_0}\right)^{1/2}\right\}^{3/2}, \\ L &= \widehat{\mu}_0 \max\left\{\left\{1 - \frac{2}{3} z_\alpha \left(\frac{\widehat{\tau}}{\widehat{\mu}_0}\right)^{1/2}\right\}^{3/2}, 0\right\}, \end{aligned}$$

where z_α is the $100(1-\alpha)$-percentile of the normal distribution. If the frequency count of infected individuals is outside this interval it is considered as unusual, and if it is above the threshold U it is considered a possible outbreak [6].

Past outbreaks

When calculating the baseline count, the calculations are based on historical data. If there has been an outbreak in the historical data used, this must be considered in the model. If past outbreaks are included in the calculations, then the threshold will be to high and the sensitivity will be reduced. Past outbreaks are included in the model by using a reweighting procedure that reduces the influence of high baseline counts. Residuals are given by

$$s_i\left(\widehat{\phi}\right) = \frac{3}{2\widehat{\phi}^{1/2}} \frac{y_i^{2/3} - \widehat{\mu}_i^{2/3}}{\widehat{\mu}_i^{1/6}(1-h_{ii})^{1/2}}$$

where h_{ii} are the elements on the diagonal of the hat matrix. The hat matrix is a matrix that maps the vector of observed values into the vector of fitted values. If the data is Poisson distributed, where $\phi = 1$, the residuals are known as the standardised Anscombe residuals. The weights are given by

$$\omega_i = \begin{cases} \gamma s_i(1)^{-2} & \text{if } s_i > 1, \\ \gamma & \text{otherwise ,} \end{cases}$$

i.e. corresponding to $\widehat{\phi} = 1$, and where γ is a constant which satisfy $\sum \omega_i = n$. Empirical data is used to give low weights to counts with large residuals. The reweighting reduces the effect of past outbreaks, but it does not eliminate it [6].

The algorithm

When a new observation is available, the following algorithm is applied to the vector of counts. First an initial model is fitted, and the initial estimated $\widehat{\mu}_i$ and $\widehat{\phi}$ are calculated. Then the weights are calculated and the model is fitted once more. The dispersion parameter ϕ is estimated again, and the model is rescaled. The trend is left out if it is not significant and the procedure is repeated. The threshold value is computed using historical data.

The analysis is carried out with the following parameters. For baseline calculations the number of years back in time used are $b = 5$, if available, and half of the chosen window is $w = 3$, which gives maximum $n = 35$ baseline weeks. The data did not show a significant trend, which was therefore omitted. A $\frac{2}{3}$-power transformation was used for threshold calculations, and a weight function was used to reduce the influence of past outbreaks. To reduce the number of sporadic cases detected as possible outbreaks for organisms with low frequency counts, the restriction that the number of infected within the last 4 weeks has to exceed 5 for an alarm to occur, is implemented [6].

2.4 The dynamic linear model

In this section the dynamic linear model used in the analysis is presented. The model developed is based on the theory presented in section 7.1 on page 54. In the dynamic linear model it is assumed that the data is normally distributed. Because the data is count data, it is assumed to be Poisson distributed, therefore a $\frac{2}{3}$-power transform is applied to the data for skewness correction [21]. This is the same transform used for normalization in the threshold calculations in Farringtons algorithm.

Model structure

The dynamic linear model used is defined as the sum of two independent components: A random walk plus noise component and a seasonal component. A dynamic linear model, as defined in definition 7.2 on page 54, is then given by a normal prior distribution

$$\theta_0 \sim N(m_0, C_0),$$

and by the observation and state equations

$$\begin{aligned} Y_t &= F_t\theta_t + v_t \quad , \quad v_t \sim N(0, V) \\ \theta_t &= G_t\theta_{t-1} + w_t \quad , \quad w_t \sim N(0, W), \end{aligned}$$

where m_0, C_0, V and W are estimated using the observation in the steady periods without outbreaks, i.e. observation 0 to 200 and 250 to 500. The parameters in the model have the following parameters

$$\begin{aligned} m_0 &= \begin{bmatrix} 1.887749085 \\ 0.021707174 \\ 0.001223346 \end{bmatrix}, \\ C_0 &= \begin{bmatrix} 0.30877187 & -0.15352297 & 0.08173705 \\ -0.15352297 & 0.19556025 & -0.01283034 \\ 0.08173705 & -0.01283034 & 0.23292694 \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} F &= \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.9927089 & 0.1205367 \\ 0 & -0.1205367 & 0.9927089 \end{bmatrix}, \\ V &= 0.7947231 \text{ and } W = \begin{bmatrix} 0.0403078 & 0 & 0 \\ 0 & 0.005419615 & 0 \\ 0 & 0 & 0.005419615 \end{bmatrix}. \end{aligned}$$

A significant level of $\alpha = 0.01$ was used, and if an observation was above the 99% prediction interval when running the Kalman filter, then an alarm was given. If an observation was above the 99% prediction interval it was changed to *NA* in the data, and the Kalman filter was run again. The standard innovations is analysed to determine whether the observations in the steady periods are normally distributed for checking of the model assumptions.

2.5 The multi-process dynamic linear model

In this section the multi-process dynamic linear model used for analysis is presented. The model developed is based on the theory presented in section 8.2 on page 78. As for the dynamic linear model a $\frac{2}{3}$-power transform of the observation is used for skewness correction.

Model structure

A multi-process model class II is used for the analyses, where the probabilities with which the model is selected are first-order Markov. Three different model states are defined:

1. Steady state
2. Outlier
3. Possible outbreak

The three different states are modeled using a dynamic linear model, and the models are defined as the sum of two independent components: A random walk plus noise component and a seasonal component. The matrices F_t and G are for all three models given as

$$G = I_3 \quad \text{and} \quad F_t = \left[\begin{array}{ccc} 1 & \cos(2\pi \frac{7}{365.25} \cdot t) & \sin(2\pi \frac{7}{365.25} \cdot t) \end{array} \right].$$

The steady state model is defined as model 1 with variance parameters

$$V = 0.7947231 \quad \text{and} \quad W = \left[\begin{array}{ccc} 0.0403078 & 0 & 0 \\ 0 & 0.005419615 & 0 \\ 0 & 0 & 0.005419615 \end{array} \right],$$

which is the same parameters as in the dynamic linear model described in the last section.

The outlier model is defined as model 2, and it is assumed that the observation variance is 10 times the observation variance of the steady state model, but the state variance is the same. Thus the variance parameters are

$$V = 7.947231 \quad \text{and} \quad W = \begin{bmatrix} 0.0403078 & 0 & 0 \\ 0 & 0.005419615 & 0 \\ 0 & 0 & 0.005419615 \end{bmatrix}.$$

The model for possible outbreaks is defined as model 3. The variance parameters are the same as for the outlier model, but the two models differ in the transition probabilities, which are defined as

	Steady	Outlier	Outbreak
Steady	0.985	0.985	0.090
Outlier	0.010	0.010	0.010
Outbreak	0.005	0.005	0.900

.

The transition probabilities should be read as there is a 98.5% probability of staying in steady state if the previous model was steady state. If the previous state was outbreak, then there is a 9% probability that the new state is steady. The transition probabilities are chosen so 98.5% of the time series is in steady state.

For each time t the filtering distribution is approximated using the multi-process Kalman filter, proposition 8.2 presented on page 82, but instead of using mixture collapses the aim is to retain the most probable model sequences. The number of possible model sequences at time k is 3^k. Suppose that at time $t \geq k$ there are stored 3^k model sequences. Then for $t+1$ the likelihood for all 3^{k+1} possible model sequences are calculated, and the 3^k model sequences with highest likelihood are saved, i.e. if $k = 4$ then 81 model sequences are saved. So at a given time $t \geq k$ there are the model sequences

$$M_j = (\alpha_{1j}, \ldots, \alpha_{tj})$$

and their likelihoods L_j for $j = 1, \ldots, 3^k$. Let

$$I_m = \{j | \alpha_{tj} = m\} \quad , \quad m = 1, 2, 3,$$

then the posterior probability of model m at time t is approximated by

$$\widehat{\Pr}(\alpha_t = m) = \frac{\sum_{j \in I_m} L_j}{\sum L_j}.$$

The most likely model sequences are selected at each time, because of the large number of mixture components as time progress, which increases the complexity of the calculations, thus the components where the posterior probabilities are small are ignored.

Chapter 3

Results

In this chapter the results of the analysis using the three different methods are presented. Statens Serum Institut has identified two outbreaks of Mycoplasma pneumoniae in the period Juli 1st 1994 to Juli 29th 2005, the first in 1998/1999 and the second in 2004/2005 [16],[17]. The years 1998/1999 correspond to the week 132 to 208, and the years 2004/2005 correspond to the week 445 to 525, i.e. it is expected to detect an outbreak in each of these time periods. It is assumed, that there are no outbreaks in the weeks outside these periods, i.e. there are no outbreaks in week 1 to 131 and week 209 to 444. If there is an alarm in the periods with no outbreak, it is considered a false positive alarm.

Farringtons algorithm

In figure 3.2 the results of the analysis using Farringtons algorithm are presented, where the number of infected, the threshold and alarms are shown. In the first years there are a number of alarms indicating a possible outbreak without there being a high number of infected. As time passes more data become available for baseline values, thereby increasing the amount of data used for threshold calculations. The threshold is affected by the first outbreak up to 5 years after, even though a weight function is used to reduce the effect of past outbreaks by giving low weights to counts with high residuals. The threshold values during the second outbreak are not affected by the first outbreak, because the threshold is only calculated using historical data up to 5 years back.

Dynamic linear model

Figure 3.3 shows the result of the analysis using the dynamic linear model, indicating the number of infected, the threshold and the alarms. There are no

alarms in the first two years and only one false alarm before the period, where the first outbreak is expected. The threshold values are not affect by past outbreaks, and it easily adapt to the seasonal variations.

The dynamic linear model rely on the assumption that the observations are normally distributed. This can be examined by checking that the standard innovations in the steady periods are normally distributed. Figure 3.1 is a QQ-plot of the standard innovations in the steady periods, where it is shown, that the standard innovations deviate slightly from the normal distribution, which could indicate systematic deviation. This is, however, disregarded, and it is assumed that the standard innovations are normally distributed.

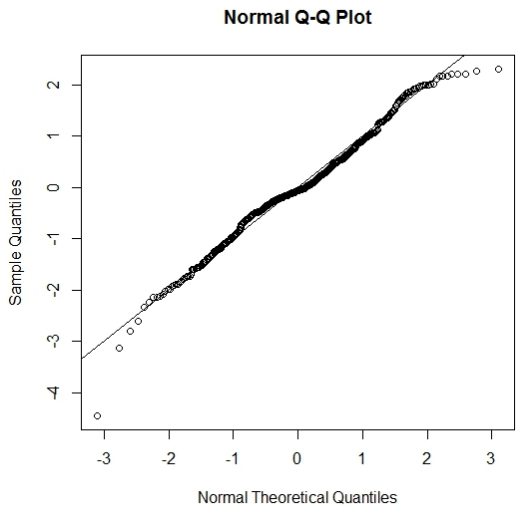


Figure 3.1: *QQ-plot of the standard innovations in the steady periods.*

Multi-process dynamic linear model

In figure 3.4 the results of the analysis using a multi-process dynamic linear model are presented. The most likely sequence of models is shown, where the number of infected, outliers and the alarms are indicated. This shows that there are two clusters of alarms corresponding to the two periods where outbreaks are expected. Two outliers are identified, where the first outlier in week 152 correspond to an unusual high number of infected that week, and the second in week 339 is an unusual low number of infected that week.

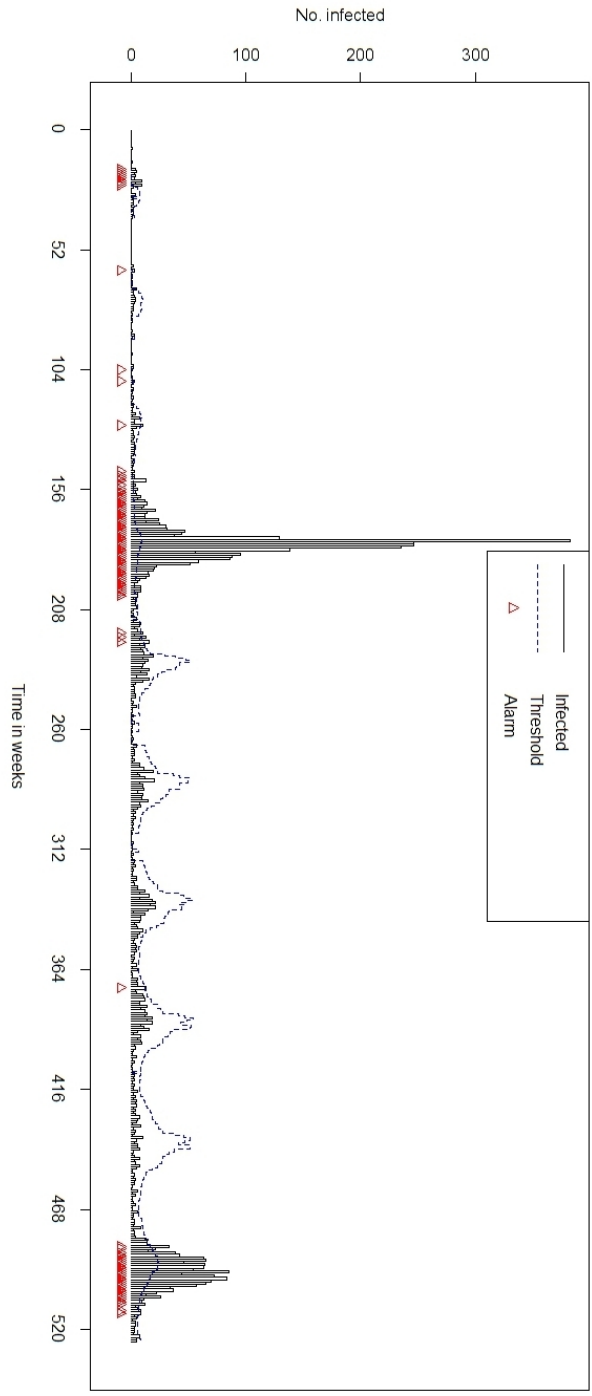


Figure 3.2: *Result of analysis using Farringtons algorithm.*

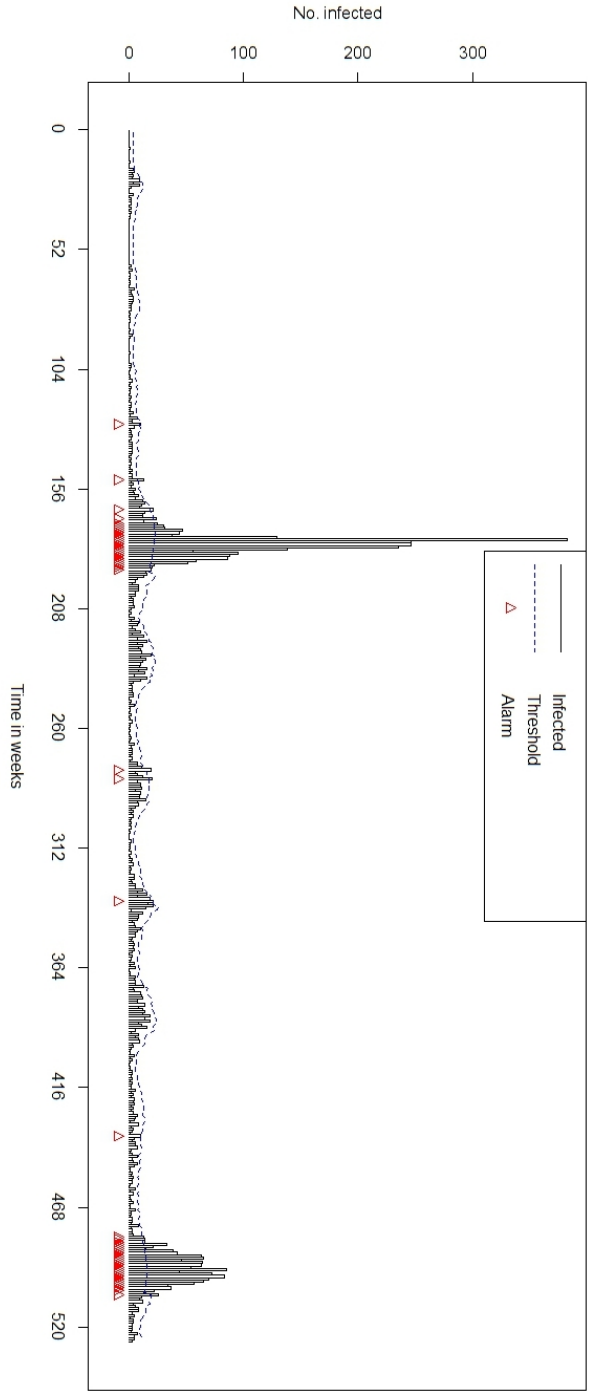


Figure 3.3: *Result of analysis using the dynamic linear model.*

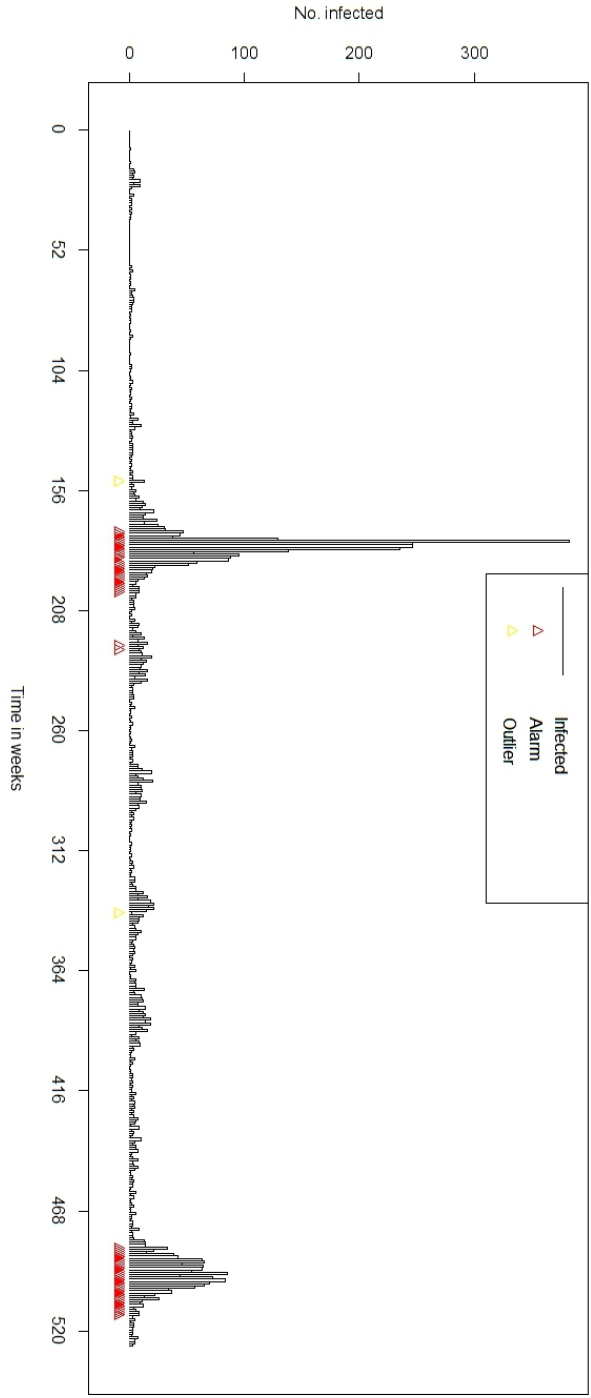


Figure 3.4: *Results of analysis using multi-process dynamic linear model.*

Comparison

Table 3.1 shows week 145 to 178 along with the number of infected, the threshold values and the alarms using Farringtons algorithm and the dynamic linear model, DLM, and the weeks with alarms using the multi-process dynamic linear model, MDLM. These weeks correspond to part of the years 1998/1999, where the first alarms are indicated by the three methods. Farringtons algorithm gives the first alarm in week 148, but the number of infected is only 3 this week, and it therefore questionable whether this is the beginning of the outbreak. The same argument applies to the alarms indicated by Farringtons algorithm in the following weeks, even thought the number of infected become more frequent. In week 176 the number of infected is 38, but in the following week 177 the number is 129, which is well above the threshold values given by Farringtons algorithm.

The first alarm given by the dynamic linear model is in week number 152, but there are no alarms in the next 12 weeks, i.e. the next alarm is in week 165. This could indicate that the alarm in week 152 is a false positive, but further epidemiological investigations need to be carried out to determine this. There are no alarms in the three weeks following week 165, and it is not until week 171, that there are alarms in each week until the number of infected decreases again. The threshold values for the dynamic linear model are higher than the threshold for Farringtons algorithm.

The multi-process dynamic linear model gives the first alarm in week 174, and there are alarms from this week to week 200 except week 186.

Week	No. infected	Farringtons algorithm		DLM		MDLM
		Threshold	Alarm	Threshold	Alarm	Alarm
145	2	2.180402	No	6.379180	No	No
146	1	1.443721	No	6.596930	No	No
147	1	1.443721	No	6.364031	No	No
148	3	1.443721	Yes	6.206510	No	No
149	1	1.443721	No	6.888621	No	No
150	3	1.443721	Yes	6.658264	No	No
151	3	1.915426	Yes	7.298282	No	No
152	13	2.321480	Yes	7.832231	Yes	No
153	2	2.321480	No	8.100284	No	No
154	4	2.268024	Yes	8.071196	No	No
155	2	2.038544	No	8.897144	No	No
156	6	2.154036	Yes	8.729422	No	No
157	4	2.249601	Yes	10.067662	No	No
158	5	2.294297	Yes	10.450053	No	No

159	8	1.952740	Yes	11.097567	No	No
160	6	2.087300	Yes	12.561153	No	No
161	12	2.466892	Yes	13.078961	No	No
162	14	2.905565	Yes	15.290707	No	No
163	11	2.905565	Yes	17.625298	No	No
164	9	2.905565	Yes	18.576575	No	No
165	21	2.871391	Yes	18.627691	Yes	No
166	14	2.871391	Yes	19.233855	No	No
167	12	2.887439	Yes	20.711052	No	No
168	12	2.887439	Yes	21.022993	No	No
169	24	2.742993	Yes	21.189470	Yes	No
170	13	3.520094	Yes	21.608446	No	No
171	25	4.202538	Yes	21.790446	Yes	No
172	30	4.924282	Yes	22.074671	Yes	No
173	31	5.715024	Yes	22.292990	Yes	No
174	47	7.056105	Yes	22.443546	Yes	Yes
175	44	8.247032	Yes	22.525668	Yes	Yes
176	38	8.382453	Yes	22.539931	Yes	Yes
177	129	8.883971	Yes	22.488170	Yes	Yes
178	383	8.858443	Yes	22.373464	Yes	Yes

Table 3.1: *Results of analysis using the three different methods. Weeks, the number of infected, the threshold and alarms for weeks 145 to 178 are shown.*

In table 3.2 week 478 to 492 are shown and again the number of infected, the threshold values and the alarms for Farringtons algorithm and the dynamic linear model, and the alarms for the multi-process dynamic linear model are given. These weeks corresponds to part of the years 2004/2005, which are equivalent to the weeks where the second outbreak is expected. The first alarm given by Farringtons algorithm is in week 484, and except for week 486 there is an alarms each week until the number of infected decreases again.

The dynamic linear model gives the first alarm in the period, where the second outbreak is expected, in week 481, and there are alarms each week until the number of infected decreases again. The threshold values for the dynamic linear model are lower than the threshold values for Farringtons algorithm.

For the multi-process dynamic linear model alarms occur from week 484 to 513.

Week	No. infected	Farringtons algorithm		DLM		MDLM
		Threshold	Alarm	Threshold	Alarm	Alarm
478	3	11.45655	No	10.75522	No	No
479	3	11.97124	No	10.71925	No	No
480	4	12.39311	No	10.66690	No	No
481	13	14.32504	No	10.97631	Yes	No
482	14	14.92616	No	11.47859	Yes	No
483	14	16.89924	No	11.97497	Yes	No
484	33	16.75521	Yes	12.45902	Yes	Yes
485	21	18.76398	Yes	12.92430	Yes	Yes
486	15	19.46819	No	13.36456	Yes	Yes
487	39	21.59929	Yes	13.77388	Yes	Yes
488	42	22.87034	Yes	14.14687	Yes	Yes
489	63	23.26654	Yes	14.47881	Yes	Yes
490	65	22.80209	Yes	14.76576	Yes	Yes

Table 3.2: *Results of the analysis of the three different methods. Weeks, the number of infected, the threshold and alarms for weeks 478 to 490 are shown.*

During the entire period of analysis Farringtons algorithm gives 95 alarms, and 16 of them are in the periods with no outbreaks. This means that at least 16.8% of the alarms are false positive. The false positive alarms in the first years of analysis, where the number of baseline values is limited, could just be seasonal variation detected as possible outbreaks. Some of the alarms in the beginning of the period, where the first outbreak is expected, could also be false positive, but it is not possible to determine when the outbreak actually begins without further epidemiological investigations.

The dynamic linear model gives 54 alarms in the entire period, where 5 alarms are in the periods with no outbreaks i.e. at least 9.3% are false positive alarms.

For the multi-process dynamic linear model there are 58 alarms throughout the period of analysis, and 2 alarms are in the period with no outbreaks, i.e. 3.4% are false positive alarms. The number of infected in the weeks with the two false positive alarms are unusually low.

Chapter 4

Discussion

In this project different methods for detection of outbreaks were compared by analysing weekly counts of Mycoplasma pneumoniae. The different methods presented in this project are intended as an aid to automatic detection of outbreaks, because the demands from diseases under surveillance exceeds the capabilities for manual scanning. The three methods are Farringtons algorithm, the dynamic linear model and the multi-process dynamic linear model. For each method the weeks with alarms were presented, and the number of alarms that are clearly false positive were given. The weeks, where an outbreak was expected to occur, were further investigated to evaluate the possible onset of the outbreak.

The data analysed in this project was collected by Statens Serum Institut, which carries out surveillance in Denmark [8]. In the analyses it is assumed that the data is collected from the same region and that the diagnoses are validated. However, if these assumptions are not true it should be taken into account in formulation of the models. It is also assumed that there are no variation in the reporting delay throughout the period of analysis. The date of report is used as reference date thereby ignoring the reporting delay. Another approach could have been to apply a correction factor to the data based on an estimate of the delay distribution. This, however, requires the date of infection, which are not available. The disadvantage of estimating the reporting delay is that additional uncertainty is introduced into the system. The timeliness of the system is affected by the mean of the reporting delay, because the longer the delay is, the longer it takes for an outbreak to be detected. The sensitivity of the detection system is affected by the variance of the reporting delay, since the variability reduces the change that a threshold will be exceeded [2].

In this project data is aggregated from days into weeks to reduce the variability throughout the week. This also reduces the number of observations with

no infected and small counts. An alternative approach, if analysis on daily basis is desired, could be to use a weight function that account for weekends and holidays.

There is not adjusted for an increase in the population throughout the period either, and the influence of the change in population size has not been further analysed, but an increase would also affect the results. There was an increase of about 214000 people in the population size in the period January 1st 1994 to January 1st 2005 [22].

The distribution of the data is assumed to be Poisson, because it is count data. Farringtons algorithm accommodate this assumption by using a log-linear regression model, but a $\frac{2}{3}$-power transformation for normalization is used in the threshold calculation, which is based on a normal prediction interval. The dynamic linear model and the multi-process dynamic linear model rely on the assumption that the data is normally distributed. To achieve normally distributed data a $\frac{2}{3}$-power transformation is used for skewness correction. To validate the assumption that the data is normally distributed the standard innovations during the steady period have to be normally distributed. The QQ-plot of the standard innovations shows that the standard innovations deviate from the normal distribution and the assumption of normality may not be meet. Alternatively a generalized dynamic linear model or a multi-process generalized dynamic linear model could be used. These types of models allows the observations and the state process to follow other distributions than the normal distribution e.g. the Poisson distribution.

The threshold value in Farringtons algorithm is highly affected by the baseline values used in the calculations. When a small amount of data is available for baseline values, Farringtons algorithm is likely to detect seasonal variations as possible outbreaks resulting in false positive alarms. Past outbreaks also have an effect on the threshold value even though a reweighting procedure is used to give low weights to counts with high residuals. Both Farringtons algorithm and the dynamic linear model are based on a forecast interval when defining the threshold, but where Farringtons algorithm is highly affected by the baseline used, the dynamic linear model is better at adapting to the underlying expected seasonal variation.

The parameters in the dynamic linear model and the multi-process dynamic linear model are estimated based on the steady period of the time series. This introduces bias, because the time series under analysis is also used for estimation of the unknown parameters. Ideally the parameters should be estimated based on a period not under analysis. The threshold defined in Farringtons algorithm is only based on the previous history of the time series.

The dynamic linear model and the multi-process dynamic linear model are defined as consisting of two components: a random walk plus noise and a seasonal component, where the seasonal component is defined as having a period of 52 week. There was not included a trend in Farringtons algorithm, because it was not significant. Therefore a trend component was not used when formulating the dynamic linear model and the multi-process dynamic linear model.

For both Farringtons algorithm and the dynamic linear model it is assumed that there are 52 weeks in a year i.e. 364 day. In Farringtons algorithm this is used to identify the corresponding baseline weeks previous years for threshold calculations, and in the dynamic linear model it is used in definition of the seasonal component. Thus the assumption of 52 weeks in a year shift the baseline values and the estimated seasonal variation, since there are 365 days in a year and leap years is not taken into account. In the multi-process dynamic linear model, however, it is assumed that there are 365.25 days in a year, thereby accounting for the 365 days in a year and leap years.

A 99% prediction interval is used to define the threshold for Farringtons algorithm and the dynamic linear model. The size of the prediction interval was chosen to reduce the number of false positives, but the size of the interval also affect how early an outbreak is detected. A smaller prediction interval could mean that outbreaks are detected earlier, but it could also give more false positive alarms.

It is not possible to define the precise onset of the outbreaks, and which model is the first to detect the two outbreaks, but the alarms still gives an indication of when the outbreaks are detected by the three methods. For the first outbreak Farringtons algorithm indicates an outbreak before the dynamic linear model and the multi-process dynamic linear model. The number of infected in the first weeks, where Farringtons algorithm gives alarms, is on the other hand low, and it is therefore questionable when the outbreak is detected at early. The possible false positive alarms in the first years of analysis could be caused by the limited baseline values available for threshold calculations. In the period where the first outbreak is expected, the dynamic linear model has few sporadic alarms in the week 160 to 170, but it is not until week 171, that there is alarm each week until the number of incidences decreases again. The multi-process dynamic linear indicate an outlier in the beginning of the period, where the first outbreak is expected, and indicate that the onset of the outbreak is week 174 i.e. three weeks after the dynamic linear model. The second outbreak is indicated by Farringtons algorithm to begin in week 484, but then there is a week without alarm in week 486. The first alarm given by the dynamic linear model is three weeks before Farringtons algorithm in week 481, and the multi-process dynamic linear model indicates an outbreak from week 484.

Farringtons algorithm gave 16 identified false positive alarms, and the dynamic linear model indicated 5 false positive alarms. The multi-process dynamic linear model gave 2 false positive alarms in the steady period between the two identified outbreaks, but these alarms were in weeks, where the number of infected was unusually low. The number of false positive is considerable higher for Farringtons algorithm than the dynamic linear model and the multi-process dynamic linear model. The multi-process dynamic linear model is defined so it detect observations, where the variance of the state process is high, as outliers or outbreaks. This is why both unusual high counts or unusual low counts is detected as possible outliers or outbreaks.

The multi-process dynamic linear models has the advantages over both Farringtons algorithm and the dynamic linear model that it can differentiate between several possible models; in this case steady state, outlier and outbreaks. This means that it is possible to detect outliers separately, whereas both Farringtons algorithm and the dynamic linear model only distinguish between no outbreak and possible outbreak. The results reflect this, where week 152 is marked as a possible outbreak by Farringtons algorithm and the dynamic linear model, but is identified as an outlier by the multi-process dynamic linear model.

In the multi-process dynamic linear model the different models are selected at each time with known probability. The dependence structure of the models is first-order Markov, i.e. the model at each time t only depend on the model at time $t-1$. It is, however, possible that the model at time t depends on models prior to time $t-1$. This dependence structure could be achieved by using a higher-order Markov structure of the models. The transition probabilities are a qualified guess based on the assumption that 98.5% of the time the time series is in steady state. The probabilities should, however, be analysed further, since an outbreak of Mycoplasma pneumoniae occur every 4 to 6 years with a duration of 3 to 4 months according to Statens Serum Institut, which correspond to more than 1.5% of the time. It is also assumed that the transition probabilities are fixed throughout the year, but the time of the year, where outbreaks occur, could depend on the seasonal variation. Thus an outbreak could be more likely to occur in a period, if there is a higher number of infected because of seasonal variation.

The complexity of the calculations of the multi-process dynamic linear model is reduced by ignoring possible model sequences with low posterior probability. It is, however, possible that sequences with low posterior probability in the beginning, which is ignored, later could be relevant.

The comparisons of Farringtons algorithm, the dynamic linear model and the multi-process dynamic linear model are only performed by analysing the number of samples tested positive for Mycoplasma pneumoniae. One of the requirements of the detection system is that it should be able to handle a wide variety of organisms with varying organism count. Mycoplasma pneumoniae is one of the more common microorganisms, the methods should also be compared using other data with lower and higher organism count than Mycoplasma pneumoniae.

Chapter 5

Conclusion

The aim of this project was to compare different methods for detection of disease outbreaks. The three different methods were Farringtons algorithm, the dynamic linear model and the multi-process dynamic linear model. Farringtons algorithm is currently being used by Statens Serum Institut.

Analysis of Mycoplasma pneumoniae using the three different methods indicates that the dynamic linear model and the multi-process dynamic linear model are superior to the log-linear regression model presented by Farrington et al. Farringtons algorithm is highly affected by the baseline values used for threshold calculations, where the dynamic linear model and the multi-process dynamic linear model are better at adapting to the underlying expected seasonal variation. The dynamic linear model seems to detect the onset of the outbreaks slightly before the multi-process dynamic linear model. The multi-process dynamic linear model has the advantage over the dynamic linear model that the posterior probabilities of each model is given. Thus it is possible to differentiate between the models; steady state, outliers and possible outbreaks. Farringtons algorithm and the dynamic linear model only differentiate between steady state and possible outbreaks.

Part II

Theory

Chapter 6

Generalized linear models

In the following chapter generalized linear models are presented. Generalized linear models allow the response variables to have other distributions than the normal distribution, and they are not restricted to be continuous. The response variable Y in a generalized linear model belongs to a distribution in the exponential family, which is introduced in the next section [23, p. 45].

6.1 Exponential family

A probability distribution of a random variable Y depending on a single parameter θ belongs to the exponential family, if the probability function is on the form

$$f(y|\theta) = \exp(a(y)b(\theta) + c(\theta) + d(y)), \tag{6.1}$$

where $a(y)$, $b(\theta)$, $c(\theta)$ and $d(y)$ are known functions, and $b(\theta)$ is called the natural parameter. The distribution is on canonical form if the function $a(y) = y$. If there are more parameters, than the parameter of interest θ, then they are considered as nuisance parameters, and are part of the functions $a(y)$, $b(\theta)$, $c(\theta)$ and $d(y)$, which are known [23, p. 46].

6.1.1 Expected value and variance of $a(Y)$

For the exponential family distribution the expected value of $a(Y)$ can be obtained by differentiation of the probability density function, and using that the

differential of a density function must integrate to 0. This gives

$$\begin{aligned}
&\int \frac{\partial f(y|\theta)}{\partial b} dy = 0 \Leftrightarrow \\
&\int \frac{\partial}{\partial b} \exp(a(y)b + c(\theta(b)) + d(y)) dy = 0 \Leftrightarrow \\
&\int \left(a(y) + \frac{\partial c(\theta(b))}{\partial b} \right) f(y|\theta) dy = 0 \Leftrightarrow \\
&\int a(y) f(y|\theta) dy + \int \frac{\partial c(\theta(b))}{\partial b} f(y|\theta) dy = 0 \Leftrightarrow \\
&\mathrm{E}[a(Y)] = \frac{-\partial c(\theta(b))}{\partial b} \Leftrightarrow \\
&\mathrm{E}[a(Y)] = \frac{-\partial c(\theta)}{\partial \theta} \cdot \frac{\partial \theta}{\partial b} \Leftrightarrow \\
&\mathrm{E}[a(Y)] = \frac{\frac{-\partial c(\theta)}{\partial \theta}}{\frac{\partial b}{\partial \theta}} \Leftrightarrow \\
&\mathrm{E}[a(Y)] = \frac{-c'(\theta)}{b'(\theta)}.
\end{aligned} \tag{6.2}$$

To obtain the variance of $a(Y)$ the density function is differentiated twice and because the second differential of a density function must integrate to zero, then

$$\begin{aligned}
&\frac{\partial^2}{\partial b^2} \int f(y|\theta) dy = 0 \Leftrightarrow \\
&\int \frac{\partial}{\partial b} \left(\left(a(y) + \frac{\partial c(\theta(b))}{\partial b} \right) f(y|\theta) \right) dy = 0 \Leftrightarrow \\
&\int \frac{\partial^2 c(\theta(b))}{\partial b^2} f(y|\theta) + \left(a(y) + \frac{\partial c(\theta(b))}{\partial b} \right) \frac{\partial}{\partial b} f(y|\theta) dy = 0 \Leftrightarrow \\
&\int \frac{\partial^2 c(\theta(b))}{\partial b^2} f(y|\theta) + \left(a(y) + \frac{\partial c(\theta(b))}{\partial b} \right) \left(a(y) + \frac{\partial c(\theta(b))}{\partial b} \right) f(y|\theta) dy = 0 \Leftrightarrow \\
&\int \frac{\partial^2 c(\theta(b))}{\partial b^2} f(y|\theta) dy + \int \left(a(y) - \mathrm{E}[a(Y)] \right)^2 f(y|\theta) dy = 0 \Leftrightarrow \\
&\mathrm{Var}[a(Y)] = -\frac{\partial^2 c(\theta(b))}{\partial b^2} \Leftrightarrow \\
&\mathrm{Var}[a(Y)] = -\left(\frac{\partial b}{\partial \theta} \right)^{-1} \frac{\partial}{\partial \theta} \left(\frac{\partial c(\theta(b))}{\partial b} \right) \Leftrightarrow \\
&\mathrm{Var}[a(Y)] = -\left(\frac{\partial b}{\partial \theta} \right)^{-1} \frac{\partial}{\partial \theta} \left(\frac{\frac{\partial c(\theta)}{\partial \theta}}{\frac{\partial b}{\partial \theta}} \right) \Leftrightarrow
\end{aligned}$$

$$\begin{aligned}\mathrm{Var}[a(Y)] &= -\frac{1}{b'(\theta)}\left(\frac{c''(\theta)b'(\theta)-c'(\theta)b''(\theta)}{(b'(\theta))^2}\right) \Leftrightarrow \\ \mathrm{Var}[a(Y)] &= \frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{(b'(\theta))^3} \qquad (6.3)\end{aligned}$$

[23, pp. 48-49].

6.1.2 The score statistic and the information

For a distribution of the exponential family the log-likelihood function is given by

$$\ell(\theta;y) = a(y)b(\theta) + c(\theta) + d(y).$$

The score statistic U is the derivative of the log-likelihood function $\ell(\theta;y)$

$$U(\theta;y) = \frac{d\ell(\theta;y)}{d\theta} = a(y)b'(\theta) + c'(\theta).$$

Because the score statistic is dependent on y it can be considered as a random variable

$$U = a(y)b'(\theta) + c'(\theta).$$

The expected value of U is given as

$$\begin{aligned}\int \exp(\ell(y;\theta))dy &= 1 \Leftrightarrow \\ \frac{\partial}{\partial\theta}\int \exp(\ell(y;\theta))dy &= 0 \Leftrightarrow \\ \int U\exp(\ell(y;\theta))dy &= 0 \Leftrightarrow \\ \mathrm{E}[U] &= 0.\end{aligned}$$

The information $\mathfrak{I}$ is the variance of U, which is given by

$$\begin{aligned}\mathfrak{I} &= \mathrm{Var}[U] \\ &= (b'(\theta))^2\mathrm{Var}[a(Y)] \\ &= (b'(\theta))^2\left(\frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{(b'(\theta))^3}\right) \\ &= \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta).\end{aligned}$$

The variance of U can also be written as

$$\begin{aligned}\mathrm{Var}[U] &= \mathrm{E}[U^2] + (\mathrm{E}[U])^2 \\ &= \mathrm{E}[U^2],\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\int \exp(\ell(y;\theta))dy &= 0 \Leftrightarrow \\
\int \frac{\partial}{\partial\theta}(U\exp(\ell(y;\theta)))dy &= 0 \Leftrightarrow \\
\int U'\exp(\ell(y;\theta))dy + \int U^2\exp(\ell(y;\theta))dy &= 0 \Leftrightarrow \\
\mathrm{E}[U'] + \mathrm{E}[U^2] &= 0 \Leftrightarrow \\
\mathrm{Var}[U] &= -\mathrm{E}[U']
\end{aligned}$$

[23, p. 50].

6.2 Generalized linear models

A generalized linear is defined from independent random variables $Y_1, \ldots, Y_N$, where each variable belongs to a model on the same form from the exponential family. The distribution of the variables Y_i only depend on a parameter θ_i, and it is on canonical form. The probability density function of Y_i is given by

$$f(y_i|\theta_i) = \exp(y_i b(\theta_i) + c(\theta_i) + d(y_i)),$$

and the joint probability density function of $Y_1, \ldots, Y_N$ is

$$\begin{aligned}
f(y_1, \ldots, y_N|\theta_1, \ldots, \theta_N) &= \prod_{i=1}^{N} \exp(y_i b(\theta_i) + c(\theta_i) + d(y_i)) \\
&= \exp\left(\sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)\right).
\end{aligned}$$

Let $\mathrm{E}[Y_i] = \mu_i$, where μ_i is a function of θ_i, which may depend on some explanatory variables x_i. Then for a generalized linear model it is assumed that there is a transformation of μ_i so that

$$g(\mu_i) = \mathbf{x}_i^T\beta = \eta_i.$$

This function is called the link function and is a monotone, differentiable function. The vector $\mathbf{x}_i^T$ represent the ith row of the design matrix $\mathbf{X}$, and it is a $p \times 1$ vector, which contains the explanatory variables. The vector β is a $p \times 1$ vector containing the parameters of interest $\beta_1, \ldots, \beta_p$, $p < N$.

Therefore a generalized linear model contain three elements:

1. Independent random response variables $Y_1, \ldots, Y_N$, which belongs to a distribution of the same form from the exponential family.

2. The parameters of interest

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},$$

and the explanatory variables

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{bmatrix}.$$

3. The link function

$$g(\mu_i) = \mathbf{x}_i^T \beta = \eta_i, \quad \text{where} \quad \mu_i = \mathrm{E}[Y_i].$$

[23, pp. 51-52].

6.3 Maximum likelihood estimation

Given independent random variables $Y_1, \ldots, Y_N$ which satisfy the properties of a generalized linear model. Maximum likelihood estimation is used to estimate the parameters β, which are connected to $Y_1, \ldots, Y_N$ through the expected value $E[Y_i] = \mu_i$ and the link function $g(\mu_i) = \mathbf{x}_i^T \beta$. The log-likelihood function for each Y_i is given by

$$\ell_i = y_i b(\theta_i) + c(\theta_i) + d(y_i),$$

where the function b, c and d are given by the exponential family defined by equation (6.1). The expected value of Y_i is defined in equation (6.2), the variance of Y_i is given by equation (6.3), and the link function is $g(\mu_i) = \mathbf{x}_i^T \beta = \eta_i$, where $\mathbf{x}_i^T$ is the ith row of the design matrix $\mathbf{X}$ with elements x_{ij} for $j = 1, \ldots, p$. Because all the Y_i's is independent the log-likelihood function is given by

$$\ell = \sum_{i=1}^{N} \ell_i = \sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i).$$

The score vector U is used to find the maximum likelihood estimator, $\widehat{\beta}$, for the parameter β

$$U_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{N} \left(\frac{\partial \ell_i}{\partial \beta_j} \right) = \sum_{i=1}^{N} \left(\frac{\partial \ell_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j} \right) \tag{6.4}$$

for $j = 1, \ldots, p$. The maximum likelihood estimator, $\widehat{\beta}$, is given by the solution of the equation $U(\beta) = 0$. The differential with respect to θ_i of the log-likelihood function is

$$\begin{aligned} \frac{\partial \ell_i}{\partial \theta_i} &= y_i b'(\theta_i) + c'(\theta_i) \\ &= y_i b'(\theta_i) - b'(\theta_i) \left(\frac{-c'(\theta_i)}{b'(\theta_i)} \right) \\ &= b'(\theta_i)(y_i - \mu_i), \end{aligned} \tag{6.5}$$

the differential of θ_i with respect to μ_i is

$$\begin{aligned} \frac{\partial \theta_i}{\partial \mu_i} &= 1 \Big/ \left(\frac{\partial \mu_i}{\partial \theta_i} \right) \\ &= 1 \Big/ \left(\frac{-c''(\theta_i)b'(\theta_i) + c'(\theta_i)b''(\theta_i)}{(b'(\theta_i))^2} \right) \\ &= 1 \,/ b'(\theta_i) \mathrm{Var}[Y_i] \,, \end{aligned} \tag{6.6}$$

and the differential of μ_i with respect to β_j is

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ij} \tag{6.7}$$

using the definition of the link function. Combining equation (6.4), (6.5), (6.6) and (6.7) the score function is

$$U_j = \sum_{i=1}^{N} \left(\frac{(y_i - \mu_i)}{\mathrm{Var}[Y_i]} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right) \tag{6.8}$$

for $j = 1, \ldots, p$. The elements of the information matrix $\mathfrak{I}$ is then defined as

$$\begin{aligned} \mathfrak{I}_{jk} &= \mathrm{E}[U_j U_k] \\ &= \mathrm{E}\left[\sum_{i=1}^{N} \left(\frac{(Y_i - \mu_i)}{\mathrm{Var}[Y_i]} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \right) \sum_{l=1}^{N} \left(\frac{(Y_l - \mu_l)}{\mathrm{Var}[Y_l]} x_{lk} \frac{\partial \mu_l}{\partial \eta_l} \right) \right] \\ &= \sum_{i=1}^{N} \frac{\mathrm{E}[(Y_i - \mu_i)^2] x_{ij} x_{ik}}{(\mathrm{Var}[Y_i])^2} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \\ &= \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{\mathrm{Var}[Y_i]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2, \end{aligned} \tag{6.9}$$

since the Y_i's are independent and $\mathrm{E}[(Y_i - \mu_i)(Y_l - \mu_l)] = 0$ for $i \neq l$. The method of scoring is used to find the maximum likelihood estimate of β, which

is the solution to $U(\beta) = 0$. A numerical solution is obtained using the Taylor series approximations

$$\begin{aligned} U(\beta) &\approx U\left(\widehat{\beta}\right) + \left\{\frac{\partial^2 \ell}{\partial\beta_i \partial\beta_j}\right\}\left(\beta - \widehat{\beta}\right) \\ &= U\left(\widehat{\beta}\right) + \mathrm{E}\left[\frac{\partial^2 \ell}{\partial\beta_i \partial\beta_j}\right]\left(\beta - \widehat{\beta}\right) \\ &= U\left(\widehat{\beta}\right) - \mathfrak{I}\left(\beta - \widehat{\beta}\right) \\ &= -\mathfrak{I}\left(\beta - \widehat{\beta}\right). \end{aligned}$$

This means that $\widehat{\beta} = \mathfrak{I}^{-1}U(\beta) + \beta$, which leads to the general estimating equation

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} + \left(\mathfrak{I}^{(m-1)}\right)^{-1}\mathbf{U}^{(m-1)},$$

where the vector of estimates of the parameters $\beta_1, \ldots, \beta_p$ is $\mathbf{b}^{(m)}$ at the mth iteration, $\left(\mathfrak{I}^{(m-1)}\right)^{-1}$ is the inverse of the information matrix with elements $\mathfrak{I}_{jk}$ defined in equation (6.9), and $\mathbf{U}^{(m-1)}$ is a vector of elements defined in equation (6.8) evaluated at $\mathbf{b}^{(m-1)}$ [23, pp. 64-65].

6.4 Inference

In this section inference for generalized linear models will be described.

Sampling distribution for score statistics

Let $Y_1, \ldots, Y_N$ be independent random variables in a generalized linear model with parameters β, where $\mathrm{E}[Y_i] = \mu_i$ and $g(\mu_i) = \mathbf{x}_i^T\beta = \eta_i$. The score statistics defined in equation (6.8) is given by

$$U_j = \sum_{i=1}^{N}\left(\frac{(y_i - \mu_i)}{\mathrm{Var}[Y_i]}x_{ij}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)\right), \tag{6.10}$$

for $j = 1, \ldots, p$. Equation (6.10) is a sum of independent terms, which may be approximated by a normal distribution. The expected value is $\mathrm{E}[U_j] = 0$ for $j = 1, \ldots, p$, because $\mathrm{E}[Y_i] = \mu_i$ for all i, and the variance-covariance matrix for the score statistic is given by the information matrix $\mathfrak{I}$ with elements $\mathfrak{I}_{jk} = \mathrm{E}[U_jU_k]$. For one parameter β the asymptotic sampling distribution for the score statistic is

$$\frac{U - \mathrm{E}[U]}{\sqrt{\mathrm{Var}[U]}} = \frac{U}{\sqrt{\mathfrak{I}}} \sim N(0, 1)$$

implying

$$\frac{(U - \mathrm{E}[U])^2}{\mathrm{Var}[U]} = \frac{U^2}{\mathfrak{I}} \sim \chi^2(1),$$

since $\mathrm{E}[U] = 0$ and $\mathrm{Var}[U] = \mathfrak{I}$.
For a vector of parameters $\beta = [\beta_1 \cdots \beta_p]^T$ the score vector $U = [U_1 \cdots U_p]^T$ has the asymptotic multivariate normal distribution $U \sim N_p(0, \mathfrak{I})$ and for large samples

$$\Big[U - \mathrm{E}[U]\Big]^T V^{-1} \Big[U - \mathrm{E}[U]\Big] = U^T \mathfrak{I}^{-1} U \sim \chi^2(p)$$

[23, pp. 74-75].

Sampling distribution for maximum likelihood estimators

The sampling distribution of the maximum likelihood estimator $b = \widehat{\beta}$ can be obtained using Taylor approximation of the score function for a vector parameter β given by

$$\begin{aligned} U(\beta) &\approx U(b) + U'(b)(\beta - b) \\ &\approx -\mathfrak{I}(b)(\beta - b), \end{aligned}$$

where it is used that the derivative of the score function can be approximated by its expected value $\mathrm{E}[U'(b)] = -\mathfrak{I}$, evaluated at $\beta = b$. Given that the information $\mathfrak{I}$ is invertible, this can be written as

$$(b - \beta) \approx -\mathfrak{I}^{-1} U,$$

and if $\mathfrak{I}$ is regarded as a constant, then $\mathrm{E}[b - \beta] = 0$, since $\mathrm{E}[U] = 0$. This means that asymptotically $\mathrm{E}[b] = \beta$ and b is a consistent estimator of β. The variance-covariance matrix V for b is given by

$$\begin{aligned} V &= \mathrm{E}\left[(b - \beta)(b - \beta)^T\right] \\ &= \mathfrak{I}^{-1} \mathrm{E}\left[UU^T\right] \mathfrak{I}^{-1} \\ &= \mathfrak{I}^{-1}, \end{aligned}$$

since $\mathrm{E}\left[UU^T\right] = \mathfrak{I}$ and $\mathfrak{I}$ is symmetric i.e. $\left(\mathfrak{I}^{-1}\right)^T = \mathfrak{I}^{-1}$. Then for b the asymptotic sampling distribution is $b \sim N_p(\beta, \mathfrak{I}^{-1})$ and

$$\Big[b - \mathrm{E}[b]\Big]^T V^{-1} \Big[b - \mathrm{E}[b]\Big] = \Big[b - \beta\Big]^T \mathfrak{I}(b) \Big[b - \beta\Big] \sim \chi^2(p), \tag{6.11}$$

which is the Wald statistic. Equation (6.11) is an exact result if the response variables in a GLM are normally distributed [23, pp. 77-78].

The likelihood ratio

Let Y be a random vector with density function $f(y;\beta)$. The hypothesis to be tested is

$$\begin{array}{ll} H_0: & \beta \in \mathcal{B}_0 \\ H_1: & \beta \in \mathcal{B}_1 \end{array} \quad , \quad \mathcal{B} = \mathcal{B}_0 \cup \mathcal{B}_1,$$

where $\mathcal{B}_0$ and $\mathcal{B}_1$ are disjoint parameter sets. Generally the likelihood ratio test at level α has the rejection region

$$R = \{y | \lambda(y) \leq \lambda_\alpha\},$$

where the likelihood ratio is given by

$$\lambda(y) = \frac{\sup_{\beta \in \mathcal{B}_0} L(\beta; y)}{\sup_{\beta \in \mathcal{B}} L(\beta; y)},$$

and the critical value λ_α is selected so

$$\sup_{\beta \in \mathcal{B}_0} \Pr\{\lambda(y) \leq \lambda_\alpha; \beta\} = \alpha.$$

An equivalent test statistic, which is still called the likelihood ratio, is

$$W(y) = -2\log(\lambda(y)) = -2\left[\ell\left(\widehat{\beta_0}; y\right) - \ell\left(\widehat{\beta}; y\right)\right].$$

This test statistic measures the difference between the log-likelihood at $\widehat{\beta}$ and $\widehat{\beta_0}$. If $\mathcal{B}$ is defined as the set of parameters $\beta = (\beta_1, \ldots, \beta_p)^T$, and $\mathcal{B}_0$ is obtained by the $p - q$ equations,

$$\begin{cases} g_1(\beta_1, \ldots, \beta_p) = 0, \\ \vdots \\ g_{p-q}(\beta_1, \ldots, \beta_p) = 0, \end{cases}$$

where $g_1, \ldots, g_{p-q}$ are regular functions, then $W(y) \xrightarrow{d} \chi^2_{p-q}$ [20, pp. 112-113,116].

Deviance

The deviance can be used to analyse how well the model fit the data by comparing two GLMs, M_1 and M_2. The two models have the same distribution and link function, but $M_2 \subset M_1$ i.e. M_2 is nested within M_1. First consider the situation where β_1 is the parameter vector for M_1 with p parameters, and β_2 is the parameter vector for M_2 containing q parameters, where $q < p$. Then using the likelihood ratio

$$W(y) = -2\log\left(\frac{L(\widehat{\beta}_2; y)}{L(\widehat{\beta}_1; y)}\right) \xrightarrow{d} \chi^2(p - q). \tag{6.12}$$

Second consider the situation where M_1 is the saturated model, which is the model with the maximum number of parameters, N, that can be estimated, and M_2 is the model of interest, which is a restriction of M_1, with q parameters. Let $\beta_{\max}$ be the parameter vector for M_1, and let $\widehat{\beta}_{\max}$ be the maximum likelihood estimate of $\beta_{\max}$. The deviance is then defined as

$$D = -2[\ell(\widehat{\beta_2}; y) - \ell(\widehat{\beta}_{\max}; y)],$$

and $D \sim \chi^2(N - q)$. Then equation (6.12) can be rewritten using the deviance

$$\begin{aligned} W(y) &= -2\log\left(\frac{L(\widehat{\beta}_2; y)}{L(\widehat{\beta}_1; y)}\right) \\ &= -2\log\left(\frac{L(\widehat{\beta}_2; y)}{L(\widehat{\beta}_{\max}; y)} \cdot \frac{L(\widehat{\beta}_{\max}; y)}{L(\widehat{\beta}_1; y)}\right) \\ &= -2\log\left(\frac{L(\widehat{\beta}_2; y)}{L(\widehat{\beta}_{\max})}\right) + 2\log\left(\frac{L(\widehat{\beta}_1; y)}{L(\widehat{\beta}_{\max})}\right) \\ &= D(M_2) - D(M_1), \end{aligned}$$

where $D(M_1)$ is the deviance of M_1, and $D(M_2)$ is the deviance of M_2 [20, pp. 242-244].

6.4.1 Confidence interval

Confidence intervals can be used for statistical inference. The width of a confidence interval provides a measure of precision with which inference can be made. For an estimate $\widehat{\mu}$ it is desirable to assess the degree of accuracy of the estimate and provide a set of possible values of the parameter μ, that contains $\widehat{\mu}$. This set of possible values is called an interval estimate. The set $C(y)$ is a confidence set of level $1 - \alpha$ for μ if the probability that the random set $C(Y)$ includes the true parameter μ is $1 - \alpha$ for any possible value of μ, i.e.

$$\Pr\{C(Y) \ni \mu; \mu\} = 1 - \alpha.$$

This statement holds a priori, when $C(Y)$ is a random set, but a posteriori the statement does not hold for the observed set $C(y)$. A posteriori it is said that a specific interval (c_1, c_2) is a $(1 - \alpha)\%$ confidence interval.
Let $\overline{y}$ be the sample mean of a sample of size n from a normally distributed,

$N(\mu, \sigma^2)$, random variable. Then $\overline{Y} - \mu \sim N(0, \frac{\sigma^2}{n})$ and

$$\begin{aligned} 1-\alpha &= \Pr\left\{ \left(\overline{Y}-\mu\right) \frac{\sqrt{n}}{\sigma} \in \left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right) \right\} \\ &= \Pr\left\{ -z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu - \overline{Y} < z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right\} \\ &= \Pr\left\{ \overline{Y} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{Y} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right\} \\ &= \Pr\left\{ \left(\overline{Y} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{Y} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) \ni \mu \right\}, \end{aligned}$$

where $\Phi\left(-z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$. Thus the interval $\left(\overline{y} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{y} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$ is a confidence interval of level $1-\alpha$ [20, pp. 115,141-142].

6.5 Poisson distribution

A discrete random variable Y is Poisson distributed if the probability function is given by

$$\begin{aligned} f(y;\theta) &= \frac{\theta^y e^{-\theta}}{y!} \\ &= \exp(y\log(\theta) - \theta - \log y!), \end{aligned} \tag{6.13}$$

where $y = 0, 1, 2, \ldots$ is the number of events and θ is the average number of events as shown below. The Poisson distribution is part of the exponential family, where $a(y) = y$, $b(\theta) = \log(\theta)$, $c(\theta) = -\theta$ and $d(y) = -\log(y!)$, and it is on canonical form. The expected value for the Poisson distribution is

$$\mathrm{E}[Y] = \mathrm{E}[a(Y)] = \frac{-c'(\theta)}{b'(\theta)} = \frac{-(-1)}{1/\theta} = \theta,$$

where it is used that the expected value of $a(Y)$ is given by equation (6.2). Similarly the variance calculated using the variance of $a(Y)$ is given by equation (6.3),

$$\begin{aligned} \mathrm{var}[Y] &= \mathrm{var}[a(Y)] = \frac{b''(\theta)c'(y) - c''(\theta)b'(\theta)}{(b'(\theta))^3} \\ &= \frac{\left(-\frac{1}{\theta^2}\right)\cdot(-1) - 0\cdot\left(\frac{1}{\theta}\right)}{\left(\frac{1}{\theta}\right)^3} = \frac{\left(\frac{1}{\theta}\right)^2}{\left(\frac{1}{\theta}\right)^3} = \theta \end{aligned}$$

[23, p. 47, 165].

6.6 Poisson regression

Consider $Y_1, \ldots, Y_N$ independent random variables, where Y_i is the number of events observed from exposure n_i for the ith covariate pattern. For example could Y_i be the number of incidences of a disease in a specific area, which would depend on the total number of individuals in the area, n_i, and other variables, such as previous medical history. The subscript i indicates the different combinations of disease and medical history. The expected value of Y_i is given by

$$\mathrm{E}\left[Y_i\right] = \mu_i = n_i\theta_i,$$

where θ_i may depend on some other explanatory variables x_i, which is modelled by

$$\theta_i = \exp(x_i^T\beta).$$

The generalized linear model is

$$\mathrm{E}[Y_i] = \mu_i = n_i \exp(x_i^T\beta),$$

where $Y_i \sim \mathrm{Po}(\mu_i)$ and the logarithmic function is the link function, which gives a linear component

$$\log(\mu_i) = \log(n_i) + x_i^T\beta.$$

The term $\log(n_i)$ is the offset and is a known constant, and x_i is the covariate pattern and β is the parameters. The Poisson regression model is sometimes called a log-linear model.

The fitted values of Y_i are

$$\hat{Y_i} = \hat{\mu}_i = n_i \exp(x_i^T\mathbf{b}),$$

where $i = 1, \ldots, N$ and $\mathbf{b}$ is the maximum likelihood estimate of β. The fitted values are estimates of the expected values and is therefore also denoted e_i. The Pearson residuals are given by

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}},$$

where o_i is the observed value of Y_i and the standard error is estimated by $\sqrt{e_i}$ because the variance and the expected value are equal for the Poisson distribution. The Pearson residuals and the χ^2 goodness of fit statistic are related for the Poisson distribution by

$$X^2 = \sum r_i^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

[23, pp. 166-167].

Deviance for a Poisson model

Let the response variables $Y_i \sim \text{Po}(\lambda_i)$, $i = 1, \ldots, N$, be independent, then the log-likelihood function is

$$\ell(\beta; y) = \sum y_i \log(\lambda_i) - \sum \lambda_i - \sum \log(y_i!).$$

The maximum value of the likelihood function is

$$\ell(\mathbf{b}_{\max}; y) = \sum y_i \log(y_i) - \sum y_i - \sum \log(y_i!),$$

where the maximum likelihood estimate of $\hat{\lambda}_i = y_i$. The log-likelihood function of the model of interest with p parameters evaluated at the maximum likelihood estimate for β, $\mathbf{b}$, is

$$\ell(\mathbf{b}; y) = \sum y_i \log(\hat{y_i}) - \sum \hat{y_i} - \sum \log(y_i!)$$

The deviance for the Poisson model is defined as

$$\begin{aligned} D &= 2[\ell(\mathbf{b}_{\max}; y) - \ell(\mathbf{b}; y)] \\ &= 2\left[\sum y_i \log(y_i/\hat{y_i}) - \sum (y_i - \hat{y_i})\right] \\ &= 2\sum \left[o_i \log\left(\frac{o_i}{e_i}\right) - (o_i - e_i)\right], \end{aligned}$$

where o_i is the observed value y_i and e_i is the estimated expected value $\hat{y_i}$. The deviance residuals are

$$d_i = sign(o_i - e_i)\sqrt{\left[o_i \log\left(\frac{o_i}{e_i}\right) - (o_i - e_i)\right]},$$

for $i = 1, \ldots, N$. The deviance and the deviance residuals are related as $D = \sum d_i^2$, and the deviance and the χ^2 goodness of fit statistic are also related approximately in the following way

$$\begin{aligned} D &\approx 2\sum_{i=1}^{N} \left[(o_i - e_i) + \frac{1}{2}\frac{(o_i - e_i)^2}{e_i} - (o_i - e_i)\right] \\ &= \sum_{i=1}^{N} \frac{(o_i - e_i)^2}{e_i} \\ &= X^2, \end{aligned}$$

where the Taylor series expansion $o \log\left(\frac{o}{e}\right) = (o - e) + \frac{1}{2}\frac{(o-e)^2}{e} + \ldots$ is used. Both the deviance and the χ^2 statistic can be used as a measure of goodness of fit. They can be obtained from the data and the fitted model and are compared with the χ^2 distribution with $N - q$ degrees of freedom, where q is the number of estimated parameters [23, pp. 83-84,166-168].

Chapter 7

State space models

With state space models we consider time series, which are the output of a dynamic system affected by random noise. State space models are a flexible class of models, that can be used in a wide variety of applications including, when the time series are non-stationary, have structural changes or display irregular patterns. The model can consists of a combination of several components; for instance trend, seasonal or regressive components, that gives a natural interpretation of the parameters and the output. Recursive computations of the conditional distribution of the parameters of interest given the available information are used for estimation and forecasting [24, p. 31].

Let $(Y_t)_{t\geq 1}$ be a time series. The joint distribution of $(Y_1, \ldots, Y_t)$ for any $t \geq 1$ is defined using Markovian dependence structure,because assumptions of independence or exchangeability would make time irrelevant, which is often not true for most time series. The time series $(Y_t)_{t\geq 1}$ is a Markov chain if

$$\pi(y_t|y_{1:t-1}) = \pi(y_t|y_{t-1})$$

for any $t > 1$, and the joint finite-dimensional distribution is given by

$$\pi(y_{1:t}) = \pi(y_1) \cdot \prod_{j=2}^{t} \pi(y_j|y_{j-1}).$$

In many situations it is not appropriate to assume a Markovian structure for the observations. Therefore it is assumed in state space models that there is a state process, which is an unobservable Markov chain (θ_t), and that Y_t is a measurement of θ_t affected by random disturbance. Figure 7.1 shows a directed acyclic graph, which illustrates the dependence structure in a state space model. The graph can be used to deduce conditional independence properties of random variables, for example variables A and B are conditionally independent given a third set C, if and only if C separates A and B [24, pp. 39-41].

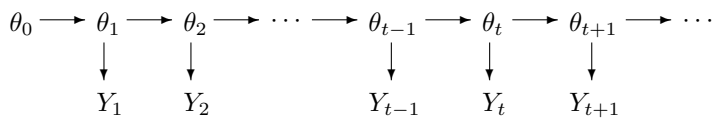


Figure 7.1: *Dependence structure in a state space model [24, p. 41]*

Definition 7.1 (State space model)
A state space model consists of to time series: An $\mathbb{R}^p$-valued time-series $(\theta_t : t = 0, 1, \ldots)$, and $(Y_t : t = 1, 2, \ldots)$ an $\mathbb{R}^m$-valued time series. The time series satisfy the following assumptions:

1. *(θ_t) is a Markov chain.*
2. *Conditionally on (θ_t), the Y_t's are independent and Y_t depends on θ_t only.*

[24, p. 40]. ◇

The joint distribution of $(\theta_0, \theta_1, \ldots, \theta_t, Y_1, \ldots, Y_t)$ for any $t > 0$ is then given by

$$\pi(\theta_{0:t}, y_{1:t}) = \pi(\theta_0) \cdot \prod_{j=1}^{t} \pi(\theta_j|\theta_{j-1})\pi(y_j|\theta_j). \tag{7.1}$$

This distribution can be used to derive distributions of interest by conditioning or marginalization. If the states are discrete-valued random variables the state space models is sometimes called hidden Markov models [24, pp. 40-41].

7.1 Dynamic linear models

Dynamic linear models, also known as Gaussian linear state space models, are a particular class of state space models. Let $(Y_t)_{t\geq 1}$ be a time series, then a dynamic linear model is defined as follows.

Definition 7.2 (Dynamic linear model)
A dynamic linear model (DLM) is defined by a Normal prior distribution

$$\theta_0 \sim N_p(m_0, C_0) \tag{7.2a}$$

for the p-dimensional state space vector at time $t = 0$ and the pair of equations for each time $t \geq 1$,

$$Y_t = F_t\theta_t + v_t \quad , \quad v_t \sim N_m(0, V_t) \tag{7.2b}$$
$$\theta_t = G_t\theta_{t-1} + w_t \quad , \quad w_t \sim N_p(0, W_t) \tag{7.2c}$$

where F_t is a known $m \times p$-matrix and G_t is a known $p \times p$-matrix, and $(v_t)_{t\geq 1}$ and $(w_t)_{t\geq 1}$ are two independent sequences of independent Gaussian random vectors with zero mean and known variance matrices $(V_t)_{t\geq 1}$ and $(W_t)_{t\geq 1}$, respectively. The state vector at time $t = 0$, θ_0 is independent of (v_t) and (w_t). Equation (7.2b) *is the observation equation and equation* (7.2c) *is the state equation or system equation [24, p. 41].* ◇

In dynamic linear models it is assumed, that the distribution is Gaussian, which is true in many applications [24, p. 42].

7.2 State estimation and forecasting

State space models can be used in many applications. Let a model be specified by the densities $\pi(y_t|\theta_t)$ and $\pi(\theta_t|\theta_{t-1})$. The main objective of state space models is to make inference on unobserved states and predict future events based on past events. Calculating the conditional distribution given the previous informations is used for estimation and forecasting. The conditional densities $\pi(\theta_s|y_{1:t})$ are calculated to estimate the states. This involves three different tasks, *filtering* is when $s = t$, state *predicting* is when $s > t$, and *smoothing* is when $s < t$. Filtering is done by calculating the conditional density $\pi(\theta_t|y_{1:t})$, which for a DLM is carried out by the Kalman filter, that update the inference of the states when new observations are obtained. Filtering can be be implemented as a recursive algorithm. When smoothing the conditional distribution of $\theta_{1:t}$ given $y_{1:t}$ is calculated, which also can be implemented as a recursive algorithm.

The one-step-ahead forecasting is prediction of θ_{t+1} and Y_{t+1} based on the previous observations $y_{1:t}$. First the state θ_{t+1} is estimated and then the observation Y_{t+1} is estimated. The state predictive density of the one-step-ahead forecast is given by $\pi(\theta_{t+1}|y_{1:t})$, and the one-step-ahead predictive density is $\pi(y_{t+1}|y_{1:t})$. Similarly the k-step-ahead forecasts Y_{t+k} can be calculated by first estimating the state θ_{t+k} for $k \geq 1$ and then predicting Y_{t+k}. The state predictive density of the k-step-ahead forecast is given by $\pi(\theta_{t+k}|y_{1:t})$, and the k-step-ahead predictive density is $\pi(y_{t+k}|y_{1:t})$. The forecasts become more uncertain for large values of k [24, pp. 49-51].

7.2.1 Filtering

In this section the recursive steps in calculations of the filtering densities $\pi(\theta_t|y_{1:t})$ in a state space model is described. The filtering and predictive densities can be calculated recursively because of the assumptions in definition 7.1. The following proposition present the filtering recursive steps.

PROPOSITION 7.1 (FILTERING RECURSIONS)
For a general state space model specified by definition 7.1 the following is true

(i) The one-step-ahead predictive density for θ_t is calculated using the filtered density $\pi(\theta_{t-1}|y_{1:t-1})$, which gives

$$\pi(\theta_t|y_{1:t-1}) = \int \pi(\theta_t|\theta_{t-1})\pi(\theta_{t-1}|y_{1:t-1})d\theta_{t-1}. \tag{7.3a}$$

(ii) The one-step-ahead predictive density for Y_t is calculated using the predictive density for the states as

$$\pi(y_t|y_{1:t-1}) = \int \pi(y_t|\theta_t)\pi(\theta_t|y_{1:t-1})d\theta_t. \tag{7.3b}$$

(iii) The filtering density can be calculated using (i) and (ii)

$$\pi(\theta_t|y_{1:t}) = \frac{\pi(y_t|\theta_t)\pi(\theta_t|y_{1:t-1})}{\pi(y_t|y_{1:t-1})}. \tag{7.3c}$$

Proof. The state θ_t is conditionally independent of $Y_{1:t-1}$ given θ_{t-1}. Then (i) is proved by

$$\begin{aligned}
\pi(\theta_t|y_{1:t-1}) &= \int \pi(\theta_{t-1}, \theta_t|y_{1:t-1})d\theta_{t-1} \\
&= \int \pi(\theta_t|\theta_{t-1}, y_{1:t-1})\pi(\theta_{t-1}|y_{1:t-1})d\theta_{t-1} \\
&= \int \pi(\theta_t|\theta_{t-1})\pi(\theta_{t-1}|y_{1:t-1})d\theta_{t-1}.
\end{aligned}$$

The observation Y_t is conditionally independent of $Y_{1:t-1}$ given θ_t and (ii) is proved by

$$\begin{aligned}
\pi(y_t|y_{1:t-1}) &= \int \pi(y_t, \theta_t|y_{1:t-1})d\theta_t \\
&= \int \pi(y_t|\theta_t, y_{1:t-1})\pi(\theta_t|y_{1:t-1})d\theta_t \\
&= \int \pi(y_t|\theta_t)\pi(\theta_t|y_{1:t-1})d\theta_t.
\end{aligned}$$

The proof of (iii) uses Bayes' rule and the conditional independence of Y_t and $Y_{1:t-1}$ given θ_t. Then

$$\begin{aligned}
\pi(\theta_t|y_{1:t}) &= \frac{\pi(\theta_t|y_{1:t-1})\pi(y_t|\theta_t, y_{1:t-1})}{\pi(y_t|y_{1:t-1})} \\
&= \frac{\pi(\theta_t|y_{1:t-1})\pi(y_t|\theta_t)}{\pi(y_t|y_{1:t-1})}.
\end{aligned}$$

□

[24, pp. 51-53].

7.2.2 Kalman filter for DLM

Proposition 7.1 showed the principle of filtering and forecasting, but calculating the conditional distributions required is not always easy. One of the advantages of the DLM is that all the relevant distributions are Gaussian, which means that they can be completely determined by their means and variances. The following proposition shows this result.

Proposition 7.2 (Kalman filter)
Let a DLM be defined by definition 7.2, let the prior distribution be given by $\theta_0 \sim N(m_0, C_0)$*, and let*

$$\theta_{t-1}|y_{1:t-1} \sim N\left(m_{t-1}, C_{t-1}\right).$$

Then the following statements are true.

(i) The one-step-ahead predictive distribution of θ_t *given* $y_{1:t-1}$ *is Gaussian with parameters*

$$\begin{aligned} a_t &= E[\theta_t|y_{1:t-1}] = G_t m_{t-1}, \\ R_t &= Var[\theta_t|y_{1:t-1}] = G_t C_{t-1} G_t^T + W_t. \end{aligned}$$

(ii) The one-step-ahead predictive distribution of Y_t *given* $y_{1:t-1}$ *is Gaussian with parameters*

$$\begin{aligned} f_t &= E[Y_t|y_{1:t-1}] = F_t a_t, \\ Q_t &= Var[Y_t|y_{1:t-1}] = F_t R_t F_t^T + V_t. \end{aligned}$$

(iii) The filtering distribution of θ_t *given* $y_{1:t}$ *is Gaussian with parameters*

$$\begin{aligned} m_t &= E[\theta_t|y_{1:t}] = a_t + R_t F_t^T Q_t^{-1} e_t, \\ C_t &= Var[\theta_t|y_{1:t}] = R_t - R_t F_t^T Q_t^{-1} F_t R_t, \end{aligned}$$

where $e_t = Y_t - f_t$ *is the forecast error.*

Proof. The joint distribution of $(\theta_0, \theta_1, \ldots, \theta_t, Y_1, \ldots, Y_t)$ is given by equation (7.1), where the marginal and conditional distributions are Gaussian. Thus it follows that the joint distribution for the random vector $(\theta_0, \theta_1, \ldots, \theta_t, Y_1, \ldots, Y_t)$ is also Gaussian for any $t \geq 0$, and the distribution of any subvector or conditional distribution of some components given others is also Gaussian. This means, that it is enough to compute the means and variances of the predictive distributions and the filtering distributions, because they are Gaussian distributed.

Let $\theta_t|y_{1:t-1} \sim N(a_t, R_t)$, then (i) is proved using equation (7.2c), and a_t and R_t can be specified as

$$\begin{aligned} a_t &= \mathrm{E}[\theta_t|y_{1:t-1}] = \mathrm{E}\Big[\mathrm{E}[\theta_t|\theta_{t-1}, y_{1:t-1}]\Big|y_{1:t-1}\Big] \\ &= \mathrm{E}[G_t\theta_{t-1}|y_{1:t-1}] = G_t m_{t-1}, \end{aligned}$$

and

$$\begin{aligned} R_t &= \mathrm{Var}[\theta_t|y_{1:t-1}] \\ &= \mathrm{E}\Big[\mathrm{Var}[\theta_t|\theta_{t-1}, y_{1:t-1}]\Big|y_{1:t-1}\Big] + \mathrm{Var}\Big[\mathrm{E}[\theta_t|\theta_{t-1}, y_{1:t-1}]\Big|y_{1:t-1}\Big] \\ &= W_t + \mathrm{Var}[G_t\theta_{t-1}|y_{1:t-1}] \\ &= W_t + G_t C_{t-1} G_t^T. \end{aligned}$$

Let $Y_t|y_{1:t-1} \sim N(f_t, Q_t)$, then (ii) is proved using equation (7.2b), and f_t and Q_t can be specified as

$$\begin{aligned} f_t &= \mathrm{E}[Y_t|y_{1:t-1}] = \mathrm{E}\Big[\mathrm{E}[Y_t|\theta_t, y_{1:t-1}]\Big|y_{1:t-1}\Big] \\ &= \mathrm{E}[F_t\theta_t|y_{1:t-1}] = F_t a_t, \end{aligned}$$

and

$$\begin{aligned} Q_t &= \mathrm{Var}[Y_t|y_{1:t-1}] \\ &= \mathrm{E}\Big[\mathrm{Var}[Y_t|\theta_t, y_{1:t-1}]\Big|y_{1:t-1}\Big] + \mathrm{Var}\Big[\mathrm{E}[Y_t|\theta_t, y_{1:t-1}]\Big|y_{1:t-1}\Big] \\ &= V_t + \mathrm{Var}[F_t\theta_t|y_{1:t-1}] \\ &= V_t + F_t R_t F_t^T. \end{aligned}$$

The conditional covariance of Y_t and θ_t is given by

$$\begin{aligned} \mathrm{Cov}[Y_t, \theta_t|y_{1:t-1}] &= \mathrm{Cov}[F_t\theta_t + v_t, \theta_t|y_{1:t-1}] \\ &= F_t\mathrm{Cov}[\theta_t, \theta_t|y_{1:t-1}] + F_t\mathrm{Cov}[v_t, \theta_t|y_{1:t-1}] \\ &= F_t\mathrm{Var}[\theta_t|y_{1:t-1}] \\ &= F_t R_t, \end{aligned}$$

because $\mathrm{Cov}[v_t, \theta_t|y_{1:t-1}] = 0$ since v_t and θ_t are conditionally independent. From multivariate Gaussian theory the distribution of Y_t and θ_t given $y_{1:t-1}$ is

$$\begin{pmatrix} Y_t \\ \theta_t \end{pmatrix} \Big| y_{1:t-1} \Big) \sim N\left(\begin{pmatrix} f_t \\ a_t \end{pmatrix}, \begin{Bmatrix} Q_t & F_t R_t \\ R_t F_t^T & R_t \end{Bmatrix} \right).$$

This means that θ_t is Gaussian with the following expected value and variance conditionally on $y_{1:t}$

$$m_t = \mathrm{E}[\theta_t|y_t, y_{1:t-1}] = a_t + R_t F_t^T Q_t^{-1}(y_t - f_t),$$

and

$$C_t = \text{Var}[\theta_t | y_t, y_{1:t-1}] = R_t - R_t F_t^T Q_t^{-1} F_t R_t.$$

□

The predictive and filtering distributions can be calculated recursively using the Kalman filter beginning at $\theta_0 \sim N(m_0, C_0)$ and then calculate $\pi(\theta_1)$, $\pi(y_1)$ and $\pi(\theta_1|y_1)$, when new data is obtained [24, pp. 53-55].

Filtering with missing observations

Sometimes time series contain missing observations, which needs to be taking into account in the filtering recursions. A missing observation does not contain any information, so the observation at time t is missing, $y_t = NA$, and

$$\pi(\theta_t | y_{1:t}) = \pi(\theta_t | y_{1:t-1}). \tag{7.5}$$

Thus the filtering distribution at time t is the one-step-ahead predictive distribution at time $t-1$. This means that in the filtering recursion, proposition 7.1, equation (7.3c) has to be replaced with equation (7.5). The filtering distribution at time t for a DLM is given by setting the mean value $m_t = a_t$ and the variance $C_t = R_t$, because $\theta_t | y_{1:t-1} \sim N(a_t, R_t)$ [24, p. 59].

7.2.3 Smoothing

In state space models estimation and forecasting can be applied sequentially, when new data are collected. Often information about a time series Y_t is available for a period of time, and it is desired to retrospectively reconstruct the behavior of the system. This can be done using a backward recursive algorithm, which compute the conditional distribution of θ_t given $y_{1:T}$ for any $t < T$, starting from the filtering distribution $\pi(\theta_T | y_{1:T})$ and estimating the states backwards. This method is called smoothing and is presented for general state space models in the following proposition.

Proposition 7.3 (Smoothing recursion)
For a general state space model specified by definition 7.1 the following is true

(i) The state system $(\theta_0, \ldots, \theta_T)$ given $y_{1:T}$ has backward transition probabilities specified by

$$\pi(\theta_t | \theta_{t+1}, y_{1:T}) = \frac{\pi(\theta_{t+1} | \theta_t) \pi(\theta_t | y_{1:t})}{\pi(\theta_{t+1} | y_{1:t})}.$$

(ii) Conditional on $y_{1:T}$ the smoothing distribution of θ_t can be calculated using the backward recursion in t starting from $\pi(\theta_T|y_{1:T})$

$$\pi(\theta_t|y_{1:T}) = \pi(\theta_t|y_{1:t}) \int \frac{\pi(\theta_{t+1}|\theta_t)}{\pi(\theta_{t+1}|y_{1:t})} \pi(\theta_{t+1}|y_{1:T}) d\theta_{t+1}.$$

Proof. The state θ_t and the observation $Y_{t+1:T}$ are conditionally independent given θ_{t+1}, and given θ_t, θ_{t+1} and $Y_{1:T}$ are conditionally independent. Then (i) is proved using Bayes formula

$$\begin{aligned} \pi(\theta_t|\theta_{t+1}, y_{1:T}) &= \pi(\theta_t|\theta_{t+1}, y_{1:t}) \\ &= \frac{\pi(\theta_t|y_{1:t})\pi(\theta_{t+1}|\theta_t, y_{1:t})}{\pi(\theta_{t+1}|y_{1:t})} \\ &= \frac{\pi(\theta_t|y_{1:t})\pi(\theta_{t+1}|\theta_t)}{\pi(\theta_{t+1}|y_{1:t})}. \end{aligned}$$

The density $\pi(\theta_t, \theta_{t+1}|y_{1:T})$ with respect to θ_{t+1} is marginalized to prove (ii)

$$\begin{aligned} \pi(\theta_t|y_{1:T}) &= \int \pi(\theta_t, \theta_{t+1}|y_{1:T}) d\theta_{t+1} \\ &= \int \pi(\theta_{t+1}|y_{1:T}) \pi(\theta_t|\theta_{t+1}, y_{1:T}) d\theta_{t+1} \\ &= \int \pi(\theta_{t+1}|y_{1:T}) \frac{\pi(\theta_{t+1}|\theta_t)\pi(\theta_t|y_{1:t})}{\pi(\theta_{t+1}|y_{1:t})} d\theta_{t+1} \\ &= \pi(\theta_t|y_{1:t}) \int \pi(\theta_{t+1}|\theta_t) \frac{\pi(\theta_{t+1}|y_{1:T})}{\pi(\theta_{t+1}|y_{1:t})} d\theta_{t+1}. \end{aligned}$$

□

The smoothing recursion can be expressed using the means and variances of the smoothing distributions for a DLM. This is presented in the following proposition.

Proposition 7.4 (Kalman smoother)
Given a DLM as defined in definition 7.2, if $\theta_{t+1}|y_{1:T} \sim N(s_{t+1}, S_{t+1})$, then $\theta_t|y_{1:T} \sim N(s_t, S_t)$, where

$$\begin{aligned} s_t &= m_t + C_t G_{t+1}^T R_{t+1}^{-1}(s_{t+1} - a_{t+1}) \\ S_t &= C_t - C_t G_{t+1}^T R_{t+1}^{-1}(R_{t+1} - S_{t+1}) R_{t+1}^{-1} G_{t+1} C_t. \end{aligned}$$

Proof. The conditional distribution of θ_t given $y_{1:T}$ is Gaussian, because of properties of the multivariate Gaussian distribution, which means that it is

sufficient to calculate the mean and the variance. The mean value of θ_t given $y_{1:T}$ is

$$\begin{aligned} s_t = \mathrm{E}[\theta_t|y_{1:T}] &= \mathrm{E}\Big[\mathrm{E}[\theta_t|\theta_{t+1}, y_{1:T}]\Big|y_{1:T}\Big] \\ &= \mathrm{E}\Big[\mathrm{E}[\theta_t|\theta_{t+1}, y_{1:t}]\Big|y_{1:T}\Big], \end{aligned}$$

since θ_t and $y_{t+1:T}$ are conditional independent given θ_{t+1}, thus

$$\pi(\theta_t|\theta_{t+1}, y_{1:T}) = \pi(\theta_t|\theta_{t+1}, y_{1:t}).$$

The variance of θ_t given $y_{1:T}$ is

$$\begin{aligned} S_t = \mathrm{Var}[\theta_t|y_{1:T}] &= \mathrm{E}\Big[\mathrm{Var}[\theta_t|\theta_{t+1}, y_{1:T}]\Big|y_{1:T}\Big] + \mathrm{Var}\Big[\mathrm{E}[\theta_t|\theta_{t+1}, y_{1:T}]\Big|y_{1:T}\Big] \\ &= \mathrm{E}\Big[\mathrm{Var}[\theta_t|\theta_{t+1}, y_{1:t}]\Big|y_{1:T}\Big] + \mathrm{Var}\Big[\mathrm{E}[\theta_t|\theta_{t+1}, y_{1:t}]\Big|y_{1:T}\Big]. \end{aligned}$$

The likelihood $\pi(\theta_{t+1}|\theta_t, y_{1:t}) = \pi(\theta_{t+1}|\theta_t)$ is $\theta_{t+1}|\theta_t \sim N(G_{t+1}\theta_t, W_{t+1})$ given by the state equation (7.2c) and the prior $\theta_t|y_{1:t} \sim N(m_t, C_t)$. Because both the likelihood and prior are Gaussian distributed the posterior is also Gaussian. The expected value of θ_t given θ_{t+1} and $y_{1:t}$ can then be written as

$$\begin{aligned} &\mathrm{E}[\theta_t|\theta_{t+1}, y_{1:t}] \\ &\quad = \mathrm{E}[\theta_t|y_{1:t}] + \mathrm{Cov}[\theta_t, \theta_{t+1}|y_{1:t}]\mathrm{Var}[\theta_{t+1}|y_{1:t}]^{-1}(\theta_{t+1} - \mathrm{E}[\theta_{t+1}|y_{1:t}]) \\ &\quad = m_t + C_t G_{t+1}^T R_{t+1}^{-1}(\theta_{t+1} - a_{t+1}), \end{aligned}$$

since

$$\begin{aligned} \mathrm{Cov}[\theta_t, \theta_{t+1}|y_{1:t}] &= \mathrm{Cov}[\theta_t, G_{t+1}\theta_t + v_t|y_{1:t}] \\ &= \mathrm{Var}[\theta_t|y_{1:t}]G_{t+1}^T \\ &= C_t G_{t+1}^T. \end{aligned}$$

The variance of θ_t given θ_{t+1} and $y_{1:t}$ is

$$\begin{aligned} &\mathrm{Var}[\theta_t|\theta_{t+1}, y_{1:t}] \\ &\quad = \mathrm{Var}[\theta_t|y_{1:t}] - \mathrm{Cov}[\theta_t, \theta_{t+1}|y_{1:t}]\mathrm{Var}[\theta_{t+1}|y_{1:t}]^{-1}\mathrm{Cov}[\theta_t, \theta_{t+1}|y_{1:t}]^T \\ &\quad = C_t - C_t G_{t+1}^T R_{t+1}^{-1} G_{t+1} C_t^T, \end{aligned}$$

and

$$\mathrm{Var}\Big[\mathrm{E}[\theta_t|\theta_{t+1}, y_{1:t}]\Big|y_{1:T}\Big] = C_t G_{t+1}^T R_{t+1}^{-1}\mathrm{Var}[\theta_{t+1}|y_{1:T}]R_{t+1}^{-1} G_{t+1} C_t^T.$$

It is assumed that $\mathrm{E}[\theta_{t+1}|y_{1:T}] = s_{t+1}$ and $\mathrm{Var}[\theta_{t+1}|y_{1:T}] = S_{t+1}$. The mean value and variance of θ_t given $y_{1:T}$ is then

$$\begin{aligned} s_t &= \mathrm{E}\Big[\mathrm{E}[\theta_t|\theta_{t+1}, y_{1:t}]\Big|y_{1:T}\Big] \\ &= m_t + C_t G_{t+1}^T R_{t+1}^{-1}(s_{t+1} - a_{t+1}) \end{aligned}$$

and

$$\begin{aligned} S_t &= \mathrm{E}\Big[\mathrm{Var}[\theta_t|\theta_{t+1}, y_{1:t}]\Big|y_{1:T}\Big] + \mathrm{Var}\Big[\mathrm{E}[\theta_t|\theta_{t+1}, y_{1:t}]\Big|y_{1:T}\Big] \\ &= C_t - C_t G_{t+1}^T R_{t+1}^{-1} G_{t+1} C_t + C_t G_{t+1}^T R_{t+1}^{-1} S_{t+1} R_{t+1}^{-1} G_{t+1} C_t \\ &= C_t - C_t G_{t+1}^T R_{t+1}^{-1} (R_{t+1} - S_{t+1}) R_{t+1}^{-1} G_{t+1} C_t. \end{aligned}$$

□

The distribution of $\theta_t|y_{1:T}$ can be computed recursively backwards beginning at $t = T - 1$ and then proceed with $t = T - 2$ and so on, given that $\theta_t|y_{1:T} \sim N(s_T = m_T, S_T = C_T)$ [24, pp. 60-62],[25, p. 114].

7.2.4 Forecasting

In state space models forecasting future values of the observations, Y_{t+k}, or of the state vectors, θ_{t+k}, of a time series, when knowing the time series up to time t is one of the main tasks. The forecast can be computed recursively as new observations are obtained. The one-step-ahead predictive distribution for a general state space model was presented in proposition 7.1 and for a DLM in proposition 7.2. Sometimes one is interested in forecasting k-steps-ahead into the future. The following proposition, proposition 7.5, present the distributions of the state and the observation at time $t + k$. The filtering distribution at time t acts as an initial distribution for the forecast distribution. This means that in a state space model the distribution containing information about present and future states $(\theta_{t+k})_{k\geq 0}$ and future observations $(Y_{t+k})_{k\geq 1}$ are the conditional distributions $\pi(\theta_{t+k}|\theta_{t+k-1})$ and $\pi(y_{t+k}|\theta_{t+k})$, and the initial distribution $\pi(\theta_t|y_{1:t})$. For a DLM it is sufficient to calculate the mean and the variance of $\pi(\theta_t|y_{1:t})$, m_t and C_t, for prediction of future values. Figure 7.2 illustrates, that the observations

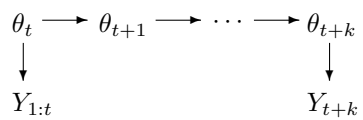


Figure 7.2: *Forecasting of states and observations*

$Y_{1:t}$ gives information about θ_t, which gives information about future values of the time series. In the next proposition, proposition 7.5, the forecast recursions for a general state space model is presented.

PROPOSITION 7.5 (FORECASTING RECURSION)
For a general state space model specified by definition 7.1 the following is true for any $k > 0$.

(i) The k-step-ahead forecast distribution of the state is

$$\pi(\theta_{t+k}|y_{1:t}) = \int \pi(\theta_{t+k}|\theta_{t+k-1})\pi(\theta_{t+k-1}|y_{1:t})d\theta_{t+k-1}.$$

(ii) The k-step-ahead forecast distribution of the observation is

$$\pi(y_{t+k}|y_{1:t}) = \int \pi(y_{t+k}|\theta_{t+k})\pi(\theta_{t+k}|y_{1:t})d\theta_{t+k}.$$

Proof. The state θ_{t+k} is conditionally independent of $Y_{1:t}$ given θ_{t+k-1}, and (i) is proved by

$$\begin{aligned}
\pi(\theta_{t+k}|y_{1:t}) &= \int \pi(\theta_{t+k}, \theta_{t+k-1}|y_{1:t})d\theta_{t+k-1} \\
&= \int \pi(\theta_{t+k}|\theta_{t+k-1}, y_{1:t})\pi(\theta_{t+k-1}|y_{1:t})d\theta_{t+k-1} \\
&= \int \pi(\theta_{t+k}|\theta_{t+k-1})\pi(\theta_{t+k-1}|y_{1:t})d\theta_{t+k-1}.
\end{aligned}$$

The observation Y_{t+k} is conditionally independent of $Y_{1:t}$ given θ_{t+k}, and (ii) is proved by

$$\begin{aligned}
\pi(y_{t+k}|y_{1:t}) &= \int \pi(y_{t+k}, \theta_{t+k}|y_{1:t})d\theta_{t+k} \\
&= \int \pi(y_{t+k}|\theta_{t+k}, y_{1:t})\pi(\theta_{t+k}|y_{1:t})d\theta_{t+k} \\
&= \int \pi(y_{t+k}|\theta_{t+k})\pi(\theta_{t+k}|y_{1:t})d\theta_{t+k}.
\end{aligned}$$

□

For DLM it is sufficient to calculate the means and variances of the forecast distributions, since all the distributions are Gaussian. The following proposition present the forecast recursions for a DLM.

PROPOSITION 7.6
Let a DLM be defined by definition 7.2, then the following is true for $k \geq 0$.

(i) The state $\theta_{t+k}|y_{1:t} \sim N(a_{t+k}, R_{t+k})$, with

$$\begin{aligned}
a_{t+k} &= E[\theta_{t+k}|y_{1:t}] = G_{t+k}a_{t+k-1}, \\
R_{t+k} &= Var[\theta_{t+k}|y_{1:t}] = G_{t+k}R_{t+k-1}G_{t+k}^T + W_{t+k},
\end{aligned}$$

where $a_{t+0} = m_t$ and $R_{t+0} = C_t$.

(ii) The observation $Y_{t+k}|y_{1:t} \sim N(f_{t+k}, Q_{t+k})$*, with*

$$\begin{aligned} f_{t+k} &= E[Y_{t+k}|y_{1:t}] = F_{t+k}a_{t+k}, \\ Q_{t+k} &= Var[Y_{t+k}|y_{1:t}] = F_{t+k}R_{t+k}F_{t+k}^T + V_{t+k}. \end{aligned}$$

Proof. The proposition is proved using induction. For $k = 1$ the proposition is equivalent to proposition 7.2 (i) and (ii), and thereby the result holds. For $k > 1$ (i) and (ii) is proved by

$$\begin{aligned} a_{t+k} &= \mathrm{E}[\theta_{t+k}|y_{1:t}] \\ &= \mathrm{E}\Big[\mathrm{E}[\theta_{t+k}|\theta_{t+k-1}, y_{1:t}]\Big|y_{1:t}\Big] \\ &= \mathrm{E}[G_{t+k}\theta_{t+k-1}|y_{1:t}] \\ &= G_{t+k}a_{t+k-1}, \\ R_{t+k} &= \mathrm{Var}[\theta_{t+k}|y_{1:t}] \\ &= \mathrm{Var}\Big[\mathrm{E}[\theta_{t+k}|\theta_{t+k-1}, y_{1:t}]\Big|y_{1:t}\Big] + \mathrm{E}\Big[\mathrm{Var}[\theta_{t+k}|\theta_{t+k-1}, y_{1:t}]\Big|y_{1:t}\Big] \\ &= \mathrm{Var}[G_{t+k}\theta_{t+k-1}|y_{1:t}] + W_{t+k} \\ &= G_{t+k}R_{t+k-1}G_{t+k}^T + W_{t+k}, \\ f_{t+k} &= \mathrm{E}[Y_{t+k}|y_{1:t}] \\ &= \mathrm{E}\Big[\mathrm{E}[Y_{t+k}|\theta_{t+k}, y_{1:t}]\Big|y_{1:t}\Big] \\ &= \mathrm{E}[F_{t+k}\theta_{t+k}|y_{1:t}] \\ &= F_{t+k}a_{t+k}, \\ Q_{t+k} &= \mathrm{Var}[Y_{t+k}|y_{1:t}] \\ &= \mathrm{Var}\Big[\mathrm{E}[Y_{t+k}|\theta_{t+k}, y_{1:t}]\Big|y_{1:t}\Big] + \mathrm{E}\Big[\mathrm{Var}[Y_{t+k}|\theta_{t+k}, y_{1:t}]\Big|y_{1:t}\Big] \\ &= \mathrm{Var}[F_{t+k}\theta_{t+k}|y_{1:t}] + V_{t+k} \\ &= F_{t+k}R_{t+k}F_{t+k}^T + V_{t+k}. \end{aligned}$$

□

The forecasts become more imprecise as k gets larger, because more uncertainty enters the system. In the following the forecast error is defined.

7.3 The innovation process and model checking

The one-step-ahead forecasts $f_t = \mathrm{E}[Y_t|Y_{1:t-1}]$ for DLMs are given by the Kalman filter, proposition 7.2, and the forecast error is defined as

$$\begin{aligned} e_t &= Y_t - \mathrm{E}[Y_t|Y_{1:t-1}] \\ &= Y_t - f_t \\ &= F_t\theta_t + v_t - F_t a_t \\ &= F_t(\theta_t - a_t) + v_t. \end{aligned}$$

In the following proposition different properties of the sequence $(e_t)_{t\geq 1}$ of forecast errors are presented.

Proposition 7.7
The sequence $(e_t)_{t\geq 1}$ of forecast errors of a DLMs has the following properties.

(i) The expected value of e_t is zero, i.e. $E[e_t] = 0$

(ii) The random error vector e_t is uncorrelated with any function of $Y_1, \ldots, Y_{t-1}$, i.e. $Cov[e_t, Z] = 0$, where $Z = g(Y_1, \ldots, Y_{t-1})$.

(iii) The error e_t and the observation Y_s are uncorrelated for any $s < t$.

(iv) The errors e_t and e_s are uncorrelated for any $s < t$.

(v) The error e_t is a linear function of $Y_1, \ldots, Y_t$.

(vi) The process $(e_t)_{t\geq 1}$ of forecast errors is a Gaussian process.

Proof. *(i)* The expected value of e_t is

$$\mathrm{E}[e_t] = \mathrm{E}\Big[\mathrm{E}[Y_t - f_t|Y_{1:t-1}]\Big] = 0$$

using that $\mathrm{E}[Y] = \mathrm{E}\Big[\mathrm{E}[Y|X]\Big]$.

(ii) The covariance of e_t and Z is

$$\begin{aligned} \mathrm{Cov}[e_t, Z] &= \mathrm{E}[e_t Z] \\ &= \mathrm{E}\Big[\mathrm{E}[e_t Z|Y_{1:t-1}]\Big] \\ &= \mathrm{E}\Big[\mathrm{E}[e_t|Y_{1:t-1}]Z\Big] \\ &= 0. \end{aligned}$$

(iii) For univariate observations it follows from *(ii)* by setting $Z = Y_s$. Or else use *(ii)* on each component of Y_s.

(*iv*) Again for univariate observations it follows from (*ii*) by setting $Z = e_s$. Or else use (*ii*) on each component of e_s.

(*v*) The error e_t is a linear function of $Y_1, \ldots, Y_t$, because $f_t = \mathrm{E}[Y_t|Y_{1:t-1}]$ is a linear function of $Y_1, \ldots, Y_{t-1}$, since $Y_1, \ldots, Y_t$ have a joint Gaussian distribution.

(*vi*) From (*v*) it follows that for any t the sequence $(e_1, \ldots, e_t)$ is a linear transformation of $(Y_1, \ldots, Y_t)$, which has a joint Gaussian distribution. Then $(e_1, \ldots, e_t)$ has a joint Gaussian distribution, and the process $(e_t)_{t\geq 1}$ is Gaussian, since all finite distributions are Gaussian.

□

The observation $Y_t = f_t + e_t$ is the sum of a component f_t, which is predictable from past observations, and a component, e_t, independent of the past and containing the new information given by Y_t. Because of this the forecast errors e_t are also called innovations.

The sequence of standardized innovations for univariate observations is given by

$$\widetilde{e}_t = \frac{e_t}{\sqrt{Q_t}},$$

which is Gaussian white noise. Gaussian white noise are independent, identically distributed random variables with zero mean. The standardized innovations can be used to check the model assumptions. This means that, if the model assumptions are correct, then the sequence $\widetilde{e}_1, \ldots, \widetilde{e}_t$ should be a sample of size t from a standard normal distribution. A QQ-plot can be used to check if the standard innovations are normal distributed, and the empirical autocorrelation function to check if they are uncorrelated [24, pp. 73-75].

7.4 Model specification

In this section different classes of DLMs are presented. These classes can be used alone or added together to model univariate time series. Specification of the model can be difficult, but one approach is to consider a time series as a combination of different components for instance trend or seasonality. Each component of the time series is represented by a DLM and then added together in a DLM. A univariate time series (Y_t) can be produced as the sum of independent components

$$Y_t = Y_{1,t} + \ldots + Y_{h,t},$$

where $Y_{i,t}$ represent a component such as trend or seasonality. A DLM can be used to describe the ith component $Y_{i,t}$ for $i = 1, \ldots, h$ as

$$\begin{aligned} Y_{i,t} &= F_{i,t}\theta_{i,t} + v_{i,t} \quad , \quad v_{i,t} \sim N(0, V_{i,t}) \\ \theta_{i,t} &= G_{i,t}\theta_{i,t-1} + w_{i,t} \quad , \quad w_{i,t} \sim N(0, W_{i,t}), \end{aligned}$$

where each of the p_i-dimensional state vectors $\theta_{i,t}$ are unique and for all $i \neq j$ the time series $(Y_{i,t}, \theta_{i,t})$ and $(Y_{j,t}, \theta_{j,t})$ are mutually independent. To achieve the DLM for the time series (Y_t) the independent components are added so $Y_t = \sum_{i=1}^{h} Y_{i,t}$ is expressed by the DLM

$$\begin{aligned} Y_t &= F_t\theta_t + v_t \quad , \quad v_t \sim N(0, V_t) \\ \theta_t &= G_t\theta_{t-1} + w_t \quad , \quad w_t \sim N_h(0, W_t), \end{aligned}$$

where the state vector is given by $\theta_t = [\theta_{1,t} \cdots \theta_{h,t}]^T$, $F_t = [F_{1,t}| \cdots |F_{h,t}]$, G_t and W_t are given by the block diagonal matrices

$$G_t = \begin{bmatrix} G_{1,t} & & \\ & \ddots & \\ & & G_{h,t} \end{bmatrix}, \quad W_t = \begin{bmatrix} W_{1,t} & & \\ & \ddots & \\ & & W_{h,t}, \end{bmatrix}$$

and $V_t = \sum_{i=1}^{j} V_{i,t}$ [24, pp. 88-89].

7.4.1 Trend models

Trend can be described as a smooth development of the series over time, which can be modeled using polynomial DLMs. A polynomial model of order n is defined as a DLM with constant matrices $F_t = F$ and $G_t = G$, which are determined by a forecast function. The forecast function gives the expected trend at time t and is defined as

$$f_{t+k} = \mathrm{E}[Y_{t+k}|y_{1:t}] = a_{t+0} + a_{t+1}k + \ldots + a_{t+n-1}k^{n-1} \ , \quad k \geq 0 \tag{7.6}$$

where the parameteres $a_{t+0}, \ldots, a_{t+n-1}$ are linear functions of m_t and are independent of k. This means that the forecast function is a polynomial of order $n-1$, and the forecast function can also be written as

$$\begin{aligned} f_{t+k} &= \mathrm{E}[Y_{t+k}|y_{1:t}] \\ &= Fa_{t+k} \\ &= FGa_{t+k-1} \\ &= \ldots \\ &= FG^{k-1}a_{t+1}. \end{aligned} \tag{7.7}$$

Usually small values of n are used. The polynomial model of order $n = 1$ is the random walk plus noise model, and the polynomial model of order $n = 2$ is the linear growth model [24, p. 89].

Random walk plus noise

The random walk plus noise model, also known as a *local level model*, is a polynomial model of order $n = 1$, given by the equations

$$\begin{aligned} Y_t &= \theta_t + v_t \quad , \quad v_t \sim N(0, V) \\ \theta_t &= \theta_{t-1} + w_t \quad , \quad w_t \sim N(0, W), \end{aligned}$$

where (v_t) and (w_t) are independent, and $F_t = G_t = 1$. The observations (Y_t) are modeled as a level θ_t plus noise v_t, which is affected by random changes over time [24, p. 42].

The Kalman filter of the random walk plus noise model is given by the one-step-ahead predictive distribution of θ_t given $y_{1:t-1}$

$$\theta_t | y_{1:t-1} \sim N(m_{t-1}, R_t = C_{t-1} + W),$$

the one-step-ahead predictive distribution of Y_t given $y_{1:t-1}$

$$Y_t | y_{1:t-1} \sim N(f_t = m_{t-1}, Q_t = R_t + V),$$

and the filtering distribution of θ_t given $y_{1:t}$

$$\theta_t | y_{1:t} \sim N(m_t = m_{t-1} + K_t e_t, C_t = K_t V),$$

where $e_t = Y_t - f_t$ and $K_t = R_t Q_t^{-1}$. The behavior of the process (Y_t) is affected by the signal-to-noise ratio, which is the ratio between the two error variances $r = W/V$. The structure of the estimation and forecasting reflect this ratio. The mean value of the filtering distribution can be written as

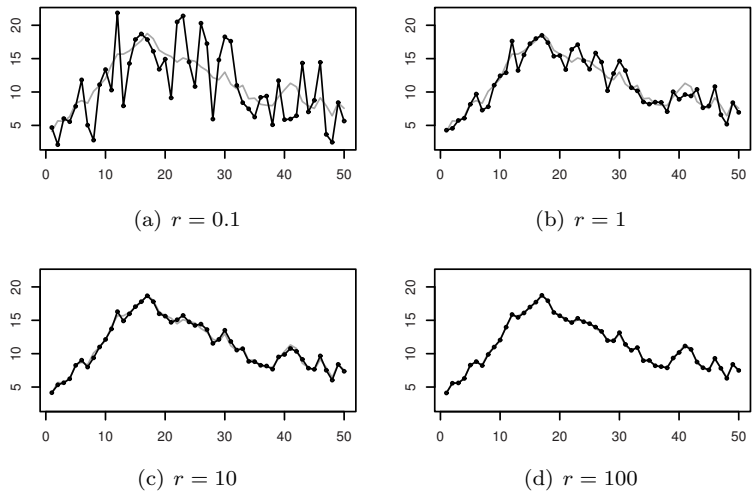
$$\begin{aligned} m_t &= m_{t-1} + K_t e_t = m_{t-1} + K_t(y_t - f_t) \\ &= m_{t-1} + K_t(y_t - m_{t-1}) = m_{t-1} + K_t y_t - K_t m_{t-1} \\ &= K_t y_t + (1 - K_t) m_{t-1}. \end{aligned}$$

This shows that m_t is a weighted average of y_t and m_{t-1}. The weight of the current observation y_t is called the adaptive coefficient and is given by

$$K_t = \frac{R_t}{Q_t} = \frac{C_{t-1} + W}{C_{t-1} + W + V},$$

where $0 < K_t < 1$. For any C_0, the adaptive coefficient K_t is small and the observation y_t receives little weight, if the signal-to-noise ratio is small. On the other hand if the variance of the error of the observation $V = 0$, then $K_t = 1$, and the one-step-ahead forecast is given by the previous data point, thus $m_t = y_t$ [24, pp. 55-56].

In summary if the signal-to-noise ratio is small, the model is trusted more than new data points, and if the signal-to-noise ratio is large, new data points are



(a) $r = 0.1$ (b) $r = 1$

(c) $r = 10$ (d) $r = 100$

Figure 7.3: *Random walk plus noise models for different values of the signal-to-noise ratio. The observations* (Y_t) *is the black line and the state* (θ_t) *is the gray line [24, p. 90].*

trusted more than the model. This is illustrated in figure 7.3, where a random walk plus noise model is shown with different signal-to-noise ratios. In figure 7.3(a) the signal-to-noise ratio is small, and the model is trusted more than the data, but in figure 7.3(d) the signal-to-noise ratio is large, and the data is trusted more than the model.

The k-step-ahead predictive distribution is given by

$$Y_{t+k}|y_{1:t} \sim N(m_t, Q_{t+k}) \quad , \; k \geq 1,$$

where

$$\begin{aligned} Q_{t+k} &= C_t + \sum_{j=1}^{k} W_{t+j} + V_{t+k} \\ &= C_t + kW + V, \end{aligned}$$

since the terms in the error sequence (w_t) are independent. The forecast function of the model is constant, $f_{t+k} = \mathrm{E}[Y_{t+k}|y_{1:t}] = m_t$, which is why the model is also called a steady model. The variance Q_{t+k} contain the uncertainty on the future observations, which increases linearly as k gets larger. The random walk plus noise model can be used to model data without any clear trend or seasonal variability [24, p. 90].

7.4.2 Linear growth model

A linear growth model is also called a local linear trend model. The model is similar to the random walk plus noise model, but a time-varying slope is included in the state dynamics. This is defined in the following equations

$$\begin{aligned} Y_t &= \mu_t + v_t, & v_t &\sim N(0, V) \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + w_{t,1}, & w_{t,1} &\sim N(0, \sigma_\mu^2), \\ \beta_t &= \beta_{t-1} + w_{t,2}, & w_{t,2} &\sim N(0, \sigma_\beta^2), \end{aligned}$$

where μ_t is the local level, β_t is the local growth rate and the errors (v_t), $(w_{t,1})$ and $(w_{t,2})$ are independent. The local level μ_t is a cumulated random walk. This model is a polynomial DLM of order 2, where the parameters are given by

$$\theta_t = \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix} \quad , \quad G = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad , \quad W = \begin{bmatrix} \sigma_\mu^2 & 0 \\ 0 & \sigma_\beta^2 \end{bmatrix} \quad , \quad F = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

The variances of the system σ_μ^2 and σ_β^2 can be equal to zero. In the model a current level μ_t is assumed to change linearly over time, and the growth rate can also change. Let $m_{t-1} = \begin{bmatrix} \widehat{\mu}_{t-1} & \widehat{\beta}_{t-1} \end{bmatrix}^T$, then the one-step-ahead point forecasts and filtering state estimates are

$$\begin{aligned} a_t &= \mathrm{E}[\theta_t | y_{1:t-1}] = G m_{t-1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \widehat{\mu}_{t-1} \\ \widehat{\beta}_{t-1} \end{bmatrix} = \begin{bmatrix} \widehat{\mu}_{t-1} + \widehat{\beta}_{t-1} \\ \widehat{\beta}_{t-1} \end{bmatrix}, \\ f_t &= \mathrm{E}[Y_t | y_{1:t-1}] = F_t a_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \widehat{\mu}_{t-1} + \widehat{\beta}_{t-1} \\ \widehat{\beta}_{t-1} \end{bmatrix} = \widehat{\mu}_{t-1} + \widehat{\beta}_{t-1}, \end{aligned}$$

and

$$m_t = a_t + K_t e_t = \begin{bmatrix} \widehat{\mu}_{t-1} + \widehat{\beta}_{t-1} \\ \widehat{\beta}_{t-1} \end{bmatrix} + \begin{bmatrix} k_{t1} \\ k_{t2} \end{bmatrix} e_t = \begin{bmatrix} \widehat{\mu}_{t-1} + \widehat{\beta}_{t-1} + k_{t1} e_t \\ \widehat{\beta}_{t-1} + k_{t2} e_t \end{bmatrix}$$

where $K_t = \begin{bmatrix} k_{t1} & k_{t2} \end{bmatrix}^T$. The forecast function is a linear function of k given by equation (7.7)

$$f_{t+k} = F G^{k-1} a_{t+1} = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & k-1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \widehat{\mu}_t + \widehat{\beta}_t \\ \widehat{\beta}_t \end{bmatrix} = \widehat{\mu}_t + k\widehat{\beta}_t$$

[24, pp. 42,96].

7.4.3 Seasonal models

Time series showing cyclical behavior can be modeled using Fourier-form seasonal models. Consider a discrete-time periodic function with period s, which is

described by the values it takes at time $t = 1, 2, \ldots, s$. Let $g_t = \alpha_t$, $t = 1, \ldots, s$, be a function, where the values of g_t is repeated every s-times, so $g_{s+1} = \alpha_1$, $g_{s+2} = \alpha_2$ and so on. Thus the periodic function g_t is related to a s-dimensional vector $\alpha = [\alpha_1, \ldots, \alpha_s]^T$. Assume that s is even and let the Fourier frequencies be defined as

$$\omega_j = \frac{2\pi j}{s}\,, \qquad j = 0, 1, \ldots, \frac{s}{2},$$

and let s s-dimensional vectors be given by

$$\left.\begin{aligned} e_0 &= [1, 1, \ldots, 1]^T \\ c_1 &= [\cos(\omega_1), \cos(2\omega_1), \ldots, \cos(s\omega_1)]^T \\ s_1 &= [\sin(\omega_1), \sin(2\omega_1), \ldots, \sin(s\omega_1)]^T \\ &\vdots \\ c_j &= [\cos(\omega_j), \cos(2\omega_j), \ldots, \cos(s\omega_j)]^T \\ s_j &= [\sin(\omega_j), \sin(2\omega_j), \ldots, \sin(s\omega_j)]^T \\ &\vdots \\ c_{s/2} &= [\cos(\omega_{s/2}), \cos(2\omega_{s/2}), \ldots, \cos(s\omega_{s/2})]^T, \end{aligned}\right\} \tag{7.8}$$

where $s_{s/2}$ is a vector of zeros and is therefore not considered, and $c_{s/2} = [-1, 1, -1, \ldots, -1, 1]^T$. The vectors in equation (7.8) are orthogonal and thereby a basis, which means that every vector in $\mathbb{R}^s$ can be written as a linear combination of $e_0, c_1, s_1, \ldots, c_{s/2}$, so

$$\alpha = a_0 e_0 + \sum_{j=1}^{s/2-1} (a_j c_j + b_j s_j) + a_{s/2} c_{s/2}.$$

This representation is helpful, since the basis vectors can be extended to periodic functions very easily because of the trigonometric functions. For example the tth component of s_j is $s_j(t) = \sin(2\pi tj/2)$ for any $1 \leq t \leq s$. The function s_j extended to a periodic function is $s_j(t) = s_j(t + ks)$, so

$$\sin\left(\frac{2\pi(t+ks)j}{s}\right) = \sin\left(\frac{2\pi tj}{s} 2\pi kj\right) = \sin\frac{2\pi tj}{s}.$$

This means that the extension of $s_j(t)$ to any integer t is given by putting t in the trigometric expression defining s_j. The representation of the basis vectors in equation (7.8) goes from smooth to rough, where the constant vector e_0 correspond to a constant periodic function and $c_{s/2}$ is the maximal oscillation, where the periodic function goes back and forth between -1 and 1. The expression $\frac{2\pi tj}{s}$, for any j, extends the interval $(0, 2\pi j]$ and the cosine function c_j runs through a complete period j times as t goes from 1 to s. This means that as j gets higher the rougher c_j is, and the function oscillates more frequent. The

functions s_j and c_j with the same oscillation frequency can be grouped together, and the jth harmonic of g_t can be defined as

$$S_j(t) = a_j \cos(t\omega_j) + b_j \sin(t\omega_j), \quad j = 1, \ldots, s/2,$$

where $b_{s/2} = 0$. Because the mean is modeled separately from the seasonal component, a_0 is set to zero, and since the basis vectors are orthogonal the sum of any harmonic over an entire period is zero. Then the function g_t can be written as

$$g_t = \sum_{j=1}^{s/2} S_j(t). \tag{7.9}$$

Let j be fixed, then the development of S_j as time goes from t to $t+1$ is

$$S_j(t) \longmapsto S_j(t+1) = a_j \cos((t+1)\omega_j) + b_j \sin((t+1)\omega_j).$$

If $j < s/2$ then it is impossible to calculate $S_j(t+1)$ from $S_j(t)$ without knowing a_j and b_j. But if the conjugate harmonic

$$S_j^*(t) = -a_j \sin(t\omega_j) + b_j \cos(t\omega_j)$$

is also known, then $S_j(t+1)$ can be calculated explicitly. So

$$\begin{aligned}
S_j(t+1) &= a_j \cos((t+1)\omega_j) + b_j \sin((t+1)\omega_j) \\
&= a_j \cos(t\omega_j + \omega_j) + b_j \sin(t\omega_j + \omega_j) \\
&= a_j(\cos(t\omega_j)\cos(\omega_j) - \sin(t\omega_j)\sin(\omega_j)) \\
&\quad + b_j(\sin(t\omega_j)\cos(\omega_j) + \cos(t\omega_j)\sin(\omega_j)) \\
&= a_j \cos(t\omega_j)\cos(\omega_j) - a_j \sin(t\omega_j)\sin(\omega_j) \\
&\quad + b_j \sin(t\omega_j)\cos(\omega_j) + b_j \cos(t\omega_j)\sin(\omega_j) \\
&= (a_j \cos(t\omega_j) + b_j \sin(t\omega_j))\cos(\omega_j) \\
&\quad + (-a_j \sin(t\omega_j) + b_j \cos(t\omega_j))\sin(\omega_j) \\
&= S_j(t)\cos(\omega_j) + S_j^*(t)\sin(\omega_j).
\end{aligned} \tag{7.10}$$

Similarly the conjugate harmonic S_j^* can be computed as

$$S_j^*(t+1) = S_j^*(t)\cos(\omega_j) + S_j(t)\sin(\omega_j). \tag{7.11}$$

By merging equation (7.10) and equation (7.11) then

$$\begin{bmatrix} S_j(t+1) \\ S_j^*(t+1) \end{bmatrix} = \begin{bmatrix} \cos(\omega_j) & \sin(\omega_j) \\ -\sin(\omega_j) & \cos(\omega_j) \end{bmatrix} \begin{bmatrix} S_j(t) \\ S_j^*(t) \end{bmatrix}.$$

The jth harmonic fits in a DLM by considering the bivariate state vector $[S_j(t), S_j^*(t)]^T$ with evolution matrix

$$H_j = \begin{bmatrix} \cos(\omega_j) & \sin(\omega_j) \\ -\sin(\omega_j) & \cos(\omega_j) \end{bmatrix}$$

and observation matrix $F = \begin{bmatrix} 1 & 0 \end{bmatrix}$. When $j = s/2$, then $S_{s/2}$ changes sign for every unit time increase, since

$$S_{s/2}(t+1) = \cos((t+1)\pi) = -\cos(t\pi) = -S_{s/2}(t).$$

In a DLM this can be seen as s univariate state vector with evolution matrix $H_{s/2} = [-1]$ and observation matrix $F = [1]$. To obtain a DLM representation the different harmonics can be combined to attain equation (7.9). This is managed by considering the state vector

$$\theta_t = (S_1(t), S_1^*, \ldots, S_{\frac{s}{2}-1}(t), S_{\frac{s}{2}-1}(t)^*, S_{\frac{s}{2}}(t))^T, \quad t = 0, 1, \ldots$$

with the evolution matrix

$$G = \begin{bmatrix} H_1 & & 0 \\ & \ddots & \\ 0 & & H_{\frac{s}{2}} \end{bmatrix},$$

and the observation matrix

$$F = \begin{bmatrix} 1 & 0 & 1 & 0 & \ldots & 0 & 1 \end{bmatrix}.$$

By letting all the evolution and observation variances be zero, a DLM representation of the seasonal component is achieved [24, pp. 102-106].

7.5 Estimation of unknown parameters

Until now it has been assumed that the system matrices F_t, G_t, V_t and W_t were known. This is not always the case, and if not the model matrices are set to depend on a vector of unknown parameters ψ. The vector ψ can be estimated using maximum likelihood estimation, MLE.

Let $Y_1, \ldots, Y_n$ be random vectors having distributions depending on an unknown parameter ψ. The joint density of the observations for a specific value of the parameters is given by $p(y_1, \ldots, y_n; \psi)$ and the likelihood of ψ is

$$L(\psi) = c \cdot p(y_1, \ldots, y_n; \psi),$$

where c is a constant. For a DLM the joint density of the observations can be written in terms of the conditional density of y_t given $y_{1:t-1}$ and the unknown parameter ψ, so

$$p(y_1, \ldots, y_n; \psi) = \prod_{t=1}^{n} p(y_t | y_{1:t-1}; \psi).$$

The conditional densities are Gaussian with mean f_t and variance Q_t, which implicitly depend on ψ, thus the log-likelihood can be expressed as

$$\ell(\psi) = -\frac{1}{2}\sum_{t=1}^{n}\log|Q_t| - \frac{1}{2}\sum_{t=1}^{n}(y_t - f_t)^T Q_t^{-1}(y_t - f_t). \tag{7.12}$$

Equation (7.12) is numerically maximized to attain the MLE of the unknown parameters

$$\widehat{\psi} = \underset{\psi}{\operatorname{argmax}}\ \ell(\psi).$$

Choosing the right starting point is essential in the numerical optimization. Because the log-likelihood function for a DLM expressed in equation (7.12) can have many local maxima, the MLE should be obtained using different starting values and then compared [24, pp. 143-144]. Other methods can also be used for estimation of the unknown parameters e.g. the expectation-maximization algorithm, EM-algorithm, or Markov chain Monte Carlo, MCMC, methods.

Chapter 8

Multi-process models

In the previous chapter it is assumed that a single DLM is sufficient to represent the behavior of the entire time series, but in some situations one DLM is insufficient for describing the behavior of the time series. Then various DLMs can be considered, and a combination of different DLMs is called a multi-process model. Multi-process models are also called mixture models, because they combine effects using mixtures of DLMs.

A DLM is defined in definition 7.2 by the matrices F_t, G_t, V_t and W_t, which here are denoted M_t. When defining a DLM there can be uncertainty about some parameters in the model if the time series contain growth or outliers. A parameter α includes these uncertainties, and the choice of α then defines the model. The dependence of the model on these uncertainties is denoted $M_t = M_t(\alpha)$, $t = 1, 2, \ldots$, and for each time t and α, $M_t(\alpha)$ is a standard DLM as defined in definition 7.2. Let the set of all possible values of α be denoted $\mathcal{A}$, then the set of all possible DLMs at time t is

$$\{M_t(\alpha) \ : \ \alpha \in \mathcal{A}\}.$$

There are two different types of multiprocess models. The first model, multi-process model class I, satisfy the following

- $M_t(\alpha_0)$ holds for all t, $\alpha_0 \in \mathcal{A}$.

Here α is constant over time, thus a single DLM is appropriate for all time, but the exact value of α is uncertain. The second model, multi-process model class II, satisfy

- $M_t(\alpha_t)$ holds at time t for some sequence of values $\alpha_t \in \mathcal{A}$.

In this model different values of α are appropriate at different times, and thereby no single DLM can describe the entire time series [25, pp. 427-428]. The two types of multi-process models are presented in the following sections.

8.1 Multi-process model class I

Definition 8.1 (Multi-process model class I)
Let the discrete set $\mathcal{A}$ be the parameter space for α, and assume that there exists an $\alpha \in \mathcal{A}$ so $M_t(\alpha)$ holds for all t. The time series Y_t then follow a multi-process model class I. ◇

Here α is constant over time, but the exact value is not known. The DLM $M_t(\alpha_t)$ can be analysed as described in chapter 7 given any specific value of $\alpha_t \in \mathcal{A}$. The filtering distribution of θ_t is given by $\pi(\theta_t|\alpha_t, y_{1:t})$, the smoothing distribution of θ_t is $\pi(\theta_t|\alpha_t, y_{1:T})$ and the predictive distribution of Y_{t+k}, $k > 0$, is $\pi(Y_{t+k}|\alpha_{t+k}, y_{1:t})$. The problem is that the exact value of α_t is not known, but this can be solved using the marginal distribution.
The marginal distribution of the filtering distribution is

$$\pi(\theta_t|y_{1:t}) = \sum_{\alpha_t \in \mathcal{A}} \pi(\theta_t|\alpha_t, y_{1:t})\pi(\alpha_t|y_{1:t}), \tag{8.1a}$$

where $\pi(\theta_t|\alpha_t, y_{1:t})$ is the distribution given model α_t, and the probability $\pi(\alpha_t|y_{1:t})$ is a weight. For the smoothing distribution the marginal distribution is given by

$$\pi(\theta_t|y_{1:T}) = \sum_{\alpha_t \in \mathcal{A}} \pi(\theta_t|\alpha_t, y_{1:T})\pi(\alpha_t|y_{1:T}), \tag{8.1b}$$

and the marginal distribution of the predictive distribution is

$$\pi(Y_{t+k}|y_{1:t}) = \sum_{\alpha_{t+k} \in \mathcal{A}} \pi(Y_{t+k}|\alpha_{t+k}, y_{1:t})\pi(\alpha_{t+k}|y_{1:t}). \tag{8.1c}$$

Inference about α_t and α_{t+k} is obtained using the probabilities $\pi(\alpha_t|y_{1:t})$ and $\pi(\alpha_{t+k}|y_{1:t})$ at time t and $t+k$. These probabilities are recursively calculated beginning with a prior distribution of α_t and updated using Bayes theorem

$$\begin{aligned} \pi(\alpha_t|y_{1:t}) &= \pi(\alpha_t|y_t, y_{1:t-1}) \\ &= \frac{\pi(\alpha_t|y_{1:t-1})\pi(y_t|\alpha_t, y_{1:t-1})}{\sum_{\alpha \in \mathcal{A}} \pi(y_t|\alpha, y_{1:t-1})\pi(\alpha|y_{1:t-1})}, \end{aligned} \tag{8.2}$$

where $\pi(y_t|\alpha_t, y_{1:t-1})$ is the one-step-ahead predictive distribution given by the Kalman filter, proposition 7.2.

To summarize the uncertainty about α requires the following calculations.

i) Calculate each of the quantities of interest, i.e. the filtering distribution $\pi(\theta_t|\alpha_t, y_{1:t})$, the smoothing distribution $\pi(\theta_t|\alpha_t, y_{1:T})$ or the predictive distribution $\pi(Y_{t+k}|\alpha_{t+k}, y_{1:t})$, for all $\alpha_t, \alpha_{t+k} \in \mathcal{A}$.

ii) Calculate equation (8.2) for each $\alpha_t \in \mathcal{A}$.

iii) Calculate for each of the quantities of interest the corresponding marginal distribution i.e. the marginal filtering distribution, equation (8.1a), the marginal smoothing distribution, equation (8.1b) or the marginal predictive distribution, equation (8.1c).

The set $\mathcal{A}$ in definition 8.1 is a discrete set, and the calculations in i) to iii) can be carried out without problems. If the parameter space $\mathcal{A}$ is continuous, there are an infinite number of possibilities for α_t and the calculations in i) to iii) can be difficult. When $\mathcal{A}$ is continuous, a discrete set that spans the parameter space can be chosen, and the calculations can be performed.

The distributions in equations (8.1) are discrete probability mixtures as they are sums of multiple distributions, and because the distributions are normal, they are probability mixtures of k normal distributions. The mean value and the variance of the marginal filtering distribution can be calculated using the properties of probability mixtures. Let $h_t = \mathrm{E}[\theta_t|y_{1:t}]$, $H_t = \mathrm{Var}[\theta_t|y_{1:t}]$, $h_t(\alpha_t) = \mathrm{E}[\theta_t|\alpha_t, y_{1:t}]$ and $H_t(\alpha_t) = \mathrm{Var}[\theta_t|\alpha_t, y_{1:t}]$ for $\alpha_t \in \mathcal{A}$. The mean value for probability mixtures is calculated as

$$\mathrm{E}[\theta_t|y_{1:t}] = \sum_{\alpha_t \in \mathcal{A}} \mathrm{E}[\theta_t|\alpha_t, y_{1:t}]\pi(\alpha_t|y_{1:t}),$$

thus the mean value of the marginal filtering distribution is given by

$$h_t = \mathrm{E}[\theta_t|y_{1:t}] = \sum_{\alpha_t \in \mathcal{A}} h_t(\alpha_t)\pi(\alpha_t|y_{1:t}). \tag{8.3}$$

The variance of probability mixtures is given by

$$\mathrm{Var}[\theta_t|y_{1:t}] = \sum_{\alpha_t \in \mathcal{A}} \Big(\mathrm{Var}[\theta_t|\alpha_t, y_{1:t}] + \mathrm{Var}\Big[\mathrm{E}[\theta_t|\alpha_t, y_{1:t}]\Big]\Big)\, \pi(\alpha_t|y_{1:t})$$

and thereby the variance of the marginal filtering distribution is

$$\begin{aligned} H_t &= \mathrm{Var}[\theta_t|y_{1:t}] \\ &= \sum_{\alpha_t \in \mathcal{A}} \Big(H_t(\alpha_t) + \mathrm{Var}\Big[\mathrm{E}[\theta_t|\alpha_t, y_{1:t}]\Big]\Big)\, \pi(\alpha_t|y_{1:t}) \\ &= \sum_{\alpha_t \in \mathcal{A}} \big(H_t(\alpha_t) + [h_t(\alpha_t) - h_t][h_t(\alpha_t) - h_t]^T\big)\, \pi(\alpha_t|y_{1:t}). \end{aligned} \tag{8.4}$$

Similarly the means and variances of the marginal smoothing distribution and the marginal predictive distribution can be calculated.

In the posterior probabilities, equation (8.1), the supported values of α are identified, and those with high weights, i.e. $\pi(\alpha_j|y_{1:t})$ is high, influence the multi-process more than the values with low weights. These probabilities can change over time, but under general conditions all the posterior probabilities will converge to zero except one value in $\mathcal{A}$, this probability will go towards one. A large number of samples are often necessary to identify a single value of α, thus a single DLM. When α is continuous, the value in the chosen discrete set that spans $\mathcal{A}$ closest to the true value is identified. A mixture model is

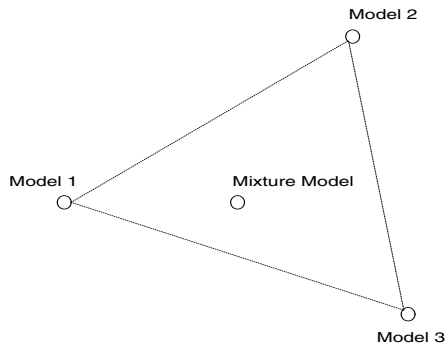


Figure 8.1: *Illustration of a mixture model [25, p. 433].*

illustrated in figure 8.1, where the generic model $M_t(\alpha)$ is in the middle of the three mixture components $M_t(\alpha_j)$, $j = 1, 2, 3$, and is recognized as a mixture of the three components with respect to particular values of the posterior model probabilities. This means that, as is often the case, no single DLM constructs the time series, but a mixture of components let the probabilities change to adapt to changes in the process [25, pp. 427-433].

8.2 Multi-process model class II

In this section multi-process models class II and their properties will be presented.

Definition 8.2 (Multi-process model class II)
Let the discrete set $\mathcal{A}$ be the parameter space for α, and suppose that at each time t, there exists an $\alpha \in \mathcal{A}$ so $M_t(\alpha)$ holds. The time series Y_t then follow a multi-process model class II. ◇

In this type of model α can take different values at different times, and it is assumed that no single DLM can describe the entire time series. Instead a discrete collection of DLMs can be used, which is often the situation in practice. In theory α can take a different value for each observation and the possibilities for multi-process model class II are enormous, but often a small number of α values are used for easy interpretation and calculations.
In the definition of a multi-process model class II, definition 8.2, it is not defined how α should be selected at each time. Here the focus is on a type of model called multi-process model class II mixture, where α is selected at time t with known probability.

Definition 8.3 (Multi-process model class II mixture)
Suppose that in a multi-process model class II, definition 8.2, $\alpha = \alpha_t$ at time t, which defines the model $M_t(\alpha_t)$, is selected with known probability

$$\pi\left(\alpha_t|\alpha_{1:t-1}, y_{1:t-1}\right) = \pi\left(M_t\left(\alpha_t\right)|\{M_s(\alpha_s)\}_{s=1}^{t-1}, y_{1:t-1}\right),$$

where $\{M_s(\alpha_s)\}_{s=1}^{t-1}$ are the models from time 1 to $t-1$. The time series Y_t then follow a multi-process model class II mixture. ◇

Here the probability $\pi(\alpha_t|\alpha_{1:t-1}, y_{1:t-1})$ depends on the history of the time series.

In the following special cases of the multi-process model class II mixture are presented.

1) Fixed probability model:
The probabilities with which the model is selected are fixed so

$$\pi(\alpha_t|\alpha_{1:t-1}, y_{1:t-1}) = \pi(\alpha_t), \quad \alpha_t \in \mathcal{A},$$

for all t. This means that the probabilities are constant and are independent of what has previously occurred. The dependence structure of a fixed probability model is illustrated in figure 8.2.

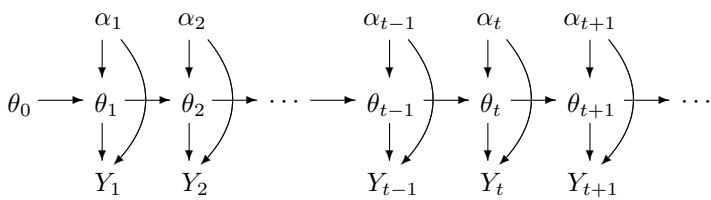


Figure 8.2: *Dependence structure of the fixed probability model.*

2) First-order Markov model:
The probabilities with which the model is selected are first-order Markov i.e. the model at time t depends only on the model at time $t-1$. The transition probabilities are known constants, so

$$\pi(\alpha_t|\alpha_{1:t-1}, y_{1:t-1}) = \pi(\alpha_t|\alpha_{t-1}), \quad \alpha_t, \alpha_{t-1} \in \mathcal{A},$$

and the prior probabilities

$$\pi(\alpha_1) = \pi(M_1(\alpha_1))$$

are known for $\alpha_1 \in \mathcal{A}$. The dependence structure of the first-order Markov model is illustrated in figure 8.3.

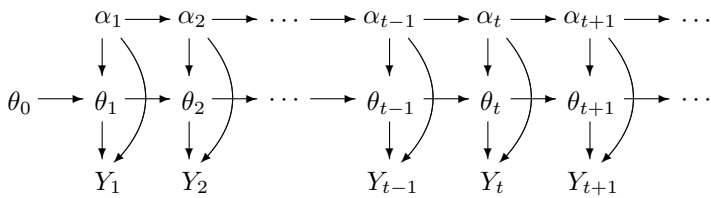


Figure 8.3: *Dependence structure of the first-order Markov model.*

The subsequent analysis will focus on the second type of model [25, pp. 443-445].

8.2.1 First-order Markov

The joint distribution of $(\theta_0, \theta_1, \ldots, \theta_t, Y_1, \ldots, Y_t, \alpha_1, \ldots, \alpha_t)$ for any $t > 0$ is then given by

$$\pi(\theta_{0:t}, y_{1:t}, \alpha_{1:t}) = \pi(\theta_0) \cdot \prod_{j=1}^{t} \pi(y_j|\theta_j, \alpha_j)\pi(\theta_j|\theta_{j-1}, \alpha_j)\pi(\alpha_j|\alpha_{j-1}), \qquad (8.5)$$

where $\pi(\alpha_1|\alpha_0) = \pi(\alpha_1)$. This distribution can be used to derive distributions of interest by conditioning or marginalization.

Multi-process filtering

In the following proposition the filtering recursions for multi-process model class II mixture is presented.

PROPOSITION 8.1
For a multi-process model class II mixture, where the probabilities with which the model is selected are first-order Markov, the following statements hold.

(i) The one-step-ahead joint predictive density for θ_t and α_t given $y_{1:t-1}$ is calculated using the filtered density $\pi(\theta_{t-1}, \alpha_{t-1}|y_{1:t-1})$, which gives

$$\begin{aligned}\pi(\theta_t, \alpha_t|y_{1:t-1}) &= \int\int \pi(\theta_t|\theta_{t-1}, \alpha_t)\pi(\alpha_t|\alpha_{t-1}) \\ &\quad \cdot\pi(\theta_{t-1}, \alpha_{t-1}|y_{1:t-1})d\theta_{t-1}d\alpha_{t-1}.\end{aligned}$$

(ii) The one-step-ahead predictive density for Y_t given $y_{1:t-1}$ is

$$\pi(y_t|y_{1:t-1}) = \int\int \pi(y_t|\theta_t, \alpha_t)\pi(\theta_t, \alpha_t|y_{1:t-1})d\theta_t d\alpha_t.$$

(iii) The filtering density for θ_t and α_t given $y_{1:t}$ can be calculated using (i) and (ii)

$$\pi(\theta_t, \alpha_t|y_{1:t}) = \frac{\pi(y_t|\theta_t, \alpha_t)\pi(\theta_t, \alpha_t|y_{1:t-1})}{\pi(y_t|y_{1:t-1})}.$$

Proof. The state θ_t is conditionally independent of $Y_{1:t-1}$ given θ_{t-1}, and α_t is conditionally independent of $Y_{1:t-1}$ given α_{t-1}. Then (i) is proved by

$$\begin{aligned}\pi(\theta_t, \alpha_t|y_{1:t-1}) &= \int\int \pi(\theta_t, \theta_{t-1}, \alpha_t, \alpha_{t-1}|y_{1:t-1})d\theta_{t-1}d\alpha_{t-1} \\ &= \int\int \pi(\theta_t, \alpha_t|\theta_{t-1}, \alpha_{t-1}, y_{1:t-1})\pi(\theta_{t-1}, \alpha_{t-1}|y_{1:t-1})d\theta_{t-1}d\alpha_{t-1} \\ &= \int\int \pi(\theta_t, \alpha_t|\theta_{t-1}, \alpha_{t-1})\pi(\theta_{t-1}, \alpha_{t-1}|y_{1:t-1})d\theta_{t-1}d\alpha_{t-1} \\ &= \int\int \pi(\theta_t|\theta_{t-1}, \alpha_t)\pi(\alpha_t|\alpha_{t-1})\pi(\theta_{t-1}, \alpha_{t-1}|y_{1:t-1})d\theta_{t-1}d\alpha_{t-1}.\end{aligned}$$

The observation Y_t is conditionally independent of $Y_{1:t-1}$ given θ_t and α_t, and (ii) is proved by

$$\begin{aligned}\pi(y_t|y_{1:t-1}) &= \int\int \pi(y_t, \theta_t, \alpha_t|y_{1:t-1})d\theta_t d\alpha_t \\ &= \int\int \pi(y_t|\theta_t, \alpha_t, y_{1:t-1})\pi(\theta_t, \alpha_t|y_{1:t-1})d\theta_t d\alpha_t \\ &= \int\int \pi(y_t|\theta_t, \alpha_t)\pi(\theta_t, \alpha_t|y_{1:t-1})d\theta_t d\alpha_t.\end{aligned}$$

Bayes' rule and the conditional independence of Y_t and $Y_{1:t-1}$ given θ_t and α_t is used to prove (iii). Then

$$\begin{aligned}\pi(\theta_t, \alpha_t|y_{1:t}) &= \frac{\pi(\theta_t, \alpha_t|y_{1:t-1})\pi(y_t|\theta_t, \alpha_t, y_{1:t-1})}{\pi(y_t|y_{1:t-1})} \\ &= \frac{\pi(\theta_t, \alpha_t|y_{1:t-1})\pi(y_t|\theta_t, \alpha_t)}{\pi(y_t|y_{1:t-1})}.\end{aligned}$$

□

[24, pp. 51-53],[25, pp. 443-445].

The multi-process Kalman filter

In the following it is assumed, that the model $M_t(\alpha_t)$ is a DLM, and the matrices G_t, F_t, V_t and W_t are given by specific values for each value of $\alpha_t \in \mathcal{A}$, i.e. $G_t(\alpha_t)$, $F_t(\alpha_t)$, $V_t(\alpha_t)$ and $W_t(\alpha_t)$. The exact predictive and filtering distributions are probability mixtures. If $\mathcal{A} = \{1, \ldots, k\}$, then at time $t = 1$ there are k components, giving k possible models, at $t = 2$ there are again k components, which gives k^2 possible model sequences, and at time t there are k^t possible model sequences. This can give computational difficulties because of the large number of mixture components as time proceeds. The complexity of the computations can be reduced by ignoring components where the posterior probabilities are small. The multi-process Kalman filter is presented in the next proposition using mixture collapse. Mixture collapse is a method, where probability mixtures is represented by an approximating distribution.

Proposition 8.2
Consider a multi-process model class II mixture, where the probabilities with which the model is selected are first-order Markov. Let the prior distribution $\theta_0 \sim N(m_0, C_0)$ *and* $\pi(\alpha_{t-1}|y_{1:t-1})$ *be given, and assume the approximation*

$$\theta_{t-1}|y_{1:t-1} \approx N(m_{t-1}, C_{t-1}).$$

Then the following statements hold.

(ia) The one-step-ahead predictive distribution of θ_t *given* $y_{1:t-1}$ *and* α_t *is approximately normal with parameters*

$$\begin{aligned}a_t(\alpha_t) &= E[\theta_t|y_{1:t-1}, \alpha_t] = G_t(\alpha_t)m_{t-1}, \\ R_t(\alpha_t) &= Var[\theta_t|y_{1:t-1}, \alpha_t] = G_t(\alpha_t)C_{t-1}G_t(\alpha_t)^T + W_t(\alpha_t).\end{aligned}$$

(ib) The one-step-ahead predictive distribution of θ_t given $y_{1:t-1}$ is approximately normal with parameters

$$\begin{aligned} a_t &= E[\theta_t|y_{1:t-1}] = \sum_{\alpha_t \in \mathcal{A}} \pi(\alpha_t|y_{1:t-1}) a_t(\alpha_t), \\ R_t &= Var[\theta_t|y_{1:t-1}] \\ &= \sum_{\alpha_t \in \mathcal{A}} \left(R_t(\alpha_t) + (a_t(\alpha_t) - a_t)\,(a_t(\alpha_t) - a_t)^T \right) \pi(\alpha_t|y_{1:t-1}), \end{aligned}$$

where

$$\pi(\alpha_t|y_{1:t-1}) = \sum_{\alpha_{t-1} \in \mathcal{A}} \pi(\alpha_t|\alpha_{t-1}) \pi(\alpha_{t-1}|y_{1:t-1}) \tag{8.6}$$

is the one-step-ahead approximation of the predictive distribution of α_t given $y_{1:t-1}$.

(iia) The one-step-ahead predictive distribution of Y_t given $y_{1:t-1}$ and α_t is approximately normal with parameters

$$\begin{aligned} f_t(\alpha_t) &= E[Y_t|y_{1:t-1}, \alpha_t] = F_t(\alpha_t) a_t, \\ Q_t(\alpha_t) &= Var[Y_t|y_{1:t-1}, \alpha_t] = F_t(\alpha_t) R_t F_t(\alpha_t)^T + V_t(\alpha_t). \end{aligned}$$

(iib) The one-step-ahead predicitve distribution of Y_t given $y_{1:t-1}$ is approximately normal with parameters

$$\begin{aligned} f_t &= E[Y_t|y_{1:t-1}] = \sum_{\alpha_t \in \mathcal{A}} \pi(\alpha_t|y_{1:t-1}) f_t(\alpha_t), \\ Q_t &= Var[Y_t|y_{1:t-1}] \\ &= \sum_{\alpha_t \in \mathcal{A}} \left(Q_t(\alpha_t) + (f_t(\alpha_t) - f_t)\,(f_t(\alpha_t) - f_t)^T \right) \pi(\alpha_t|y_{1:t-1}), \end{aligned}$$

where $\pi(\alpha_t|y_{1:t-1})$ are given as in equation (8.6).

(iiia) The filtering distribution of θ_t given $y_{1:t}$ and α_t is approximately normal with parameters

$$\begin{aligned} m_t(\alpha_t) &= E[\theta_t|y_{1:t}, \alpha_t] = a_t + R_t F_t(\alpha_t)^T Q_t^{-1} e_t \\ C_t(\alpha_t) &= Var[\theta_t|y_{1:t}, \alpha_t] = R_t - R_t F_t(\alpha_t)^T Q_t^{-1} F_t(\alpha_t) R_t, \end{aligned}$$

where $e_t = Y_t - f_t$ is the forecast error.

(iiib) The filtering distribution of θ_t given $y_{1:t}$ is approximately normal with parameters

$$\begin{aligned} m_t &= E[\theta_t|y_{1:t}] = \sum_{\alpha_t \in \mathcal{A}} \pi(\alpha_t|y_{1:t}) m_t(\alpha_t) \\ C_t &= Var[\theta_t|y_{1:t}] \\ &= \sum_{\alpha_t \in \mathcal{A}} \left(C_t(\alpha_t) + (m_t(\alpha_t) - m_t)\,(m_t(\alpha_t) - m_t)^T \right) \pi(\alpha_t|y_{1:t}), \end{aligned}$$

and the filtering distribution of α_t given $y_{1:t}$ is approximated by

$$\pi(\alpha_t|y_{1:t}) \approx \frac{\pi(\alpha_t|y_{1:t-1})\pi(y_t|y_{1:t-1},\alpha_t)}{\sum_{\alpha\in\mathcal{A}}\pi(\alpha|y_{1:t-1})\pi(y_t|y_{1:t-1},\alpha)},$$

where $\pi(\alpha|y_{1:t-1})$ is given by equation (8.6), and $\pi(y_t|y_{1:t-1},\alpha_t)$ is given by (iia).

Proof. The joint distribution of $(\theta_0,\theta_1,\ldots,\theta_t,Y_1,\ldots,Y_t,\alpha_1,\ldots,\alpha_t)$ is given by equation (8.5). The marginal and conditional distributions and the distribution of any subvector or conditional distributions of some components given others can be approximated with a Gaussian distribution using mixture collapses.
Since the model $M_t(\alpha_t)$ is given by a DLM for any value of $\alpha_t \in \mathcal{A}$, and the matrices $G_t(\alpha_t)$, $F_t(\alpha_t)$, $V_t(\alpha_t)$ and $W_t(\alpha_t)$ are known for all values of $\alpha_t \in \mathcal{A}$, then (*ia*), (*iia*) and (*iiia*) are proved using the Kalman filter, proposition 7.2. The means and variances in (*ib*), (*iib*) and (*iiib*) are given by the mean and variance for probability mixtures, which are equivalent to equation (8.3) and equation (8.4) for multi-process model class I. The probability $\pi(\alpha_t|y_{1:t})$ is given using Bayes' theorem

$$\begin{aligned}\pi(\alpha_t|y_{1:t}) &= \pi(\alpha_t|y_{1:t-1},y_t)\\ &= \frac{\pi(\alpha_t|y_{1:t-1})\pi(y_t|y_{1:t-1},\alpha_t)}{\sum_{\alpha\in\mathcal{A}}\pi(\alpha|y_{1:t-1})\pi(y_t|y_{1:t-1},\alpha)}.\end{aligned}$$

□

[24, pp. 53-55],[25, pp. 443-445,448,450].

Multi-process smoothing

The multi-process smoothing recursions are presented in the following proposition.

Proposition 8.3
For a multi-process model class II mixture, where the probabilities with which the model is selected are first-order Markov, the following statements hold.

(i) The state system $(\theta_0,\ldots,\theta_T)$ and $(\alpha_1,\ldots,\alpha_T)$ given $y_{1:T}$ have backward transition probabilities specified by

$$\pi(\theta_t,\alpha_t|\theta_{t+1},\alpha_{t+1},y_{1:T}) = \frac{\pi(\theta_{t+1}|\theta_t,\alpha_{t+1})\pi(\alpha_{t+1}|\alpha_t)\pi(\theta_t,\alpha_t|y_{1:t})}{\pi(\theta_{t+1},\alpha_{t+1}|y_{1:t})}.$$

(ii) Conditional on $y_{1:T}$ the smoothing distribution of θ_t and α_t can be calculated using the backward recursion starting from $\pi(\theta_T|y_{1:T})$

$$\begin{aligned}\pi(\theta_t,\alpha_t|y_{1:T}) &= \pi(\theta_t,\alpha_t|y_{1:t})\int\int \frac{\pi(\theta_{t+1}|\theta_t,\alpha_{t+1})\pi(\alpha_{t+1}|\alpha_t)}{\pi(\theta_{t+1},\alpha_{t+1}|y_{1:t})} \\ &\quad \cdot\pi(\theta_{t+1},\alpha_{t+1}|y_{1:T})d\theta_{t+1}d\alpha_{t+1}.\end{aligned}$$

Proof. The state θ_t and α_t are conditionally independent of $Y_{t+1:T}$ given θ_{t+1} and α_{t+1}, and given θ_t and α_t, θ_{t+1} and α_{t+1} are conditionally independent of $y_{1:T}$. Then (i) is proved using Bayes formula

$$\begin{aligned}\pi(\theta_t,\alpha_t|\theta_{t+1},\alpha_{t+1},y_{1:T}) &= \pi(\theta_t,\alpha_t|\theta_{t+1},\alpha_{t+1},y_{1:t}) \\ &= \frac{\pi(\theta_{t+1},\alpha_{t+1}|\theta_t,\alpha_t,y_{1:t})\pi(\theta_t,\alpha_t|y_{1:t})}{\pi(\theta_{t+1},\alpha_{t+1}|y_{1:t})} \\ &= \frac{\pi(\theta_{t+1},\alpha_{t+1}|\theta_t,\alpha_t)\pi(\theta_t,\alpha_t|y_{1:t})}{\pi(\theta_{t+1},\alpha_{t+1}|y_{1:t})} \\ &= \frac{\pi(\theta_{t+1}|\theta_t,\alpha_{t+1})\pi(\alpha_{t+1}|\alpha_t)\pi(\theta_t,\alpha_t|y_{1:t})}{\pi(\theta_{t+1},\alpha_{t+1}|y_{1:t})}.\end{aligned}$$

The density $\pi(\theta_t,\alpha_t,\theta_{t+1},\alpha_{t+1}|y_{1:T})$ with respect to θ_{t+1} and α_{t+1} is marginalized to prove (ii)

$$\begin{aligned}\pi(\theta_t,\alpha_t|y_{1:T}) &= \int\int \pi(\theta_t,\alpha_t,\theta_{t+1},\alpha_{t+1}|y_{1:T})d\theta_{t+1}d\alpha_{t+1} \\ &= \int\int \pi(\theta_{t+1},\alpha_{t+1}|y_{1:T})\pi(\theta_t,\alpha_t|\theta_{t+1},\alpha_{t+1},y_{1:T})d\theta_{t+1}d\alpha_{t+1} \\ &= \int\int \frac{\pi(\theta_{t+1}|\theta_t,\alpha_{t+1})\pi(\alpha_{t+1}|\alpha_t)\pi(\theta_t,\alpha_t|y_{1:t})}{\pi(\theta_{t+1},\alpha_{t+1}|y_{1:t})} \\ &\quad \cdot\pi(\theta_{t+1},\alpha_{t+1}|y_{1:T})d\theta_{t+1}d\alpha_{t+1} \\ &= \pi(\theta_t,\alpha_t|y_{1:t})\int\int \frac{\pi(\theta_{t+1}|\theta_t,\alpha_{t+1})\pi(\alpha_{t+1}|\alpha_t)}{\pi(\theta_{t+1},\alpha_{t+1}|y_{1:t})} \\ &\quad \cdot\pi(\theta_{t+1},\alpha_{t+1}|y_{1:T})d\theta_{t+1}d\alpha_{t+1}.\end{aligned}$$

□

[24, pp. 60-61],[25, pp. 443-445].

Multi-process Kalman smoothing

In the following proposition it is assumed that the smoothing distributions are approximately normal and that the model $M_t(\alpha_t)$ is a DLM. Again the matrices G_t, F_t, V_t and W_t are given by specific values for each value of $\alpha_t \in \mathcal{A}$, i.e. $G_t(\alpha_t)$, $F_t(\alpha_t)$, $V_t(\alpha_t)$ and $W_t(\alpha_t)$.

PROPOSITION 8.4
Consider a multi-process model class II mixture, where the probabilities with which the model is selected are first-order Markov. Assume the approximation

$$\theta_{t+1}|y_{1:T} \approx N(s_{t+1}, S_{t+1})$$

is given, then the following statements hold.

(i) The smoothed distribution of θ_t given $y_{1:T}$ and α_t is approximately normal with parameters

$$\begin{aligned} s_t(\alpha_t) &= E[\theta_t|y_{1:T}, \alpha_t] = m_t(\alpha_t) + C_t(\alpha_t)G_{t+1}^T R_{t+1}^{-1}(s_{t+1} - a_{t+1}) \\ S_t(\alpha_t) &= Var[\theta_t|y_{1:T}, \alpha_t] \\ &= C_t(\alpha_t) - C_t(\alpha_t)G_{t+1}^T R_{t+1}^{-1}(R_{t+1} - S_{t+1})R_{t+1}^{-1}G_{t+1}C_t(\alpha_t). \end{aligned}$$

(ii) The smoothed distribution of θ_t given $y_{1:t-1}$ is then approximately normal with parameters

$$\begin{aligned} s_t &= E[\theta_t|y_{1:T}] = \sum_{\alpha_t \in \mathcal{A}} \pi(\alpha_t|y_{1:T}) s_t(\alpha_t) \\ S_t &= Var[\theta_t|y_{1:T}] \\ &= \sum_{\alpha_t \in \mathcal{A}} \left(S_t(\alpha_t) + (s_t(\alpha_t) - s_t)(s_t(\alpha_t) - s_t)^T \right) \pi(\alpha_t|y_{1:T}), \end{aligned}$$

where it is assumed that $\pi(\alpha_t|y_{1:T})$ can be approximated by

$$\pi(\alpha_t|y_{1:T}) \approx \sum_{\alpha_{t+1} \in \mathcal{A}} \pi(\alpha_t|\alpha_{t+1})\pi(\alpha_{t+1}|y_{1:T}).$$

Proof. Because the model $M_t(\alpha_t)$ is given by a DLM for any value of $\alpha_t \in \mathcal{A}$, and the matrices $G_t(\alpha_t)$, $F_t(\alpha_t)$, $V_t(\alpha_t)$ and $W_t(\alpha_t)$ are known for all $\alpha_t \in \mathcal{A}$, then (i) is proved using the Kalman smoother, proposition 7.4. The mean and variance in (ii) given by the mean and variance for probability mixtures, which are equivalent to equation (8.3) and equation (8.4) for multi-process model class I. □

[24, pp. 61-62],[25, pp. 443-445].

Multi-process forecasting

The multi-process filtering recursions presented in proposition 8.1 gives the one-step-ahead predictive distribution. In the following proposition the k-step-ahead predictive distribution is presented.

PROPOSITION 8.5
For a multi-process model class II mixture, where the probabilities with which the model is selected are first order Markov, the following statements hold for any $k > 0$.

(i) The k-step-ahead forecast distribution of the state is

$$\begin{aligned}\pi(\theta_{t+k}, \alpha_{t+k}|y_{1:t}) &= \int\int \pi(\theta_{t+k}|\theta_{t+k-1}, \alpha_{t+k})\pi(\alpha_{t+k}|\alpha_{t+k-1}) \\ &\quad \cdot\pi(\theta_{t+k-1}, \alpha_{t+k-1}|y_{1:t})d\theta_{t+k-1}d\alpha_{t+k-1}.\end{aligned}$$

(ii) The k-step-ahead forecast distribution of the observation is

$$\pi(y_{t+k}|y_{1:t}) = \int\int \pi(y_{t+k}|\theta_{t+k}, \alpha_{t+k})\pi(\theta_{t+k}, \alpha_{t+k}|y_{1:t})d\theta_{t+k}d\alpha_{t+k}.$$

Proof. The state θ_{t+k} is conditionally independent of $y_{1:t}$ given θ_{t+k-1}, and α_{t+k} is conditionally independent of $y_{1:t}$ given α_{t+k-1}. Then (i) is proved by

$$\begin{aligned}\pi(\theta_{t+k}, \alpha_{t+k}|y_{1:t}) &= \int\int \pi(\theta_{t+k}, \theta_{t+k-1}, \alpha_{t+k}, \alpha_{t+k-1}|y_{1:t})d\theta_{t+k-1}d\alpha_{t+k-1} \\ &= \int\int [\pi(\theta_{t+k}, \alpha_{t+k}|\theta_{t+k-1}, \alpha_{t+k-1}, y_{1:t}) \\ &\quad \cdot\pi(\theta_{t+k-1}, \alpha_{t+k-1}|y_{1:t})]\, d\theta_{t+k-1}d\alpha_{t+k-1} \\ &= \int\int [\pi(\theta_{t+k}, \alpha_{t+k}|\theta_{t+k-1}, \alpha_{t+k-1}) \\ &\quad \cdot\pi(\theta_{t+k-1}, \alpha_{t+k-1}|y_{1:t})]\, d\theta_{t+k-1}d\alpha_{t+k-1} \\ &= \int\int [\pi(\theta_{t+k}|\theta_{t+k-1}, \alpha_{t+k})\pi(\alpha_{t+k}|\alpha_{t+k-1}) \\ &\quad \cdot\pi(\theta_{t+k-1}, \alpha_{t+k-1}|y_{1:t})]\, d\theta_{t+k-1}d\alpha_{t+k-1}.\end{aligned}$$

The observation Y_{t+k} is conditionally independent of $y_{1:t}$ given θ_{t+k} and α_{t+k}, and (ii) is proved by

$$\begin{aligned}\pi(y_{t+k}|y_{1:t}) &= \int\int \pi(y_{t+k}, \theta_{t+k}, \alpha_{t+k}|y_{1:t})d\theta_{t+k}d\alpha_{t+k} \\ &= \int\int \pi(y_{t+k}|\theta_{t+k}, \alpha_{t+k}, y_{1:t})\pi(\theta_{t+k}, \alpha_{t+k}|y_{1:t})d\theta_{t+k}d\alpha_{t+k} \\ &= \int\int \pi(y_{t+k}|\theta_{t+k}, \alpha_{t+k})\pi(\theta_{t+k}, \alpha_{t+k}|y_{1:t})d\theta_{t+k}d\alpha_{t+k}.\end{aligned}$$

□

[24, p. 70],[25, pp. 443-445].

Multi-process Kalman forecasting

If it is assumed that the forecasting distributions are approximately normal, the multi-process Kalman forecasting recursions can be given as in the following proposition.

PROPOSITION 8.6
Consider a multi-process model class II mixture, where the probabilities with which the model is selected are first-order Markov, then the following is true for $k \geq 0$.

(ia) The distribution of θ_{t+k} *given* $y_{1:t}$ *and* α_{t+k} *is approximately normal with parameters*

$$\begin{aligned} a_{t+k}(\alpha_{t+k}) &= E[\theta_{t+k}|y_{1:t}, \alpha_{t+k}] = G_{t+k}(\alpha_{t+k})a_{t+k-1} \\ R_{t+k}(\alpha_{t+k}) &= Var[\theta_{t+k}|y_{1:t}, \alpha_{t+k}] \\ &= G_{t+k}(\alpha_{t+k})R_{t+k-1}G_{t+k}(\alpha_{t+k})^T + W_{t+k}(\alpha_{t+k}). \end{aligned}$$

(ib) The state forecast θ_{t+k} *given* $y_{1:t}$ *is approximately normal distributed, i.e.* $\theta_{t+k}|y_{1:t} \approx N(a_{t+k}, R_{t+k})$, *with parameters*

$$\begin{aligned} a_{t+k} &= E[\theta_{t+k}|y_{1:t}] = \sum_{\alpha_{t+k} \in \mathcal{A}} \pi(\alpha_{t+k}|y_{1:t})a_{t+k}(\alpha_{t+k}) \\ R_{t+k} &= Var[\theta_{t+k}|y_{1:t}] \\ &= \sum_{\alpha_{t+k} \in \mathcal{A}} \pi(\alpha_{t+k}|y_{1:t}) \cdot \\ &\quad \left(R_{t+k}(\alpha_{t+k}) + (a_{t+k}(\alpha_{t+k}) - a_{t+k})(a_{t+k}(\alpha_{t+k}) - a_{t+k})^T\right), \end{aligned}$$

where the model forecast is approximated by

$$\pi(\alpha_{t+k}|y_{1:t}) \approx \sum_{\alpha_{t+k-1} \in \mathcal{A}} \pi(\alpha_{t+k}|\alpha_{t+k-1})\pi(\alpha_{t+k-1}|y_{1:t}). \tag{8.7}$$

(iia) The distribution of Y_{t+k} *given* $y_{1:t}$ *and* α_{t+k} *is approximately normal with parameters*

$$\begin{aligned} f_{t+k}(\alpha_{t+k}) &= E[Y_{t+k}|y_{1:t}, \alpha_{t+k}] = F_{t+k}(\alpha_{t+k})a_{t+k} \\ Q_{t+k}(\alpha_{t+k}) &= Var[Y_{t+k}|y_{1:t}, \alpha_{t+k}] \\ &= F_{t+k}(\alpha_{t+k})R_{t+k}F_{t+k}(\alpha_{t+k})^T + V_{t+k}(\alpha_{t+k}). \end{aligned}$$

(iib) The observation Y_{t+k} given $y_{1:t}$ is approximately normal distributed, i.e. $Y_{t+k}|y_{1:t} \approx N(f_{t+k}, Q_{t+k})$, with parameters

$$\begin{aligned} f_{t+k} &= E[Y_{t+k}|y_{1:t}] = \sum_{\alpha_{t+k}\in\mathcal{A}} \pi(\alpha_{t+k}|y_{1:t}) f_{t+k}(\alpha_{t+k}) \\ Q_{t+k} &= Var[Y_{t+k}|y_{1:t}] \\ &= \sum_{\alpha_{t+k}\in\mathcal{A}} \pi(\alpha_{t+k}|y_{1:t}) \cdot \\ &\quad \left(Q_{t+k}(\alpha_{t+k}) + (f_{t+k}(\alpha_{t+k}) - f_{t+k})(f_{t+k}(\alpha_{t+k}) - f_{t+k})^T\right), \end{aligned}$$

where $\pi(\alpha_{t+k}|y_{1:t})$ is given as in equation (8.7).

Proof. Since the model $M_t(\alpha_t)$ is given by a DLM for any value of $\alpha_t \in \mathcal{A}$, and the matrices $G_t(\alpha_t)$, $F_t(\alpha_t)$, $V_t(\alpha_t)$ and $W_t(\alpha_t)$ are known for all values of $\alpha_t \in \mathcal{A}$, then (ia) and (iia) is proved using proposition 7.6. The means and variances in (ib) and (iib) are given by the mean and variance for probability mixtures, which are equivalent to equation (8.3) and equation (8.4) for multi-process model class I. □

[24, p. 71],[25, pp. 443-445].

Bibliography

[1] Ron Brookmeyer, William D. Kalsbeek, and Donna F. Stroup. *Public Health Surveillance in Action: A Framework, Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*. Oxford University Press, 2003.

[2] Paddy Farrington and Nick Andrews. *Application to Infectious Disease Surveillance, Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*. Oxford University Press, 2003.

[3] Statens Serum Institut. Surveillance in denmark. http://www.ssi.dk/ English/PublicHealth/Surveillance%20in%20Denmark.aspx, 2011.

[4] Statens Serum Institut. Mission statement. http://www.ssi.dk/English/ Service/AboutSSI/Maal%20og%20opgaver.aspx, 2011.

[5] Statens Serum Institut. Sygdomsudbrud. http://www.ssi.dk/ Smitteberedskab/Sygdomsudbrud.aspx, 2012.

[6] C. P. Farrington, N. J. Andrews, A. D. Beale, and M. A. Catchpole. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society. Series A*, 159(3, pp. 547-563), 1996.

[7] Frederic H. Martini. *Fundamentals of Anatomy and Physiology, Seventh edition*. Pearson, 2006.

[8] Statens Serum Institut. Beregning af udbrudsstatus. http://www.ssi.dk/ graphics/UdbrudMonitor/Siderne/om-algoritmen.htm, 2011.

[9] Statens Serum Institut. Tarminfektions monitor. http://www.ssi.dk/ graphics/UdbrudMonitor/Siderne/Hovedindex.htm, 2012.

[10] Steffen Unkel, C. Paddy Farington, Paul H. Garthwaite, Chris Robertson, and Nick Andrews. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 175(1, pp. 49-82).

[11] Benjamin J. Cowling, Irene O. L. Wong, Lai-Ming Ho, Steven Riley, and Gabriel M. Leung. Methods for monitoring influenza surveillance data. *International Journal of Epidemiology*, 35(pp. 1314-1321).

[12] A. F. M. Smith and M. West. Monitoring renal transplants: An application of the multiprocess kalman filter. *Biometrics*, 39(4, pp. 867-878).

[13] Joe Whittaker and Sylvia Frühwirth-Schnatter. A dynamic changepoint model for detecting the onset of growth in bacteriological infections. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(4, pp. 625-640).

[14] Statens Serum Institut. Tema om mycoplasma pneumoniae. http://www.ssi.dk/Aktuelt/Temaer/Mycoplasma%20pneumoniae.aspx, 2011.

[15] Lægehåndbogen. Mycoplasmapneumoni. http://laegehaandbogen.dk/lunger/tilstande-og-sygdomme/mycoplasmapneumoni-2165.html, 2012.

[16] C. Akehurst. Mycoplasma pneumoniae epidemics. *Eurosurveillance*, 2(50), 1998.

[17] Statens Serum Institut Peter Henrik Andersen, Department of Epidemiology. Epi-news. *EPI-NEWS National Surveillance of Communicable Diseases*, 51, 2004.

[18] M. Höhle, T. Correa, M. Hofmann, C. Lang, M. Paul, A. Riebler, S. Steiner, M. Virtanen, and V. Wimmer. Modeling and monitoring discrete response time series. http://cran.r-project.org/web/packages/surveillance/surveillance.pdf, 2010.

[19] Giovanni Petris. Bayesian and likelihood analysis of dynamic linear models. http://cran.r-project.org/web/packages/dlm/dlm.pdf, 2010.

[20] Adelchi Azzalini. *Statistical Inference Based on the likelihood*. Chapman and Hall/CRC, 1996.

[21] FJ Anscombe. Contribution to discussion on paper by h. hotelling. *Journal of the Royal Statistical Society. Series B*, 15(2, pp. 229-230), 1953.

[22] Danmarks statistik. Nøgle tal om befolkningen. http://www.statistikbanken.dk/statbank5a/default.asp?w=1366, 2012.

[23] Annette J. Dobson and Adrian G. Barnett. *An Introduction to Generalized Linear Models, Third Edition*. Chapman and Hall/CRC, 2008.

[24] Giovanni Petris, Sonia Petrone, and Patrizia Campagnoli. *Dynamic Linear Models with R*. Springer, 2009.

[25] Mike West and Jeff Harrison. *Bayesian Forecasting and Dynamic Models, Second Edition*. Springer, 1997.

Appendix A

Codes

A.1 Data preprocessing

```
##Load of original data##
#csv file containing the observation number (time),
#the date (dato) and the number of infected (mycoplasma)
Data<-read.csv(file="mycoplasmaCounts.csv",header=TRUE)

##Plot of number of infected in days##
plot(Data$mycoplasma,type="h",main="",xaxt="n",xlab="Time in days",
ylab="No. infected")
axis(1,at=c(365*(0:length(Data$time))))

##Total number of infected##
sum(Data$mycoplasma)

##Number of samples recorded on each week day##
n<-floor(length(Data$time)/7)
a<-0;b<-0;c<-0;d<-0;e<-0;f<-0;g<-0
for (i in (7*(1:n)-6)){
a<-a+Data$mycoplasma[i]    #Friday
b<-b+Data$mycoplasma[i+1]  #Saturday
c<-c+Data$mycoplasma[i+2]  #Sunday
d<-d+Data$mycoplasma[i+3]  #Monday
e<-e+Data$mycoplasma[i+4]  #Tuesday
f<-f+Data$mycoplasma[i+5]  #Wednesday
g<-g+Data$mycoplasma[i+6]  #Thursday
}
```

```
Weeksum<-c(d,e,f,g,a,b,c)
plot(Weeksum,type="h",xlab="",ylab="No. infected",ylim=c(0,1300),
xaxt="n")
n<-length(Weeksum)
axis(1,at=c(1,2,3,4,5,6,7),labels=FALSE)
labels=c("Monday","Tuesday","Wednesday","Thursday","Friday",
"Saturday","Sunday")
text(1:n,par("usr")[3]-5,labels=labels,srt=45,adj=c(1.1,1.4),
xpd=TRUE,cex=.9)
labelmonday=c("1052","","","","","","")
text(1:n,par("usr")[1]-10,labels=labelmonday,adj=c(0.5,-24),
xpd=TRUE,cex=.9)
labeltuesday=c("","1187","","","","","")
text(1:n,par("usr")[1]-10, labels=labeltuesday,adj=c(0.5,-27),
xpd=TRUE,cex=.9)
labelwednesday=c("","","1061","","","","")
text(1:n,par("usr")[1]-10, labels = labelwednesday,
adj=c(0.5,-24.3),xpd=TRUE,cex=.9)
labelthursday=c("","","","901","","","")
text(1:n,par("usr")[1]-10,labels=labelthursday,adj=c(0.5,-20.6),
xpd=TRUE,cex=.9)
labelfriday=c("","","","","922","","")
text(1:n,par("usr")[1]-10,labels=labelfriday,adj=c(0.5,-21.1),
xpd=TRUE,cex=.9)
labelsaturday=c("","","","","","110","")
text(1:n,par("usr")[1]-10,labels=labelsaturday,adj=c(0.5,-3),
xpd=TRUE,cex=.9)
labelsunday=c("","","","","","","62")
text(1:n,par("usr")[1]-10,labels=labelsunday,adj=c(0.5,-1.9),
xpd=TRUE,cex=.9)

##No. infected from December 11th 1998 to December 29th 1998##
Section<-Data$mycoplasma[1625:1643]
plot(Section,type="h",xlab="",ylab="No. Infected",xaxt="n")
axis(1,at=1:length(Section),labels=FALSE)
labels=c("Fri Dec. 11th","Sat Dec. 12th","Sun Dec. 13th",
"Mon Dec. 14th","Tue Dec. 15th","Wed Dec. 16th","Thu Dec. 17th",
"Fri Dec. 18th","Sat Dec. 19th","Sun Dec. 20th","Mon Dec. 21th",
"Tue Dec. 22th","Wed Dec. 23th","Thu Dec. 24th","Fri Dec. 25th",
"Sat Dec. 26th","Sun Dec. 27th","Mon Dec. 28th","Tue Dec. 29th")
text(1:length(Section),par("usr")[3]-0.2,labels=labels,srt=45,
adj=c(1.1,1.1),xpd=TRUE,cex=.9)
```

```
##Change data from dayly counts to weekly counts##
n<-floor(length(Data$time)/7)
obsorg<-c()
for(i in (7*(1:n)-6)){
y<-Data$mycoplasma[i:(i+6)]
weekobs<-sum(y)
obsorg<-c(obsorg,weekobs)
}

##Plot of the number of infected in weeks##
plot(obsorg,type="h",main="",xaxt="n",xlab="Time in weeks",
ylab="No. infected")
axis(1,at=c(52*(0:length(obsorg))))

##Frequency of observations##
obsorg[230]<-NA  #383 infected in week 230
obsorg[231]<-NA  #246 infected in week 231
obsorg[232]<-NA  #246 infected in week 232
obsorg[233]<-NA  #235 infected in week 233
hist(obsorg,breaks=140,main = paste(""),xlab = "No. infected",
ylim=c(0,150),xlim=c(0,140),right=FALSE)
box(lty="solid")
```

A.2 Farringtons algorithm

```
##Load of surveillance package##
#which includes Farringtons algorithm
library(surveillance)

##Change of function from surveillance package ##
#changed to make reweight function work correctly
algo.farrington.fitGLM.fast <- function(response, wtime,
timeTrend = TRUE, reweight = TRUE){
if (timeTrend) {
design <- cbind(intercept = 1, wtime)
Formula <- response ~ wtime
}
else {
design <- matrix(1, nrow = length(wtime))
Formula <- response ~ 1
}
model <-glm.fit(design,response,family=quasipoisson(link="log"))
```

```
if (!model$converged) {
if (timeTrend) {
model <- glm.fit(design[, 1, drop = FALSE], response,
family = quasipoisson(link = "log"))
Formula <- response ~ 1
cat("Warning: No convergence with timeTrend -- trying without.\n")
}
}
class(model) <- c("glm", "lm")
phi <- 1  #max(summary.glm(model)$dispersion, 1) #Changed here
if (reweight) {
s <- anscombe.residuals(model, phi)
omega <-algo.farrington.assign.weights(s)
model <-glm.fit(design,response,family=quasipoisson(link="log"),
weights = omega)
phi <- max(summary.glm(model)$dispersion, 1)
}
model$phi <- phi
model$wtime <- wtime
model$response <- response
model$terms <- terms(Formula)
class(model) <- c("algo.farrington.glm", "glm")
return(model)
}

##Disease prognosis object created##
week<-1:578
state<-rep(0,length(week))
Mycoplasma<-create.disProg(week=week,observed=obsorg,
state=state,freq=52)

##Analysis using Farringtons algorithm##
control<-list(b=5,w=3,range=53:length(week),reweight=TRUE,
trend=FALSE,verbose=FALSE,alpha=0.01,limit54=c(5,4))
result<-algo.farrington(Mycoplasma,control=control)

##Plot of results##
plot(result,xaxis.years=FALSE,xaxt="n",xlab="Time in weeks",
legend.opts=list(x="top",legend=c("Infected", "Threshold", "Alarm"),
lty=NULL,pch=NULL,col=NULL))
axis(1,at=c(52*(0:length(week))))
```

```
##Result tables##
beg<-145
end<-178
tableoutbreak.one<-data.frame(week=beg:end,
Infected=obsorg[(beg+52):(end+52)],
Threshold=result$upperbound[beg:end],
Alarm=result$alarm[beg:end])
View(tableoutbreak.one)
beg<-478
end<-490
tableoutbreak.two<-data.frame(week=beg:end,
Infected=obsorg[(beg+52):(end+52)],
Threshold=result$upperbound[beg:end],
Alarm=result$alarm[beg:end])
View(tableoutbreak.two)
```

A.3 The dynamic linear model

```
##Skewness correction##
obs<-obsorg^(2/3)

##Load of dlm package##
library(dlm)

##Function for estimation of unknown parameters##
build<-function(parm){
dlmModPoly(1, dV = exp(parm[1]), dW = exp(parm[2]))+
dlmModTrig(s = 52, q = 1, dV = 0,dW = exp(rep(parm[3],2)))
}

##Steady period for estimation of unknown parameters##
est<-obs[0:200]
est<-append(est,rep(NA,49),after=length(est))
est<-append(est,obs[250:500],after=length(est))

##Estimation of unknown parameters using MLE##
fit<-dlmMLE(est,rep(0,3),build)

##Model with estimated parameters##
model<-dlmModPoly(order=1,dV=exp(fit$par[1]),dW=exp(fit$par[2]))+
dlmModTrig(s=52,q=1,dV=0,dW=exp(rep(fit$par[3],2)))
```

```
##Kalman filter##
filter<-dlmFilter(est,model)

##The variances##
variance<-dlmSvd2var(filter$U.C,filter$D.C)

##The mean value of the prior##
model$m0<-apply(filter$m,2,mean)

##The variance of the prior##
model$C0<-variance[[125]]

##Analysis function##
OneStep<-function(X,lev=0.99){
L<-length(X$res)
if(L==length(X$y)) stop("The End")
f<-dlmFilter(X$y[L+1],X$mod)
fc <- dlmForecast(X$mod)
up<-(fc$f[1,1]+qnorm(lev)*sqrt(fc$Q[[1]][1,1]))^(3/2)
X$upper<-c(X$upper,up)
r<-(X$y[L+1]-fc$f[1,1])/sqrt(fc$Q[[1]][1,1])
if(r>qnorm(lev)){
X$res<-c(X$res,NA)
f<-dlmFilter(NA,X$mod)
}
else{
X$res<-c(X$res,r)
}
f<-dlmSmooth(f)
X$mod$m0<-f$s[2,]
X$mod$C0<-dlmSvd2var(f$U.S,f$D.S)[[2]]
X
}

##List containing the model and observations##
surv<-list(mod=model,y=obs,res=NULL)

##Analysis##
for(i in 1:length(obs)) surv<-OneStep(surv)

##Weeks with alarms##
ii<-(1:578)[is.na(surv$res)]
```

```
##Plot of results##
plot(obsorg[53:578],type="h",main="",xaxt="n",xlab="Time in weeks",
ylab="No. infected")
axis(1,at=c(52*(0:length(obsorg))))
legend("top",c("Infected","Threshold","Alarm"),
lty=c(1,2,NA_integer_),pch=c(NA_integer_,NA_integer_,2),
col=c("black","blue","red"))
for(i in ii) points(c(i-52,i-52),c(-10,-10),col="red",pch=2)
lines(surv$upper[53:578],lty=2,col="blue")

##QQ-plot of the standard innovations##
qqnorm(surv$res,xlab="Normal Theoretical Quantiles")
abline(0,1)

##Tables with results##
Alarm<-is.na(surv$res)
beg<-145
end<-178
tableoutbreak.one<-data.frame(week=beg:end,
Infected=obsorg[(beg+52):(end+52)],
Threshold=surv$upper[(beg+52):(end+52)],
Alarm=Alarm[(beg+52):(end+52)])
View(tableoutbreak.one)
beg<-478
end<-490
tableoutbreak.two<-data.frame(week=beg:end,
Infected=obsorg[(beg+52):(end+52)],
Threshold=surv$upper[(beg+52):(end+52)],
Alarm=Alarm[(beg+52):(end+52)])
View(tableoutbreak.two)
```

A.4 The multi-process dynamic linear model

```
##Skewness correction##
obs<-obsorg^(2/3)

##Load of dlm package##
library(dlm)

##Function for estimation of unknown parameters##
build<-function(parm){
dlmModPoly(1, dV = exp(parm[1]), dW = exp(parm[2]))+
```

```
dlmModTrig(s = 52, q = 1, dV = 0,dW = exp(rep(parm[3],2)))
}

##Steady period for estimation of unknown parameters##
est<-obs[0:200]
est<-append(est,rep(NA,49),after=length(est))
est<-append(est,obs[250:500],after=length(est))

##Estimation of unknown parameters using MLE##
fit<-dlmMLE(est,rep(0,3),build)

##The estimated variances##
vars<-exp(fit$par)

##Filter function and calculation of 2*log-likelihood##
filter.likelihood <- function(y,Ft,mod,hist,trans){
at <- hist$mt
Rt <- hist$Ct+mod$w
ft <- sum(Ft[1,]*at)
Qt <- (Ft%*%Rt%*%t(Ft))[1,1]+mod$v
hist$mt <- at+Rt%*%t(Ft)*(y-ft)/Qt
hist$Ct <- Rt-Rt%*%t(Ft)%*%Ft%*%Rt/Qt
hist$Qt <- Qt
hist$ft <- ft
L <- length(hist$al)
al0 <- hist$al[L]
hist$lp <- hist$lp-log(Qt)-(y-ft)^2/Qt+trans[mod$al,al0]
hist$alpha <- c(hist$al,mod$al)
hist
}

##Calculate the filtering distribution of the models##
probAlpha<- function(H,nM){
lps <- unlist(lapply(H,function(x) x$lp))
lps <- lps-lps[1]
ps <- exp(lps)
ps <- ps/sum(ps)
L <- length(H[[1]]$al)
m0 <- unlist(lapply(H,function(x) x$al[L]))
res <- c()
for(i in 1:nM) {
ii <- m0==i
if(sum(ii)==0) res <- c(res,0) else res <- c(res,sum(ps[ii]))
```

```
}
res
}

##The design matrix is defined##
Design <- model.matrix(~cos(2*pi*time*7/365)+sin(2*pi*time*7/365),
data=data.frame(time=1:length(obs)))
dimTheta <- dim(Design)[2]

##The variances##
vars <- c(vars,vars[3])

##The steady state model##
steady <- list(v=vars[1],w=diag(vars[-1]),al=1)

##The outlier model##
outlier <- steady
outlier$v=steady$v*10
outlier$al <- 2

##The outbreak model##
outbreak <- outlier
outbreak$al <- 3

##(2*log) The transition probabilities##
transition <- 2*log(matrix(c(.985,.01,.005,.985,.01,.005,
.09,.01,.9),3))

##The models##
models <- list(steady=steady,outlier=outlier,
outbreak=outbreak)

##The number of models##
numMod <- length(models)

##The history##
history <- list()
for(i in 1:numMod) history<-c(history,
list(list(lp=transition[i,1],
mt=rep(0,3),Ct=diag(rep(1,3)),alpha=i)))

##The lag##
lag<-3
```

```
##The number of models saved##
savenum <- numMod^(lag+1)

##Probability of each model##
alpha.probability <- c()

##Analysis##
for(i in 1:lag){
H <- list()
Ft <- matrix(Design[i,],1)
y <- obs[i]
for(mod in models){
for(hist in history)
H <- c(H,list(filter.likelihood(y,Ft,mod,hist,transition)))
}
history <- H
alpha.probability <- cbind(alpha.probability,
round(probAlpha(history,numMod),3))
}
for(i in (lag+1):length(obs)){
H <- list()
Ft <- matrix(Design[i,],1)
y <- obs[i]
for(mod in models){
for(hist in history)
H <- c(H,list(filter.likelihood(y,Ft,mod,hist,transition)))
}
lps <- unlist(lapply(H,function(x) x$lp))
history <- H[order(lps,decreasing=T)][1:savenum]
alpha.probability <- cbind(alpha.probability,
round(probAlpha(history,numMod),3))
}

##The most likely sequence of models##
pickMax<-function(x){
mx<-max(x)
ii<-(1:length(x))[x==mx]
min(ii)
}
mostLikely<-apply(alpha.probability,2,pickMax)
```

```
##Alarms and outliers##
alarms<-(1:578)[mostLikely==3]
outlier<-(1:578)[mostLikely==2]

##Plot of results##
plot(obsorg[53:578],type="h",main="",xaxt="n",
xlab="Time in weeks",ylab="No. infected")
axis(1,at=c(52*(0:length(obsorg))))
legend("top",c("Infected", "Alarm", "Outlier"),
lty=c(1,NA_integer_,NA_integer_),pch=c(NA_integer_,2,2),
col=c("black","red","yellow"))
for(i in alarms) points(c(i-52,i-52),c(-10,-10),col="red",pch=2)
points(c(outlier-52,outlier-52),c(-10,-10),col="yellow",pch=2)
```