12-2014

Online Detection of Outliers and Structural Breaks using Sequential Monte Carlo Methods

Richard Wanjohi
University of Arkansas, Fayetteville

Online Detection of Outliers and Structural Breaks using Sequential Monte Carlo Methods

Online Detection of Outliers and Structural Breaks
using Sequential Monte Carlo Methods

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy in Mathematics

by

Richard Wanjohi
Kenyatta University
Bachelor of Education (Science), 2007
University of Arkansas
Master of Science in Statistics, 2009

December 2014
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

Dr. Giovanni Petris
Dissertation Director

Dr. Edward Gbur
Committee Member

Dr. Mark Arnold
Committee Member

Dr. Avishek Chakraborty
Committee Member

Abstract

Outliers and structural breaks occur quite frequently in time series data. Whereas outliers often contain valuable information about the process under study, they are known to have serious negative impact on statistical data analysis. Most obvious effect is model misspecification and biased parameter estimation which results in wrong conclusions and inaccurate predictions. Structural time series consist of underlying features such as level, slope, cycles or seasonal components. Structural breaks are permanent disruptions of one or more of these components and might be a signal of serious changes in the observed process. Detecting outliers and estimating the location of structural breaks has progressively become monumental both as a theoretical research problem and an essential part of applied data analysis. Among numerous applications include finance, industrial manufacturing, medical informatics, severe weather prediction. Given that these data arrive rather frequently and sequentially in time, fast reliable and accurate detection techniques are required. We propose a model from class of state-space models of the form $y_t = f(X_t, \psi, v_t)$ and $X_t = g(X_{t-1}, \psi, w_t)$ where $\left\{X_t\right\}_{t\geq 0}$ is a hidden Markov state process. The inference of $\left\{X_t\right\}_{t\geq 0}$ depends on the observation process $\{y_t\}_{t\geq 1}$ and the parameter vector ψ, whose elements are usually unknown. The innovations v_t and w_t are conditionally *Gaussian* given the precision parameter λ and auxiliary state ω. We employ sequential Monte Carlo techniques to approximate the joint target distribution $p(X_{0:t}, \psi|y_{1:t})$. The posterior estimates for the auxiliary states ω will be used to identify outliers and structural breaks. The results prove that the algorithm is comparable to traditional and computationally expensive MCMC and superior to regular techniques such as Exponentially Weighted Moving Average (EWMA), Shewhart, and cumulative sum (CUSUM) control charts

Acknowledgements

I would like to thank my advisor Dr. Giovanni Petris for his patience, support, and encouragement throughout my time at University of Arkansas. His guidance has made this a thoughtful and rewarding journey.

I would like to thank my dissertation committee for all their support and much needed insights. I also like to thank the department of Mathematical Sciences for granting me graduate assistantship, throughout my study, without which it would have been very difficult to realise my goal.

I would like to thank most sincerely my wife Lydia, sons Henry and Alan for their unconditional love, patience and support throughout the course.

Finally, I would like to thank all my family members and friends for their prayers and for expecting nothing less than completion from me.

Contents

List of Figures

List of Abbreviations and Symbols

Symbols

$y_{1:t}$	indicates the observed data process $(y_1, y_2, \cdots y_t)$
$\theta_{0:t}$	indicates the Hidden Markov state process $(\theta_0, \theta_1, \cdots, \theta_t)$
ω_t	indicates the auxiliary state variable
ν_t	indicates the degrees of freedom auxilliary variable defined on a set of finite integers by associated probability vector π
$X_{0:t}$	indicates the state vector at time t, consisting of $\theta_{0:t}, \nu_{0:t}$ and $\omega_{0:t}$
ψ	parameter vector whose components include the state and observation precision parameter λ, its associated mean and variances a and b respectively, and probability vector π
$\mathcal{N}(\mu, \sigma^2)$	indicates a *Gaussian* distribution with mean μ and variance σ^2
$\mathcal{N}(x\|\mu, \sigma^2)$	indicates the density function of *Gaussian* distribution with mean μ and variance σ^2 evaluated at x.
$\mathcal{G}am(\alpha, \beta)$	indicates a Gamma distribution with shape parameter α and rate parameter β
$p(\gamma\|\cdots)$	indicate the probability density function for γ given all other unknowns and the data
$p(X_{0:t}, \psi\|y_{1:t})$	indicate the joint posterior distribution for state vector $X_{0:t}$ and the parameter vector ψ

Abbreviations

ARMA	Autoregressive Moving Average
CUSUM	Cumulative Sum; see Chapter 2
DLM	Dynamic linear model or *Gaussian* state space model; see Chapter 3

ESS Effective sample size; see Chapter 4

EWMA Exponentially Weighted Moving Average; see Chapter 2

FFBS Forward Filtering Backward Sampling

GARCH Generalized Autoregressive Conditional Heteroscedasticity

MCMC Markov Chain Monte Carlo

SISR Sequential importance sampling with resampling

SMC Sequential Monte Carlo

SPC Statistical Process Control; see Chapter 2

Chapter 1

Introduction

Time series analysis involves data - often continuous measurements - collected or observed sequentially in time. We denote the univariate data by $y_t \in \mathbb{R}$ where $t \in \mathcal{T}$ is the time indexing when the data was observed. The time $t \in \mathcal{T}$ can be discrete in which case $\mathcal{T} = \mathbb{Z}$ or continuous time where now $\mathcal{T} = \mathbb{R}$. For a multivariate data we have $\mathbf{y_t} \in \mathbb{R}^m$ where $\mathbf{y_t} = (y_{t,1}, y_{t,2}, \cdots y_{t,m})$. For simplicity of the analysis we will consider only discrete time series. Examples of time series data include stock market returns, oil production see Figure 1.1, data obtained from `http://www.bp.com/en/global/corporate/about-bp/statistical-review-of-world-energy-2013.html`

The main aim of time series analysis is to create a mathematical model which will capture the underlying features of the observed data and help increase the understanding of generating probabilistic mechanisms and the dynamic of the observed series. Once the model fits the data, the analyst most times may be interested also in parameter estimation and forecasting. During the analysis, stochastic homogeneity of the data is assumed. Disruptions of stochastic homogeneity of the data might be a signal of serious changes in the process observed. Such changes therefore, need to be detected as soon as data is obtained. Time series data are often faced with such abrupt disruptions, some of which are temporary while others are permanent.

Structural time series model constitute of underlying states which include level, slope, seasonal, cyclic and irregular random components. The trend component represent long

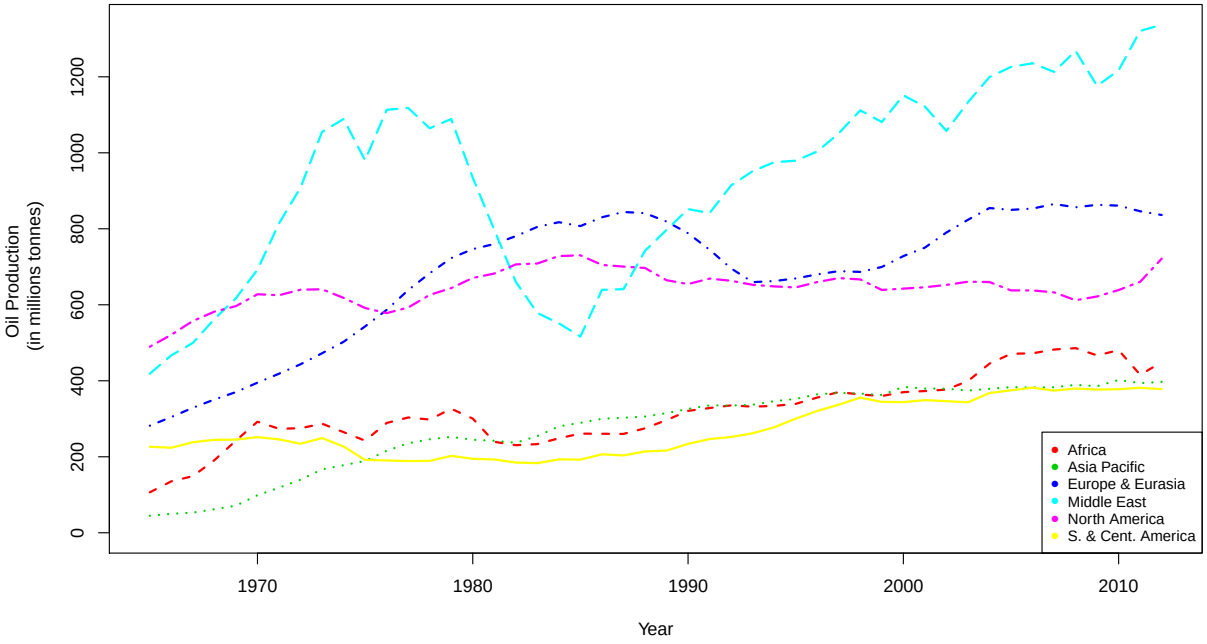

Figure 1.1: Global oil production from 1965 to 2012

term trend and usually constitute slope and intercept components. The seasonal component is the seasonal variation, cyclic component is repeated but non-periodic fluctuations and the residual make up the irregular random components. Time series of UK gas consumptions from 1960 to 1980 is displayed in Figure 1.2.

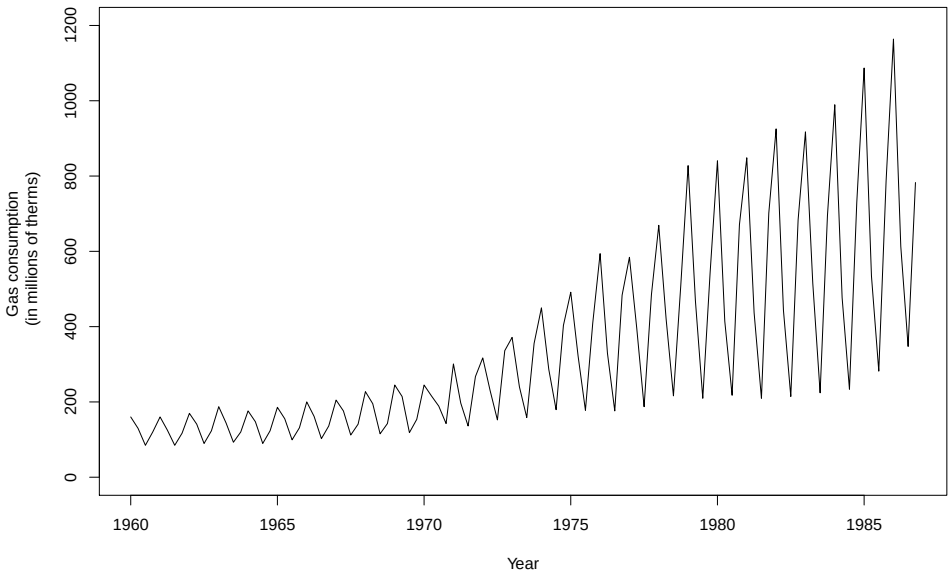

Figure 1.2: UK Quarterly gas consumption, from 1960 to 1980 in Millions of therms

To demonstrate time series decomposition,the UK gas series is used broken down into various components. Results are shown in Figure 1.3. The first chart is the observed data process, a quarterly time series of length 108. The second chart is the trend of the data and the third is the seasonal components. The last chart is the remaining components after the trend and seasonal factors have been removed, usually referred to as irregular components.

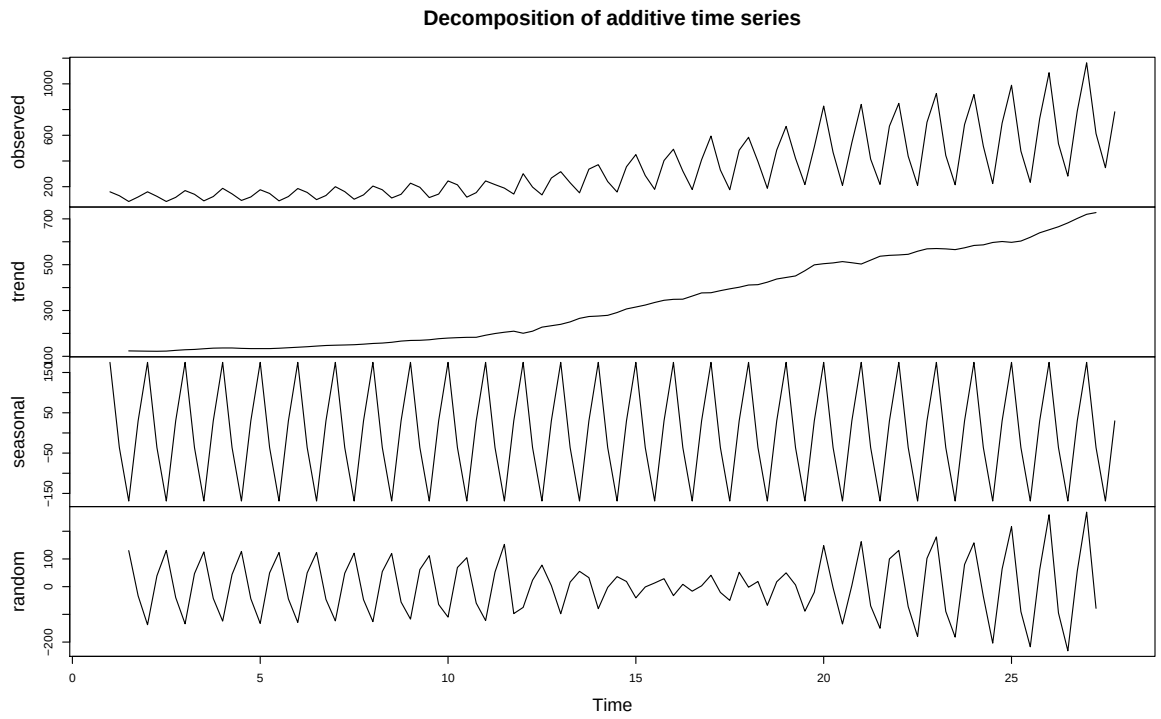

Figure 1.3: Time Series Decomposition

A basic structural time series model is of the form

$$y_t = \mathcal{T}_t + C_t + S_t + \varepsilon_t, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

where $\mathcal{T}$ is the Trend, C is cyclic, and S is the seasonal component. Any structural times series model can easily be represented in *state space form* (West & Harrison, 1989) and the model could be formulated as a regression with time varying coefficients (Petris, Petrone, & Campagnoli, 2009). Details of the state space representation are presented in Chapter 3. Structural breaks (Harvey & Koopman, 2005) are permanent shifts which occur whenever there is a change or disruptions in one or more of these components(Perron, 2006). Figure 1.4 shows permanent upward and downward shifts in a time series.

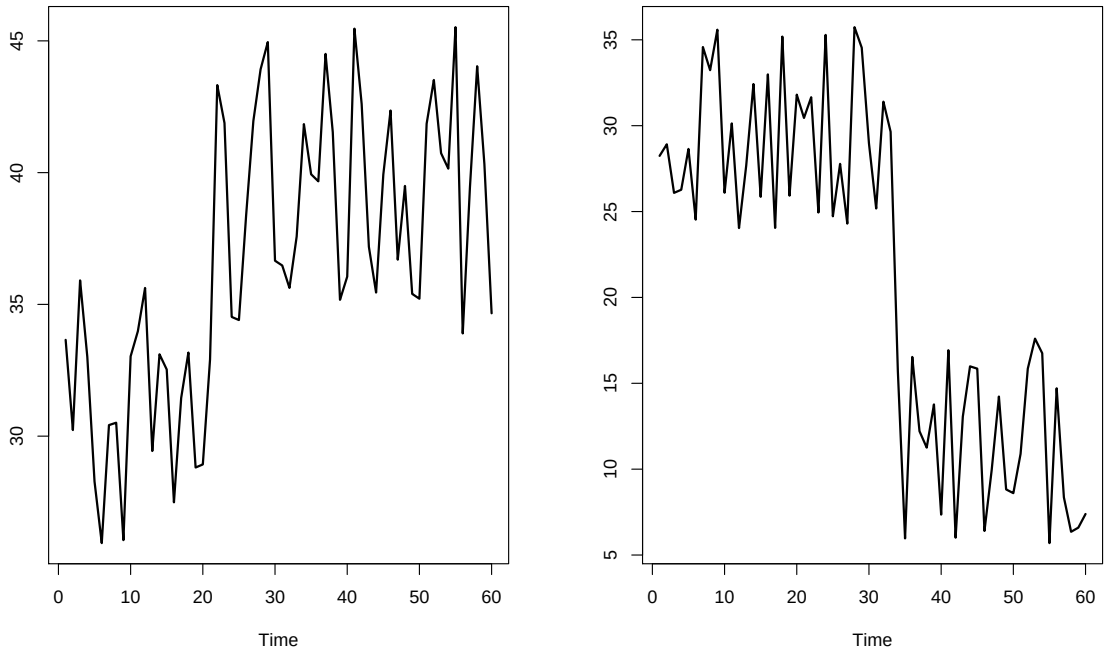

Figure 1.4: Permanent upward shift (left) and downward shift (right) in a Time Series

These permanent disruptions might be a signal of serious changes in the observed process and they are of considerable importance in the analysis of time series variables. The structural breaks occurs for any number of reasons including economic crises, changes in institutional arrangements, war, policy changes and regime shifts. For example, the combined effects of the Iranian revolution and the Iraq-Iran War in 1979 and 1980 caused a major drop in oil production in the Middle East, see Figure 1.1.

Figure 1.5 is plot of series of annual volume of discharge of the Nile River at Aswan from 1871 - 1970, given by (Cobb, 1978), which reveal a permanent drop of annual volume of discharge of the Nile River at Aswan from the year 1899. This sudden drop was largely due to the effect of Aswan dam, that was completed at that time

From the plot of UK quarterly gas consumption in Figure 1.2 there is a noticeable disruptions in the quarterly or seasonal components in the third quarter of the year 1970.

Detecting and estimating the location of structural breaks in time series has become

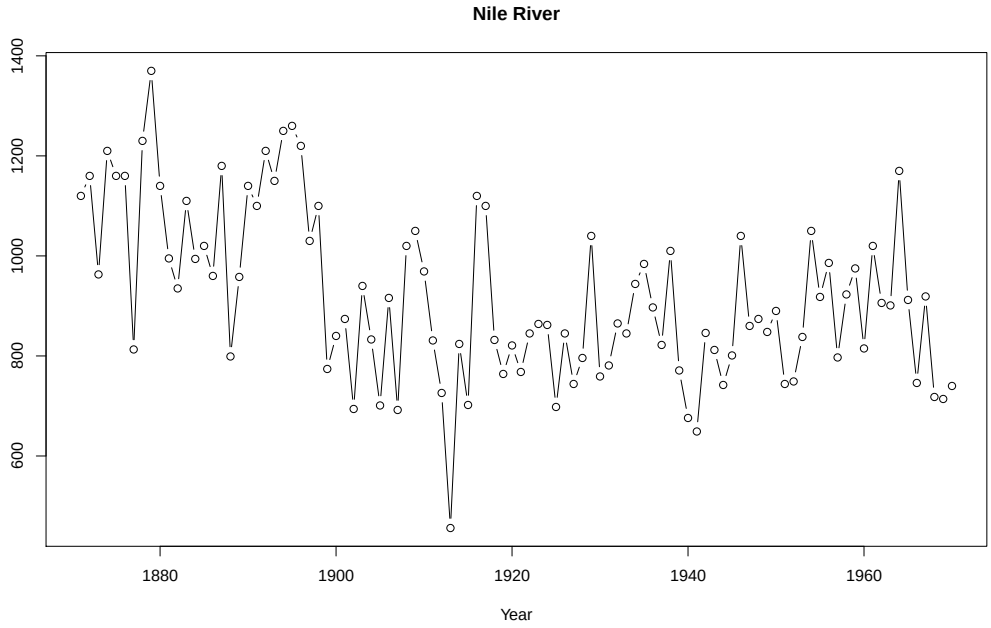

Figure 1.5: Annual volume of the Nile River from 1871 to 1970

increasingly important as both a theoretical research problem and a necessary component of applied data analysis.

An outlier or anomalous data point is defined as an observation in a dataset which appears to be inconsistent with the remainder of that set of data (Barnett & Lewis, 1994). Extreme values in a series may or may not be outliers and may arise as a result of either gross errors –due to faulty measurements, recording or typing errors –or they are true outliers.

Outliers occur frequently in measurement data and may have severe effects on model fitting and parameter estimation, leading to a mistaken conclusion and inaccurate predictions. It is therefore very important to identify them before modeling and analysis. Having said that, it is also worthy noting that, outliers often contain valuable information about the process under study or the data generating mechanisms. The outliers therefore require careful investigation and before considering possibly removing them, the researcher ought to understand why they occurred and the likelihood of their recurrence.

Outlier and structural break detection is a very important undertaking in many fields

especially in safety critical environments (Ardelean, 2012). An outlier may indicate anomaly through which a significant flawed outcome may arise. In analysis of vital variables of patients in intensive care for example, a small fault may lead to life threatening consequences. In manufacturing industries it is important to detect flaw in production line. While monitoring the usage of credit card, a sudden change in usage pattern may indicate credit card fraud. Similarly, a sudden change in monitoring process of mobile phone usage may indicate stolen mobile phone airtime. Other fields where outlier and structural breaks detection methods have been suggested include finance (Andreou & Ghysels, 2009) clinical trials (Penny & Jolliffe, 2001) and medical informatics (Laurikkala et al., 2000), voting irregularity analysis, severe weather prediction (Zhao, Lu, & Kou, 2003), geographic information systems (Shekhar, Lu, & Zhang, 2003), economics (Koop & Potter, 2000),(Perron, 2006).

In order to obtain a lucid statistical data analysis it is important, as an initial step, to detect outlying observations and structural breaks, if any, in the data.

Most existing outlier and structural detection techniques, however, deal with static data or are done off-line. There are different strategies to detect outlying observations including clustering algorithms, regression based statistics, likelihood ratio test and cumulative sum of observed residuals.

There is voluminous amount of work on structural breaks and outlier detection over the last 50 years in the statistics and other related fields literature, although the on-line approach - where the goal is to detect an outlier or whether a structural change has occurred, in real time- is minimal.

In today's world where, in most cases, data arrive rather frequently and sequentially in time a fast, reliable and accurate detection methods are required for this online analysis. Methods that can be able to handle any level of noise and provide robustness against outliers. To that end, this dissertation's focus is on sequential data and online detection of outliers and structural break, and my main objectives was primarily to

- Design an algorithm to make inference sequentially for the model for outliers and structural breaks. The model used is decribed in details in Chapter 5
- Implement the algorithm in a statistical software and
- To obtain good results in practice on simulated and real data.

The dissertation is organized as follows: Chapter 2, we discuss various method that are available for detecting outliers and structural breaks in array of data and online. Chapter 3 emphasizes the State space models. We review the existing literature including the highly celebrated Kalman filter and Forward Filtering Backward Sampling (FFBS) algorithms . In Chapter 4 we discuss Sequential Monte Carlo methods, Chapter 5 we introduce the model for structural breaks and outlier for online data, Chapter 6 we evaluate our algorithm with simulated data and real data and provide some results. These results are compared with output from MCMC algorithm and finally the dissertation ends with some discussion in Chapter 7.

Chapter 2

Outliers and Structural breaks: Review

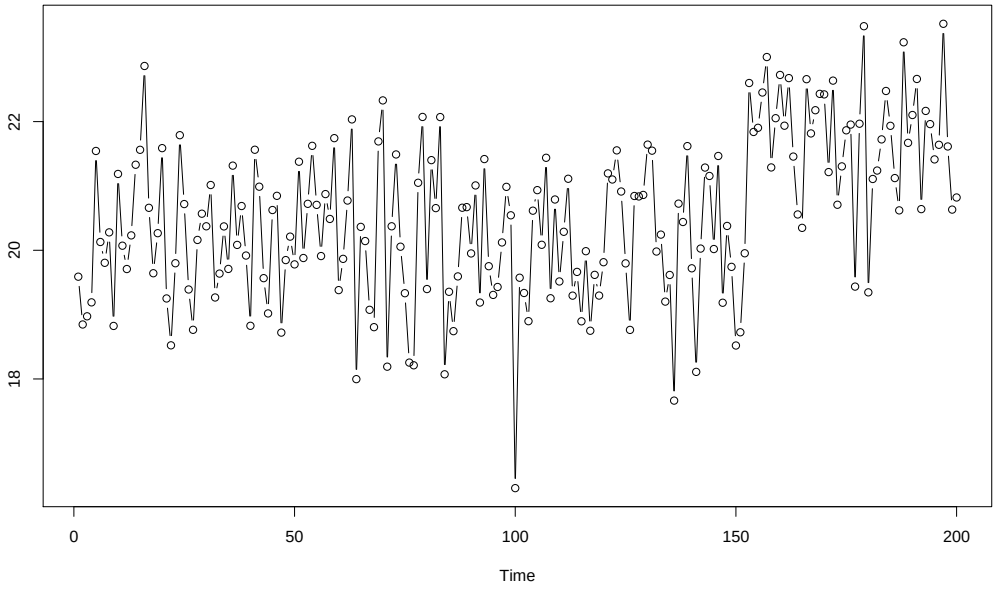

Figure 2.1: Time series data showing an outlier at time t= 100 and possible structural break at time $t = 153$

Most studies on outliers classify them as either (i) additive outlier - where only one observation is affected and after which the series return to its normal path, or (ii)innovative outlier which influences subsequent observations from its initial position. There are various strategies and approaches which has been developed to deal with outliers and structural breaks in statistical data. Most of these methods are static batch-type techniques which employs the full data set in detecting existence of outliers or breaks.

Sequential detection methods have also been proposed in some analysis, albeit minimally. Detection of outliers is simple when there is no serial dependence on the observations. In this case the procedures that involves detecting instabilities in mean and variances are of vital importance. The extreme observations -the very large or very small- are often treated to be inconsistent with the assumed generating mechanism or distribution and hence require to be tested for outlyingness. A number of outlier detection techniques over static data have been proposed (Barnett & Lewis, 1984), (Hodge & Austin, 2004).

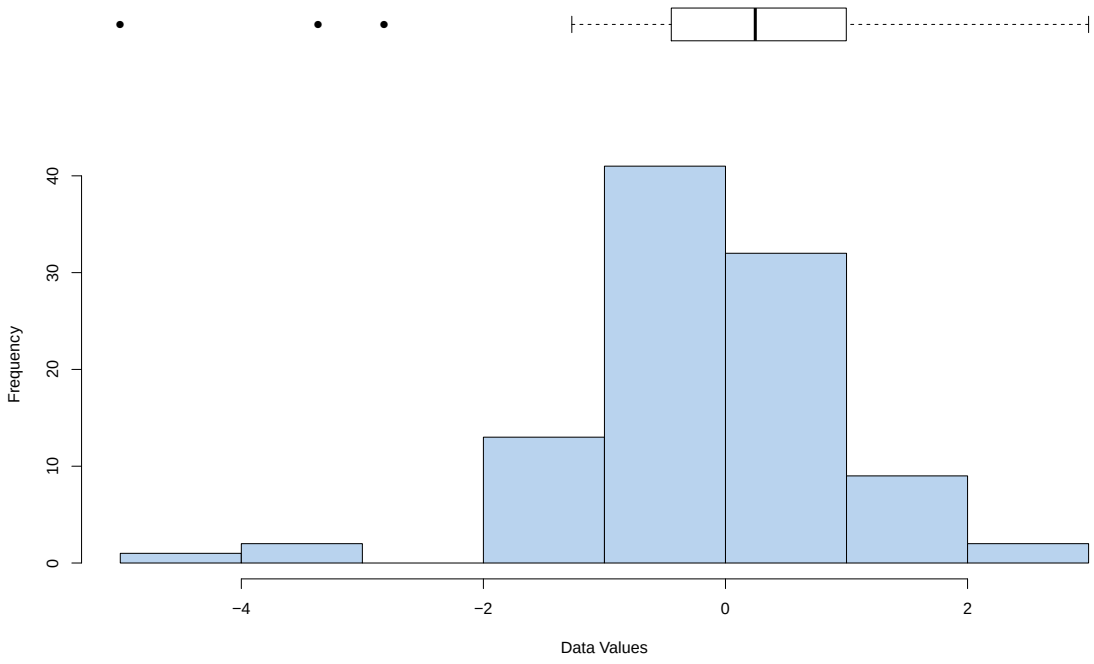

Figure 2.2: Univariate data: The box plot identifies the outlying observations

The simplest and traditional statistical outlier detection technique is the use of the box plots. They provide graphical representation that allows the researcher to identify outlying observations in both univariate and multivariate data sets. The box plots make no assumption of the distribution of the data and they plot, among others, the lower and upper extreme values. The outliers are identified as observations beyond these two extreme values. A univariate data set is displayed in Fig 2.2 by a skewed histogram and an overlaid box plot from which the three outlying observations are clearly identifiable. Figure 2.3

shows a basic scatterplot from bivariate data obtained from selected cities in US. A negative correlation between mortality and education level is well captured and one of the superimposed box plots clearly uncovers the presence of an outlier.

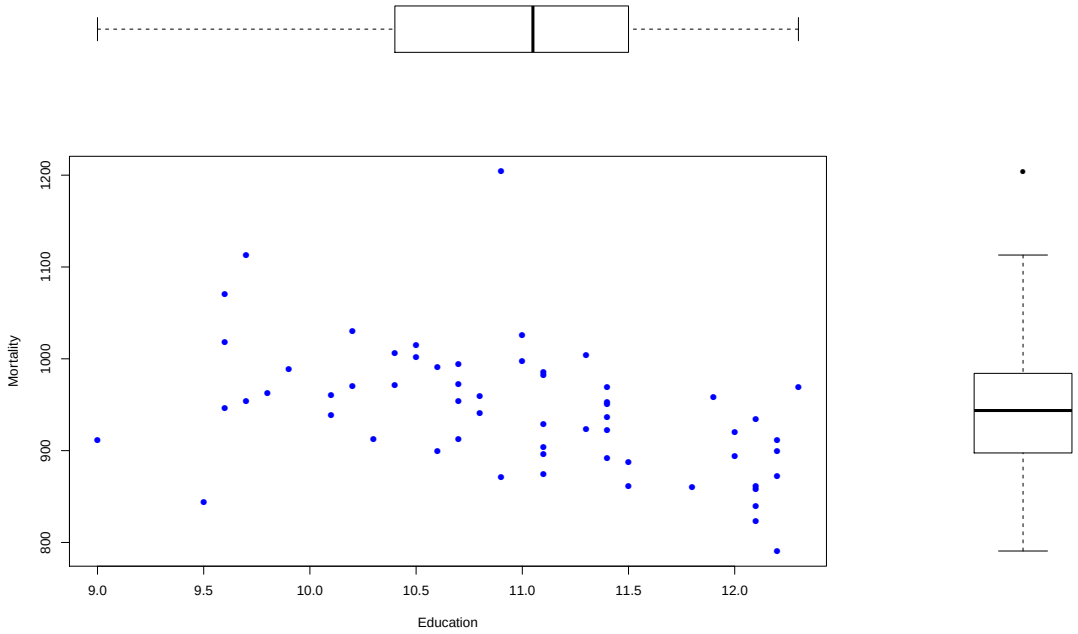

Figure 2.3: Outlier in a bivariate data

The distance-based approaches (Knorr, Ng, & Tucakov, 2000) computes and utilizes distance between two data points or examines the spatial proximity of each data point in the data space and if its proximity deviates considerably from the proximity of the other data then a data point is considered an outlier. These techniques do not make prior assumptions of the data distribution model. They are simple to implement but, since they are based on calculation of distances between all observations, they suffer from *Curse of Dimensionality*; that is, computational complexity increases as the dimension of data m and number of observations T increases. The clustering algorithms such as k-Nearest Neighbors heirechical and k-means algorithms features prominently in this approach. Figure 2.4 shows some outliers from a popular dataset *Iris* using k-means clustering approach.

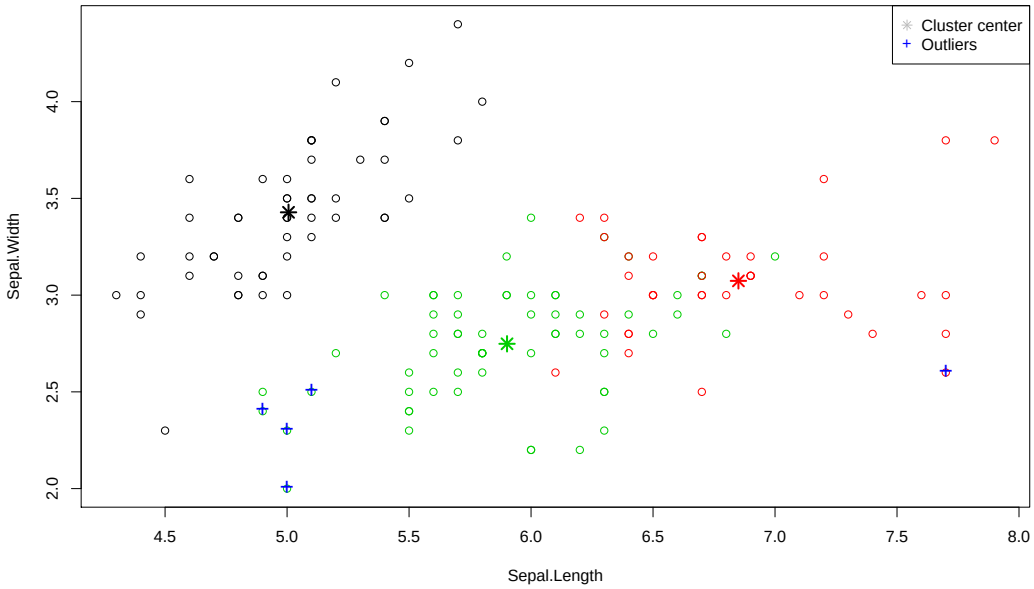

Figure 2.4: Outliers with k-Means Clustering

Statistical-based approach (Rousseeuw & Leroy, 2005), assume data follows a certain statistical model. In this case, the probabilistic tests, based on the model, are carried out and outliers are identified as say, points that have a low probability to be generated by the overall distribution.

In many applications it is common for time series data to be serially dependent. There is high interest in current time series research to incorporate structural dependence of the observations in the analysis. This is the fundamental concept of this dissertation.

A significant literature exist which tests for structural breaks or non-linearity in time series. As hypothesis testing problem for detecting structural breaks, the null is set up to describe series with structural stability while the alternative contains one or more structural breaks.

2.1 Statistical Process Control (SPC)

The study of structural breaks in time series stemmed from quality control, but is now an integral part of a wide variety of fields. These applications include economics (Rodrigues & Rubia, 2011) and finance (Severin & Schmid, 1998), education, medicine (Bottle & Aylin, 2008), health services (Woodall, 2006).

SPC methods are used to detect when a stable process- one with fixed mean level and fixed variation- departs from stability. Most traditional diagnostics tools, like Shewhart control charts, popular in Statistical Process Control are used to define a standard of quality for manufacturing process and to determine whether the determined quality is being maintained by the process. The most important factor is the variability in the quality of the finished product. No matter how much attention is paid towards quality of a product, a certain amount of variability is unavoidable and is a function of random forces and likely to be beyond control. Other methods include Change point detection (CPD) models whose goal is recognizing regime change events and adapting the predictive model appropriately. The Bayesian change point analysis assume a change point model of the parameters, integrating out the uncertainty in the parameters, rather than using a point estimate.

2.1.1 Quality Control Charts

The quality control charts, suggested by (Page, 1954) and detailed in (Hawkins & Olwell, 1998) and (Montgomery, 2007) were originally designed for industrial and manufacturing processes to define a standard of quality and determine whether that standard is being maintained by the process over time. The idea of standard control charts is to take the individual quality measures or statistics- usually means- of subsamples of these measures and plot them on a marked chart with control limits from the target value. It is on these plots that the unusual patterns or departures from state of statistical control will be discovered. We devote the next section to review some of the most common control charts.

2.1.2 Shewhart charts

One of the most widely used is the Shewhart control charts (Shewhart, 1926) which monitor the production process and detect any significant deviation from a chosen quality characteristic of the products. A sample of fixed size is drawn at each regular time interval and desired statistic is computed. A sequence of such statistics are represented graphically in the form of a control chart. When a statistic falls outside of pre-determined control limits e.g. 'three-sigma control limits', the production process, at that time, is said to be out-of-control and a warning sign is raised. See figure 2.5.

Shewhart charts are sensitive to large process shifts however the probability of these charts detecting small shift fast is quite small.

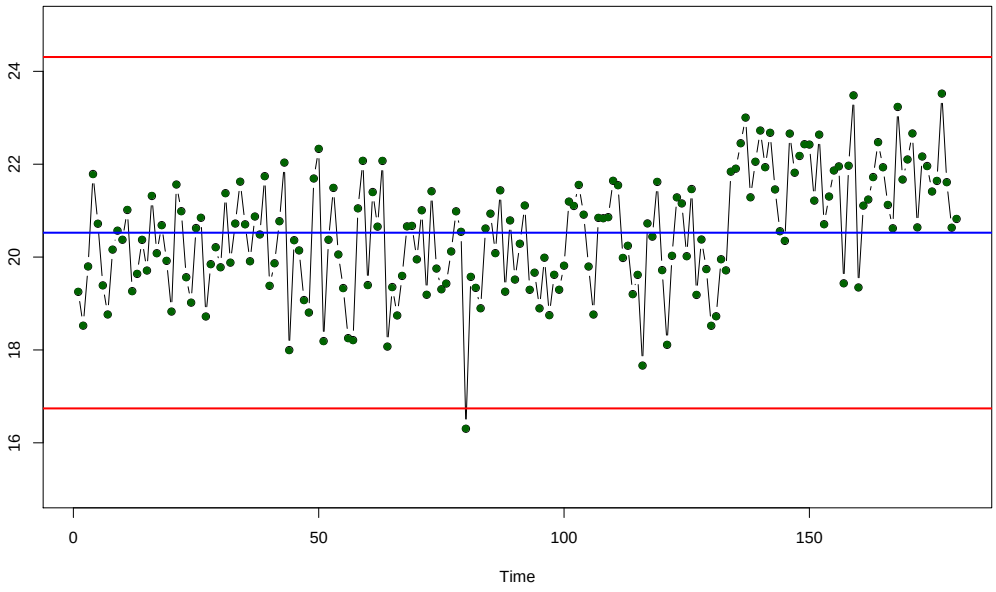

Figure 2.5: Mean QC Chart

2.1.3 Cumulative Sum (CUSUM) charts

Another most widely used control charts is the cumulative sum (CUSUM) (Page, 1954) charts. They are procedures for mean and uses cumulative history, or the past information

of the process, to help in detecting small systematic departures from its normal and stable condition. These changes are detected easily and faster than in the standard Shewhart charts, however for large, abrupt shifts Shewhart chart detect much faster. The CUSUM are non-parametric and do not make use of a particular time series model fit.

The CUSUM charts are build on principles of Maximum Likelihood Estimation (MLE). The standard CUSUM chart for controlling the process mean takes samples from the process at a fixed interval and uses a control statistic based on cumulative sum of differences between the sample mean and the target value. The procedure for CUSUM is as follows: Suppose the quality measurements $X_1, X_2, \ldots$ are taken sequentially with time, and assume that X_i is normally distributed with mean μ and variance σ^2, that is $X_i \sim N(\mu_i, \sigma^2)$ $i = 1, 2, \ldots$ and the variance σ^2 is known and remains constant. The idea is that, if the process is in control then any mean μ_i is equal to the target mean μ_0. This is the condition that need be monitored.

The 2-sided CUSUM chart is based on the cumulative statistics, S_i and T_i, $i = 1, 2, \ldots$, where

$$S_i = max(0, S_{i-1} + Z_i - k)$$

$$T_i = min(0, T_{i-1} + Z_i + k),$$

where $Z_i = \frac{(X_i - \mu_0)}{\sigma}$ and $k > 0$. The cumulative sums is given by $C_i = \sum_{j=1}^{i} Z_j, i = 1, 2, \ldots$ As the number of measurements are taken the probability that the CUSUM value may drift into extreme values increases. This is corrected by the reference value $k = \frac{\delta\sigma}{2}$ where δ the amount of shift in the process mean that we wish to detect. The process is out of control if either S_i or T_i exceed the control limit determined by a value $h > 0$. The choice of h is dependent on how sensitive the method is meant to be. The smaller it is the quicker will any departure from target be detected but also the more likely a false alarm will occur. In most cases h is chosen to be five times the process standard deviation or computed by

$h = \frac{\sigma}{\delta}\ln\left(\frac{1-\beta}{\alpha}\right)$ where α is the probability of false alarm and β is probability of failing to detect a shift in the mean when it has actually occurred. The CUSUM is expected to signal whenever $S_N \geq h$ or $T_N \leq -h$. The value N is popularly known as the run length and defined as the number of measurements between each false alarm when the process is still in control. Its average value is known as the average run length (ARL).

Example 1 *To illustrate this consider series of 20 observations whose first 15 are sampled from standard normal distribution and the rest are drawn from a normal distribution with mean* $\mu = 1$ *and* $\sigma = 1$. *We want to detect the upward shift in mean, so* $\delta = 1$ *and therefore* $k = 0.5$. *From previous discussion* $h = 5$

The simulated values are shown in the Table 2.1

The CUSUM chart for this data is shown in Figure 2.6 and it is clear that the out-of-control signal is given after the 18th observation.

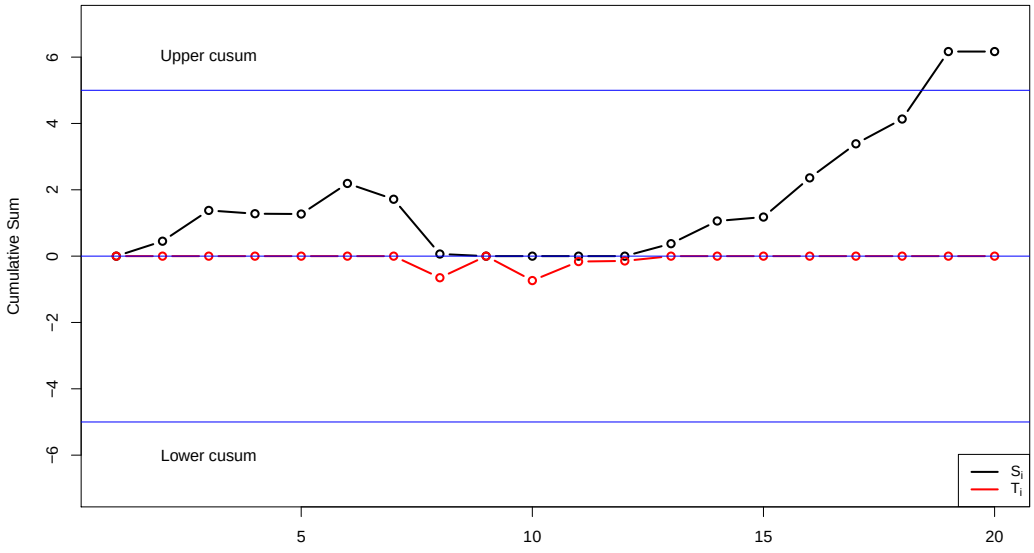

Figure 2.6: CUSUM Chart

Table 2.1: Simulated data and associated CUSUM statistics

i	X_i	C_i	S_i	T_i
1	-2.15	-5.15	0.00	0.00
2	0.95	-1.20	0.45	0.00
3	1.43	2.38	1.38	0.00
4	0.40	1.83	1.28	0.00
5	0.49	0.89	1.27	0.00
6	1.42	1.91	2.19	0.00
7	0.02	1.45	1.72	0.00
8	-1.15	-1.13	0.06	-0.74
9	0.28	-0.87	0.00	-0.16
10	-1.24	-0.96	0.00	-0.14
11	0.07	-1.16	0.00	0.00
12	-0.48	-0.41	0.00	0.00
13	0.88	0.39	0.38	0.00
14	1.18	2.06	1.06	0.00
15	0.62	1.80	1.18	0.00
16	1.68	2.30	2.36	0.00
17	1.53	3.21	3.39	0.00
18	1.25	2.77	4.13	0.00
19	2.53	3.78	6.17	0.00
20	0.50	3.03	6.17	0.00

2.1.4 Exponentially Weighted Moving Average (EWMA)

This control scheme, introduced by (Roberts, 1959), utilizes the statistic A_t,

$$A_t = \phi y_t + (1-\phi)A_{t-1}, \qquad 0 < \phi \leq 1, \qquad t = 1, 2, \ldots$$

and some determined upper and lower limits. The sequentially observed data y_t can be the actual observed value or the sample mean from designed sampling strategy from the process. A_0 is often taken to be the process target value, μ_0 . The control limits are determined as follows:

$$\mu_0 \pm L\sigma\sqrt{\frac{\phi}{(2-\phi)}(1-(1-\phi)^{2t})}$$

where L is the width of the control limits. Both ϕ and L are chosen after specifying desired ARL and the shifted anticipated. The EWMA have proved to be effective against small shifts but, just like CUSUM, does not react quickly to large shifts as compared to Shewhart chart. The comparison of the 3 charts is displayed in Figure 2.7.

Example 2 *We use the same data as in Example 1, and let* $\mu_0 = 1, \phi = 0.1, L = 2.7$. *The results are displayed in Figure 2.7*

In a serially dependent processes, parametric models are often used to describe explicitly the structural dependence assumed in the data while at the same time seek potential structural breaks. Most commonly used model are class of Autoregressive Moving Average (ARMA) Tsay (1988) and Generalized Autoregressive conditional Heteroscedasticity (GARCH)(Bollerslev, 1986) type models.

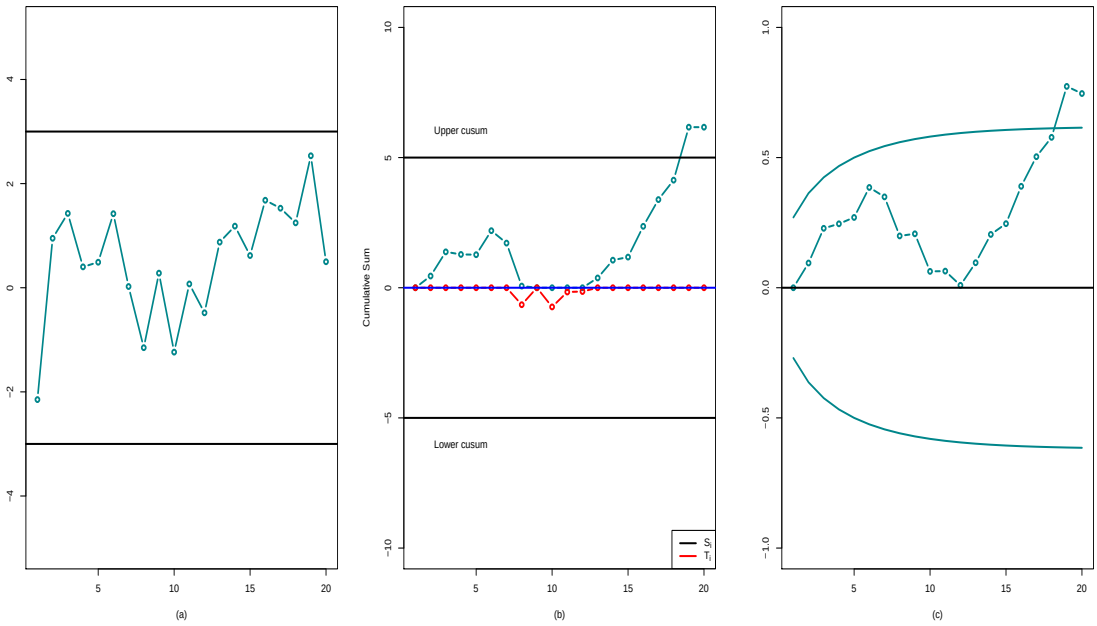

Figure 2.7: (a)Shewhart,(b) CUSUM and (c) EWMA Chart

2.2 ARMA and GARCH models

The ARMA model for outliers proposed by Box and Tiao (1975) involves the unobservable Z_t related to observed series y_t by the function

$$y_t = f(t) + Z_t$$

where $f(t)$ is parametric function that represent exogenous disturbance of Z_t, and the Z_t is modelled by ARMA model

$$\phi(B)Z_t = \varphi(B)a_t$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2, \ldots \phi_p B^p$ and $\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 \ldots \varphi_q B^q$ are Autoregressive and Moving Averages polynomials in B of degrees p and q respectively. B is backwardshift operator such that $BZ_t = Z_{t-1}$, $\{a_t\}$ is a sequence of independent normally distributed variables with mean zero and variance σ_a^2

The function $f(t)$ is designed

$$f(t) = \omega_0 \frac{\omega(B)}{\delta(B)} \xi_t^{(d)}$$

where $\omega(B) = 1 - \omega_1 B - \omega_2 B^2 - \ldots - \omega_s B^s$ and $\delta(B) = 1 - \delta_1 B - \delta_1 B - \ldots - \delta_r B^r$ are polynomials in B with degrees s and r respectively. ω_0 is the magnitude of the outlier or the initial jump of the series and $\xi_t^{(d)}$, is an indicator variable that signifies the occurrence of outlier or structural break at point d .

For an outlier (additive) model, $\omega(B) = \delta(B)$

$$\xi_t^{(d)} = \begin{cases} 1, & t = d \\ 0, & t \neq d \end{cases}$$

To detect a structural change, the $\delta(B)$ is taken to be equal to 1 and

$$\xi_t^{(d)} = \begin{cases} 1, & t \geq d \\ 0, & t < d \end{cases}$$

Other special cases for $\omega(B)/\delta(B)$ are discussed in (Box & Tiao, 1975) and (Tsay, 1988) In most cases the parameters involved in these models are usually unknown and practically they are estimated from the data. Outliers and structural breaks problems have been considered as hypothesis tests with null describing the model with no outlier or breaks. The alternative contains one or multiple outliers or breaks. Under null hypothesis, the maximum likelihood estimates (MLE) are consistent, and often suggested, and can therefore be used to estimate the parameters and to design relevant test statistics (Aue & Horváth, 2013). These test statistics are used to identify outliers or structural breaks if any. The Weighted Likelihood estimation method also provide efficient and robust estimators for ARMA models. The idea of outlier detection using likelihood ratio test -

when the location and the type is known- in autoregressive processes was first proposed by (Fox, 1972)

(Ardelean, 2012) proposed the use of the (GARCH) process in detecting outliers and structural breaks in time series. A real-valued discrete time stochastic process $(X_t)_{t\in\mathbb{Z}}$ is a GARCH(p,q) process if:

$$X_t|\mathcal{F}_{t-1} = \sigma_t\varepsilon_t,$$
$$\sigma_t^2 = (\sigma_t(\psi))^2 = \alpha_0 + \sum_{i=1}^{p}\alpha_i X_{t-i}^2 + \sum_{i=1}^{q}\beta_i\sigma_{t-i}^2$$

where $\mathcal{F}_t$ denote the information set of the process up to time t, and the innovations ε_t is sequence of $i.i.d$ random variables from some distribution $\mathcal{G}$ with $E_{\mathcal{G}}(\varepsilon_t) = 0$ and $E_{\mathcal{G}}(\varepsilon_t^2 = 1)$. $\psi = (\alpha_0, \alpha_1, \ldots \alpha_0, \beta_1, \ldots \beta_q), \alpha_0 > 0, \alpha_i \geq 0, i = 1, \ldots p$ and $\beta_j \geq 0, j = 1, \ldots q$

Following (Ardelean, 2012), both the outliers can be modeled as follows:

additive:

$$Y_t = X_t + \varepsilon I_t(\tau)$$
$$X_t|\mathcal{F}_{t-1} \sim N(0, \sigma_{t-1}^2)$$
$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p}\alpha_i X_{t-i}^2 + \sum_{i=1}^{q}\beta_i\sigma_{t-i}^2$$

innovational

$$Y_t = X_t + \varepsilon I_t(\tau)$$
$$X_t|\mathcal{F}_{t-1} \sim N(0, \sigma_{t-1}^2)$$
$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p}\alpha_i Y_{t-i}^2 + \sum_{i=1}^{q}\beta_i\sigma_{t-i}^2$$

where $(X_t)_{t\in\mathbb{Z}}$ is the underlying and unobserved GARCH(1,1) process and $(Y_t)_{t\in\mathbb{Z}}$ is the

observed process, $\varepsilon \in \mathbb{R}$ is the size of the outlier occurring at time time $\tau \in \mathbb{Z}$, and $I_t(\tau)$ is an indicator function which is equal to 1 if $\tau = t$ and 0 otherwise. It is obvious from the relation that GARCH(1,1) parameterizes the conditional variance in terms of ARMA(1,1) To test for occurrence of outlier and structural breaks simultaneously the model is modified such that

$$Y_t = X_t + \varepsilon_1 I_t(\tau)$$
$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 + \sum_{i=1}^{p} \varepsilon_{1+i} I_{\tau-i}$$

where ε_1 is the size of the outlier occurring at time τ and $\varepsilon_2, \ldots \varepsilon_{p+1}$ is the size of each structural break. Due to the ARMA representation, the parameters in the model $\psi = (\alpha_0, \alpha_1, \ldots \alpha_p, \beta_1, \ldots \beta_q, \varepsilon_1, \ldots \varepsilon_{p+1})$ can be estimated using MLE. By taking $\hat{\psi}_0 = (\hat{\alpha}_0, \hat{\alpha}_1, \ldots \hat{\alpha}_p, \hat{\beta}_1, \ldots, \hat{\beta}_q)$ to be the restricted MLE and $\hat{\psi}_1 = (\hat{\alpha}_0, \hat{\alpha}_1, \ldots \hat{\alpha}_p, \hat{\beta}_1, \ldots, \hat{\beta}_q, \hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_{p+1})$ to be the unrestricted MLE, the likelihood ratio test statistic, for testing the hypothesis of no outlier or break at time τ, that is

$$Ho : \varepsilon_1 = \varepsilon_2 = \ldots = \varepsilon_{p+1} = 0$$

can be computed, for every $\tau \leq T$, as

$$\lambda_\tau = 2(\log L(\hat{\psi}_0) - \log L(\hat{\psi}_1)) \sim \chi^2_{(p+1)}$$

2.3 Regression models

Regression models have also been used in modeling outliers and structural breaks. Consider the general linear regression model

$$Y_t = \mathbf{X}_t \boldsymbol{\beta} + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \qquad t = 1, 2, \ldots T.$$

where $\mathbf{X}_t = (1, X_{1,t}, \ldots, X_{p,t})$ is a vector of the intercept and p non-random explanatory variables and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\mathsf{T}$ are the regression coefficients.

These unknown coefficients are usually estimated by ordinary least-squares method

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}Y, \quad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_T \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X_1} \\ \vdots \\ \mathbf{X}_T \end{bmatrix}$$

Outliers with respect to the explanatory variables are called the leverage points; they can have adverse effect on the regression model. Leverage points do not necessarily correspond to outliers and also their response variable need not be outliers.

The predicted or the fitted values, $\hat{Y}$ are computed using the data matrix and the estimated coefficients

$$\hat{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

The ordinary residuals $\hat{\epsilon}$, is the difference between the predicted and the observed values

$$\hat{\epsilon} = Y - \hat{Y}$$

are the widely used measures in identifying outliers in regression models. The techniques available involve deleting rows with suspicious observation or leverage point and compute statistics thereafter. Examination is then done on the effect of each row deletion on the estimated coefficients and their estimated covariance structure, the predicted values, and the residuals. Most common outlier diagnostics involve statistics mostly computed using the estimated regression coefficients, are briefly discussed below

2.3.1 Hat Matrix

The hat matrix denoted by H and defined as

$$H = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$$

plays an important role in identifying outliers or leverage points. The diagonal elements h_{ii} of H being the amount of leverage exerted by the i^{th} observation on the i^{th} fitted value. The i^{th} observation is an influential point when h_{ii} exceeds $2p/T$, where p is the rank of the $\mathbf{X}$ matrix.

2.3.2 Cook's Distance

Cook's Distance statistic proposed by (Cook, 1979)

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})^\mathsf{T}(\hat{Y}_{(i)} - \hat{Y})}{ps^2}$$

follows $F_{p,T-p}$ distribution, where $\hat{Y}_{(i)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(\boldsymbol{i})}$ with $\hat{\boldsymbol{\beta}}_{(\boldsymbol{i})}$ as the vector of estimated regression coefficients with the i^{th} row deleted, s^2 is the estimator of σ^2. The D_i statistic has a cut-off of 4/p and large values indicates an outlier or leverage point.

2.3.3 DFFITS

The $DFFITS_i$ is the difference between the fitted response variable, $\hat{Y}_i$ from the full model and the predicted values $\hat{Y}_i(i)$ obtained after removing the i^{th} observation from the dataset.

$$DFITTS_i = \frac{\hat{Y}_i - \hat{Y}_i(i)}{\sqrt{\hat{\sigma}^2_{(i)}h_{ii}}}$$

where h_{ii} is the i^{th} diagonal element of the hat matrix, H. A value is considered suspicious if $|DFITTS| > 1$ for small to medium data sets and for large data sets, if

$|DFITTS| > 2\sqrt{p/T}$

2.3.4 DFBETAS

This is the change in the estimate of regression coefficients that would occur if the i^{th} row is removed.

$$DFBETAS_{j(i)} = \frac{\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_{j(i)}}{\sqrt{\hat{\sigma}^2_{(i)}(X^\intercal X)^{-1}_{jj}}}$$

The cut-off value for $DFBETAS$ for small to medium data set is 1 while in large data sets a value $|DFBETAS| > 2/\sqrt{T}$ is considered suspicious.

These numerical measures though effective in case of existence of single outlier, may fail if more than one outliers exists. Moreover, when data is collected over time, serial dependence is a significant component and therefore model assumption of independent errors is violated and model can't be used.
However if we allow for time-varying coefficients the now generalized regression, discussed in Chapter 3, can be used to detect outliers and structural breaks in time series data.

2.4 Some Multivariate Outlier Detection Methods

2.4.1 Static data

In a multivariate data the classical approach in detecting outliers is to consider the distance of a each observation as well as the shape and the size of the data. The shape and size of multivariate data are expressed by the covariance matrix.
The basic statistical measure for outliers detection and which takes also into account the covariance matrix is the Mahalanobis distance (Mahalanobis, 1936). The statistic is computed using the estimate of multivariate location- usually the mean- and the sample covariance matrix. If m-dimensional multivariate sample $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ is a random

sample from multivariate normally distributed data with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the Mahalanobis' distance

$$MD_t = \left((\mathbf{y}_t - \hat{\boldsymbol{\mu}})^\mathsf{T} \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{y}_t - \hat{\boldsymbol{\mu}})\right)^{\frac{1}{2}}$$

identifies the observation that are very far from the centre of the data cloud or the centroid. A test statistic for MD_t is given by

$$\frac{(T-m)T}{(T^2-1)m} MD_t$$

which is approximate F distribution with degrees of freedom m and $T-m$.
Since the sample estimates $\hat{\mu}$ and $\hat{\Sigma}$ are very sensitive to outliers, which the MD_t is meant to identify, they need to be estimated using a robust procedure in order to provide a credible and reliable criterion. There is significant literature on robust estimation of MD_t (Franklin, Thomas, & Brodeur, 2000), (Peña & Prieto, 2001)
Due to calculations of the covariance matrix estimate, $\hat{\Sigma}$, the Mahalanobis distance is computationally expensive, with runtime $O(T^2 m)$, for large and high dimensional data sets.
Another popular statistical measure is the Euclidean distance

$$d_{x,y} = \sqrt{\sum_{i=1}^{T} (x_i - y_i)^2}$$

Both Mahalanobis and Euclidean distance measures are important ingredients in proximity based techniques for outlier detection, such as clustering and k-Nearest Neighbors algorithms.

2.4.2 Time series data

(Galeano, Peña, & Tsay, 2006) used m-dimensional vector $\mathbf{Z}_t = (Z_{1,t}, Z_{2,t}, \ldots, Z_{m,t})^\mathsf{T}$

following the vector ARMA (VARMA)model

$$\Phi(B)\mathbf{Z}_t = \boldsymbol{\varphi}(B)a_t$$

where $\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \ldots - \Phi_p B^p$ and $\boldsymbol{\varphi}(B) = 1 - \boldsymbol{\varphi}_1 B - \boldsymbol{\varphi}_2 B^2 - \ldots - \boldsymbol{\varphi}_q B^q$ are $m \times m$ matrix polynomial of degrees p and q. B, like in the univariate case, is a backward shift operator such that $B\mathbf{Z}_t = \mathbf{Z}_{t-1}$ and $a_t = (a_{1,t}, a_{2,t}, \ldots a_{m,t})^{\mathsf{T}}$ is a sequence of uncorrelated Gaussian random vectors with mean $\mathbf{0}$ and positive-definite covariance matrix $\boldsymbol{\Sigma}$ $\mathbf{Z}_t$ are related to the observed series $\mathbf{Y}_t = (Y_{1,t}, Y_{2,t}, \ldots, Y_{m,t})^{\mathsf{T}}$ by the function

$$\mathbf{Y}_t = \mathbf{Z}_t + \boldsymbol{\alpha}(B)\boldsymbol{\omega}\xi_t^{(d)}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_m)^{\mathsf{T}}$ is the size of the outlier, $\xi_t^{(d)}$ is the indicator variable such that $\xi_t^{(d)} = 1$ if $t = d$ and zero otherwise. The matrix $\boldsymbol{\alpha}(B)$ define the type of outlier with $\boldsymbol{\alpha}(B) = \boldsymbol{I}$ indicating an additive outlier and if $\boldsymbol{\alpha}(B) = \boldsymbol{\varphi}(B)/\Phi(B)$ indicates a multivariate structural break. Other special cases of $\boldsymbol{\alpha}(B)$ are discussed in (Galeano et al., 2006). (Atkinson, Koopman, & Shephard, 1997) used the *Gaussian State space model*, details given in Chapter 3, with regression variables through which shocks- outliers or breaks- were introduced.

Chapter 3

State Space Model

3.1 General State Space Model

State space models provide an effective basis for practical time series analysis and forecasting (Durbin & Koopman, 2012), (Harrison & West, 1997), (Aoki, 1990).
The models are highly applicable in various fields and disciplines including computer vision (i.e. tracking), control theory, econometrics, population dynamics. The state space model involves two processes: the latent or unobserved Markov state process, $\{\theta_t\}_{t\geq 1}$, $\theta_t \in \mathbb{R}^p$ and the noisy observation process $\{y_t; t \in \mathbb{N}\}$, $y_t \in \mathbb{R}^m$ that is related to the state process. The state space model is specified through descriptions of the sampling distribution, the state vector evolution, and the initialization of the state vector. See equations 3.1 and 3.2
The state vector contains all relevant information required to describe the system under investigation. It may contain regression variables or components of time series such as level, trend, seasonal or cyclic components. In tracking problems, for example, this information could be related to the kinematic characteristics of the target object. In an econometric problem, it could be related to monetary flow, interest rates, inflation, stock markets etc.

Conditional probability

The conditional probability of a variable a given b is defined

$$p(a|b) = \frac{p(a,b)}{p(b)}$$

from where Bayes' rule follows quickly

$$p(a|b) = \frac{p(b|a)p(a)}{p(b)}$$

or more conceptually

$$Posterior = \frac{Likelihood \times Prior}{Marginal\ likelihood\ (evidence)}$$

Conditional Independence

A variable a is conditionally independent of b given c, denoted by $a \perp b \mid c$, if $p(a \mid b, c) = p(a \mid c)$.

Let $y_{1:t} := (y_1, y_2 \ldots, y_t)$ represent all the data or information up to and including time t, and $\theta_{0:t} := (\theta_0, \ldots., \theta_t)$ be state representation up to time t. The state space model are based on two very important assumptions:

- conditional on the parameter ψ state process $\{\theta_t\}_{t\geq 0}$ is a Markov process; that is $p(\theta_t|\theta_{0:t-1}, \psi) = p(\theta_t|\theta_{t-1}, \psi)$.
- y_t depends only on θ_t and conditional on the state process $\{\theta_t\}_{t\geq 0}$, the $\{y_t\}$'s are independent. $p(y_t|\theta_{0:t}, y_{1:t-1}) = p(y_t|\theta_t)$

This conditional dependence is demonstrated in Figure 3.1.
The general state space model is defined by these two equations

$$y_t|\theta_t, \psi \sim p(y|\theta_t, \psi) \tag{3.1}$$

$$\theta_t|\theta_{t-1}, \psi \sim p(\theta_t|\theta_{t-1}, \psi) \tag{3.2}$$

with initial density $p(\theta_0|\psi)$ and the prior $p(\psi)$ where ψ is a vector of parameters, usually unknown.

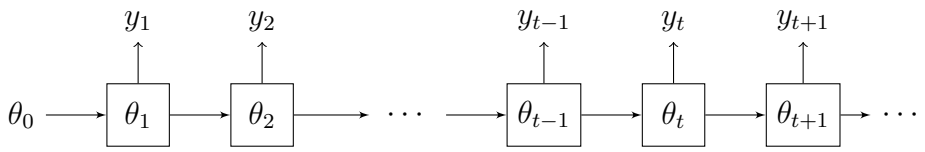

Figure 3.1: Structural dependence of state space model

The goal in statistical inference on state space models is, based on the available data, to estimate the unobserved states and the unknown parameters in the model and predict the states and/or future observations. The estimation of the state vector entails filtering and smoothing problem. This inference is achieved by computing conditional and or marginal distributions based on the joint distribution

$$p(\theta_{0:t}, \psi, y_{1:t}) = p(\psi) \underbrace{p(\theta_0|\psi) \prod_{j=1}^{t} p(\theta_j|\theta_{j-1}, \psi)}_{p(\theta_{0:t}|\psi)} \underbrace{\prod_{j=1}^{t} p(y_j|\theta_j, \psi)}_{p(y_{1:t}|\theta_{0:t}, \psi)} \tag{3.3}$$

Given data up to time t and assuming that ψ is known, the marginal distribution $\{p(\theta_t|y_{1:t})\}_{t\geq 1}$ also known as the filtering density is obtained via Bayes' rule as

$$p(\theta_t|y_{1:t}) = \frac{p(y_t|\theta_t)p(\theta_t|y_{1:t-1})}{p(y_t|y_{1:t-1})}$$

To obtain an estimate of the states joint distribution $p(\theta_{0:t}|y_{1:t})$, again, by Bayes' rule we have,

$$p(\theta_{0:t}|y_{_{1:t}}) = \frac{p(\theta_{0:t}, y_{1:t})}{p(y_{1:t})} = \frac{p(y_t|\theta_t)p(\theta_t|\theta_{t-1})p(\theta_{0:t-1}|y_{1:t-1})}{p(y_t|y_{1:t-1})}$$

The marginal likelihood $p(y_{1:t})$ can be obtained as

$$p(y_{1:t}) = \int \cdots \int p(\theta_{0:t}, y_{1:t}) d_{\theta_{0:t}}$$

State smoothing involves going back in time and deriving the states values using all the

available data. The smoothing is achieved by the density $p(\theta_t|y_{1:T})$ for $t < T$,

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

For predicting or forecasting future states and observations, the $k-$steps ($k \geq 1$) predictive densities for the states and observation respectively, is given by

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

Parameter learning is achieved via the density $p(\psi|y_{1:t})$.

For linear Gaussian models, all posteriors are Gaussian and the above quantities can be computed analytically by using well established algorithms which include Kalman filter and smoother (Kalman et al., 1960) and the Forward Filtering Backward Sampling (FFBS) (Frühwirth-Schnatter, 1994). For non-linear and non-Gaussian models, computing the above quantities in closed form is analytically intractable, and numerical approximation, in particular Markov Chain Monte Carlo (MCMC), is required. However for online inferences –where data arrive rapidly and frequently and hence fast and efficient updates of posterior quantities is required –MCMC are ineffective.

3.2 Dynamic Linear Model (DLM)

Also known as *Gaussian State Space Model*, DLM is a class of state space models where equations (3.1) and (3.2) both are linear and Gaussian (West & Harrison, 1989), (Harrison & West, 1991) (Harrison & West, 1997). The model is specified by initial distribution

$\theta_0 \sim \mathcal{N}(m_0, C_0)$ and the equations

$$y_t = F_t\theta_t + v_t \qquad v_t \sim \mathcal{N}(0, V_t) \tag{3.4}$$

$$\theta_t = G_t\theta_{t-1} + w_t \qquad w_t \sim \mathcal{N}(0, W_t) \tag{3.5}$$

where F_t and G_t are known $m \times p$ and $p \times p$ transition matrices. The possible time-dependent quantities F_t, G_t, V_t and W_t may depend on a parameter vector ψ.
By allowing for time-varying coefficients we can show that DLM is a generalization of linear regression model,

$$y_t = \mathbf{X}_t\boldsymbol{\beta}_t + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_t^2), \qquad t = 1, 2, \ldots, T$$

where $\mathbf{X}_t = (1, X_{1,t}, \ldots, X_{p,t})$ is a vector of the intercept and p non-random explanatory variables and $\boldsymbol{\beta}_t = (\beta_{0,t}, \beta_{1,t}, \ldots, \beta_{p,t})^\mathsf{T}$ and we model evolution of coefficients

$$\beta_{j,t} = \beta_{j,t-1} + w_{j,t}, \quad j = 0, 1, \ldots, p$$

which is a DLM with $F_t = [X_t],\ \theta_t = [\beta_{0,t}, \beta_{1,t}, \ldots, \beta_{p,t}]^\mathsf{T},\ V_t = \sigma_t^2$, and $G = I_p$, identity matrix. (Petris et al., 2009)
The random walk plus noise also known as the local level model

$$\begin{aligned} y_t &= \gamma_t + v_t & v_t &\sim \mathcal{N}(0, V) \\ \gamma_t &= \gamma_{t-1} + w_t & w_t &\sim \mathcal{N}(0, W) \end{aligned}$$

is the simplest DLM with $m = p = 1$ hence $F = G = 1$ and $\theta = \gamma$ and $\psi = (V,\ W)$

3.2.1 Structural Time Series

Structural time series model is a linear combination of a random error component, ε –with zero mean and a constant variance σ^2–and at least one of the three structural components, namely; the trend ($\mathtt{T}$), cycle ($\mathtt{C}$), and seasonal ($\mathtt{S}$) components.
The basic structural time series model is shown below

$$y_t = \mathtt{T}_t + \mathtt{C}_t + \mathtt{S}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2) \;\; t = 1, \ldots, T$$

Any structural model can be represented as DLM. For example, the locally linear trend model which is of the form

$$y_t = \mathtt{T}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2) \quad t = 1, \ldots, T \tag{3.6}$$

and the linear trend is quickly derived from the deterministic function

$$\mathtt{T}_t = \mathtt{T}_{t-1} + \rho_{t-1} + \vartheta_t \tag{3.7a}$$

$$\rho_t = \rho_{t-1} + \xi_t \tag{3.7b}$$

where the innovations ϑ_t with zero mean and variance σ^2_ϑ account for vertical or the upward and downward shift of the trend. The innovations ξ_t have zero means and variance σ^2_ξ and they account for the trend's change in slope. These innovations ϑ_t and ξ_t as well as ε_t are mutually uncorrelated.
Using the equations (3.6), (3.7a) and (3.7b) we have a DLM with

$$G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \theta_t = \begin{pmatrix} \mathtt{T}_t \\ \rho_t \end{pmatrix}, \quad F = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad W = \begin{pmatrix} \sigma^2_\vartheta & 0 \\ 0 & \sigma^2_\xi \end{pmatrix}, \quad V = \sigma^2$$

and $\psi = (\sigma^2, \sigma^2_\vartheta, \sigma^2_\xi)$

3.2.2 ARMA representation

A significant number of state space representations for ARMA models exist. For detailed discussion on these representations, see (Petris et al., 2009), (Brockwell & Davis, 2009), (Kitagawa & Gersch, 1996), (Kedem & Fokianos, 2002). To illustrate, lets consider the data $y_t = x_t + v_t,\ \ v_t \sim N(0, V_t)$ where x_t is unobserved autoregressive process of order p ; $x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t$, with $\epsilon_t \sim N(0, \sigma^2_{\epsilon_t})$. This is a DLM with

$$G_t = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & \cdots & \cdots & 1 & 0 \end{pmatrix}, \quad \theta_t = \begin{pmatrix} x_t \\ x_{t-1} \\ x_{t-2} \\ \vdots \\ x_{t-p+1} \end{pmatrix}, \quad w_t = \begin{pmatrix} \epsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$F_t = [1, 0, 0, \ldots, 0, 0]$$

and $\psi = (\phi_1, \phi_2, \ldots, \phi_p, V_t, \sigma^2_{\epsilon_t})$

Since many linear models including ARMA models admit state space representation, the statistical inference on state space models can be applied to both stationary and non-stationary data. Moreover, the state space models provides components that are easier to interpret unlike those from ARMA models.

When DLM is fully specified the state estimation, smoothing and or predictions as well as observation predictions, can be carried out by using the Kalman Filter and Smoother and FFBS algorithms. However when ψ is unknown, numerical methods −and in particular the Monte Carlo methods−are required.

3.2.3 Kalman Filter

Given the information available at time t, the Kalman filter(Kalman et al., 1960) is a set of recursion equations- the predictions and updating equations- for determining optimal estimates of the state vector θ_t.

First, let $m_t = E(\theta_t|y_{1:t})$ be the optimal estimator of θ_t based on information up to time t, and $C_t = E[(\theta_t - m_t)(\theta_t - m_t)'|y_{1:t}]$ be the mean square error (MSE) matrix of m_t.

The prediction step takes place prior to arrival of the data at time t, and involves predicting the states

$$p(\theta_t|y_{1:t-1}) = \int p(\theta_{t-1}|y_{1:t-1})p(\theta_t|\theta_{t-1})d\theta_{t-1}$$

Given, at time $t-1$, that $\theta_{t-1} \sim N(m_{t-1}, C_{t-1})$, then from (3.4) and (3.5), it follows quickly that the parameters for the predictive distribution of θ_t, given the information up to time $t-1$, will be

$$p(\theta_t|y_{1:t-1}) \sim \mathcal{N}(m_t^*, C_t^*)$$

where

$$\begin{aligned} m_t^* &= E(\theta_t|y_{1:t-1}) = G_t m_{t-1} \\ C_t^* &= E[(\theta_t - m_{t-1})(\theta_t - m_{t-1})^{\mathsf{T}}|y_{1:t-1}] \\ &= G_t C_{t-1} G_t^{\mathsf{T}} + W_t \end{aligned}$$

and the corresponding optimal predictor of y_t given all the information up to time $t-1$ is

$$\begin{aligned} y_{t|t-1} &= E(y_t|y_{1:t-1}) = F_t m_t^* \\ &= F_t G_t m_{t-1} \end{aligned}$$

Once the new observation y_t become available, the prediction error
$e_t = y_t - y_{t|t-1} = y_t - F_t G_t m_{t-1}$ and its MSE

$$E(e_t e_t^{\mathsf{T}}) = Q_t = F_t C_t^* F_t^{\mathsf{T}} + V_t$$

and the states updating step is defined, by Bayes formula

$$\begin{aligned} p(\theta_t|y_{1:t}) &= \frac{p(y_t|\theta_t)p(\theta_t|y_{1:t-1})}{p(y_t|y_{1:t-1})} \\ &\propto \mathcal{N}(\theta_t|m_t, C_t) \end{aligned}$$

with optimal predictor of θ_t and its MSE matrix computed as follow:

$$\begin{aligned} m_t &= m_t^* + C_t^* F_t^{\mathsf{T}} Q_t^{-1} e_t \\ C_t &= C_t^* - C_t^* F_t^{\mathsf{T}} Q_t^{-1} F_t C_t^* \end{aligned}$$

3.2.4 Forward Filtering Backward sampling (FFBS)

Given all the data up to time $t = T$, we may be interested in computing $p(\theta_{0:T}|y_{1:T})$.

- Forward Sampling

  This is achieved through Kalman Filter and computes the normal distribution $p(\theta_t|y_{1:t})$ at each $t = 1, 2, \ldots, T$

- Backward-sampling:

At time $t = T$: we sample θ_T^* from $p(\theta_T|y_{1:T})$

For $t = T-1, T-2, \ldots, 0$: sample θ_t^* from normal distribution $p(\theta_t|y_{1:t}, \theta_{t+1}^*)$

The desired output from the FFBS is the sequence $\theta_T^*, \theta_{T-1}^*, \ldots, \theta_0^*$

3.2.5 MCMC in DLM

Inference on DLM with unknown parameters can be carried out by using MCMC approach. Again we let ψ, be the vector of unknown parameters in the DLM. The inference will be based on the posterior distribution $p(\theta_{0:T}, \psi|y_{1:T})$ whose decomposition is as follows

$$p(\theta_{0:T}, \psi|y_{1:T}) = p(\theta_{0:T}|y_{1:T}, \psi)p(\psi|y_{1:T})$$

It is logical, therefore, to use Gibbs sampling technique to sample iteratively from this posterior distribution,

- starting with prior $\psi = \psi^*$
- Apply FFBS algorithm to draw smoothed state vector $\Theta^* = (\theta_T^*, \theta_{T-1}^*, \ldots, \theta_0^*)$ from $p(\theta_{0:T}|\psi^*, y_{1:T})$
- Draw new value of ψ^* from $p(\psi|\Theta^*, y_{1:T})$
- Iterate for large number of times.

Chapter 4

Sequential Monte Carlo (SMC) Methods

In many time series the data do arrive rather frequently and sequentially in time and one is interested in estimating recursively, in real time, the evolving posterior distribution. Markov Chain Monte Carlo (MCMC) methods, though useful for off-line or batch inferences, are ineffective or of limited use for online inferences.

SMC are Monte Carlo technique that have been developed to deal with sequential or online inferences (Doucet, De Freitas, Gordon, et al., 2001). SMC techniques have been developed in a wide range of disciplines (e.g. missile tracking, stock market, medical monitoring) and go under many names: Particle filtering, Bootstrap filtering, the condensation algorithm, Interacting particle approximations, Survival of the fittest among others.

SMC is a 'divide and conquer' approach that evaluates the full posterior by dividing it up into one time step at a time. That is, we want to compute $p(\theta_{0:t}, \psi|y_{1:t})$ sequentially in time t. First we compute $p(\theta_{0:1}, \psi|y_1)$ at time $t = 1$, then $p(\theta_{0:2}, \psi|y_{1:2})$ at time $t = 2$ and so on. Each target distribution is approximated by weighted Monte Carlo samples known as *particles*. This relation is denoted as

$$\left\{\theta_{0:t}^{(i)}, \psi^{(i)}, w_t^{(i)}\right\}_{i=1}^{N} \sim p(\theta_{0:t}, \psi|y_{1:t})$$

where $N \gg 1$, $w_t^{(i)} > 0, \quad \sum_N^{i=1} w_t^{(i)} = 1$.

The idea is that the empirical distribution of this collection converges asymptotically to p

as $N \to \infty$. That is, for any p-integrable function $\Phi_t : \mathbb{R}^{t \times p} \to \mathbb{R}$

$$\sum_{i=1}^{N} w_t^{(i)} \Phi_t(\theta_{0:t}^{(i)}) \xrightarrow{a.s} E_p(\Phi_t) = \int \Phi_t(\theta_{0:t}) p(\theta_{0:t}|y_{1:t}) d\theta_{0:t} \tag{4.1}$$

as $N \to \infty$.

This strategy makes SMC technique fast and thus ideal for online inferences where fast and efficient updates of posterior quantities and forecasts are necessary to deal with high frequency incoming data.

The fundamental concepts in SMC are Bayesian inference, Monte Carlo samples, importance sampling, and resampling.

We now briefly describe two fundamental concepts in SMC; the importance sampling and the resampling technique. The mechanisms through which the particles evolve.

4.1 Importance sampling

It is generally impossible to sample from $p(.)$ therefore we approximate our target distribution $p(.)$ with a proposal density $q(.)$, also called the importance density, which is easy to sample from. The goal is to approximate the expected value of an arbitrary function $g(x)$ using the underlying probability density p(.) and the proposal density. Assume that at time t-1, we have particles $\{\theta_{0:t-1}^{i}\}$ which have been sampled from a proposal density $q_{t-1}(\theta_{0:t-1})$. Since they are not samples from the target density, they are weighted with weights given by

$$w_{t-1}^{(i)} \propto \frac{p_{t-1}(\theta_{0:t-1}^{(i)})}{q_{t-1}(\theta_{0:t-1}^{(i)})}$$

Assume we are interested in finding the expected value of an arbitrary function $\mathrm{g}(\theta)$, with respect to $p(\theta)$. By definition

$$E_{p(\theta)}[\mathrm{g}(\theta)] = \int \mathrm{g}(\theta)p(\theta)d\theta = \int \mathrm{g}(\theta)\frac{p(\theta)}{q(\theta)}q(\theta)d\theta \tag{4.2}$$

$$= E_{q(\theta)}[\mathrm{g}(\theta)\hat{w}(\theta)] \tag{4.3}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} \mathrm{g}(\theta^{(i)})\hat{w}^{(i)} \tag{4.4}$$

where $\theta^{(i)}, i = 1, \ldots, N$ is drawn from $q(\theta)$ and $\hat{w}(\theta) = \frac{p(\theta)}{q(\theta)}$ is the importance weights function. The importance weights function is usually known up to proportionality constant, $\bar{w}(\theta) = k\hat{w}(\theta)$ where k is independent of θ

$$\begin{aligned} k &= \int kp(\theta)d\theta \\ &= \int \bar{w}(\theta)q(\theta)d\theta \\ &= E_{q(x)}[\bar{w}(\theta)] \end{aligned}$$

Now we can re-write equation (4.3) as

$$\begin{aligned} E_{p(\theta)}[\mathrm{g}(\theta)] &= E_{q(\theta)}[\mathrm{g}(\theta)\frac{\bar{w}(\theta)}{k}] = \frac{E_{q(\theta)}[\mathrm{g}(\theta)\bar{w}(\theta)]}{E_{q(x)}[\bar{w}(\theta)]} \\ &\approx \frac{\sum_{i=1}^{N} \mathrm{g}(\theta^{(i)})\bar{w}^{(i)}}{\sum_{i=1}^{N} \bar{w}^{(i)}} \\ &= \sum_{i=1}^{N} \mathrm{g}(\theta^{(i)})w^{(i)} \end{aligned}$$

where $w = \bar{w}^{(i)} / \sum_{j=1}^{N} \bar{w}^{(j)}$.

A good approximation of $p(\theta)$ is therefore

$$\hat{p}(\theta) = \sum_{i=1}^{N} w^{(i)}\delta_{\theta^{(i)}} \tag{4.5}$$

where δ_x is the Dirac delta mass at x.

At time t, the weighted Monte Carlo approximation to $p(\theta_{0:t}, \psi|y_{1:t})$ is given by

$$\hat{p}(\theta_{0:t}, \psi) = \sum_{i=1}^{N} w_t^{(i)} \delta_{(\theta_{0:t}^{(i)}, \psi^{(i)})}$$

where:
$w_t^{(i)} = \bar{w}_t^{(i)} / \sum_{j=1}^{N} \bar{w}_t^{(j)}$ and $\bar{w}_t = \frac{p(\theta_{0:t}^{(i)}, \psi^{(i)}|y_{1:t})}{q(\theta_{0:t}^{(i)}, \psi^{(i)}|y_{1:t})}$

If we use a structured proposal distribution, and assuming that ψ is known

$$\begin{aligned} q_t(\theta_{0:t}|y_{1:t}) &= q_0(\theta_0) q_1(\theta_1|\theta_0, y_1) q_2(\theta_2|\theta_1, \theta_0, y_{1:2}), \ldots, q_t(\theta_t|\theta_{0:t-1}, y_{1:t}) \\ &= q_{t-1}(\theta_{0:t-1}|y_{1:t-1}) q_t(\theta_t|\theta_{0:t-1}, y_{1:t}) \end{aligned}$$

Since $\theta_{0:t-1}^{(i)}|y_{1:t-1} \sim q_{t-1}(\theta_{0:t-1}|y_{1:t-1})$ is available, then we only need to sample

$$\theta_t^{(i)}|\theta_{0:t-1}^{(i)}, y_{1:t} \sim q_t\big(\theta_t|\theta_{0:t-1}^{(i)}, y_{1:t}\big)$$

to obtain $\theta_{0:t}^{(i)}|y_{1:t} \sim q_t(\theta_{0:t}|y_{1:t})$

and the un-normalized weights, $\bar{w}_t^{(i)}$ are updated according to

$$\begin{aligned} \bar{w}_t^{(i)} &= \frac{p(\theta_{0:t}^{(i)}|y_{1:t})}{q(\theta_{0:t}^{(i)}|y_{1:t})} = \frac{p(y_t, \theta_t|\theta_{0:t-1}, y_{1:t-1}) p(\theta_{0:t-1}|y_{1:t-1})}{q_{t|t-1}(\theta_t|\theta_{0:t-1}, y_{1:t}) q_{t-1}(\theta_{0:t-1}|y_{1:t-1})} \\ &= w_{t-1}^{(i)} \frac{p(y_t|\theta_t) p(\theta_t|\theta_{t-1})}{q_{t|t-1}(\theta_t|\theta_{t-1}, y_{1:t})} \end{aligned}$$

4.2 Resampling and auxiliary index

One major problem with particle filter is particle degeneracy. After a few iterations, most particles have negligible weight and the weight is concentrated on few particles only. The other problem is the loss of diversity or sample impoverishment where particles with high

weight are selected more and more often while the others die out slowly. To counter these problems resampling of the particles is recommended. The technique involves sampling new particles, with replacement from a weighted empirical measure related to the current particles. Resampling eliminates particles with low weights and chooses more particles in high probability regions of the state space.

Resampling too often will decreases the number of distinct particles and do increase the Monte Carlo variance of the estimator. However, resampling also reduces the variances of estimates at a later stage. It is important therefore to resample but do so only when it is absolutely necessary. Deciding when to resample is a crucial part of the algorithm and its usually determined by assessing the quality of the current particles. The resampling is done whenever this criterion is above or below a certain predetermined threshold. By doing this, a reasonable number of contributing particles is maintained.

The most commonly used criterion is the effective sample size(ESS). This approach was originally proposed in (Kong, Liu, & Wong, 1994) with idea that the need of resampling increases with increase of the variance of importance weights. Since this variance is unknown, the ESS is used in its place. ESS provides an approximation of the number of independent samples from the target distribution that would be required to provide an estimate of comparable variance.

The simplest way to perform resampling consists of sampling the N new particles from the weighted distribution $\hat{p}_t^N$; the resulting N_t^N are distributed according to a multinomial distribution of parameters w_t^N.

Other sampling schemes include Stratified resampling (Kitagawa, 1996) and residual resampling (Douc & Cappé, 2005). These reduce the variance of N_t^N relatively to that of the multinomial scheme.

4.3 Convergence results

Numerous convergence results are available for SMC methods (Crisan & Doucet, 2002) (Del Moral, 2004).

Again we take $\Phi_t : \mathbb{R}^{t \times p} \to \mathbb{R}$

$$\bar{\Phi}_t = \mathbb{E}_p(\Phi_t) = \int \Phi_t(\theta_{0:t}) p(\theta_{0:t}|y_{1:t}) d\theta_{0:t}$$
$$\hat{\Phi}_t = \int \Phi_t(\theta_{0:t}) \hat{p}(\theta_{0:t}|y_{1:t}) d\theta_{0:t} = \sum_{i=1}^{N} w_t^{(i)} \Phi_t(\theta_{0:t}^{(i)})$$

Under very week assumptions, there exists a constant C such that for any $r > 0$

$$\mathbb{E}[|\hat{\Phi}_t - \bar{\Phi}_t|^r]^{1/r} \leq \frac{C_t}{\sqrt{N}}$$

and

$$\lim_{N \to \infty} \sqrt{N}(\hat{\Phi}_t - \bar{\Phi}_t) \Rightarrow \mathcal{N}(0, \sigma_t^2)$$

These results are however weak since C grows exponentially with time t, in which case we will have to use a time-increasing N to achieve a fixed precision. Stronger convergence results arises when exponential stability is assumed. If the model has properties where for any $\theta_1, \theta_1' \in \Phi$

$$\int |p(\theta_t|y_{1:t}, \theta_1) - p(\theta_t|y_{1:t}, \theta_1')| d\theta_t \leq \rho^{t-1}$$

with $0 \leq \rho < 1$, there exist constants D and $M < \infty$, exponential in $\dim(\theta_t)$, such that for

any $r > 0$

$$\mathbb{E}[|\hat{\Phi}_t - \bar{\Phi}_t|^r]^{1/r} \leq \frac{M}{\sqrt{N}}$$

and

$$\lim_{N\to\infty} \sqrt{N}(\hat{\Phi}_t - \bar{\Phi}_t) \Rightarrow \mathcal{N}(0, \sigma_t^2)$$

where $\sigma_t^2 \leq D$

4.4 Particle Filters

Different Particle fileter algorithms exists and simply differ in the way the importance function $q(\theta_t|\theta_{t-1})$ is chosen.

4.4.1 Bootstrap Filter (BF)

Among the most popular filters is the bootstrap filter (BF), also known as the sequential importance sampling with resampling (SISR) filter,(Gordon, Salmond, & Smith, 1993). The filter uses the state equation (3.2) for state prediction and then the particles are resampled using observation equation (3.1). The algorithm can be summarised as

- Prediction:

  To approximate the density $p(\theta_t|y_{1:t-1}) = \int p(\theta_t|\theta_{t-1}, y_{1:t-1})p(\theta_{t-1}|y_{1:t-1})d\theta_{t-1}$ particles $\tilde{\theta}_t^{(i)}$ are drawn from $p(\theta_t|\theta_{t-1}^{(i)})$ for $i = 1, 2, \ldots, N$

  Here we realise that the importance function is chosen as the prior density of the hidden state.

- Update:

  The particles $\{\tilde{\theta}_t^{(i)}\}_{i=1}^N$ are resampled with weights proportional to their likelihoods, i.e. $w_t^{(i)} \propto p(y_t|\tilde{\theta}_t^{(i)})$

4.4.2 Auxiliary Particle Filter (APF)

Proposed by Pitt and Shephard (1999), the filter resamples previous particles with weights proportional to the proposal density $p(y_t|g(\theta_t))$ for some function g- mean or mode of $p(\theta_t|\theta_{t-1}^{(i)})$

- Draw $\{\tilde{\theta}_{t-1}^{(j)}\}_{j=1}^{N}$ from $\{\theta_{t-1}^{(i)}\}_{i=1}^{N}$ with weights $w_t \propto p(y_t|g(\theta_t^{(i)}))$, where $g(\theta_t^{(i)}) = E(\theta_t|\theta_{t-1})$
- Draw $\{\theta_t^{(i)}\}_{i=1}^{N}$ from $p(\theta_t|\tilde{\theta}_{t-1}^{(i)})$
- Resample with weights $w_t \propto \frac{p(y_t|\theta_t^{(i)})}{p(y|g(\theta_t^{(j)}))}$

4.4.3 The Auxiliary Particle filter with parameter estimation

Proposed by (Liu & West, 1999), this popular filter assumes that for a fixed parameter vector ψ, the set of i.i.d particles $\{\theta_t, \psi^{(i)}\}$ approximate $p(\theta_t, \psi|y_{1:t})$ such that $p(\psi|y_{1:t})$ can be approximated by

$$p(\psi|y_{1:t}) \approx \sum_{i=1}^{N} f_N(\psi; m^{(i)}, h^2\Sigma)$$

Where h^2 is smoothing factor associated with shrinkage factor a, such that $h^2 = (1 - a^2)$ and m and Σ are defined in the summary below:

For $t = 1, 2, \ldots$

Input: Monte Carlo sample $(\theta_{t-1}^{(i)}, \psi_{t-1}^{(i)})$ and weights $w_{t-1}^{(i)}, i = 1, 2, \ldots, N$

- Compute $\bar{\psi}, \Sigma$, the posterior mean and variance matrix of ψ, respectively, from $\psi_{t-1}^{(i)}$ and $w_{t-1}^{(i)}$
- Compute $g(\theta_t^{(i)}) = E(\theta_t|\theta_{t-1}^{(i)}, \psi_{t-1}^{(i)})$ and $m^{(i)} = a\psi_{t-1}^{(i)} + (1 - a)\bar{\psi}$

- sample integer $k \in \{1, 2 \ldots, N\}$ with probability
  $Pr^{(i)} \propto w_{t-1}^{(i)} p(y_t | g(\theta_t^{(i)}), m^{(i)})$
- sample $\psi_t^{(i)} \sim N(.|m^{(k)}, h^2\Sigma)$
- sample $\theta_t^{(i)} \sim p(.|\theta_{t-1}^{(k)}, \psi_t^{(i)})$
- evaluate weights: $w_t \propto \frac{p(y_t|\theta_t^{(i)}, \psi_t^{(i)})}{p(y_t|g(\theta_t^{(k)}), m^{(k)})}$

4.4.4 Particle filtering and learning using sufficient statistics

The distribution of parameter in many models depends on a low dimensional set of conditionally sufficient statistics S_t such that $p(\psi|\theta_{0:t}, y_{1:t})$ is equivalent to $p(\psi|S_t)$, where S_t is easily updated recursively by $S_t = \mathcal{S}(S_{t-1}, \theta_t, y_t)$. This approach have been used in SMC methods in class of state-space models to learn sequentially the parameter vector (Storvik, 2002), (Fearnhead, 2002), (Carvalho, Johannes, Lopes, & Polson, 2010), (Polson, Stroud, & Müller, 2008)
By decomposing the joint conditional distribution

$$\begin{aligned} p(\theta_{0:t}, \psi|y_{1:t}) &\propto p(\theta_{0:t}, \psi, y_t|y_{1:t-1}) \\ &= p(y_t|\theta_t, \psi)p(\theta_t|\theta_{0:t-1}, y_{1:t-1}, \psi)p(\psi|\theta_{0:t-1}, y_{1:t-1})p(\theta_{0:t-1}|y_{1:t-1}) \\ &= p(y_t|\theta_t, \psi)p(\theta_t|\theta_{t-1}, \psi)p(\psi|S_{t-1})p(\theta_{0:t-1}|y_{1:t-1}) \end{aligned}$$

the general idea of this approach is to sample the parameter vector ψ from $p(\psi|S_{t-1})$, then θ_t form $p(\theta_t|\theta_{t-1}, \psi)$ and re-weight with weights being proportional to $p(y_t|\theta_t, \psi)$ and finally update the sufficient statistics.

In summary:
for i in $1, \ldots, N$

- Sample $\psi^{(i)}$ from $p(\psi|S_{t-1}^{(i)})$

- Draw $\theta_t^{(i)}$ form $p(\theta_t|\theta_{t-1}^{(i)}, \psi^{(i)})$
- Resample with weights $w_t^{(i)} \propto p(y_t|\theta_t^{(i)}, \psi^{(i)})$
- Update sufficient statistics $S_t^{(i)} = \mathcal{S}(S_{t-1}^{(i)}, \theta_t^{(i)}, y_t)$

Chapter 5

Model for structural breaks and Outliers

5.1 Fat-tailed t-distribtution & Mixture of Normals

The class of conditionally linear Gaussian state-space models offers a general and convenient framework for parameter learning, state filtering and detection of outliers and structural breaks(Petris et al., 2009) The state-space representations of such model is a linear dynamic mixture models, in the sense that it is linear, conditional on a vector of latent random variables and scale parameters.

To account for outliers and structural breaks, we modify the DLM by using heavy-tailed Student-t distribution (Petris et al., 2009; Shephard, 1994).The Student-t-distribution admits representation of scale mixture of Normal distribution and can also accommodate, through its degree of freedoms parameter, different degrees of heaviness in the tails.

Replace the *Gaussian* distribution of v_t and w_t in equations (3.4) and (3.5) respectively, with *Student-t* distribution with scale parameter λ^{-1} and degrees of freedom ν as shown below.

Letting $\nu_{y,t}\omega_{y,t} \sim \chi^2_{\nu_{y,t}}$ and $Z \sim \mathcal{N}(0, \lambda_y^{-2})$ and expressing v_t as

$$v_t = \frac{Z}{\sqrt{\omega_{y,t}}}$$

then v_t follows *Student-t* distribution with scale parameter λ_y^{-1} and degrees of freedom ν_y;

$$v_t|\lambda_y\nu_{y,t} \sim t_{\nu_y}(0, \lambda_y^{-1})$$

Consequently, conditional on λ_y and $\omega_{y,t}$, v_t is now *Gaussian* with $V_t = (\lambda_y\omega_{y,t})^{-1}$

$$v_t|\lambda_y\omega_{y,t} \sim \mathcal{N}(0, (\lambda_y\omega_{y,t})^{-1}) \tag{5.1}$$

Using the same argument as above, we have also that $w_{t,j}$ is conditionally *Gaussian* given $\lambda_{\theta,j}$ and $\omega_{\theta,j,t}$ with $W_{t,j} = (\lambda_{\theta,j}\omega_{\theta,j,t})^{-1}$

$$w_{t,j}|\lambda_{\theta,j}\omega_{\theta,j,t} \sim \mathcal{N}(0, (\lambda_{\theta,j}\omega_{\theta,j,t})^{-1}), \qquad j = 1, ...p. \tag{5.2}$$

From now the parameter vector will be presented as

$$\psi = \left(a_y, b_y, \pi_y, \lambda_y, (a_{\theta,j}, b_{\theta,j}, \pi_{\theta,j}, \lambda_{\theta,j}; j = 1...p)\right)$$

and state vector,

$$X_t = \left(\nu_{y,t}, \omega_{y,t}, (\nu_{\theta,j,t}, \omega_{\theta,j,t}, \theta_{j,t}; j = 1, ...p)\right)$$

In the modified V_t and $W_{t,j}$, the parameter λ represent precision of observation and state evolution respectively. The expected value of $\omega = 1$ if there is no outlier or structural break in the series. The posterior mean of the latent variable, ω can be used to flag possible outliers and /or structural breaks.

Small values of $\omega_{y,t}$ correspond to large variances V_t, making a large innovations v_t to be accounted for by the model. A small value of $\omega_{y,t}$ will signal an outlier in the series. Similarly a small value of $\omega_{\theta,j,t}$ correspond to a large variance $W_{t,j}$ (the jth diagonal element of W_t) and therefore a small value of $\omega_{\theta,j,t}$ will flag a break or a jump in the jth

component of the state vector.

In summary, a $p-$dimensional state space model can be expressed as

$$y_t = F_t\theta_t + v_t \qquad v_t|\lambda_y\omega_{y,t} \sim \mathcal{N}(0, V_t)$$
$$\theta_t = G_t\theta_{t-1} + w_t \qquad w_t|\lambda_\theta\omega_{\theta,t} \sim \mathcal{N}(0, W_t)$$

where $V_t = (\lambda_y\omega_{y,t})^{-1}$ and $W_{t,j} = (\lambda_{\theta,j}\omega_{\theta,j,t})^{-1} \quad j = 1, \ldots, p$ such that

$$W_t = \begin{pmatrix} W_{t,1} & & & \\ & W_{t,2} & & \\ & & \ddots & \\ & & & W_{t,p} \end{pmatrix}, \qquad \theta_t = \begin{pmatrix} \theta_{t,1} \\ \theta_{t,2} \\ \vdots \\ \theta_{t,p} \end{pmatrix}$$

and F_t and G_t are known transition matrices of order $m \times p$ and $p \times p$ respectively.

In this study only univariate observation data is considered and hence $m = 1$

5.2 Prior specifications

In this section we discuss and specify the prior for hierarchical structure of the observation variances $V_t = (\lambda_y\omega_{y,t})^{-1}$ and each diagonal element, $W_{t,j} = (\lambda_{\theta,i}\omega_{\theta,t,j}), \quad j = 1, \ldots, p$, of state innovation variances W_t.

The precision parameter λ follows Gamma distribution with prior mean and variance equal to a and b respectively drawn uniformly over a large interval. That is

$$a_y \sim Unif(0, A_y) \qquad a_{\theta,j} \sim Unif(0, A_{\theta,j})$$
$$b_y \sim Unif(0, B_y) \qquad b_{\theta,j} \sim Unif(0, B_{\theta,j})$$

and

$$\lambda_y|a_y, b_y \sim \mathcal{G}am\left(\tfrac{a_y^2}{b_y}, \tfrac{a_y}{b_y}\right) \qquad \lambda_{\theta,j}|a_{\theta,j}, b_{\theta,j} \sim \mathcal{G}am\left(\tfrac{a_{\theta,j}^2}{b_{\theta,j}}, \tfrac{a_{\theta,j}}{b_{\theta,j}}\right)$$

As noted previously, the latent variable ω follows a Chi-square distribution with ν degrees of freedom, which is a special Gamma distribution. Therefore the prior for ω is Gamma distribution with scale and shape parameters equal to $\nu/2$.

$$\omega_{y,t}|\nu_{y,t} \sim \mathcal{G}am\Big(\tfrac{\nu_{y,t}}{2}, \tfrac{\nu_{y,t}}{2}\Big) \qquad \omega_{\theta,j,t}|\nu_{\theta,j,t} \sim \mathcal{G}am\Big(\tfrac{\nu_{\theta,j,t}}{2}, \tfrac{\nu_{\theta,j,t}}{2}\Big)$$

The auxiliary variable ν can take any positive real value, but for simplicity, we restrict its range to a set of finite integers $\boldsymbol{n} = (n_1, n_2, \ldots, n_K)$ with corresponding probabilities $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$.

$$\nu_{y,t} \sim Multi(1; \pi_y) \qquad \nu_{\theta,j,t} \sim Multi(1; \pi_{\theta,j})$$

where the probabilities π are drawn from a Dirichlet distribution with specified parameters ξ.

$$\pi_y \sim Dir(\xi_y) \qquad \pi_{\theta,j} \sim Dir(\xi_{\theta,j}), \quad j = 1, \ldots, p$$

5.3 Parameter Estimation

5.3.1 Kernel Mixture approximation

We extend Liu and West approach, in section(4.4.3), of kernel approximation to normal distribution to other distributions of interest to approximate $p(\varphi|y_{1:t})$

$$p(\psi|y_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} f_N(\psi; \alpha^{(i)}, \beta^{(i)})$$

The idea is to replace the delta masses see Equation 4.5 with continuous distribution of interest, whose mean μ and variance σ^2 are obtained respectively through shrinkage and

smoothing modification as shown below

$$\mu(\psi^{(i)}) = \ r\psi_{t-1}^{(i)} + (1-r)\bar{\psi} = \ m(\alpha^{(i)}, \beta^{(i)}) \tag{5.3}$$

$$\sigma^{2(i)} = \ h^2\Sigma = \ \sigma^2(\alpha^{(i)}, \beta^{(i)}) \tag{5.4}$$

where $\bar{\varphi}$ is the posterior mean and Σ is the variance matrix of φ from $\varphi_{t-1}^{(i)}$ and $w_{t-1}^{(i)}$, that is

$$\bar{\psi} = \sum_{i=1}^{N} w_{t-1}^{(i)} \psi_{t-1}^{(i)}$$

$$\Sigma = \sum_{i=1}^{N} w_{t-1}^{(i)} (\psi_{t-1}^{(i)} - \bar{\psi}_{t-1})(\varphi_{t-1}^{(i)} - \bar{\psi}_{t-1})^{'}$$

r is the shrinkage parameter associated with smoothing parameter h such that $r^2 + h^2 = 1$ The parameter α and β required in this estimation are quickly obtained by solving equations 5.3 and 5.4.

5.3.2 MCMC moves

Another approach to explore parameter space is to incorporate MCMC moves that target the parameter posterior in the particle filter (Gilks & Berzuini, 2002). The idea is to build a Markov kernel $K_t(\psi, \psi^{'})$ such that

$$p(\psi^{'}|\theta_t, y_{1:t}) = \int p(\psi|\theta_t, y_{1:t}) K_t(\psi, \psi^{'}) d\psi$$

With a properly designed Markov kernel, samples from the particle filter can be 'jittered', reducing degeneracy caused by successive resampling, restore variability and hence improve quality of posterior estimates.

5.3.3 Sufficient Statistics

In this approach we use low dimension set of conditionally sufficient statistics S_t such that the distribution $p(\psi|\theta_{0:t}, y_{1:t})$ is equivalent to $p(\psi|S_t)$, where S_t is easily updated recursively by $S_t = \mathcal{S}(S_{t-1}, \theta_t, y_t)$ (Storvik, 2002), (Fearnhead, 2002)

5.3.4 Hybrid of Kernel approximation and sufficient statistics approaches

In this study a hybrid of both kernel approximation method in section 5.3.1 and sufficient statistics approach, section 5.3.3 was used. The parameter vector ψ is decomposed into two vector components ϕ and φ

$$\begin{aligned}\phi =& \big(\lambda_y, \lambda_{\theta,j}; j = 1, ...p\big) \\ \varphi =& \Big(a_y, b_y, \pi_y, (a_{\theta,j}, b_{\theta,j}, \pi_{\theta,j}; j = 1...p)\Big)\end{aligned}$$

where, conditional on φ, ϕ admits recursive conditional sufficient statistics; that is $p(\phi|X_{0:t}, y_{1:t}, \varphi) = p(\phi|S_t, \varphi)$ and S_t is sufficient statistics which can be updated recursively by

$$S_t = \mathcal{S}\big(S_{t-1}, X_t, y_t\big)$$

and the prior for ψ is given as follows

$$\begin{aligned}p(\psi) &= p(\phi, \varphi) = p(\phi|X_0, \varphi)p(\varphi) \\ &= p(\phi|S_0, \varphi)p(\varphi)\end{aligned}$$

The distribution $p(\varphi|y_{1:t})$ is approximated by Kernel mixture approximation method. This parameter decomposition decreases the number of parameters in the vector φ, thereby reducing the Monte Carlo error in the kernel approximation.

We now give the specific solutions of α and β required for the kernel approximation to $p(\varphi|y_{1:t})$ for all the elements in parameter vector φ

- *Parameters a and b*

The unknown parameters a and b are positive and sampled over large interval we rescale them to interval (0, 1) and the respective $\bar{\varphi}$ and Σ are computed thereafter. The idea of rescaling is so that we can use standard distribution, beta say, which we are sure will always give random variates in the interval (0, 1) and whose mean and variance are well defined. By definition, for a beta distribution with mean m and variance σ^2, and parameters α and β,

$$m^{(i)} = \frac{\alpha^{(i)}}{\alpha^{(i)} + \beta^{(i)}}$$
$$\sigma^{2(i)} = \frac{\alpha^{(i)}\beta^{(i)}}{(\alpha^{(i)} + \beta^{(i)})^2(\alpha^{(i)} + \beta^{(i)} + 1)}$$

Using the equations above and basic algebra we have;

$$\alpha^{(i)} = \frac{(1 - m^{(i)})(m^{(i)})^2}{\sigma^{2(i)}} - m^{(i)}$$
$$\beta^{(i)} = \frac{(1 - m^{(i)})^2 m^{(i)}}{\sigma^{2(i)}} - (1 - m^{(i)})$$

The mixture obtained

$$\sum_{i=1}^{N} w_{t-1}\boldsymbol{\mathcal{B}}(\psi; \alpha^{(i)}, \beta^{(i)})$$

has mean $\bar{\varphi}$ and variance Σ. Before they are used for inference, the values of a and b obtained above must be scaled back to their original intervals.

- *Parameter π*

Since π is a probability vector whose elements sum to 1 and $\pi_k > 0$, we use Dirichlet

distribution

$$p(P = \pi_k|\alpha_k) \sim \mathcal{D}(\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$, with $\alpha_k > 0$ for each k and whose mean

$$E(\pi_k) = \frac{\alpha_k}{\alpha_0} \tag{5.5}$$

and variance

$$Var(\pi_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \tag{5.6}$$

We need to estimate the parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ using equations (5.5) and (5.6) above, and the posterior mean $\bar{\psi}_\pi$ and variance matrix Σ_π of ψ_π from $\psi_{t-1}^{(i)}$ and $w_{t-1}^{(i)}$. With $m_k^{(i)}$ and $\sigma_k^{2(i)}$ computed as described previously, we then have

$$m_k^{(i)} = \frac{\alpha_k^{(i)}}{\alpha_0^{(i)}} \tag{5.7}$$

$$\sigma_k^{2(i)} = \frac{\alpha_k^{(i)}(\alpha_0^{(i)} - \alpha_k^{(i)})}{\alpha_0^{2(i)}(\alpha_0^{(i)} + 1)} = \frac{m_k^{(i)}(1 - m_k^{(i)})}{(\alpha_0^{(i)} + 1)} \tag{5.8}$$

It follows quickly from equations (5.7) and (5.8) that

$$\alpha_k^{(i)} = m_k^{(i)} \alpha_0^{(i)}$$

where

$$\alpha_0^{(i)} = \frac{m_k^{(i)}(1 - m_k^{(i)})}{\sigma_k^{2(i)}}$$

5.4 State Estimation

The target density of interest $p(X_t|X_{0:t-1}, \psi, y_{1:t})$ can be decomposed

$$p(y_t|\theta_t, V_t)p(\theta_t|\theta_{t-1}, \omega_t, \psi) \times p(\omega_t|\nu_t, \psi) \times p(\nu_t|\psi)$$

In particular to estimate the p-dimensional states θ_t, we have the data $\mathbf{y}_t = (y_1, y_2, \ldots, y_T)$;

$$y_t \sim \mathcal{N}(F\theta_t, V_t), \quad \text{with} \;\; V_t = (\lambda_y \omega_{y,t})^{-1}$$

which is a multivariate normal distribution, and the likelihood function $p(y_t|F, \theta_t, V)$

$$p(\mathbf{y}_t|F, \theta_t, V_t) = \frac{1}{(2\pi)^{\frac{T}{2}}|V|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{y_t} - F\theta_t)^\intercal V_t^{-1}(\mathbf{y_t} - F\theta_t)\right\}$$

where $|V|$ is the determinant of covariance matrix V_t. The states

$$\theta_t \sim \mathcal{N}(G\theta_{t-1}, W_t), \text{with} \;\; W_t = \text{diagonal}((\lambda_1 \omega_{\theta,1,t})^{-1}, \ldots, (\lambda_p \omega_{\theta,p,t})^{-1})$$

$$\begin{aligned} p(\theta_t|G, \theta_{t-1}, W_t) &= \frac{1}{(2\pi)^{\frac{T}{2}}|W|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\theta_t - G\theta_{t-1})^\intercal W_t^{-1}(\theta_t - G\theta_{t-1})\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\theta_t - G\theta_{t-1})^\intercal W_t^{-1}(\theta_t - G\theta_{t-1})\right\} \end{aligned}$$

The posterior distribution of the states θ_t can be determined by using Bayes theorem

$$\begin{aligned} p(\theta_t|.) &\propto \exp\left\{-\frac{1}{2}(\mathbf{y} - F\theta_t)^\intercal V_t^{-1}(\mathbf{y} - F\theta_t)\right\} \times \\ &\quad \exp\left\{-\frac{1}{2}(\theta_t - G\theta_{t-1})^\intercal W_t^{-1}(\theta_t - G\theta_{t-1})\right\} \\ &\sim \mathcal{N}(m, C) \end{aligned}$$

where

$$C = (W_t^{-1} + F^{\mathsf{T}} V_t^{-1} F_t)^{-1}$$
$$m = C(F^{\mathsf{T}} V_t^{-1} y_t + W_t^{-1} G_t \theta_{t-1})$$

Clearly the posterior estimates of θ_t depends on $\omega_{y,t}$ and $\omega_{\theta,i,t}$, through V_t and W_t respectively.

On the other hand the posterior estimates of $\omega_{y,t}$ and $\omega_{\theta,i,t}$ depends on, among others, θ_t and ν_t. For instance, the posterior estimate for $\omega_{y,t}$ can be quickly derived,

$$\begin{aligned}
p(\omega_{y,t}|y_{1:t}, \theta_{0:t}, \psi) &\propto p(y_{1:t}|\theta_{0:t}, \psi) p(\omega_{y,t}|\nu_t, \psi) \\
&\propto \prod_{t=1}^{T} \omega_{y,t} \exp\Big\{ -\frac{\omega_{y,t}\lambda_y}{2}(y_t - F_t\theta_t)^2 \Big\} \times \omega_{y,t}^{\frac{\nu_{y,t}}{2}-1} \exp\Big\{ \omega_{y,t}\frac{\nu_{y,t}}{2} \Big\} \\
&\propto \omega^{\frac{\nu_{y,t}+1}{2}-1} \exp\Big\{ -\frac{\omega_{y,t}}{2}[\lambda_y(y_t - F_t\theta_t)^2 + \nu_{y,t}] \Big\} \\
&\sim \mathcal{G}am\left(\frac{\nu_{y,t}+1}{2}, \frac{\lambda_y(y_t - F_t\theta_t)^2 + \nu_{y,t}}{2} \right)
\end{aligned}$$

Due to this dependence structure on these states in our model, we propose the following importance sampling strategies to sample them.

5.4.1 Sequential Bridge Sampling

We propose the use, in our algorithm, of a nested sequential bridge sampling by utilizing a SMC sampling approach which allows the partition of function of a non-analytically-normalizable distribution $p(X)$ to be estimated in an unbiased fashion through a chain of Markov transitions.

For each particle i for $i = 1, 2 \ldots, N$ and at each time t we extend the state space $X_t^{(i)}$ to a sequence of distributions $p_k(X_t^{(i)})$ $\;0 \leq k \leq d$, with $p_d(X_t^{(i)}) = p(X_t^{(i)})$.

The sequence $\{p_k(X_t^{(i)})\}_{k \geq 0}$ forms a bridge of successive approximations from initial

density $p_0(X_t^{(i)})$, which is diffuse and easier to sample from, to $p_d(X_t^{(i)}) = p(X_t^{(i)})$

To obtain efficient IS for target p_{k+1}, it is expected that, p_k differs only slightly from p_{k+1}. This is achieved by expressing

$$\begin{aligned} p_k(X_t^{(i)}|\psi^{(i)}) &= p_0(X_t^{(i)}|\psi^{(i)})^{1-b_k} p(X_t^{(i)}|\psi^{(i)})^{b_k} \\ &\propto p_0(X_t^{(i)}|\psi^{(i)}) p(y_t|X_t^{(i)}, \psi^{(i)})^{b_k} \end{aligned}$$

for $0 = b_0 < b_1 < \ldots < b_D = 1$

A random draw of $X_t^{(i)}$ is made by sequentially drawing $X_{t,(k)}^{(i)}|\psi^{(i)}, \quad 0 \leq d \leq D$, and equating $X_t^{(i)} = X_{t,(D)}^{(i)}|\psi^{(i)}$

In summary, for each time t and for every particle $i, i = 1, 2, \ldots, N$

- Draw $X_{t,0}^{(i)}$ from $p_0(X_t^{(i)}|\psi^{(i)})$
- Set $w_{t,0} = 1$
- Then for $k = 1, 2, \ldots, D$
    - Resample particle if ESS $<$ R, a pre-determined threshold value, and set $w_{t,k-1} = 1/N$
    - Sample $X_{t,k}^{(i)}$ from $q(.|X_{t,k-1}, \psi^{(i)}, y_t)$
    - Compute weights $w_{t,k} \propto w_{t,k-1} \dfrac{p_k(X_{t,k}^{(i)}|\psi_t^{(i)}) L_{k-1}(X_k^{(i)}, X_{k-1}^{(i)})}{p_{k-1}(X_{t,k-1}^{(i)}|\psi_t^{(i)}) q_k(X_{k-1}^{(i)}, X_k^{(i)})}$

      where $L()$ is an arbitrary backward Markov Kernel. If we choose

      $$L_{k-1}(X_k, X_{k-1}) = \frac{p_k(X_{t,k-1}^{(i)}|\psi_t^{(i)}) q_k(X_{k-1}, X_k)}{p_k(X_k^{(i)}|\psi_t^{(i)})}$$

      then the weights above will be

      $$w_{t,k} \propto w_{t,k-1} \frac{p_k(X_{t,k-1}^{(i)}|\psi_t^{(i)})}{p_{k-1}(X_{t,k-1}^{(i)}|\psi_t^{(i)})}$$

5.4.2 Scoring the data

We also propose an approach where the auxiliary states $\nu_t^{(i)}$ are selected uniformly from a set of finite integers according to indicator variable $Z_t^{(i)}$

$$Z_t^{(i)} = \frac{|y_t - F_t G_t \theta_{t-1}^{(i)}|}{\sqrt{\lambda_y^{-1(i)}}}$$

This stem from the well known concept of regular Z-scores where extreme values in a normal distribution are associated with high Z-scores in absolute sense. Large values of $Z_t^{(i)}$ are assigned to the small values in the set and vice-versa.
Once $\nu_t^{(i)}$ is obtained, we draw $\omega_t^{(i)} \sim p(\omega_t|\nu_t^{(i)}, \psi^{(i)})$ and consequently
$\theta_t^{(i)} \sim p(\theta_t|y_t, \theta_{t-1}^{(i)}, \omega_t^{(i)}, \nu_t^{(i)}, \psi^{(i)})$

Fixed interval smoothing

The idea is, with all the data up to the current time t, we go back in time for some fixed steps, ℓ say, and infer the state vector $\{X_{t-\ell:t}\}$. In this case the smoothing distribution of interest is defined by $p(X_{t-\ell:t}|y_{t-\ell:t}, \psi)$ which, we can approximate by using the following decomposition

$$p(X_{t-\ell:t}|y_{t-\ell:t}, \psi) = p(X_t|y_{t-\ell:t}, \psi) \prod_{s=t-\ell}^{t-1} p(X_s|X_{s+1:t}, y_{t-\ell}, \psi)$$

The approximation to $p(X_t|y_{t-\ell:t}, \psi)$ is given via particle filtering and following the approach in (Godsill, Doucet, & West, 2004) we have

$$p(X_s|X_{s+1:t}, y_{t-\ell}, \psi) \propto p(X_s|y_{t-\ell:s}) p(X_{s+1}|X_t)$$

and the smoothing algorithm can be summarized as follow

- with probability $w_t^{(i)}$ obtain $\hat{X}_t = X_t^{(i)}$

- for $s = t-1, t-2, \ldots, t-\ell$:
  1. compute weights $w_{s|s+1} \propto w_s^{(i)} p(\hat{X}_{s+1}|X_s^{(i)}, \psi^{(i)})$
  2. with probability $w_{s|s+1}^{(i)}$ select $\hat{X}_s = X_s^{(i)}$
- set $\hat{\boldsymbol{X}}_t = (\hat{X}_{t-\ell}, \hat{X}_{t-\ell+1}, \ldots, \hat{X}_t)$

$\hat{\boldsymbol{X}}_t$ is the required approximation to $p(X_{t-\ell:t}|y_{t-\ell:t}, \psi)$

5.5 Algorithm Summary

Our focus at each time t is to approximate the target distribution $p(X_{0:t}, \psi|y_{1:t})$. We can decompose this joint posterior distribution as follows

$$
\begin{aligned}
p(X_{0:t}, \psi|y_{1:t}) &= p(X_{0:t}, \psi|y_{1:t}) \\
&\propto p(X_{0:t}, \phi, \varphi, y_t|y_{1:t-1}) \\
&= p(y_t|X_{0:t}, y_{1:t-1}, \phi, \varphi) p(X_t|X_{0:t-1}, y_{1:t-1}, \phi, \varphi) p(\phi|X_{0:t-1}, \varphi, y_{1:t-1}) p(X_{0:t-1}, \varphi|y_{1:t-1}) \\
&= p(y_t|X_t, \phi, \varphi) p(X_t|X_{t-1}, \phi, \varphi) p(\phi|S_{t-1}, \varphi) p(X_{0:t-1}, \varphi|y_{1:t-1}) \\
&\approx p(y_t|X_t, \phi, \varphi) p(X_t|X_{t-1}, \phi, \varphi) p(\phi|S_{t-1}, \varphi) \hat{p}(X_{0:t-1}, \varphi) \\
&= \sum w_{t-1}^{(i)} p(y_t|X_t, \phi, \varphi) p(X_t|X_{t-1}^{(i)}, \phi, \varphi) p(\phi|S_{t-1}, \varphi) f_N(\varphi; m^{(i)}, h^2\Sigma) \delta_{X_{0:t-1}^{(i)}}
\end{aligned}
$$

for $i = 1, 2, \ldots, N$.

This decomposition will allow us to use a hybrid approach for parameter estimation discussed in section 5.3.4

Our SMC algorithm is summarized as follows:

- We have particles $\left\{(X_{0:t-1}, \varphi_{t-1}, \phi, S_{t-1}, w_{t-1})^{(i)}\right\}_{i=1}^N$ approximating $p(X_{0:t-1}, \varphi, \phi|y_{1:t-1})$ at time $t-1$
- Compute $\tilde{w}_t^{(i)} \propto w_{t-1}^{(i)} q\Big(y_t|g(X_{t-1}^{(i)}), \mu(\varphi_{t-1}^{(i)}), \phi^{(i)}\Big)$ where $g(X_{t-1}^{(i)}) = E(X_t|X_{t-1}^{(i)}, \mu(\varphi_{t-1}^{(i)}), \phi^{(i)})$

- Resample $\{(X_{0:t-1}, \varphi_{t-1}, \phi, S_{t-1})^{(i)}\}_{i=1}^{N}$ with the weights equal to $\tilde{w}_t^{(i)}$ to get new set $\{(\tilde{X}_{0:t-1}, \tilde{\varphi}_{t-1}, \tilde{\phi}, \tilde{S}_{t-1})^{(i)}\}_{i=1}^{N}$

- Draw the parameter vector $\tilde{\varphi}_t^{(i)} \sim q(\mu(\tilde{\varphi}_{t-1}^{(i)}), h^2\Sigma)$ according to Kernel mixture approximation detailed in section 5.3.4

- Draw states $\tilde{X}_t^{(i)}$ according to one of the approaches discussed in section 5.4
  $\tilde{X}_t^{(i)} \sim q(X_t|\tilde{X}_{t-1}^{(i)}, \tilde{\varphi}_t^{(i)}, \tilde{\phi}^{(i)}, \tilde{S}_{t-1}^{(i)}, y_t)$

  and set $\tilde{X}_{0:t}^{(i)} = (\tilde{X}_{0:t-1}^{(i)}, \tilde{X}_t^{(i)})$

  - For example, using the indicator variable $Z_t^{(i)}$ approach

  (a) Draw $\tilde{\nu}_t^{(i)}$ from a set of finite integers according to $Z_t^{(i)}$

  (b) Draw $\tilde{\omega}_t$ given $\tilde{\nu}_t$

  $$\tilde{\omega}_{y,t}^{(i)}|\tilde{\nu}_{y,t}^{(i)} \sim \mathcal{G}am(\tfrac{\tilde{\nu}_{y,t}^{(i)}}{2}, \tfrac{\tilde{\nu}_{y,t}^{(i)}}{2})$$

  and similarly draw the two auxiliary, $\tilde{\nu}_{\theta_j,t}$ and $\tilde{\omega}_{\theta_j,t}$ states for the state $\theta_j; \quad j = 1, 2, \ldots p$

  (c) Now with $V_t^{(i)} = (\tilde{\omega}_{y,t}^{(i)}\tilde{\lambda}_{y,t-1}^{(i)})^{-1}$ and $W_{t,j}^{(i)} = (\tilde{\omega}_{\theta_j,t}^{(i)}\tilde{\lambda}_{\theta_j,t-1}^{(i)})^{-1}$ we draw the state

  $$\tilde{\theta}_t^{(i)} \sim \mathcal{N}(\theta_t|y_t, \tilde{\theta}_{t-1}^{(i)}, V_t^{(i)}, W_t^{(i)})$$

- Compute the weights

  $$w_t^{(i)} \propto \frac{p(\tilde{X}_{0:t}, \tilde{\varphi}_t^{(i)}, \tilde{\phi}^{(i)}|y_{1:t})}{q(\tilde{X}_{0:t}, \tilde{\varphi}_t^{(i)}, \tilde{\phi}^{(i)}|y_{1:t})}$$

  - Again as an example, having drawn the states $\tilde{X}_{0:t}^{(i)}$ using approach outlined above, the weights can be obtained explicitly as follows

$$w_t^{(i)} \propto \frac{p(y_t|\tilde{X}_t^{(i)}, \tilde{\varphi}_t^{(i)}, \tilde{\phi}^{(i)})}{p(y_t|\mu(\tilde{\varphi}_{t-1}^{(i)}), g(\tilde{X}_{t-1}^{(i)}), \tilde{\phi}^{(i)})} \times \frac{\sum_{k=1}^{K} \mathcal{G}am\left(\tilde{\omega}_{y,t}^{(i)}; \frac{\nu}{2}, \frac{\nu}{2}\right)\pi_y(\nu)}{\mathcal{G}am\left(\tilde{\omega}_{y,t}^{(i)}; \frac{\tilde{\nu}_{y,t}^{(i)}}{2}, \frac{\tilde{\nu}_{y,t}^{(i)}}{2}\right)} \times \frac{\prod_{j=1}^{p}\left(\sum_{k=1}^{K} \mathcal{G}am\left(\tilde{\omega}_{\theta,j,t}^{(i)}; \frac{\nu}{2}, \frac{\nu}{2}\right)\pi_{\theta,j}(\nu)\right)}{\prod_{j=1}^{p}\left(\mathcal{G}am\left(\tilde{\omega}_{\theta,j,t}^{(i)}; \frac{\tilde{\nu}_{\theta,t,j}^{(i)}}{2}, \frac{\tilde{\nu}_{\theta,t,j}^{(i)}}{2}\right)\right)}$$

- To account for the discrepancies between the target distribution $p(.)$ and proposal distribution $q(.)$ we re-weight the particles to obtain the posterior draw. That is: Draw the set $\{(X_t, \varphi_t, S_{t-1}^{(i)})^{(i)}\}_{i=1}^{N}$ from $\{(\tilde{X}_t, \tilde{\varphi}_t, \tilde{S}_{t-1})^{(i)}\}_{i=1}^{N}$ with weights $w_t^{(i)}$
- With the posterior draw we update sufficient statistics $S_t^{(i)} = \mathcal{S}\big(S_{t-1}^{(i)}, \varphi_t^{(i)}, X_t^{(i)}, y_t\big)$
- Sample the parameter vector ϕ

$$\phi_t^{(i)} \sim p(\phi|S_t^{(i)}, \varphi_t^{(i)})$$

- Smoothing, if needed

  For a fixed interval ℓ, we obtain an approximation to $p(X_{t-l:t}|y_{t-l:t})$ based on factorization

$$p(X_{t-l:t}|y_{t-l:t}) = p(X_t|y_{1:t}) \prod_{s=t-l}^{t-1} p(X_s|X_{s+1:t}, y_{t-l:t})$$

Chapter 6

Application and Results

6.1 Nile River problem

The annual readings of volume discharge from the Nile River at Aswan, Egypt from 1871 to 1970 are given in (Cobb, 1978). Many time series studies have referred to this data.(See time series for the data in Figure 6.1). Cobb used the data to demonstrate conditional inference about change point and found that there was a permanent decline in volume in 1899, the year the first Aswan dam was completed. In his study though, he couldn't identify conspicuous outlier in the year 1913 nor the mild ones in year 1888 and 1964 .

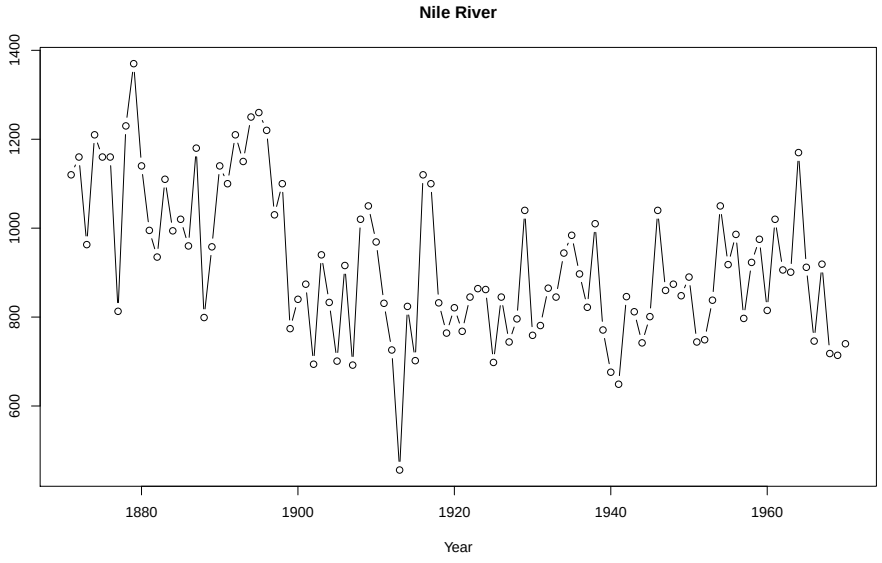

Figure 6.1: Annual volume of the Nile River from 1871 to 1970

We fitted the data the model for structural breaks and outliers, summarized in page 50, with one dimension state vector i.e. $p = 1$ and hence $F = G = 1$. The model is commonly known as the local level model

$$y_t = \theta_t + v_t \qquad v_t|\lambda_y\omega_{y,t} \sim \mathcal{N}(0, V_t)$$
$$\theta_t = \theta_{t-1} + w_t \qquad w_t|\lambda_\theta\omega_{\theta,t} \sim \mathcal{N}(0, W_t)$$

where $V_t = (\lambda_y\omega_{y,t})^{-1}$ and $W_t = (\lambda_\theta\omega_{\theta,t})^{-1}$.

We set the number of particles $N = 20000$ and fixed smoothing interval ℓ to be 5. The results using both MCMC, off-line analysis, and the SMC approach are compared. Recall, from our discussion in section 5.1, that the distribution of $\omega_{y,t}$ and $\omega_{\theta,t}$ will be used to identify, if any, outlying observations and structural breaks respectively.

The posterior estimates for $\omega_{y,t}$ and $\omega_{\theta,t}$ from the MCMC output are displayed in Figure 6.2. It is quite clear, from the plot of $\omega_{y,t}$, that the outlying observation in the year 1913 has been captured. In the same figure, plot of the estimated values of $\omega_{\theta,t}$ indicate the change in level of River Nile in the year 1899 with the value of ω_θ at that time being very close to zero, while all the others are at or very close to 1.

Results from the SMC approach shows comparable output. The plot of the Nile River data, filtered and smoothed values of the states θ_t from 1891 to 1970 are shown in Figure 6.3. The change in level of the river from the year 1899 is very clear from plots of both smoothed and filtered values of θ_t. The posterior estimates of the $\omega_{y,t}$ and $\omega_{\theta,t}$ are shown in Figure 6.4. From these plots, we can see that the extreme outlier in the year 1913 and mild ones in the year 1888 and 1964 have been captured by both the filtered and smoothed values of the ω_y. The structural break in the year 1899 has been identified by the very small smoothed value of ω_θ at that time.

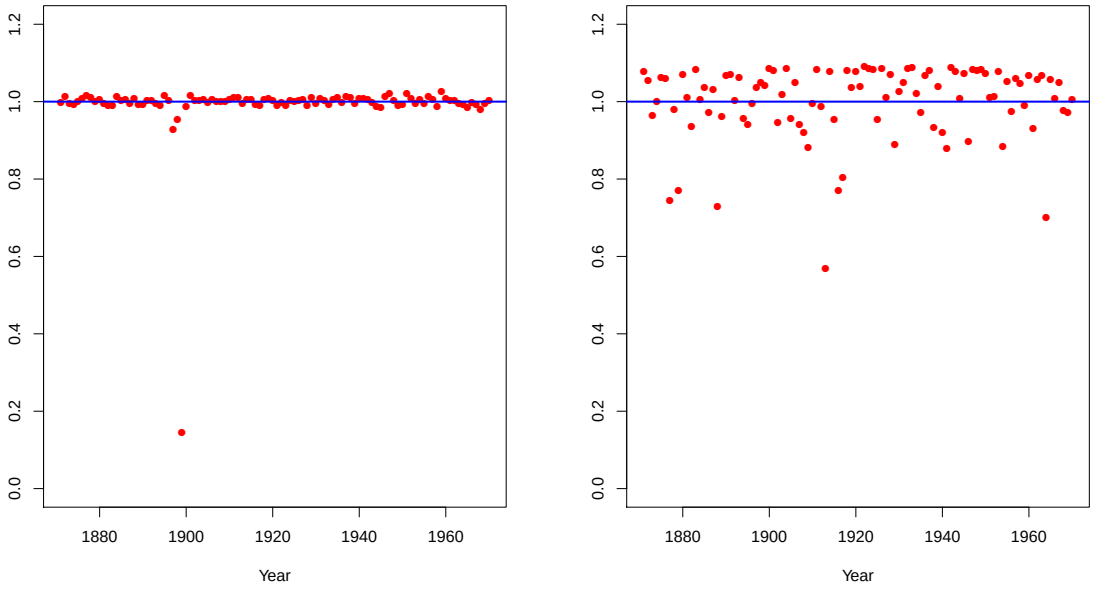

Figure 6.2: Estimated ω_θ (left) and ω_y (right) using MCMC approach.
Small value of $\omega_{\theta,1899}$ signal the break and small values of ω_y in 1888 and 1964 signals outlying observation at the time

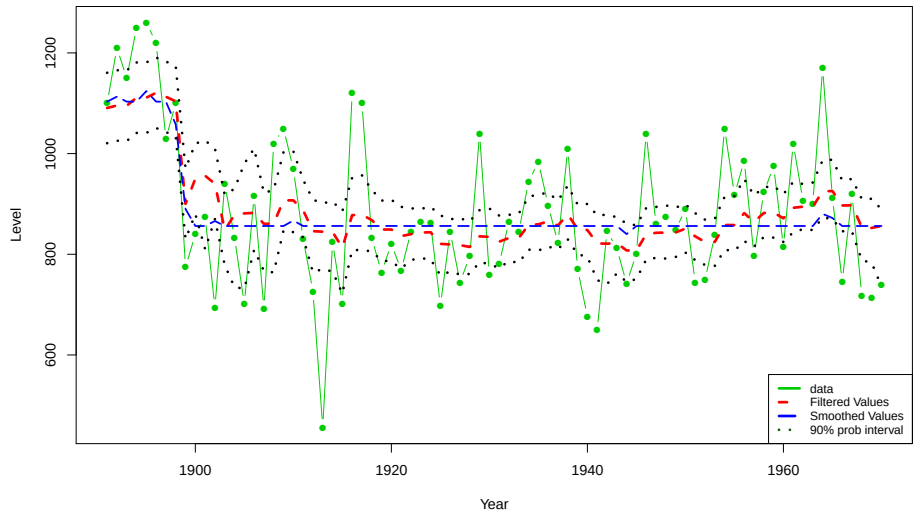

Figure 6.3: Plot of filtered and smoothed values from Nile River data using SMC algorithm

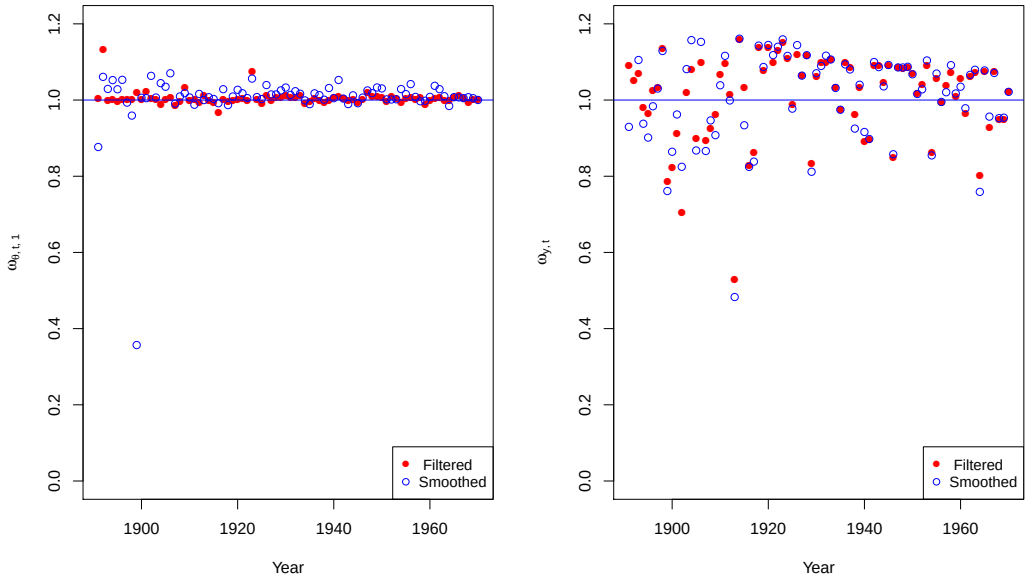

Figure 6.4: Posterior estimates of ω_θ (left) and ω_y (right) from Nile River data using SMC algorithm

6.2 Simulated data

6.2.1 Local level model

We use simulated data with 200 data points from a $\mathcal{N}(20, 4)$ which we manipulate in such a way that there is a potential outlier and structural break. In particular the data is shifted upwards from time $t = 155$ and a outlying observation created at time $t = 100$.

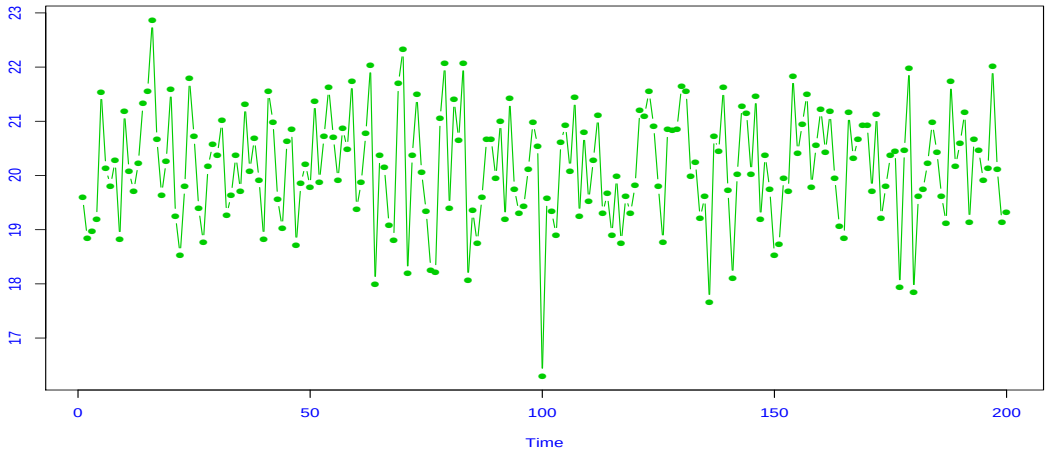

Figure 6.5: Simulated time series with a potential outlier and structural break

We apply the SMC algorithm to the last 180 data points this data using the local level model, where $p = 1$ hence $F = G = 1$, and the number of particles N set to 20000. The first 20 data points were used in MCMC to generate the prior estimates for the particle filter. The data used in this inference is plotted in Figure 6.5. The plot of the filtered values of the states and smoothed values, see Figure 6.6, clearly indicate there is a huge jump or structural break in the series.

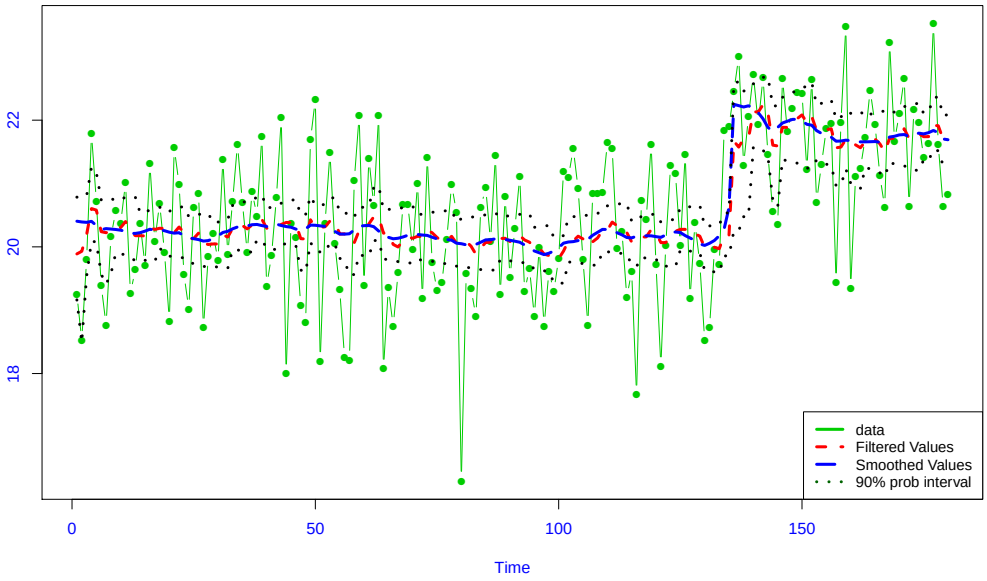

Figure 6.6: Plot of simulated data, filtered and smoothed values of the sates, θ, obtained via SMC approach

To verify this observation, we plot the posterior estimates for the auxiliary variables ω_y and ω_θ. See Figure 6.19. Obviously, from the plots we can see that the extreme value at time $t = 80$ has been identified as an outlier. The plot of posterior estimates of ω_θ, clearly shows that the structural break at time $t = 135$ has been captured.

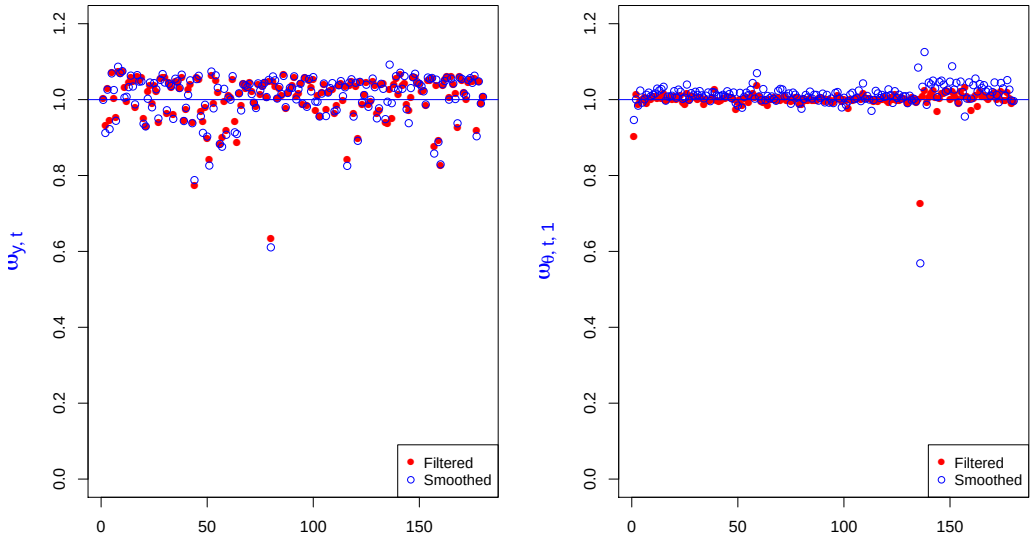

Figure 6.7: Posterior estimates of ω's from a simulated time series with a potential outlier and structural break obtained using SMC approach

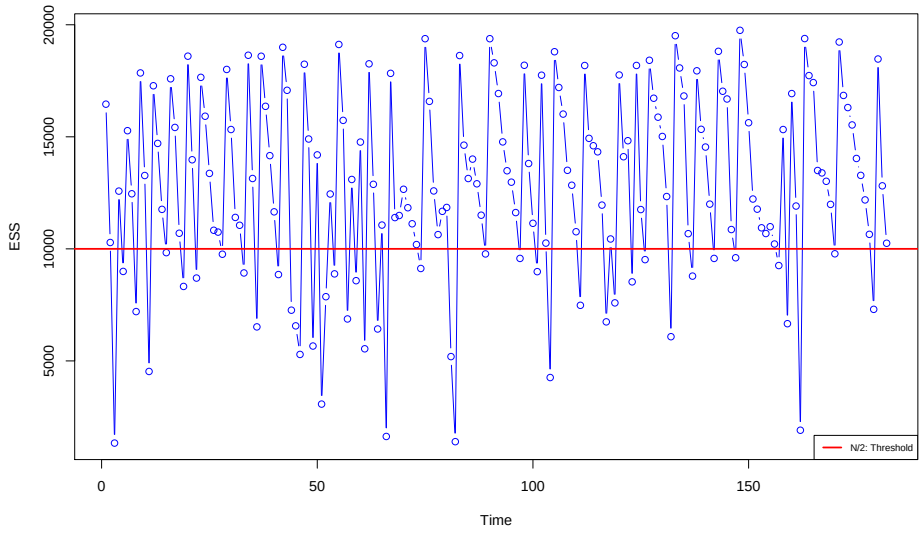

Figure 6.8: Monitoring the Effective Sample Size

We diagnosed the efficiency of our particle filter by monitoring the ESS. The output is shown in Figure 6.8

To compare our results we employ MCMC approach on the same data set and the same model. The number of Monte Carlo samples were set to 20500, the first 500 taken as burn-in. The MCMC ran with all the 200 data points available, took approximately 40 minutes to complete. Our SMC approach took less than 12 minutes to produce the results presented.

To determine computational cost for online inference, a sequential MCMC was run. That is, starting with the first observation we run an MCMC, and thereafter a new MCMC was run each time the next observation was included in the data until all 200 data points were included. Despite taking more than 39 hours to run the entire data set, the results, which are shown in Figure 6.9 & 6.10, are quite comparable to our SMC approach.

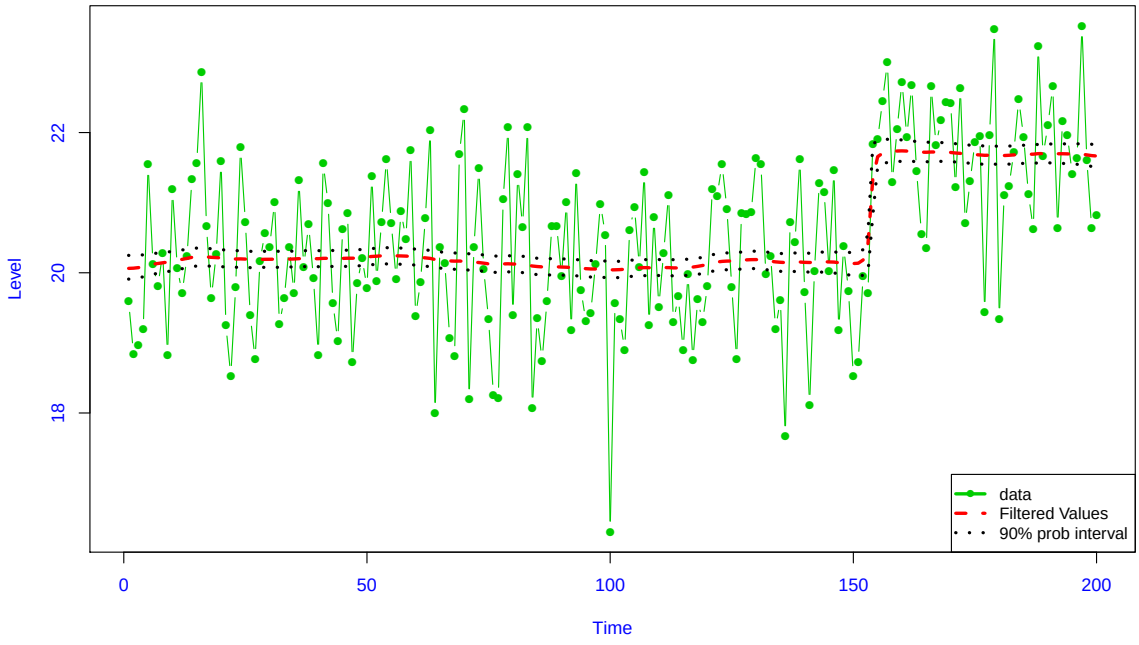

Figure 6.9: Plot of simulated data, filtered and smoothed values of the sates, θ, obtained using sequential MCMC

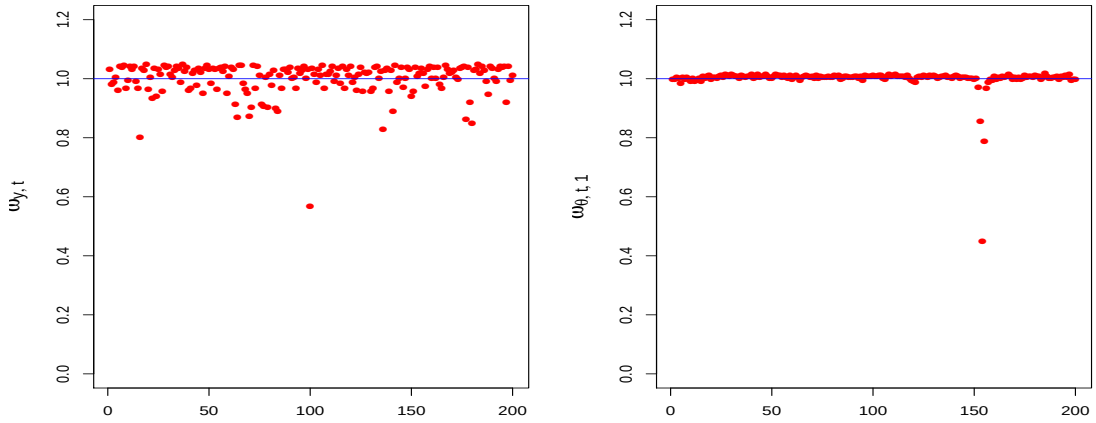

Figure 6.10: Posterior estimates of ω's from a simulated time series with a potential outlier and structural break obtained using sequential MCMC

Again we compare our approach to other regular outlier detection techniques. The plots from Shewhart, CUSUM and EWMA charts are shown in Figure 6.11. From these plots we can see that Shewhart is able to capture the outlier at $t = 80$ but not the break at $t = 135$. The CUSUM on the other hand is able to detect the jump in the series but after 7 time steps. However it is not able to capture the outlier at time $t = 80$. The EWMA is more sensitive and it is able to capture the break, much faster than CUSUM, after the third time step. The results prove that these techniques are not effective for online detection of outliers and structural break.

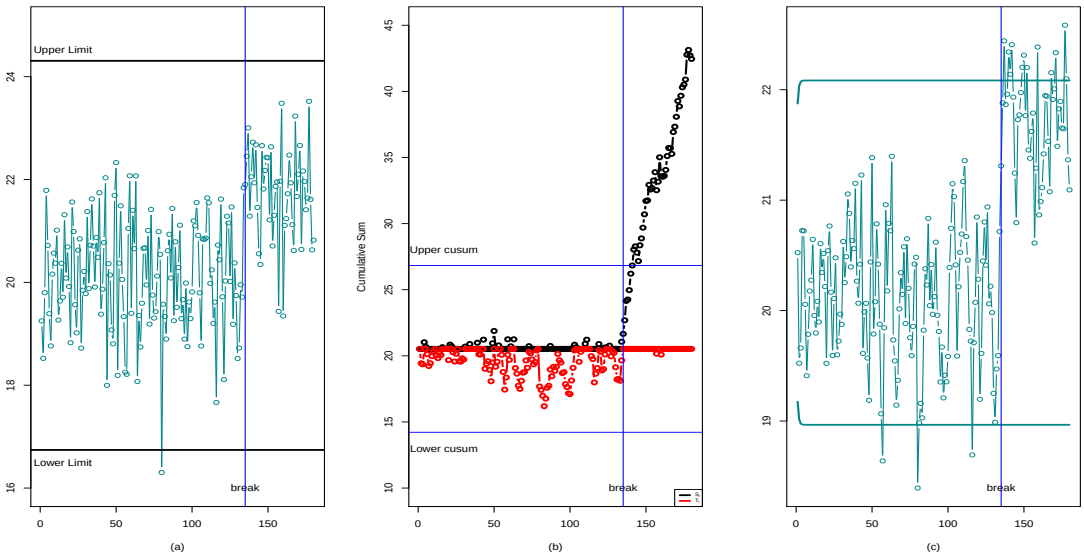

Figure 6.11: Detection of breaks and outliers from simulated data using (a)Shewhart,(b) CUSUM and (c) EWMA Chart

6.2.2 Linear trend model

For a high order polynomial, we use a data with linear trend simulated from ARIMA(1,1,1) which was manipulated to include a break. The data which consist of 40 observations and break at time $t = 26$ is shown Figure 6.12

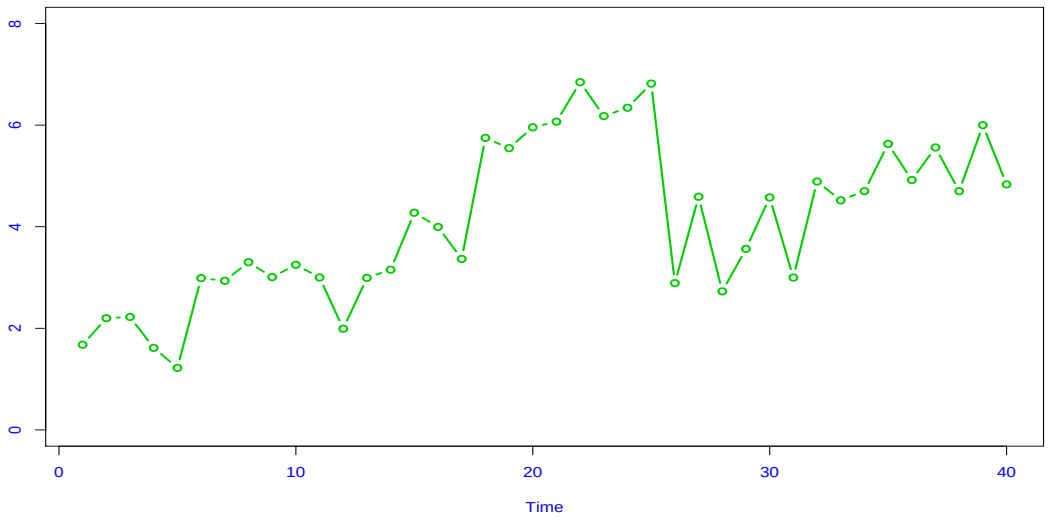

Figure 6.12: Simulated data with linear trend and possible structural break

To account for both the upward and downward shift of the trend and for the trend's change in slope, we employed a second order model, $p = 2$, detailed on page 33

$$y_t = F_t\theta_t + v_t \qquad v_t|\lambda_y\omega_{y,t} \sim \mathcal{N}(0, V_t)$$

$$\theta_t = G_t\theta_{t-1} + w_t \qquad w_t|\lambda_\theta\omega_{\theta,t} \sim \mathcal{N}(0, W_t)$$

where $V_t = (\lambda_y\omega_{y,t})^{-1}$ and $W_{t,j} = (\lambda_{\theta,j}\omega_{\theta,j,t})^{-1} \quad j = 1, 2.$

$$W_t = \begin{bmatrix} W_{t,1} & 0 \\ 0 & W_{t,2} \end{bmatrix}, \quad \theta_t = \begin{bmatrix} \theta_{t,1} \\ \theta_{t,2} \end{bmatrix}, G = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, F = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

The state elements $\theta_{t,1}$ and $\theta_{t,2}$ correspond to intercept and slope components, respectively, of the series. The $\omega_{\theta,t,1}$ in the variance $W_{t,1}$ will flag, if any, structural break in the intercept component while the $\omega_{\theta,t,2}$ in the variance $W_{t,2}$ will flag, if any, structural break in the slope component of the state vector. As usual, $\omega_{y,t}$ will identify any outlier in the series.

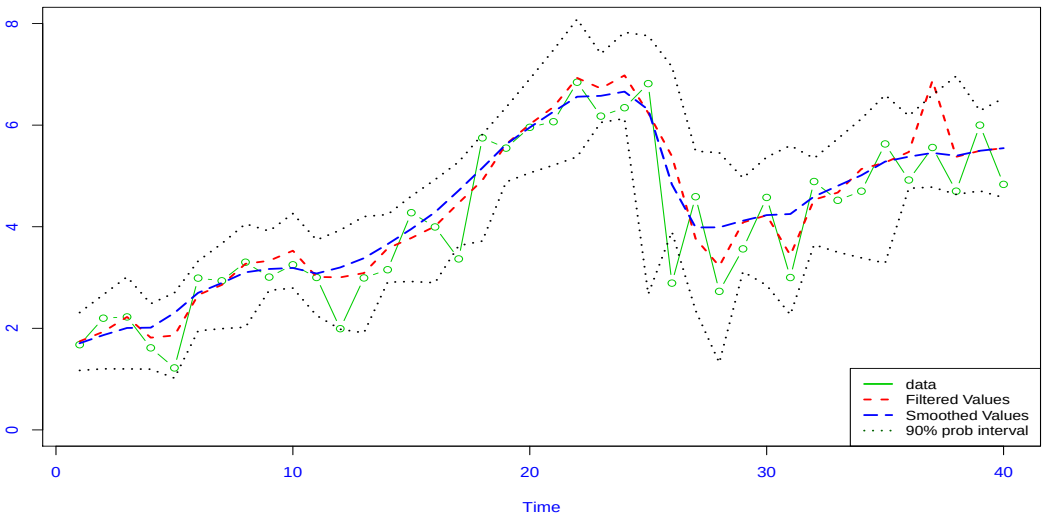

Figure 6.13: Plot of simulated data, filtered and smoothed values

The plot in Figure 6.13 shows that the slope if fairly stable and that there is a downward shift in the trend at time $t = 26$. We expect the $\omega_{\theta,t,1}$ to capture this break.

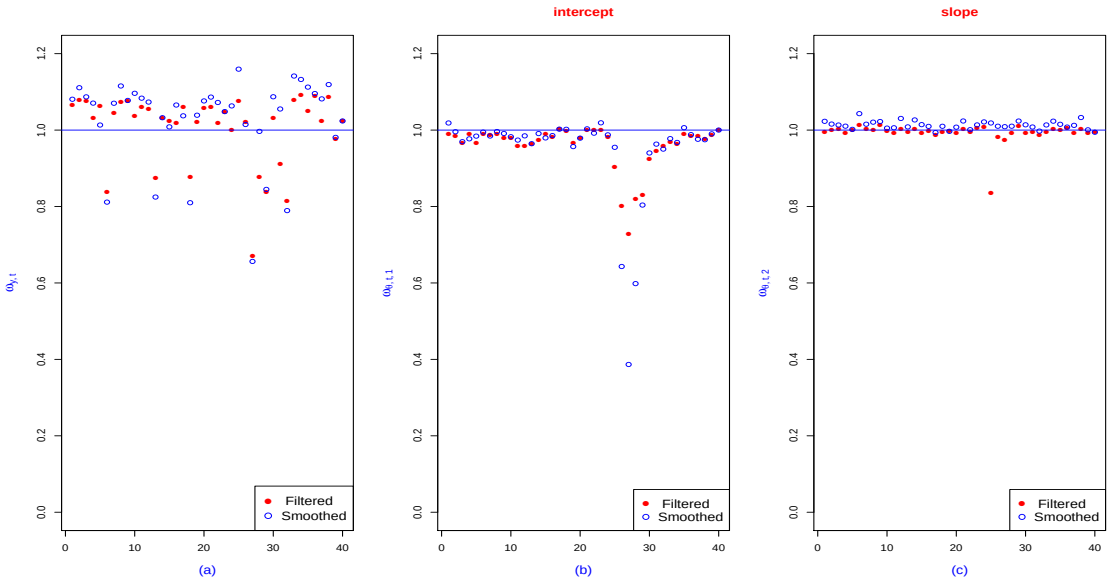

Figure 6.14: Posterior estimates of (a)$\omega_{y,t}$, (b) $\omega_{\theta,t,1}$ and (c) $\omega_{\theta,t,2}$

From Figure 6.14 we can see that the both the smoothed and filtered values of $\omega_{\theta,t,1}$ confirms a break in the intercept component of the state vector. The smoothed values of

the slope component are quite stable, as expected. Also the smoothed value at $t = 26$ confirms a false alert by the filtered value.

6.3 An outlier or structural break?

Assuming that, through sequential data update, we obtain a data value that is extreme. It would be very important for us to know if this data point is actually an outlier or a structural break has really occurred. Would the model distinguish between the presence of an outlier and an occurrence of structural break? If it is a structural break, how long do we have to wait to really tell? To answer these question, we will examine three different scenario and study the behaviour of the posterior estimate of the ω's. In all the three cases we will use the simplest model, the local level

6.3.1 Scenario 1

We use simulated data and considered a case where the most current, at time $t = 181$ [1], data value recorded is an extreme or a potential outlier. See Figure 6.15
The posterior estimates of the auxiliary variables ω's from this series are shown in Figure 6.16. From this graph, it is very clear from the plot of ω_y that the two 'serious' outliers, at time $t = 80$ and $t = 181$ have been captured. On the other hand, the posterior estimates of ω_θ do not signal any structural break, as logically expected.

[1] The number 181 is an arbitrary choice with no practical or statistical significance attached

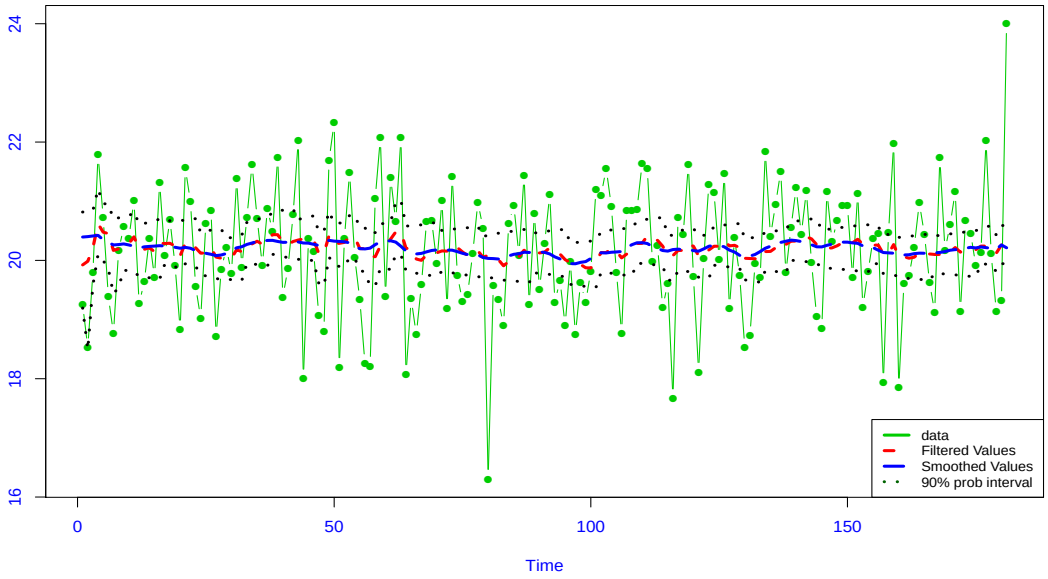

Figure 6.15: Filtered and smoothed values of a time series with two potential outliers, one at the current time $t = 181$

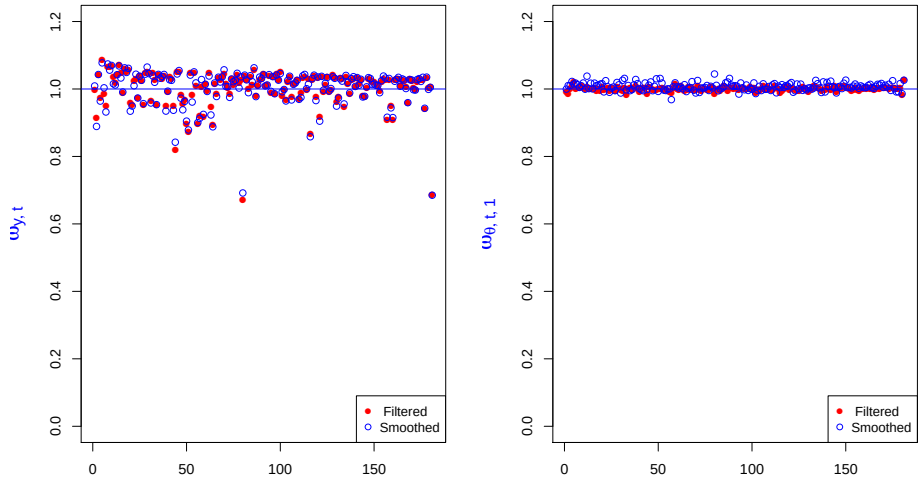

Figure 6.16: Posterior estimates of $\omega_{y,t}$ (left) and $\omega_{\theta,t}$ (right) from a time series with distinct outliers at time $t = 80$ and the current time $t = 181$

The same plot as in Figure 6.16 but including probability interval is shown in Figure 6.17

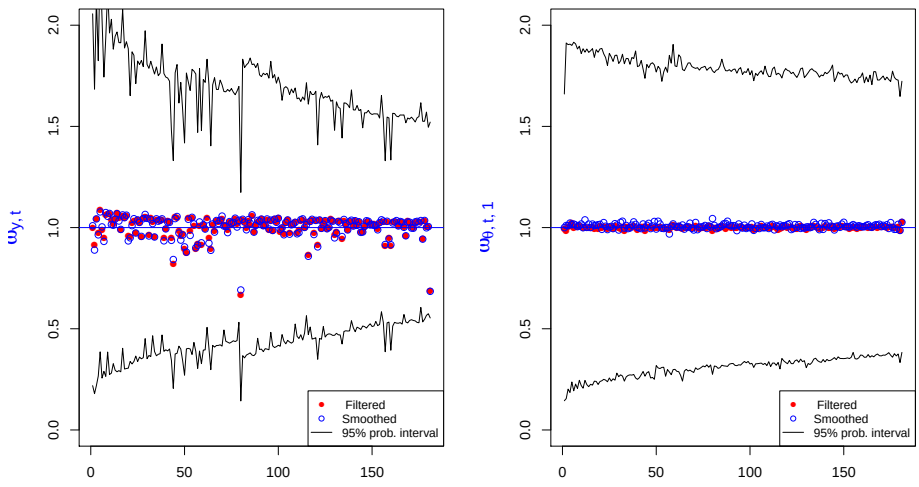

Figure 6.17: Posterior estimates of $\omega_{y,t}$ (left) and $\omega_{\theta,t}$ (right) from a time series with distinct outliers at time $t = 80$ and the current time $t = 181$

6.3.2 Scenario 2

We use the entire data from section 6.3.1 and assume we have a new data now received, at time $t = 182$, which is *within* the range of other data values. The filtered and smoothed values from this series are shown in Figure 6.18

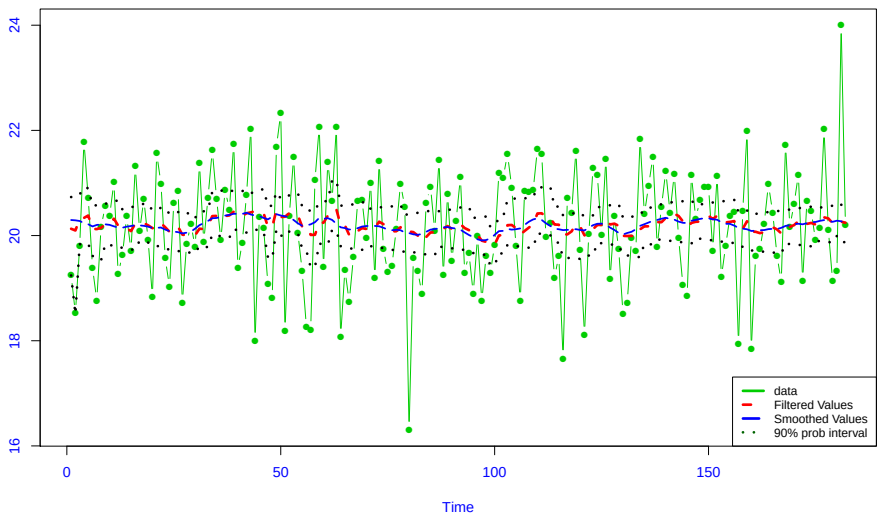

Figure 6.18: Filtered and smoothed values of a time series with current at time $t = 182$ within the expected level

The posterior estimates for the auxiliary variables ω_y and ω_θ are displayed in Figure 6.19. Obviously, from the plots we can see that the two extreme values at time $t = 80$ and $t = 181$ have been identified as outliers. The posterior estimates of ω_θ, on the other hand are quite stable as expected.

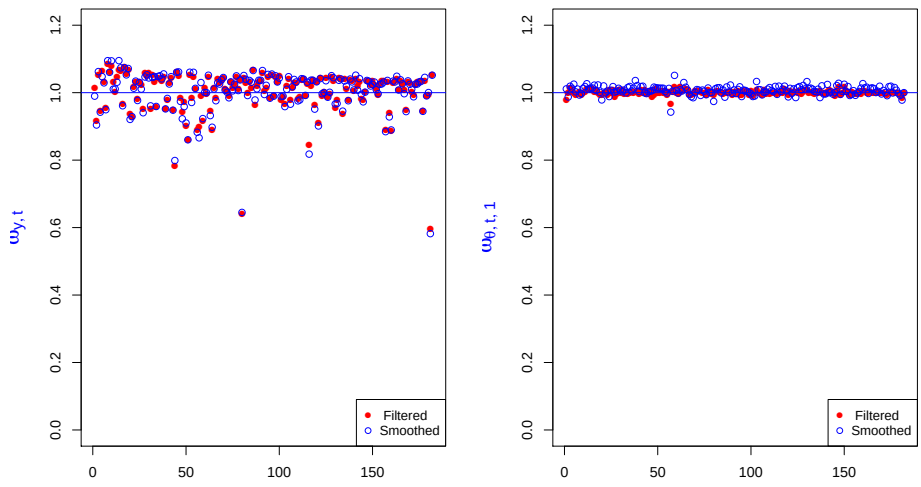

Figure 6.19: Posterior estimates of $\omega_{y,t}$ (left) and $\omega_{\theta,t}$ (right) from a time series with distinct outliers at time $t = 80$ and at time $t = 181$.

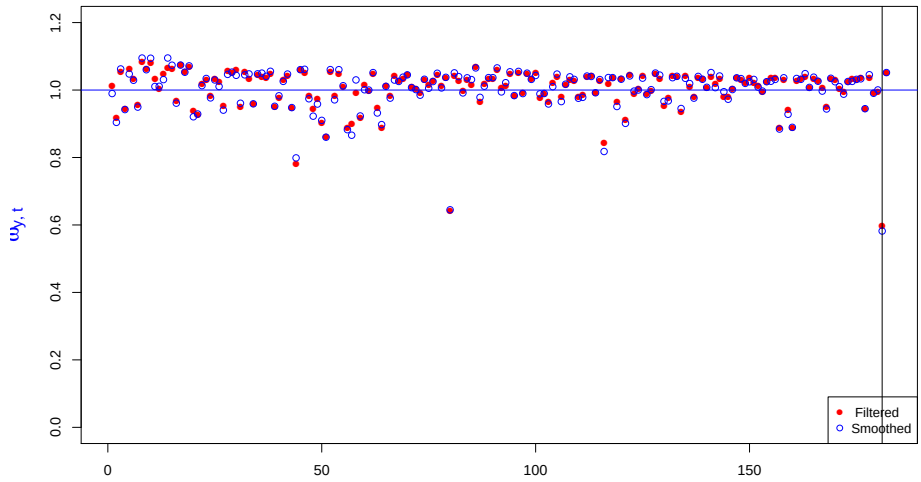

Figure 6.20: An elaborate plot of posterior estimates of $\omega_{y,t}$ showing presence of outliers at time $t = 80$ and time $t = 181$

The same plot as in Figure 6.19 but including probability interval is shown in Figure 6.21

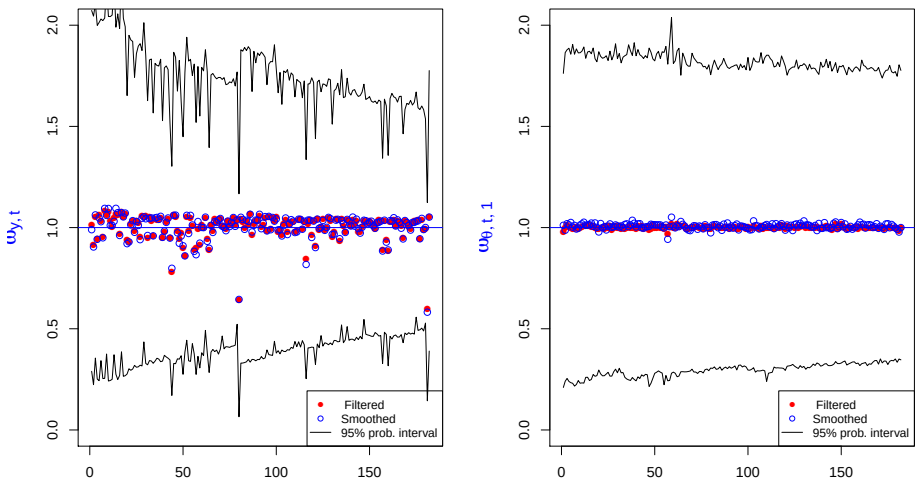

Figure 6.21: Posterior estimates of $\omega_{y,t}$ (left) and $\omega_{\theta,t}$ (right) from a time series with distinct outliers at time $t = 80$ and at time $t = 181$.

6.3.3 Scenario 3

Again, we use the entire data set from section 6.3.1 and assume this time that the new data currently received, that is at time $t = 182$, is also an extreme. So in this case we have two consecutive extreme data values. This may be interpreted as either two consecutive outliers or onset of structural break at the previous time step. The filtered and smoothed values from this series are shown in Figure 6.22

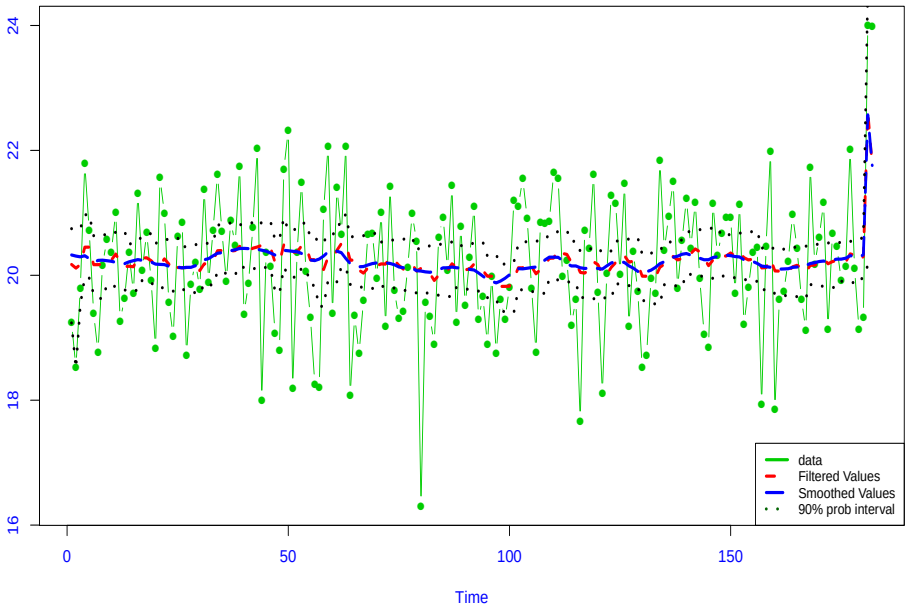

Figure 6.22: Filtered and smoothed values of a time series with the two most current values ($t = 181, 182$) *far* from their expected values

The posterior estimates of the auxiliary variable ω from this series are plotted in Figure 6.23. From the graph, the model strongly suggest a structural break in the series at time $t = 181$. The estimate for ω_θ at that time is conspicuously low. Unlike in scenario two, the model this time indicate presence of outlier only at time $t = 80$.

These results shows that our algorithm is able to distinguish between the outlier and structural break, and in this data set we only have to wait one time step to do that.

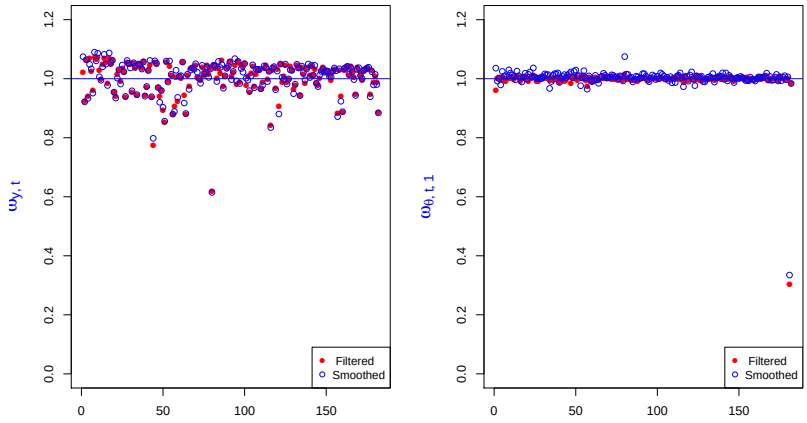

Figure 6.23: Posterior estimates of $\omega_{y,t}$ (left) and $\omega_{\theta,t}$ (right) from a time series with distinct outlier at time $t = 80$ and two most current values ($t = 181, 182$) *far* from their expected values

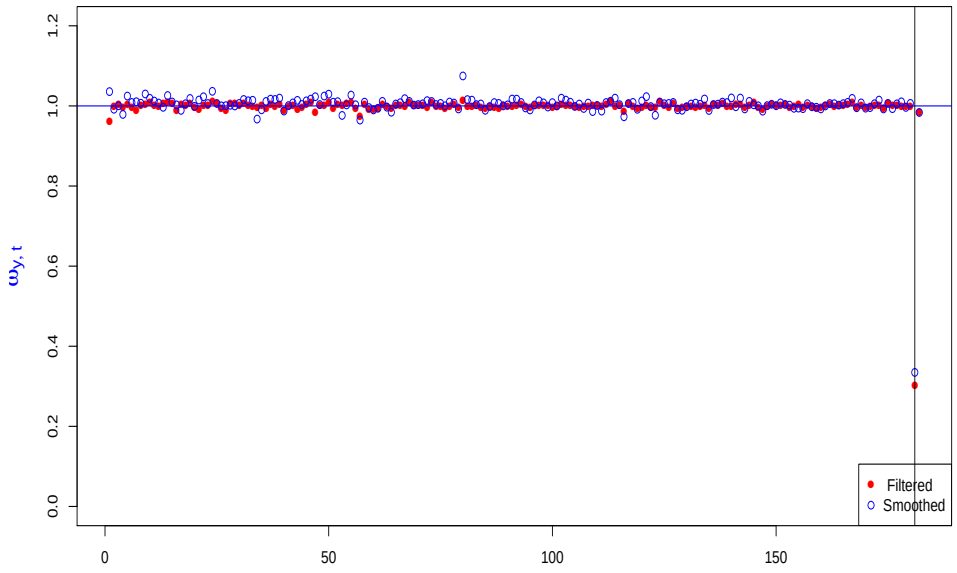

Figure 6.24: An elaborate plot of posterior estimates of $\omega_{\theta,t}$ showing potential structural break at time $t = 181$

The same plot as in Figure 6.23 but including probability interval is shown in Figure 6.25

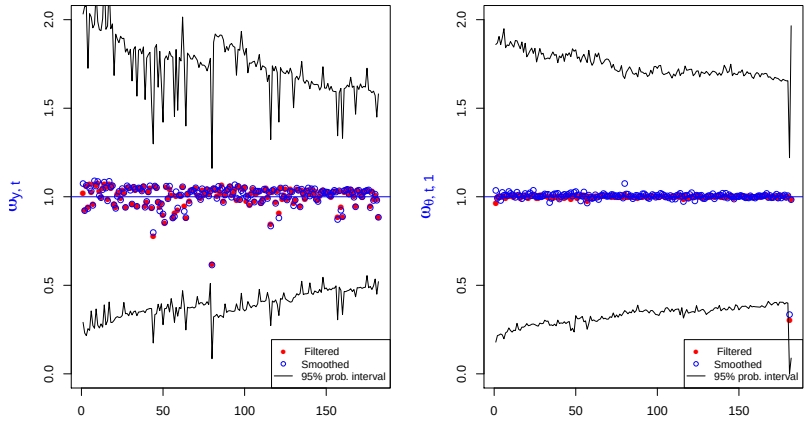

Figure 6.25: Posterior estimates of $\omega_{y,t}$ (left) and $\omega_{\theta,t}$ (right) from a time series with distinct outlier at time $t = 80$ and two most current values ($t = 181, 182$) *far* from their expected values

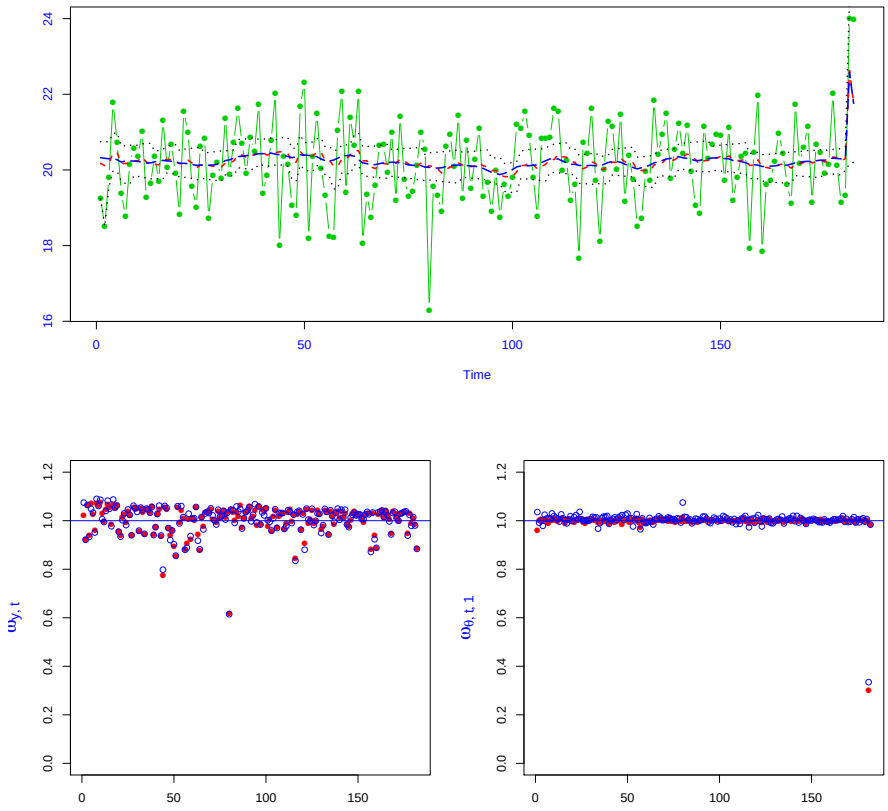

Figure 6.26: Top: A plot of filtered, smoothed and data from a time series with potential outlier at t= 80, and two most current data *far* from their expected values
Bottom left: Posterior estimates of $\omega_{y,t}$ showing a distinct outlier at time $t = 80$
Bottm right: Posterior estimates of $\omega_{\theta,t}$ indicate a potential structural break at time $t = 181$

Chapter 7

Discussion

When our particle filter is employed in the model for outliers and structural breaks, and with large number of particles, it can estimate jointly the model parameters and filter the states. The results also prove that the algorithm is accurately able to flag possible observation outliers and structural breaks in real time.

From both the design of observation variance V_t and the results shown, we realise that the further the observation is from the rest of the data, the larger the value of V_t is and consequently the smaller the value of $\omega_{y,t}$. Therefore, the closer the value of $\omega_{y,t}$ is to zero, the more the outlyingness on observation y_t. The threshold value of $\omega_{y,t}$, for an observation to be considered serious outlier, is still under study but a value less than 0.7 is significant. When there is no outlying observations in the data, the values of ω_y are expected to be equal to 1. Similarly when there is no structural break in the series, the values of $\omega_{\theta,j,t}$ are expected to be 1. Value of $\omega_{\theta,j,t}$ less than 1 indicate a structural break in the j^{th} component of the state vector at time t, and the smaller the value is the larger the break. It is worth noting that, like many other statistical inferences, the more data we have the more effective the model is. We expect better results from data set of length 100, say, as compared to data set of length 10. Similarly the output from the smoothing density, which utilizes ℓ number of most current observations at a time, is more accurate than results from the filtering density which utilizes only the most current observation.

Both the computational complexity and memory requirements for the particle filter is

$\mathcal{O}(N)$ whereas that of MCMC is $\mathcal{O}(tN)$. Every time a new data value is observed requires an update of posterior distribution, that is a move from $p(X_{0:t-1}, \psi|y_{1:t-1})$ to $p(X_{0:t}, \psi|y_{1:t})$. Consequently, a new MCMC run will be required which make the MCMC approach impractical for online inferences. The results from the SMC algorithm are comparable with those from MCMC output. However, the SMC algorithm is quite fast while the running time for MCMC is unbearable, especially as the of length the data set and the dimension of state-space increases

Again, our results outperform those from regular outlier detection strategies such as the Shewhart, CUSUM and EWMA. The Shewhart chart were unable to detect any structural break, but are good in detecting large outlying observations. The CUSUM charts are able to detect the break but after a substantial delay. The EWMA is more sensitive to small jumps in the series than CUSUM and thus able to detect the jump much faster although one or two time steps after our SMC approach. Also, the parameters involved in EWMA procedure requires very careful choice or need to be estimated from the data. This may be challenging and requires a lot of expertise. Lastly, when it comes to structural time series of higher order, i.e. $p > 1$, these techniques will not be able to identify structural breaks, if any, in individual components in the state space.

The beauty of our model is being able to distinguish between an outlier and structural break. Once an observation arrive and it is far from its expected value, logically we have to wait for at least one more observation to certainly determine if it is an outlier or if it is an onset of a structural break in the series. This was demonstrated by results in section 6.3. From the data used, the onset of structural break was confirmed after our second observation, from where the break occurred, was received. Prior to getting the second observation, the data value where the break occurred was flagged as an outlier. This is important in keeping the analyst alert of likely possibility of a structural break. The sensitivity of SMC to changes in prior specifications is still under study and left as future work.

References

Andreou, E., & Ghysels, E. (2009). Structural breaks in financial time series. In *Handbook of financial time series* (pp. 839–870). Springer.

Aoki, M. (1990). *State space modeling of time series.* Cambridge Univ Press.

Ardelean, V. (2012). *Detecting outliers in time series* (Tech. Rep.). IWQW Discussion Paper series.

Atkinson, A., Koopman, S., & Shephard, N. (1997). Detecting shocks: outliers and breaks in time series. *Journal of Econometrics*, *80*(2), 387–422.

Aue, A., & Horváth, L. (2013). Structural breaks in time series. *Journal of Time Series Analysis*, *34*(1), 1–16.

Barnett, V., & Lewis, T. (1984). Outliers in statistical data. *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, Chichester: Wiley, 1984, 2nd ed.*, *1*.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (Vol. 3). Wiley New York.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, *31*(3), 307–327.

Bottle, A., & Aylin, P. (2008). Intelligent information: a national system for monitoring clinical performance. *Health services research*, *43*(1p1), 10–31.

Box, G., & Tiao, G. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, *70*(349), 70–79.

Brockwell, P. J., & Davis, R. A. (2009). *Time series: theory and methods.* Springer.

Carvalho, C., Johannes, M., Lopes, H., & Polson, N. (2010). Particle learning and smoothing. *Statistical Science*, *25*(1), 88–106.

Cobb, G. W. (1978). The problem of the nile: conditional solution to a changepoint problem. *Biometrika*, *65*(2), 243–251.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, *74*(365), 169–174.

Crisan, D., & Doucet, A. (2002). A survey of convergence results on particle filtering methods for practitioners. *Signal Processing, IEEE Transactions on*, *50*(3), 736–746.

Del Moral, P. (2004). *Feynman-kac formulae*. Springer.

Douc, R., & Cappé, O. (2005). Comparison of resampling schemes for particle filtering. In *Image and signal processing and analysis, 2005. ispa 2005. proceedings of the 4th international symposium on* (pp. 64–69).

Doucet, A., De Freitas, N., Gordon, N., et al. (2001). *Sequential monte carlo methods in practice* (Vol. 1). Springer New York.

Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press.

Fearnhead, P. (2002). Markov chain monte carlo, sufficient statistics, and particle filters. *Journal of Computational and Graphical Statistics*, *11*(4), 848–862.

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 350–363.

Franklin, S., Thomas, S., & Brodeur, M. (2000). Robust multivariate outlier detection using mahalanobis distance and modified stahel-donoho estimators. In *Proceedings of the second international conference on establishment surveys* (pp. 697–706).

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of time series analysis*, *15*(2), 183–202.

Galeano, P., Peña, D., & Tsay, R. S. (2006). Outlier detection in multivariate time series by projection pursuit. *Journal of the American Statistical Association*, *101*(474), 654–669.

Gilks, W., & Berzuini, C. (2002). Following a moving targetmonte carlo inference for dynamic bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(1), 127–146.

Godsill, S., Doucet, A., & West, M. (2004). Monte carlo smoothing for nonlinear time series. *Journal of the American Statistical Association*, *99*(465), 156–168.

Gordon, N. J., Salmond, D. J., & Smith, A. F. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. In *Radar and signal processing, iee proceedings f* (Vol. 140, pp. 107–113).

Harrison, J., & West, M. (1991). Dynamic linear model diagnostics. *Biometrika*, *78*(4), 797–808.

Harrison, J., & West, M. (1997). *Bayesian forecasting and dynamic models.* Springer Verlag, New York.

Harvey, A., & Koopman, S. (2005). Structural time series models. *Encyclopedia of Biostatistics*.

Hawkins, D. M., & Olwell, D. H. (1998). *Cumulative sum charts and charting for quality improvement*. Springer.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126.

Kalman, R., et al. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, *82*(1), 35–45.

Kedem, B., & Fokianos, K. (2002). *Regression models for time series analysis* (Vol. 323). Wiley-Interscience.

Kitagawa, G. (1996). Monte carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of computational and graphical statistics*, *5*(1), 1–25.

Kitagawa, G., & Gersch, W. (1996). *Smoothness priors analysis of time series* (Vol. 116). Springer.

Knorr, E. M., Ng, R. T., & Tucakov, V. (2000). Distance-based outliers: algorithms and applications. *The VLDB JournalThe International Journal on Very Large Data Bases*, *8*(3-4), 237–253.

Kong, A., Liu, J. S., & Wong, W. H. (1994). Sequential imputations and bayesian missing data problems. *Journal of the American statistical association*, *89*(425), 278–288.

Koop, G., & Potter, S. (2000). Nonlinearity, structural breaks, or outliers in economic time series. *Nonlinear Econometric Modeling in Time Series Analysis*, 61–78.

Laurikkala, J., Juhola, M., Kentala, E., Lavrac, N., Miksch, S., & Kavsek, B. (2000). Informal identification of outliers in medical data. In *Fifth international workshop on intelligent data analysis in medicine and pharmacology* (pp. 20–24).

Liu, J., & West, M. (1999). *Combined parameter and state estimation in simulation-based filtering*. Institute of Statistics and Decision Sciences, Duke University.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, *2*, 49–55.

Montgomery, D. C. (2007). *Introduction to statistical quality control*. John Wiley & Sons.

Page, E. (1954). Continuous inspection schemes. *Biometrika*, 100–115.

Peña, D., & Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, *43*(3).

Penny, K. I., & Jolliffe, I. T. (2001). A comparison of multivariate outlier detection methods for clinical laboratory safety data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *50*(3), 295–307.

Perron, P. (2006). Dealing with structural breaks. *Palgrave handbook of econometrics*, *1*, 278–352.

Petris, G., Petrone, S., & Campagnoli, P. (2009). *Dynamic linear models with r*. Springer.

Polson, N. G., Stroud, J. R., & Müller, P. (2008). Practical filtering with sequential parameter learning. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(2), 413–428.

Roberts, S. (1959). Control chart tests based on geometric moving averages. *Technometrics*, *1*(3), 239–250.

Rodrigues, P., & Rubia, A. (2011). The effects of additive outliers and measurement errors when testing for structural breaks in variance*. *Oxford Bulletin of Economics and Statistics*, *73*(4), 449–468.

Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection* (Vol. 589). John Wiley & Sons.

Severin, T., & Schmid, W. (1998). *Statistical process control and its application in finance*. Springer.

Shekhar, S., Lu, C.-T., & Zhang, P. (2003). A unified approach to detecting spatial outliers. *GeoInformatica*, *7*(2), 139–166.

Shephard, N. (1994). Partial non-gaussian state space. *Biometrika*, *81*(1), 115–131.

Shewhart, W. A. (1926). Quality control charts1. *Bell System Technical Journal*, *5*(4), 593–603.

Storvik, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *Signal Processing, IEEE Transactions on*, *50*(2), 281–289.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of forecasting*, *7*(1), 1–20.

West, M., & Harrison, J. (1989). Subjective intervention in formal models. *Journal of Forecasting*, *8*(1), 33–53.

Woodall, W. H. (2006). The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology*, *38*(2), 89–104.

Zhao, J., Lu, C.-T., & Kou, Y. (2003). Detecting region outliers in meteorological data. In *Proceedings of the 11th acm international symposium on advances in geographic information systems* (pp. 49–55).