Università di Firenze, Università di Perugia, INdAM consorziate nel CIAFM

**DOTTORATO DI RICERCA**
**IN MATEMATICA, INFORMATICA, STATISTICA**

CURRICULUM IN STATISTICA
CICLO XXXI

**Sede amministrativa Università degli Studi di Firenze**
Coordinatore Prof. Graziano Gentili

# Robust time series analysis with the Forward Search

Settore Scientifico Disciplinare SECS-S/01


**Dottorando**:
Alessandro Ghiretti

**Tutore**
Prof. Fabrizio Laurini

**Coordinatore**
Prof. Graziano Gentili


Anni 2015/2018

Life is what happens to you while you're busy making other plans.

---

*Jhon Lennon*

# Ringraziamenti

Sono stati tre anni intensi, in cui ho vissuto momenti stupendi e di difficoltà, e durante i quali ho avuto la fortuna di conoscere tante persone eccezionali da cui ho imparato molto.
Il primo ringraziamento speciale lo dedico ai miei due splendidi genitori Daniela e Paolo che mi hanno appoggiato nel modo migliore che un figlio potesse desiderare. Vorrei ringraziarli per avere sostenuto economicamente la mia educazione, per avermi concesso l'opportunità di proseguire i miei studi in questo stupendo campo che è la statistica e per essermi stati vicino in questi anni durante tutte le situazioni difficili. Non potrei desiderare genitori migliori.
Il secondo ringraziamento lo dedico a Fabrizio Laurini per la sua guida e la pazienza che ha dimostrato durante il mio lavoro a questa tesi. Il suo supporto è stato fondamentale per diversi motivi: mi ha appassionato e spinto ad approfondire il tema delle time series, dandomi molti suggerimenti e spunti utili per iniziare e terminare questo lavoro. Lo ringrazio anche fortemente per essere stato di conforto dopo diversi fallimenti.
Ringrazio Andrea Cerioli, che mi ha incoraggiato nel proseguire i miei studi e nel coltivare la mia passione per la statistica. E' stato un ottimo mentore.
Ringrazio i restanti membri della sezione di Statistica dell'Università di Parma. In particolare, Marco Riani, che mi ha aiutato nel migliorare le mie competenze di programmazione, specialmente durante il primo anno quando ero un "pivello". Gianluca Morelli e Aldo Corbellini, con cui ho condiviso tanti momenti di divertimento e di svago.
Vorrei ringraziare Giovanni M. Marchetti, che mi ha insegnato come approcciarmi alla statistica, tirandomi su il morale in alcuni momenti in cui non mi sentivao all'altezza della situazione.
Infine, un caloroso ringraziamento lo dedico a tutte le persone che non compaiono nella lista ma che mi sono state vicine e mi hanno supportato e motivato durante questo percorso.

# Acknowledgements

I spent three intense years, in which I experienced wonderful and difficult moments, and I was lucky enough to meet many exceptional people from whom I learned a lot.
I dedicate my first special thanks to my parents, Daniela and Paolo, who supported me in the best way that a son could wish for. I would like to thank them for having economically supported my education, for allowing me to continue my studies further in this amazing field which statistics is and for being close to me during these years in all the difficult situations. I could not wish for better parents.
I would like to dedicate my second thanks to Fabrizio Laurini, for his guidance and patience that he has shown during my work on this thesis. His support was fundamental for several reasons: he made me passionate and guided me through the topic of time series analysis, giving me many suggestions and useful ideas to start and finish this work. I also thank him very much for being reassuring after several failures.
I would like to thank Andrea Cerioli who encouraged me in proceeding my education and in cultivating my passion for statistics. He has been a great mentor.
I thank the members of the Statistics section at the University of Parma. In particular, Marco Riani, who helped me to improve my programming skills, especially during the first year when I was a rookie.
Gianluca Morelli and Aldo Corbellini with whom I shared many moments of fun and entertainment.
I would like to thank Giovanni M. Marchetti, who provided useful suggestions about how to open up to statistics and cheered me up on several occasions when I felt not good enough.
To conclude, I would like to thank all the people whom names are not mentioned but who supported and motivated me during this path.

# Abstract

In this thesis I consider the problem of outliers detection in univariate time series, in particular when the assumed underlying data generating process belongs to the class of seasonal linear processes.
In statistical literature several outliers detection techniques have been proposed in the context of linear seasonal time series models, and lately, further techniques have been developed in the computer science literature, however, these are not base on an underlying statistical model and will not be discussed here.
Traditional outliers detection technique may fail when a pattern of outliers contaminate the data. In this case two effects can be observed: swamping and masking. Swamping happens when the effect of the outlying units is so strong that the procedure classifies good units as outlying data points. Masking, instead, results when few data points are not detected by the diagnostics because they appear "masked" from the other outliers units.
This work consider an extension of the Forward Search procedure to the aforementioned class of models, with the aim of producing a detection method not affected by the masking nor by the swamping effects. In order to extend the Forward Search, the models are first represented in their State Space form, and to deal with the missing units generated through the search, the Kalman filter is used. During the search a series of quantities, relevant for outliers detection and which results insensitive to the swamping and masking effect are monitored.
The aim of this work is threefold. First, review the theory behind the class of linear stationary seasonal models and the theory behind outliers in the time series context. Second, to extend the Forward Search to the class of seasonal linear time series models, introducing a series of new quantities whose monitoring can reveal spurious units. Third, suggest an automatic outliers detection technique that might be used as an exploratory tool to identify in a straightforward way anomalies in the data. In order to show the potential of the Forward Search two real examples are considered.

# Contents

# Introduction

This thesis is concerned with the application of the Forward Search to linear models for the analysis of time series data. In particular we refer to the class of $SARIMA$ models, which allow to describe stationary as well as non stationary series that may contain a trend, a seasonal component or both.
As in standard parametric techniques developed for independent data, classical time series methods relies on the assumption of normality.
By assuming Gaussian innovations it is possible to simplify the computational aspects of the estimation process, and take advantages of some well known results concerning the maximum likelihood estimator and the derived quantities.
Nevertheless, when the data deviates from normality, due to few atypical observations, traditional estimation and inferential procedures might result inconsistent and strongly biased.
In the statistical literature, units that may badly affect the estimates and the inferential results are called outliers. In the literature, many different definitions of outliers have been proposed, but in general units that differ from the majority of the data or that result inconsistent with the assumed model are identified as outliers.
Some recent references that discuss the impact of the outliers and the related statistical methods in a detailed way are Maronna et al. (2006), Huber and Ronchetti (2009) and Hampel et al. (2011).
In statistics the problem of outliers has been addressed in two different ways. Through the robust approach and through the detection and identification of atypical units.

- *Robust approach*: in the robust approach the major aim is to develop estimators and inferential methods that are not affected by a certain number of outliers. Thus, the objective of a robust estimator is to obtain an estimate as close as possible to the estimate obtained in absence of outliers. Likewise, the aim of a robust inferential test is to obtain a statistical test that maintain its good properties when the sample is contaminated.

- *Outliers detection*: outliers detection procedures have the objective to identify the units that represent potential outliers. Generally these methods work in two steps.

First a measure of deviation between the data and the assumed model is computed. Second a statistical test is performed to check for statistical evidence against a given units. Once the units have been flagged as potential outliers further analysis can be done on the data.

Often the difference between robust and outliers detection procedures is very mild, however we want to emphasis that the two techniques pursue two different objectives.
Following this further, robust procedures, depending on the way they are performed can be classified into three different categories:

- *Soft Trimming*: soft trimming techniques refer mainly to $M$, $S$, $MM$ estimators and derived methods (Huber and Ronchetti (2009)). The aim here is that observations near the center of the distribution retain fully their value, but a function $\rho(\cdot)$, which determines the form and the amount of trimming, ensures that increasingly remote observations receive a weight that decreases with the distance from the center.

- *Hard Trimming*: in hard trimming procedures observations are weighted according to a $\{0, 1\}$ scheme. Units that conform with the specified model receive weight 1 and are included in the estimation process, whereas observations that do not conform with the assumed model, and are potential outliers, receive weight 0 and are excluded from the estimation. The most widely used hard trimming estimators are the $LTS$ estimator and its extension the $RLTS$ (Rousseeuw (1984)) that retains high robustness while resulting more efficient.

- *Adaptive Trimming*: in the Forward Search ($FS$) (Hadi (1992), Atkinson et al. (2010), Atkinson and Riani (2012)), the observations are again hard trimmed, but the value of $h$, that is to say the number of units on which the fit is performed, is determined by the data, being found adaptively by the search. Data analysis starts from a very robust fit to a few, carefully selected, observations found by $LMS$ or $LTS$ with the minimum value of $h$. The number of observations used in fitting then increases until all are included.

A dated, but broad survey of robust theory applied to time series models is given by Stockinger and Dutter (1987).
Differently from the *iid* context, in time series data, depending on the type of the contamination model assumed, it is possible to identify different types of outliers. A first classification of outliers in time series was given by Fox (1972), who introduced the innovative and additive outliers, respectively abbreviate by IO and AO, and two different models that can be used to describe them and their effects on the time series.
Further types of outliers have been discussed by Tsay (1988) and Peña (1990). A detailed discussion of the different forms out outliers and on their effects will be given in Chapter

5. The identification and the study of the outliers in time series is an integrating part of the analysis process. In fact, by considering, the type, the magnitude and the instant at which an outlier occurred, it is possible to collect useful insights about the underlying series. For example, an outlier might indicate a change of regime, suggesting that a different model might be more appropriate for that problem. Further, the presence of different outliers in a series might reveal that particular events of interest took place at those instants. For all of these reasons the treatment of outliers is a complex subject that requires as much common sense as theory. Methods developed for dealing with outliers should not be seen as mere procedures that classify an observation as conform or not with respect to the assumed model, but as an exploratory tool that may reveal interesting characteristics of the underlying data. In my personal opinion, the exploratory nature of the Forward Search fits really well in the context of time series.

The $FS$ represent a powerful and flexible tool that can be applied in many different context and for different purposes. Other applications beyond outliers detection are robust inference and monitoring.

In this thesis I propose a possible extension of the Forward Search method to the class of $SARIMA$ models. In time series analysis, the Forward Search has been first introduced by Grossi and Riani (2002), in the context of non-stationary structural time series models and subsequently Grossi and Laurini (2009) and Grossi (2004) extended it to conditional volatility models ($ARCH$ and $GARCH$) models.

In particular, Riani (2004), used the Kalman filter to obtain the exact maximum likelihood estimates of the hyper-parameters, pointing out that the missing observations generated at the different steps of the $FS$ are easily handled by the Kalman filter and its recursions. Whereas, F. Laurini and L. Grossi proved how the $FS$ can be used as a useful tool to improve the performance of a statistical test.

In Grossi and Laurini (2009), they showed how the standard $ARCH$ test can be greatly affected by atypical units and next they suggested the $FS$ as a flexible tool that can be used to robustify the $ARCH$ test. Recently Crosato and Grossi (2017), extended the Forward Search methodology to $GARCH(1,1)$ models. In that setting, where the Kalman filter can not be used to handle the missing observations they adopted an adaptive soft trimming strategy. That is, a weight function assigns a weight to each unit at each step of the search. Next, the estimation is performed on all the units taking into account the different wights received. On one hand, as pointed out by the authors, this procedure decrease the robustness properties of the search, while, on the other hand, it allows to keep intact the natural time dependence structure of the data which is generally not considered in the $FS$ performed with the Kalman filter. When dealing with general $ARMA(p,q)$ processes one of the difficulties that might be encountered is parameters estimation.

For this reason many outlier detection techniques developed for this class of models relied on the $AR$ approximation. By doing so it is possible to obtain conditional maximum likelihood estimates in an easy way. However, the $AR$ approximations despite it is correct

from mathematical reasoning might be inconvenient in many practical applications where a parsimonious model is preferred, or when the observed time series is short.
By representing the $ARMA$ process in the state space form we overcome this difficulty. The use of the state space representation and the Kalman filter has two advantages. First it enable us to treat exact $ARMA(p, q)$ process without approximating the model of interest with an high order $AR$ model. Second following Riani (2004), the Kalman filter results a formidable technique to handle the missing unit generated during the $FS$.
Moreover, since Kalman filter can be used to obtain maximum likelihood estimators for every linear and non linear model that has a state space representation, the $FS$ applied to state space models can potentially be applied to any linear model that admits a state space representation.
The thesis will be structured as follows. In Chapter 1 I will introduce some basic concepts about time series models and time series analysis that lay the foundations of the topics discussed in the subsequent chapters. Chapters 2 presents the class of $ARMA$ processes and their related time domain properties and characteristics. Chapter 3 presents the class of $SARIMA$ models that are the natural extension of $ARMA$ model to deal with non stationary processes that exhibit trend, seasonal component or both. In Chapter 4 I explain the state space representation of the models discussed in the previous chapters and that are used in the Forward Search. Moreover, I will develop the theory behind the Kalman filter which plays a key role in the Forward Search in obtaining the maximum likelihood estimators
Next, in Chapter 5 I will present the different types of outliers that can be encountered in time series data and in Chapter 6 I will review the outlier detections techniques, that inspired the extension of the Forward Search to the class of $SARIMA$ models.
Finally, in Chapter 7 starting from the linear regression model I will present the Forward Search method and its extensions to the class of $SARIMA$ family of processes. In order to show the potential of this technique two real time series are analyzed and discussed.

# Chapter 1

# Introduction to stochastic processes

The objective of this first chapter is to introduce formally the concept of a time series and review the statistical and probabilistic tools that will be used in the next chapters. Starting from the definition of stochastic process, the properties of stationarity and ergodicity will next be presented. Subsequently we will shortly review the main stochastic processes generally encountered in time series literature.

## 1.1 Stochastic process: definition and properties

A time series is a set of ordered observations $y_t$, taken often at equal spaced instants in time. In order to discuss formally the properties and the characteristics that a time series have, it is customary to consider a time series as a finite realization of an underlying process that generates the data.
Formally, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where $\Omega$ is the sample space, $\mathcal{F}$ is the $\sigma$-algebra defined on the sample space and $\mathbb{P}$ is the probability measure. Take $\mathcal{T}$ as an index, and without loss of generality consider $\mathcal{T}$ as the time.

**Definition 1** *A **stochastic process** is a collection of indexed random variables* $\{Y_t\} = \{Y_t(\omega) : t \in \mathcal{T}, \omega \in \Omega\}$*, that is* $Y_t : \Omega \to \mathbb{R}$*.*

Mathematically, the index set $\mathcal{T}$ can be any set, but for practical applications it is usually taken as $\{1, 2, \ldots, n\}$, as $\mathbb{N}$, the set of natural numbers, or as $\mathbb{R}$, the set of real numbers. Each random variable $Y_t(\omega)$ in the stochastic process is a function of the two elements $t$ and $\omega$, and depending on them it can assume different meanings. In particular:

1. When $Y_t(\omega)$ is a function of $t \in \mathcal{T}$ and $\omega \in \Omega$, $Y_t(\omega)$ is a stochastic process
2. When $Y_t(\omega)$ is a function of $t \in \mathcal{T}$ and we fix $\omega = \omega_0$, then $Y_t(\omega_0)$ represents a single realization of the stochastic process, which is called **sample path**

3. When $Y_t(\omega)$ is a function of $\omega$ given $t = t_0$, $Y_t(\omega)$ is a random variable
4. When $t = t_0$ and $\omega = \omega_0$, we have a real number that represents the value assumed by $Y_t(\omega)$ when $\omega = \omega_0$ at instant $t = t_0$.

Given the previous notions it is now possible to formally define a time series.

**Definition 2** *A time series* $\{y_t\}, t = 1, 2, \ldots, n$ *is defined as an observed sample path of the underlying stochastic process* $\{Y_t\}$ *which is assumed to generate the data.*

In order to fully describe the probability structure of a stochastic process $\{Y_t\}$, all the finite dimensional marginal distributions of the process should be known. That is to say, we should know the distribution function $F_{Y_{t_i}}(y_{t_i})$ at each instant $t_i$, the distribution function $F_{Y_{t_i},Y_{t_j}}(y_{t_i}, y_{t_j})$ at each pair of instants $t_i, t_j$, and so on until the joint distribution of the process$F_{\{Y\}}(y_{t_1}, y_{t_2}, \ldots, y_{t_n})$ is fully specified.
In practice this is hardly possible and it is customary to define the stochastic process in terms of its first and second moments. Define the expectation operator as $\mathbb{E}[Y_t] = \int_{\mathcal{S}} y_t dF_{Y_t}(y_t)$, where $\mathcal{S}$ is the support of $Y_t$. The first and second moments are given by

$$\mathbb{E}[Y_t] = \mu_t \tag{1.1}$$

$$\mathbb{V}(Y_t) = \mathbb{E}[(Y_t - \mu_t)^2] = \sigma_t^2 \tag{1.2}$$

$$\mathbb{C}\text{ov}(Y_t, Y_k) = \mathbb{E}[(Y_t - \mu_t)(Y_k - \mu_k)] = \gamma(t, k) \tag{1.3}$$

If several realizations are available, the above quantities could be estimated by "ensemble" estimators. For example an ensemble estimate for equation (1.1) is given by

$$\tilde{\mu}_t = \frac{1}{m} \sum_{i=1}^{m} y_t^{(i)}.$$

where $y_t^{(i)}$ denotes the $i$th realization of $Y_t$ and $m$ is the total number of realizations.
However, for each random variable $Y_t$ at time $t$, we have at hand only a single observation and some restrictions must be imposed on the process that generate the data. These restrictions are necessary both, to restrict the number of parameters needed to characterize the probabilistic structure of the process, and to obtain meaningful estimates. Restrictions come in two forms: stationarity and ergodicity.

1. *Stationarity*: restrictions on the time heterogeneity of the process, that allows to reduce the number of parameters needed to describe the process.

2. *Ergodicity*: restrictions on the memory of the process such that the moment of the process can be estimated in a consistent way.

### 1.1.1 Stationarity

**Definition 3** *Let* $\{Y_t\}$ *be a stochastic process and let* $F_{\{Y\}}(y_{t_1}, y_{t_2}, \ldots, y_{t_n})$ *denote the distribution function of the process at times* $t_1, t_2, \ldots, t_n$*. The process* $\{Y_t\}$ *is defined a* ***strictly stationary*** *process if for each integer* $\{t_1, t_2, \ldots, t_k\} \in \mathcal{T}$ *and* $k$

$$F_{\{Y\}}(y_{t_1}, ..., y_{t_k}) = F_{\{Y\}}(y_{t_1+k}, ..., y_{t_m+k}).$$

Strict stationarity requires that the joint distribution of the process $\{Y_t\}$ is invariant with respect to time shifts. This assumption is particularly strong, and in fact, assuming strict stationarity is equivalent to assume that all the moments of the process do not change over time.
In many situations it can be extremely difficult, or even impossible, to verify that a stochastic process is strictly stationary. As a consequence, this assumption is relaxed by introducing a weaker form of stationarity that imposes restrictions only on the first two moments of the process.

**Definition 4** *A process* $\{Y_t\}$ *is said to be* ***weakly stationary*** *or* ***covariance stationary*** *if for all* $t \in \mathcal{T}$ *and every* $k$

$$\mathbb{E}[Y_t] = \mu \tag{1.4}$$

$$\mathbb{V}(Y_t) = \mathbb{E}[(Y_t - \mu)^2 = \sigma^2 \tag{1.5}$$

$$\mathbb{C}\text{ov}(Y_t, Y_{t-k}) = \mathbb{E}[(Y_t - \mu)(Y_{t-k} - \mu)] = \gamma(k) \tag{1.6}$$

Literally, weak stationarity requires that the first two moments of the process are not a function of time.
It is important to note that a strict stationary process with finite second moments is covariance stationary, whilst a weakly stationary process where all moments remain constant in time is also a strictly stationary process.
From a mathematical point of view the stationarity assumption allows us to reduce the number of parameters needed to describe the underlying process. While, from a practical point of view it places some regularities conditions on the process, in such a way that we are able to predict, with a certain margin of uncertainty, future values of it.

### 1.1.2 Ergodicity

Stationarity restrictions are introduced in order to decrease the number of elements needed to characterize the distribution of the process.
Once the number of parameters has been restricted it is important to identify under what conditions it is possible to obtain consistent estimates of the parameters using time series averages. It turns out that these conditions are met when the process is ergodic.
In order for a process to be ergodic some measure of dependence between random variables $Y_t$ and $Y_{t-k}$ must vanish as $k$ increase. If this is satisfied a law of large numbers can be applied.

**Theorem 1 (Weak Ergodic Theorem)** *Let $\{Y_t\}$ be a covariance stationary process with finite mean $\mathbb{E}[Y_t] = \mu$, and $\mathbb{C}\text{ov}(Y_t, Y_{t-k}) = \gamma(k)$. If*

$$\sum_{t=1}^{T} |\gamma(k)| < \infty \tag{1.7}$$

*then*

$$\frac{1}{T}\sum_{t=1}^{T} Y_t \xrightarrow{p} \mu \quad as \quad T \to \infty$$

The condition $\sum_{k=1}^{T} |\gamma(k)|$ is known as the **absolute summability** of the autocovariance function and it is obviously satisfied for *iid* processes. Absolute summability is sufficient for weak ergodicity but is by no means necessary for it. A stronger result can in fact be obtained by imposing the weaker condition $\frac{1}{T}\sum_{k=1}^{T} \gamma(k) \to 0$. Processes that satisfy this weaker condition are called mixing processes.

**Definition 5** *A process $\{Y_t\}$ is said to be **mixing** if for all $t \in \mathcal{T}$ and for every $k$, $\lim_{k\to\infty} \lambda(k) = 0$, where $\lambda(k)$ is defined as*

$$|\mathbb{C}\text{ov}(Y_t, Y_{t-k})| \leq \lambda(k). \tag{1.8}$$

The coefficient $\lambda(k)$ is commonly called the mixing coefficient. From a practical point of view the above inequality tells us that as we consider random variables far apart in the series their covariance becomes approximately zero. Serial dependence that is ruled out by the standard LLN, is allowed in the ergodic theorem, provided that it disappears in the long run. Moreover, since for each measurable function $g(\cdot)$, $g(Y_t)$ is ergodic and stationary whenever $Y_t$ is, the ergodic theorem implies that any finite moment, if it exists, can be consistently be estimated by the corresponding sample moment.

### 1.1.3 Autocovariance, autocorrelation and partial autocorrelation functions

When the process $\{Y_t\}$ is stationary its dependence properties can be summarized by $\gamma(k)$, whose expression for a stationary process was given in equation (1.6).
Equation (1.6) seen as a function of the time lag $k$ is called the autocovariance function of the process (ACVF). When the process is stationary it possible to prove that the ACVF enjoys the following properties:

**Theorem 2** *Let $\{Y_t\}$ be a stationary process with autocovariance function $\gamma(k)$. Then $\gamma(k)$ satisfies the following properties*

1. $\sigma^2 = \gamma(0) > 0$
2. $|\gamma(k)| < \gamma(0)$ *for all* $k \geq 1$
3. $\gamma(k) = \gamma(-k)$

The last property allows us to consider the ACVF only for positive lags since it is a symmetric function of the lag value $k$.
In order to compare processes measured on different scales a useful transformation can be obtained dividing $\gamma(k)$ by the variance of the process. This standardized ACVF is called the autocorrelation function ($ACF$). The $ACF$ is a normalized measure of the autocovariance, that is

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\mathbb{C}\text{ov}(Y_t, Y_{t-k})}{\mathbb{V}(Y_t)}. \tag{1.9}$$

For a stationary process the autocorrelation satisfies the following properties

**Theorem 3** *Let $\{Y_t\}$ be a stationary process with autocorrelation function $\rho(k)$. Then $\rho(k)$ satisfies the following properties*

1. $\rho(0) = 1$
2. $|\rho(k)| < 1$ *for all* $k$
3. $\rho(k) = \rho(-k)$

Another further summary of the linear dependence of a stationary process can be obtained from the partial autocorrelation function ($PACF$). The partial autocorrelation function measures the correlation between $Y_t$ and $Y_{t-k}$ after their linear dependence on the intervening random variables $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-k+1}$ has been removed.

**Definition 6** *Let $\{Y_t\}$ be a stationary process. The partial autocorrelation function of the process s defined by*

$$\alpha(k) = \mathbb{C}\text{or}(Y_1 - P_{\{1,Y_1,Y_2,\ldots,Y_k\}}Y_1, Y_{k-1} - P_{\{1,Y_1,Y_2,\ldots,Y_{k-1}\}}Y_k)$$

where $P_{\{1,Y_1,Y_2,\ldots,Y_k\}}Y_1$ and $P_{\{1,Y_1,Y_2,\ldots,Y_{k-1}\}}Y_k$ are the projections of $Y_1$ and $Y_k$ on the space spanned by $\{1, Y_1, Y_2, \ldots, Y_{k-1}\}$.
The value of $\alpha(k)$ is the partial autocorrelation at lag $k$. The partial autocorrelation therefore, corresponds to the correlation between the two sets of residuals obtained by regressing $Y_t$ and $Y_k$ on the intermediate observations $\{1, Y_1, Y_2, \ldots, Y_{k-1}\}$.
When dealing with real data the true $ACVF$, $ACF$ and $PACF$ are commonly unknown and they are replaced by their sample counterparts.

## 1.2 Common stochastic processes

In this section we will briefly present three of the most common stochastic processes that are often encountered in time series literature. These are the white noise process, the martingale process and the linear process.
The white noise process and the linear process represent the building blocks for the $ARMA$ models discussed in the next chapter.

### 1.2.1 White noise process

**Definition 7** *A sequence of random variables $\{Y_t\}$ is a white noise process with mean $\mu$ and variance $\sigma^2$, if for each $t \in \mathcal{T}$ and $k$*

1. $\mathbb{E}[Y_t] = \mu$
2. $\begin{cases} \gamma(0) = \sigma^2 \\ \gamma(k) = 0 \end{cases}$

A white noise sequence of random variables will be denoted by $Y_t \sim WN(\mu, \sigma^2)$.
Literally, a white noise sequence indicates a sequence of uncorrelated random variables with mean and variance constant over time. Assuming incorrelation is much weaker than assuming independence, as, independence imply incorrelation while the converse does not hold.
When a white noise process is also independent it is called an independent white noise process or strong white noise.

### 1.2.2 Martingale and martingale difference processes

Martingale are commonly encountered in time series when proving asymptotic results and theorems that involves dependence among the random variables.
To begin, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with additionally, $\{\mathcal{F}_t\}$, an increasing sequence of sub-$\sigma$-algebra of $\mathcal{F}$. $\{\mathcal{F}_t\}$ is defined a **filtration**.
Literally the filtration represent the information that we gain by observing the process for an increasing amount of time. If $t$ is interpreted as a discrete time index, then $\mathcal{F}_t$ contains the information up to time $t$.

**Definition 8** *A sequence $\{Y_t\}$ of random variables is $\mathcal{F}_t$ adapted if $Y_t$ is measurable with respect to $\mathcal{F}_t$ for all $t$. If $\mathcal{F}_t = \sigma\{Y_1, Y_2, \ldots, Y_t\}$ we define $\{\mathcal{F}_t\}$ as the* ***natural filtration*** *of the series.*

Having defined the filtration we can give a formal definition of martingale.

**Definition 9** *An $\mathcal{F}_t$ adapted sequence of random variables $\{Y_t\}$ is a martingale if*

$$\mathbb{E}[Y_t|\mathcal{F}_t] < \infty \quad \textit{for all} \quad t$$

*and*

$$\mathbb{E}[Y_{t+1}|\mathcal{F}_t] = Y_t \quad a.s \quad \textit{for all} \quad t.$$

In econometrics literature the $\sigma$-algebra $\mathcal{F}_t = \sigma\{Y_1, Y_2, \ldots, Y_t\}$ is commonly called the information set and heuristically it contains the information provided by the random variables $Y_1, Y_2, \ldots, Y_t$.
A white noise process is a particular case of martingale, however the converse it is not necessarily true. While in a martingale the first conditional moment is assumed constant over time, the other moments might be functions of time. Strictly connected to the martingale there is the **martingale difference**.

**Definition 10** *An $\mathcal{F}_t$ adapted sequence of random variables $\{Y_t\}$ is a martingale difference if*

$$\mathbb{E}[Y_t|\mathcal{F}_t] < \infty \quad \textit{for all} \quad t$$

*and*

$$\mathbb{E}[Y_{t+1}|\mathcal{F}_t] = 0 \quad a.s \quad \textit{for all} \quad t$$

### 1.2.3 Linear process

**Definition 11** *A zero mean process $\{Y_t\}$ is linear if it can be written as an infinite linear combination of white noise terms. That is*

$$Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \qquad \epsilon_t \sim WN(0, \sigma_\epsilon^2) \tag{1.10}$$

where for convention $\psi_0 = 1$.
The process is said linear because it does not contain mixed components of the form $\epsilon_{t-j}\epsilon_{t-i}$. The white noise process $\epsilon_t$ may be regarded as a series of shocks that drive the system.
When $\mathbb{E}[Y_t] = \mu$ a linear representation of the process can be obtained for $\tilde{Y}_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$, where $\tilde{Y}_t = (Y_t - \mu)$ is the deviation of the process from its mean.
Introducing the backshift operator $B$, defined as $B\epsilon_t = \epsilon_{t-1}$, and the linear filter associated to it $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ the process can be rewritten as

$$Y_t = \psi(B)\epsilon_t.$$

The importance of the linear process relies on the fundamental result of Wold (1938) who established the theorem.

**Theorem 4 (Wold Theorem)** *Any zero-mean purely non deterministic stationary process $Y_t$ can be represented as*

$$Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \qquad \epsilon_t \sim WN(0, \sigma_\epsilon^2)$$

*with $\sum_{j=0}^{\infty} \psi_j^2 < \infty$.*

The $\epsilon_t$ are uncorrelated with zero mean and common variance $\sigma_\epsilon^2$ but need not be independent. Throughout we will refer to linear processes for processes of the form (1.10) in which the $\epsilon_t$ are independent. A linear process is stationary and it properties may be expressed in terms of its first two moments.
The unconditional mean and the variance of the process are given by

$$\mathbb{E}[\sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}] = \sum_{j=0}^{\infty} \psi_j \mathbb{E}[\epsilon_{t-j}] = 0$$

$$\mathbb{V}(\sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}) = \sum_{j=0}^{\infty} \psi_j^2 \mathbb{V}(\epsilon_{t-j}) = \sigma_\epsilon^2 \sum_{j=0}^{\infty} \psi_j^2$$

In order order for the process to be stationary, a sufficient condition is that $\sum_{j=0}^{\infty} \psi^2 < \infty$, which ensures that the variance of the process is bounded.
The autocovariance function is

$$\begin{aligned}\gamma(k) &= \mathbb{E}[(\sum_{j=0}^{\infty} \psi_j \epsilon_{t-j})(\sum_{i=0}^{\infty} \psi_i \epsilon_{t-i-k})] \\ &= \sum_{i,j=0}^{\infty} \psi_j \psi_i \mathbb{E}[\epsilon_{t-j} \epsilon_{t-i-k}] \\ &= \sigma_\epsilon^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k}\end{aligned}$$

and the autocorrelation function is

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\sigma_\epsilon^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k}}{\sigma_\epsilon^2 \sum_{j=0}^{\infty} \psi_j^2} = \frac{\sum_{j=0}^{\infty} \psi_j \psi_{j+k}}{\sum_{j=0}^{\infty} \psi_j^2}.$$

Under suitable conditions the linear process $Y_t$ can be expressed as a weighted sum of past $Y_t$'s and an added shock $\epsilon_t$, that is

$$\begin{aligned}Y_t &= \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \cdots + \epsilon_t \\ &= \sum_{j=1}^{\infty} \pi_j Y_{t-j} + \epsilon_t \end{aligned} \tag{1.11}$$

In this alternative form the current value $Y_t$ or the current deviation $\tilde{Y}_t$ is "regressed" on the past values $Y_{t-1}, Y_{t-2}, \ldots$ of the process.
A linear process that can be represented as (1.11) is said to be **invertible**.

### Relationship between $\psi$ and $\pi$ weights

The relationship between the weights of the two forms can be established by using the backshift operator.
Given the process $Y_t = \psi(B)\epsilon_t$ where

$$\psi(B) = (1 + \sum_{j=1}^{\infty} \psi_j B^j) = \sum_{j=0}^{\infty} \psi_j B^j$$

we assume that the process is invertible and can be represented as

$$\epsilon_t = \pi(B)Y_t$$

where

$$\pi(B) = (1 - \sum_{j=1}^{\infty} \pi_j B^j) = \sum_{j=0}^{\infty} \pi_j B^j.$$

From the two representations the following equality can be derived

$$\psi(B)\pi(B)Y_t = \psi(B)\epsilon_t = Y_t.$$

Hence, $\psi(B)\pi(B) = 1$ that leads to the relationship

$$\pi(B) = \psi^{-1}(B).$$

This relationship can be used to recover the weights $\pi$ knowing the weights $\psi$ and vice versa.

**Stationarity and invertibility conditions for a linear process**

**Stationarity**: The necessary condition for a linear process to be stationary is that the series defined in (1.10) is convergent. This is guaranteed by the condition that $\sum_{j=0}^{\infty} |\psi_j| < \infty$. A sufficient condition for this to hold is that $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ which ensures that the variance of the process is bounded.
Alternatively the stationary condition can be embodied using the backshift operator. Given the polynomial $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ we require that the series is convergent for $|B| < 1$.

**Invertibility**: A similar restriction applied to the weights $\psi$ can be applied to the $\pi$ weights to ensure that the process is invertible, that is to say, that the representation (1.11) is admissible.
In order for the process to be invertible we require that the series $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$ converges for all $|B| < 1$, which is equivalent to the condition that $\sum_{j=0}^{\infty} |\pi_j| < \infty$.

# Chapter 2

# Linear stationary models

The aim of this chapter is to introduce the class of stationary linear time series models, which are used to describe the covariance structure in the time series.
There are three sub-groups of linear models, the autoregressive models, the moving average models and the mixed autoregressive-moving average models that are obtained as a combination of the first two. The existence and the importance of this three classes of models depends on the Wold theorem. In the following the time domain properties of these three classes of models will be presented and discussed.

## 2.1 Autoregressive models

The class of autoregressive models can be broadly considered an extension of the linear regression model, with the difference that instead of a set of explanatory regression variables the regression is performed on the lagged values $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$. The number of lagged variables used in the equation defines the order of the model.
In an autoregressive process of order $p$, the random variables separated by $p$ lags are conditionally independent.

### 2.1.1 Autoregressive model of order p

The autoregressive model of order $p$, shortly indicated with $AR(p)$, can be obtained from the linear representation (1.11) by setting the weights $\pi_j$, for $j > p$ equal to zero. Formally,

**Definition 12** *$\{Y_t\}$ is an autoregressive process of order $p$ if for every $t$,*

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t \tag{2.1}$$

*where* $\epsilon_t \sim \text{WN}(0, \sigma_\epsilon^2)$ *and* $\{\phi_0, \phi_1, \ldots, \phi_p\}$ *are unknown parameters.*

Using the backshift operator, the $AR(p)$ model can be written in the equivalent form

$$\phi(B)Y_t = \phi_0 + \epsilon_t \tag{2.2}$$

where $\phi(B) = (1 - \sum_{j=1}^{p} \phi_j B^j)$.

This implies the following relation for the model

$$\begin{aligned} \phi(B)Y_t = \phi_0 + \epsilon_t \Rightarrow Y_t &= \phi^{-1}(B)\phi_0 + \phi^{-1}(B)\epsilon_t \\ &= \frac{\phi_0}{1 - \sum_{j=1}^{p} \phi_j} + \psi(B)\epsilon_t, \end{aligned} \tag{2.3}$$

where $\phi^{-1}(B) = \psi(B)$.
Taking expectation of (2.3) the unconditional mean of the process is

$$\mathbb{E}[Y_t] = \mathbb{E}[\frac{\phi_0}{1 - \sum_{j=1}^{p} \phi_j} + \sum_{j=0}^{\infty} \phi_j \epsilon_{t-j}] = \frac{\phi_0}{1 - \sum_{j=1}^{p} \phi_j} = \mu.$$

Substituting $\phi_0 = \mu(1 - \sum_{j=1}^{p} \phi_j)$ in (2.1) and rearranging the terms, we obtain the mean centered representation of the process, that is,

$$(Y_t - \mu) = \psi(B)\epsilon_t \tag{2.4}$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$.

**Stationary and invertibility conditions for autoregressive processes**

In order for an autoregressive process to be stationary the parameters $\phi_1, \phi_2, \ldots, \phi_p$ must satisfy certain conditions. These stationarity conditions are generally derived from the roots of the polynomial in the lag operator $\phi(B)$.
The general $AR(p)$ model can be written as (2.4) provided that the series $\sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$ is convergent. By factorizing $\phi(B)$ as

$$\phi(B) = (1 - \lambda_1 B)(1 - \lambda_2 B) \ldots (1 - \lambda_p B)$$

where $\lambda_1^{-1}, \ldots, \lambda_p^{-1}$ are the roots of $\phi(B)$, and following Box et al. (2015), $\phi^{-1}(B)$ can be expanded in partial fractions, yielding

$$(Y_t - \mu) = \phi^{-1}(B) = \sum_{i=1}^{p} \frac{K_i}{1 - \lambda_i B} \epsilon_t.$$

Consequently, in order for $\phi^{-1}(B)$ to be a convergent series the weights $\psi_j = \sum_{i=1}^{p} K_i \lambda_i^j$ must be absolutely summable, so that we must have $|\lambda_i| < 1$ for $i = 1, \ldots, p$.
Equivalently, the roots of the polynomial $\phi(B) = 0$ must lie outside the unit circle, that is they must be greater than 1 in absolute value.
Equation $\phi(B) = 0$ is called the **characteristic equation** of the process and the roots of the characteristic equation are generally referred as the **zeros** of the autoregressive polynomial.
Alternatively, the stationarity conditions can be verified in terms of the polynomial

$$z^p - \phi_1 z^{p-1} - \cdots - \phi_p = 0.$$

By noting that the roots of $\phi(B)$ are the reciprocals of the roots of the polynomial in $z$, the stationarity condition that all the roots of $\phi(B) = 0$ must lie outside the unit circle is equivalent to the requirement that all the roots of $z^p - \phi_1 z^{p-1} - \cdots - \phi_p = 0$ lie inside the unit circle. Given that $\pi(B) = \phi(B) = (1 - \sum_{j=1}^{p} \phi_j B^j)$ is finite, an autoregressive process is always invertible.

### Autocovariance and autocorrelation function

To obtain the autocovariance of the process we consider the mean centered representation of the model, that is

$$(Y_t - \mu) = \sum_{j=1}^{p} \phi_j (Y_{t-j} - \mu) + \epsilon_t.$$

The autocovariance is then obtained multiplying the above equation by $(Y_{t-k} - \mu)$ and taking expectation. Doing so we obtain

$$\begin{aligned} \gamma(k) = \mathbb{E}[(Y_t - \mu)(Y_{t-k} - \mu)] &= \mathbb{E}\{[\sum_{j=1}^{p} \phi_j (Y_{t-j} - \mu) + \epsilon_t](Y_{t-k} - \mu)\} \\ &= \mathbb{E}[\sum_{j=1}^{p} \phi_j (Y_{t-j} - \mu)(Y_{t-h} - \mu) + \epsilon_t (Y_{t-h} - \mu)] \end{aligned} \tag{2.5}$$

where the expectation $\mathbb{E}[\epsilon_t(Y_{t-k} - \mu)]$ is zero for $k > 0$ since $(Y_{t-k} - \mu)$ can only involve

the socks up to time $t-k$.
Thus, the covariance function results equal to

$$\gamma(k) = \begin{cases} \sum_{j=1}^{p} \phi_j \gamma(k-j) & k > 0 \\ \sum_{j=1}^{p} \phi_j \gamma(j) + \sigma_\epsilon^2 & k = 0. \end{cases} \tag{2.6}$$

The autocorrelation is given by

$$\rho(k) = \sum_{j=1}^{p} \phi_j \rho(k-j) \quad k > 0. \tag{2.7}$$

Using the backshift operator, equation (2.7) can be rewritten as

$$\rho(k) = \sum_{j=1}^{p} \phi_j B^j \rho(k) \Rightarrow \phi(B)\rho(k) = 0$$

where $\phi(B) = (1 - \sum_{j=1}^{p} \phi_j B^j)$.
The process generates for the different lags a system of equations which is commonly known as the **Yule-Walker equations**. This system of equations can be solved to obtain the values of the autoregressive parameters starting from the autocorrelation function and vice versa.
Next, using the factorization

$$\phi(B) = (1 - \lambda_1 B)(1 - \lambda_2 B) \dots (1 - \lambda_p B)$$

it is possible to show (Box et al. (2015) for a proof) that the general solution for $\phi(B)\rho(k) = 0$ is

$$\rho(k) = A_1 \lambda_1^k + A_2 \lambda_2^k + \dots + A_p \lambda_p^k$$

where $A_i$'s are constants.
Depending on the value assumed by the roots $\lambda_i$ two situations can emerge.

1. A root $\lambda_i$ is real in that case the term $A_i \lambda_i$ decays to zero geometrically as $k$ increase. In literature this is referred to **damped exponential**.

2. A pair of roots $\lambda_i$ and $\lambda_j$ are complex conjugates, in this case their contribution follows a **damped sine wave**.

### 2.1.2 First-order autoregressive process

As a particular example we consider the $AR(1)$ model that together with the $AR(2)$ process represents two of the autoregressive models most adopted in practice. Moreover the results can be easily extended to processes of higher order.

**Definition 13** *$Y_t$ is an autoregressive process of order 1 if for every $t$,*

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t \tag{2.8}$$

*where $\epsilon_t \sim \mathrm{WN}(0, \sigma_\epsilon^2)$ and $\{\phi_0, \phi_1\}$ are unknown parameters.*

The unconditional mean of the process is

$$\begin{aligned} \mu = \mathbb{E}[Y_t] &= \mathbb{E}[\frac{\phi_0}{1-\phi_1} + \sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}] \\ &= \frac{\phi_0}{1-\phi_1} + \sum_{j=0}^{\infty} \phi_1^j \mathbb{E}[\epsilon_{t-j}] \\ &= \frac{\phi_0}{1-\phi_1} \end{aligned} \tag{2.9}$$

and $\phi_1$ has to satisfy the restriction $|\phi_1| < 1$ in order for the process to be stationary.

#### Autocovariance and autocorrelation function

The autocovariance function of the $AR(1)$ process is

$$\begin{aligned} \gamma(k) &= \mathbb{E}[(Y_t - \mu)(Y_{t-k} - \mu)] \\ &= \mathbb{E}[(\sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j})(\sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j-k})] \\ &= \sum_{j=0}^{\infty} \phi_1^j \phi_1^{j+k} \mathbb{E}(\epsilon_{t-j}^2) \\ &= \sigma_\epsilon^2 \phi_1^k \sum_{j=0}^{\infty} \phi_1^{2j} = \frac{\sigma_\epsilon^2}{1-\phi_1^2} \phi_1^k \end{aligned}$$

and setting $k$ equal to zero the variance results

$$\sigma^2 = \gamma(0) = \frac{\sigma_\epsilon^2}{1-\phi_1^2}.$$

Finally, the autocorrelation can be obtained as

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\phi_1^k \gamma(0)}{\gamma(0)} = \phi_1^k$$

Since $|\phi_1| < 1$, the autocorrelation decays exponentially to zero when $\phi_1$ is positive and oscillates in sign when $\phi_1$ is negative. In particular $\rho(1) = \phi_1$.

## 2.2 Moving-average processes

A moving-average process ($MA$) is obtained by setting, in the linear process (1.10), the parameters $\psi_j$, for $j > q$ equal to zero. The number of parameters, $q$, different from zero define the order of the moving average process process.
Conversely from autoregressive processes in a moving average process of order $q$, the random variables after $q$ lags are independent from each other. In practice moving average models of order 1 and 2 are the most commonly used.
We will start by considering the time domain properties of a generic moving average process of order $q$, $MA(q)$, and next we will state the results for the $MA(1)$ model.

### 2.2.1 Moving average process of order q

**Definition 14** *$Y_t$ is a moving average process of order $q$ with mean $\mu$ if for every $t$,*

$$(Y_t - \mu) = \epsilon_t - \theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2} - \cdots - \theta_q\epsilon_{t-q} \tag{2.10}$$

*where $\epsilon_t \sim \mathrm{WN}(0, \sigma_\epsilon^2)$ and $\{\theta_1, ..., \theta_q\}$ are unknown parameters.*

Using the backshift operator defined previously, the model can be written in the alternative form

$$(Y_t - \mu) = \theta(B)\epsilon_t,$$

where the polynomial $\theta(B)$ is equal to

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q = 1 - \sum_{j=1}^{q} \theta_j B^j$$

and for convention we might assume that $\theta_0 = -1$.
Rearranging the term and taking expectation of (2.10) the unconditional mean of the process results equal to $\mu$. That is,

$$\mathbb{E}[Y_t] = \mu + \mathbb{E}[\sum_{j=0}^{q} -\theta_j \epsilon_{t-j}] = \mu + \sum_{j=0}^{q} -\theta_j \mathbb{E}[\epsilon_{t-j}] = \mu.$$

**Stationarity and invertibility conditions for moving average processes**

A key difference between $AR$ and $MA$ processes is that for the $MA$ processes no restrictions must be imposed on the parameters to ensure stationarity. This results by noting that the series

$$\phi(B) = 1 - \theta_1 B - \theta_2 B^2 - \ \cdots - \theta_q B^q$$

is finite.
The condition of invertibility for a moving average process can be obtained by writing the process as

$$\epsilon_t = \theta^{-1}(B)(Y_t - \mu).$$

where $\theta^{-1}(B) = \pi(B)$.
Expanding $\theta^{-1}(B)$ in partial fractions we obtain

$$\theta^{-1}(B) = \sum_{i=1}^{q} \frac{M_i}{(1 - \lambda_i B)} \tag{2.11}$$

where again $\lambda_i$ are the roots of the polynomial $\theta^{-1}(B) = 0$.
The series defined in (2.11) converges if $|\lambda_i| < 1$, for $i = 1, 2, \ldots, q$. Thus, the invertibility condition of a moving average process of order $q$ is that the roots $\lambda_i^{-1}$ of the characteristic equation

$$\theta(B) = 1 - \theta_1 B - \theta_2 B - \cdots - \theta_q B$$

lie outside the unit circle or alternatively that all the roots of the polynomial

$$z^q - \theta_1 z^{q-1} - \cdots - \theta_q = 0,$$

lie inside the unit circle.
The weights $\pi_j$ of the linear representation can be obtained from the relation $\theta(B)\pi(B) = 1$. From the equality the weights $\pi_j$ satisfy the equation

$$\pi_j = \theta_1\pi_{j-1} + \theta_2\pi_{j-2} + \cdots + \theta_q\pi_{j-q},$$

with the convention that $\theta_0 = -1$ and $\pi_j = 0$ for $j < 0$. The weights $\pi_j$ can thus be recursively obtained in terms of the $\theta_i$.

**Autocovariane and autocorrelation function**

The autocovariance of the $MA(q)$ process is

$$\begin{aligned}\gamma(k) &= \mathbb{E}[(\sum_{j=0}^{q} -\theta_j\epsilon_{t-j})(\sum_{i=0}^{q} -\theta_i\epsilon_{t-i-k})] \\ &= \mathbb{E}(\sum_{i,j=0}^{q} \theta_j\theta_i\epsilon_{t-j}\epsilon_{t-i-k}) \\ &= \sum_{i=0}^{q} \theta_{i+k}\theta_i\mathbb{E}(\epsilon_{t-i-k}^2) \\ &= \sigma_\epsilon^2 \sum_{i=0}^{q-k} \theta_{i+k}\theta_i. \end{aligned} \tag{2.12}$$

Terms involving $\epsilon$'s at different lags have been dropped because they are uncorrelated, while $\theta_0$ is assumed to be equal to $-1$.
Setting $k = 0$ in (2.12) the variance of the process is

$$\gamma(0) = \sigma_\epsilon^2 \sum_{i=0}^{q} \theta_i^2.$$

while for $k > 0$, the autocovariance is

$$\gamma(k) = \sigma_\epsilon^2 \sum_{i=0}^{q-k} \theta_{i+k}\theta_i.$$

Finally, the autocorrelation is

$$\rho(k) = \frac{\sum_{i=0}^{q-k} \theta_i \theta_{i+k}}{\sum_{i=0}^{q} \theta_i^2} \quad k = 1, 2, \ldots, q$$

Therefore, the autocorrelation function of an $MA(q)$ process is zero after the lag $q$. Because of this, the autocorrelation function of a moving average process is generally said to **cut off after lag q**.

### 2.2.2 First-order moving average process

In many real applications the $MA(1)$ and $MA(2)$ processes provides good results. Moreover, the properties stated for a first-order moving average model can be easily extended to the $MA(2)$ model. Formally,

**Definition 15** *$Y_t$ is a moving average process of order 1 with mean $\mu$ if for every $t$,*

$$(Y_t - \mu) = \epsilon_t - \theta \epsilon_{t-1} \tag{2.13}$$

*where $\epsilon_t \sim WN(0, \sigma_\epsilon^2)$ and $\theta$ is an unknown parameter.*

In order for the process to be invertible the value of $\theta$ must satisfy $|\theta| < 1$. Whereas the process is stationary for all the values of $\theta$.

#### Autocovariance and autocorrelation functions

The autocovariance of the $MA(1)$ process is defined as

$$\gamma(k) = \mathbb{E}[(Y_t - \mu)(Y_{t-k} - \mu)] = \mathbb{E}[(\epsilon_t - \theta\epsilon_{t-1})(\epsilon_{t-k} - \theta\epsilon_{t-k-1})]$$

$$= \mathbb{E}(\epsilon_t \epsilon_{t-k} - \theta \epsilon_t \epsilon_{t-k-1} - \theta \epsilon_{t-1} \epsilon_{t-k} + \theta^2 \epsilon_{t-1} \epsilon_{t-k-1}). \tag{2.14}$$

From (2.14) the variance of the process results equal to

$$\gamma(0) = \mathbb{E}(\epsilon_t^2 - 2\theta\epsilon_t\epsilon_{t-1} + \theta^2\epsilon_{t-1}^2)$$

$$= \sigma^2 + \theta^2\sigma^2 = \sigma_\epsilon^2(1+\theta^2)$$

while the lag 1 autocovariance is

$$\gamma(1) = \mathbb{E}(\epsilon_t\epsilon_{t-1} - \theta\epsilon_t\epsilon_{t-2} - \theta\epsilon_{t-1}^2 + \theta^2\epsilon_{t-1}\epsilon_{t-2})$$

$$= -\theta\sigma_\epsilon^2. \tag{2.15}$$

Therefore, the behavior of the autocovariance function of the $MA(1)$ process can be summarized as

$$\gamma(k) = \begin{cases} \sigma_\epsilon^2(1+\theta^2) & k = 0 \\ -\theta\sigma_\epsilon^2 & k = 1 \\ 0 & k > 1. \end{cases}$$

From (2.14) and (2.15) the lag 1 autocorrelation is given by

$$\rho(1) = \frac{-\theta\sigma_\epsilon^2}{(1+\theta^2)\sigma_\epsilon^2} = \frac{-\theta}{(1+\theta^2)}$$

For $k > 1$ autocorrelations are of course equal to zero. As anticipated, the autocorrelation of the $MA(1)$ process cut off after lag 1 and is equal to

$$\rho(k) = \begin{cases} 1 & k = 0 \\ \frac{-\theta}{(1+\theta^2)} & k = 1 \\ 0 & k > 1. \end{cases}$$

The correlation of the $MA(1)$ process is completely determined by the value of $\theta$. Positive values of $\theta$ induces positive correlation, while negative values induce negative correlation. The maximum possible value of $\rho(1)$ is 0.5, which occurs if $\theta = 1$. The smallest value is $-0.5$, which occurs for $\theta = -1$.

For any value of $\rho(1)$ between $-0.5$ and $0.5$ there are two values of $\theta$ that can produce that autocorrelation. This is explained by noting that the value of $\theta/(1+\theta^2)$ is unchanged if $\theta$ is replaced by $1/\theta$. That is

$$\rho(1) = \frac{1/\theta}{1+(1/\theta)^2} = \frac{\theta^2(1/\theta)}{\theta^2[1+(1/\theta)^2]} = \frac{\theta}{\theta^2+1}.$$

## 2.3 Relationships between $AR$ and $MA$ processes

Having discussed the properties of the autoregressive and moving average processes it is possible to establish a duality between the two. In particular, an autoregressive process of order $p$ can be represented as an infinite weighted sum of past innovations, that is

$$(Y_t - \mu) = \phi^{-1}(B)\epsilon_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}.$$

Conversely, an invertible moving average process of order $q$ can be written as an infinite weighted sum of pasts $(Y_t - \mu)$, that is

$$\theta^{-1}(B)(Y_t - \mu) = \sum_{j=0}^{\infty} \pi_j (Y_{t-j} - \mu) = \epsilon_t.$$

Thus, provided that the series $\theta^{-1}(B)$ and $\phi^{-1}(B)$ are convergent, the above duality allows to approximate a moving average process with an autoregressive one of infinite order, and vice versa.

This duality is reflected also in the structure of the autocorrelation and the partial autocorrelation functions of the two processes. The autocorrelation function of the $MA(q)$ process is zero beyond $q$ lags, but since it is equivalent to an infinite $AR$ process, its partial autocorrelation function is infinite and it is characterized by damped sine and cosine waves. On the other hand, the $AR(p)$ process has a partial autocorrelation that is zero after $p$ lags but its autocorrelation consists of a mixture of damped exponentials and/or sine waves.

## 2.4 Mixed autoregressive-moving average processes

Autoregressive and moving average processes can be unified together in a more general model, the mixed autoregressive moving average process ($ARMA$). Its attractiveness and wide use is given by its ability to describe a wide variety of situations with a very parsimonious number of parameters. That is to say, an $ARMA$ process of a low order is generally able to describe an observed time series better, than an autoregressive or a moving average process of higher order.
The number of autoregressive and moving average components in the equation determine the order of the $ARMA$ process.

**Definition 16** *$Y_t$ is a autoregressive-moving average process of order $(p, q)$, if for every $t$*

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + .. + \phi_p Y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q} \tag{2.16}$$

*where $\epsilon_t \sim \mathrm{WN}(0, \sigma_\epsilon^2)$ and $\{\phi_0, \phi_1, \ldots, \phi_p, \theta_1, \theta_2 \ldots, \theta_q\}$ are unknown parameters. It will be written $ARMA(p, q)$.*

In terms of the backshift operator the model can be specified as

$$\phi(B) Y_t = \phi_0 + \theta(B)\epsilon_t, \tag{2.17}$$

where

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p, \quad \text{and} \quad \theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q.$$

By taking expectation of (2.16) and assuming stationarity the unconditional mean of the process is

$$\begin{aligned}
\mu &= \mathbb{E}[\phi_0 + \phi_1 Y_{t-1} + .. + \phi_p Y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q}] \\
\mu &- \sum_{j=1}^{p} \phi_j B^j \mathbb{E}[Y_t] = \phi_0 \\
\mu &- (1 - \sum_{j=1}^{p} \phi_j))\phi_0 \\
\mu &= \frac{\phi_0}{1 - \sum_{j=1}^{p} \phi_j}
\end{aligned}$$

which corresponds to the mean of an autoregressive process of order $p$.
Substituting $\phi_0 = \mu(1 - \sum_{j=1}^{p} \phi_j)$ in (2.16) we obtain the mean centered version of the process, that leads to the following formulation,

$$\phi(B)(Y_t - \mu) = \theta(B)\epsilon_t.$$

It possible to think of the $ARMA(p, q)$ process in two ways:

1. As a $p$th order autoregressive process where $u_t$ follows a $q$th order moving average process, that is

   $$\phi(B)(Y_t - \mu) = u_t$$

   with $u_t = \theta(B)\epsilon_t$

2. As a $q$th order moving average process where $a_t$ follows an autoregressive process of order $p$, that is

   $$(Y_t - \mu) = \theta(B)a_t$$

   where $\phi(B)a_t = \epsilon_t$, so that $\phi(B)(Y_t - \mu) = \theta(B)\phi(B)a_t = \theta(B)\epsilon_t$.

**Stationarity and invertibility conditions for the autoregressive-moving average process**

In order for the model (2.16) to be stationary the roots of $\phi(B) = 0$ must lie outside the unit circle, which is identical to the condition encountered for an autoregressive process to be stationary. Likewise, the roots of $\theta(B) = 0$ must lie outside the unit circle if the process is to be invertible.
When both these constraints are satisfied the stationary and invertible $ARMA(p, q)$ process has both an infinite moving average representation given by

$$(Y_t - \mu) = \psi(B)\epsilon_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j},$$

where $\psi(B) = \phi^{-1}(B)\theta(B)$, and an infinite autoregressive representation, that is

$$\pi(B)(Y_t - \mu) = \sum_{j=0}^{\infty} \pi_j (Y_{t-j} - \mu) = \epsilon_t,$$

where $\pi(B) = \theta^{-1}(B)\phi(B)$.

The weights $\psi_j$ of the autoregressive representation can be determined from the relation $\phi(B)\psi(B) = \theta(B)$ while the weights $\pi_j$ can be obtained from the relation $\theta(B)\pi(B) = \phi(B)$.

**Autocovariance and autocorrelation functions for autoregressive-moving average processes**

The $ARMA(p, q)$ model defined in (2.16) can be rewritten in the mean centered version, that is

$$(Y_t - \mu) = \sum_{j=1}^{p}(Y_{t-j} - \mu) - \sum_{i=0}^{q}\theta_i\epsilon_{t-i}.$$

where $\theta_0 = -1$.
The autocovariance function can then be obtained multiplying by $(Y_{t-k} - \mu)$ and taking expectation

$$\begin{aligned}\gamma(k) &= \mathbb{E}[(Y_t - \mu)(y_{t-k} - \mu)] = \mathbb{E}[\sum_{j=1}^{p}\phi_j(Y_{t-j} - \mu) - \sum_{i=0}^{q}\theta_i\epsilon_{t-i}](Y_{t-k} - \mu) \\ &= \mathbb{E}[\sum_{j=1}^{p}\phi_j(Y_{t-j} - \mu)(Y_{t-k} - \mu)] - \mathbb{E}[\sum_{i=0}^{q}\theta_i\epsilon_{t-i}(Y_{t-k} - \mu)] \\ &= \sum_{j=1}^{p}\phi_j\gamma(k-j) - \underbrace{\mathbb{E}[\sum_{i=0}^{q}\theta_i\epsilon_{t-i}(Y_{t-k} - \mu)]}_{C}.\end{aligned}$$

Next, replacing $(Y_{t-k} - \mu)$ in $C$ with its moving average representation, $Y_{t-k} = \psi(B)\epsilon_{t-k}$, we obtain

$$\mathbb{E}[\sum_{i=0}^{q}\theta_i\epsilon_{t-i}\sum_{j=0}^{\infty}\psi_j\epsilon_{t-k-j}].$$

By the properties of white noise processes it follows that $\epsilon_t$ at different lags are uncorrelated and the preceding equation for $\gamma(k)$ becomes

$$\gamma(k) = \sum_{j=1}^{p}\phi_j\gamma(k-j) - \sigma_\epsilon^2\sum_{i=k}^{q}\theta_i\psi_{i-k}.$$

Hence, the autocovariance function is

$$\gamma(k) = \begin{cases} \sum_{j=1}^{p} \phi_j \gamma(k-j) - \sigma_\epsilon^2 \sum_{i=k}^{q} \theta_i \psi_{i-k} & k \geq p+1 \\ \sum_{j=1}^{p} \phi_j \gamma(j) - \sigma_\epsilon^2 \sum_{i=0}^{q} \theta_i \psi_i & k = 0 \\ \sum_{j=1}^{p} \phi_j \gamma(k-j) & k \geq q+1. \end{cases} \tag{2.18}$$

(2.18) implies that the autocorrelation is

$$\rho(k) = \begin{cases} \sum_{j=1}^{p} \phi_j \rho(k-j) - \frac{\sigma_\epsilon^2 \sum_{i=k}^{q} \theta_i \psi_{i-k}}{\gamma(0)} & k \geq p+1 \\ \sum_{j=1}^{p} \phi_j \rho(k-j) & k \geq q+1 \end{cases}$$

and hence $\phi(B)\rho(k) = 0$ for $k \geq q+1$.
Therefore, for the $ARMA(p,q)$ process beyond a certain number of lags, that is $q-p$ the $ACF$ displays the same shape of that of an $AR(p)$ process.
In particular if $p > q$ the whole autocorrelation function $\rho(k)$ for $k = 0, 1, 2, \ldots$ will consist of a mixture of damped exponentials and/or damped sine waves whose structure is defined by the roots of the polynomial $\phi(B)$. Conversely if $q \geq p$ there will be $q-p+1$ initial values of $\rho(k)$ that do not follow this general pattern.

# Chapter 3

# Linear non stationary models

The class of $ARMA$ models introduced in the previous chapter to describe stationary time series relies on the assumption that the underlying process is weakly stationary, that is, that the mean and the variance are not functions of time and that the autocovariance is a function only of the time lag. However, many time series encountered in practice are not stationary and exhibits time-changing means and/or variances. It is hence necessary to introduce further statistical techniques that are used when the stationarity assumption is not met. In literature there are several approaches to model non stationary time series, here we will follow the Box Jenkins approach, which relies on the $ARIMA$ and $SARIMA$ models. These two further class of models represent a direct extension of the $ARMA$ models to non stationary time series and therefore fit well in our purpose of extending the Forward Search to time series models.
We will start by considering the two different cases of non-stationarity, non-stationarity in variance, and non-stationarity in mean. Next we will introduce the class of $ARIMA$ and $SARIMA$ models introduced to deal with non-stationary time series through differencing.

## 3.1 Non stationarity in variance

We begin by considering a time series that is non stationary in variance, i.e the variance $\sigma_Y^2$ of the process is a function of time.
We assume that the time series is generate by a process $Y_t$ that can be decomposed into a function that describes the behavior of the mean over time plus a random error component with mean 0

$$Y_t = \mu_t + \epsilon_t. \tag{3.1}$$

Moreover, we assume that the variance of the innovations $\epsilon_t$, is related to the mean level

$\mu_t$ by some known function $h(\cdot)$, that is

$$\sigma_Y^2 = \mathbb{V}(\epsilon_t) = h^2(\mu_t)\sigma_\epsilon^2$$

The objective in this case, is to find a transformation of the data $g(Y_t)$ that will stabilize the variance over time. That is, the variance of the transformed variable $g(Y_t)$ should be approximately constant over $t$.
Expanding $g(Y_t)$ as a first-order Taylor expansion around the mean $\mu_t$ yields

$$g(Y_t) \approx g(\mu_t) + (Y_t - \mu_t)g'(\mu_t).$$

Next, it is possible to approximate the variance of $g(Y_t)$ as

$$\mathbb{V}[g(Y_t)] = \mathbb{V}[g(\mu_t) + (Y_t - \mu_t)g'(\mu_t)]$$

$$= [g'(\mu_t)]^2\mathbb{V}(Y_t)$$

$$= [g'(\mu_t)]^2h^2(\mu_t)\sigma_\epsilon^2.$$

Hence, in order to stabilize the variance it is necessary to chose the transformation $g(\cdot)$ such that,

$$g'(Y_t) = h^{-1}(\mu_t)$$

In real applications, when the series exhibits a non stationary behavior, it customary to transform it taking the natural logarithm or the square root. The natural logarithm corresponds to the case when $Y_t$ is proportional to its level, that is $h(\mu_t) = \mu_t$. The variance-stabilizing transformation has then to satisfy $g'(\mu_t) = \mu_t^{-1}$. This implies that $g(\mu_t) = \log(\mu_t)$. Alternatively, when $h(\mu_t) = \mu_t^{1/2}$ the transformation has to satisfy $g'(\mu_t) = \mu_t^{-1/2}$. Thus, since $g(\mu_t) = 2\mu_t^{1/2}$ the transformation $\sqrt{Y_t}$ is a proper variance-stabilizing transformation.
These two examples are particular cases of the class of power transformations introduced by Box and Cox (1964)

$$g(Y_t) = \frac{Y_t^\lambda - 1}{\lambda} \tag{3.2}$$

where the logarithm transformation is obtained by $\lim_{\lambda\to 0}[(Y_t^\lambda - 1)/\lambda] = \log(Y_t)$, while

the square root is obtained by setting $\lambda = 1/2$ in (3.2).

## 3.2 Non stationarity in mean

The non constant mean $\mu_t$ in equation (3.1) can be modeled in many different ways.
It is customary to extend the representation (3.1) to

$$Y_t = L_t + S_t + u_t, \tag{3.3}$$

where $\mu_t = L_t + S_t$ and $u_t$ is a stationary invertible zero mean process.
$L_t$ is a slowly varying function known as the **trend component** while $S_t$ is a function with known period $s$ known as the **seasonal component**.
When the series does not exhibits a seasonal pattern (3.3) simply reduces to

$$Y_t = L_t + u_t. \tag{3.4}$$

We will first consider the case when the underlying process of the series can be represented as (3.4), i.e no seasonal component is observed and the mean of the process corresponds to the trend.
Moreover, restricting the attention to the model (3.4) two cases can emerge:

1. The component $L_t$ is a **deterministic trend**.
2. The component $L_t$ is a **stochastic trend**.

### 3.2.1 Deterministic trend

When the series is non stationary in mean and the trend $L_t$ is assumed to be a deterministic function of time, $g(t)$, the general approach followed is to **detrend** the series.
In order to detrend the series it is necessary to assume a particular model for $g(t)$. If $g(t)$ is represented as a linear function of time, that is

$$g(t) = \alpha + \beta t,$$

the process $Y_t$ becomes

$$Y_t = \alpha + \beta t + u_t$$

and the trend component $L_t$ is a **linear deterministic** trend (increasing in $t$ for $\beta > 0$).
The process can be easily extended to cases where the deterministic trend is a nonlinear function of time. As an example by using the polynomial function $g(t) = \alpha + \sum_{j=1}^{p} \beta_j t^j$ it is

possible to specify a **polynomial** trend. Likewise it is possible to define an **exponential** trend, $g(t) = \exp(\alpha+\beta t)$. In the majority of real applications polynomials of order greater than two are rarely used. Once a proper function $g(t)$ is identified to describe the trend the parameters it can be fitted by standard OLS method and subsequently subtracted from the process $Y_t$ such that an $ARMA$ model can be used to describe the remaining stationary component $u_t$.
In particular by the Wold theorem the stochastic component $u_t$ can be represented as $u_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} = \psi(B)\epsilon_t$ and the process $Y_t$ has the form

$$\begin{aligned} Y_t &= g(t) + \psi(B)\epsilon_t \\ &= g(t) + \epsilon_t + \psi_1\epsilon_{t-1} + \psi_2\epsilon_{t-2} + \dots \end{aligned} \tag{3.5}$$

where $\epsilon_t \sim WN(0, \sigma_\epsilon^2)$.
Equation (3.5) implies the following results for the process $Y_t$

$$\mathbb{E}(Y_t|\mathcal{F}_{t-1}) = g(t) + \epsilon_t + \psi_1\epsilon_{t-1} + \psi_2\epsilon_{t-2} + \dots$$

$$\mathbb{V}(Y_t|\mathcal{F}_{t-1}) = \mathbb{V}(\epsilon_t) = \sigma_\epsilon^2.$$

Then, once the estimated trend component $g(t)$ is removed from the series we refer at the difference $Y_t - g(t)$ as the detrended series.

### 3.2.2 Stochastic trend

As an alternative to the trend stationary representation (3.5) we assume that the process has the following expression

$$Y_t = Y_{t-1} + g(t) + u_t. \tag{3.6}$$

The value of the process at time $t$ is equal to the value of the process at previous time plus a deterministic component $g(t)$ and a stationary invertible zero mean process.
By noting that the process $g(t) + u_t$ is trend stationary, the process defined in (3.6) can be reduced to a stationary process by taking the first difference, that is

$$\Delta Y_t = Y_t - Y_{t-1} = g(t) + u_t$$

where $\Delta$ is called the **difference operator** and is defined as $\Delta = (1 - B)$. Processes that can be made stationary after differencing the series are called **difference stationary** processes. The two most common examples of stochastic trend process are the **random walk** process and the **random walk with drift** process We will briefly consider these two processes

### Random Walk process

The simplest stochastic trend model can be obtained by setting in equation (3.6) $g(t) = 0$ and $u_t = \epsilon_t$, where $\epsilon_t \sim WN(0, \sigma_\epsilon^2)$, to obtain

$$Y_t = Y_{t-1} + \epsilon_t.$$

By repeated substitution up to time 0 we obtain

$$Y_t = y_0 + \sum_{j=1}^{t} \epsilon_j. \tag{3.7}$$

The process given in equation (3.7) is defined as the **random walk** process.
Conditionally on the starting value $Y_0 = y_0$ which is assumed to be deterministic, the mean and the variance of the process are given by

$$\mathbb{E}(Y_t|y_0) = y_0 + \sum_{j=1}^{t} \mathbb{E}(\epsilon_j) = y_0$$

$$\mathbb{V}(Y_t|y_0) = \sum_{j=1}^{t} \mathbb{V}(\epsilon_j) = t\sigma_\epsilon^2$$

and the process is not stationary.
Likewise, we can consider the ACVF of the process, that is

$$\gamma(k) = \mathbb{E}[(Y_t - y_0)(Y_{t-k} - y_0)]$$

$$= \mathbb{E}(\sum_{j=1}^{t} \epsilon_j \sum_{j=1}^{t-k} \epsilon_j) = \sum_{j=1}^{t-k} \mathbb{E}(\epsilon_j^2)$$

$$= (t-k)\sigma_\epsilon^2$$

Finally, dividing the ACVF by the variance of the process we obtain the autocorrelation function

$$\rho(k) = \frac{(t-k)\sigma^2}{t\sigma_\epsilon^2} = \frac{t-k}{t}. \tag{3.8}$$

Letting $t \to \infty$ the $ACF$ (3.8) has limit equal to 1. This means that the process is asymptotically perfectly correlated and as a consequence the memory of the process is infinite.
A further interpretation of the random walk process can be obtained by considering an $AR(1)$ process

$$(1 - \phi B)Y_t = \epsilon_t.$$

The random walk process can be obtained by setting $\phi = 1$. Next, solving the characteristic equation $\phi(z) = 1 - z = 0$ we obtain that the characteristic root is $z = 1$ which lies on the unitary circle, and therefore the process is not stationary. This fact it is generally expressed by saying that the process is a **unit-root process**.

### Random walk with drift

By setting $g(t) = \mu$ in (3.6) we obtain the **random walk with drift**

$$Y_t = \mu + Y_{t-1} + \epsilon_t.$$

Assuming that the first term $y_0$ is constant and by repeating substitution the random walk with drift process can be rewritten as

$$
\begin{aligned}
Y_t &= \mu + (\mu + Y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\
&\vdots \quad \text{keep substituting} \\
&= y_0 + t\mu + \sum_{j=1}^{t} \epsilon_j,
\end{aligned}
$$

where $y_0 + t\mu$ is a deterministic linear trend component. In addition to the deterministic trend the term $\sum_{j=1}^{t} \epsilon_j$ is generally referred as **stochastic trend** which determines the non stationary part of the process.
Conditioning on $y_0$ the mean and variance of the process are given by

$$
\begin{aligned}
\mathbb{E}(Y_t|y_0) &= y_0 + t\mu, \\
\mathbb{V}(Y_t|y_0) &= t\sigma_\epsilon^2,
\end{aligned}
$$

and both diverges to infinity as the time $t$ increase. Therefore, we can conclude that the process is non stationary in mean and in variance.

## 3.3 Autoregressive Integrated Moving average processes

The models discussed so far have all been introduced with the assumption that the underlying process is stationary. That is to say that the roots of the polynomial $\phi(B)$ lie outside the unit circle.
As introduced previously many empirical time series are non stationary. Besides non stationarity in mean or in variance Box and Jenkins introduce the concept of **homogeneous stationarity** to refers to time series that apart from local level, one part of the series behaves much like any other part. Processes that describe this type on non stationarity can be obtained by assuming that some suitable difference $d$ of the process is stationary. Processes where the $d$th difference of the series is stationary are called autoregressive integrated moving average processes ($ARIMA$).
In order to introduce this class of processes lets consider the model defined as

$$
\varphi(B)(Y_t - \mu) = \theta(B)\epsilon_t \tag{3.9}
$$

where $\varphi(B)$ is a non stationary autoregressive operator, defined in such a way that $d$ roots of $\varphi(B) = 0$ are unity and the reminder lies outside the unit circle.

The model can then be written as

$$\varphi(B)(Y_t - \mu) = \phi(B)(1 - B)^d(Y_t - \mu) = \theta(B)\epsilon_t.$$

By writing $(1 - B)^d$ in terms of the difference operator $\nabla^d$, defined as $\nabla^d(Y_t - \mu) = (1 - B)^d(Y_t - \mu)$ we obtain that

$$\phi(B)\nabla^d Y_t = \theta(B)\epsilon_t$$

or setting $\upsilon_t = \nabla^d Y_t$

$$\phi(B)\upsilon_t = \theta(B)\epsilon_t.$$

The model, therefore corresponds to assume that the $d$th difference of the process can be represented by an invertible stationary $ARMA$ process.

**Definition 17** *$Y_t$ is an autoregressive integrated moving average process of order $(p, d, q)$, if for every $t$*

$$\phi(B)\nabla^d Y_t = \theta(B)\epsilon_t \tag{3.10}$$

*where $\nabla^d Y_t$ is a stationary process and $\epsilon_t \sim WN(0, \sigma_\epsilon^2)$.*

In what follows we will assume that the process is invertible and causal. In terms of the backshift polynomials that is equivalent to say that

1. The roots $\phi(B) = 0$, where $\phi(B)$ is defined the autoregressive operator, lie outside the unit circle.
2. $d$ of the roots of $\varphi(B) = 0$ are equal to unity, where $\varphi(B)$ is defined the generalized autoregressive operator.
3. The roots of $\theta(B) = 0$ lie outside the unit circle, where $\theta(B)$ is defined the moving average operator.

Of course, when the order of integration $d$ is equal to 0 the process reduces to an autoregressive moving average model.
An extended representation of model (3.10) can be obtained by expanding the autoregressive operator. Thus if,

$$\varphi(B) = \phi(B)(1 - B)^d = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_{p+d} B^{p+d},$$

model (3.10) can be written as

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_{p+d} Y_{t-p-d} + \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q}.$$

### 3.3.1 Stochastic and Deterministic Trends

Sometimes it is useful to consider an extension of the standard $ARIMA$ model presented in (3.10) by adding a constant term $\theta_0$, that is

$$\phi(B)\nabla^d Y_t = \theta_0 + \theta(B)\epsilon_t.$$

When the constant term $\theta_0$ is omitted, the model, is able to represent time series that exhibits a stochastic trend characterized by random changes in the level and in the slope of the series.
Conversely, by allowing $\theta_0$ to be non-zero we automatically allow for the presence a deterministic polynomial trend of degree $d$. Allowing $\theta_0$ to be nonzero is equivalent to

$$\mathbb{E}(\nabla^d Y_t) = \mathbb{E}(\upsilon) = \mu_\upsilon = \frac{\theta_0}{1 - \phi_1 - \phi_2 - \cdots - \phi_p},$$

to be nonzero. In many real applications, where the reason for a deterministic component in the trend does not exists, the mean of $\upsilon$ is assumed to be equal to zero. This is equivalent to the assumption that in the data is present a stochastic trend.

### 3.3.2 Different forms of the $ARIMA$ model

Like the class of $ARMA$ model the $ARIMA$ model can be represented in three different forms. Each of these allows to consider and appreciate some special aspects of this class of models. In particular, the current value $Y_t$ of the process can be represented as:

1. In term of the pas values of $Y$'s and previous values of the innovations $\epsilon$'s, like in equation (3.10).
2. In terms of the current and past innovations $\epsilon_t$
3. In terms of a weighted sum of previous values $Y_{t-k}$ of the process and the current shock $\epsilon_t$.

#### Random shock form

As we have specified in chapter 1 a process is linear if it can be represented as (1.10), that is as a linear combination of white noise terms $\epsilon_t$.

However, as pointed out by Box et al. (2015), "since the non stationary $ARIMA$ models are not in statistical equilibrium over time, they cannot be assumed to extend infinitely into the past, and hence an infinite linear representation will not be possible". But a related finite truncated form , which will be discussed subsequently always exists.
In order to find the linear representation we start from the linear process (1.10) and operate on it with the operator $\varphi(B)$ we obtain

$$\varphi(B)Y_t = \varphi(B)\psi(B)\epsilon_t.$$

Next, noting that $\varphi(B)Y_t = \theta(B)\epsilon_t$, we obtain

$$\varphi(B)\psi(B) = \theta(B).$$

Therefore, the weights $\psi_j$ of an $ARIMA$ process can be obtained by equating the coefficients of $B$ in

$$(1 - \varphi_1 B - \cdots - \varphi_{p+d}B^{p+d})(1 + \phi_1 B + \phi_2 B^2 + \dots) = (1 - \theta_1 B - \cdots - \theta_q B^q)$$

and finally the weights $\psi_j$ can be determined recursively through the equations

$$\psi_j = \varphi_1\psi_{j-1} + \varphi_2\psi_{j-2} + \cdots + \varphi_{p+d}\psi_{j-p-d} \quad j > 0$$

with $\psi_0 = 1$, $\psi_j = 0$ for $j < 0$ and $\theta_j = 0$ for $j > q$.
It is possible to prove that the form $Y_t = \sum_{j=0}^{\infty}\psi_j\epsilon_t$ is strictly non convergent for an $ARIMA$ model, that is to say the weights $\psi_j$ are not absolutely summable. Nevertheless, a truncated version of the linear form is always valid.
For a detailed explanation of the truncated form see Box et al. (2015).

### Inverted form of the model

As previously presented for the class of $ARMA$ models a linear model $(Y_t - \mu) = \psi(B)\epsilon_t$ can also be represented in its inverted form

$$\psi^{-1}(B)(Y_t - \mu) = \epsilon_t \tag{3.11}$$

that is equal to

$$\pi(B)(Y_t - \mu) = (1 - \sum_{j=1}^{\infty} \pi_j B^j)(Y_t - \mu) = \epsilon_t.$$

The weights of the $ARIMA$ model can be derived by substituting (3.11) in (3.10), to obtain

$$\varphi(B)Y_t = \theta(B)\pi(B)Y_t.$$

Next, the weights $\pi$ can be be obtained by equating the coefficients of $B$ in $\varphi(B) = \theta(B)\pi(B)$, that is,

$$(1 - \varphi_1 B - \cdots - \varphi_{p+d} B^{p+d}) = (1 - \theta_1 B - \cdots - \theta_q B^q)(1 - \pi_1 B - \pi_2 B^2 - \dots).$$

Finally, the $\pi$ weights can be obtained recursively solving

$$\pi_j = \theta_1 \pi_{j-1} + \theta_2 \pi_{j-2} + \cdots + \theta_q \pi_{j-q} + \varphi_j \quad j > 0$$

with $\pi_0 = -1$, $\pi_j = 0$ for $j < 0$ and $\varphi_j = 0$ for $j > p + d$.

## 3.4 Seasonal Integrated Autoregressive Moving Average process

The class of $ARIMA$ processes has been developed with the purpose of modeling processes of the type (3.4), when the the underlying stochastic process can be represented by (3.3) the class of $SARIMA$ processes can be used.
Time series of the type (3.3) exhibits a trend and a seasonal pattern. Examples of seasonal data are sales data or geological data, such as temperature data.
In general a time series with a seasonal component exhibits a periodic behavior with period $s$, that is to say, similarities in the series occur after $s$ time intervals.
The direct extension of the $ARIMA$ models to handle seasonal time series is given by the $SARIMA$ models.
The fundamental aspect about seasonal time series with period $s$ is that observations located at $s$ intervals apart are similar.
When a series exhibits a certain seasonal behavior we expect that the relationships between the observations occur:

- Between observations inside the same period.
- Between the observations located in two different periods.

In order to link $Y_t$ to $Y_{t+s}$, that is the between periods relationships, a model of the following form is adopted

$$\Phi(B^s)\nabla_s^D Y_t = \Theta(B^s)u_t$$

where $\nabla_s^D = (1-B^s)(1-B)^D$ and $\Phi(B^s)$,$\Theta(B^s)$ are polynomials in $B^s$ of degree $P$ and $Q$ that satisfy the invertibility and stationarity conditions postulated in the previous chapter.
The error components $u_t$ will in general be correlated, that is $u_t$ will in general be correlated with $u_{t-1}, u_{t-2}$ and so on. In order to account for this relationships, i.e the within periods relationships, the model

$$\phi(B)\nabla^d u_t = \theta(B)\epsilon_t,$$

is introduced.
Here $\epsilon_t$ is assumed to be a white noise process, and $\phi(B)$ and $\theta(B)$ are polynomials in $B$ of degrees $p$ and $q$ that satisfy the invertibility and stationarity conditions and are referred as the non seasonal polynomials.
Substituting the new defined model in the previous one we obtain the general seasonal multiplicative model

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D Y_t = \theta_q(B)\Theta_Q(B^s)\epsilon_t$$

Formally

**Definition 18** *$Y_t$ is a seasonal autoregressive integrated moving average process of order $(p,d,q)\times(P,D,Q)_s$, if for every $t$*

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D Y_t = \theta_q(B)\Theta_Q(B^s)\epsilon_t \tag{3.12}$$

*where $\nabla^d\nabla_s^D Y_t$ is a stationary process and $\epsilon_t \sim WN(0,\sigma_\epsilon^2)$*

### 3.4.1 The relation between structural component and $ARIMA$ models

In general a structural component time series model relies on the assumption that the process $\{Y_t\}$ can be formulated as (3.3), that is as the sum of three latent components; a trend, a seasonal and a noise component.
Even if the latent components are unobservable and cannot be specified in a unique way they have a clear and meaningful interpretation.
In order to illustrate the nature of a structural time series model we consider the basic structural model generally abbreviated a BSM as formulate by Harvey (1990).
The model is defined as

$$Y_t = L_t + S_t + \epsilon_t \quad \epsilon_t \sim WN(0,\sigma_\epsilon^2)$$

where $L_t$ is the trend term and follows the local linear trend model

$$\begin{aligned} L_t &= L_{t-1} + \beta_{t-1} + \eta_t \quad \eta \sim WN(0,\sigma_\eta^2) \\ \beta &= \beta_{t-1} + \xi_t \quad \xi_t \sim WN(0,\sigma_\xi^2) \end{aligned}$$

and $S_t$ the seasonal component follows the dummy variable seasonal model, that is

$$(1 + B + B^2 + \cdots + B^{s-1})S_t = \omega_t \quad \omega_t \sim WN(0,\sigma_\omega^2)$$

where $\eta_t, \xi_t, \omega_t$ and $\epsilon_t$ are mutually uncorrelated.
The random component $\eta_t$ in the trend model allows the level of the trend to shift up and down, while the random component $\xi_t$ allows the slope to change in a stochastic way.

As limiting cases when $\sigma_\xi^2 = 0$ then $\beta_t = \beta_{t-1}$ and so the slope is a fixed constant for all $t$, and the trend reduces to a random walk with drift $(1-B)L_t = \beta + \eta_t$. Conversely, if $\sigma_\eta^2 = 0$ the linear trend model reduces to the deterministic trend model $L_t = L_{t-1} + \beta$.
The dummy variable seasonal model require that the seasonal effects $S_t$ sum to zero over $s$ consecutive values, subject to a random disturbance with mean zero that allows the seasonal effects to change gradually over time.
Again, a special limiting case can be obtained when $\sigma_\omega^2 = 0$. In this case $S_t = S_{t-1}$ and $S_t + S_{t-1} + \cdots + S_{t-s+1} = 0$.
The main advantage of the BSM model is that it generalizes a regression model where the trend is represented by a fixed straight line and the seasonality by a fixed seasonal effect using the indicator variables, by allowing the trend and seasonality to vary over time.
Structural models like the BSM model have an equivalent $ARIMA$ representation. In this particular case, the local linear linear trend model satisfies

$$(1-B)^2 L_t = (1-B)\beta_{t-1} + (1-B)\eta_t = \xi_{t-1} + (1-B)\eta_t$$

Now, following Box et al. (2015) p.333 it can be proved that $\xi_{t-1} + (1-B)\eta_t$ can be represented as an $MA(1)$ process $(1-\theta B)a_t$ so that $(1-B)^2 L_t = (1-\theta B)a_t$ and $L_t$ has an $ARIMA(0,2,1)$ representation.
Sometimes the equivalent $ARIMA$ representation of a structural model is referred to as its reduced form. In this thesis we will use the term reduced form with a different meaning, that is to indicate the most parsimonious representation of a given model in its state space form.

# Chapter 4

# State space form and the Kalman filter

In this chapter we will present the state space representation of a linear model and the connected Kalman filter.
The Kalman filter in the context of the Forward Search is particularly important for two reasons, first it provides a very efficient way to compute the maximum likelihood estimates in the presence of several missing units in the series. Second, once the unknown parameters have been estimated the Kalman filter gives the residuals of the model, which are key for the Forward Search.

## 4.1 State space representation

The state space form is a very general representation that can be applied to many statistical models. In particular, every linear model has a state space representation which is not unique.
In the sequel we will consider the generic form of the state space representation and of the Kalman filter which are given for multivariate time series. This does not represent a restriction since the results can be readily extended to the univariate case where of course some vectors and matrices degenerate to scalars.
Denote by $\mathbf{y}_t$ an observed time series, that can be univariate as in our case or multivariate. $\mathbf{y}_t$ is called the **observation vector**. The series $\mathbf{y}_t$ is connected to the **state vector** $\mathbf{x}_t$, through the **observation equation**

$$\mathbf{y}_t = \mathbf{Z}_t\mathbf{x}_t + \boldsymbol{\nu}_t \qquad \boldsymbol{\nu}_t \sim N(\mathbf{0}, \mathbf{H}_t) \tag{4.1}$$

where $\mathbb{E}(\boldsymbol{\nu}_t) = \mathbf{0}$ and $\mathbb{V}(\boldsymbol{\nu}_t) = \mathbf{H}_t$.
The elements of the state vector $\mathbf{x}_t$ are not observable, however, they are assumed to be

generated by a first order Markov process,

$$\mathbf{x}_{t+1} = \mathbf{T}_t\mathbf{x}_t + \mathbf{R}_t\boldsymbol{\eta}_t \qquad \boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{Q}_t) \tag{4.2}$$

where $\mathbb{E}(\boldsymbol{\eta}_t) = \mathbf{0}$ and $\mathbb{V}(\boldsymbol{\eta}_t) = \mathbf{Q}_t$. Equation (4.2) is called **state equation**.
The idea underlying this model is that the evolution of the system over time is determined by $\mathbf{x}_t$ according to the state equation. However, because $\mathbf{x}_t$ cannot be observed directly, the statistician must base the analysis on the observation vector $\mathbf{y}_t$.
The model defined by equation (4.1) and (4.2) is defined as the general linear Gaussian state space model.
In many circumstances the vector $\mathbf{x}_t$ does not have a direct interpretation and it is determined by construction. The major aim of the state space form is to define a model such that $\mathbf{x}_t$ has as few elements as possible and contains all the relevant information of the system at time $t$. A state space form that minimize the number of components in the vector $\mathbf{x}_t$ is defined to be a **minimal realization**. In order to complete the specification of the system three further assumptions needs to be specified:

(A1) The initial state vector $\mathbf{x}_1$ is normally distributed with mean $\mathbb{E}(\mathbf{x}_1) = \mathbf{m}_1$ and variance $\mathbb{V}(\mathbf{x}_1) = \mathbf{P}_1$, that is $\mathbf{x}_1 \sim N(\mathbf{m}_1, \mathbf{P}_1)$

(A2) The error terms $\boldsymbol{\nu}_t$ and $\boldsymbol{\eta}_t$ are uncorrelated with the initial state, that is $\mathbb{C}\text{ov}(\boldsymbol{\nu}_t, \mathbf{x}_1) = 0$, $\mathbb{C}\text{ov}(\boldsymbol{\eta}_t, \mathbf{x}_1) = 0$ for $t = 1, \ldots, T$

(A3) The error terms $\boldsymbol{\nu}_t$ and $\boldsymbol{\eta}_t$ are uncorrelated in all time periods, that is $\mathbb{C}\text{ov}(\boldsymbol{\nu}_t, \boldsymbol{\eta}_k) = 0$ for all $k, t = 1, \ldots, T$.

The matrices $\mathbf{Z}_t, \mathbf{T}_t, \mathbf{R}_t, \mathbf{Q}_t$ are called **system matrices**, and although they may change with time they do so in a deterministic way and therefore they are non stochastic.
When the system matrices are invariant over time the model is said to be **time invariant** and the stationary models that we will discuss represent a special case.
Examples of models where the matrices change with time are regression models with dynamic parameters, however in the following the matrices will be assumed to be time invariant and the index $t$ will be dropped.
In real applications, some of the matrices $\mathbf{Z}_t, \mathbf{T}_t, \mathbf{R}_t, \mathbf{Q}_t$ will depend on one ore more unknown parameters. In the state space literature these parameters are referred as **hyperparameters** and will denoted by the vector $\boldsymbol{\theta}$.

## 4.2 The state space form for the class of $ARMA$ processes

Before to proceed to the discussion of the Kalman filter and how it can be used to obtain the exact likelihood function we give the state space representation of the different models that we have discussed so far.
The state space representations illustrated in this section are the same adopted to run the Forward Search.

### 4.2.1 State space form of $AR(p)$ and $MA(q)$ processes

Given an $AR(p)$ process in its mean centered form,

$$(Y_t - \mu) = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \cdots + \epsilon_t,$$

the resulting state space representation is

$$\underbrace{(Y_t - \mu)}_{\mathbf{y}_t} = \underbrace{(1 \quad \dots \quad 0)}_{\mathbf{Z}} \underbrace{\begin{pmatrix} Y_t - \mu \\ \vdots \\ Y_{t-p} - \mu \end{pmatrix}}_{\mathbf{x}_t}$$

$$\underbrace{\begin{pmatrix} Y_{t+1} - \mu \\ \vdots \\ Y_{t-p+1} - \mu \end{pmatrix}}_{\mathbf{x}_{t+1}} = \underbrace{\begin{pmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}}_{\mathbf{T}} \underbrace{\begin{pmatrix} Y_t - \mu \\ \vdots \\ Y_{t-p} - \mu \end{pmatrix}}_{\mathbf{x}_t} + \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 0 \end{pmatrix}}_{\mathbf{R}} \underbrace{\epsilon_{t+1}}_{\boldsymbol{\eta}_t}$$

As a simple example we consider an $AR(2)$ process. In this case the system matrices reduces to

$$\underbrace{(Y_t - \mu)}_{\mathbf{y}_t} = \underbrace{(1 \quad 0)}_{\mathbf{Z}} \underbrace{\begin{pmatrix} Y_t - \mu \\ Y_{t-1} - \mu \end{pmatrix}}_{\mathbf{x}_t}$$

$$\underbrace{\begin{pmatrix} Y_{t+1} - \mu \\ Y_t - \mu \end{pmatrix}}_{\mathbf{x}_{t+1}} = \underbrace{\begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}}_{\mathbf{T}} \underbrace{\begin{pmatrix} Y_t - \mu \\ Y_{t-1} - \mu \end{pmatrix}}_{\mathbf{x}_t} + \underbrace{\begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\mathbf{R}} \underbrace{\epsilon_{t+1}}_{\boldsymbol{\eta}_t}$$

The state space representation can next be extended to moving-average processes. Given a mean centered $MA(q)$ process

$$(Y_t - \mu) = \epsilon_t - \theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2} - \cdots - \theta_q\epsilon_{t-q},$$

its state space representation is given by

$$\underbrace{(Y_t - \mu)}_{\mathbf{y}_t} = \underbrace{(1 \quad \theta_1 \quad \theta_2 \quad \dots \quad \theta_{q-1} \quad \theta_q)}_{\mathbf{Z}} \underbrace{\begin{pmatrix} \epsilon_t \\ \vdots \\ \epsilon_{t-q} \end{pmatrix}}_{\mathbf{x}_t}$$

$$\underbrace{\begin{pmatrix} \epsilon_{t+1} \\ \vdots \\ \epsilon_{t-q+1} \end{pmatrix}}_{\mathbf{x}_{t+1}} = \underbrace{\begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}}_{\mathbf{T}} \underbrace{\begin{pmatrix} \epsilon_t \\ \vdots \\ \epsilon_{t-q} \end{pmatrix}}_{\mathbf{x}_t} + \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 0 \end{pmatrix}}_{\mathbf{R}} \underbrace{\epsilon_{t+1}}_{\boldsymbol{\eta}_t}$$

For instance when a simple $MA(1)$ process is considered the observation and state equation reduces to

$$\underbrace{(Y_t - \mu)}_{\mathbf{y}_t} = \underbrace{(1 \quad \theta)}_{\mathbf{Z}} \underbrace{\begin{pmatrix} \epsilon_t \\ \epsilon_{t-1} \end{pmatrix}}_{\mathbf{x}_t}$$

$$\underbrace{\begin{pmatrix} \epsilon_{t+1} \\ \epsilon_t \end{pmatrix}}_{\mathbf{x}_{t+1}} = \underbrace{\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}}_{\mathbf{T}} \underbrace{\begin{pmatrix} \epsilon_t \\ \epsilon_{t-1} \end{pmatrix}}_{\mathbf{x}_t} + \underbrace{\begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\mathbf{R}} \underbrace{\epsilon_{t+1}}_{\boldsymbol{\eta}_t}$$

### 4.2.2 $ARMA(p, q)$ model

There are several alternatives representations of $ARMA(p, q)$ processes, here we will consider the form proposed in Harvey (1990) and adopted also in Durbin and Koopman (2012). This form has the advantage that the error terms in the observation and state equations are independent from each other. In order to express an $ARMA(p, q)$ processes in state space form, it is first necessary to rewrite the model as

$$(Y_t - \mu) = \sum_{j=1}^{r} \phi_j(Y_{t-j} - \mu) + \epsilon_t + \sum_{j=1}^{r-1} \theta_j\epsilon_{t-j}$$

where $r = \max(p, q+1)$, $\phi_j = 0$ for $j > p$ and $\theta_j = 0$ for $j > q$. The system matrices for the $ARMA(p,q)$ model are given by

$$\mathbf{Z}_{1\times r} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \end{pmatrix},$$

$$\mathbf{x}_{t\ r\times 1} = \begin{pmatrix} Y_t - \mu \\ \phi_2(Y_{t-1} - \mu) + \cdots + \phi_r(Y_{t-r+1} - \mu) + \theta_1\epsilon_t + \cdots + \theta_{r-1}\epsilon_{t-r+2} \\ \phi_3(Y_{t-1} - \mu) + \cdots + \phi_r(Y_{t-r+2} - \mu) + \theta_2\epsilon_t + \cdots + \theta_{r-1}\epsilon_{t-r+3} \\ \vdots \\ \phi_r(Y_{t-1} - \mu) + \theta_{r-1}\epsilon_t \end{pmatrix},$$

$$\mathbf{T}_{r\times r} = \begin{pmatrix} \phi_1 & 1 & \cdots & 0 \\ \vdots & 0 & \cdots & 0 \\ \phi_{r-1} & 0 & \ddots & 1 \\ \phi_r & 0 & \cdots & 0 \end{pmatrix},$$

$$\mathbf{R}_{r\times 1} = \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{r-1} \end{pmatrix},$$

$$\boldsymbol{\eta}_t = \epsilon_{t+1}.$$

Therefore the full system becomes

$$\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t$$

$$\mathbf{x}_{t+1} = \mathbf{T}\mathbf{x}_t + \mathbf{R}\boldsymbol{\eta}_t.$$

### 4.2.3 $ARIMA(p, d, q)$ and seasonal extension

To obtain the state space representation for the $ARIMA(p, d, q)$ model we follow the same steps used in the $ARMA(p, q)$ case. First, the $ARIMA(p, d, q)$ process can be rewritten as

$$\dot{Y}_t = \sum_{j=1}^{r} \phi_j \dot{Y}_{t-j} + \epsilon_t + \sum_{j=1}^{r-1} \theta_j \epsilon_{t-j},$$

where now $\dot{Y}_t = \nabla^d (Y_t - \mu)$, $r = \max(p, q+1)$ and, for convenience, $\phi_0 = 0$.
The system matrices for the general $ARIMA(p, d, q)$ model are given by

$$\mathbf{Z}_{1\times r+d} = \begin{pmatrix} \mathbf{1}'_{1\times d+1} & 0 & \cdots & 0 \end{pmatrix}$$

$$\mathbf{x}_{r+d\times 1} = \begin{pmatrix} Y_{t-1} - \mu \\ \nabla(Y_{t-1} - \mu) \\ \vdots \\ \dot{Y}_t \\ \phi_2 \dot{Y}_{t-1} + \cdots + \phi_r \dot{Y}_{t-r+1} + \theta_1 \epsilon_t + \cdots + \theta_{r-1} \epsilon_{t-r+2} \\ \phi_3 \dot{Y}_{t-1} + \cdots + \phi_r \dot{Y}_{t-r+2} + \theta_2 \epsilon_t + \cdots + \theta_{r-1} \epsilon_{t-r+3} \\ \vdots \\ \phi_r \dot{Y}_{t-1} + \theta_{r-1} \epsilon_t \end{pmatrix}$$

$$\mathbf{T}_{r+d\times r+d} = \begin{pmatrix} & \mathbf{Z} & & & \\ 0 & 1 & 1 & \cdots & 0 \\ \mathbf{0}'_{1\times d} & \phi_1 & 1 & \cdots & 0 \\ \mathbf{0}'_{1\times d} & \vdots & 0 & \cdots & 0 \\ \mathbf{0}'_{1\times d} & \phi_{r-1} & 0 & \ddots & 1 \\ \mathbf{0}'_{1\times d} & \phi_r & 0 & \cdots & 0 \end{pmatrix}$$

$$\mathbf{R}_{r+d\times 1} = \begin{pmatrix} \mathbf{0}_{d\times 1} \\ 1 \\ \theta_1 \\ \vdots \\ \theta_{r-1} \end{pmatrix}$$

$$\boldsymbol{\eta}_t = \epsilon_{t+1}$$

As an example we can consider the state space representation of the $ARIMA(2,1,1)$ model. That is,

$$\underbrace{(Y_t - \mu)}_{\mathbf{y}_t} = \underbrace{(1 \quad 1 \quad 0)}_{\mathbf{Z}} \underbrace{\begin{pmatrix} Y_{t-1} - \mu \\ \dot{Y}_t \\ \phi_2 \dot{Y}_{t-1} + \theta_1 \epsilon_t \end{pmatrix}}_{\mathbf{x}_t}$$

$$\underbrace{\begin{pmatrix} Y_t - \mu \\ \dot{Y}_{t+1} \\ \phi_2 \dot{Y}_t + \theta_1 \epsilon_{t+1} \end{pmatrix}}_{\mathbf{x}_{t+1}} = \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 0 & \phi_1 & 1 \\ 0 & \phi_2 & 0 \end{pmatrix}}_{\mathbf{T}} \underbrace{\begin{pmatrix} Y_{t-1} - \mu \\ \dot{Y}_t \\ \phi_2 \dot{Y}_{t-1} + \theta_1 \epsilon_t \end{pmatrix}}_{\mathbf{x}_t} + \underbrace{\begin{pmatrix} 0 \\ 0 \\ 1 \\ \theta_1 \end{pmatrix}}_{\mathbf{R}} \underbrace{\epsilon_{t+1}}_{\boldsymbol{\eta}_t}$$

When the time series of interest exhibits a seasonal behavior, the seasonality and the trend are removed by differencing repeatedly the series as presented in Chapter 3. The final difference time series $\tilde{Y}_t = \nabla^d \nabla_s^D Y_t$ can then be modeled as a proper $ARMA$ model, therefore the results obtained for the non seasonal $ARIMA$ models can be extended directly to seasonal $ARIMA$ model by substituting $\tilde{Y}_t$ instead of $Y_t$.

## 4.3 The Kalman filter

Once the model of interest has been placed in the state space form it is possible to conduct three different kind of inference with it: **filtering**, **prediction** and **smoothing**.
Denote by $\mathcal{F}_t$ the sigma algebra generated by the observations up to time $t$, that is $\mathcal{F}_t = \sigma(Y_1, \ldots, Y_t)$, that for simplicity it can be interpreted as the information available up to time $t$. Further, we assume that the model parameters are known.
The three different inferences mentioned above can be described as follows:

(I1) **Filtering**. Filtering corresponds to recover the state vector $\mathbf{x}_t$ given $\mathcal{F}_t$, i.e all the information available up to time $t$.

(I2) **Prediction**. Prediction refers to forecast of $\mathbf{x}_h$ or $\mathbf{y}_h$ given $\mathcal{F}_t$, the information observed up to time $t$. $h$ is the forecast horizon, typically $h = T+1, \ldots, T+k$.

(I3) **Smoothing**. Smoothing refers to estimating the state vector $\mathbf{x}_k$ given $\mathcal{F}_T$ where $k < T$.

To attain the above results a major role is performed by the Kalman filter, which is based at least in the Gaussian case on some properties of the multivariate normal distribution. The main role of the Kalman filter is to update the knowledge of the state vector recursively when a new observation becomes available. In other words by knowing the distribution of the state $\mathbf{x}_t$ conditional on the information set $\mathcal{F}_{t-1}$, the Kalman filter is used to update the conditional distribution of $\mathbf{x}_t$ with the updated information set $\mathcal{F}_t$.

### 4.3.1 Some results on the multivariate normal distribution

The derivation of the Kalman filter is derived by applying important properties of the multivariate normal distribution. Therefore, before discussing the Kalman filter we briefly recall these results. The proofs are taken form Durbin and Koopman (2012).
Suppose, that two random vectors $\mathbf{y}$ and $\mathbf{x}$ are jointly normally distributed

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim N(\boldsymbol{\mu}_{xy}, \boldsymbol{\Xi})$$

where,

$$\mathbb{E}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix} = \boldsymbol{\mu}_{xy}, \quad \text{and} \quad \mathbb{V}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}'_{xy} & \boldsymbol{\Sigma}_{yy} \end{pmatrix} = \boldsymbol{\Xi}.$$

The conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ is still a normal distribution,

$$\mathbf{x}|\mathbf{y} \sim N(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) \tag{4.3}$$

with mean vector

$$\mathbb{E}(\mathbf{x}|\mathbf{y}) = \boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y) \tag{4.4}$$

and covariance matrix

$$\mathbb{V}(\mathbf{x}|\mathbf{y}) = \boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}'_{xy}.$$

Now, if we define a transformation from $\mathbf{x}$ to the variable $\mathbf{z} = \mathbf{x} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)$, since

the transformation is linear and $(\mathbf{x}, \mathbf{y})$ is normally distributed, by the properties of normal random variables, $(\mathbf{z}, \mathbf{y})$ is still normally distributed with mean

$$\mathbb{E}(\mathbf{z}) = \boldsymbol{\mu}_x,$$

variance

$$\mathbb{V}(\mathbf{z}) = \mathbb{E}[(\mathbf{z} - \boldsymbol{\mu}_x)(\mathbf{z} - \boldsymbol{\mu}_x)'] = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{xy}'$$

and covariance

$$\begin{aligned} \mathbb{C}\text{ov}(\mathbf{z}, \mathbf{y}) &= \mathbb{C}\text{ov}[\mathbf{x} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y), \mathbf{y}] \qquad (4.5) \\ &= \mathbb{C}\text{ov}(\mathbf{x}, \mathbf{y}) - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\mathbb{C}\text{ov}[(\mathbf{y} - \boldsymbol{\mu}_y), \mathbf{y}] \\ &= \boldsymbol{\Sigma}_{xy} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yy} = 0. \end{aligned}$$

Equation (4.5) shows clearly that the random vectors $\mathbf{z}$ and $\mathbf{y}$ are uncorrelated. Using the result that if two uncorrelated normal random vectors are also independent we can state that $\mathbf{z}$ is independent from $\mathbf{y}$. Extending this result we obtain that $\mathbf{x}|\mathbf{y}$ is independent from $\mathbf{y}$. This result is well known in multivariate regression theory, as eq. (4.4) states that $\mathbf{x}$ can be predicted via regression, when $\mathbf{x}$ is unknown and $\mathbf{y}$ is known.
We take as our estimate of $\mathbf{x}$ the conditional expectation $\mathbb{E}(\mathbf{x}|\mathbf{y})$, that is

$$\widehat{\mathbf{x}} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xx}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)$$

with estimation error equal to $\widehat{\mathbf{x}} - \mathbf{x}$. It is possible to show that the proposed estimator is conditionally unbiased, $\mathbb{E}(\widehat{\mathbf{x}} - \mathbf{x}|\mathbf{y}) = 0$ and unconditionally unbiased $\mathbb{E}(\widehat{\mathbf{x}} - \mathbf{x}) = 0$. For further properties of the estimator see Durbin and Koopman (2012).

### 4.3.2 Derivation of the Kalman filter

The objective of the Kalman filter is to obtain the conditional distribution of the state vectors $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ given the information set $\mathcal{F}_t$ for $t = 1, \ldots, T$. In order to introduce the Kalman filter some notation is first necessary. We will denote the conditional expectations of the state vector given the information set available at time $t$ by $\mathbf{m}_{t|t} = \mathbb{E}(\mathbf{x}_t|\mathcal{F}_t)$, $\mathbf{m}_{t+1} = \mathbb{E}(\mathbf{x}_{t+1}|\mathcal{F}_t)$, and the corresponding conditional variances by $\mathbf{P}_{t|t} = \mathbb{V}(\mathbf{x}_t|\mathcal{F}_t)$, $\mathbf{P}_{t+1} = \mathbb{V}(\mathbf{x}_{t+1}|\mathcal{F}_t)$.
Since all distributions are assumed normal it follows from the previous results on the multivariate normal distribution that conditional distributions of subset of variables given other subsets are also normal. In particular the following two results hold

$$\mathbf{x}_t|\mathcal{F}_t \sim N(\mathbf{m}_{t|t}, \mathbf{P}_{t|t}) \qquad \text{and} \qquad \mathbf{x}_{t+1}|\mathcal{F}_t \sim N(\mathbf{m}_{t+1}, \mathbf{P}_{t+1}).$$

Moreover, we assume that the initial state vector is normally distributed with mean and variance known, that is $\mathbf{x}_1 \sim N(\mathbf{m}_1, \mathbf{P}_1)$.
Now, given a value of $\mathbf{m}_t$, by applying the multivariate normal result we can estimate $\mathbf{y}_t$ using the conditional expectation, that is

$$\begin{aligned}\widehat{\mathbf{y}}_t &= \mathbb{E}(\mathbf{y}_t|\mathcal{F}_{t-1}) \\ &= \mathbb{E}(\mathbf{Z}_t\mathbf{x}_t + \boldsymbol{\nu}_t|\mathcal{F}_{t-1}) = \mathbf{Z}_t\mathbf{m}_t\end{aligned}$$

with prediction error given by

$$\begin{aligned}\mathbf{v}_t &= \mathbf{y}_t - \widehat{\mathbf{y}}_t = \mathbf{y}_t - \mathbb{E}(\mathbf{y}_t|\mathcal{F}_{t-1}) \\ &= \mathbf{y}_t - \mathbb{E}(\mathbf{Z}_t\mathbf{x}_t + \boldsymbol{\nu}_t|\mathcal{F}_{t-1}) = \mathbf{y}_t - \mathbf{Z}_t\mathbf{m}_t.\end{aligned}$$

The quantity $\mathbf{v}_t$ is the one step ahead forecast error of $\mathbf{y}_t$ given the information set $\mathcal{F}_{t-1}$, in the case of $ARMA$ processes $\mathbf{v}_t$ corresponds to the the traditional residual.
Now, to derive the Kalman filter it is sufficient to consider the joint distribution of $\mathbf{v}_t$ and $\mathbf{x}_t$ given $\mathcal{F}_{t-1}$.
The distribution of $\mathbf{v}_t$ conditional on $\mathcal{F}_{t-1}$ is a normal distribution

$$\mathbf{v}_t|\mathcal{F}_{t-1} \sim N(\mathbf{0}, \mathbf{F}_t)$$

where the zero mean comes by

$$\mathbb{E}(\mathbf{v}_t|\mathcal{F}_{t-1}) = \mathbb{E}(\mathbf{y}_t|\mathcal{F}_{t-1}) - \mathbb{E}(\mathbf{y}_t|\mathcal{F}_{t-1}) = \mathbf{0},$$

and

$$\mathbf{F}_t = \mathbb{V}(\mathbf{v}_t|\mathcal{F}_{t-1}) = \mathbb{V}(\mathbf{Z}_t\mathbf{x}_t + \boldsymbol{\nu}_t - \mathbf{Z}_t\mathbf{m}_t|\mathcal{F}_{t-1}) = \mathbf{Z}_t\mathbf{P}_t\mathbf{Z}_t' + \mathbf{H}_t.$$

Furthermore, the distribution of the state vector $\mathbf{x}_t$ conditional on $\mathcal{F}_{t-1}$ is also normal,

$$\mathbf{x}_t|\mathcal{F}_{t-1} \sim N(\mathbf{m}_t, \mathbf{P}_t)$$

and the joint distribution of $(\mathbf{x}_t, \mathbf{v}_t)$ conditional on $\mathcal{F}_{t-1}$ is still a normal distribution, given by

$$\begin{pmatrix} \mathbf{x}_t \\ \mathbf{v}_t \end{pmatrix} |\mathcal{F}_{t-1} \sim N\left[ \begin{pmatrix} \mathbf{m}_t \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{P}_t & \mathbf{P}_t\mathbf{Z}_t' \\ \mathbf{P}_t'\mathbf{Z}_t & \mathbf{F}_t \end{pmatrix} \right]$$

where,

$$\mathbf{P}_t\mathbf{Z}_t' = \mathbb{C}\text{ov}(\mathbf{x}_t, \mathbf{v}_t|\mathcal{F}_{t-1}).$$

Since knowing $\mathcal{F}_{t-1}$ and $\mathbf{v}_t$ is equivalent to know $\mathcal{F}_t$, it follows that

$$\mathbf{m}_{t|t} = \mathbb{E}(\mathbf{x}_t|\mathcal{F}_t) = \mathbb{E}(\mathbf{x}_t|\mathcal{F}_{t-1}, \mathbf{v}_t).$$

This quantity can be obtained using the results of the multivariate normal distribution by regressing $\mathbf{x}_t$ on $\mathbf{v}_t$, which gives

$$\begin{aligned} \mathbf{m}_{t|t} &= \mathbb{E}(\mathbf{x}_t|\mathcal{F}_{t-1}) + \mathbb{C}\text{ov}(\mathbf{x}_t, \mathbf{v}_t|\mathcal{F}_{t-1})[\mathbb{V}(\mathbf{v_t}|\mathcal{F}_{t-1})]^{-1}\mathbf{v}_t \qquad (4.6) \\ &= \mathbf{m}_t + \mathbf{P}_t\mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{v}_t. \end{aligned}$$

Next, it is possible to obtain the conditional variance of the state vector $\mathbf{x}_t$ conditional on the information set $\mathcal{F}_t$, that is

$$\begin{aligned}\mathbf{P}_{t|t} &= \mathbb{V}(\mathbf{x}_t|\mathcal{F}_t) = \mathbb{V}(\mathbf{x}_t|\mathcal{F}_{t-1}, \mathbf{v}_t) \\ &= \mathbb{V}(\mathbf{x}_t|\mathcal{F}_{t-1}) - \mathbb{C}\text{ov}(\mathbf{x}_t, \mathbf{v}_t|\mathcal{F}_{t-1})[\mathbb{V}(\mathbf{v}_t|\mathcal{F}_{t-1})]^{-1}\mathbb{C}\text{ov}(\mathbf{v}_t, \mathbf{x}_t|\mathcal{F}_{t-1}) \\ &= \mathbf{P}_t - \mathbf{P}_t\mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{Z}_t\mathbf{P}_t'.\end{aligned} \tag{4.7}$$

Equations (4.6) and (4.7) are called the **updating equations** of the Kalman filter since they update the distribution of the state vector as new information becomes available. Finally, we can make use of the knowledge of $\mathbf{x}_t$ given $\mathcal{F}_t$ to predict the distribution of $\mathbf{x}_{t+1}$. We have,

$$\begin{aligned}\mathbf{m}_{t+1} &= \mathbb{E}(\mathbf{x}_{t+1}|\mathcal{F}_t) = \mathbb{E}(\mathbf{T}_t\mathbf{x}_t + \mathbf{R}_t\boldsymbol{\eta}_t|\mathcal{F}_t) \\ &= \mathbf{T}_t\mathbb{E}(\mathbf{x}_t|\mathcal{F}_t) = \mathbf{T}_t\mathbf{m}_{t|t} \\ &= \mathbf{T}_t(\mathbf{m}_t + \mathbf{P}_t\mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{v}_t) = \mathbf{T}_t\mathbf{m}_t + \mathbf{K}_t\mathbf{v}_t\end{aligned} \tag{4.8}$$

where $\mathbf{K}_t = \mathbf{T}_t\mathbf{P}_t\mathbf{Z}_t'\mathbf{F}_t^{-1}$, and

$$\begin{aligned}\mathbf{P}_{t+1} &= \mathbb{V}(\mathbf{x}_{t+1}|\mathcal{F}_t) = \mathbb{V}(\mathbf{T}_t\mathbf{x}_t + \mathbf{R}_t\boldsymbol{\eta}_t|\mathcal{F}_t) \\ &= \mathbf{T}_t\mathbb{V}(\mathbf{x}_t|\mathcal{F}_t)\mathbf{T}_t' + \mathbf{R}_t\mathbf{Q}_t\mathbf{R}_t' \\ &= \mathbf{T}_t(\mathbf{P}_t - \mathbf{P}_t\mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{Z}_t\mathbf{P}_t')\mathbf{T}_t' + \mathbf{R}_t\mathbf{Q}_t\mathbf{R}_t' \\ &= \mathbf{T}_t\mathbf{P}_t\mathbf{T}_t' - \underbrace{\mathbf{T}_t\mathbf{P}_t\mathbf{Z}_t'\mathbf{F}_t^{-1}}_{\mathbf{K}_t}\mathbf{Z}_t\mathbf{P}_t'\mathbf{T}_t' + \mathbf{R}_t\mathbf{Q}_t\mathbf{R}_t' \\ &= \mathbf{T}_t\mathbf{P}_t(\mathbf{T}_t - \mathbf{K}_t\mathbf{Z}_t)' + \mathbf{R}_t\mathbf{Q}_t\mathbf{R}_t'.\end{aligned} \tag{4.9}$$

Equations (4.8) and (4.9) are referred as the **prediction equations**, while the matrix $\mathbf{K}_t$ is called the **Kalman gain**.

The above system of recursions constitute the Kalman filter. The main power of this filter is that it enables to update the knowledge of the system in iterative way, each time a new observation is acquired.

**Kalman Filter Algorithm** In summary the Kalman filter computes, for $t = 1, \ldots, T$ the following set of recursions

1. $\mathbf{v}_t = \mathbf{y}_t - \mathbf{Z}_t\mathbf{m}_t$
2. $\mathbf{F}_t = \mathbf{Z}_t\mathbf{P}_t\mathbf{Z}_t' + \mathbf{H}_t$
3. $\mathbf{m}_{t|t} = \mathbf{m}_t + \mathbf{P}_t\mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{v}_t$
4. $\mathbf{P}_{t|t} = \mathbf{P}_t - \mathbf{P}_t\mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{Z}_t\mathbf{P}_t$
5. $\mathbf{K}_t = \mathbf{T}_t\mathbf{P}_t\mathbf{Z}_t'\mathbf{F}_t^{-1}$
6. $\mathbf{m}_{t+1} = \mathbf{T}_t\mathbf{m}_{t|t} + \mathbf{K}_t\mathbf{v}_t$
7. $\mathbf{P}_{t+1} = \mathbf{T}_t\mathbf{P}_{t|t}(\mathbf{T}_t - \mathbf{K}_t\mathbf{Z}_t)' + \mathbf{R}_t\mathbf{Q}_t\mathbf{R}_t'$

When time invariant systems are considered, like in the $ARMA$ case the subscript $t$ from the system matrices can be dropped and the set of recursions become

1. $\mathbf{v}_t = \mathbf{y}_t - \mathbf{Z}\mathbf{m}_t$
2. $\mathbf{F}_t = \mathbf{Z}\mathbf{P}_t\mathbf{Z}' + \mathbf{H}$
3. $\mathbf{m}_{t|t} = \mathbf{m}_t + \mathbf{P}_t\mathbf{Z}'\mathbf{F}_t^{-1}\mathbf{v}_t$
4. $\mathbf{P}_{t|t} = \mathbf{P}_t - \mathbf{P}_t\mathbf{Z}'\mathbf{F}_t^{-1}\mathbf{Z}\mathbf{P}_t$
5. $\mathbf{K}_t = \mathbf{T}\mathbf{P}_t\mathbf{Z}'\mathbf{F}_t^{-1}$
6. $\mathbf{m}_{t+1} = \mathbf{T}\mathbf{m}_{t|t} + \mathbf{K}_t\mathbf{v}_t$
7. $\mathbf{P}_{t+1} = \mathbf{T}\mathbf{P}_{t|t}(\mathbf{T} - \mathbf{K}_t\mathbf{Z})' + \mathbf{R}\mathbf{Q}\mathbf{R}'$

A useful shortcut to the standard Kalman filter can be obtained when we restrict our interest to time-invariant state space models as in our case. When the system matrices $\mathbf{Z}_t$,$\mathbf{H}_t$,$\mathbf{T}_t$,$\mathbf{R}_t$ and $\mathbf{Q}_t$ are constant over time, the recursion for $\mathbf{P}_{t+1}$ converges to a constant matrix, $\overline{\mathbf{P}}$ which is obtained as

$$\overline{\mathbf{P}} = \mathbf{T}\overline{\mathbf{P}}\mathbf{T}' - \mathbf{T}\overline{\mathbf{P}}\mathbf{Z}'\overline{\mathbf{F}}^{-1}\mathbf{Z}\overline{\mathbf{P}}\mathbf{T}' + \mathbf{R}\mathbf{Q}\mathbf{R}'$$

where $\overline{\mathbf{F}} = \mathbf{Z}\overline{\mathbf{P}} + \mathbf{H}$. The advantage of this result is that after the convergence of $\mathbf{P}$ to the steady space the recursive updates of $\mathbf{F}_t$,$\mathbf{K}_t$,$\mathbf{P}_{t|t}$ and $\mathbf{P}_{t+1}$ are no longer required.

**Missing observations** One of the most appealing characteristics of the Kalman filter is the possibility of dealing with missing observations in an easy and straightforward way. Suppose that observations $\mathbf{y}_j$ for $j = \tau, \ldots, \tau^*$, with $1 < \tau < \tau^* < T$, are missing.
The Kalman filter can be obtained for the missing cases simply by placing $\mathbf{Z}_t = 0$ for $t = \tau, \ldots, \tau^* - 1$.
Therefore when missing observations are present in the series, the Kalman filter results equal to the standard case with the exception that the matrix $\mathbf{Z}_t$ is taken equal to 0 at those time points where observations are missing.
Therefore, the recursions for $t = \tau, \ldots, \tau^* - 1$ becomes:

1. $\mathbf{v}_\tau = \mathbf{y}_\tau$
2. $\mathbf{F}_{\tau|t-1} = \mathbf{H}_t$
3. $\mathbf{m}_{\tau|t} = \mathbf{m}_{\tau|t-1}$
4. $\mathbf{P}_{\tau|t} = \mathbf{P}_{\tau|t-1}$
5. $\mathbf{m}_{\tau+1|t} = \mathbf{T}_t \mathbf{m}_{\tau|t}$
6. $\mathbf{P}_{\tau+1|t} = \mathbf{T}_t \mathbf{P}_{\tau|t} \mathbf{T}_t' + \mathbf{R}_t \mathbf{Q}_t \mathbf{R}_t'$

### 4.3.3 Initialization of the filter

In the derivation of the Kalman filter it was assumed that $\mathbf{x}_1 \sim N(\mathbf{m}_1, \mathbf{P}_1)$ and $\mathbf{m}_1$ and $\mathbf{P}_1$ where known. In all practical applications the elements of $\mathbf{m}_1$ and $\mathbf{P}_1$ are unknown. When this occurs it is necessary to have rules to define these initial vector for the Kalman filter, this process is called **initialization** of the filter.

**Stationary Initialization** When the process is stationary the initialization of the Kalman filter can be obtained with a simple closed form formula.
In order for a multivariate process to be stationary the eigenvalues of the matrix $\mathbf{T}_t$ must lie all inside the unit circle. In this case it is possible to initialize the filter by taking the unconditional mean and variance of the process $\mathbf{x}_t$. It is important to recall that stationarity implies a time-invariant state space model and therefore the subscripts from the system matrices can be dropped.
The unconditional mean is given by taking expectation of (4.2), that is

$$\mathbb{E}(\mathbf{x}_{t+1}) = \mathbf{T}\mathbb{E}(\mathbf{x}_t)\mathbf{R}\mathbb{E}(\boldsymbol{\eta}_t)$$

and since the process is covariance stationary

$$(\mathbf{I} - \mathbf{T})\mathbb{E}(\mathbf{x}_t) = \mathbf{0}.$$

Now, since unity is not an eigenvalue of $\mathbf{T}$, see Hamilton (1994), the matrix $(\mathbf{I} - \mathbf{T})$ is nonsingular. As a consequence, the unique solution is $\mathbb{E}(\mathbf{x}_t) = \mathbf{0}$.
Next, the unconditional variance of $\mathbf{x}_t$ can be found similarly. We have

$$\mathbb{E}(\mathbf{x}_{t+1}\mathbf{x}'_{t+1}) = \mathbb{E}[(\mathbf{T}\mathbf{x}_{t+1} + \mathbf{R}\boldsymbol{\eta}_t)(\mathbf{T}\mathbf{x}_{t+1} + \mathbf{R}\boldsymbol{\eta}_t)']$$

$$= \mathbf{T}\mathbb{E}(\mathbf{x}_t\mathbf{x}'_t)\mathbf{T}' + \mathbb{E}(\boldsymbol{\eta}_t\boldsymbol{\eta}'_t).$$

Letting $\boldsymbol{\Omega}$ be the variance matrix of $\mathbf{x}_t$ the previous equation becomes

$$\boldsymbol{\Omega} = \mathbf{T}\boldsymbol{\Omega}\mathbf{T}' + \mathbf{Q}$$

whose solution is given by

$$vec(\boldsymbol{\Omega}) = (\mathbf{I} - \mathbf{T} \otimes \mathbf{T})^{-1}\, vec(\mathbf{Q})$$

where $\otimes$ indicates the Kronecker product and $vec(\mathbf{A})$ denotes the operator that stacks the columns of the matrix $\mathbf{A}$ into a vector.
As a result, provided that the process of interest is stationary the Kalman filter can be initialized by taking $\mathbf{m}_1 = \mathbf{0}$ and $\mathbf{P}_1 = vec(\mathbf{\Omega})$.

**Diffuse Initialization** When the process $\mathbf{x}_t$ is non stationary the closed form solution obtained for the initial state mean and variance in the stationary case cannot be recovered and the **diffuse** initialization of the filter is used. This procedure comes necessary when the $ARIMA$ and $SARIMA$ models are considered.
Following Durbin and Koopman (2012) the exact initialization of the Kalman filter starts with the specification of a general model for the initial state vector $\mathbf{x}_1$, that is

$$\mathbf{x}_1 = \mathbf{m} + \mathbf{A}\boldsymbol{\delta} + \mathbf{R}_0\boldsymbol{\eta}_0, \qquad \boldsymbol{\eta}_0 \sim N(0, \mathbf{Q}_0). \tag{4.10}$$

The vector $\mathbf{m}$ is known, $\boldsymbol{\delta}$ is a vector of unknown quantities, and the matrices $\mathbf{A}$ and $\mathbf{R}_0$ are selection identity matrices, for some columns. They are selected in such a way that $\mathbf{A}'\mathbf{R}_0 = 0$.
The objective of the above representation is to decompose the initial state vector in three components; a constant part $\mathbf{m}$, a non stationary part $\mathbf{A}\delta$, and a stationary part $\mathbf{R}_0\boldsymbol{\eta}_0$.
The vector $\boldsymbol{\delta}$ is taken to be normally distributed, that is $\boldsymbol{\delta} \sim N(0, \kappa\mathbf{I})$, where $\kappa \to \infty$.
The normality assumption of the vector $\boldsymbol{\delta}$ act as a non-informative prior distribution for the process $\mathbf{x}_1$, where the non informative part is obtained by letting the variance tending to infinity. According to model (4.10), the expected value and the variance of the initial state vector are given by

$$\begin{aligned}\mathbb{E}(\mathbf{x}_1) &= \mathbb{E}(\mathbf{m} + \mathbf{A}\boldsymbol{\delta} + \mathbf{R}_0\boldsymbol{\eta}_0) \\ &= \mathbf{m}.\end{aligned}$$

$$\begin{aligned}\mathbb{V}(\mathbf{x}_1) &= \mathbb{V}(\mathbf{m} + \mathbf{A}\boldsymbol{\delta} + \mathbf{R}_0\boldsymbol{\eta}_0) \\ &= \kappa\mathbf{A}\mathbf{A}' + \mathbf{R}_0\mathbf{Q}_0\mathbf{R}' \\ &= \kappa\mathbf{P}_\infty + \mathbf{P}^*.\end{aligned}$$

Having defined the diffuse assumption of the state process two different initialization procedures can be adopted. The first one, that we will call **approximated diffuse**, and is discussed in Harvey (1990), consists in initializing the filter by assigning a large value to $\kappa$.

However, despite this procedure is simple, and might provide good results in preliminary analysis, it may be affected by misleading rounding errors as pointed out in Durbin and Koopman (2012). The second initialization method, introduced by Koopman (1997), is obtained by running the **exact Kalman filter**. The idea behind the exact initialization is that after $d$ ad-hoc steps the procedure converge to the standard Kalman filter that can next be adopted for the remaining $T - d$ units. A detailed proof of the converge result can be found in Koopman (1997) and Durbin and Koopman (2012).
The exact Kalman filter is obtained by running for $d$ steps a modified version of the Kalman filter, whose recursions depends on the properties of the matrix $\mathbf{F}_{\infty,t}$.
Once the convergence criterion is met the standard Kalman Filter is runned for the remaining $T - d$ units.
The exact Kalman filter is structured as follow.
Initialize the two components of the state variance

$$\mathbf{P}_{\infty,1} = \mathbf{A}\mathbf{A}'$$
$$\mathbf{P}_{*,1} = \mathbf{R}_0\mathbf{Q}_0\mathbf{R}_0'$$

Case 1. $\mathbf{F}_{\infty,t}$ is not singular

1. $\mathbf{v}_t^{(0)} = \mathbf{y}_t - \mathbf{Z}_t\mathbf{m}_t^{(0)} \qquad \mathbf{m}_1^{(0)} = \mathbf{m}$
2. $\mathbf{F}_{\infty,t} = \mathbf{Z}_t\mathbf{P}_{\infty,t}\mathbf{Z}_t'$
3. $\mathbf{F}_{*,t} = \mathbf{Z}_t\mathbf{P}_{*,t}\mathbf{Z}_t' + \mathbf{H}_t$
4. $\mathbf{M}_{\infty,t} = \mathbf{P}_{\infty,t}\mathbf{Z}_t'$
5. $\mathbf{M}_{*,t} = \mathbf{P}_{*,t}\mathbf{Z}_t'$
6. $\mathbf{F}_t^{(1)} = \mathbf{F}_{\infty,t}^{-1}$
7. $\mathbf{F}_t^{(2)} = -\mathbf{F}_{\infty,t}^{-1}\mathbf{F}_{*,t}\mathbf{F}_{\infty,t}^{-1}$
8. $\mathbf{K}_t^{(0)} = \mathbf{T}_t\mathbf{M}_{\infty,t}\mathbf{F}_t^{(1)} = \mathbf{T}_t\mathbf{P}_{\infty,t}\mathbf{Z}_t'\mathbf{F}_t^{(1)}$
9. $\mathbf{K}_t^{(1)} = \mathbf{T}_t\mathbf{M}_{*,t}\mathbf{F}_t^{(1)} + \mathbf{T}_t\mathbf{M}_{\infty,t}\mathbf{F}_t^{(2)} = \mathbf{T}_t(\mathbf{P}_{*,t}\mathbf{Z}_t'\mathbf{F}_t^{(1)} + \mathbf{P}_{\infty,t}\mathbf{Z}_t'\mathbf{F}_t^{(2)})$
10. $\mathbf{L}_t^{(0)} = \mathbf{T}_t - \mathbf{K}_t^{(0)}\mathbf{Z}_t$
11. $\mathbf{L}_t^{(1)} = -\mathbf{K}_t^{(1)}\mathbf{Z}_t$
12. $\mathbf{m}_{t+1}^{(0)} = \mathbf{T}_t\mathbf{m}_t^{(0)} + \mathbf{K}_t^{(0)}\mathbf{v}_t^{(0)}$
13. $\mathbf{P}_{\infty,t+1} = \mathbf{T}_t\mathbf{P}_{\infty,t}\mathbf{L}_t^{(0)'}$
14. $\mathbf{P}_{*,t+1} = \mathbf{T}_t\mathbf{P}_{\infty,t}\mathbf{L}_t^{(1)'} + \mathbf{T}_t\mathbf{P}_{*,t}\mathbf{L}_t^{(0)'} + \mathbf{R}_t\mathbf{Q}_t\mathbf{R}_t'$

Case 2. $\mathbf{F}_{\infty,t} = 0$

1. $\mathbf{v}_t^{(0)} = \mathbf{y}_t - \mathbf{Z}_t\mathbf{a}_t^{(0)}$
2. $\mathbf{F}_t = \mathbf{F}_{*,t} = \mathbf{Z}_t\mathbf{P}_{*,t}\mathbf{Z}_t'$
3. $\mathbf{M}_t = \mathbf{M}_{*,t} = \mathbf{P}_{*,t}\mathbf{Z}_t'$
4. $\mathbf{K}_t^{(0)} = \mathbf{T}_t\mathbf{M}_{*,t}\mathbf{F}_{*,t}^{-1} = \mathbf{T}_t\mathbf{P}_{*,t}\mathbf{Z}_t'(\mathbf{Z}_t\mathbf{P}_{*,t}\mathbf{Z}_t')^{-1}$
5. $\mathbf{L}_t^{(0)} = \mathbf{T}_t - \mathbf{K}_t^{(0)}\mathbf{Z}_t$
6. $\mathbf{m}_{t+1}^{(0)} = \mathbf{T}_t\mathbf{m}_t^{(0)} + \mathbf{K}_t^{(0)}\mathbf{v}_t^{(0)}$
7. $\mathbf{P}_{\infty,t+1} = \mathbf{T}_t\mathbf{P}_{\infty,t}\mathbf{T}_t'$
8. $\mathbf{P}_{*,t+1} = \mathbf{T}_t\mathbf{P}_{*,t}\mathbf{L}_t^{(0)'} + \mathbf{R}_t\mathbf{Q}\mathbf{R}_t'$

After $d$ steps the quantity $\mathbf{P}_{\infty,t}$ becomes equal to 0 and the filter switch to the traditional Kalman filter recursions which is initialized with

$$\begin{aligned}\mathbf{m}_{d+1} &= \mathbf{m}_{d+1}^{(0)}\\ \mathbf{P}_{d+1} &= \mathbf{P}_{*,d+1}.\end{aligned}$$

and performed for the remaining $T - d$ steps.

## 4.4 The prediction error decomposition

In the derivation of the Kalman filter we have assumed that the parameters of the model were known. In the majority of statistical applications the parameters of the model are unknown and must be estimated.
The estimation of the parameters is generally performed by maximizing the likelihood function and the resulting estimator is the well-known maximum likelihood estimator.
When dealing with time series models the maximization of the exact likelihood is not straightforward and few shortcuts are generally introduced. For instance when dealing with autoregressive processes it is common practice to maximize the conditional likelihood which, in the Gaussian case, reduces to the ordinary least square estimator. The same principle can be applied to causal moving average processes, in this case it is possible to approximate the $MA$ process with an $AR$ process of higher order and obtain the least squares estimates. Next, the estimates of the $MA$ parameters can be recovered.
The approximation provided by the conditional likelihood can affect the final estimate, especially when the time series is short or has only few observations. Moreover, when there are many missing observations in the time series, as in the initial steps of the Forward Search, the estimation becomes particularly difficult, if not impossible. One of the

main advantages of the state space representation is the possibility to obtain exact maximum likelihood estimators.The Kalman filter therefore results extremely useful because it allows to obtain the exact likelihood function through the so called prediction error decomposition.
Assume that we observe a time series $\mathbf{y}_t$ for $t = 1, \ldots, T$ and that it can be represented with the Gaussian state space model. The likelihood function is then given by

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{T} f(\mathbf{y}_t; \boldsymbol{\theta} | \mathcal{F}_{t-1})$$

where $\boldsymbol{\theta}$ is the vector of the unknown hyperparameters and $f(\mathbf{y}_t; \theta | \mathcal{F}_{t-1})$ is the density function of $\mathbf{y}_t$ conditional on the information set $\mathcal{F}_{t-1}$.
Under the assumption that the error term in the observation equation has a Gaussian distribution, the distribution of $\mathbf{y}_t$ given the information set $\mathcal{F}_{t-1}$ is itself a normal distribution with mean equal to $\mathbb{E}(\mathbf{y}_t | \mathcal{F}_{t-1}) = \mathbf{Z}\mathbf{m}_{t|t-1} = \widehat{\mathbf{y}}_t$ and variance$\mathbb{V}(\mathbf{y}_t | \mathcal{F}_{t-1}) = \mathbf{F}_t$. Moreover, from the results obtained in the derivation of the Kalman filter the distribution of $\mathbf{x}_t$ conditional on the information set $\mathcal{F}_{t-1}$ is again a normal distribution with mean $\mathbf{m}_t$ and variance $\mathbf{P}_t$.
The Kalman filter is an extremely useful tool that can be used to obtain the exact maximum likelihood for every model that can be represented in state space form. Once the likelihood has been computed it can be maximized with respect to the unknown parameters. The main advantage therefore is that a single routine can be adopted to estimate a huge variety of models. We will first consider the stationary case, that is to say when the Kalman filter can be initialized with the closed form formula. Next we will report the exact likelihood for the non-stationary case, i.e. when the exact diffuse initialization is performed.

**Stationary case**

When the model considered is stationary with a Gaussian distribution, the likelihood function results equal to

$$L(\boldsymbol{\theta}) = 2\pi^{-\frac{NT}{2}} \prod_{t=1}^{T} |\mathbf{F}_t|^{-1/2} \exp - \left[ \frac{\sum_{t=1}^{T} (\mathbf{v}_t' \mathbf{F}_t^{-1} \mathbf{v}_t)}{2} \right]$$

and by taking the logarithm the resulting log-likelihood is

$$\ell(\boldsymbol{\theta}) = -\frac{NT}{2}\log\ 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log|\mathbf{F}_t| - \frac{1}{2}\sum_{t=1}^{T}\mathbf{v}_t'\mathbf{F}_t^{-1}\mathbf{v}_t \tag{4.11}$$

$$= -\frac{NT}{2}\log\ 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log|\mathbf{F}_t| - \frac{1}{2}\sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{Z}_t\mathbf{m}_t)'\mathbf{F}_t^{-1}(\mathbf{y}_t - \mathbf{Z}_t\mathbf{m}_t)$$

Equation (4.11) is referred as the **prediction error decomposition** and encompasses in a general form the likelihoods for all the models presented in Chapter 2. This is the reason why the Kalman filter algorithm turns out to be so useful.
Once the model has been expressed in its state space form and the Kalman Filter has been runned it is possible to maximize the exact log-likelihood with respect to the vector of hyperparameters $\boldsymbol{\theta}$.
The maximization problem is non linear and an iterative algorithm is employed. The standard algorithm adopted in econometrics literature is the BFGS and a detailed description of its application to the state space models is provided in Durbin and Koopman (2012).
In order to improve the stability of the algorithm it might be important to reduce the dimension of the vector $\boldsymbol{\theta}$.
When dealing with univariate time series it is generally possible to re parametrize the model such that $\boldsymbol{\theta} = (\boldsymbol{\theta}_*', \sigma_*^2)'$ where $\boldsymbol{\theta}_*$ contains $k-1$ parameters and $\sigma_*^2$ is the error variance. After the reparametrization, the error variances can be expressed $\mathbb{V}(\boldsymbol{\nu}_t) = \sigma_*^2\mathbf{H}_t$ and $\mathbb{V}(\boldsymbol{\eta}_t) = \sigma_*^2\mathbf{Q}_t$.
The prediction error decomposition of the likelihood and log-likelihood functions of the reparametrized model are given by:

$$L(\boldsymbol{\theta}) = (2\pi\sigma_*^2)^{-\frac{T}{2}}\prod_{t=1}^{T}\mathbf{F}_t^{-1/2}\exp - \left[\frac{\sum_{t=1}^{T}\mathbf{v}_t^2}{2\sigma_*^2\mathbf{F}_t}\right]$$

and

$$\ell(\boldsymbol{\theta}) = -\frac{T}{2}\log\ 2\pi - \frac{T}{2}\log\ \sigma_*^2 - \frac{1}{2}\sum_{t=1}^{T}\log\ \mathbf{F}_t - \frac{1}{2\sigma_*^2}\sum_{t=1}^{T}\mathbf{v}_t^2/\mathbf{F}_t \tag{4.12}$$

Since $\mathbf{v}_t$ and $\mathbf{F}_t$ do not depend on $\sigma_*^2$, differentiating (4.12) with respect to $\sigma_*^2$ yields to

$$\hat{\sigma}^2(\boldsymbol{\theta}_*) = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbf{v}_t^2}{\mathbf{F}_t}$$

By substituting $\hat{\sigma}_*^2(\boldsymbol{\theta}_*)$ in (4.12) we obtain the concentrated or profile log-likelihood, that must be maximized with respect to $\boldsymbol{\theta}_*$. That is,

$$\ell_c(\boldsymbol{\theta}_*) = -\frac{T}{2}(\log\ 2\pi + 1) - \frac{1}{2}\sum_{t=1}^{T}\log \mathbf{F}_t - \frac{T}{2}\log\ \hat{\sigma}_*^2(\boldsymbol{\theta}_*)$$

$$\propto -\frac{1}{2}\sum_{t=1}^{T}\log \mathbf{F}_t - \frac{T}{2}\log\ \hat{\sigma}_*^2(\boldsymbol{\theta}_*).$$

**Non stationary case**

When some of the elements o the initial state vector are diffuse the exact Kalman filter is performed. As a result, the likelihood needs to take into account the first $d$ steps in which the exact initialization is performed.
The resulting log-likelihood function to be maximized when the exact Kalman filter is performed is

$$\ell(\theta) = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{d}\varpi_t - \frac{1}{2}\sum_{t=d+1}^{T}(\log|\mathbf{F}_t| + \mathbf{v}_t'\mathbf{F}_t\mathbf{v}_t)$$

where,

$$\varpi_t = \begin{cases} \log|\mathbf{F}_{\infty,t}| & \text{if} \quad \mathbf{F}_{\infty,t} \quad \text{is positive definite} \\ \log|\mathbf{F}_{*,t}| + \mathbf{v}_t^{(0)'}\mathbf{F}_{*,t}\mathbf{v}_t^{(0)} & \text{if} \quad \mathbf{F}_{\infty,t} = 0 \end{cases}$$

The above expression of the log likelihood is presented in Koopman (1997).
The maximization of the log likelihood function is next performed with the $BFGS$ algorithm in the same way as in the stationary case.

# Chapter 5

# Outliers and influential observations in time series

Time series data, as many other type of statistical data, are often subject to outliers or atypical observations.
In statistics the term outliers has a long history and generally it is referred to those units that presents characteristics not shared by the majority of the data. Therefore, intuition suggests that the outlying observations are those that look discordant from the bulk of data with respect to some prescribed characteristic or measure of deviation.
In time series, and more generally in other statistical contexts the study of outliers has been conducted with two different aims that brought to two different approaches: the **diagnostic approach** and the **robust approach**.

1. In the **diagnostic approach**, diagnostic methods, based on models residuals, are employed to identify possible outliers or atypical observations, that are tested afterwards.
   Once the outliers are discovered it is possible to formulate a model which accounts for them or remove them from the analysis.
   The target of this strategy is to estimate the effect of the outliers, that may be the main objective of the analysis. Subsequently, it is possible to obtain an estimate of the parameters which is updated with the updated knowledge of the outlying units.

2. In the **robust approach** standard estimation methods are modified or substituted by robust ones, with the aim of obtaining a final estimate of the parameters which is unaffected by the atypical units.
   Once the robust estimate has been obtained the outlying units can be identified and tested using the robust residuals or alternative robust measures of deviation.

In the time series context, outliers have been first discussed by Fox (1972) who introduced the innovative outliers and proposed a maximum likelihood ratio tests for their detection. Further works by Abraham and Box (1979), Bruce and Martin (1989) and Chang et al. (1988) extended the definition of outliers, proposing two different types of deviation the **innovation outliers** ($IO$), that coincides with the innovative outliers proposed by Fox, and the **additive outliers** ($AO$).
Tsay (1988), using the intervention model, generalized further the definition of outliers including **level shifts** ($LS$) and **change of variance** ($CV$) points.
Despite $LS$ and $CV$ do not represent outliers in the strict sense, these can affect and alter the final fit of a model significantly.
As a consequence, being able to recognize which units are potential outliers and the corresponding type of outliers might result of great relevance in several applications.

## 5.1 Types of outliers

In order to describe the effect that the different outliers have on a time series we will use the $ARIMA$ model.
Thus, we assume that a time series $v_t$ is generated by an underlying $ARIMA(p, d, q)$ model

$$\phi(B) \bigtriangledown^d v_t = \theta(B)\epsilon_t$$

which have been presented in Chapter 3.
We recall that when the invertibility and causality conditions are satisfied the model can be specified in its $AR(\infty)$ form

$$\pi(B)v_t = \epsilon_t$$

where $\pi(B) = \bigtriangledown^d\Phi(B)/\theta(B) = 1 - \pi_1 B - \pi_2 B^2 - \dots$ or in the $MA(\infty)$ form as

$$v_t = \psi(B)\epsilon_t$$

where $\psi(B) = \theta(B)/\phi(B)\bigtriangledown^d$.
In order to take into account the possible contaminations that can occur we suppose that we do not observe the real time series $v_t$ but $y_t$ which allows for different types of contamination. Therefore, the series $y_t$ will differ from $v_t$ at those instants at which an outlier occurred.

### 5.1.1 Innovation outliers

The first type of outliers that we consider, is the $IO$, that was first introduced by Fox (1972) and represents the most harmless one.
The innovation outliers corresponds to the assumption that the innovation series $\epsilon_t$, $t = 1, \ldots, T$ is contaminated. The outliers are generated by some change or endogenous effect on the innovations of the process and the model to represent the effect of an $IO$ can be obtained by adding an impulse effect to the innovations of the original process.
The model for the observed series contaminated by an innovation outlier is

$$\phi(B) \bigtriangledown^d y_t = \theta(B)(\omega_I \delta_t^{(k)} + \epsilon_t).$$

where $\omega_I$ is the magnitude of the outlier and $\delta_t^{(k)}$ is an indicator variable that is zero except at time $t = k$ in which is equal to 1. Rearranging the model in the linear form we obtain

$$y_t = \psi(B)(\omega_I \delta_t^{(k)} + \epsilon_t)$$

where $y_t$ is the observed time series. The model shows that an innovation outlier propagates to all the innovations of the process through the weights $\psi(B)$, that depends on the particular $ARIMA(p, d, q)$ model.
It is also possible to rewrite the model in its causal form as

$$\pi(B) y_t = \omega_I \delta_t^{(k)} + \epsilon_t.$$

In general $IO$ outliers have less damaging effects on the time series than additive outliers have. The impact of an innovation outlier is represented in Figure (5.1)
An alternative method used to introduce innovation outliers is with the contamination model, defined as

$$(1 - \alpha) N(0, \sigma_\epsilon^2) + \alpha N(0, \tau \sigma_\epsilon^2)$$

where $\tau$ is a scalar greater than 1. The contamination model is widely used in robust statistics and it assumes that a $(1 - \alpha)\%$ of the data is generated from the true assumed model, while the $\alpha\%$ one is generated by a different distribution, that induces the contamination.

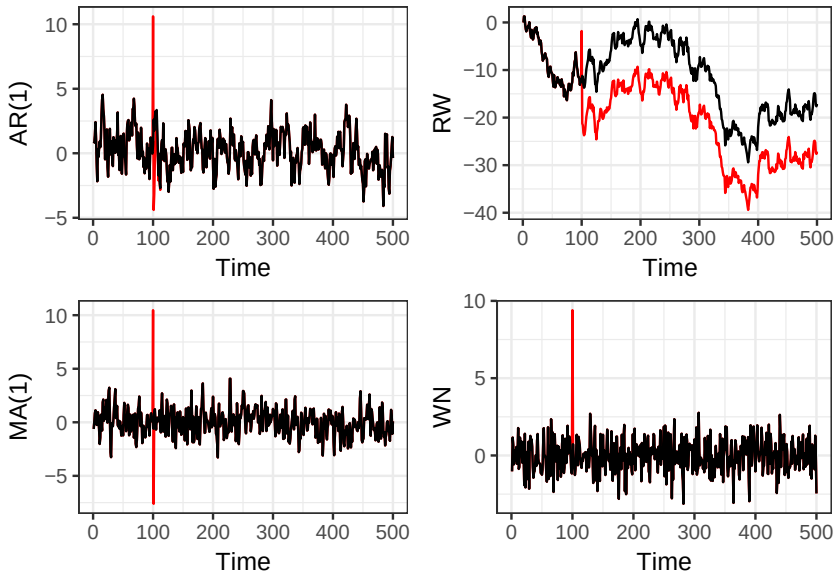


Figure 5.1: The impact on an $IO$ with magnitude $\omega_I = 10$ on four different models. Starting from the upper left corner: $Y_t = 0.7Y_{t-1} + \epsilon_t$, $Y_t = \sum_{t=1}^{t} \epsilon_t$,$Y_t = \epsilon_t - 0.7\epsilon_t$, $\epsilon_t \sim N(0, \sigma_\epsilon^2)$

### 5.1.2 Additive outliers

An additive outlier corresponds to the assumption that there is a direct contamination on the observed series. This could be given by an external error or an exogenous change of the value of the observed time series at a particular time. In this case, instead of observing the series $v_t$ we observe $y_t$ that is related to the true one as follows

$$y_t = v_t + \omega_A \delta_t^{(k)}.$$

where $\omega_A$ denotes the magnitude of the additive outlier.
The model for the observed series contaminated by an additive outlier is given by

$$\phi(B) \bigtriangledown^d (y_t - \omega_A \delta_t^{(k)}) = \theta(B)\epsilon_t.$$

The $MA(\infty)$ representation is

$$y_t = \omega_A \delta_t^{(k)} + \psi(B)\epsilon_t.$$

while, the $AR(\infty)$ is

$$\pi(B)(y_t - \omega_A \delta_t^{(k)}) = \epsilon_t.$$

An additive outlier can be interpreted as a gross error that occurs in the series at time $k$.

Compared to innovation outliers an additive outliers may have severe impact on the properties of the observed time series, in particular under the $AO$ contamination the estimates of the parameters are biased and inconsistent.
In particular, a large additive outlier will bias the estimated parameters toward zero. Moreover it can be proved in general that an additive outlier will push all the autocorrelation coefficients toward zero.

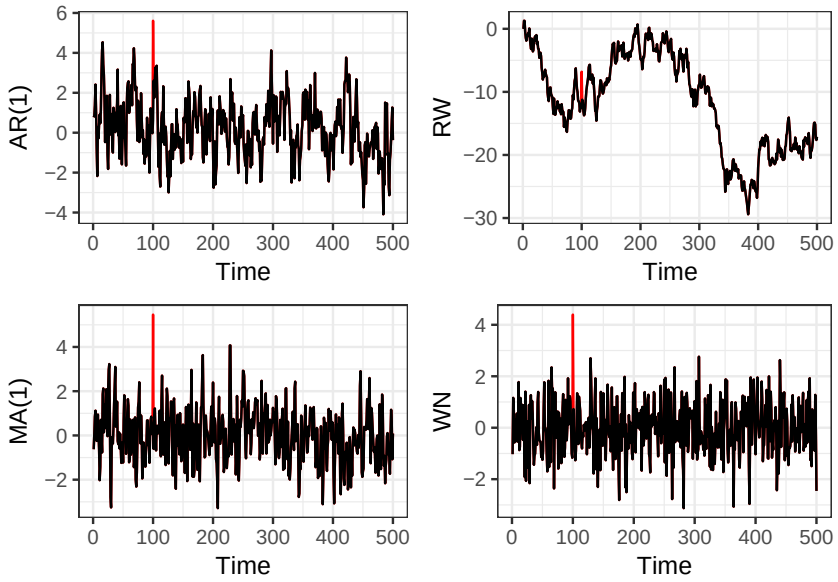


Figure 5.2: The impact on an $AO$ with magnitude $\omega_A = 5$ on four different models. Starting from the upper left corner: $Y_t = 0.7Y_{t-1} + \epsilon_t$, $Y_t = \sum_{t=1}^{t} \epsilon_t$,$Y_t = \epsilon_t - 0.7\epsilon_t$, $\epsilon_t \sim N(0, \sigma_\epsilon^2)$

### 5.1.3 Level shifts

A level shift ($LS$), represents a modification of the mean of the process, starting from a point $k$ and continuing until the end of the period observed. When the process is stationary a level shift corresponds to a change in the mean of the process after some time index and, as a consequence, the process becomes non stationary.
In presence of a level shift the observed series is related to the true one by

$$y_t = v_t + \omega_L S_t^{(k)}$$

where $S_t^{(k)}$ is a step function that takes value 0 before $k$ and 1 when $t \geq k$.
Thus, the level shift can be seen as a sequence of additive outliers all of the same size, starting at some point and lasting until the end of the observed time series.
The model for the series in the presence of level shifts is given by

$$\phi(B) \bigtriangledown^d (y_t - \omega_L S_t^{(k)}) = \theta(B)\epsilon_t.$$

Rearranging the terms leads to

$$y_t = \omega_L S_t^{(k)} + \psi(B)\epsilon_t$$

that can also be written as

$$\pi(B)(y_t - \omega_L S_t^{(k)}) = \epsilon_t.$$

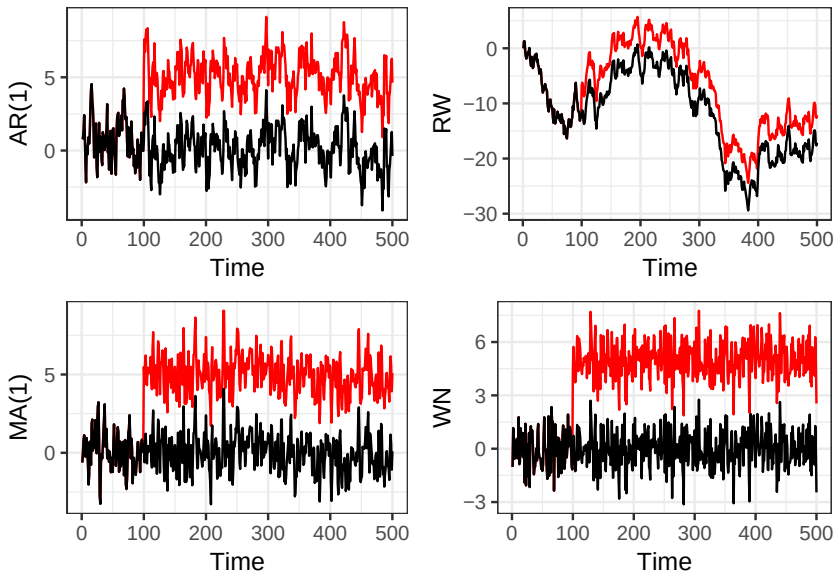


Figure 5.3: The impact on an $LS$ with magnitude $\omega_L = 5$ on four different models. Starting from the upper left corner: $Y_t = 0.7Y_{t-1} + \epsilon_t$, $Y_t = \sum_{t=1}^{t} \epsilon_t$,$Y_t = \epsilon_t - 0.7\epsilon_t$, $\epsilon_t \sim N(0, \sigma_\epsilon^2)$

### 5.1.4 Intervention model

The different type of outliers presented so far, can be considered as particular cases of interventions or deterministic effects in a time series.
The intervention model proposed by Box and Tiao (1975), to model the effect of a dynamic change in a time series at known time, can be used as a general framework to describe the impact that the different types of outliers have on the observed time series.
In particular, Box and Tiao (1975), studied the time series of pollution level in Los Angeles and investigated if a known intervention, in that case diversion of traffic induced by the opening of the Golden Gate freeway, had a significant effect on the pollution level.
Following Peña et al. (2011), the intervention model is given by

$$y_t = \omega V(B)\delta_t^{(k)} + \psi(B)\epsilon_t \tag{5.1}$$

where $\omega$ is a constant and $V(B)$ is the transfer function of the intervention at time $k$.

Using equation (5.1), all the types of outliers discussed before can be described.

*AO*. The additive outlier can be obtained by setting $V(B) = 1$. In this case the intervention does not have any dynamic propagation and the model reduces to

$$y_t = \omega\delta_t^{(\tau)} + \psi(B)\epsilon_t$$

*IO*. The innovation outlier is obtained when $V(B) = \psi(B)$ that leads to

$$y_t = \omega\psi(B)\delta_t^{(k)} + \psi(B)\epsilon_t$$

$$= \psi(B)(\omega\delta_t^{(k)} + \epsilon_t)$$

*LS*. A level shift can be obtained by setting $V(B) = 1/(1-B)$ that leads to

$$y_t = \frac{\omega}{(1-B)}\delta_t^{(k)} + \psi(B)\epsilon_t$$

$$= \omega S_t^{(k)} + \psi(B)\epsilon_t$$

where $S_t^{(k)} = (1-B)^{-1}\delta_t^{(k)}$.

Further, adopting the intervention model it is possible to specify other types of outliers. For instance, the effect of a level shift might be assumed to decrease with time and that after some time it disappears. A way to model this kind of behavior is with a **transitory change** ($TC$) proposed by Tsay (1988), with the target to describe a $LS$ that dies out exponentially. This effect is obtained by setting $V(B) = (1-\gamma B)^{-1}$, that is

$$y_t = \frac{\omega}{(1-\gamma B)}\delta_t^{(k)} + \psi(B)\epsilon_t$$

As a special case we hat that when $\gamma = 1$ the transitory change reduces to $LS$, while if $\gamma = 0$ it reduces to an additive outlier.

Another type of outlier that can be obtained is the so called **ramp-shift** outlier proposed by Chen and Tiao (1990). The intervention model in this case becomes

$$y_t = \omega R_t^{(k)} + \psi(B)\epsilon_t$$

where $R^{(k)}$ is the ramp effect, that is $R^{(k)} = 0$ for $t < k$, and $R_t^{(k)} = j$ for $t = k + j$, $j = 1, 2, \ldots$. The ramp shift outlier will produce a change in the slope of the series after $k$ times.

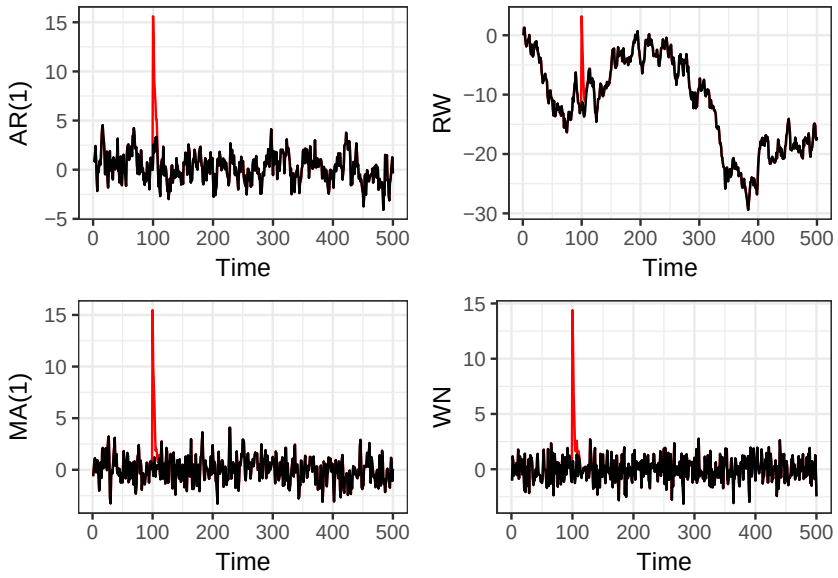


Figure 5.4: The impact on an $TC$ with magnitude $\omega_{TC} = 15$ on four different models. Starting from the upper left corner: $Y_t = 0.7Y_{t-1} + \epsilon_t$, $Y_t = \sum_{t=1}^{t} \epsilon_t$,$Y_t = \epsilon_t - 0.7\epsilon_t$, $\epsilon_t \sim N(0, \sigma_\epsilon^2)$

It is also possible to consider the presence of multiple type of outliers in the same series as well as the presence of several outliers at the same time. This case has been studied by Muirhead (1986) and Abraham and Yatawara (1988). In the case of a stationary model considering innovation and additive outliers at the same index leads to a more general form of intervention model.

## 5.2 Estimate the magnitude of the outlier

As we have presented the intervention model provides a general framework that can be used to evaluate the impact of an outlier on an $ARIMA$ model and its extensions.
In particular, using the intervention model Chang et al. (1988) showed that the information about a possible $IO$ or $AO$ is completely contained in the residuals.
Assuming that the parameters of the $ARIMA$ model are known, then it is possible to consider how to estimate the size of the outlier at a given time $k$ for the different type of outliers considered so far, respectively, $IO$, $AO$, $LS$ and $TC$.
The model for a residual as proposed by Peña is given by

$$e_t = \omega_i x_t + \epsilon_t \tag{5.2}$$

where $e_t = \pi(B)$ and $i = \{I, A, LS, TC\}$.
Depending on the structure of $\omega_i$ and $x_t$ the model can be used to represent the residuals with the presence of different types of contamination.
In particular:

*AO*. For an additive outlier we have

$$e_t = \omega_A \pi(B)\delta_t^{(k)} + \epsilon_t$$

where, $\omega_i = \omega_A$ and $x_t = \pi(B)\delta_t^{(k)}$.

*IO*. For an innovation outlier we have

$$e_t = \omega_I \delta_t^{(k)} + \epsilon_t$$

where, $\omega_i = \omega_I$ and $x_t = \delta_t^{(k)}$.

*LS*. For a level-shift outlier we have

$$e_t = \omega_L \frac{\pi(B)}{(1-B)} + \epsilon_t$$

where, $\omega_i = \omega_L$ and $x_t = \pi(B)(1-B)^{-1}$.

*TC*. Finally, for a temporary change the model becomes

$$e_t = \omega_{TC}\frac{\pi(B)}{(1-\gamma B)} + \epsilon_t$$

where, $\omega_i = \omega_{TC}$ and $x_t = \pi(B)(1-\gamma B)^{-1}$.

Assuming the model parameters known the magnitude of the outlier $\omega_i$, can be simply estimated by least squares.
That is,

$$\hat{\omega}_i = \arg\min_{\omega_i} \sum_{t=1}^{T} (e_t - \omega_i x_t)^2. \tag{5.3}$$

Differentiating (5.3) and setting the derivatives equal to zero leads to

$$\hat{\omega}_i = \frac{\sum_{t=1}^{T} e_t x_t}{\sum_{t=1}^{T} x_t^2},$$

with variance given by

$$\mathbb{V}(\hat{\omega}_i) = \sigma_\epsilon^2 \Big( \sum_{t=1}^{T} x_t^2 \Big)^{-2}.$$

Following the above procedure it is possible to obtain the estimators for the magnitude of the different type of outliers considered.
These are given by

*AO*. In the case of an additive outlier the estimator is

$$\hat{\omega}_A = \rho_A^2 \pi(F) e_t$$

where $\rho^2 = (1 + \pi_1^2 + \cdots + \pi_{T-k}^2)^{-1}$ and $F$ is the forward operator defined by $Fy_t = y_{t+1}$.

*IO*. For innovation outliers the only residual that has information about the contamination is the one at $k$ and the corresponding estimator is

$$\hat{\omega}_I = e_k$$

*LS*. For a level shift outlier the estimator of the magnitude is

$$\hat{\omega}_{LS} = \rho^2_{LS}\, l(F)e_t$$

where $l(F) = \pi(F)/(1-F)$ and $\rho^2_{LS} = (1 + l_1^2 + l_2^2 + \cdots + l^2_{T-k})^{-1}$.

$TC$. Last but not least for the $TC$ case we have

$$\hat{\omega}_{TC} = \rho^2_{TC}\, \beta(F)e_t$$

where $\rho^2_{TC} = (1 + \beta_1^2 + \beta_2^2 + \cdots + \beta^2_{T-k})^{-1}$, in which $\beta_i$ are the coefficients of $\beta(F) = \pi(F)/(1-\gamma F)$.

### 5.2.1 Tests for outliers detection

A direct consequence of the possibility of estimating the magnitude of the outlier is the possibility of constructing statistical tests.
In order to test whether an outlier of a specific type as occurred at time $k$ the test is

$$H_0 : \omega_i = 0$$
$$H_1 : \omega_i \neq 0$$

where $i = \{A, I, LS, TC\}$. When the parameters are known we can use the residuals $e_k$ derived previously, and the test under the model (5.2) is equivalent to testing the slope in a simple regression model.
Therefore, the likelihood ratio criterion leads to the general test statistic

$$\lambda_{i,t} = \frac{\hat{\omega}_{i,t}}{\rho_{i,t}\sigma_\epsilon} \tag{5.4}$$

which has a student $t$ distribution.
When the location of the outlier is unknown it is possible to use the alternative statistics $\lambda_t^* = \max\{\lambda_{j,t}\}$. However, the sample distribution of $\lambda^*_{j,t}$ is particularly complicated because of the correlation between the $\lambda_{j,t}$. The percentiles of the distribution have been obtained by simulations by Chang et al. (1988), while subsequently Ljung (1993) suggested some approximations to this distribution.
Several procedures based on these tests, or on some variation,s have been developed in order to identify potential outliers inside a time series. The next chapter reviews the most

common ones.

## 5.3 The impact of $AO$ on the exact maximum likelihood estimates: a simulation study

Having discussed, so far, the theoretical characteristics of the different outliers, we conducted a simulation study to evaluate the impact that the $AO$ might have on the estimates of the parameters when the latter are obtained maximizing the exact likelihood.
The simulation study considers three different levels of contamination, three different sample sizes and three different outliers magnitude. For each scenario we estimated via Monte-Carlo method the bias of the exact maximum likelihood estimator. The estimates are based on 10000 Monte-Carlo replicates and the simulation has been conducted for $AR$ and $MA$ processes
The three autoregressive models considered in the simulation study are:

1. $Y_t = 0.6Y_{t-1} + \epsilon_t$
2. $Y_t = 0.4Y_{t-1} + 0.2Y_{t-2} + \epsilon_t$
3. $Y_t = 0.2Y_{t-1} + 0.3Y_{t-2} + 0.2Y_{t-3} + \epsilon_t$

while the three moving average models are:

1. $Y_t = 0.6\epsilon_{t-1} + \epsilon_t$
2. $Y_t = 0.4\epsilon_{t-1} + 0.2\epsilon_{t-2} + \epsilon_t$
3. $Y_t = 0.2\epsilon_{t-1} + 0.3\epsilon_{t-2} + 0.2\epsilon_{t-3} + \epsilon_t$

where $\epsilon_t \sim WN(0, \sigma_\epsilon^2)$.
We will first consider the bias introduced by the introduction of several additive outliers in the autoregressive model and next we will consider the moving average case.

### 5.3.1 Autoregressive models

Table 5.1: $Y_t = 0.6Y_{t-1} + \epsilon_t$

$\upsilon = 0$

| $T$ | $Bias(\hat{\phi}_1)$ |
|---|---|
| 100 | 0.012 |
| 300 | 0.004 |
| 500 | 0.003 |

$\upsilon = 0.05$

| | $T$ | $Bias(\hat{\phi}_1)$ |
|---|---|---|
| $\omega_A = 5$ | 100 | 0.229 |
| | 300 | 0.221 |
| | 500 | 0.221 |
| $\omega_A = 10$ | 100 | 0.390 |
| | 300 | 0.384 |
| | 500 | 0.385 |
| $\omega_A = 15$ | 100 | 0.461 |
| | 300 | 0.456 |
| | 500 | 0.457 |

$\upsilon = 0.1$

| | $T$ | $Bias(\hat{\phi}_1)$ |
|---|---|---|
| $\omega_A = 5$ | 100 | 0.283 |
| | 300 | 0.279 |
| | 500 | 0.277 |
| $\omega_A = 10$ | 100 | 0.416 |
| | 300 | 0.414 |
| | 500 | 0.412 |
| $\omega_A = 15$ | 100 | 0.468 |
| | 300 | 0.466 |
| | 500 | 0.465 |

$\upsilon = 0.2$

| | $T$ | $Bias(\hat{\phi}_1)$ |
|---|---|---|
| $\omega_A = 5$ | 100 | 0.308 |
| | 300 | 0.305 |
| | 500 | 0.304 |
| $\omega_A = 10$ | 100 | 0.409 |
| | 300 | 0.407 |
| | 500 | 0.406 |
| $\omega_A = 15$ | 100 | 0.447 |
| | 300 | 0.445 |
| | 500 | 0.444 |

Table 5.3: $Y_t = 0.4Y_{t-1} + 0.2Y_{t-2} + \epsilon_t$

$\upsilon = 0$

| $T$ | $Bias(\hat{\phi}_1)$ | $Bias(\hat{\phi}_2)$ |
|---|---|---|
| 100 | 0.005 | 0.015 |
| 300 | 0.001 | 0.006 |
| 500 | 0.001 | 0.004 |

$\upsilon = 0.05$

| | $T$ | $Bias(\hat{\phi}_1)$ | $Bias(\hat{\phi}_2)$ |
|---|---|---|---|
| $\omega_A = 5$ | 100 | 0.149 | 0.049 |
| | 300 | 0.146 | 0.039 |
| | 500 | 0.146 | 0.036 |
| $\omega_A = 10$ | 100 | 0.251 | 0.103 |
| | 300 | 0.249 | 0.094 |
| | 500 | 0.249 | 0.092 |
| $\omega_A = 15$ | 100 | 0.298 | 0.134 |
| | 300 | 0.296 | 0.126 |
| | 500 | 0.297 | 0.124 |

$\upsilon = 0.1$

| | $T$ | $Bias(\hat{\phi}_1)$ | $Bias(\hat{\phi}_2)$ |
|---|---|---|---|
| $\omega_A = 5$ | 100 | 0.183 | 0.064 |
| | 300 | 0.181 | 0.055 |
| | 500 | 0.181 | 0.053 |
| $\omega_A = 10$ | 100 | 0.268 | 0.114 |
| | 300 | 0.267 | 0.106 |
| | 500 | 0.267 | 0.104 |
| $\omega_A = 15$ | 100 | 0.302 | 0.137 |
| | 300 | 0.302 | 0.129 |
| | 500 | 0.302 | 0.128 |

$\upsilon = 0.2$

| | $T$ | $Bias(\hat{\phi}_1)$ | $Bias(\hat{\phi}_2)$ |
|---|---|---|---|
| $\omega_A = 5$ | 100 | 0.199 | 0.071 |
| | 300 | 0.197 | 0.063 |
| | 500 | 0.196 | 0.061 |
| $\omega_A = 10$ | 100 | 0.263 | 0.110 |
| | 300 | 0.261 | 0.102 |
| | 500 | 0.260 | 0.100 |
| $\omega_A = 15$ | 100 | 0.288 | 0.127 |
| | 300 | 0.286 | 0.119 |
| | 500 | 0.285 | 0.117 |

Table 5.5: $Y_t = 0.2Y_{t-1} + 0.3Y_{t-2} + 0.2Y_{t-3} + \epsilon_t$

$\upsilon = 0$

| $T$ | $Bias(\hat{\phi}_1)$ | $Bias(\hat{\phi}_2)$ | $Bias(\hat{\phi}_3)$ |
|---|---|---|---|
| 100 | 0.005 | 0.018 | 0.007 |
| 300 | 0.001 | 0.007 | 0.002 |
| 500 | 0.000 | 0.004 | 0.002 |

$\upsilon = 0.05$

| | $T$ | $Bias(\hat{\phi}_1)$ | $Bias(\hat{\phi}_2)$ | $Bias(\hat{\phi}_3)$ |
|---|---|---|---|---|
| $\omega_A = 5$ | 100 | 0.044 | 0.112 | 0.046 |
| | 300 | 0.041 | 0.099 | 0.042 |
| | 500 | 0.041 | 0.098 | 0.042 |
| $\omega_A = 10$ | 100 | 0.093 | 0.185 | 0.095 |
| | 300 | 0.092 | 0.174 | 0.092 |
| | 500 | 0.092 | 0.173 | 0.092 |
| $\omega_A = 15$ | 100 | 0.121 | 0.222 | 0.124 |
| | 300 | 0.121 | 0.212 | 0.121 |
| | 500 | 0.121 | 0.211 | 0.122 |

$\upsilon = 0.1$

| | $T$ | $Bias(\hat{\phi}_1)$ | $Bias(\hat{\phi}_2)$ | $Bias(\hat{\phi}_3)$ |
|---|---|---|---|---|
| $\omega_A = 5$ | 100 | 0.060 | 0.136 | 0.060 |
| | 300 | 0.057 | 0.124 | 0.057 |
| | 500 | 0.056 | 0.122 | 0.057 |
| $\omega_A = 10$ | 100 | 0.106 | 0.199 | 0.106 |
| | 300 | 0.103 | 0.189 | 0.102 |
| | 500 | 0.103 | 0.187 | 0.103 |
| $\omega_A = 15$ | 100 | 0.128 | 0.226 | 0.128 |
| | 300 | 0.125 | 0.217 | 0.125 |
| | 500 | 0.125 | 0.215 | 0.125 |

$\upsilon = 0.2$

| | $T$ | $Bias(\hat{\phi}_1)$ | $Bias(\hat{\phi}_2)$ | $Bias(\hat{\phi}_3)$ |
|---|---|---|---|---|
| $\omega_A = 5$ | 100 | 0.067 | 0.144 | 0.070 |
| | 300 | 0.065 | 0.135 | 0.065 |
| | 500 | 0.064 | 0.134 | 0.064 |
| $\omega_A = 10$ | 100 | 0.106 | 0.199 | 0.106 |
| | 300 | 0.100 | 0.185 | 0.101 |
| | 500 | 0.099 | 0.184 | 0.099 |
| $\omega_A = 15$ | 100 | 0.118 | 0.212 | 0.120 |
| | 300 | 0.116 | 0.205 | 0.116 |
| | 500 | 0.115 | 0.204 | 0.115 |

For all the three sample size considered the exact maximum likelihood estimates results essentially unbiased when no outliers are introduced in the series, i.e when $\upsilon = 0$.
For a fixed sample size, when a mild 5% contamination is introduced immediately the bias increase considerably, this is a common and general phenomena observed for all the three autoregressive models considered in the simulations.
Moreover, as the contamination becomes larger i.e $\upsilon = 0.1$ or $\upsilon = 0.2$, the bias tend to increase more slowly. That is to say, for a given sample size, we expect to see similar bias across different level of contamination, if the latter is particularly large. As an example, in the $AR(1)$ model when $n = 100$ and a 5% contamination is introduced the bias increases to 0.309 while when the contamination is increased from 10% to 20% the bias increases only by a small amount, approximately 0.20.
It is important to note that the main increase on the bias of the estimator it is not provoked by the amount of contamination but by the magnitude of the additive outliers introduced in the series.
To see this consider the $AR(1)$ model and focus for example on the case with $n = 300$. When $\upsilon = 0.05$ and the magnitude of the additive outliers introduced is $\omega_A = 10$ the bias is equal to 0.384. When the contamination is equal to $\upsilon = 0.2$ but the magnitude considered is $\omega_A = 5$ the bias results 0.305.
To conclude the main effect on the bias of the estimator is generated not by the amount of outliers in the series but by their magnitudes.
Next, it is possible to consider the results for the $AR(2)$ and $AR(3)$ models. In these cases it is interesting to note the bias it is different for the different parameters considered.
In the first case the bias is larger for the $\hat{\phi}_1$ estimator while in the latter case the bias is larger for the $\hat{\phi}_2$ estimator.
As a general pattern across all the models as the sample size increase the bias for a given proportion of contamination and outliers magnitude tends to decrease.
To conclude we mention an interesting tendency that emerges in all the autoregressive models considered. For a given sample size, when the magnitude of the additive outliers is either $\omega_A = 10$ or $\omega_A = 15$ and the contamination $\upsilon$ is increased from 0.1 to 0.2 the bias tends to sensibly decrease. However this effect might be provoked by rounding errors.

### 5.3.2 Moving Average Models

Table 5.7: $Y_t = 0.6\epsilon_{t-1} + \epsilon_t$

$\upsilon = 0$

| $T$ | $Bias(\hat{\theta}_1)$ |
|---|---|
| 100 | -0.006 |
| 300 | -0.002 |
| 500 | -0.001 |

$\upsilon = 0.05$

| | $T$ | $Bias(\hat{\phi}_1)$ |
|---|---|---|
| $\omega_A = 5$ | 100 | 0.309 |
| | 300 | 0.310 |
| | 500 | 0.309 |
| $\omega_A = 10$ | 100 | 0.449 |
| | 300 | 0.449 |
| | 500 | 0.448 |
| $\omega_A = 15$ | 100 | 0.502 |
| | 300 | 0.501 |
| | 500 | 0.501 |

$\upsilon = 0.1$

| | $T$ | $Bias(\hat{\theta}_1)$ |
|---|---|---|
| $\omega_A = 5$ | 100 | 0.358 |
| | 300 | 0.360 |
| | 500 | 0.360 |
| $\omega_A = 10$ | 100 | 0.466 |
| | 300 | 0.468 |
| | 500 | 0.468 |
| $\omega_A = 15$ | 100 | 0.504 |
| | 300 | 0.506 |
| | 500 | 0.506 |

$\upsilon = 0.2$

| | $T$ | $Bias(\hat{\theta}_1)$ |
|---|---|---|
| $\omega_A = 5$ | 100 | 0.379 |
| | 300 | 0.381 |
| | 500 | 0.382 |
| $\omega_A = 10$ | 100 | 0.460 |
| | 300 | 0.461 |
| | 500 | 0.462 |
| $\omega_A = 15$ | 100 | 0.489 |
| | 300 | 0.489 |
| | 500 | 0.490 |

Table 5.9: $Y_t = 0.4\epsilon_{t-1} + 0.2\epsilon_{t-1} + \epsilon_t$

$\upsilon = 0$

| $T$ | $Bias(\hat{\theta_1})$ | $Bias(\hat{\theta_2})$ |
|---|---|---|
| 100 | -0.004 | -0.004 |
| 300 | -0.001 | -0.001 |
| 500 | -0.001 | -0.001 |

$\upsilon = 0.05$

| | $T$ | $Bias(\hat{\theta_1})$ | $Bias(\hat{\theta_2})$ |
|---|---|---|---|
| $\omega_A = 5$ | 100 | 0.174 | 0.096 |
| | 300 | 0.176 | 0.098 |
| | 500 | 0.176 | 0.097 |
| $\omega_A = 10$ | 100 | 0.277 | 0.146 |
| | 300 | 0.278 | 0.147 |
| | 500 | 0.278 | 0.147 |
| $\omega_A = 15$ | 100 | 0.318 | 0.165 |
| | 300 | 0.319 | 0.166 |
| | 500 | 0.319 | 0.166 |

$\upsilon = 0.1$

| | $T$ | $Bias(\hat{\theta_1})$ | $Bias(\hat{\theta_2})$ |
|---|---|---|---|
| $\omega_A = 5$ | 100 | 0.209 | 0.115 |
| | 300 | 0.212 | 0.116 |
| | 500 | 0.211 | 0.115 |
| $\omega_A = 10$ | 100 | 0.291 | 0.155 |
| | 300 | 0.293 | 0.154 |
| | 500 | 0.292 | 0.154 |
| $\omega_A = 15$ | 100 | 0.321 | 0.169 |
| | 300 | 0.323 | 0.167 |
| | 500 | 0.321 | 0.167 |

$\upsilon = 0.2$

| | $T$ | $Bias(\hat{\theta_1})$ | $Bias(\hat{\theta_2})$ |
|---|---|---|---|
| $\omega_A = 5$ | 100 | 0.224 | 0.120 |
| | 300 | 0.224 | 0.122 |
| | 500 | 0.225 | 0.122 |
| $\omega_A = 10$ | 100 | 0.285 | 0.149 |
| | 300 | 0.284 | 0.150 |
| | 500 | 0.285 | 0.150 |
| $\omega_A = 15$ | 100 | 0.307 | 0.159 |
| | 300 | 0.306 | 0.160 |
| | 500 | 0.306 | 0.160 |

Table 5.11: $Y_t = 0.3\epsilon_{t-3} + 0.2\epsilon_{t-2} + 0.3\epsilon_{t-1} + \epsilon_t$

$\upsilon = 0$

| $T$ | $Bias(\hat{\theta_1})$ | $Bias(\hat{\theta_2})$ | $Bias(\hat{\theta_3})$ |
|---|---|---|---|
| 100 | -0.001 | -0.007 | -0.007 |
| 300 | -0.001 | -0.001 | -0.002 |
| 500 | -0.001 | -0.001 | -0.001 |

$\upsilon = 0.05$

| | $T$ | $Bias(\hat{\theta_1})$ | $Bias(\hat{\theta_2})$ | $Bias(\hat{\theta_3})$ |
|---|---|---|---|---|
| $\omega_A = 5$ | 100 | 0.069 | 0.134 | 0.092 |
| | 300 | 0.071 | 0.136 | 0.096 |
| | 500 | 0.071 | 0.136 | 0.096 |
| $\omega_A = 10$ | 100 | 0.123 | 0.212 | 0.144 |
| | 300 | 0.125 | 0.212 | 0.146 |
| | 500 | 0.124 | 0.211 | 0.146 |
| $\omega_A = 15$ | 100 | 0.147 | 0.242 | 0.164 |
| | 300 | 0.149 | 0.242 | 0.164 |
| | 500 | 0.148 | 0.241 | 0.165 |

$\upsilon = 0.1$

| | $T$ | $Bias(\hat{\theta_1})$ | $Bias(\hat{\theta_2})$ | $Bias(\hat{\theta_3})$ |
|---|---|---|---|---|
| $\omega_A = 5$ | 100 | 0.089 | 0.162 | 0.113 |
| | 300 | 0.089 | 0.163 | 0.113 |
| | 500 | 0.089 | 0.163 | 0.114 |
| $\omega_A = 10$ | 100 | 0.134 | 0.223 | 0.152 |
| | 300 | 0.133 | 0.222 | 0.152 |
| | 500 | 0.133 | 0.223 | 0.153 |
| $\omega_A = 15$ | 100 | 0.151 | 0.244 | 0.166 |
| | 300 | 0.150 | 0.243 | 0.165 |
| | 500 | 0.150 | 0.244 | 0.166 |

$\upsilon = 0.2$

| | $T$ | $Bias(\hat{\theta_1})$ | $Bias(\hat{\theta_2})$ | $Bias(\hat{\theta_3})$ |
|---|---|---|---|---|
| $\omega_A = 5$ | 100 | 0.095 | 0.173 | 0.117 |
| | 300 | 0.096 | 0.173 | 0.121 |
| | 500 | 0.096 | 0.173 | 0.120 |
| $\omega_A = 10$ | 100 | 0.134 | 0.223 | 0.152 |
| | 300 | 0.129 | 0.217 | 0.150 |
| | 500 | 0.129 | 0.217 | 0.149 |
| $\omega_A = 15$ | 100 | 0.141 | 0.235 | 0.157 |
| | 300 | 0.141 | 0.232 | 0.160 |
| | 500 | 0.142 | 0.232 | 0.159 |

The conclusions obtained for the moving average models are the same as the one already discussed for the autoregressive models.
Also in this case the main effect on the bias is provoked by the magnitude of the outlier and few additive outliers with a large magnitude have a greater effect than a large number of additive outliers but with a small magnitude.
When the $MA(2)$ and $MA(3)$ models are considered also in this case the bias depends on the parameter that is estimated. In the first case $\hat{\theta}_1$ is the estimator with the largest bias, while in the latter $\hat{\theta}_2$ is the estimator with the largest bias.

# Chapter 6

# Review of outliers diagnostic procedures

In this chapter we provide a review of some of the most common outlier detection techniques that have been proposed in the statistical literature. On the basis of them we decided to try to extend the Forward Search to the time series context, in particular to the class of $SARIMA$ models.
The outlier detection techniques will be presented in chronological order. We want to remark that a detailed description of these techniques is beyond the scope of this thesis and for further results and/or explanations the reader is referred to Tsay (1988), Abraham and Chuang (1989), Chen and Liu (1993b), Hendry (1999) and Marczak and Proietti (2016).

## 6.1 Ruey Tsay - Journal of Forecasting (1988)

The method proposed by Tsay can be used to detect several types of outliers. In particular he considered, $AO$, $IO$, $LS$, $TC$ and $VC$.
The number of outlying points and the corresponding time points are unknown. In particular he proposed two different techniques; one that can be applied to detect innovation, additive and level shift outliers and the other to identify possible change of variance points. Both of these techniques are composed by two iterative procedures for detecting and modeling the atypical observations. The scheme elaborated by Tsay resumes in the following steps:

(S1) **Specification**

(S2) **Estimation**

(S3) **Detection**

(S4) **Removal**

were, at each iteration, the maximum of a selected test statistics is selected as the candidate test for that type of disturbance, and the grand maximum across the statistics is identified as the most likely type of outlier.
The grand maximum, is then tested against a specified critical value so the existence of the disturbance can be confirmed by statistical evidence. The procedure to detect change of variance relies on the same structure but some modifications are made. We will start by considering this last.

### 6.1.1 Procedure V: variance changes

In order to detect variance changes Tsay proposed the following scheme.

Step 1. Under the assumptions that no variance changes occurred, specify an $ARMA$ model for the observed time series $y_t$ and obtain the parameters estimate and the residuals for the assumed model.

Step 2: Compute the variance ratio using the residuals obtained from the parameters estimated at Step 1 and obtain the minimum and the maximum of the variance ratio, call them $\lambda_{min}$ and $\lambda_{max}$. Where

$$\lambda_{min} = Min\{\hat{r}_d : h \leq d \leq n-h\}$$
$$\lambda_{max} = Max\{\hat{r}_d : h \leq d \leq n-h\}$$

and $h$ is a positive integer denoting the minimum number of residuals used to estimate the variance.

Step 3: Let $\lambda^{\dagger} = Max\{\lambda_{max}, \lambda_{min}^{-1}\}$ and compare $\lambda^{\dagger}$ with a critical value $C_{\alpha}$. If $\lambda^{\dagger} < C_{\alpha}$, there is no significance evidence of variance change at level $\alpha$ . Conversely, if $\lambda^{\dagger} \geq C_{\alpha}$ a variance change is detected. The time point at which the variance change occurs is denoted $t^{\dagger}$.

Step 4: Adjust the variance change in the observed time series $y_t$ as

$$y_t^* = \begin{cases} y_t & if \quad t < t^{\dagger} \\ \overline{y}_t + (\lambda^{\dagger})^{-1/2}(y_t - \overline{y}_t) & if \quad t \geq t^{\dagger} \end{cases}$$

where $\overline{y}_t$ is the sample mean of the series $y_t$, and $y_t^*$ is the resulting corrected time series.

Next, go to Step 1 and consider $y_t^*$ as the observed time series and repeat the process until no more change of variance are detected.

Since the distribution of $\lambda^\dagger$ is particularly hard due to the correlation structure, based on simulations Tsay suggests the values $3.5, 3.0$ or $2.5$ as possible candidate values for the threshold $C_\alpha$.

### 6.1.2 Procedure M: exogenous disturbances

Procedure M is concerned about the detection of disturbances that affects the mean level of $v_t$, the true unobservable series.
With the aim of simplifying the computation, the iteration here is divided in inner and outer loops.

**Outer Iteration:**

Step A. Under the assumption that no disturbances are observed, specify an $ARMA$ model for the observed time series $y_t$ and obtain the parameter estimates for the postulated model. Further, select a critical value $C_\alpha$

**Inner Iteration:**

Step A1. Using the parameter estimates obtained at Step A compute the residuals and the residual variance.

Step A2. On the basis of the residuals and the residual variance obtained at step A1, compute the test statistics presented in Chapter 5 for every time $t$.
Next, for every test statistic, locate its maximum in absolute value, and let $t_i$, $i = \{I, A, LS, TC\}$ be the time point where the maximum for the $i$th statistics occurred.

Step A3. Find the grand maximum $\lambda^\dagger = Max\{\lambda_{I,max}, \lambda_{A,max}, \lambda_{LS,max}, \lambda_{TC,max}\}$, and compare it with the critical value $C_\alpha$. If $\lambda^\dagger < C_\alpha$, no significant disturbance is found and the procedure skips directly to Step B. On the other hand if $\lambda^\dagger \geq C_\alpha$ a disturbance is detected.

Step A4. Remove the effect of the identified disturbance by subtracting from $y_t$ the magnitude $\hat{\omega}_i$ and denote the adjusted series as $y_t^*$. Go to Step A and treat the adjusted series as the observed one.

Step B. If there is no significance disturbance found in the inner iteration stop the procedure. Otherwise, go to Step A with $y_t^*$ as the observed series.

As in the procedure V the distribution of $\lambda^{\dagger}$ is extremely complicated and candidates values for the threshold $C_{\alpha}$ have been found by simulations. In particular, the values $4.0, 3.5$ and $3.0$, selected from simulation results of Chang et al. (1988), have provided satisfactory results.

## 6.2 Abraham and Chuang - Technometrics (1989)

The procedure proposed by Abraham and Chuang relies on the assumption that a given $ARMA(p,q)$ process can be approximated relatively well by an $AR(p+q)$ process.
Their model building strategy starts by fitting a sufficiently large $AR(p)$ model, subsequently, a series of iteration are repeated to identify and clean the observed time series.
A key difference from the procedure proposed by Tsay (1988), is that the method of Abraham and Chuang focuses only on the $AO$ and $IO$ outliers types, furthermore, the conditionals least squares are used to estimate the unknown parameters.
The procedure and the statistics that they use rely on the regression representation of an autoregressive process. On one hand this simplifies the estimation of the parameters since it allows to use the ordinary least squares and on the other hand it allows to extend the algebra of linear regression to time series data.

Like the method proposed by Tsay (1988) also the procedure introduced by Abrahamn and Chuang relies on a series of steps that are repeated iteratively.
The steps are performed as follows.

Step 1: **Identification**. Use any model-selection technique to identify a tentative model $ARMA(p',q')$ which may not be the true order, next approximate the identified $ARMA$ process with an $AR(p^*)$ model, where $p^* \geq p' + q'$.

Step 2: **Outlier Detection**. Estimate the parameters of the $AR(p^*)$ process with the conditional least squares method (CLS) and compute the statistics $Q_k$ (and or $Q_{k2}$) for $k = 1$ and $k = p^* + 1$. The interested reader is referred to the article for the details of these statistics. Obtain $Q_{max}$, and perform a significance test. If an outlier is detected move to Step 3, otherwise skip directly to Step 4.

Step 3: **Cleaning the series**. Let $t^{\dagger}$ be the position of the outlier detected at step 2. If the outlier is an $AO$ then delete the equations from $(t^{\dagger} - p^*)$ to $t^{\dagger}$ from the autoregressive model and obtain a new estimate of the parameters. The $t^{\dagger}$th unit is then corrected by substituting its value with $\mathbb{E}(y_{t^{\dagger}} | \mathcal{F}_{-t^{\dagger}})$.
If the outlier detected is an $IO$, delete only the $t^{\dagger}$th equation, estimate a new set of parameters and adjust the original time series as

$$y_t^* = \begin{cases} y_t - \tilde{e}_t^* & if \quad t = t^\dagger \\ y_t - \tilde{\psi}_{t-d_0}\tilde{e}_{d_0}^* & if \quad t > t^\dagger \end{cases}$$

where $\tilde{e}_t^*$ is the residual obtained from the new parameter estimates and $\tilde{\psi}_j^*$ is the coefficient of $B^j$ in the causal representation of the process.

Step 4: **Specification**. Use the cleaned series in the last iteration to specify a new tentative model. This model is estimated using maximum likelihood and the iterative strategy is repeated.

The advantage of this procedure compared to the one discussed previously is that it takes into account the fact that a model might be misspecified due to the effect of outlying units.
However this automatic method presents several drawbacks. First, when an outlier is not identified, that is to say, a good observation in the series is wrongly classified as an outlier or an $IO$ outlier is identified as an $AO$ outlier, or vice-versa, the adjusted series may lead to a misspecified model.
Second, the use of the autoregressive approximation used to simplify the estimation process may lead to biased estimates especially when short time series are considered.
Last but not least, the mere use of information criteria for the specification of a model may lead to poor results. In fact, in practice it may happen that different information criteria lead to different model specifications.

## 6.3 Chen and Liu - Journal of the American Statistical Association (1993)

Despite the procedures discussed previously are quite effective in detecting the location and estimating the effects of the outliers, some issues still remain. In particular, Chen and Liu (1993a), pointed out the following remarks:

(a) The presence of outliers may result in the specification of an inappropriate model.

(b) Even if the model is appropriately specified outliers in a time series may still produce bias in parameter estimates and hence may affect the efficiency of outlier detection.

(c) Some outliers might not be identified due to a masking effect.

In their work Chen and Liu attempt to resolve the problems in points b and c above.
The iterative procedure that they proposed is articulated in 3 steps and the inclusion of

an inner loop resemble an extension of Tsay's procedure.
Furthermore, like the one introduced by Tsay (1988) this procedure does not relies on the autoregressive approximation it makes use of some extensions of the outlying statistics discussed in Chapter 5.

### 6.3.1 The Detection and Estimation Procedure

**Stage 1: Initial Parameter Estimation and Outlier Detection**

I.1. Compute the maximum likelihood estimates of the model parameters and obtain the corresponding residuals $e_t$, $t = 1, \ldots, T$.

**Inner Loop of Outlier Detection for Fixed Model Parameter Estimates**

I.2. For $t = 1, \ldots, T$, compute $\hat{\lambda}_I(t), \hat{\lambda}_A(t), \hat{\lambda}_{LS}(t)$ and $\hat{\lambda}_{TC}(t)$ using the residuals $e_t$. and let $\hat{\lambda}_{max}(t)$ be

$$\hat{\lambda}_{max}(t) = max\{|\hat{\lambda}_I(t)|, |\hat{\lambda}_A(t)|, |\hat{\lambda}_{LS}(t)|, |\hat{\lambda}_{TC}(t)|\}$$

If $\hat{\lambda}_{max}(t) > C_\alpha$, where $C_\alpha$ is a predetermined critical value then there is the possibility of an outlier of a given type.

I.3. If no outlier is found skip directly to step I.4. otherwise correct the observation accordingly to the type of outlier identified, and repeat step I.2. until no remaining outliers are identified.

I.4. If no outliers are found in the inner loop the series does not contain any outlier and the procedure can be stopped. If outliers are found in this inner loop move back to step I.1. and revise the parameters estimate. If one or more outliers are detected in the inner loops and no more outliers are found, move to step II.1.

**Stage II: Joint Estimation of Outlier Effects and Model Parameters**

II.1. Suppose that $m$ time points $t_1, \ldots, t_m$ are identified as possible outliers of different kinds. The outliers effects are then estimated jointly.

II.2. Compute the $\hat{\lambda}$ statistics for the estimated $\hat{\omega}_j$. If $min|\hat{\lambda}_j| \leq C$ , where $C$ is the same critical value used in step I.2, then delete the outlier at time point $t_0$ from the set of the identified outliers and go to step II.1 with the remaining $m - 1$ outliers, otherwise move to step II.3.

II.3. Obtain the adjusted series by removing the outlier effects that resulted significant at iteration II.1 and II.2.

II.4. Compute the maximum likelihood estimates of the model parameters based on the adjusted time series. If the relative change of the residual standard error from the previous estimate is greater than $\alpha$ go to step II.1 for further iterations; otherwise go to step III.1. The tolerance $\alpha$ is chosen at the beginning by the user.

**Stage III: Detection of Outliers Based on the Final Parameter Estimates**

III.1. Compute the residuals by filtering the original series based on the parameter estimates obtained at step II.4.

III.2. Use the residuals obtained at step III.1. and iterate through Stages I and II with the modifications that the parameter estimates used in the inner loop of Stage I are fixed to those obtained at step II.4 and Steps II.3 and II.4 are omitted in Stage II. The estimated $\hat{\omega}_j$ of the last iteration at step II.1 are the final estimates of the effects of the detected outliers.

## 6.4 Hendry - Methodology and Tacit Knowledge: Two Experiments in Econometrics (1999)

The statistical procedures discussed until know were all concerned about outlier detection in $ARIMA$ models. However, because of its similarities with the Forward Search the Impulse Indicator Saturation (IIS) represent an interesting method that has been first introduced by Hendry to detect a random number of breaks of unknwon duration and location in dynamic regression models.
The IIS and its extension, the Step Indicator Saturation (SIS), relies on a general-to-specific approach, where a dummy variable is added for each unit in the sample, to take into account the fact that the given unit might represent a potential outlier. The final breaks are identified as those units for which the dummy variable result significant. Since the number of parameters to estimate in the final model exceed or might be equal to the number of observations, a 3-step approach is used to obtain the final fit of the model.
Formally, a dynamic linear model in its simplest form can be represented as:

$$Y_t = \mu + \sum_{t=1}^{T} \delta_t d_t + \epsilon_t \quad \epsilon_t \sim N(0, \sigma^2)$$

where $\mu$ denotes the overall process mean, $\delta_t$ is the dummy variable added for the $t$th observation and $d_t$ is the regression coefficient associated to it.
However, since the number of parameters is greater than the number of observations an iterative procedure, the so called split-half approach, is employed.
The split-half algorithm performs the following steps.

Step 1. First, impulse indicators are added to the first $[T/2]$ half of the samples, i.e. impulse indicators are added to time points $t_1, \ldots, t_{[T/2]}$. The regression model with the $[T/2]$ indicator variables is estimated using the entire sample and dummies for wich the t-statistic resulted significant at a nominal $\alpha$ level are then flagged as potential outliers. The set of unit flagged as potential outliers in this first step is the given by:

$$\mathcal{O}_1 = \{t : t^2_{stat,t} \leq c^2_\alpha,\ t = 1, \ldots, [T/2]\}$$

where $t_{stat,t}$ denotes the t-statistic for the $t$th dummy and $c^2_\alpha$ is the t-distribution critical value corresponding to the nominal significance level $\alpha$.

Step 2. Step 1 is repeated, this time including dummies for time points $t_{T-[T/2]}, \ldots, t_T$. The dynamic model is estimated as before using the entire sample and the set $\mathcal{O}_2$, composed by the units for which the dummy resulted significant is generated. That is,

$$\mathcal{O}_2 = \{t : t^2_{stat,t} \leq c^2_\alpha,\ t = T - [T/2], \ldots, T\}$$

Step 3. Impulse dummies are added to the units defined by the set $\mathcal{O}_1 \cup \mathcal{O}_2$, and the model is re estimated using the entire sample. The set of final breaks is given by:

$$\mathcal{O}^* = \{t : t^2_{stat,t} \leq c^2_\alpha,\ t \in \mathcal{O}_1 \cup \mathcal{O}_2\}$$

On average in abscence of outliers, we would expect $\alpha T$ indicators to be retained in the final stage by chance, so that setting $\alpha = 1/T$ leads to an average misclassification of one osbervation.

Several extensions of the IIS have been proposed in literature. First several different type of dummies might be considered such as in the SIS (Doornik et al. (2013)) where step indicators are added, next different iterative procedures to estimate the final model ant the breaks might be considered such as the split-half with sequential selection (Hendry and Krolzig (2004)) which relies on the iterative elimination of the least significatng indicators in each split until only the significant ones are retained. Finally, Marczak and Proietti (Marczak and Proietti (2016)) considered an extension of the IIS the BSM structural time series model.

## 6.5 Marczak, Proietti and Grassi - Econometrics and Statistics (2018)

Marczak, Proietti and Grassi introduced for the general class of state space models a robust Kalman filter that can be used to clean the sample data and obtain unbiased estimates of the parameters as well as robust state and innovation estimates.
The idea behind this extension is to use an M-estimator to shrink the observations towards their one-step-ahead prediction by bounding the effect introduced by a new observation. In order to control for the effect of outlying units an M-estimator is used to down weight atypical units. Let $\boldsymbol{\theta}$ be a vector of parameters, an M-estimator is then defined as the minimizer of a prescribed $\rho$ function, that is

$$\hat{\boldsymbol{\theta}} = \min_{\theta} \sum_{t=1}^{T} \rho(\epsilon_t)$$

where $\epsilon_t = (y_t - \mu_t)\sigma_t^{-1}$. If the $\rho$ function is differentiable the M-estimator is thus obatined as the solution of the system of equations:

$$\sum_{t=1}^{T} \psi\left(\epsilon_t\right) \frac{\partial \epsilon_t}{\partial \theta_j}$$

where $\psi$ is the derivative of $\rho$. Strictly connected to the $\rho$ and $\psi$ function is the weight function, defined as $w(\epsilon_t) = \psi(\epsilon_t)/\epsilon_t$ which represent the weight given to each residual by the estimator during the estimation process.
Moreover, depending on the characteristics of the $\psi$ function M-estimators can be classified in to re-descending or monotone M-estimators.
In the context of structural time series models Marczak, Proietti and Grossi used the weight function derived from an M-estimator to down weight spurious observations.
The main idea behind this procedure is as follows. Let $\mathbf{v}_t$ denote the one step ahead prediction error observed from the standard Kalman filter, and let $\mathbf{u}_t = \mathbf{F}_t^{-1/2}\mathbf{v}_t$ be the orthogonalized innovation. Furthermore, denote with $\boldsymbol{\psi}(\mathbf{u}_t)$ the $\psi$ function applied element wise to the vector $\mathbf{u}_t$ and let $\boldsymbol{\Delta}_t$ be the diagonal matrix that contains the weight of the elements of the vector $\mathbf{u}_t$.
Define $\tilde{\mathbf{u}}_t = \boldsymbol{\Delta}_t\mathbf{u}_t$ and assume that $\tilde{\mathbf{u}}_t$ $N(\mathbf{0}, \sigma^2\mathbf{I}_N)$. Under this assumption $\mathbf{u}_t \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Delta}_t^{-2})$ and as a final result the cleaned innovations, $\tilde{\mathbf{v}}_t = \mathbf{F}_t^{1/2}\tilde{\mathbf{u}}_t$ are $iid$ $N(\mathbf{0}, \mathbf{F}_t)$, whereas the observed contaminations, after writing $\mathbf{v}_t = \mathbf{F}_t^{1/2}\boldsymbol{\Delta}_t^{-1}\mathbf{F}_t^{-1/2}\tilde{\mathbf{v}}_t$ have a normal distribution with mean zero and covariance matrix $= \overline{\mathbf{F}}_t = \mathbf{F}_t^{1/2}\boldsymbol{\Delta}_t^{-2}\mathbf{F}_t^{1/2}$.
The robustification step is hence achieved by replacing $\mathbf{F}_t^{-1}$ with $\overline{\mathbf{F}}_t$ in the updating equations, in this way, the underlying innovation is distilled from the contaminated observable innovation $\mathbf{u}_t$.

Of course in the case of univariate time series models the quantities reduce to scalars.

# Chapter 7

# The Forward Search

As previously discussed in Chapter 5, atypical units can be treated in two different ways; using proper diagnostics based on traditional estimation techniques or, alternatively, with robust procedures.
The main issue, in the diagnostic approach, is its failure when data are contaminated by groups or patches of outliers, leading to masking and/or swamping effects. On the other hand, the properties of the robust estimators, that couple with the issues of swamping and masking, do not match those of maximum likelihood estimators.
The Forward Search ($FS$) is an iterative procedure that seeks to merge highly efficient estimators with diagnostics which are not affected by the masking and swamping effects. The idea behind the Forward Search was first introduced by Hadi (1992) for the detection of outliers in a linear regression fit using, approximately, half of the observations, and different versions of the procedure are described Hadi and Simonoff (1993). Subsequently, Atkinson, Riani and Cerioli extended this techniques. A modern treatment of the forward search and its extensions can be found in Atkinson and Riani (2004), Atkinson et al. (2010) and Atkinson and Riani (2012).
Our purpose is to extend the $FS$ procedure to to time series data, in particular, to the class of $SARIMA$ models.
In this chapter we will first present the main idea behind the Forward Search, next we will review its application to the linear regression model and, finally, we will extend the results to the class of $SARIMA$ models.

## 7.1 The Forward Search method

The Forward Search is a general method that can be used to detect outliers in structured data or to monitor the impact that single observations have on the fitting process of a model.
Given a set of $n$ observations $y_1, y_2, \ldots, y_n$, and a true generating model $F_\theta^\star$ assumed to belong to a parametric family $\mathcal{G} = \{F_\theta : \theta \in \Theta\}$, the method starts from a subset $\mathcal{S}^{(0)}$ of dimension $|\mathcal{S}^{(0)}| = m^\dagger << n$, that is robustly chosen to contain observations from a postulated model $F_\theta$. This subset is next used to obtain a first fit to the model. Here $|\cdot|$ denotes the cardinality of the given set, i.e the number of elements inside the specific set. Next, a prescribed measure of divergence, $d_i^{FS}(\cdot)$, between the observations and the postulated model $F_\theta$ is computed for each observation. The next subset $\mathcal{S}^{(1)}$, is obtained by taking the $m^\dagger + 1$ observations that resulted with the smallest deviation measure. The steps are repeater until the postulated model is fitted to all the units.
To summarize, the steps performed by the $FS$ are three:

1. Identification of the initial subset.
2. Adding observations.
3. Monitoring the Search.

In order to introduce the Forward Search we will first examine the general *iid* case and in particular its application to the linear regression model.

### 7.1.1 Forward Search - the linear regression model

Consider a sample $y_i, i = 1, \ldots, n$ of *iid* observations. The observations are assumed to be generated according the following regression model $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i$. In this case $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$ and $\epsilon_i \sim iidN(0, \sigma^2)$.
Under the normality and independence assumptions the maximum likelihood estimator of $\boldsymbol{\beta}$ corresponds to the $OLS$ estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

As a natural measure of deviation we consider $e_i^2 = (y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2$, the squared residuals.

#### Step 1: identification of the initial subset

If the length of the vector $\boldsymbol{\beta}$ is equal to $k$, the the $FS$ starts by selecting a subset of $m^\dagger = k$ units.

In order for the Forward Search to proceed in the right direction the starting subset has to be outlier free, therefore an highly robust estimator has to be used in the initial step. To initialize the $FS$ the least median of squares ($LMS$),Rousseeuw (1984) or the least trimmed squares ($LTS$), Leroy and Rousseeuw (1987), are commonly employed.
The $LMS$ is estimator is defined as

$$\hat{\boldsymbol{\beta}}^{LMS} = \arg\min_{\boldsymbol{\beta}} Me(e_i^2(\hat{\boldsymbol{\beta}}))$$

while the $LTS$ estimator is

$$\hat{\boldsymbol{\beta}}^{LTS} = \arg\min_{\beta} \sum_{i=1}^{h} e_{(i)}^2(\hat{\boldsymbol{\beta}}).$$

where $e_{(i)}^2$ denotes the order statistics of the squared regression residuals and Me denotes the median.
The $LTS$ estimator might be preferred since it is more efficient than the $LMS$, however, empirical evidence confirms that as long as the initial subset does not contain any outliers, the $FS$ will not be affected by the estimator used in the initialization.
If the number of observations $n$ is small, the identification of the initial subset can be brought out by complete enumeration of all the possible subsets of size $k$. That is, by enumerating all of the $\binom{n}{k}$ distinct $k$tuples. If $\binom{n}{k}$ is too large some large number of subset is considered instead, for example 1000.
Let $\mathcal{M}_k^n$ be the set of all the $k$tuples obtained from the $n$ observations. The initial subsets $\mathcal{S}^{(0)}$ is then defined as the set of $k$ observations that minimize the $LMS$ or $LTS$ criterion. Therefore, when the $LMS$ criterion is adopted the initial subset is composed by the units

$$\mathcal{S}^{(0)} = \{i : Me(e_i^2(\hat{\boldsymbol{\beta}})) = min\}$$

while, when the $LTS$ criterion is used the resulting initial subset is given by

$$\mathcal{S}^{(0)} = \{i : \sum_{i=1}^{h} e_{(i)}^2(\hat{\boldsymbol{\beta}}) = min\}.$$

where $|\mathcal{S}^{(0)}| = k = m^{\dagger}$.

**Step 2: adding observations during the search**

Denote with $\mathcal{S}^{(1)}$ the subset generated at the first step of the $FS$. The set $\mathcal{S}^{(1)}$ contains $m^{\dagger} + 2$ observations.
The Forward Search moves to the new subset $\mathcal{S}^{(2)}$ of dimension $m^{\dagger} + 2$ by selecting the

$m^\dagger + 2$ units that have the smallest squared residual.
Formally, define with $\mathcal{S}^{(m)}$ the set generated at the $m$th step of the search, with $e_i^2(\hat{\boldsymbol{\beta}}^{(m)}) = (y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}^{(m)})^2,\ i = 1, \ldots, n$, the squared residuals obtained by estimating the parameters with the units included in the subset $\mathcal{S}^{(m)}$ and with $e_{(i)}^2(\hat{\boldsymbol{\beta}}^{(m)})$ the corresponding order statistics.
The new subset $\mathcal{S}^{(m+1)}$ is then given by

$$\mathcal{S}^{(m+1)} = \{i : e_i^2(\hat{\boldsymbol{\beta}}^{(m)}) \leq e_{(m+1)}^2(\hat{\boldsymbol{\beta}}^{(m)})\}.$$

From the new subset we can obtain a new set of estimates $\hat{\boldsymbol{\theta}}^{(m+1)} = \{\hat{\boldsymbol{\beta}}^{(m+1)}, (\hat{\sigma}^{(m+1)})^2\}$ estimates, that is

$$\hat{\boldsymbol{\beta}}^{(m+1)} = (\sum_{i \in \mathcal{S}^{(m+1)}} \boldsymbol{x}_i'\boldsymbol{x}_i)^{-1}(\sum_{i \in \mathcal{S}^{(m+1)}} \boldsymbol{x}_i' y_i) \qquad (\hat{\sigma}^{(m+1)})^2 = \frac{1}{m}\sum_{i \in \mathcal{S}^{(m+1)}} e_i^2(\hat{\boldsymbol{\beta}}^{(m+1)})$$

The Forward Search estimator $\hat{\boldsymbol{\theta}}^{FS}$ is then defined as the collection of estimators generated at each step of the search, that is

$$\hat{\boldsymbol{\theta}}^{FS} = \{\hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{\theta}}^{(m+1)}, \ldots, \hat{\boldsymbol{\theta}}^{(n-m^\dagger)}\}.$$

The $FS$, avoiding the inclusion of outliers in the initial step, seeks to provides a natural ordering of the data in accordance to the assumed model. In the Forward Search method a robust estimator is used only to initialize the algorithm, while in all the other steps an highly efficient estimator is used.
In this context, the fact that the $ML$ estimator has breakdown point equal to zero does not represent a problem, since the inclusion of an outlier is indicated by a marked change in the monitored quantities. The $ML$ estimator obtained by the $FS$ at a given step corresponds to

$$\hat{\boldsymbol{\theta}}^{(m)} = \arg\min_{\theta \in \Theta}\ \ell^{(m)}(\theta)$$

where $\ell^{(m)}(\theta)$ is the log-likelihood function obtained from the units included in the $m$th set, that is $\ell^{(m)}(\theta) = \sum_{i \in \mathcal{S}^{(m)}} \ln f(y_i; \theta)$.
In the Gaussian case considered here, the model likelihood obtained in the $m$th step of the search is given by

$$\mathcal{L}(\boldsymbol{\theta})^{(m)} = \prod_{i \in \mathcal{S}^{(m)}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \left[\frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right].$$

Making use of the indicator functions the likelihood cam be rewritten as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta})^{(m)} &= \prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp-\left[\frac{(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right]\right\}^{\mathbb{1}(i\in\mathcal{S}^{(m)})} \\ &= \prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp-\left[\frac{(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right]\right\}^{\mathbb{1}[e_i^2(\hat{\boldsymbol{\beta}}^{(m-1)})\leq e^2_{(m+1)}(\hat{\boldsymbol{\beta}}^{(m-1)})]},\end{aligned}$$

and by taking the logarithm the log-likelihood results equal to

$$\ell(\boldsymbol{\theta})^{(m)} = \sum_{i=1}^{n}\left(-\frac{1}{2}\ln(2\pi)-\frac{1}{2}\ln(\sigma^2)-\frac{(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right)\mathbb{1}[e_i^2(\hat{\boldsymbol{\beta}}^{(m-1)})\leq e^2_{(m+1)}(\hat{\boldsymbol{\beta}}^{(m-1)})]$$

where the contribution of a single unit to the log-likelihood is given by

$$\ell(\boldsymbol{\theta})_i^{(m)} = -\frac{1}{2}\ln(2\pi)-\frac{1}{2}\ln(\sigma^2)-\frac{(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}.$$

Therefore, the likelihood that is maximized at each step of the Forward Search can be considered as a weighted likelihood, with weights 1 if the unit is included in the $m$th set and 0 otherwise. The weight function at a given step $m$ is given by

$$w_i^{(m)} = \begin{cases}1 & \text{if} \quad e_i^2(\hat{\boldsymbol{\beta}}^{(m-1)})\leq e^2_{(m+1)}(\hat{\boldsymbol{\beta}}^{(m-1)}) \\ 0 & \text{else}\end{cases}$$

As a direct consequence the $OLS$ estimator in the regression case discussed until now can be generalized as a weighted least squares estimator, where the hard trimming weights at each step are assigned by $w(i)^{(m)}$. The estimators can be written as

$$\hat{\boldsymbol{\beta}}^{(m+1)} = (\sum_{i=1}^{n}\boldsymbol{x}_i'\boldsymbol{x}_i w_i^{(m)})^{-1}(\sum_{i=1}^{n}\boldsymbol{x}_i' y_i w_i^{(m)})$$

where now the sum is taken over all the units in the sample.
It is important to note that the sequence of subsets generated by the $FS$ are not in general monotone, that is

$$\mathcal{S}^{(m)} \not\subset \mathcal{S}^{(m+1)}$$

does not generally hold as a unit which is included in the $m$th step can be discarded in

the $m+1$th step and rejoin the search in the further steps.

### 7.1.2 Step 3: monitoring the Forward Search

Step 2 is repeated until all the units forming the dataset are included into the search.
The monitoring of at which step of the Forward Search a unit is included provides an ordering of the data with respect to the assumed null model, with observations furthest from it joining the search in the last steps.
As units are included into the search, different quantities are monitored by means of different plots. The first plot monitors all the squared standardized residuals $(s_i^{(m)})^2 = e_i^2(\hat{\boldsymbol{\beta}}^{(m)})/(\hat{\sigma}^{(n-k)})^2,\ i = 1, \ldots, n.$ obtained at each step of the search. Since $(\hat{\sigma}^{(m)})^2$ is highly affected by $m$ all the residuals are standardized by the final estimate $(\hat{\sigma}^{(n-k)})^2$. Large values of the residuals among cases not in the subset indicate the presence of possible outliers.
Two further plots of great importance are those that monitor the maximum standardized residual in the subset, that is

$$e^2_{(m)}(\hat{\boldsymbol{\beta}}^{(m)})/(\hat{\sigma}^{(n-k)})^2 \tag{7.1}$$

and the plot that monitors the smallest standardized residual of the units non included in the set

$$e^2_{(m+1)}(\boldsymbol{\beta}^{(m)})/(\hat{\sigma}^{(n-k)})^2 \tag{7.2}$$

A sharp peak in the quantity (7.1) reveals that an outlier has been included in the set $\mathcal{S}^{(m+1)}$, while a sharp increase in the quantity (7.2) reveals that an outlier will be included in the set $\mathcal{S}^{(m+2)}$.
Further quantities of interests such as estimators and test statistics can be monitored as well. The interest here might be in evaluating the impact of one or more outliers on the monitored quantities.

## 7.2 Extending the Forward Search to $SARIMA$ models

The extension of the Forward Search to time series models, that we will denote $FSTS$ can be obtained following different strategies. These strategies differ in the method of estimation and in the way the subsets of the $FS$ are formed. The idea of extending the $FS$ to linear time series models derive from the similarity between the traditional regression model and the autoregressive model. In particular, when the error terms are assumed to be independent, identically distributed with a Gaussian distribution, the the maximum likelihood estimator, becomes the OLS estimator. However, in order to apply the $FSTS$ method to the complete class of $SARIMA$ models that encompass $AR$ and $MA$ models as particular cases, a likelihood approach has to be followed. Doing this, the use of the Kalman filter to compute the prediction error decomposition of the likelihood plays a key role. The advantages gained from a likelihood approach are threefold. First, the Kalman filter provides a general method that can be applied to the entire class of linear time series models (potentially to every linear model that admits a state space representation). As a consequence the same routine can be used for the entire class of linear time series models ($SARIMA$). Second, when the size of the sample is small, the estimates obtained by maximizing the exact likelihood have a smaller bias compared to the conditional maximum likelihood estimates. Third, the Kalman filter provides a natural way to handle the missing observations generated by the $FSTS$ in the time series case.

### 7.2.1 Initialization of the $FSTS$

Differently from the *iid* case where the initial subset is chosen among all the possible subsets of dimension $k$, where $k$ denotes the number of parameters of interests, in the $FSTS$ the a robust subset is chosen among $q$ blocks of contiguous observations of fixed dimension $m^{\dagger}$. In the most general case the blocks can be overlapping. The idea of window, or block sampling, Heagerty and Lumley (2000), is to retain the same dependence structure as in the original data set.
In order to find the initial robust subset Riani (2004) and Grossi (2004) suggested to perform an exhaustive search over all the possible contiguous blocks of dimension $k$ and select the one that satisfies an $LMS$ or $LTS$ criterion.
Given a time series of length $T$ let $\mathcal{S}^{(m^{\dagger})}$ be a set formed by $m^{\dagger}$ contiguous units and let $\boldsymbol{\theta}$ be the vector of unknown parameters. The starting set corresponding the $LMS$ or $LTS$ criterion are then given by

$$\mathcal{S}^{(0)}_{LMS} = \{S^{(k)} : Me(e_t^2(\hat{\theta}_{\mathcal{S}^{(k)}})) = min\},$$

and

$$\mathcal{S}^{(0)}_{LTS} = \{S^{(k)} : \sum_{t=1}^{T} e^2_{(t)}(\hat{\theta}_{\mathcal{S}^{(k)}}) = min\}.$$

As the number of subsets increase the probability of having at least an initial subset which does not contains outliers also increase. On the other hand, if the initial subset is too small the estimator is inconsistent. As a rule of thumb, to find a balance between the robustness and the other statistical properties of the estimator we suggest a value $m^\dagger \approx \sqrt{T}$.

### 7.2.2 The Kalman filter in the $FSTS$

Denote by $\mathcal{S}^{(m)}$ the subset obtained at the $m$th step of the search and by $\boldsymbol{y}^{(-m)}$ the observed time series where the units not included in the $m$th step have been settled as missing. Following Riani (2004), ince $|\mathcal{S}^{(0)}| = m^\dagger$ the number of units settled at missing at the $m$th step is equal to $m + m^\dagger$. Applying the Kalman filter we obtain the prediction error decomposition of the log-likelihood, which is subsequently maximized with respect to the hyperparameters.
For a generic step of the $FSTS$ the Kalman filter runs as follow:

1. Initialize $\boldsymbol{m}_0 = 0$ and $\boldsymbol{\Omega} = (\mathbf{I} - \mathbf{T} \otimes \mathbf{T})^{-1}\, vec(\mathbf{Q})$
2. Check if $\mathbb{1}(t \in \mathcal{S}^{(m)}) = 1$, then
   (a) $\mathbf{v}_t = \mathbf{y}_t - \mathbf{Z}_t\mathbf{m}_t$
   (b) $\mathbf{F}_t = \mathbf{Z}_t\mathbf{P}_t\mathbf{Z}_t' + \mathbf{H}_t$
   (c) $\mathbf{m}_{t|t} = \mathbf{m}_t + \mathbf{P}_t\mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{v}_t$
   (d) $\mathbf{P}_{t|t} = \mathbf{P}_t - \mathbf{P}_t\mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{Z}_t\mathbf{P}_t$
   (e) $\mathbf{m}_{t+1} = \mathbf{T}_t\mathbf{m}_{t|t} + \mathbf{K}_t\mathbf{v}_t$
   (f) $\mathbf{P}_{t+1} = \mathbf{T}_t\mathbf{P}_{t|t}(\mathbf{T}_t - \mathbf{K}_t\mathbf{Z}_t)' + \mathbf{R}_t\mathbf{Q}_t\mathbf{R}_t'$
3. else if $\mathbb{1}(t \in \mathcal{S}^{(m)}) = 0$
   (a) $\mathbf{v}_\tau = \mathbf{y}_\tau$
   (b) $\mathbf{F}_{\tau|t-1} = \mathbf{H}_t$
   (c) $\mathbf{m}_{\tau|t} = \mathbf{m}_{\tau|t-1}$
   (d) $\mathbf{P}_{\tau|t} = \mathbf{P}_{\tau|t-1}$
   (e) $\mathbf{m}_{\tau+1|t} = \mathbf{T}_t\mathbf{m}_{\tau|t}$

(f) $\mathbf{P}_{\tau+1|t} = \mathbf{T}_t \mathbf{P}_{\tau|t} \mathbf{T}_t' + \mathbf{R}_t \mathbf{Q}_t \mathbf{R}_t'$

As a result the exact log-likelihood obtained on the subset $\mathcal{S}^{(m)}$ is

$$\begin{aligned}\ell(\theta)^{(m)} &= -0.5\ m\ \ln\sigma^2 + \sum_{t \in S^{(m)}} \frac{v_t^2}{f} \\ &= \left( -0.5 \sum_{t=1}^{T} \ln\sigma^2 + \sum_{t=1}^{T} \frac{v_t^2}{f} \right) \mathbb{1}(t \in \mathcal{S}^{(m)})\end{aligned}$$

where $v_t^2$ are the one step ahead prediction errors for the units composing the subset $\mathcal{S}^{(m)}$.

### 7.2.3 Progressing in the $FSTS$ search

Once the parameter $\widehat{\theta}^{(m)}$ at a given step are estimated, these are used to run the Kalman filter on the entire time series $\boldsymbol{y}$ in order to obtain the residuals. The FSTS moves to dimension $m+1$ by selecting the $m+1$ observations with the smallest squared (or absolute) residual. The observations are chosen by ordering all the squared residuals $e_t^2(\widehat{\boldsymbol{\theta}}^{(m)}), t = 1, \ldots, T$ and selecting the $m+1$ smallest ones.
Formally the new set is defined as

$$\mathcal{S}^{(m+1)} = \{t : e_t^2(\widehat{\boldsymbol{\theta}}^{(m)}) \leq e_{(m+1)}^2(\widehat{\boldsymbol{\theta}}^{(m)})\}.$$

The recursion are performed until all the units are included in the search.

### 7.2.4 Monitoring the $FSTS$

As in the *iid* case a series of plots are produced by the search. The most important plot is certainly the plot that monitors the standardized residuals of all the units in the series. However, since in the first steps of the search many observations are settled as missing, the resulting $ML$ estimates may results unstable, leading to wild trajectories of the residuals in the initial steps. As a solution, we simply suggest to discard the residuals obtained in the first steps of the search.
Differently from the standard linear regression model in the $FSTS$, the monitoring of the minimum standardized residuals for the units excluded, and the maximum for the units included, does not seem useful to identify possible outliers. On the contrary, we found that the monitoring of the diagnostics introduced in Chapter 5, might result particularly informative to discover the type of outlier of a particular unit. These statistics, as argued in the previous chapter might suffer from swamping and masking effects, nevertheless, their monitoring in the $FSTS$ context may reveal important patterns in the data.

### 7.2.5 An automatic outlier detection procedure

The main drawback of the $FSTS$ method is that it does not provide an automatic procedure to identify the possible outlying or influential units. Therefore, we suggest a possible method that can be employed to extract in an automatic manner the units more likely to be flagged as candidate outliers

In order to resume the information for a given unit obtained by running the search we define the following statistics $M^{\lambda}_{t,i} = Me_m(\hat{\lambda}^{(1)}_{t,i}, \hat{\lambda}^{(2)}_{t,i}, \ldots, \hat{\lambda}^{(M)}_{t,i})$,$i = \{A, I, LS\}$, where $Me_m(\cdot)$ denotes the median value of the statistics $, \hat{\lambda}^{(M)}_{1,i}$ taken over the $FSTS$ path. That is to say, once the $FSTS$ is terminated the selected diagnostic quantity for a given unit obtained trough the search, are ordinated and the median is taken. As a rule of thumb, we identify as possible outliers the units that have a value of the median statistic, $M^{\lambda}_{t,i}$, above a certain threshold. A natural threshold should be a quantile form the distribution of the given statistic. Nevertheless, the distribution of the statistic is not known and might depend on the type of models and on the true values of the parameters. Therefore, as a general procedure we select as a threshold the empirical quantile. Formally, let $F_{n,i}(M^{\lambda}_{t,i})$ be the cumulative distribution function for the given statistic, and let $\hat{q}_{\alpha,i} = F^{-1}_{n,i}(\alpha)$ the empirical quantile, then we can define the indicator function

$$\mathbb{1}(M^{\lambda}_{t,i} \leq \hat{q}_{\alpha,i}) = \begin{cases} 1 & \text{the } t\text{th unit is a possible outlier,} \\ 0 & \text{the } t\text{th unit is not an outlier,} \end{cases}$$

where $i = \{A, I, LS\}$.

Once the most atypical units have been identified, we tag the unit with the most likely type of outlier that resulted by means of the following statistics

$$\max_{i \in \{A,I,LS\}} |M^{\lambda}_{\tau,i}|,$$

where $t = \tau$ if $\mathbb{1}(M^{\lambda}_{t,i} \leq \hat{q}_{\alpha,i}) = 1$.

Moreover, in order to take into account that on average the diagnostic for detecting a particular kind of outlier might result larger than the others we correct $\max_{i \in \{A,I,LS\}} |M^{\lambda}_{\tau,i}|$ by means of a weight, that is proportional to the distance between $M^{\lambda}_{\tau,i}$ and the corresponding empirical quantile $\hat{q}_{\alpha,i}$. Formally, we define a weight function $w(z)$ that maps a proper distance function in the range $\{0, 1\}$. Without loss of generality we take as a proper distance function the absolute deviation, and to map the distance in the $\{0, 1\}$ range we define

$$w(z) = \frac{z_t - \min(z_t)}{\max(z_t) - \min(z_t)}.$$

The resulting weights are therefore given by

$$w_{\tau,i} = \frac{|M^{\lambda}_{\tau,i} - \hat{q}_{\alpha,i}| - \min(|M^{\lambda}_{\tau,i} - \hat{q}_{\alpha,i}|)}{\max(|M^{\lambda}_{\tau,i} - \hat{q}_{\alpha,i}|) - \min(|M^{\lambda}_{\tau,i} - \hat{q}_{\alpha,i}|)},$$

and the decision to classify an identified unit to the $i$th class of outliers is finally taken using

$$\max_{i \in \{A,I,LS\}} (w_{\tau,i}|M^{\lambda}_{\tau,i}|).$$

In the next section we will present two example of the $FSTS$ procedure applied on real data.

## 7.3 Real data applications

### 7.3.1 The Airline Passenger Data

The airline passenger data, given in Box and Jenkins (1976), has often been used in the time series analysis literature as an example of a non stationary seasonal time series. The series consists of 144 monthly total numbers of airline passengers from January 1949 to December 1960 and it has been used by Box and Jenkins to develop a two-coefficients time series model of factored form that is now known as the airline model. In this section we will analyze these data using the multiplicative seasonal model initially specified by Box and Jenkins employed in the *FSTS*.

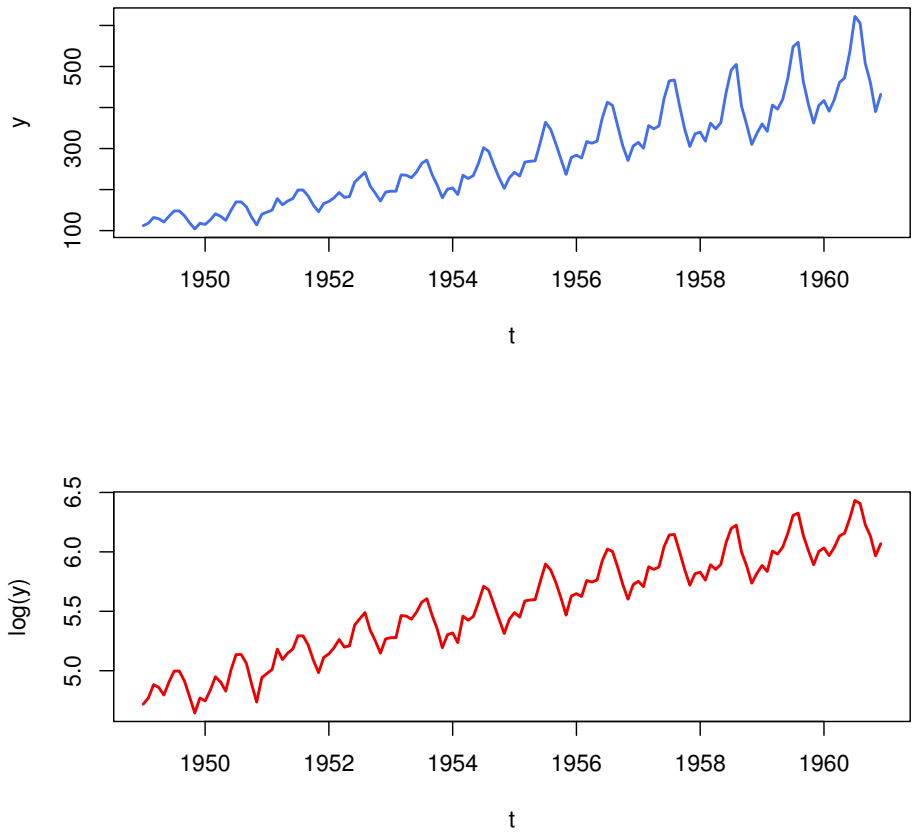


Figure 7.1: The upper panel shows the airline passenger data. As it is evident, the data show a strong seasonal behavior with period equal to 12 and non stationarity in variance, as the variance of the series increase with time. The lower panel shows the log transformed series. As it is evident the transformation works well and the variance is result stable over the period considered.

Since the series exhibits non stationarity in variance, a common step taken in the analysis of this series is to log-transform the data.
Figure 7.1 shows the series before and after the log transformation.

In order to take into account the upward trend and the seasonal component Box and Jenkins specified the following seasonal multiplicative model

$$(1-B)(1-B^{12})Y_t = (1-\theta B)(1-\Theta B^{12})\epsilon_t, \tag{7.3}$$

where $(1-B)$ takes into account the stochastic up going trend while $(1-B^{12})$ captures the stochastic monthly seasonality.
Under the assumption that model (7.3) is correctly specified, Chen and Liu (1993a) by applying the procedure of Tsay (1988), identified three outliers at time 29 (AO), 54 (LS) and 62 (AO).

### Forward Search Analysis of the Airline Passenger data

We perform the Forward Search on the logged airline passenger series. In order to initialize the $FSTS$ the initial robust subset is composed by 15 units.
The first important plot, in figure 7.2 shows the forward trajectories of the standardized residuals.
The dotted lines corresponds to the 0.95 and 0.99 quantiles of a standard normal distribution, and, as a rule of thumb, residuals that exhibit a trajectory persistently below or above the standard quantiles indicates that the corresponding unit may represent a potential outlier.

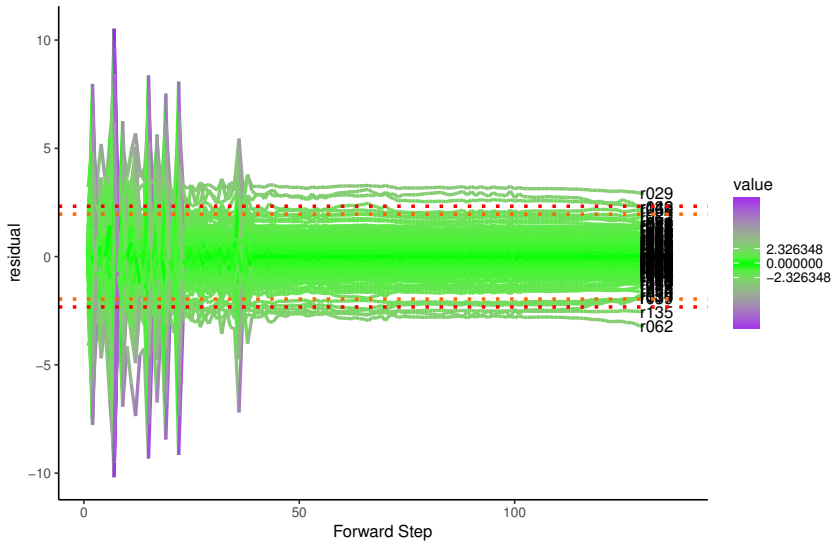


Figure 7.2: Forward trajectory plot of the standardized residuals. The dotted lines represents the 0.99 and 0.95 quantiles of the standard normal distribution. As it is evident a group of three units clearly stem out from the bulk of data. These are unit 29, 62 and 135.

A critical inspection of the plot reveals that the majority of data remains constantly between the threshold values. Since the trajectories are particularly wild, especially at the

beginning of the search, it is useful to focus the attention to the last 50% of the steps. When focusing on the last half of the search, it is possible to note that the trajectories of the units 29, 62 and 135, remain markedly outside the threshold bands for the entire path of the search. Whereas, the trajectories of the units 38, 42 and 17 decrease slightly at the end of the search. Of particular interests, are the trajectories associated to the group of three units that around step 128 decrease abruptly.
This may be the indicator of some masking effect introduced by one or several units that joined the set.

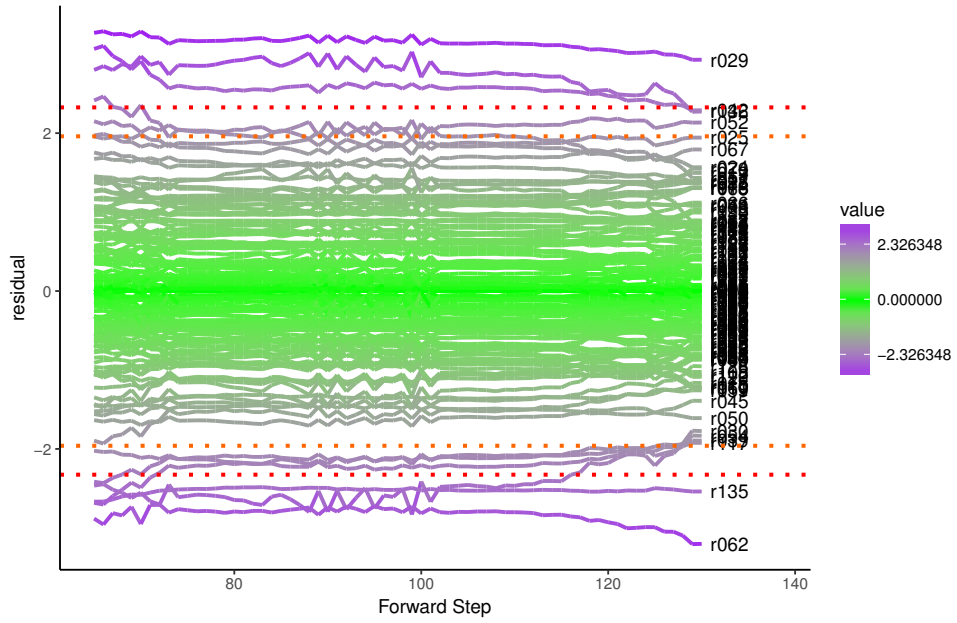


Figure 7.3: The plot shows the forward trajectories of the standardized residuals in the last 50% of the $FSTS$. As it is more clear from this plot the trajectories of the units 135, 29 and 62 remains for the entire search outside the confidence bands. Moreover the trajectories of units 38, 42 and 17 decrease slightly in the last steps of the search.

Further information can be obtained by the forward plot of the statistics employed to test whether a unit is an outlier or not.
Figures 7.4, 7.5 and 7.6 shows the forward trajectories of the outlying statistics.

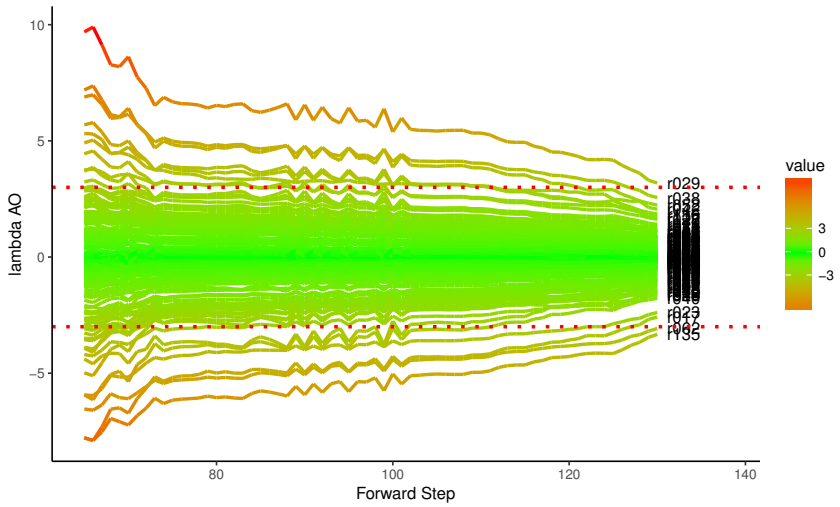


Figure 7.4: Trajectory plot of the $\hat{\lambda}_A$ statistics

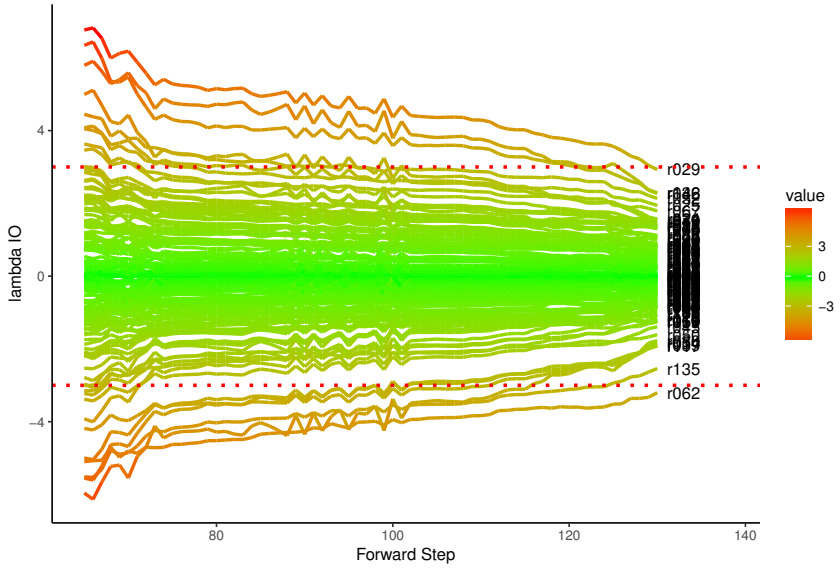


Figure 7.5: Trajectory plot of the $\hat{\lambda}_I$ statistics

The first forward plot monitors the value of the statistic used to test whether a unit is an additive outlier. Its inspection reveals that despite $\hat{\lambda}_{t,A}^{(m)}$ remains constantly above or below the threshold for the units 62, 135 and 29, there is a clear group of units, i.e 23, 17, 135 and 38 whose trajectories are not consistent with the bulk of data.

A similar conclusion is drawn from the forward plot of $\hat{\lambda}_{t,I}^{(m)}$. In general, we observe that when $\hat{\lambda}_{t,A}^{(m)}$ has a large value we will observe a large value also in $\hat{\lambda}_{t,I}^{(m)}$. The decision between additive or innovation outlier should be done depending on the absolute magnitude of the corresponding statistics.

An interesting pattern emerges when we focus on the forward plot of $\hat{\lambda}_{t,LS}^{(m)}$, shown in Figure 7.6. This plot confirms the presence of a possible level shift at time 54, the same conclusion drawn by Chen and Liu. However, other times such as 30, 39 and 24 shows trajectories distant from the majority of data that indicates the possible presence of other

level shifts.

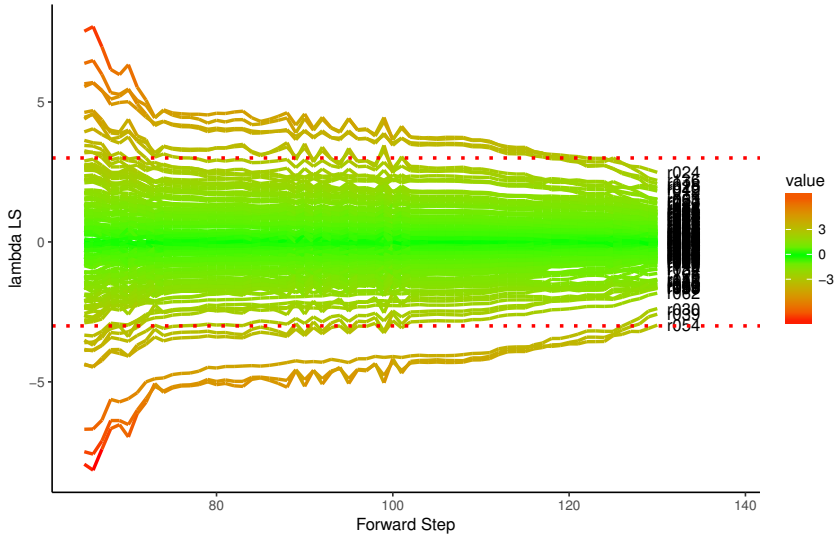


Figure 7.6: Trajectory plot of the $\hat{\lambda}_{LS}$ statistics

In order to summarize the different information provided in the different plots we suggest to compute for each $t$, the median over the different steps of the forward search.
As an example for the outlying statistics the resuming quantity that we propose is

$$M^{\lambda}_{t,i} = Me_m(\hat{\lambda}^{(1)}_{t,i}, \hat{\lambda}^{(2)}_{t,i}, \ldots, \hat{\lambda}^{(M)}_{t,i})$$

where $M$ is the total number of steps performed by the forward search and $i = \{A, I, LS\}$. We propose the median because the trajectories might be particularly unstable at the beginning of the search, and the mean might suffer from bias.
The plots for $M^{\lambda}_{A}, M^{\lambda}_{I}$ and $M^{\lambda}_{LS}$ are shown in Figure 7.7
The leaf plots in Figure 7.7 reveals a certain structure in the series. When considering this plot it is worth focusing on the units that resulted atypical in the analysis of the forward residuals plot. Units 29, 62 and 135 are identified as AO, while unit 136 results identified as IO. Particularly interesting is the plot that show $M^{\lambda}_{t,LS}$. Times 30, 39, 42 and 54 results all potential level shifts, however by taking into account the natural ordering of the series the level shift might occur at $t = 30$ and the subsequent units might be wrongly identified as $LS$ because of the swamping effect that is introduced.

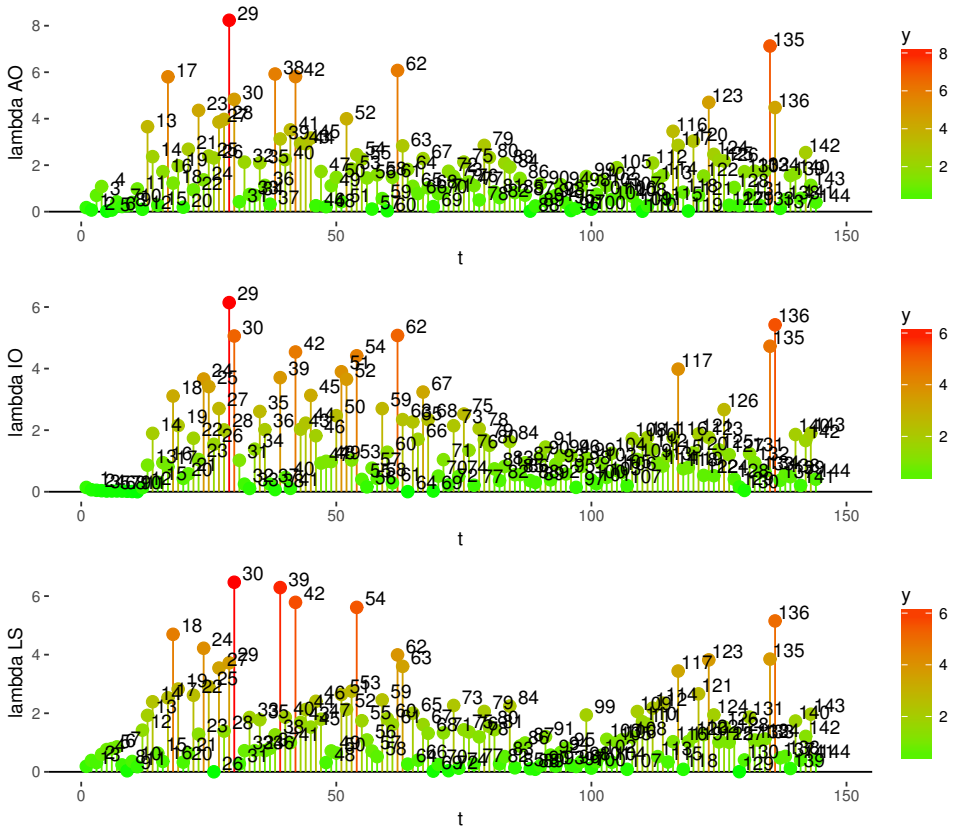


Figure 7.7: Stem plot of the $M^{\lambda}_{t,i}$ statistics, for $i = AO, IO, LS$. By preserving the natural ordering of the data and by making use of robust quantities the stem plot may reveal interesting pattern in the series. Here the stem plot confirms that units 29, 62 and 135 are influential observations. Moreover a series of other potential atypical units are identified, such as unit 38, 42 and 135. Of particular interest the stem plot in the third panel which identifies four times as possible level shifts.

Since the step at which, a given unit joins the Search, should be interpreted as a degree of outlyingness we show the last 5, 10 and 15 units that are added to the $FSTS$.
The plot confirms that units 29, 62, 135 and 136, already flagged as outliers, join the search in the last 5 steps. Surprisingly time 42 that did not seem particularly suspicious in the stem plot is one of the last units that is added to the Search. Additionally, when we move to consider the last 10 units that are added to the search, units 30, 39 and 54 are in this group. All of them were enlighten previously in the stem plot as possible level shifts.

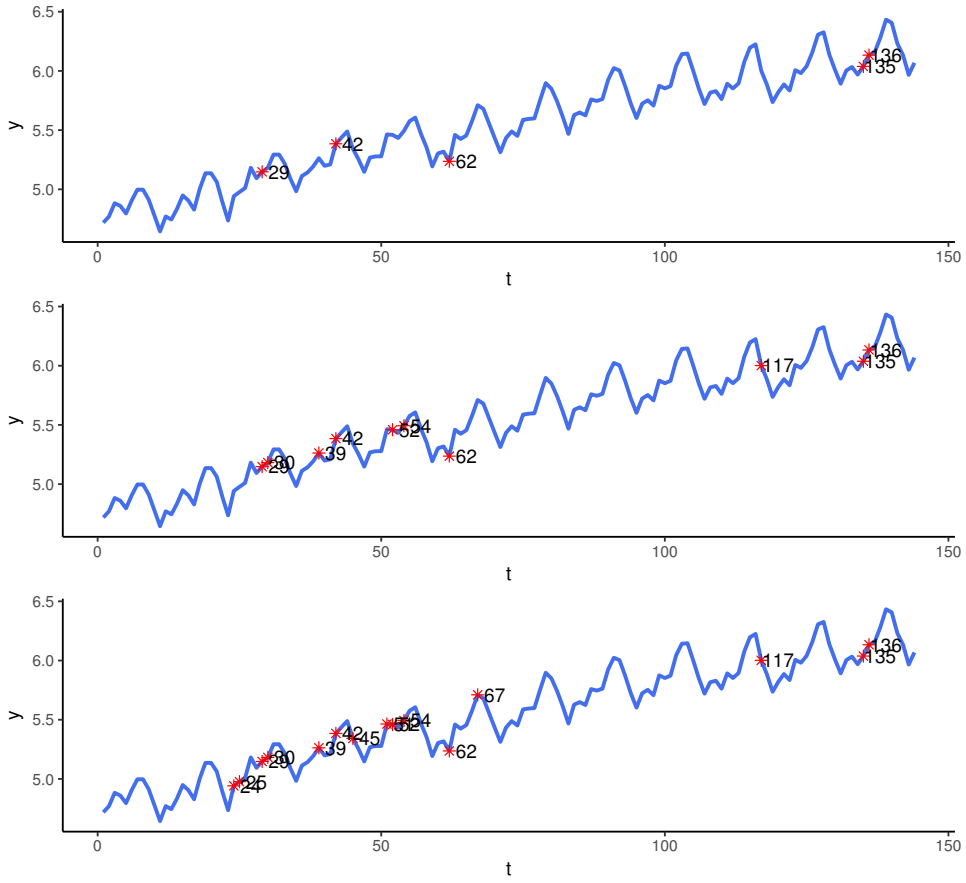


Figure 7.8: Last 5, 10 and 15 units that join the search

In order to conclude the analysis we employ the automatic outlier detection technique with three different threshold values. Formally we compare $|M^{\lambda}_{\tau,i}|$ with $\hat{q}_{0.95,i}, \hat{q}_{0.975,i}$ and $\hat{q}_{0.99,i}$.

The results are report in Table 7.1, 7.2 and 7.3.

At a 1% significance level the automatic procedure detects 5 units as outliers, including observations 29 and 135 that were flagged before as deviating units. When the statistician is willing to accept an higher degree of error and selects $\alpha = 0.05$, the automatic procedures enrich the previous set by adding 4 units. Among them observation 62 whose residual exhibits a wild trajectory for the entire $FSTS$. To conclude, overall the automatic method seems to flag as possible outliers all the units that were previously spotted from a visual inspection of different plots produced by the $FSTS$. It is important to remark that relying only in the automatic detection process might lead to inefficient results, moreover by doing so the true strength of the $FSTS$ that relies in the Forward monitoring plots it is not used.

| $\alpha = 0.01$ | | | | |
|---|---|---|---|---|
| $\tau$ | $\lvert w_{\tau,A} M^{\lambda}_{\tau,A} \rvert$ | $\lvert w_{\tau,I} M^{\lambda}_{\tau,I} \rvert$ | $\lvert w_{\tau,LS} M^{\lambda}_{\tau,LS} \rvert$ | $i$ |
| 29 | 1.071183 | **1.078213** | 1.269853 | LS |
| 30 | **1.292167** | 0.429567 | 0 | AO |
| 39 | **1.515913** | 1.026921 | 0.032233 | AO |
| 135 | 0 | 0.293607 | **1.246126** | LS |
| 136 | **1.298258** | 0 | 0.54172 | AO |

Table 7.1: Outliers detected using the weighted approach with $\alpha = 0.01$. Here $\tau$ denotes the times at which an outlier was detected while $i$ indicates the corresponding type of outlier. The significant weighted statistics are represented in bold. Surprisingly with a confidence level of 0.01 the unit 62 it is not detected by the automatic procedure.

| $\alpha = 0.025$ | | | | |
|---|---|---|---|---|
| $\tau$ | $\lvert w_{\tau,A} M^{\lambda}_{\tau,A} \rvert$ | $\lvert w_{\tau,I} M^{\lambda}_{\tau,I} \rvert$ | $\lvert w_{\tau,LS} M^{\lambda}_{\tau,LS} \rvert$ | $i$ |
| 17 | 0.018296 | 0.665099 | **1.221223** | LS |
| 29 | **3.139259** | 1.617112 | 0.991136 | AO |
| 30 | **0.932102** | 0.083338 | 0.819425 | AO |
| 39 | **1.338383** | 0.811188 | 0.859733 | AO |
| 42 | 0 | 0 | **0.079626** | LS |
| 54 | **1.327205** | 0.719166 | 0.135025 | AO |
| 62 | 0.483114 | 0.346995 | **0.822408** | LS |
| 135 | **1.722839** | 0.011623 | 0.954276 | AO |
| 136 | **0.940923** | 0.380352 | 0 | AO |

Table 7.2: Outliers detected with the weighted statistics at a confidence level equal to $\alpha = 0.025$

| $\alpha = 0.05$ | | | | |
|---|---|---|---|---|
| $\tau$ | $\lvert w_{\tau,A} M^{\lambda}_{\tau,A} \rvert$ | $\lvert w_{\tau,I} M^{\lambda}_{\tau,I} \rvert$ | $\lvert w_{\tau,LS} M^{\lambda}_{\tau,LS} \rvert$ | $i$ |
| 17 | **1.549255** | 0.630808 | 0.840064 | AO |
| 18 | 0.879295 | 0.585536 | **1.099377** | LS |
| 23 | 0 | 0.752408 | **0.840937** | LS |
| 24 | **0.927784** | 0.210627 | 0.425464 | AO |
| 29 | **6.090761** | 3.544277 | 0.249487 | AO |
| 30 | 0.079549 | 0.984486 | **3.365691** | LS |
| 38 | **0.947416** | 0.064671 | 0.824855 | AO |
| 39 | 0.918031 | 0.121577 | **3.429296** | LS |
| 42 | 1.365187 | 1.108632 | **2.17656** | LS |
| 54 | 1.062315 | 0 | **2.267336** | LS |
| 62 | **2.252213** | 1.720611 | 0 | AO |
| 117 | **1.022178** | 0.089893 | 0.395204 | AO |
| 123 | 0.107282 | **0.920041** | 0.149188 | IO |
| 135 | **4.070112** | 1.22342 | 0.177679 | AO |
| 136 | 0.094837 | **1.769571** | 1.355147 | IO |

Table 7.3: Outliers detected with the weighted statistics at a confidence level equal to $\alpha = 0.05$

### 7.3.2 Retail sales data

#### Preliminary analysis

This series consists of 150 monthly sales registered in an Supermarket shop in Parma, starting from January 2005 until June 2017. The series is reported in Figure 7.9.

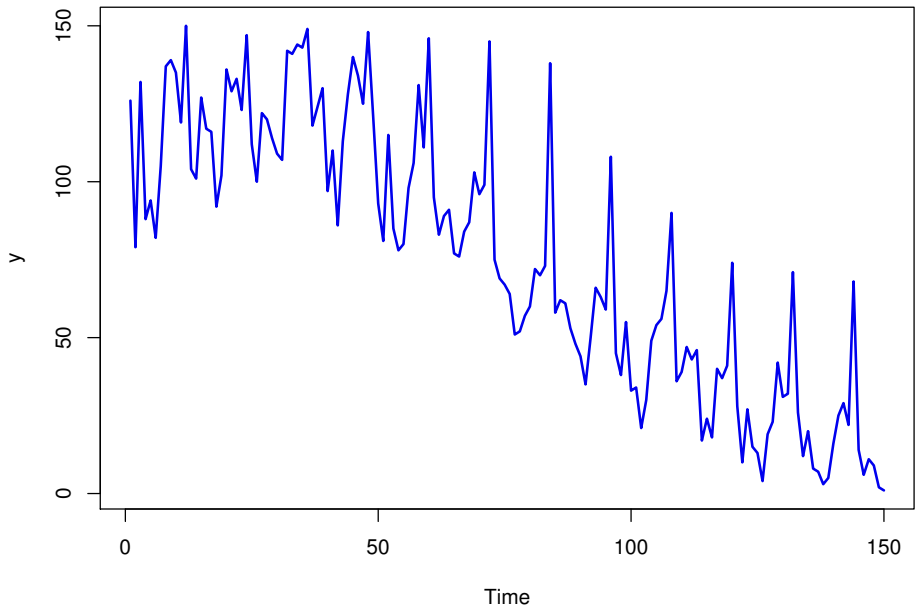


Figure 7.9: Monthly sales data of an Hypermarket shop in Parma. On the time axis 0 corresponds to the 1st of January 2005 while 150 corresponds to 1st of June 2017.

Graphical inspection reveals that the series exhibits a downward trend and a seasonal pattern with period equal to 12 months. Because of the irregularity of the seasonal pattern, especially in the first half of the series, we assume that the seasonality can be well captured by a stochastic component. While for the downward trend we take the first difference of the data.

### Model Specification

In order the identify the proper model and the respective order we proceed with the inspection of the sample autocorrelation and partial autocorrelation functions, Figure 7.10. Graphical inspection of the $SACF$ and $SPACF$ reveal a slow dampening behavior, with a regular pattern observed at every 12 lags. The slow dampening is generally uncounted in the presence of unit roots, while the 12 lags pattern confirms the seasonal component of the series.

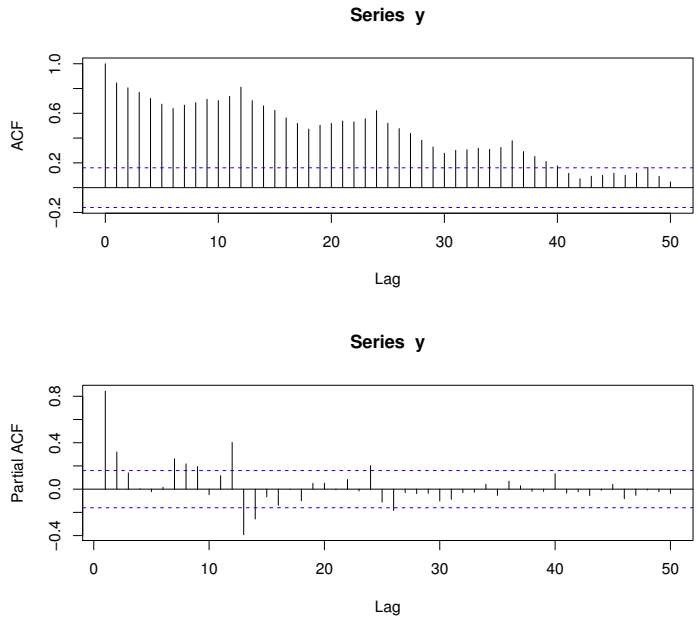


Figure 7.10: From the analysis of the estimated $ACF$ (upper panel) it is evident that the series exhibits a seasonal behavior with period equal to 12. Moreover the slow decaying may indicate the presence of possible unit roots

Furthermore we inspect the $SACF$ and $SPACF$ of $(1-B)y_t$ and of $(1-B)(1-B^{12})y_t$ respectively shown in Figures 7.11 and 7.12. These show three possible options. First, the exponential decay of the autocorrelation function at multiple lags of 12 and the cutting off of the $PAC$F after lag 13 suggests a possible seasonal autoregressive component, however the identification of non seasonal components is not clear from the $SACF$ and $SPACF$, and a first tentative model might be a $SARIMA(0,1,1)(1,0,0)_{12}$ which we will refer as $\mathcal{M}_1$.
The second strategy, could be to seasonally differentiate the data and specify a model on the transformed data. The $SACF$ an $SPACF$ for the seasonally differenced data are reported in Figure 7.12. Visual inspection suggests two candidate models: $SARIMA(0,1,1)(0,1,1)_{12}$ or $SARIMA(0,1,1)(1,1,0)_{12}$, that we will refer as $\mathcal{M}_2$ and $\mathcal{M}_3$.
Having decided a list of possible candidate models we can perform the $FSTS$ and consider the results obtained with the three specifications.

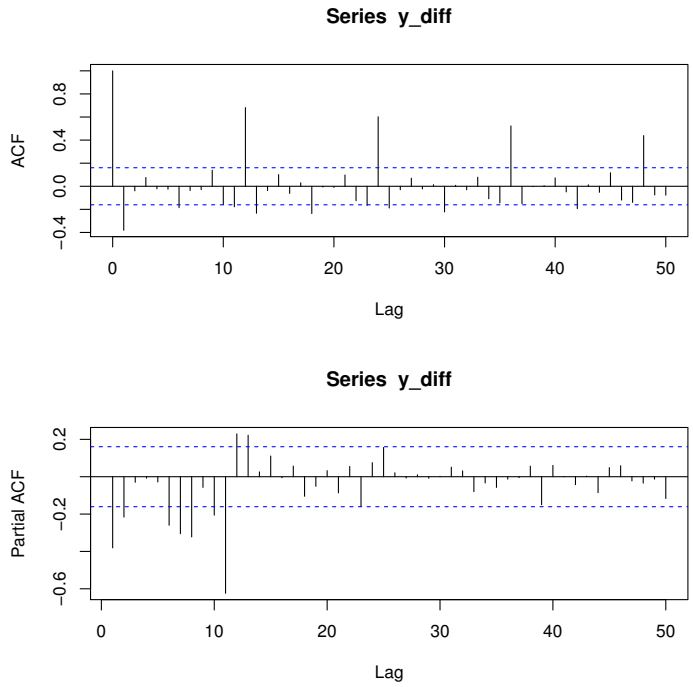


Figure 7.11: $SACF$ and $SPACF$ of the differentiated data. The first difference mitigate the slow decaying trend leaving the seasonal component of period 12. The cut off after lag 12 of the $SPACF$ suggest an a seasonal autoregressive component of order 1.

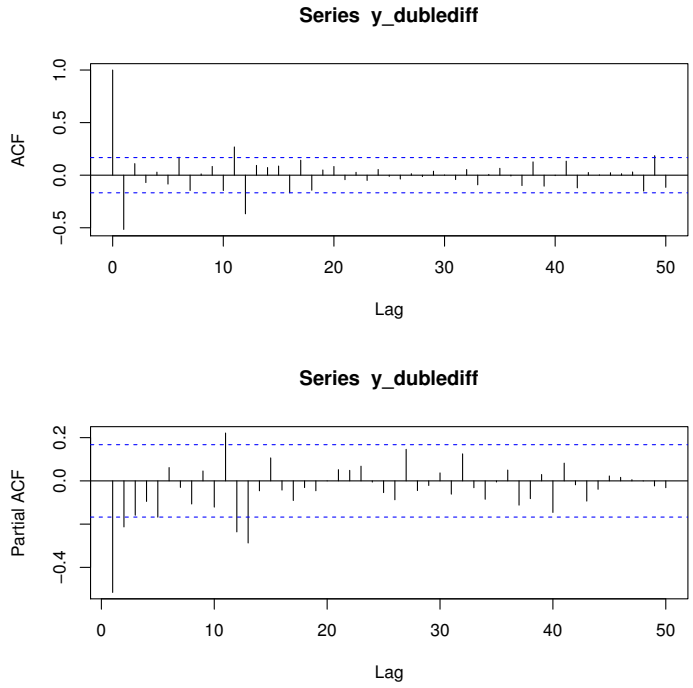


Figure 7.12: $SACF$ and $SPACF$ of after applying the filter $(1-B)(1-B^{12})$ on the retail sales data.

**Forward Search Analysis of the Retail Sales Data**

The *FSTS* as extensively discussed before provides some visual tools that can be used to analyze the residuals and shed light on the presence of atypical and important observations.
The comparison of the trajectory plot of the residuals obtained with the Forward Search performed on the different three models are reported respectively in Figure 7.13, 7.14 and 7.15.

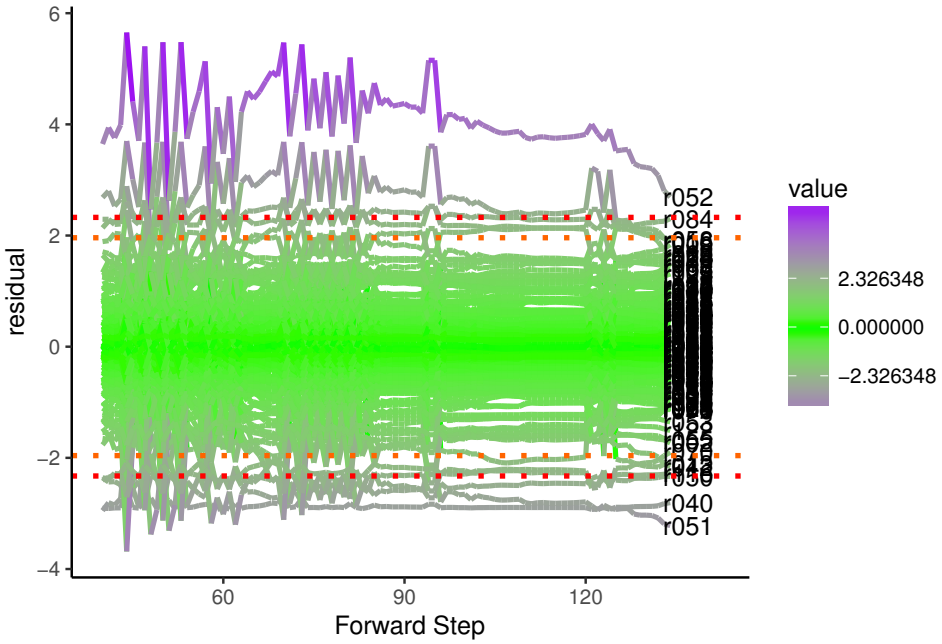


Figure 7.13: Trajectory plot of the standardized residuals when the Forward Search is performed on $\mathcal{M}_1$. Units 40, 51, 52 and 84 show trajectories outside from the threshold values for the entire Search.

As expected, the unit flagged as potential outliers depend on initial specification of the model. However, a comparison of the residuals plots obtained by the three different models might reveal a common structure between the models.
In particular, a visual inspection of the residuals trajectories shows that according to all the three models, units 40 and 51 are flagged as possible outliers.
Units 14, 16 and 116 are candidate outliers according to $\mathcal{M}_2$ and $\mathcal{M}_3$, while unit 84 results atypical in $\mathcal{M}_1$ and $\mathcal{M}_2$. Thus, from a visual analysis of the residual plots it is therefore possible to identify which are the influential units given a postulated model and, on the other hand, identify the most outlying units by a direct comparison of the plots generated by different models.
In this particular case, since the trajectories associated to units 40 and 51 appear wild according to all of the three assumed models, we can conclude with a certain degree of confidence that they are highly influential units and possible outliers.

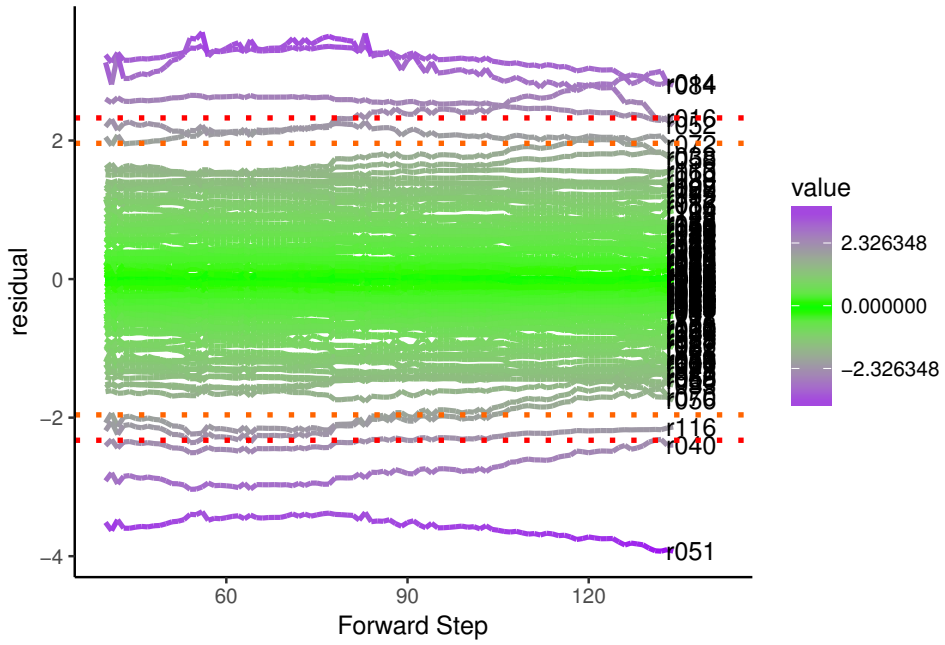


Figure 7.14: Trajectory plot of the standardized residuals when the Forward Search is performed on $\mathcal{M}_2$. In agreement with $\mathcal{M}_1$ units 40, 51 and 84 show trajectories outside the threshold values for the entire Search. Interestingly, when $\mathcal{M}_1$ is employed, the trajectory of unit 14 increase sensibly at step 110.

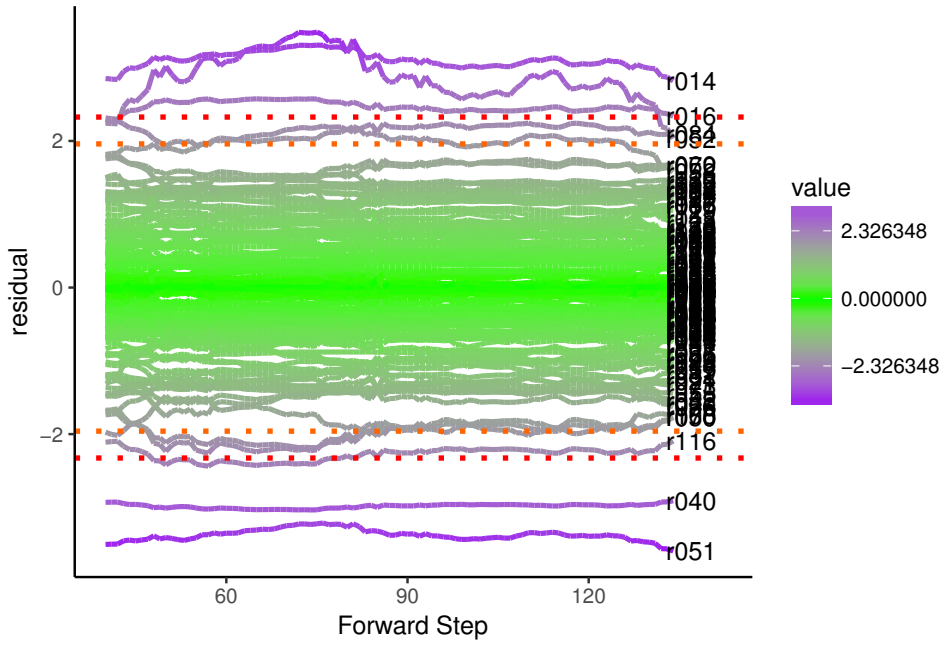


Figure 7.15: Trajectory plot of the standardized residuals when the Forward Search is performed on $\mathcal{M}_3$. In accordance with $\mathcal{M}_1$ and $\mathcal{M}_2$, units 40, 51, show trajectories outside the threshold values for the entire Search. Moreover units 14 and 16 results as possible outliers too.

Taking a step forward it is possible to compare the trajectory plots of the statistics used to identify whether a given unit is an $IO$, an $AO$ or a $LS$.

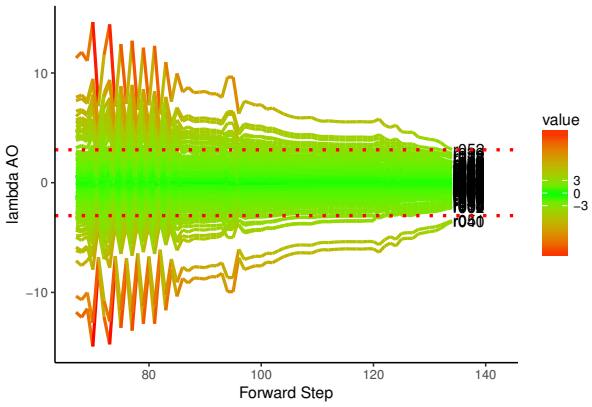


Figure 7.16

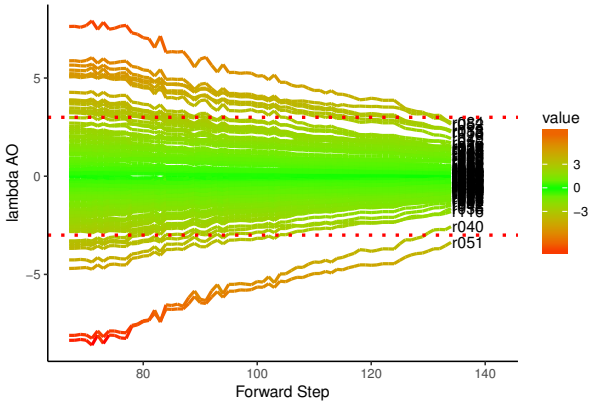


Figure 7.17

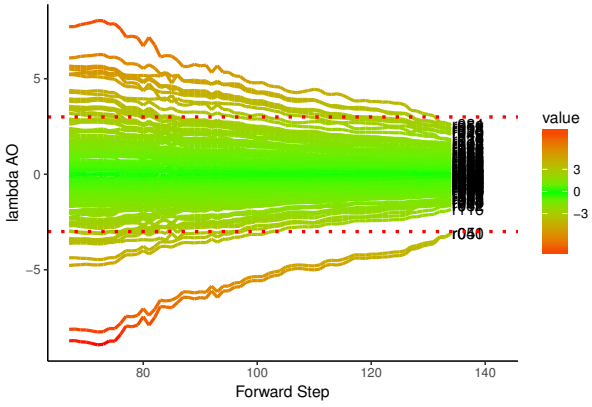


Figure 7.18: The trajectory plots of the additive outlier diagnostics for the three models employed. According to all of the three models unit 40 and 51 are flagged as possible atypical units.

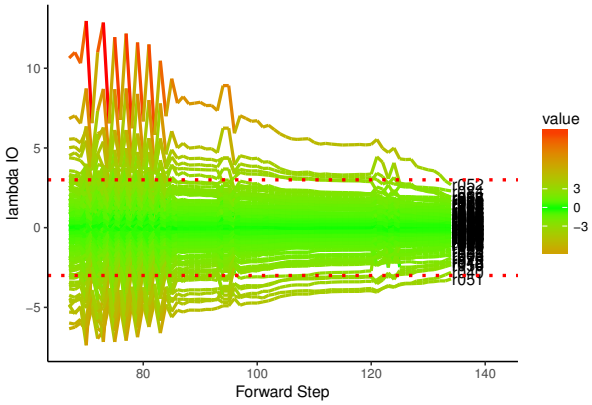


Figure 7.19

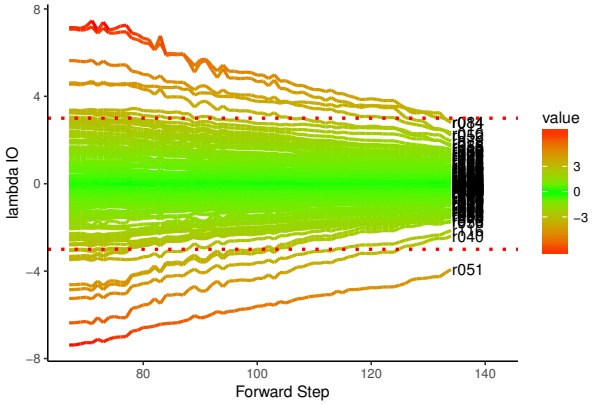


Figure 7.20

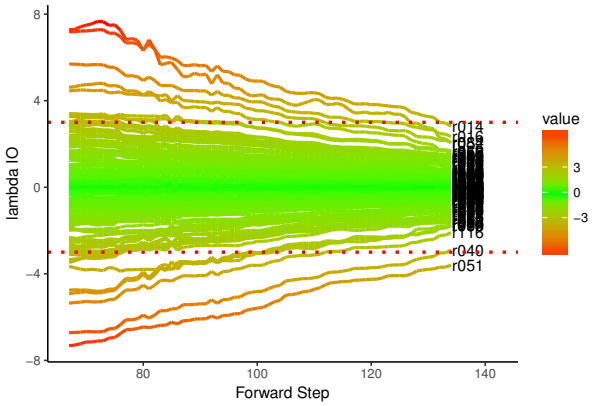


Figure 7.21: The trajectory plots of the innovation outlier diagnostics for the three models employed. According to all of the three models unit 51 is flagged as a potential innovation outlier, while the diagnostics for unit 40 result being significant only when $\mathcal{M}_3$ is used.

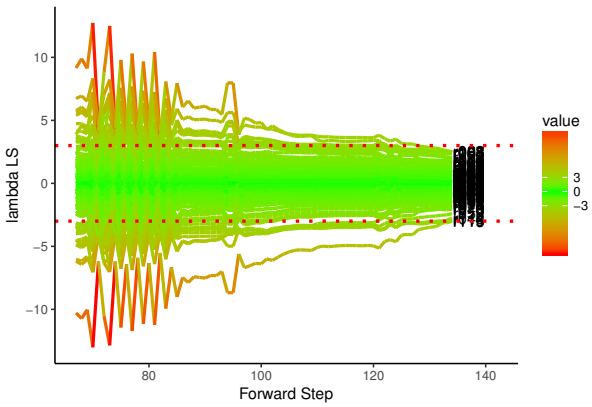


Figure 7.22

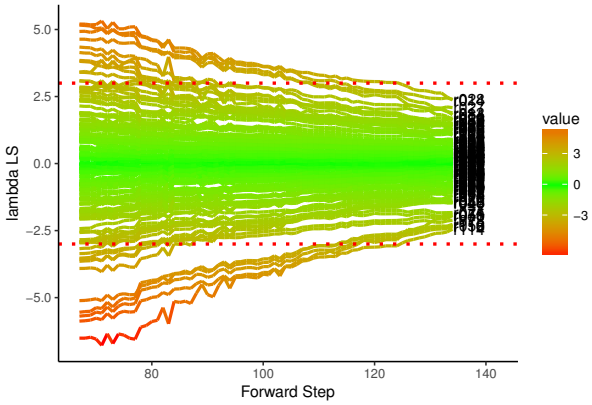


Figure 7.23

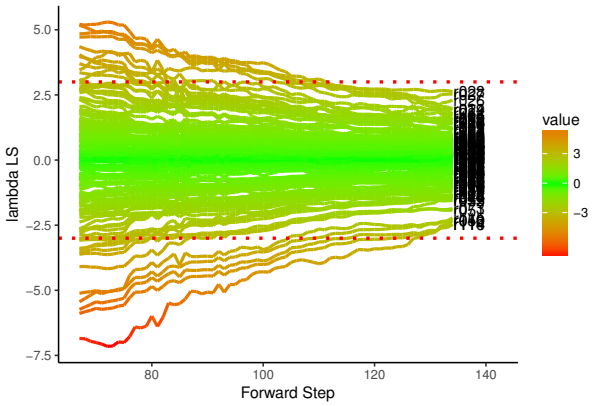


Figure 7.24: The trajectory plots of the level shift diagnostics for the three models employed. Clearly from the plots there is no evidence of any level shifts. Nevertheless when $\mathcal{M}_1$ is employed the $\hat{\lambda}_{LS,51}$ remains far from the bulk of the data for a long period in the search.

The plots confirms that the units 41 and 50 are clearly two outliers. The outlyingness of these two units is larger when model $\mathcal{M}_2$ or $\mathcal{M}_3$ are considered, as their trajectories remain markedly distant from the bulk for the entire path of the search.
For all of the three models there is no clear evidence of level shifts in the data, however, when $\mathcal{M}_1$ is assumed, the trajectory of $\hat{\lambda}_{LS,51}$ remains far from the bulk of the data for

the majority of the $FSTS$.

It is clear from this analysis that the degree of outlyingness of one unit depends heavily on the underlying assumed model, and a consequence, automatic procedures that have been developed, despite commonly employed in practical applications may lead to false discoveries in the data or to misleading specifications.

In order to discover the nature of the different units it is possible to analyze the leaf plots already proposed for the analysis of the Airline Passenger data, and further insights are provided by the automatic outlier detection procedure whose results are reported din the following tables.

Table 7.4: Outliers identified when $\mathcal{M}1$ is employed

| $\tau$ | $\lvert w_{\tau,AO}M^{\lambda}_{\tau,AO}\rvert$ | $\lvert w_{\tau,IO}M^{\lambda}_{\tau,IO}\rvert$ | $\lvert w_{\tau,LS}M^{\lambda}_{\tau,LS}\rvert$ | $i$ |
|---|---|---|---|---|
| 2 | **0.13705** | 1.13256 | 0 | IO |
| 14 | **0.8681** | 0.66 | 0.09666 | AO |
| 28 | **1.0439** | 0.75524 | 0.2731 | AO |
| 38 | 0.21289 | **1.11078** | 0.38246 | IO |
| 39 | **1.60904** | 0.86259 | 0.82045 | AO |
| 40 | **7.16624** | 2.04588 | 5.59623 | AO |
| 49 | 0 | **0.84966** | 0.08 | IO |
| 50 | 1.26513 | 0.83938 | **1.39455** | LS |
| 51 | **6.0519** | 0.90254 | 0.88458 | AO |
| 52 | 5.4566 | **6.21862** | 1.70581 | IO |
| 58 | **1.46254** | 0.19975 | 1.45933 | AO |
| 70 | **1.18849** | 0 | 0.49671 | AO |
| 114 | 0.18236 | **1.138** | 0.72721 | IO |
| 116 | **0.29456** | 0.08015 | 0.26933 | AO |

Table 7.5: Outliers identified when $\mathcal{M}2$ is employed

| $\tau$ | $\lvert w_{\tau,AO}M^{\lambda}_{\tau,AO}\rvert$ | $\lvert w_{\tau,IO}M^{\lambda}_{\tau,IO}\rvert$ | $\lvert w_{\tau,LS}M^{\lambda}_{\tau,LS}\rvert$ | $i$ |
|---|---|---|---|---|
| 14 | 0.9661 | **3.81676** | 0.06513 | IO |
| 16 | 0.86736 | 1.24312 | **1.14506** | IO |
| 28 | 1.20439 | **1.12667** | 0 | AO |
| 38 | 0.67013 | 0.91641 | **1.13919** | LS |
| 40 | **4.93824** | 2.38853 | 2.15645 | AO |
| 49 | 0 | **1.096** | 0.67089 | IO |
| 50 | **1.15743** | 0.00515 | 0.25214 | AO |
| 51 | **4.96686** | 4.28007 | 0.90189 | AO |
| 52 | **3.75322** | 3.60455 | 1.12051 | AO |
| 58 | **0.91662** | 0.01856 | 0.1795 | AO |
| 70 | **1.2476** | 0.1274 | 1.05914 | AO |
| 84 | 0.00498 | 0 | **0.03061** | LS |
| 113 | 0.08713 | **1.11721** | 0.91461 | IO |
| 114 | 0.44486 | 1.10192 | **1.18418** | LS |
| 115 | 0.19914 | **0.68908** | 0.58298 | IO |
| 116 | 0.27438 | 0.73842 | **0.98743** | LS |

Table 7.6: Outliers identified when $\mathcal{M}3$ is employed

| $\tau$ | $\lvert w_{\tau,AO}M^{\lambda}_{\tau,AO}\rvert$ | $\lvert w_{\tau,IO}M^{\lambda}_{\tau,IO}\rvert$ | $\lvert w_{\tau,LS}M^{\lambda}_{\tau,LS}\rvert$ | $i$ |
|---|---|---|---|---|
| 14 | 1.1212 | **3.9056** | 0.10162 | IO |
| 16 | 0.24242 | **1.45207** | 0.94676 | IO |
| 28 | **1.13403** | 1.07306 | 0.35805 | AO |
| 38 | 0.36856 | 0.88927 | **0.98255** | LS |
| 40 | **4.68789** | 3.67856 | 2.22688 | AO |
| 49 | 0 | **1.05542** | 0.51018 | IO |
| 50 | **0.858** | 0.00477 | 0.56909 | AO |
| 51 | 4.05254 | **4.76424** | 0.6577 | IO |
| 52 | **2.36999** | 1.34921 | 1.04626 | AO |
| 58 | **1.13229** | 0 | 0.04295 | AO |
| 84 | 0.04864 | 0.11269 | **0.13913** | LS |
| 113 | 0.00504 | **1.05093** | 0.82936 | IO |
| 114 | 0.66292 | 1.05377 | **1.48304** | LS |
| 115 | **0.50327** | 0.25638 | 0 | AO |
| 116 | 0.0741 | 0.91165 | **1.11312** | LS |

As expected despite the different models lead to different patterns of outliers it is possible to identify a group of units that results atypical in all of the three models proposed.
The common structure is reported in Table in 7.7.

Table 7.7: Common structure identified by performing the $FSTS$ on the three different assumed models. All of the three model classify units 28, 40, 49 and 58 with the same outlier type, Moreover, it is observed that $\mathcal{M}_2$ and $\mathcal{M}_3$ that contains a seasonal difference agree on the majority of types of outliers.

| $\tau$ | $\mathcal{M}1$ | $\mathcal{M}2$ | $\mathcal{M}3$ |
|---|---|---|---|
| 14 | AO | **IO** | **IO** |
| 28 | **AO** | **AO** | **AO** |
| 38 | IO | LS | LS |
| 40 | **AO** | **AO** | **AO** |
| 49 | **IO** | **IO** | **IO** |
| 50 | LS | **AO** | **AO** |
| 51 | **AO** | **AO** | IO |
| 52 | IO | **AO** | **AO** |
| 58 | **AO** | **AO** | **AO** |
| 114 | IO | **LS** | **LS** |
| 116 | AO | **LS** | **LS** |

By looking at Table 7.7 it is clear that units 28, 40 and 58 are flagged by $AO$ by all of the three models. Interestingly, models $\mathcal{M}_2$ and $\mathcal{M}_3$, both of which contains a seasonal difference, agreed on the classification of the majority of the outliers.
We conclude this analysis by showing the stem plot of the diagnostics for the three different models. As it is evident, the stem plot en lights the same results obtained with the automatic outlier detection procedure.

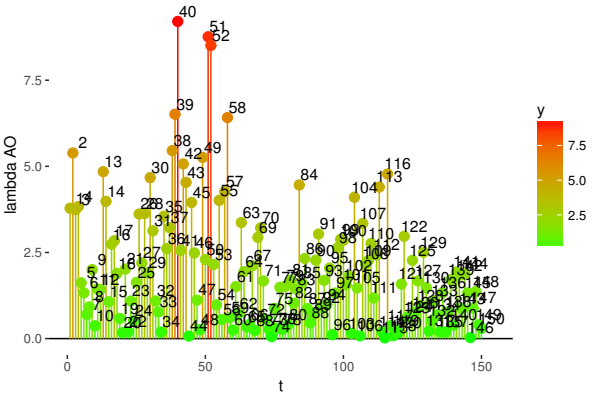


Figure 7.25

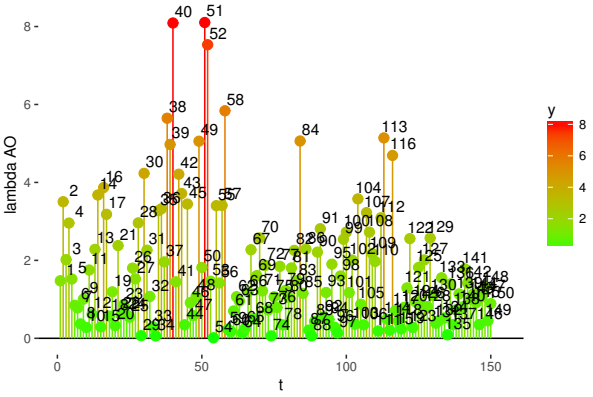


Figure 7.26

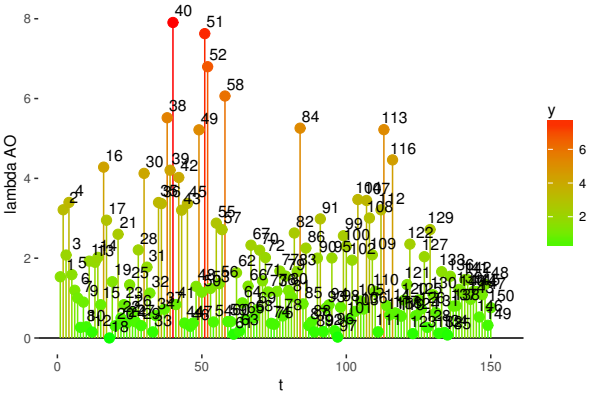


Figure 7.27

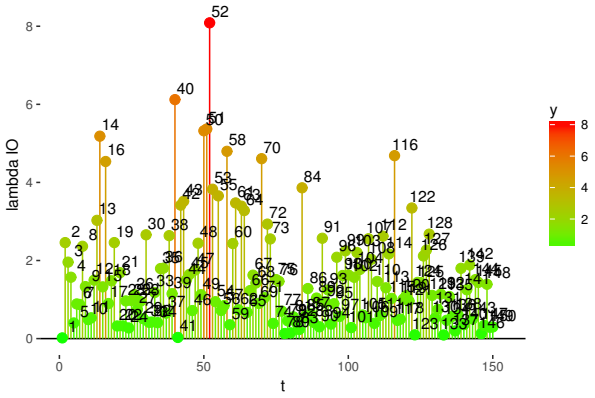


Figure 7.28

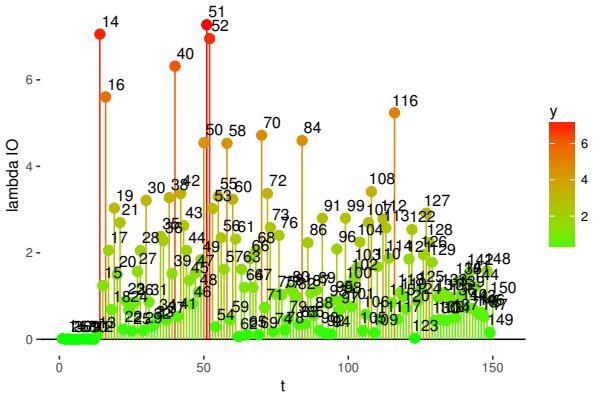


Figure 7.29

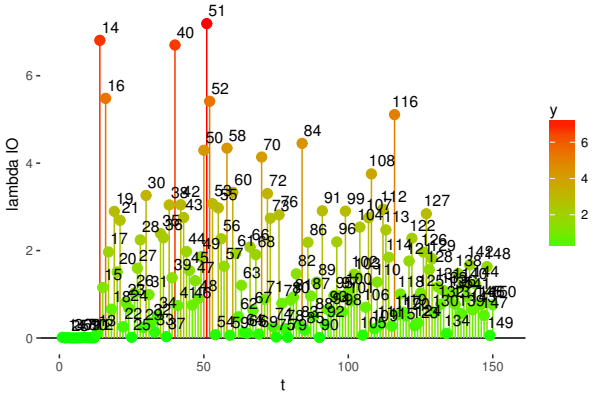


Figure 7.30

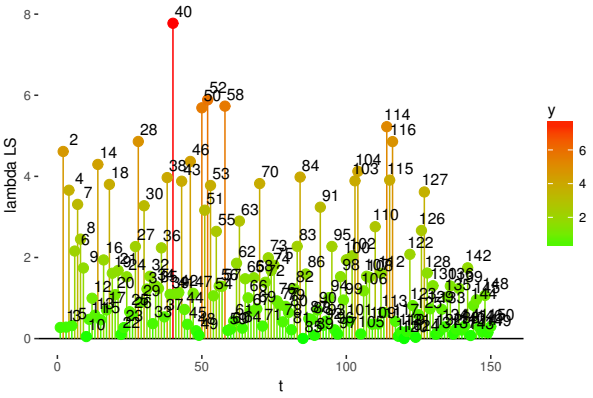


Figure 7.31

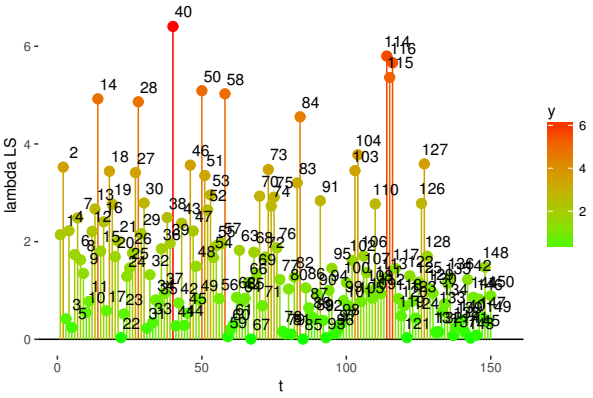


Figure 7.32

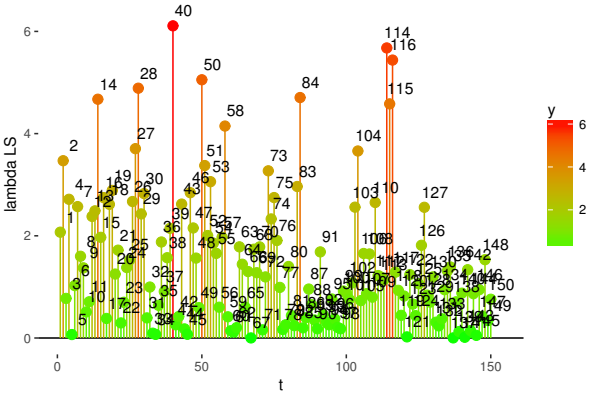


Figure 7.33

# Conclusions

In this thesis I have investigated the possibility of extending the Forward Search procedure to the class of $SARIMA$ models, where, the Kalman filter resulted a key ingredient. The applications on the two real time series, despite simple, highlighted the strength and weaknesses of this procedure applied to this class of models. Further results from a simulation study might be needed to strengthen the findings and validate the automatic outlier detection procedure.

However, in this starting phase, I believe that the main strength of the $FSTS$ relies in the series of useful plots that can be used to discover, in an exploratory analysis curious patterns in the data.

First, the information provided by the trajectory plots obtained with the $FSTS$ allow the analyst to discover possible interesting patterns in the series and highlight what can next be discovered being atypical units. Second, the leaf plot provide a robust visual tool that can be used to shed some light on the findings discovered by the visual inspection of the preliminary residuals plots. Since in the stem plot the natural ordering of the data is preserved, and because the statistics are obtained in a robust way, its interpretation is straightforward and the information that it provides is not affected by the swamping and masking effects. Finally, the automatic outliers detection procedure, despite its need of further tuning, seems an important step in extending the Forward Search as an automatic outliers detection technique.

On the other hand the application of the Forward Search to the class of $SARIMA$ models apprise some problems. In particular, I found that the Kalman filter results unstable when the units in a given step of the $FSTS$ are particularly sparse in time, that is to say when we have many missing times between one observation and the others. In extreme cases, this may result in the non convergence of the algorithm. Next, the identification of the type of outliers $AO$, $IO$ or $LS$ relies heavily on the natural time ordering of the data. As a consequence, the standard trajectory plots adopted in the search cannot be used to achieve this task. And that is why I have implemented the stem plot to try to maintain the original natural time ordering of the series.

Last but not least, the outlying nature of a given unit relies heavily on the postulated model. In order to overcome this problem two different paths seem natural. First, the initial $SARIMA$ model might be specified adopting robust procedures, such as robust

$ACF$ and $PACF$ or robust information criteria. Second, a structural model might be considered. By adopting a structural model the specification part should result simpler since depending on the visual characteristics of the series the analyst should specify which unobserved components to be included in the model.

Overall, I believe that the $FSTS$ is an extremely flexible and useful procedure that can be well adapted to many different statistical contexts as it is evident from the wide range of applications that have been proposed in literature. Moreover, the analysis of outlying units in a time series context should not be done blindly and by relying only on the automatic procedures, but the analyst should always base his judgment on external and/or personal knowledge on the data at hand.

Moreover, I would like to mention the possibility of extending the procedure proposed in Crosato and Grossi (2017) to other settings. The idea of retaining all the observations in the $FSTS$ and assign an adaptive weight is a reasonable requirement especially in time series analysis where the natural time ordering of the data seems particularly relevant for the correct identification of the type of outliers.

In this direction on one hand it might be worthed to considered different weight functions to see if the robustness properties of the search might be improved. On the other hand, it might be interesting to correct the series when an outliers is identified and proceed the $FSTS$ on the cleaned series in a manner similar to the procedures discussed in Chapter 5.

# Bibliography


Abraham, B. and Box, G. E. (1979). Bayesian analysis of some outlier problems in time series. *Biometrika*, 66(2):229–236.

Abraham, B. and Chuang, A. (1989). Outlier detection and time series modeling. *Technometrics*, 31(2):241–248.

Abraham, B. and Yatawara, N. (1988). A score test for detection of time series outliers. *Journal of time series analysis*, 9(2):109–119.

Atkinson, A. and Riani, M. (2012). *Robust diagnostic regression analysis*. Springer Science & Business Media.

Atkinson, A. C., Cerioli, A., and Riani, M. (2016a). Discussion of 'asymptotic theory of outlier detection algorithms for linear time series regression models' by johansen and nielsen. *Scandinavian Journal of Statistics*, 43(2):349–352.

Atkinson, A. C. and Riani, M. (2004). The forward search and data visualisation. *Computational Statistics*, 19(1):29–54.

Atkinson, A. C., Riani, M., and Cerioli, A. (2010). The forward search: Theory and data analysis. *Journal of the Korean Statistical Society*, 39(2):117–134.

Atkinson, A. C., Riani, M., and Torti, F. (2016b). Robust methods for heteroskedastic regression. *Computational Statistics & Data Analysis*, 104:209–222.

Balke, N. S. and Fomby, T. B. (1994). Large shocks, small shocks, and economic fluctuations: Outliers in macroeconomic time series. *Journal of Applied Econometrics*, 9(2):181–200.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

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

Brockwell, P. J. and Davis, R. A. (2016). *Introduction to time series and forecasting*. springer.

Bruce, A. G. and Martin, R. D. (1989). Leave-k-out diagnostics for time series. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 363–424.

Bustos, O. H. and Yohai, V. J. (1986). Robust estimates for arma models. *Journal of the American Statistical Association*, 81(393):155–168.

Chang, I., Tiao, G. C., and Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, 30(2):193–204.

Chen, C. and Liu, L.-M. (1993a). Forecasting time series with outliers. *Journal of Forecasting*, 12(1):13–35.

Chen, C. and Liu, L.-M. (1993b). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, 88(421):284–297.

Chen, C. and Tiao, G. C. (1990). Random level-shift time series models, arima approximations, and level-shift detection. *Journal of Business & Economic Statistics*, 8(1):83–97.

Crosato, L. and Grossi, L. (2017). Correcting outliers in garch models: a weighted forward approach. *Statistical Papers*, pages 1–32.

Doornik, J. A., Hendry, D. F., and Pretis, F. (2013). Step-indicator saturation. *University of Oxford. Working Paper*, 658.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, volume 38. Oxford University Press.

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

Franses, P. H. and Haldrup, N. (1994). The effects of additive outliers on tests for unit roots and cointegration. *Journal of Business & Economic Statistics*, 12(4):471–478.

Grossi, L. (2004). Analyzing financial time series through robust estimators. *Studies in Nonlinear Dynamics & Econometrics*, 8(2).

Grossi, L. and Laurini, F. (2009). A robust forward weighted lagrange multiplier test for conditional heteroscedasticity. *Computational Statistics & Data Analysis*, 53(6):2251–2263.

Grossi, L. and Riani, M. (2002). Robust time series analysis through the forward search. In *Compstat*, pages 521–526. Springer.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 761–771.

Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88(424):1264–1272.

Hamilton, J. D. (1994). *Time series analysis*, volume 2. Princeton university press Princeton, NJ.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons.

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge university press.

Harvey, A. C. (1993). Time series models.

Hayashi, F. (2000). *Econometrics*. Princeton university press Princeton, NJ.

Heagerty, P. J. and Lumley, T. (2000). Window subsampling of estimating functions with application to regression models. *Journal of the American Statistical Association*, 95(449):197–211.

Hendry, D. F. (1999). An econometric analysis of us food expenditure, 1931-1989.

Hendry, D. F. and Krolzig, H.-M. (2004). Resolving three 'intractable'problems using a gets approach. *Unpublished paper, Economics Department, University of Oxford*.

Huber, P. and Ronchetti, E. (2009). Robust statist., hoboken.

Johansen, S. and Nielsen, B. (2016). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics*, 43(2):321–348.

Koopman, S. J. (1997). Exact initial kalman filtering and smoothing for nonstationary time series models. *Journal of the American Statistical Association*, 92(440):1630–1638.

Leroy, A. M. and Rousseeuw, P. J. (1987). Robust regression and outlier detection. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987*.

Ljung, G. M. (1993). On outlier detection in time series. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 559–567.

Maddala, G. S. and Kim, I.-M. (1998). *Unit roots, cointegration, and structural change*. Number 4. Cambridge university press.

Marczak, M. and Proietti, T. (2016). Outlier detection in structural time series models: The indicator saturation approach. *International Journal of Forecasting*, 32(1):180–202.

Marczak, M., Proietti, T., and Grassi, S. (2018). A data-cleaning augmented kalman filter for robust estimation of state space models. *Econometrics and Statistics*, 5:107–123.

Maronna, R., Martin, R. D., and Yohai, V. (2006). *Robust statistics*, volume 1. John Wiley & Sons, Chichester. ISBN.

McQuarrie, A. D. and Tsai, C.-L. (2003). Outlier detections in autoregressive models. *Journal of Computational and Graphical Statistics*, 12(2):450–471.

Muirhead, C. R. (1986). Distinguishing outlier types in time series. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 39–47.

Peña, D. (1990). Influential observations in time series. *Journal of Business & Economic Statistics*, 8(2):235–241.

Peña, D., Tiao, G. C., and Tsay, R. S. (2011). *A course in time series analysis*, volume 322. John Wiley & Sons.

Proietti, T. and Alessandra, L. (2013). Maximum likleihood estimation of time series models: the kalman filter and beyond. In *Handbook of Research Methods and Applications in Empirical Macroeconomics*, pages 334–362. Edward Elgar Publishing.

Riani, M. (2004). Extensions of the forward search to time series. *Studies in Nonlinear Dynamics & Econometrics*, 8(2).

Riani, M. (2008). Robust transformations in univariate and multivariate time series. *Econometric Reviews*, 28(1-3):262–278.

Riani, M., Cerioli, A., and Torti, F. (2014). On consistency factors and efficiency of robust s-estimators. *Test*, 23(2):356–387.

Rousseeuw, P., Perrotta, D., Riani, M., and Hubert, M. (2018). Robust monitoring of time series with application to fraud detection. *Econometrics and statistics*.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880.

Stockinger, N. and Dutter, R. (1987). Robust time series analysis: A survey. *Kybernetika*, 23(7):1–3.

Tsay, R. S. (1986). Time series model specification in the presence of outliers. *Journal of the American Statistical Association*, 81(393):132–141.

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