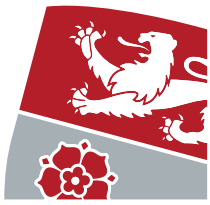


Lancaster University
Management School

Complex Exponential Smoothing

Ivan Svetunkov
Management School
Lancaster University

A thesis submitted for the degree of
Doctor of Philosophy

2016

Acknowledgements

It was one of those evenings during our holiday of the year 2012 at the small town of Antibes, not far from Cannes. The night was black as a cloak of a gloomy wizard and only a candle on the table saved us from the complete darkness. Anna and I were drinking cheap French red wine, eating cheeses and talking about our miserable lives, which obviously at that point had no bright future. Thinking about that right now, life back then was not as miserable as we thought but at least it was not as exiting as we wanted. One of the main reservations I had back then is that I could not develop skills in the field I liked and no one at the department of the university I worked in could support me.

'I don't think I have any future there.' I said, making a small sip from the glass and biting off a piece of cheese. The wine was mature and rough, while Camembert was strong, smelly but soothing.

'No, you probably don't.' Anna agreed, propping up her chin on a hand. 'You obviously don't like economics. But, let's face it, this is the only area you can work in Saint Petersburg.'

'Yes,' I agreed, 'for some reason forecasting is not popular in Russia. And no one really understands my idea of Complex Exponential Smoothing. . . '

Anna suddenly shuddered and stood up straight as if struck by ingenious thought.

‘I know what you got to do!’ she said in excitement. ’You should apply for PhD!’

And now I am sitting in my office in front of computer and I cannot fully understand, why I have a finalised thesis in front of me and how I have got here, to this moment, in this location. Obviously I would not be able to do it on my own, and the fact that I have done such a long route in such a short period: from Antibes to Lancaster, from idea to thesis in just three years – amazes me deeply.

This route would not be possible without the hard work of Nikolaos Kourentzes, my supervisor and friend, without the precious support from Robert Fildes, my second supervisor and intellectual mentor, without advices and comments of Keith Ord and detailed feedback of Rebecca Killick. I am very grateful to all of you for your guidance and support!

And I am grateful for the moral support that my wife, Anna, provided me through these years. Without her I would be in some other place, probably doing something much more boring. . . like economics.

Abstract

Exponential smoothing is one of the most popular forecasting methods in practice. It has been used and researched for more than half a century. It started as an ad-hoc forecasting method and developed to a family of state-space models. Still all exponential smoothing methods are based on time series decomposition and the usage of such components as "level", "trend", "seasonality" and "error". It is assumed that these components may vary from one time series to another and take different forms depending on data characteristics. This makes their definition arbitrary and in fact there is no single way of identifying these components.

At the same time the introduction of different types of exponential smoothing components implies that a model selection procedure is needed. This means that a researcher needs to select an appropriate type of model out of 30 different types either manually or automatically for every time series analysed. However several recent studies show that an underlying statistical model may have a form completely different than the one assumed by specific exponential smoothing models.

These modelling questions motivate our research. We propose a model without strictly defined "level", "trend" and "seasonality". The model greatly simplifies the selection procedure, distinguishing only between seasonal and non-seasonal time series. Although we call it "Complex Exponential Smoothing" (CES), due to the use of complex-valued functions, its usage simplifies the forecasting procedure.

In this thesis we first discuss the main properties of CES and propose an underlying statistical model. We then extend it in order to take seasonality into account

and conduct experiments on real data to compare its performance with several well-known univariate forecasting models. We proceed to discuss the parameters estimation for exponential smoothing and propose a "Trace Forecast Likelihood" function that allows estimating CES components more efficiently. Finally we show that Trace Forecast Likelihood has desirable statistical properties, is connected to shrinkage and is generally advisable to use with any univariate model.

Contents

Introduction

When a company needs to make a business decision, it requires information about the market, its clients' and competitors' behaviour in future. There is always a very high level of uncertainty in all of these areas. In order to decrease it companies usually produce point and interval forecasts, which allow them to make adequate decisions. The uncertainty may decrease even more if these are produced using statistical methods, ensuring that the selected model is correctly specified.

The theory of forecasting is broad and very well developed nowadays. There are basic principles that ensure that the more accurate forecasts are produced and there are several popular methods that have been shown to work effectively in practice. One of the most popular forecasting methods is "Exponential Smoothing". It was proposed independently by Brown (1956) and by Holt (2004). Due to its simplicity and general robustness it has become very popular amongst practitioners and is widely used in forecasting applications. Nevertheless exponential smoothing methods have been ignored by statisticians for many years because of the absence of an underlying statistical model. The gap between theory and practice was substantially overcome when Snyder proposed a Single Source of Error State Space model (Snyder, 1985) and then from the end of 90s to the beginning of 00s Rob J. Hyndman, Anne Koehler, Keith Ord and Ralph Snyder produced several papers, developing the theory of exponential smoothing (Ord et al., 1997; Koehler et al., 2001; Hyndman et al., 2002, 2005, 2008a). Finally they systematised all knowledge in the field and published a textbook on exponential smoothing (Hyndman et al., 2008b).

The motivation of this PhD thesis is that although exponential smoothing is very popular and has been shown to perform well in practice, it has not seen any substantial improvements since Hyndman et al. (2008b). Nevertheless some of the most popular exponential smoothing methods do not perform consistently. In a literature review presented in chapter 1 we show that there has not been a substantial change in modelling principles in time series over the last decades. Furthermore several research studies show that the existing exponential smoothing methods do not perform as accurately as expected in various cases. The explanation of this is that an underlying "true model" (if it exists at all) may be more complicated than thought, thus it cannot be properly captured by the existing techniques.

In chapter 2 we propose a new approach to time series modelling and forecasting, which is based on the notion of "information potential". The idea behind this is that each time series contains some hidden information which can not be properly revealed using conventional methods. In order to take it into account it is proposed to use a complex variables representation, using the imaginary part as some approximation of this unobservable component. Using this idea and the mechanism of exponential smoothing, we propose a model called "Complex Exponential Smoothing" (CES). The proposed model allows a smooth transition between level and trend data, and is able to model different types of time series. We discuss the statistical properties of CES and derive a state-space form. We then discuss its forecasting capabilities and show on several experiments how it performs. Finally we compare the forecast accuracy of CES with conventional Exponential smoothing (ETS) and autoregressive model with movina average (ARIMA) models on M3 monthly data, demonstrating its superiority on this dataset.

The original CES does not take possible seasonality into account, but these types of time series form a substantial part of all business data. This motivates further research and in chapter 3 we discuss an extension of the proposed model, which

allows modelling both additive and multiplicative seasonalities and introduces new form (where level of the series is constant, but the amplitude of seasonality changes). We then discuss a model selection procedure and demonstrate with a simulation that the selection between seasonal and non-seasonal CES is done more accurately than for other commonly used univariate statistical models. Finally we conduct an experiment on M3 data and show that CES with model selection outperforms ETS and ARIMA, mainly due to the increased long-term forecasting accuracy.

Knowing the properties of CES, it is essential to estimate correctly its complex smoothing parameters, which motivates the last methodological chapter of this thesis. In order to overcome the limitations of the existing estimation methods we studied the alternatives and came to conclusion that one of the possible solutions to the problem lies in the application of estimators based on multiple steps ahead forecast errors. In chapter 4 we discuss the limitation of existing methods, give explanation of the processes happening when they are applied to univariate models and propose a new technique that we call "Trace Forecast Likelihood" (TFL). We discuss its statistical and optimisation properties and demonstrate why the proposed method guarantees more accurate estimates of parameters. We then conduct a simulation in order to support our theoretical findings and demonstrate the superiority of the proposed estimator. We also show the general applicability of TFL to any univariate time series model.

The conclusions and the possible future research directions follow in the last chapter where we identify new research avenues that open based on CES, TFL and ideas behind them.

Chapter 1

Literature Review

1.1 Introduction

Ten years have passed since Gardners review (Gardner, 2006) which summarized progress in the area of exponential smoothing (ETS) since his previous review (Gardner, 1985). Those twenty years had brought a lot of advances to the field of exponential smoothing: the state-space approach using single source of errors (SSOE) has been developed (which allowed the calculation of variances and likelihood for ETS), the systematization of different exponential smoothing methods has been established and a general taxonomy, uniting all these methods, has been proposed. In addition exponential smoothing models selection procedures have been introduced. All of these were parts of the big step towards forecasting accuracy improvement in the field of exponential smoothing. The substantial contribution to the field since 2006 is the textbook by Hyndman et al. (2008b) which systematizes all the main results achieved in the exponential smoothing in a unified framework and shows how to use these models in practice in different situations.

Since then the field developed slowly introducing only minor changes in the existing models. Still there has been some progress and our literature review demonstrates the most interesting papers in the field organising the contributions in the following categories:

1. Why damped trend method works. Damped trend method has been shown to perform very well in many cases and is considered as a robust forecasting method, meaning that it can be applied to time series of different types in order to obtain good forecasts. Here we analyse the articles aiming to give an explanation of the performance of this method.

2. Combinations of forecasts. It has been shown since the first M-Competitions (Makridakis et al., 1982) that combinations of forecasts perform very well, sometimes even outperforming individual methods. Usually exponential smoothing is included in combinations exclusively or as one of the forecasting methods. Several papers in this area have been included in this section.

3. Hybrid models. Another approach used in order to increase forecasting accuracy uses hybrid models. Arguably this approach follows the same ideas as combination of forecasts but from a more sophisticated perspective, often including models with complimentary properties.

4. Temporal aggregation. There is renewed interest to this approach to time series forecasting. It employs the idea that aggregation of time series on different frequencies may reveal some hidden patterns in data.

5. High-frequency data. Methods for high-frequency data forecasting are considered to be more complicated than the methods for slow moving data. This is because this type of data contains many more patterns and information. Usually exponential smoothing and ARIMA models are modified in order to take this into account. This area has seen the substantial development over the last decade but still holds several unanswered questions.

Obviously there have been more papers that make use of exponential smoothing over the last ten years, but those discussed in this review demonstrate gaps in the field and motivate our research.

Before we proceed with more details on the new forecasting methods, we need to clarify what exponential smoothing is. In this thesis we discuss single source of error state-space model underlying all the exponential smoothing methods that can be formulated the following way (Hyndman et al., 2008b):

$$\begin{matrix} y_t = w(v_{t-l}) + r(v_{t-l})\epsilon_t \\ v_t = f(v_{t-l}) + g(v_{t-l})\epsilon_t \end{matrix}, \tag{1.1}$$

where y_t is actual value of time series, ϵ_t is error term, v_{t-l} is a state vector, $w(.)$ is measurement function, $r(.)$ is error term function, $f(.)$ is transition function and $g(.)$ is persistence function.

A special case of this model is pure additive ETS, which can be written as:

$$\begin{matrix} y_t = w' v_{t-l} + \epsilon_t \\ v_t = F v_{t-l} + g\epsilon_t \end{matrix}, \tag{1.2}$$

where w is measurement vector, F is transition matrix, g is persistence vector.

Simple Exponential Smoothing, being an important method for this thesis, has an underlying ETS(A,N,N) state-space model:

$$\begin{matrix} y_t = l_{t-1} + \epsilon_t \\ l_t = l_{t-1} + \alpha\epsilon_t \end{matrix}, \tag{1.3}$$

where l_t is level of series and α is smoothing parameter for level.

The other important method - Holt's method - has the following underlying ETS model:

$$\begin{matrix} y_t = l_{t-1} + b_{t-1} + \epsilon_t \\ l_t = l_{t-1} + b_{t-1} + \alpha\epsilon_t, \\ b_t = b_{t-1} + \beta\epsilon_t \end{matrix} \tag{1.4}$$

where b_t is trend component and β is smoothing parameter for trend. This model in (Hyndman et al., 2008b) taxonomy is called ETS(A,A,N).

The majority of exponential smoothing models can be represented this state space form. Some of them are discussed in this chapter.

1.2 Common datasets for forecasting

There are several popular datasets in forecasting that are widely used in different papers. Here we give a brief description of them.

M3 dataset refers to the data used in M3-Competition (Makridakis and Hibon, 2000). It consists of 3003 time series, including yearly, quarterly, monthly and "other" types of data. These are time series from different areas, including industry, finance, macroeconomic etc.

NN3 dataset is a subset of monthly time series from M3. It includes 111 data with large number of observations.

Australian tourism competition dataset (Athanasopoulos et al., 2011) has 518 yearly, 427 quarterly and 366 monthly time series, which were supplied by tourism agencies and various academics, who had used them in previous tourism forecasting studies

Finally, there is also macroeconomic data by Federal Reserve Bank of St. Louis (FRED), which consists mainly of yearly macroeconomic data.

1.3 Why damped trend method works

There are several papers published since 2006 explaining the good performance of damped trend exponential smoothing proposed by Gardner and McKenzie (1985). The method performed very well in M3-Competition (Makridakis and Hibon, 2000) and several authors reported that it also performed well on other time series. Several researchers have looked into this problem trying to identify the appropriate rationale of such good performance of the method.

Snyder and Koehler (2009) show that simple exponential smoothing with the embedded Triggs' tracking signal is equivalent to a restricted damped-trend exponential smoothing method. The authors claim that this idea helps in understanding

why damped-trend method performs better than other ETS methods on M3 data. They claim that it underlies high number of different methods. A small competition on yearly data of M3 is performed in this research and it is shown that restricted damped-trend method performs even better than simple damped trend.

McKenzie and Gardner (2010) give a detailed explanation of the performance of random coefficients state-space model underlying their method:

$$\begin{cases} y_t = l_{t-1} + A_t b_{t-1} + \epsilon_t \\ l_t = l_{t-1} + A_t b_{t-1} + \alpha\epsilon_t \\ b_t = A_t b_{t-1} + \beta\epsilon_t \end{cases}. \tag{1.5}$$

where A_t is a binary random variable with $P(A_t = 0) = \phi$ and $P(A_t = 0) = 1 - \phi$. The inclusion of random term A_t allows to describe the variety of different processes. For example if A_t is equal to one on every observation t, then the resulting process resembles the local trend process. If A_t is equal to zero, then the local level process is generated.

The authors also derive the equivalent ARIMA model with random coefficients for the proposed model (1.5) and give their explanation of why damped trend performs on average better than the other exponential smoothing methods (SES and Holt's methods). They show that the damped trend method underlies a wider range of processes than other exponential smoothing methods and thus damped trend is more capable of modelling and forecasting time series of different nature (stochastic or deterministic level / trend). This gives damped trend the property of being robust when using one single model for forecasting different time series.

In a following article Gardner and McKenzie (2011) once again examine damped trend method and derive several particular methods from it. They propose an heuristic estimation procedure for the initial values using local and global values (the last appears to give a little bit less accurate forecasts) and use Solver from MS Excel to minimize Mean Squared Error (MSE). The resulting symmetric Mean Absolute Percentage Error (SMAPE) differs a little bit from SMAPE published by Makridakis and

Hibon (2000) but the difference is not statistically significant. Models are estimated on M3 data and authors show that several series follow ETS(A,N,N) with damped drift and random walk with damped drift processes. As a result the damped trend method (being a more general case) performs well on M3 dataset.

In our view one of the reasons of good performance of the damped trend method is the ϕ parameter which reduces the effect of wrong estimation of trend component. This is probably the reason why damped trend performs better than Holt's method. If a different estimation technique is used, then performance of Holt's method would probably improve. It also highlights the usefulness of more flexible models to conventional SES and Holt's methods.

1.4 Forecasts combinations and hybrid models

Over the years combinations of forecasts have been reported to perform, on average, more accurately than individual forecasts (Makridakis et al., 1982; Makridakis and Hibon, 2000). There are some developments on how to best combine forecasts, but only a few papers that use exponential smoothing have appeared in the literature.

The basic method of forecast combination usually involves using either a simple mean or weighted average. The general knowledge of statistics dictates that in cases of outliers it is wiser to use median value instead of mean. Jose and Winkler (2008) show that using winsorizing and trimming instead of median usually gives an increased forecast accuracy. The rationale for their proposition is that the mean is usually influenced by outliers and median while being a robust measure usually leads to the loss of a useful information. The authors use forecasts made by 22 methods from M3 competition in the combination to show that the use of winsorizing and trimming increases forecasting accuracy. The results of combinations of forecasts made by 5, 7 and 9 models are also shown. Authors remark that the increase in forecasting accuracy from 5 to 9 models looks comparable to the increase from 9 to 22 models,

which leads to the conclusion that increase in forecasting accuracy is declining with the increase of number of models used in combination. Their proposed combination methods are also tested on Nominal Gross Domestic Product data with similar results. In conclusion the authors advise using 10% – 13% of trimming and 15% – 45% of winsorizing, giving the preference to winsorizing. These recommendations however do not have a theoretical ground and are based on empirical findings of the authors.

One of the other possible forecast combination techniques involves defining weights for different methods rather than using equal weights. Kolassa (2011) uses information criteria, such as AIC, AICc and BIC to determine weights and combine point and interval forecasts for different exponential smoothing models rather than choosing the best one among them (using similar principles as in Burnham and Anderson (2004)). The author conducts an experiment on M3 data and shows that forecasting accuracy increases in some cases when using the proposed method. One disadvantage of the proposed method is that it can only be applied to models from the same class (either exponential smoothing or autoregressive models) while combination forecasts made using different approaches could contain different information and features of analysed time series.

Another popular approach in forecasting is the construction of hybrid models as alternative to a combination of forecasts. Usually in these cases one of the models that is used in the process is exponential smoothing model.

For example Maia and de Carvalho (2011) propose combining Holt's exponential smoothing modified for interval-valued series with a neural networks method. The resulting method is checked on a stock-market series and is compared with single models. The selection of Holt's method in the article instead of alternatives like damped trend, which has been shown in several papers to work better for trend time series, is not justified. The time series the authors use could also be seen as a multivariate problem and comparison of results with multivariate ETS on a bigger

sample of time series would have been more appropriate, as it could show whether the differences in accuracy of different methods are statistically significant. Furthermore the authors did not include the Naive method in their benchmarks list, which typically is unbeatable on this type of series. Whilst the research was lacking the paper is still useful for our review as it recognises that hybrid models may perform better than single models.

Wang et al. (2012) propose a hybrid model for stock index forecasting. This model consists of SES, ARIMA and a neural network method (NN). Authors propose to select weights using the squared one-step-ahead forecasting error minimization via a genetic algorithm. The proposed hybrid model outperforms the equally weighed hybrid model, random walk, SES, ARIMA and NN used separately on two time series. The comparison of methods included in the article is not enough. Comparing the proposed hybrid model with other hybrids proposed in other papers on a larger dataset is necessary in order to justify the proposed hybrid. For now these results are not statistically significant and the selection of competing methods raises additional questions. For example, why did the authors choose SES instead of some other more robust exponential smoothing method?

Nevertheless both Maia and de Carvalho (2011) and Wang et al. (2012) highlight that ETS models may not capture all information available in a series.

Davydenko and Fildes (2012) propose a method of joint forecasting by a statistical model and expert's judgment using a Bayesian approach. They argue that by combining point forecasts of any statistical model and the standard error of experts more accurate forecasts can be obtained. This can be explained by the idea that using judgemental opinion can in general increase forecasting accuracy if an expert has additional valuable information about the future. The authors benchmark the resulting forecasts with Naive and automatic ARIMA on Australian tourism demand data. Notably state-space exponential smoothing applied to the data produced downwards

trend (based on the several last available observations) and performed the worst of all the models. This happens mainly because of a weights distribution used in exponential smoothing and the property of "short memory", where recent events have higher value than the past. As a result the model is excluded from further comparison, which shows that using judgment one can obtain more accurate forecasts with the proposed joint model.

Findings in the area of forecasting discussed in this section indicate that an underlying statistical model may have a more complicated form. Otherwise combinations of forecasts and hybrid models would not be needed; the only thing a forecaster would need to do would be to select the correct model. Still even if the model has a known form, it may be hard to select it from all the possible models known due to the large number of candidates. Furthermore there are different model selection procedures but none of them guarantees for an individual data set that the correct model will be selected. Finally even if an appropriate model is selected, it needs to be estimated properly in order to cater for limited sample. The example of Davydenko and Fildes (2012) demonstrates this point: if state-space exponential smoothing model was estimated using a different method, it could in theory perform better.

1.5 Temporal aggregation

Temporal aggregation operates by aggregating data using some function with higher frequency in order to obtain slower moving series. After the new time series are produced, the most appropriate of them is selected, the model is fitted on that level, forecasts are produced and after that – somehow reconciled.

Andrawis et al. (2011) propose a new approach for forecast combination. They propose to combine forecasts made using monthly series and then the same series aggregated to yearly data. Though the authors do not give a detailed explanation of why this should be done, they claim that aggregation to yearly data allows one

to capture long-term dependencies while monthly data models capture short-term information. Several methods of weights estimation are studied in the article including simple and geometric means and different types of weighted averages. The M3, NN3 and tourism data (inbound tourism demand for Egypt, 34 time series) are used in the benchmark of the proposed approach. The usage of NN3 by the authors is not clear as it is a subset of M3.

Kourentzes et al. (2014) propose a novel framework for time series forecasting called "Multi Aggregation Prediction Algorithm" (MAPA). Forecast using MAPA is produced in 3 steps: (1) Aggregation of original time series (which is done using different levels). (2) Fitting of ETS to each of the levels and transformation of obtained multiplicative components (if they appear) into additive components. (3) Combination of the obtained components to produce final ETS combined model. The final forecast is done after all 3 steps. M3 data and FRED are used in the experiment. MAPA is compared with Hyndman et al. (2002) ETS and a simple combination of forecasts on different levels (using mean and median values). The results indicate that MAPA is statistically significantly (on probability of 0.05) more accurate in almost all the categories of times series in M3 and FRED.

Kourentzes and Petropoulos (2016) propose a modification of the original MAPA in order to take promotions into account as exogenous variables. They tackle several problems caused by multicollinearity and conduct an experiment on simulated and real data. The authors show that forecasts of MAPAx have smaller bias and higher accuracy than ETS, MAPA, ETS with exogenous variables and regression on the used dataset.

The fact that MAPA and MAPAx perform so well indicates that conventional ETS cannot always capture correctly components in time series. There are at least two possible reasons for that: (1) The underlying model is more complicated than it is thought. MAPA in this case manages to capture non-linear relations by aggregating

data. (2) ETS components cannot be properly captured because of the problems with estimation on finite samples. Due to combination of forecasts MAPA in this case manages to mitigate this and produce better estimates of ETS components.

1.6 High-frequency time series forecasting

Finally, one of the main streams in scientific publications since 2006 is the use the of state-space approach for high-frequency time series forecasting. Taylor's double seasonal method (Taylor, 2003) has been considered as one of the main exponential smoothing based solutions to the problem of high-frequency data forecasting till 2006. The method can be represented by the following system (transforming it in to error-correction form):

$$\begin{cases} \hat{y}_{t+h|t} = (l_{t-1} + h \cdot b_{t-1}) s_{1,t-m_1} s_{2,t-m_2} \\ l_t = l_{t-1} + b_{t-1} + \alpha \frac{\epsilon_t}{s_{1,t-m_1} s_{2,t-m_2}} \\ b_t = b_{t-1} + \beta \frac{\epsilon_t}{s_{1,t-m_1} s_{2,t-m_2}} \\ s_{1,t} = s_{1,t-m_1} + \gamma_1 \frac{\epsilon_t}{l_{t-1}+b_{t-1}} \\ s_{2,t} = s_{2,t-m_2} + \gamma_2 \frac{\epsilon_t}{l_{t-1}+b_{t-1}} \end{cases}, \tag{1.6}$$

where h is forecast horizon, y_t is actual data, l_t is level component, b_t is trend component, $s_{1,t}$ is first seasonal coefficient, $s_{2,t}$ is second seasonal coefficient, m_1 is frequency of the first type of seasonality (for example, monthly), m_2 is frequency of the second type of seasonality (for example, weekly), α, β, γ_1 and γ_2 are smoothing parameters.

The Taylor (2003) method has several disadvantages. Firstly, it is based on Holt-Winters method with multiplicative seasonality and has restrictions concerning the number of seasonal coefficients (the number of coefficients is very high and is hard to estimate) and how often they are updated. Besides, Taylor's double seasonal model does not have an underlying statistical model. That is why the topic of high-frequency time series forecasting became of interest in several articles since.

Gould et al. (2008) propose a model for multiple seasonality forecasting. They address the problem of absence of an appropriate underlying model for time series having multiple seasonalities (for example, daily data can have weekly and monthly

seasonality at the same time) and propose a new approach for high-frequency time series forecasting. Their model, called MS(r, m_1, m_2) ("Multiple Seasonality"), has additive and multiplicative forms and is a generalization of Holt-Winters and Taylor's methods. Here r designates the number of different sub-cycles. For example, sub-cycles for working days can be similar and the same can hold for weekends. This may result in a decrease of number of seasonal coefficients from 7 (5 for the workdays and 2 for weekends) to $r = 2$.

MS with additive seasonality is:

$$\begin{cases} y_t = l_{t-1} + b_{t-1} + \sum_{i=1}^{r} x_{i,t} s_{i,t-m_i} + \epsilon_t \\ l_t = l_{t-1} + b_{t-1} + \alpha\epsilon_t \\ b_t = b_{t-1} + \beta\epsilon_t \\ s_{i,t} = s_{i,t-m_i} + \sum_{j=1}^{r} \gamma_{i,j} x_{j,t} \epsilon_t \end{cases}, \tag{1.7}$$

MS with multiplicative seasonality is:

$$\begin{cases} y_t = (l_{t-1} + b_{t-1}) \left(\sum_{i=1}^{r} x_{i,t} s_{i,t-m_i}\right) (1 + \epsilon_t) \\ l_t = (l_{t-1} + b_{t-1})(1 + \alpha\epsilon_t) \\ b_t = b_{t-1} + \beta(l_{t-1} + b_{t-1})\epsilon_t \\ s_{i,t} = s_{i,t-m_i} \left(1 + \left(\sum_{j=1}^{r} \gamma_{i,j} x_{j,t}\right) \epsilon_t\right) \end{cases}, \tag{1.8}$$

where $x_{i,t}$ is a dummy variable that is equal to 1 when time t occurs during the sub-cycle i and 0 otherwise, $\gamma_{i,j}$ is smoothing parameter.

The authors' model uses dummy variables $x_{i,t}$ for modelling sub-cycles which permits implementation of the idea of joint updates for seasonal coefficients inside sub-cycles. Besides, using the proposed idea the number of seed values can be substantially decreased. One additional advantage of the proposed model is that it can handle missing data without any need for additional variables. This is done via the very same dummy variables $x_{i,t}$.

The authors carry out a comparison of the model with a modification of Taylor's double seasonal and classical Holt-Winters methods on traffic data and hourly vehicle count data. This experiment demonstrates superiority of the proposed model. The authors also show that Taylor's method can be considered a special case of their new multiple seasonal model.

The same year Taylor (2008) publishes an article with a comparison of several methods on high-frequency data. The list of models includes several different ETS, ARIMA for double seasonality, regression model (demand as affected by weather), Taylor's double seasonal method with first order autocorrelation correction (Taylor, 2003) and Gould et al. (2008) model. All of them are tested on electricity demand data. Taylor remarks that simple Holt's method performs poorly and excludes it from further analysis. Different models are also optimised using different forecasting horizons (for example, including 30 observations ahead). As a result of this small competition, Taylors double seasonal model optimised for 30 minutes ahead gave the most accurate forecasts for short term, while the regression model gave the most accurate forecasts for long term.

In Taylor (2010) British and French electricity load series are studied, triple seasonal ARMA and triple seasonal Holt-Winters models with a first order autocorrelation correction (triple seasonal HWT) are proposed. These models become the most accurate during an experiment in the article amongst triple, double and single seasonal models for the horizon of 1 hour up to 1 day. The author shows that a combination of triple seasonal ARMA and HWT gives even more accurate forecasts. It should be remarked once again that all the research is done using only 2 time series and all the models are compared using MAPE.

De Livera (2010) proposes a new generalization for high-frequency data exponential smoothing models which is called "BATS($\omega, \phi, p, q, m_1, m_2, ..., m_T$) (B stands for "Box-Cox transformation", A – ARMA residuals, T – trend component, S – seasonal component), which implies the prior Box-Cox transformation of original data. Here ω is Box-Cox transformation parameter, ϕ is damping parameter, p and q are ARMA

orders and $m_1, ..., m_T$ are seasonal periods. The general form of the BATS model is:

$$\begin{cases} y_t^{(\omega)} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{M} s_{i,t-m_i} + \epsilon_t \\ l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t \\ b_t = \phi b_{t-1} + \beta d_t \\ d_t = \sum_{j=1}^{p} \varphi_j d_{t-j} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \end{cases}, \tag{1.9}$$

where d_t is an ARMA(p,q) process, φ_j and θ_j are parameters of AR(p) and MA(q) parts respectively. All the other parameters in this model correspond to the ones discussed above.

As we see the main difference between BATS and other high-frequency seasonal models is in usage of ARMA and Box-Cox transformation. This allows it to capture additional information in data and as a result the proposed BATS model outperformes Taylor's double-seasonal exponential smoothing on example of a time series. The paper does not provide an extensive empirical evaluation and it is hard to gauge the accuracy of BATS compared to other high-frequency data methods.

De Livera et al. (2011) show an underlying statistical model for BATS and propose a new model based on Fourier series. The authors show that Gould et al. (2008) and Taylor's double and triple seasonal models can be viewed as special cases of BATS. For example, Taylor's double seasonal model with an autoregressive term can be written as BATS$(1, 1, 1, 0, m_1, m_2)$. TBATS model (T stands for "trigonometric") is based on Fourier transformation which is the main difference from BATS. It uses following system of equations instead of $s_{i,t-m_i}$ in (1.9):

$$\begin{cases} s_{i,t} = \sum_{j=1}^{k_i} s_{i,j,t} \\ s_{i,j,t} = \left(s_{i,j,t-1} \cos \lambda_{i,j} + s^*_{i,j,t-1} \sin \lambda_{i,j}\right) + \gamma_{i,1} d_t \\ s^*_{i,j,t} = \left(s^*_{i,j,t-1} \cos \lambda_{i,j} - s_{i,j,t-1} \sin \lambda_{i,j}\right) + \gamma_{i,2} d_t \end{cases}, \tag{1.10}$$

where $\lambda_{i,j} = \frac{2\pi j}{m_i}$, $\gamma_{i,1}$ and $\gamma_{i,2}$ are smoothing parameters, $s_{i,j,t-1}$ is the level of the i-th seasonal component, $s^*_{i,j,t-1}$ is the trend of the i-th seasonal component, k_i is the number of harmonics required for i-th seasonal component.

Using the Fourier transformation allows a decrease in the number of parameters needed for initialization, while usage of $\gamma_{i,1}$ and $\gamma_{i,2}$ in (1.10) still allows to use the

principle of exponential smoothing in the model that results in update of seasonal components over time. Due to (1.10) the TBATS model allows us to deal with non-integer seasonality and has fewer initial values to estimate compared with BATS.

Both BATS and TBATS allow modelling complicated time series with several seasonal patterns (for example daily, weekly and monthly). State-space forms of BATS and TBATS, equivalent ARIMA and likelihood function are derived in De Livera et al. (2011). It is shown that TBATS can also be used for complex seasonal time series decomposition. TBATS is then compared with BATS on a time series, but no real competition between models is conducted.

Continuing the topic of high-frequency time series forecasting Taylor and Snyder (2012) compare Taylor's double seasonal model, double seasonal ARIMA and Gould et al. (2008) model. They show that Taylor's model while being more complex in number of parameters than Gould's is still more accurate because it assumes that parts of days can differ in contrast with Gould's model assuming that only whole days should be treated as identical. The comparison of methods accuracy is done only on three time series.

Summarizing this section, two conclusions can be reached:

1. Models for high-frequency data open a question of correct estimation of parameters. Several models have been proposed, but all of them use short-term information, being optimised using the conventional mean squared one step ahead forecast error. Only Taylor (2008) notes that forecasting accuracy increases when models are estimated using multiple steps ahead estimators, but no one else in the field of high-frequency data uses this finding.

2. Gould's and Taylor's models allow modelling of both additive and multiplicative seasonality. However selection procedures between additive and multiplicative multi-seasonal models are not established, so it is not obvious how to select an appropriate model in real data using these models.

Therefore the estimation of such models, particularly given the number of parameters and identifying and capturing the nature of the multiple seasonal cycles, remain challenging. These should be areas of further research.

1.7 Conclusions

Concluding the review, there are several open questions that motivate our research. First, the true model (if one exists) may have a much more complicated structure than it is though to have. The papers in forecast combinations, hybrid models and temporal aggregation support this hypothesis. Proposing a more flexible non-linear model could address this problem. However such a model would be desirable to retain the simple interpretation and ease of use of conventional exponential smoothing.

Second, even if we assume that a true model is known, selecting it from the pool of all the known models is not a trivial task. Combining several models allows us to partially overcome this limitation but does not completely solve the problem. Introducing a model that could side step the selection procedure could be an alternative solution to the problem. The good performance of damped trend indicates that this may be a fruitful avenue to explore.

Third, high-frequency data demonstrates that seasonality in high-frequency can be both additive and multiplicative but no mechanism for model selection has been proposed. Proposing a seasonal model that would be able to model both types without a need to switch between them could be another possible direction of research. Furthermore on high frequency data, for example in the case of solar irradiance the observed seasonality may be neither.

Finally, high-frequency data examples and damped trend method papers point out that the existing models may face a problem from a different angle: estimation rather than model selection. Using multiple steps ahead estimators seems to help solve this problem. This question will be investigated further.

This work will attempt to address these questions. As discussed above some of the open questions can be approached in multiple alternative ways. For example, we will explore the problem of model selection both from the point of view of deriving more flexible models, but also as part of parameter estimation. This way we attempt to explore the promising directions of the research we identified in this review.

Finally it is important to underline that exponential smoothing has the advantage of being relatively easy to understand and use in practice. It is desirable that further improvements in the field should retain this property, as this is crucial to allow innovations to transfer to practice. This will be central to our motivation of how we research new models and techniques in the area of exponential smoothing.

Chapter 2

Theory of Complex Exponential Smoothing

In this chapter we discuss the idea of "information potential", propose a new modelling approach in time series and a new model using this approach. All the materials in this chapter are based on an article submitted to European Journal of Operational Research (currently under review).

2.1 Introduction

Exponential smoothing is a very successful group of forecasting methods which is widely used both in theoretical research (for examples see Jose and Winkler, 2008; Kolassa, 2011; Maia and de Carvalho, 2011; Wang et al., 2012; Athanasopoulos and de Silva, 2012; Kourentzes et al., 2014) and in practice (see different forecasting comparisons by Fildes et al., 1998; Makridakis and Hibon, 2000; Gardner and Diaz-Saiz, 2008; Athanasopoulos et al., 2011).

The exponential smoothing methods are well known and popular amongst practising forecasters for more than half a century originating from the work by Brown (1956). Hyndman et al. (2002), based on work by Snyder (1985) and Ord et al. (1997), embedded exponential smoothing within a state space framework, providing its statistical rationale, resulting in ETS (short for ExponenTial Smoothing). ETS provides

a systematic framework to estimate parameters, construct prediction intervals and choose between different types of exponential smoothing.

While ETS is widely used, recent work has demonstrated that it is possible to improve upon. Research shows that various combinations of exponential smoothing models result in composite ETS forms beyond the 30 standard forms (Hyndman et al., 2008b) which leads to improvement in forecasting accuracy (Kolassa, 2011; Kourentzes et al., 2014). We argue that models in the existing taxonomy do not cover all the possible forms. A possible complicating factor for ETS is the assumption that any time series can be decomposed into several distinct components, namely level, trend and seasonality. In particular separating the level and trend of the time series presents several difficulties. The motivation in this paper is to try to avoid this artificial distinction. For this reason we focus our research on the level and trend aspects, and do not examine seasonality that can usually be easily distinguished.

We propose a new approach to time series modelling that eliminates the arbitrary distinction between level and trend components, and effectively sidesteps the model selection procedure. To do this we introduce the concept of "information potential", an unobserved time series component. We argue that this information exists in any time series and including it in a model results in superior forecasting accuracy. We encode the observed values of a time series together with the information potential as complex variables, giving rise to the proposed Complex Exponential Smoothing (CES). We demonstrate that CES has several desirable properties in terms of modelling flexibility over conventional exponential smoothing and demonstrate its superior forecasting accuracy empirically.

The rest of the article is structured as follows. The conventional and the new approaches to time series modelling are briefly discussed in section 2.2, leading to the introduction of the Complex Exponential Smoothing (CES) in section 2.3. Within that section properties of CES and connections with existing models are discussed.

Section 2.4 proposes a proxy for the information potential and discusses its implications for CES. Finally real life examples are shown and the comparison of CES performance with competing statistical models is carried out in section 2.5, followed by concluding remarks.

2.2 Conventional and complex-valued approaches

Hyndman et al. (2008b) systematised all existing exponential smoothing methods and showed that any ETS model is based on one out of five types of trends (none, additive, damped additive, multiplicative and damped multiplicative), one out of two types of errors (additive and multiplicative) and one out of three types of seasonal components (none, additive and multiplicative). This taxonomy leads to 30 exponential smoothing models that underlie different types of time series. Model parameters are optimised using maximum likelihood estimation and analytical variance expressions are available for most exponential smoothing forms. The authors proposed to use information criteria for model selection, which allowed choosing the most appropriate exponential smoothing model in each case. Hyndman et al. (2008b) argued that AIC is the most appropriate information criterion. Billah et al. (2006) demonstrated that there was no significant difference in the forecasting accuracy when using different information criteria.

Focusing on the non-seasonal cases the general ETS data generating process has the following form:

$$y_t = f(l_{t-1}, b_{t-1}, \epsilon_t), \tag{2.1}$$

where y_t is the value of the series, l_t is the level component, b_t is the trend component, ϵ_t is the error term and $f(C(\cdot))$ is some function that allows including these components in either additive or multiplicative form. However the composite forecasts discussed in Kolassa (2011) and Kourentzes et al. (2014) hint to the lack of a clear separation between level and trend components. This decomposition in (2.1)

can be considered arbitrary: depending on the chosen ETS model, its initial values and smoothing parameters, different estimates of the time series components can be obtained. However, these components are unobservable and their identification relies on an appropriate decomposition. For example it is often hard to distinguish a local-level time series from a trend series. Consider simple exponential smoothing (SES):

$$\hat{y}_t = \alpha y_{t-1} + (1-\alpha)\hat{y}_{t-1}, \tag{2.2}$$

where $\hat{y}_t$ is the estimated value of the series and α is the smoothing parameter which should lie inside the region (0, 2) (Brenner et al., 1968). Following the notation by Hyndman et al. (2002) SES has an underlying ETS(A,N,N) model:

$$\begin{cases} y_t = l_{t-1} + \epsilon_t \\ l_t = l_{t-1} + \alpha\epsilon_t \end{cases}. \tag{2.3}$$

When the smoothing parameter (α) in (2.3) increases, changes in level become so rapid that the generated time series can reveal features of trends. When the smoothing parameter becomes greater than one, the series can reveal an even clearer global trend (Hyndman et al., 2008b, p.42). Selecting the correct model in this situation becomes a challenging task. This is just one example where it is hard to identify the correct model with the standard exponential smoothing decomposition approach.

Instead of decomposing time series we propose to study two characteristics: i) the observed value of the series y_t; and ii) its "information potential" p_t. We introduce the information potential as a non-observable component of the time series that characterises it and influences the observed values y_t. We propose that although the actual value y_t and its prediction are the variables of main interest in modelling, the unobserved information potential p_t contains additional useful information about the observed series. Thus it is necessary to include it in models. In this paper we show that the information potential increases the flexibility of models allowing to capture

a wider range of behaviours, it also removes the need for the arbitrary distinction between level and trend and results in increased forecasting accuracy.

We encode these two real variables into a single complex variable $y_t + ip_t$ that permits both of them to be taken into account during the modelling process (Svetunkov, 2012). Here i is the imaginary unit which satisfies the equation: $i^2 = -1$. Thus the general data generating process using this notation has the form:

$$y_t + ip_t = f(Q, \epsilon_t), \tag{2.4}$$

where Q is a set of some complex variables, chosen for the prediction of $y_t + ip_t$. Using this idea, we propose the Complex Exponential Smoothing (CES) based on the data generating process (2.4), in analogy to the conventional exponential smoothing model. This is introduced in the detail in the following section.

2.3 Complex Exponential Smoothing

2.3.1 Method and model

Combining the simple exponential smoothing method with the idea of information potential, substituting the real variables in (2.2) by complex variables, gives us the complex exponential smoothing method:

$$\hat{y}_{t+1} + i\hat{p}_{t+1} = (\alpha_0 + i\alpha_1)(y_t + ip_t) + (1 - \alpha_0 + i - i\alpha_1)(\hat{y}_t + i\hat{p}_t), \tag{2.5}$$

where $\hat{y}_t$ is the estimated value of series, $\hat{p}_t$ is the estimated value of information potential and $\alpha_0 + i\alpha_1$ is complex smoothing parameter. Representing the complex-valued function as a system of two real-valued functions leads to:

$$\begin{cases} \hat{y}_{t+1} = (\alpha_0 y_t + (1 - \alpha_0)\hat{y}_t) - (\alpha_1 p_t + (1 - \alpha_1)\hat{p}_t) \\ \hat{p}_{t+1} = (\alpha_1 y_t + (1 - \alpha_1)\hat{y}_t) + (\alpha_0 p_t + (1 - \alpha_0)\hat{p}_t) \end{cases}. \tag{2.6}$$

It can be seen from (2.6) that the final CES forecast consists of two parts: one is produced by SES, while the second only employs the SES mechanism. It is obvious

that CES is a non-linear method in its nature as both first and second equations in (2.6) are connected with each other and change simultaneously depending on the complex smoothing parameter value.

We derive the underlying statistical model for CES to study its properties. The model can be written in the following state-space form (see Appendix A for the derivation):

$$\begin{cases} y_t = l_{t-1} + \epsilon_t \\ l_t = l_{t-1} - (1-\alpha_1)c_{t-1} - \alpha_1 p_t + \alpha_0 \epsilon_t \\ c_t = l_{t-1} + (1-\alpha_0)c_{t-1} + \alpha_0 p_t + \alpha_1 \epsilon_t \end{cases}, \tag{2.7}$$

where l_t is the level component, c_t is the information component at observation t and ϵ_t is an error term. Observe that dependencies in time series have a non-linear structure and no explicit trend component is present in the time series as this model does not need to artificially break the series into level and trend, as ETS does. This idea still allows to rewrite (2.7) in a shorter more generic way, resembling the general SSOE state-space framework:

$$\begin{cases} y_t = w' x_{t-1} + \epsilon_t \\ x_t = F x_{t-1} + q p_t + g \epsilon_t \end{cases}, \tag{2.8}$$

where $x_t = \begin{pmatrix} l_t \\ c_t \end{pmatrix}$ is the state vector, $F = \begin{pmatrix} 1 & -(1-\alpha_1) \\ 1 & 1-\alpha_0 \end{pmatrix}$ is the transition matrix, $g = \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}$ is the persistence vector, $q = \begin{pmatrix} -\alpha_1 \\ \alpha_0 \end{pmatrix}$ is the information potential persistence vector and $w = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is the measurement vector.

The state-space form (2.8) permits extending the CES in a similar ways to ETS to include additional states for seasonality or exogenous variables. The main difference between model (2.8) and the conventional ETS is that the former includes the information potential term. Furthermore the transition matrix in (2.8) includes smoothing parameters which is not a standard feature for ETS models. This form uses the generic SSOE state-space exponential smoothing notation and allows a clear understanding of the model.

2.3.2 Weights distribution in CES

Being an exponential smoothing method CES is capable of distributing weight between the observations in a different manner. In this paragraph we will show how the complex smoothing parameter influences this distribution.

Substituting the estimated values in the right side of equation (2.5) repeatedly the following recursive form of CES can be obtained:

$$\hat{y}_{t+1} + i\hat{p}_{t+1} = (\alpha_0 + i\alpha_1) \sum_{j=0}^{t} (1 - \alpha_0 + i - i\alpha_1)^j (y_{t-j} + ip_{t-j}), \tag{2.9}$$

where $t \to \infty$.

To obtain the bounds of the complex smoothing parameter the weights in (2.9) can be perceived as a geometric progression with a complex ratio $(1 - \alpha_0 + i - i\alpha_1)$; hence this series will converge to some complex number only if the absolute value of the ratio is less than one:

$$v = \sqrt{(1 - \alpha_0)^2 + (1 - \alpha_1)^2} < 1. \tag{2.10}$$

The condition (2.10) is crucial for the preservation of the stability condition (Hyndman et al., 2008a): the older observations should have smaller weights compared to the newer observations. For CES this is represented graphically on the plane as a circle with a centre at coordinates (1, 1) as in Figure 2.1. This parameter space is not usual for exponential smoothing methods.

Using (2.9) CES can also be represented in the trigonometric form of complex variables, which should help in understanding the mechanism used in the method:

$$\hat{y}_{t+1} + i\hat{p}_{t+1} = R \sum_{j=0}^{t} \left[v^j (\cos(\varphi + j\gamma) + i \sin(\varphi + j\gamma))(y_{t-j} + ip_{t-j}) \right], \tag{2.11}$$

where $R = \sqrt{\alpha_0^2 + \alpha_1^2}$, $\varphi = \arctan \frac{\alpha_1}{\alpha_0} + 2\pi k$ and $\gamma = \arctan \frac{1-\alpha_1}{1-\alpha_0} + 2\pi k$, $k \in Z$. Hereafter we assume that $k = 0$ in the calculation of all the polar angles as all other angles do not add to model understanding. This is because angles γ and $\gamma + 2\pi$

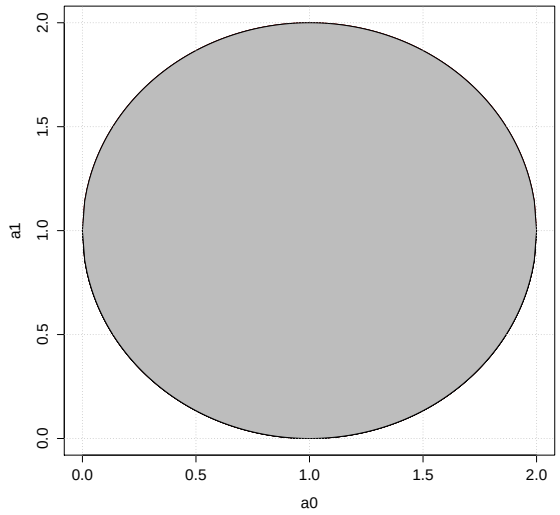


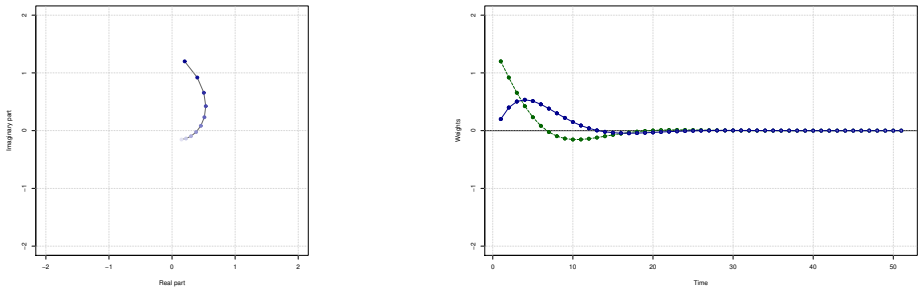
Figure 2.1: Bounds of complex smoothing parameter.

lead to exactly the same values of cosine and sine functions, but the latter does not contribute in interpretation of the complex variable.

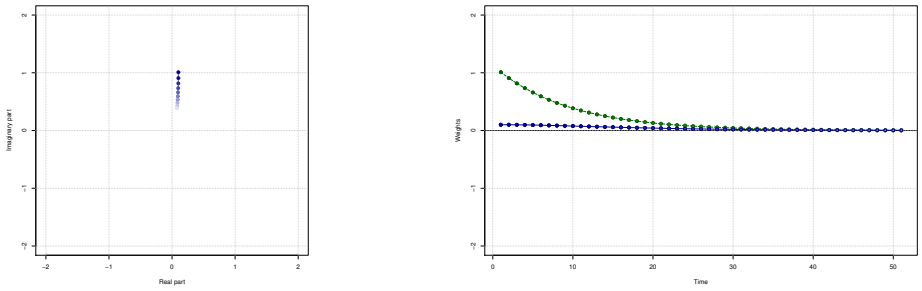
Multiplying complex variables in (2.11) shows how the real and imaginary parts of the forecast are formed separately, which can be presented in the following system:

$$\begin{cases} \hat{y}_{t+1} = R\sum_{j=0}^{T} v^j \cos(\varphi + j\gamma) y_{t-j} - R\sum_{j=0}^{T} v^j \sin(\varphi + j\gamma) p_{t-j} \\ \hat{p}_{t+1} = R\sum_{j=0}^{T} v^j \sin(\varphi + j\gamma) y_{t-j} + R\sum_{j=0}^{T} v^j \cos(\varphi + j\gamma) p_{t-j} \end{cases}. \tag{2.12}$$

Here the CES forecast depends on the previous values of the series and the information potential, which are weighed in time using trigonometric functions, depending on the value of the complex smoothing parameter. For example, when $\alpha_0 + i\alpha_1 = 0.2 + 1.2i$ weights are distributed as shown on Figures 2.2a and 2.2b: they converge to zero slowly, demonstrating a hump for 4th observation. Their absolute value decreases in time due to (2.10). This weights distribution results in forecasts based on the long term time series characteristics, taking some observation into account with higher weight. This example demonstrates the flexibility in weights distribution in CES. A more conventional case is shown in Figures 2.2c and 2.2d. The smoothing parameter is equal to $0.1 + 1.01i$, which results in a slow exponential decline of the weights in time. Plots 2.2a and 2.2b show the decrease of weights on



(a) on complex plane (b) in time



(c) on complex plane (d) in time

Figure 2.2: CES weights distribution. 2.2b and 2.2d: blue solid lines – real parts of complex weights, green dashed lines – the imaginary parts of complex weights.

complex plane. They demonstrate what is happening with the equivalent complex smoothing parameter when it is exponentiated in the power j with $j = 1, ..., T$: the original vector is rotated on the complex plane, the amount of rotation depends on value of complex smoothing parameter.

Plots on the Figure 2.2 also demonstrate that due to the complex nature of the weights their sum will usually be a complex number that can be calculated using:

$$S = (\alpha_0 + i\alpha_1) \sum_{j=0}^{\infty} (1 - \alpha_0 + i - i\alpha_1)^j = \frac{\alpha_0^2 - \alpha_1 + \alpha_1^2 + i\alpha_0}{\alpha_0^2 + (1 - \alpha_1)^2}. \tag{2.13}$$

The sum of weights depends solely on the value of the complex smoothing parameter. It can become real number only when $\alpha_0 = 0$ which contradicts condition (2.10). It can also be concluded that the real part of S can be any real number while its imaginary part is restricted to positive real numbers only. S indicates that CES is not an averaging model, diverging from the conventional exponential smoothing

interpretation. In a way it has more common features with ARIMA, rather than exponential smoothing, because parameters of the former are not restricted with bounds, guaranteeing it property of "averaging" model, while exponential smoothing is usually restricted this way.

2.3.3 Connection with other forecasting models

Underlying ARIMAX model

The majority of exponential smoothing models have equivalent underlying ARIMA models. For example, ETS(A,N,N) has underlying ARIMA(0,1,1) model (Gardner, 1985). It can be shown that CES has an underlying ARIMAX(2,0,2) model (see Appendix B for the derivations):

$$\begin{cases} (1 - \phi_1 B - \phi_2 B^2) y_t = (1 - \theta_{1,1} B - \theta_{1,2} B^2) \epsilon_t + (-\gamma_1 B - \gamma_2 B^2) p_t \\ (1 - \phi_1 B - \phi_2 B^2) \xi_t = (1 - \theta_{2,1} B - \theta_{2,2} B^2) p_t + (-\gamma_1 B - \gamma_2 B^2) \epsilon_t, \end{cases} \tag{2.14}$$

where $\xi_t = p_t - c_{t-1}$ is an information gap, the value showing the amount of the information missing in the information component c_t, while the parameters of ARIMAX are: $\phi_1 = 2 - \alpha_0$, $\phi_2 = \alpha_0 + \alpha_1 - 2$, $\theta_{1,1} = 2 - 2\alpha_0$, $\theta_{1,2} = 2\alpha_0 + 2\alpha_1 - 2 - \alpha_0^2 - \alpha_1^2$, $\theta_{2,1} = -2$, $\theta_{2,2} = 2$, $\gamma_1 = \alpha_1$, $\gamma_2 = \alpha_0 - \alpha_1$.

The first equation in (2.14) demonstrates how the actual data is generated, while the second equation shows how the unobservable part of the data is formed. The information potential variable p_t in the right hand side of the first equation of the model (2.14) functions as an external variable that contributes to the model and refines the AR and MA terms. The error term ϵ_t in the second equation in (2.14) has the same role and coefficients as the information potential in the first equation. Furthermore the AR term has the same coefficients in both first and second equations in (2.14).

Comparison with single exponential smoothing

When $p_t = 0$ and $\alpha_1 = 1$ the system (2.6) becomes:

$$\begin{cases} \hat{y}_{t+1} = \alpha_0 y_t + (1-\alpha_0)\hat{y}_t \\ \hat{p}_{t+1} = y_t + (1-\alpha_0)\hat{p}_t \end{cases}. \tag{2.15}$$

It is obvious that the first equation in (2.15) is the SES method and that the two equations in (2.15) become independent from each other under these conditions. The second equation is not needed for the purpose of forecasting.

The interpretation of $p_t = 0$ is that no additional information is used in the time series generation and the classical exponential smoothing time series decomposition view is adequate. This results from (2.7) in the following CES model (see Appendix C):

$$\begin{cases} y_t = l_{t-1} + \epsilon_t \\ l_t = l_{t-1} + \alpha_0 \epsilon_t \\ c_t = \frac{l_{t-1}}{\alpha_0} + \frac{\epsilon_t}{\alpha_0} \end{cases}. \tag{2.16}$$

We see in (2.16) that the level component becomes the same as in ETS(A,N,N). The only difference of CES is the presence of non-observable information component that does not interact with y_t.

In this case the weights in CES are distributed similarly to single exponential smoothing and converge to the complex number (2.13) which becomes equal to:

$$S = \frac{\alpha_0^2 + i\alpha_0}{\alpha_0^2} = 1 + i\frac{1}{\alpha_0}. \tag{2.17}$$

The real part of S is equal to 1, as in SES. The stability region of CES becomes equivalent to the stability region of SES, a $(0, 2)$ region for the smoothing parameter:

$$\begin{aligned} v &= \sqrt{(1-\alpha_0)^2 + (1-\alpha_1)^2} < 1, \\ v &= |1-\alpha_0| < 1, \text{ with } \alpha_1 = 0. \end{aligned} \tag{2.18}$$

From the above we see that CES encompasses SES and that it is the information potential part that gives CES its different properties.

2.4 Information potential proxy

2.4.1 State-space form

Due to the unobservability of the information potential it needs to be approximated by some characteristics of the studied time series. A proxy that leads to several useful properties is:

$$p_t = y_t - \hat{y}_t = \epsilon_t. \tag{2.19}$$

The logic behind this proxy is that ϵ_t can be seen as a gauge of the information not included in a model.

There can be other information potential proxies, that could be used instead. For example, differences of the data can also contain useful information about the original time series.

The proxy (2.19) leads to a different state-space model, based on (2.7), underlying CES:

$$\begin{cases} y_t = l_{t-1} + \epsilon_t \\ l_t = l_{t-1} - (1-\alpha_1)c_{t-1} + (\alpha_0 - \alpha_1)\epsilon_t \\ c_t = l_{t-1} + (1-\alpha_0)c_{t-1} + (\alpha_0 + \alpha_1)\epsilon_t \end{cases}. \tag{2.20}$$

Now the persistence vector $g = \begin{pmatrix} \alpha_0 - \alpha_1 \\ \alpha_0 + \alpha_1 \end{pmatrix}$, while all the other vectors and matrices values from (2.7) can be retained. The main difference between the general state-space CES (2.7) and (2.20) is that the smoothing parameter changes in both equations of the transition equation, leading to additional non-linearity in the model. The conditional variance of the model can be estimated using the new persistence vector and the transition matrix can be used to estimate the conditional mean.

2.4.2 Likelihood function

The error term in (2.20) is additive. As a result the likelihood function for CES is trivial and is similar to the likelihood function of additive exponential smoothing

models (Hyndman et al., 2008b, p.68):

$$L(g, x_0, \sigma^2|y) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^T \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left(\frac{\epsilon_t}{\sigma}\right)^2\right). \tag{2.21}$$

It was shown by Hyndman et al. (2008b) that when the error term is distributed normally, maximizing the likelihood function to estimate the parameters of CES is equivalent to minimizing the sum of squared errors: $SSE = \sum_{t=1}^{t} \epsilon_t^2$.

2.4.3 Stationarity and stability conditions for CES

The stationarity condition for general exponential smoothing in the state-space form (2.8) holds when all the eigenvalues of F lie inside the unit circle (Hyndman et al., 2008b, p.38). CES can be both stationary and not, depending on the complex smoothing parameter value, in contrast to ETS models that are always nonstationary. Calculating eigenvalues of F for CES gives the following roots:

$$\lambda = \frac{2 - \alpha_0 \pm \sqrt{\alpha_0^2 + 4\alpha_1 - 4}}{2}. \tag{2.22}$$

If the absolute values of both roots are less than 1 then the estimated CES is stationary.

When $\alpha_1 > 1$ one of the eigenvalues will always be greater than one. In this case both eigenvalues will be real numbers and CES produces a non-stationary trajectory. When $\alpha_1 = 1$ CES becomes equivalent to ETS(A,N,N). Finally, the model becomes stationary when (see Appendix D):

$$\begin{cases} \alpha_1 < 5 - 2\alpha_0 \\ \alpha_1 < 1 \\ \alpha_1 > 1 - \alpha_0 \end{cases} \tag{2.23}$$

The other important property that arises from (2.20) is the stability condition for CES. We first use $\epsilon_t = y_t - l_{t-1}$ in the transition equation in (2.20). After manipulations, the following is obtained:

$$\begin{cases} y_t = l_{t-1} + \epsilon_t \\ \begin{pmatrix} l_t \\ c_t \end{pmatrix} = \begin{pmatrix} 1 - \alpha_0 + \alpha_1 & -(1 - \alpha_1) \\ 1 - \alpha_0 - \alpha_1 & 1 - \alpha_0 \end{pmatrix} \begin{pmatrix} l_{t-1} \\ c_{t-1} \end{pmatrix} + \begin{pmatrix} \alpha_0 - \alpha_1 \\ \alpha_1 + \alpha_0 \end{pmatrix} y_t \end{cases}. \tag{2.24}$$

The matrix $D = \begin{pmatrix} 1-\alpha_0+\alpha_1 & -(1-\alpha_1) \\ 1-\alpha_0-\alpha_1 & 1-\alpha_0 \end{pmatrix}$ is called the discount matrix and can be written in the general form:

$$D = F - gw'. \tag{2.25}$$

The model is said to be stable if all the eigenvalues of (2.25) lie inside the unit circle. The eigenvalues are given by the following formula:

$$\lambda = \frac{2 - 2\alpha_0 + \alpha_1 \pm \sqrt{8\alpha_1 + 4\alpha_0 - 4\alpha_0\alpha_1 - 4 - 3\alpha_1^2}}{2}. \tag{2.26}$$

CES will be stable when the following system of inequalities is satisfied (see Appendix E):

$$\begin{cases} (\alpha_0 - 2.5)^2 + \alpha_1^2 > 1.25 \\ (\alpha_0 - 0.5)^2 + (\alpha_1 - 1)^2 > 0.25 \\ (\alpha_0 - 1.5)^2 + (\alpha_1 - 0.5)^2 < 1.5 \end{cases}. \tag{2.27}$$

Both the stationarity and stability regions are shown in Figure 2.3. The stationarity region (2.23) corresponds to the triangle. All the combinations of smoothing parameters lying below the curve in the triangle will produce the stationary harmonic trajectories, while the rest lead to the exponential trajectories. The stability condition (2.27) corresponds to the dark region, which has an unusual form. The stability region intersects the stationarity region, but in general stable CES can produce both stationary and non-stationary forecasts.

2.4.4 Conditional mean and variance of CES

The conditional mean of CES for h steps ahead with known l_t and c_t can be calculated using the state-space (2.20):

$$E\left(y_{t+h}|x_t\right) = w'F^{h-1}x_t, \tag{2.28}$$

where F is the matrix from (2.8). The conditional mean estimated using (2.28) consists of two parts: the conditional mean of the actual values and the conditional mean of the information potential. The former is the main interest in forecasting.

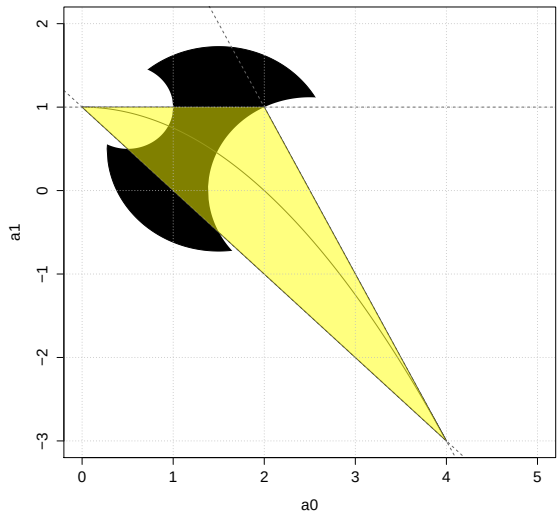


Figure 2.3: Stability and stationarity regions of CES, derived from the state-space form (2.20).

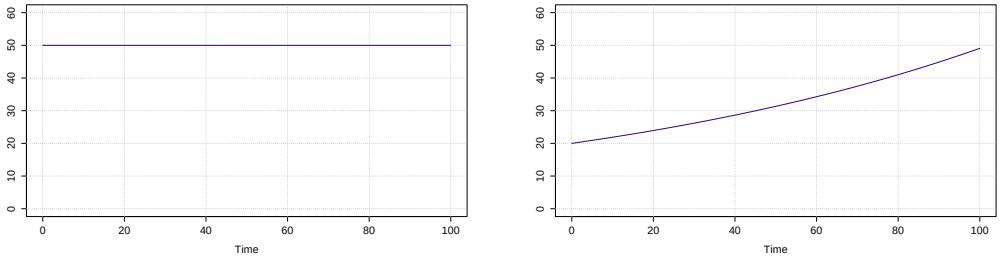
The forecasting trajectories (2.28) will differ depending on the values of l_1, c_1 and the complex smoothing parameter. The analysis of stationarity conditions shows that there are several types of forecasting trajectories of CES depending on the particular value of the complex smoothing parameter:

1. When $\alpha_1 = 1$ all the values of the forecast will be equal to the last obtained forecast, which corresponds to a flat line. This trajectory is shown on the Figure 2.4a.

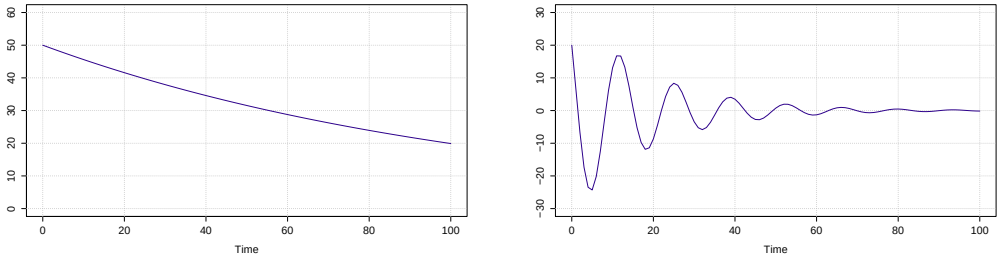
2. When $\alpha_1 > 1$ the model produces the trajectory with the exponential growth which is shown on Figure 2.4b.

3. When $\frac{4-\alpha_0^2}{4} < \alpha_1 < 1$ the trajectory becomes stationary and CES produces the exponential decline shown on Figure 2.4c.

4. When $1 - \alpha_0 < \alpha_1 < \frac{4-\alpha_0^2}{4}$ the trajectory becomes harmonic and will converge to zero (see Figure 2.4d).



(a) the flat line (b) exponential growth



(c) exponential decline (d) harmonic oscilation

Figure 2.4: Forecasting trajectories.

5. Finally, when $0 < \alpha_1 < 1 - \alpha_0$ the diverging harmonic trajectory is produced, the model becomes non-stationary. This forecasting trajectory is of no use in forecasting, that is why we do not show it on graphs.

Using (2.20) the conditional variance of CES for h steps ahead with known l_t and c_t can be calculated (Hyndman et al., 2008b, p.96):

$$V\left(y_{t+h}|x_t\right) = \begin{cases} \sigma_\epsilon^2 \left(J_2 + \sum_{j=1}^{h} (w' F^{j-1} g)^2\right) & \text{when } h > 1 \\ \sigma_\epsilon^2 & \text{when } h = 1 \end{cases} \tag{2.29}$$

This conditional variance formula corresponds in fact to the conditional variance of additive ETS models.

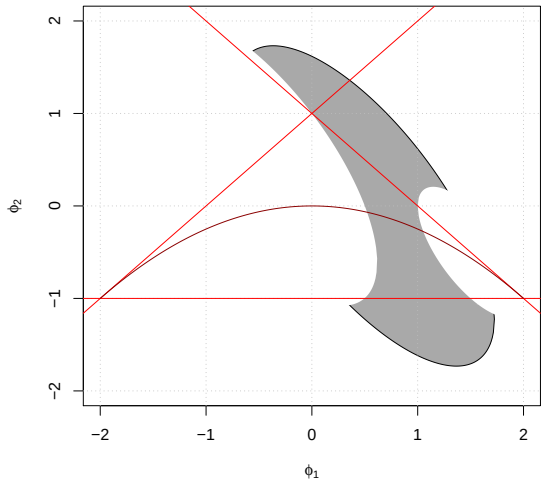


Figure 2.5: Invertibility region of CES on the plane of AR coefficients.

2.4.5 Connection with other forecasting models

Underlying ARMA model

Given the proxy (2.19) we can explore the connections with the other forecasting models further on. Model (2.14) transforms into ARMA(2,2) model:

$$\begin{cases} (1-\phi_1 B-\phi_2 B^2)y_t = (1-\theta_{1,1}B-\theta_{1,2}B^2)\epsilon_t \\ (1-\phi_1 B-\phi_2 B^2)\xi_t = (1-\theta_{2,1}B-\theta_{2,2}B^2)\epsilon_t \end{cases}, \tag{2.30}$$

where $\phi_1 = 2-\alpha_0$, $\phi_2 = \alpha_0+\alpha_1-2$, $\theta_{1,1} = 2-2\alpha_0+\alpha_1$, $\theta_{1,2} = 3\alpha_0+\alpha_1-2-\alpha_0^2-\alpha_1^2$, $\theta_{2,1} = 2+\alpha_1$ and $\theta_{2,2} = \alpha_0-\alpha_1-2$.

The coefficients of AR terms of this model are connected with the coefficients of MA terms, via the complex smoothing parameters. This connection is non-linear but it imposes restrictions on the AR terms plane. Figure 2.5 demonstrates how the invertibility region restricts the AR coefficients field. The triangle on the plane corresponds to the stationarity condition of AR(2) models, while the dark area demonstrates the invertibility region of CES.

Note that exponential smoothing models are nonstationary, but it is crucial for them to be stable or at least forecastable (Ord et al., 1997). In general the stability

condition of ETS corresponds to the invertibility condition for ARIMA. That is why the later should be preferred to the former in CES framework. This means that CES will produce both stationary and non-stationary trajectories, depending on the complex smoothing parameter value. This selection between different types of processes happens naturally in the model without the need of model selection procedure. As an example, the similar stability condition on the same plane for ETS(A,N,N) corresponds to the dot with the coordinates (1,0) while for ETS(A,Ad,N) it corresponds to the line $\phi_2 = 1 - \phi_1$ restricted by the segment $\phi_1 \in (1, 2)$.

Further observations on single exponential smoothing

Given (2.19) then several additional properties of CES become obvious after regrouping the elements of (2.6):

$$\begin{cases} \hat{y}_{t+1} = \hat{y}_t + \alpha_0 \epsilon_t - \alpha_1 p_t - (1 - \alpha_1)\hat{p}_t \\ \hat{p}_{t+1} = \hat{y}_t + \alpha_1 \epsilon_t + \alpha_0 p_t + (1 - \alpha_0)\hat{p}_t \end{cases}. \tag{2.31}$$

When α_1 is close to 1 the influence of $\hat{p}_t$ on $\hat{y}_{t+1}$ becomes minimal and the second smoothing parameter α_0 in (2.31) behaves similar to the smoothing parameter in SES: $\alpha_0 - 1$ in CES becomes close to α in SES. For example the value $\alpha_0 = 1.24$ in CES will correspond to $\alpha = 0.24$ in SES.

The insight from this is that when the series is stationary, the optimal smoothing parameter in SES should be close to zero and the optimal α_0 in CES will be close to one. At the same time the real part of the complex smoothing parameter will become close to 2 when the series demonstrates persistent changes in level.

2.4.6 Initialisation of CES

There can be several ways of initialising states of CES. One of these follows the idea with likelihood function and perceiving the initial values of state vector as parameters that need to be optimised. This refers to the ETS optimisation proposed in Hyndman et al. (2002). The other possible way of CES initialisation is using backcasting, similar

to how it is sometimes done with ARIMA (Box and Jenkins, 1976). Finally, Kalman filter can be used for similar purposes.

Among these three options we prefer the first one, because it corresponds to the modern approach used in exponential smoothing.

2.5 Empirical results

2.5.1 Examples of simulated data

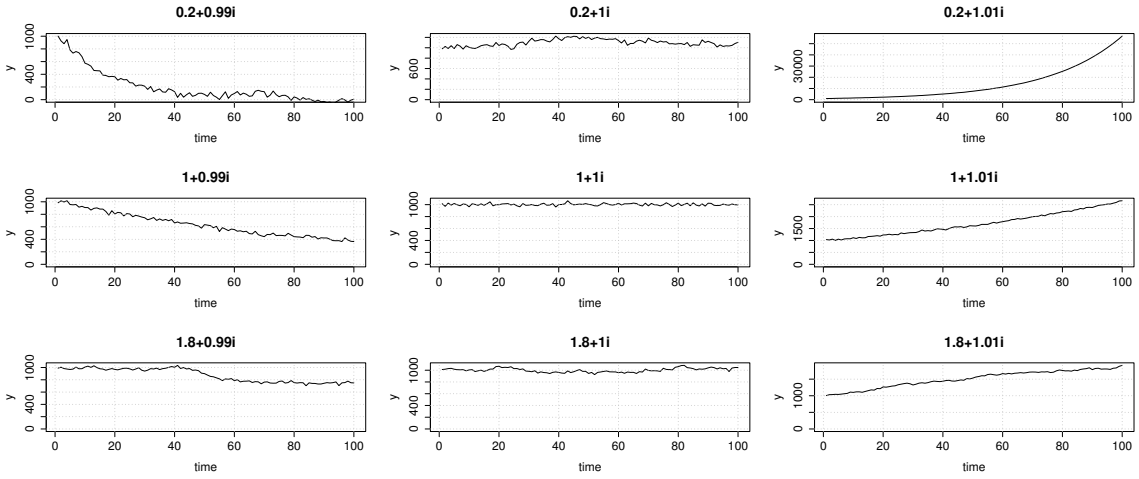
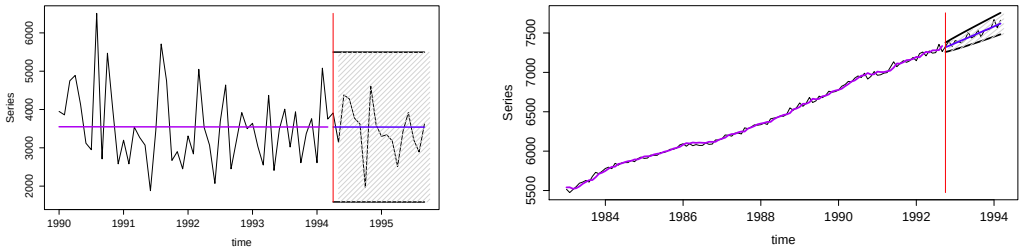


Figure 2.6: Processes simulated using CES with different complex smoothing parameter values (subplot titles). Positive initial value of level.

In order to better understand what CES is capable of producing as data generating process, we have generated several time series, which are shown in the Figure 2.6. Note that the imaginary part of the complex smoothing parameter determines the direction of the trend (or the absence of it when $\alpha_1 = 1$), while the real part influences the steepness of trend. When $\alpha_0 < 1$ the trend has obvious exponential character. When $\alpha_0 = 1$ the trend slows down but still shows the features of the exponential function. Finally, when $\alpha_0 > 1$ the series reveals features of additive trend with possible change of level and slope. One of the interesting findings is that when $\alpha_0 + i\alpha_1 = 1 + i$ the generated series becomes stationary. The other feature of CES is that the behaviour of the trend with $\alpha_1 < 1$ resembles the damped trend model. Finally if, for some reason,



(a) Series number 1664 (b) Series number 2721

Figure 2.7: Stationary and trended time series. Red line divides the in-sample with the holdout-sample.

a level of series becomes negative, then the same complex smoothing parameter values will result in the opposite trajectories direction (for example, $\alpha_0 < 1$ will result in the increase rather than decline).

These examples demonstrate flexibility of CES and its potential in dealing with level and trend time series.

2.5.2 Real time series examples

We use CES on two time series to demonstrate how it performs on real data: a level series and a trend series from M3-Competition. The first series (number 1664) is shown in Figure 2.7a and the second series (number 2721) is shown in Figure 2.7b. Historic and fitted values, point and 95% interval forecasts produced by CES are plotted. Visual and statistical (using the ADF and KPSS tests) analysis indicate that series N1664 is stationary, while N2721 exhibits a clear trend.

Estimation of CES on the first time series results in the complex smoothing parameter $\alpha_0 + i\alpha_1 = 0.99999 + 0.99996i$. This indicates that SES could potentially be efficiently used for this time series as well. All the roots of the characteristic equation for such a complex smoothing parameter lie outside the unit circle and inequality (2.23) is satisfied which means that the model produces a stationary trajectory.

CES estimated on the second time series has complex smoothing parameter $\alpha_0 +$

Table 2.1: Forecasting methods used

Name	Method
Naive	Random walk model with Naive method
SES	ETS(A,N,N) which corresponds to Simple Exponential Smoothing method
AAN	ETS(A,A,N) which corresponds to Holt's method (Holt, 2004)
MMN	ETS(M,M,N) which underlies Pegel's method (Pegels, 1969)
AAdN	ETS(A,Ad,N) which underlies additive damped trend method (Gardner and McKenzie, 1985)
MMdN	ETS(M,Md,N) which corresponds to multiplicative damped trend method (Taylor, 2003)
ZZN	ETS(Z,Z,N) – the general exponential smoothing model with the model selection procedure proposed by (Hyndman et al., 2002) using AICc.
ARIMA	ARIMA model implemented by (Hyndman et al., 2002) in R. It uses hypothesis testing and AICc for order selection.
Theta	Theta method proposed by Assimakopoulos and Nikolopoulos (2000). It performed very well on this time series and it is useful to compare CES with it. We use implementation of Theta from forecast package in R.
CES	Complex Exponential Smoothing.

$i\alpha_1 = 1.48187 + 1.00352i$. This means that the forecast of CES on the second time series is influenced by a larger number of observations compared to the first time series. There are several roots of characteristic equation lying inside the unit circle and the imaginary part of the complex smoothing parameter is greater than one. All of this indicates that the model is non-stationary in mean, producing growing trajectory. That is what we can see on the Figure 2.7b.

These two examples show that CES is capable of identifying whether the series is stationary or not and producing the appropriate forecast without the need for a model selection procedure.

2.5.3 M3 competition results

To evaluate the forecasting performance of CES against other exponential smoothing based methods we use the M3-competition dataset (Makridakis and Hibon, 2000) to conduct an empirical comparison. All 814 non-seasonal monthly time series are used. We have split all the data in two categories: level and trend series. The distinction is done using "ets" functions from "forecast" package in R: if a trend model is selected, then we consider the series to be trended. Otherwise the series is level. This should in general be favourable towards ETS models. Table 2.1 lists the benchmarks used. These benchmarks have been selected to evaluate the performance of CES against various forms of exponential smoothing forecasts that are capable of capturing level and trend components. ETS(Z,Z,N) permits selecting the most appropriate non-seasonal model for each series, allowing any type of error or trend component. We have also included ARIMA in order to see if CES with underlying ARMA(2,2) model can beat it, and Theta, which performed very well in the original M3-Competition (Makridakis and Hibon, 2000).

The estimation of all the models was done in R. All the models excluding CES were estimated using functions from "forecast" package by Hyndman and Khandakar (2008): "naive", "ets", "theta" and "auto.arima". A function "ces" from a package "smooth" (available on CRAN) was written by the authors of this paper to estimate CES.

We retain a similar setup to the original competition to facilitate comparability of the results. The last 18 observations are withheld as a test set that is forecasted by each of the methods considered. The forecasts are evaluated using the Mean Absolute Scaled Error (MASE) proposed by Hyndman and Koehler (2006).

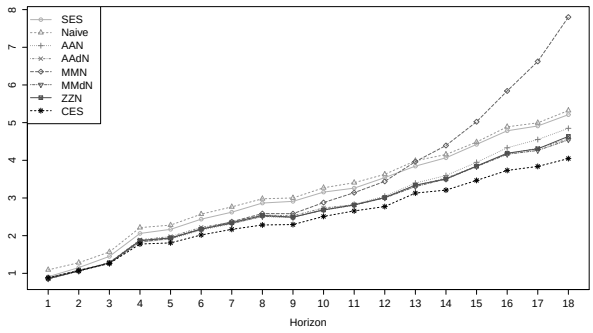
Mean and median MASE over all the horizons are shown in the Table 2.2. The table also presents split into categories of level and trend time series. The best performing method is highlighted in boldface. CES has a lowest mean and median MASE

Table 2.2: Mean ASE (MASE) and Median ASE (MdASE) values for different forecasting models

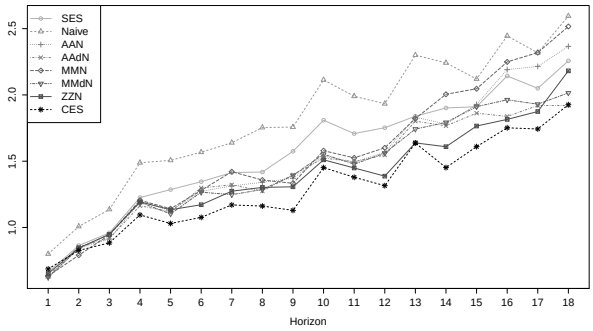
Method	Overall		Level		Trend	
	MASE	MdASE	MASE	MdASE	MASE	MdASE
Naive	3.216	1.927	1.881	1.183	4.067	3.046
SES	3.098	1.681	1.680	0.894	4.003	3.076
AAN	2.755	1.602	1.731	0.959	3.408	2.232
AAdN	2.726	1.467	1.700	0.895	3.380	2.643
MMN	3.323	1.703	2.035	0.977	4.145	2.372
MMdN	2.700	1.490	1.716	0.890	3.327	2.644
ZZN	2.716	1.466	1.682	0.901	3.375	2.395
ARIMA	2.700	1.448	1.669	0.895	3.358	2.222
Theta	2.693	1.462	**1.606**	**0.868**	3.386	2.466
CES	**2.496**	**1.387**	1.649	0.902	**3.036**	**1.990**

compared to all the other models used on this set of data overall. It outperforms ETS(A,N,N) and ETS(A,A,N) models, but also ETS(Z,Z,N) with the implemented model selection procedure. As can be seen from the column "Level" in the table CES has one of the lowest Mean ASE, being outperformed by Theta only. The situation is slightly worse for CES in Median ASE, although not substantially. As a result it is not the best method in this category, but it performs better than some other ETS models. The main advantage of CES can be seen in the last two columns, corresponding to "Trend" category of time series – CES outperforms all the other models with a substantial difference. So, we can conclude that CES managed to capture trends correctly in time series and could also produce decent forecasts in cases of level time series.

To investigate further why CES is more accurate than the other models mean and median MASE are calculated for each forecast horizon in Figure 2.8. The difference in accuracies between CES and other models increases with the increase of the forecasting horizon. For example, CES is substantially more accurate than the second best method ETS(M,Md,N) for horizons $h = 6, \dots, 18$ but it is not the best



(a) Mean MASE



(b) Median MASE

Figure 2.8: MASE values over different forecasting horizons for the competing models.

method for the shorter horizons ($h = 1, \ldots, 3$). This indicates that CES was able to capture the long-term dependencies in time series better than the other exponential smoothing models.

2.6 Conclusions

In this paper we presented a new approach to time series modelling. Instead of taking into account only the actual value of series and decomposing it into several components, we introduced the notion of the information potential variable and combined it with the actual observations in one complex variable. Adopting this approach rather than using the arbitrary decomposition of time series into several components (level, trend, error) leads to a new model, the Complex Exponential Smoothing that is introduced in this paper.

The state-space form using Single Source of Error was presented, which allowed the derivation of the stability and stationarity conditions of CES as well as conditional mean and variance of the model.

CES is a flexible model that can produce different types of trajectories. It encompasses both level and multiplicative trend series and approximates additive trend very well. One of the advantages of CES is that it smoothly moves from one trajectory to the other, depending on the complex smoothing parameter value. Furthermore, CES is able to distribute weights between different observations in time either exponentially or harmonically. This feature allows CES to capture long-term dependencies and non-linear relations in time series.

CES has an underlying ARMAX(2,2) model, where the exogenous part corresponds to the information potential variable. By using the proxy discussed in the article, the model transforms into restricted ARMA(2,2). The main difference between ARMA(2,2) underlying CES and the general ARMA(2,2) is that in CES framework the AR and MA parameters become connected with each other. This leads to a different parameter space and a different modelling of time series.

Finally, CES is empirically shown to be more accurate than various exponential smoothing benchmarks, especially ETS(A,N,N), ETS(A,A,N) and ETS with model selection. This provides evidence that CES can be used effectively to capture both level and trend time series cases, side-stepping the model selection problem that conventional exponential smoothing and similar models face. We argue that the smooth transition of CES between level and trend series, in contrast to conventional ETS that implies abrupt changes due to the model selection between level and trend forms, is advantageous. CES is also able to capture long-term dependencies which results in more accurate forecasts for the longer horizons, in comparison to the exponential smoothing benchmarks considered here.

In conclusion, the Complex Exponential Smoothing model has unique and desirable properties for time series forecasting. Using the ideas of complex variables and information potential, CES builds on the established and widely used ideas behind exponential smoothing to overcome several limitations and modelling challenges of the latter.

Chapter 3

Seasonal Complex Exponential Smoothing

In this chapter we propose an extension of the Complex Exponential Smoothing for modelling seasonality, discuss its properties and demonstrate its performance on simulated and real data. This work has been submitted as a paper to the International Journal of Forecasting (currently under review).

3.1 Introduction

Exponential smoothing (ETS) is one of the most popular family of models used both in research and practice. The variety of the exponential smoothing models is wide and allows modelling different types of time series components. Hyndman et al. (2008b) present a taxonomy which leads to 30 models with different types of error, trend and seasonal components.

ETS is not free of modelling challenges. First, the large number of model forms introduces a selection problem. Although the model variety allows to capture different types of processes, at the same time it makes it difficult to select a correct one for a time series. This is usually addressed by using an information criterion, typically the Akaike Information Criterion (Hyndman et al., 2002), though Billah et al. (2006) showed that using other information criteria did not lead to significant differences in forecasting performance. However recent research showed that choosing a single most

appropriate exponential smoothing model for a time series may not lead to the most accurate forecast. As a result various combination approaches have been proposed in the literature. For example, Kolassa (2011) investigated combination of the different ETS forecasts using Akaike weights with good results.

Second, it is assumed in ETS framework that any time series may be decomposed into level, trend and seasonal components, which in real life are arbitrary and unobservable. For example, it may not always be easy to identify whether a series exhibits a changing level or a trend. Similar complications are relevant to identifying the seasonal component and its nature: whether it is additive or multiplicative.

Third, the combination approaches highlight that there may be composite ETS forms that are not captured by the known models. The combinations described by Kolassa (2011) result in such non-customary trend and seasonality forms. Kourentzes et al. (2014) argue that full identification of trend and seasonality is not straightforward with conventional modelling, showing the benefits of using multiple levels of temporal aggregation to that purpose, demonstrating forecasting performance improvements. **?** proceed to show that this problem is more acute under the presence of extreme values, such as outliers due to promotional activities, where again a similar combination approach leads to more accurate forecasts. We argue that these composite forms of exponential smoothing perform well because the type of time series components, as well as the interaction between them, may be too restrictive under ETS for some time series.

In this paper we aim to overcome these problems by using a more general exponential smoothing formulation that has been proposed Chapter 2. It assumes that any time series can be modelled as the observed series value and an unobserved information potential. This leads to the less restrictive model without an arbitrary decomposition of time series. The authors implement this idea using complex variables, proposing the Complex Exponential Smoothing (CES). Their investigation is

limited to non-seasonal time series, where they find that CES can accurately model time series without arbitrarily separating them into level and trend ones. Another crucial advantage of CES over conventional ETS is that it can capture both stationary and non-stationary processes.

Here we extend CES for seasonal time series, which leads to a a family of CES models that can model all types of level, trend, seasonal and trend seasonal time series in the conventional ETS classification. In contrast to ETS, CES has only two forms (non-seasonal and seasonal) and hence we deal with a simpler model selection problem that allows capturing different types of components and interactions between them. To test that the extended CES models the appropriate time series structure, we conduct a simulation study and find that the simplified selection problem leads to better results in comparison with conventional ETS. We then evaluate the forecasting performance of extended CES against established benchmarks and find it to produce more accurate forecasts on the monthly M3 dataset. We argue that even though the formulation of CES appears to be more complicated than individual ETS models, the substantially simplified selection problem and its good accuracy makes it appealing for practice.

The rest of the paper is organised as follows: section 3.2 introduces the Complex Exponential Smoothing model and its seasonal extension. Section 3.3 presents the setup and provides the results of empirical evaluations on simulated and real data. This section is then followed by concluding remarks.

3.2 Complex Exponential Smoothing

3.2.1 Information Potential

A fundamental idea behind CES is the *Information Potential* (discussed in Chapter 2). Any measured time series contains some information, which may be less than the time series in all its totality and potentially unobservable due to sampling. For

example, that might be due to some unobserved long memory process or other more exotic structures.

It was shown that there is a convenient way to write the measured series and the information potential using a complex variable: $y_t + ip_t$ (Svetunkov, 2012), where y_t is the actual value of the series, p_t is the information potential on the observation t and i is the imaginary unit, the number that satisfies the equation: $i^2 = -1$. In Chapter 2 we have proposed using the value of the error term: $p_t = \epsilon_t$ as this proxy which exhibits desirable properties that are discussed below.

3.2.2 Non-seasonal CES

A complex counterpart to conventional exponential smoothing can be developed as:

$$\hat{y}_{t+1} + i\hat{p}_{t+1} = (\alpha_0 + i\alpha_1)(y_t + ip_t) + (1 - \alpha_0 + i - i\alpha_1)(\hat{y}_t + i\hat{p}_t) \tag{3.1}$$

where $\hat{y}_t$ is the forecast of the actual series, $\hat{p}_t$ is the estimate of the information potential, $\alpha_0 + i\alpha_1$ is complex smoothing parameter. Using the aforementioned proxy $p_t = \epsilon_t$, Svetunkov and Kourentzes (2015) derived the state-space model of CES that can be used to further explore its properties. It can be written in the following way:

$$\begin{cases} y_t = l_{t-1} + \epsilon_t \\ l_t = l_{t-1} - (1 - \alpha_1)c_{t-1} + (\alpha_0 - \alpha_1)\epsilon_t \\ c_t = l_{t-1} + (1 - \alpha_0)c_{t-1} + (\alpha_0 + \alpha_1)\epsilon_t \end{cases}, \tag{3.2}$$

where l_t is the level component, c_t is the information component on observation t, and $\epsilon_t \sim N(0, \sigma^2)$. This state-space form of CES can be written in a more general form:

$$\begin{cases} y_t = w' x_{t-1} + \epsilon_t \\ x_t = F x_{t-1} + g\epsilon_t \end{cases}, \tag{3.3}$$

where $x_t = \begin{pmatrix} l_t \\ c_t \end{pmatrix}$ is state vector, $F = \begin{pmatrix} 1 & -(1 - \alpha_1) \\ 1 & 1 - \alpha_0 \end{pmatrix}$ is transition matrix, $g = \begin{pmatrix} \alpha_0 - \alpha_1 \\ \alpha_0 + \alpha_1 \end{pmatrix}$ is persistence matrix and $w = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is a measurement vector.

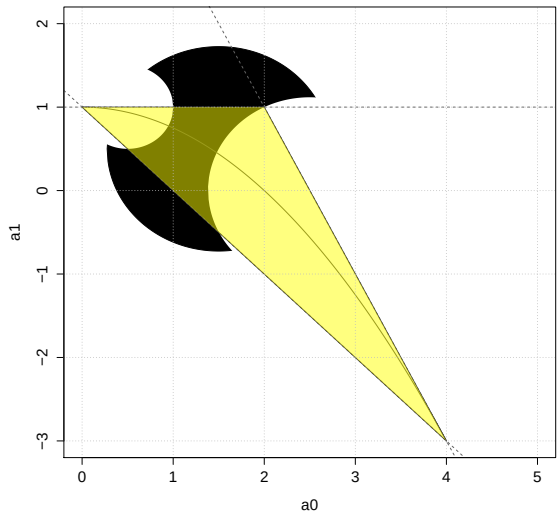


Figure 3.1: CES stability (the black area) and stationarity (the light triangle) conditions.

One of the main features of CES is that it does not contain an explicit trend: its l_t and c_t components are connected with each other and change in time depending on the value of the complex smoothing parameter. It can be shown that CES has an underlying ARIMA(2,0,2) model:

$$\begin{cases} (1 - \phi_1 B - \phi_2 B^2) y_t = (1 - \theta_{1,1} B - \theta_{1,2} B^2) \epsilon_t \\ (1 - \phi_1 B - \phi_2 B^2) \xi_t = (1 - \theta_{2,1} B - \theta_{2,2} B^2) \epsilon_t \end{cases}, \tag{3.4}$$

where $\phi_1 = 2 - \alpha_0$, $\phi_2 = \alpha_0 + \alpha_1 - 2$, $\theta_{1,1} = 2 - 2\alpha_0 + \alpha_1$, $\theta_{1,2} = 3\alpha_0 + \alpha_1 - 2 - \alpha_0^2 - \alpha_1^2$, $\theta_{2,1} = 2 + \alpha_1$, $\theta_{2,2} = \alpha_0 - \alpha_1 - 2$ and $\xi_t = p_t - c_{t-1}$ is an information gap, the value showing the amount of the information missing in the information component c_t.

However the parameter space for the autoregressive terms in this model differs from the conventional AR(2) model due to the connection of AR and MA terms via the complex smoothing parameter. This gives the model an additional flexibility and allows it to switch naturally between level and trend time series. It also should be noted that CES can be both stationary and non-stationary, depending on the complex smoothing parameter value, while all the conventional ETS models are strictly non-stationary.

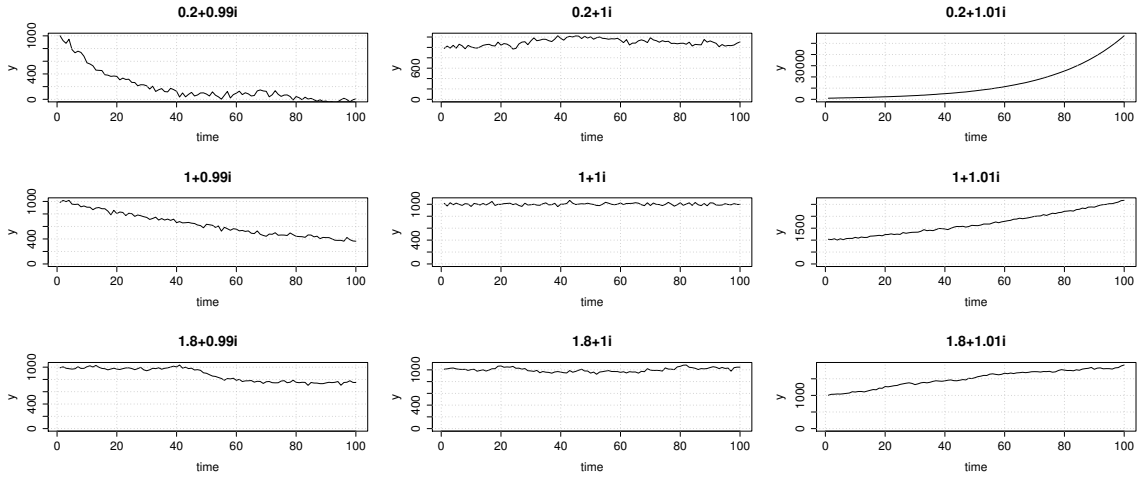


Figure 3.2: Processes simulated using CES with different complex smoothing parameter values (subplot titles). Positive initial value of level.

The stability and stationarity conditions for CES have already been discussed in Chapter 2. They are also shown on the Figure 3.1 as a reminder. They correspond to the following inequalities:

$$\begin{cases} (\alpha_0 - 2.5)^2 + \alpha_1^2 > 1.25 \\ (\alpha_0 - 0.5)^2 + (\alpha_1 - 1)^2 > 0.25 \\ (\alpha_0 - 1.5)^2 + (\alpha_1 - 0.5)^2 < 1.5 \end{cases} \text{and} \begin{cases} \alpha_1 < 5 - 2\alpha_0 \\ \alpha_1 < 1 \\ \alpha_1 > 1 - \alpha_0 \end{cases}. \tag{3.5}$$

As it was discussed in the Chapter 2, CES can produce different types of trajectories depending on the complex smoothing parameter value. These trajectories are shown in the Figure 3.2 as a reminder of what CES is capable of producing. The examples in the Figure demonstrate that CES is able to produce different types of trends as well as level series (in an ETS context). In Chapter 2 we showed empirically that a single CES model can be used instead of several ETS models with different types of trends for common forecasting tasks. This is one of the major advantages of the model. Disregarding seasonal time series no model selection is needed for CES, in contrast to conventional exponential smoothing.

3.2.3 Seasonal CES

Here we extend the CES model presented above to cater for seasonality. The simplest way to derive a seasonal model using CES is to take values of level and information components with a lag m which corresponds to the seasonality lag instead of $t-1$:

$$\begin{cases} y_t = l_{1,t-m} + \epsilon_t \\ l_{1,t} = l_{1,t-m} - (1-\beta_1)c_{1,t-m} + (\beta_0 - \beta_1)\epsilon_t \\ c_{1,t} = l_{1,t-m} + (1-\beta_0)c_{1,t-m} + (\beta_0 + \beta_1)\epsilon_t \end{cases} \tag{3.6}$$

where $l_{1,t}$ is the level component, $c_{1,t}$ is the information component on observation t, β_0 and β_1 are smoothing parameters. This formulation follows similar ideas to the reduced seasonal exponential smoothing forms by Snyder et al. (2001) and Ord and Fildes (2012).

The model (3.6) preserves all the features of the original CES (3.2) and produces all the possible CES trajectories with the same complex smoothing parameter values. These trajectories, however, will be lagged and will appear for each seasonal element instead of each observation. This means that the model (3.6) produces non-linear seasonality which can approximate both additive and multiplicative seasonality depending on the complex smoothing parameter and initial components values. Note that the seasonality produced by the model may even have these features regardless of the value of the level, thus generating seasonalities that cannot be classified as either additive or multiplicative. Naturally, no known exponential smoothing model can produce similar patterns: multiplicative seasonality implies that the amplitude of fluctuations increases with the increase of the level, while with seasonal CES the amplitude may increase without it.

However this model is not flexible enough: it demonstrates all the variety of possible seasonality changes only when the level of the series is equal to zero (meaning that a part of the data lies in the positive and another lies in the negative plane). To overcome this limitation we extend the original model (3.2) with the basic seasonal

model (3.6) in one general seasonal CES model:

$$\begin{cases} y_t = l_{0,t-1} + l_{1,t-m} + \epsilon_t \\ l_{0,t} = l_{0,t-1} - (1-\alpha_1)c_{0,t-1} + (\alpha_0 - \alpha_1)\epsilon_t \\ c_{0,t} = l_{0,t-1} + (1-\alpha_0)c_{0,t-1} + (\alpha_0 + \alpha_1)\epsilon_t \\ l_{1,t} = l_{1,t-m} - (1-\beta_1)c_{1,t-m} + (\beta_0 - \beta_1)\epsilon_t \\ c_{1,t} = l_{1,t-m} + (1-\beta_0)c_{1,t-m} + (\beta_0 + \beta_1)\epsilon_t \end{cases} \tag{3.7}$$

The model (3.7) still can be written in a conventional state-space form (3.2). It exhibits several differences from the conventional smoothing seasonal models. First, the proposed seasonal CES in (3.7) does not have a set of usual seasonal components as the ordinary exponential smoothing models do, which means that there is no need to renormalise them. The values of $l_{1,t}$ and $c_{1,t}$ correspond to some estimates of level and information components in the past and have more common features with seasonal ARIMA (Box and Jenkins, 1976, p.300) than with the conventional seasonal exponential smoothing models. Second, it can be shown that the general seasonal CES has an underlying model that corresponds to SARIMA$(2, 0, 2m + 2)(2, 0, 0)_m$ (see Appendix F), which can be either stationary or not, depending on the complex smoothing parameters values.

The general seasonal CES preserves the properties of both models (3.2) and (3.6): it can produce non-linear seasonality and all the possible types of trends discussed above, as now the original level component $l_{0,t}$ can become negative while the lagged level component $l_{1,t}$ may become strictly positive.

Finally this model retains the interesting property of independence of the original level and lagged level components, so a multiplicative (or other) shape seasonality may appear in the data even when the level of the series does not change. This could happen for example when the seasonality is either nonlinear or some other variable is determining its evolution, as for example is the case with solar power generation (Trapero et al., 2015).

3.2.4 Parameters estimation

During the estimation of the parameters of the general seasonal CES some constraints should be introduced to achieve the stability of the model. The state-space model is stable when all the eigenvalues of the discount matrix $D = F - gw'$ lie inside the unit circle. Unfortunately, it is hard to derive the exact regions for the smoothing parameters of general seasonal CES but the stability condition can be checked during the optimisation. To do that the eigenvalues of the following discount matrix should be calculated (see Appendix G):

$$D = \begin{pmatrix} 1-\alpha_0+\alpha_1 & \alpha_1-1 & \alpha_1-\alpha_0 & 0 \\ 1-\alpha_0-\alpha_1 & 1-\alpha_0 & -\alpha_1-\alpha_0 & 0 \\ \beta_1-\beta_0 & 0 & 1-\beta_0+\beta_1 & \beta_1-1 \\ -\beta_1-\beta_0 & 0 & 1-\beta_0-\beta_1 & 1-\beta_0 \end{pmatrix} \tag{3.8}$$

The estimation of the parameters of CES can be done using the likelihood function. This is possible due to the additive form of the error term in (3.7). The likelihood function turns out to be similar to the one used in ETS models (Hyndman et al., 2002):

$$L(g, x_0, \sigma^2|y) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^T \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left(\frac{\epsilon_t}{\sigma}\right)^2\right) \tag{3.9}$$

This likelihood function value can be used in calculation of information criteria. The twice the negative logarithm of (3.9) can be calculated to that purpose:

$$-2\log(L(g, x_0|y)) = T\left(\log\left(\frac{2\pi e}{T}\right) + \log\left(\sum_{t=1}^{T}\epsilon_t^2\right)\right). \tag{3.10}$$

3.2.5 Model selection

Using (3.10) the Akaike Information Criterion for both seasonal and non-seasonal CES models can be calculated:

$$AIC = 2k - 2\log(L(g, x_0|y)) \tag{3.11}$$

where k is the number of coefficients and initial states of CES. For the non-seasonal model (3.2) k is equal to 4 (2 complex smoothing parameters and 2 initial states). For the seasonal model the number of the coefficients in (3.7) becomes much greater than in the original model: $k = 4 + 2m + 2$, which is 4 smoothing parameters, $2m$ initial lagged values and 2 initial values of the generic level and information components. Naturally, other information criteria can be constructed.

Observe that the model selection problem for CES is reduced to choosing only between non-seasonal and seasonal variants, instead of the multiple model forms under conventional ETS.

3.3 Empirical evaluation

3.3.1 Model selection

To evaluate the performance of the model selection procedure we simulate series with known structure and attempt to model them with CES. Each series is simulated using ETS at a monthly frequency and contains 120 observations. All the smoothing parameters are generated using uniform distribution, restricted by the traditional bounds: $\alpha \in (0, 1), \beta \in (0, \alpha), \gamma \in (0, 1 - \alpha)$. Normal distribution with zero mean is used in the error term generation. We generate series with either additive or multiplicative errors, using error standard deviation equal to 50 and 0.05 respectively. The initial value of the level is set to 5000, while the initial values of trend is set to 0 for the additive cases and to 1 for the multiplicative cases. This allows the model used as DGP to produce either growth or decline, depending on the error term and smoothing parameter values. All the initial values of seasonal components are randomly generated and then normalised. For each of the 9 process shown in Table 3.1 1000 time series are generated.

We fit the two types of CES and chose the one with the smallest AIC corrected for small sample sizes (AICc) value. As benchmarks we fit ETS and ARIMA (both

DGP	CES	ETS			ARIMA		
		Trend	Seasonal	Overall	Trend	Seasonal	Overall
$N(5000, 50^2)$	**99.9**	97.3	99.8	97.1	96.5	45.7	44.1
ETS(ANN)	99.1	88.0	**99.7**	49.3	51.5	46.5	28.0
ETS(MNN)	99.3	85.1	**99.6**	50.9	59.7	47.3	30.0
ETS(AAN)	91.5	94.4	**99.7**	82.3	96.4	45.7	43.5
ETS(MMN)	98.9	91.6	**99.7**	68.9	92.3	35.2	32.2
ETS(ANA)	**100**	85.4	**100**	46.3	53.0	**100**	53.0
ETS(AAA)	**100**	92.1	**100**	79.1	86.3	**100**	86.3
ETS(MNM)	**100**	65.9	**100**	32.6	61.9	**100**	61.9
ETS(MMM)	98.2	88.4	**100**	52.7	70.3	**100**	70.3
Average	98.5	87.6	**99.8**	62.1	74.2	68.9	49.9

Table 3.1: The percentage of forecasting models chosen correctly for each ETS data generating process (DGP).

use AICc for model selection). The implementation of CES was done in R and is available as the "CES" package (in github: `https://github.com/config-i1/CES`). The benchmarks are produced using the "forecast" package for R (Hyndman and Khandakar, 2008).

The percentage of the correct models chosen by CES, ETS and ARIMA are shown in the Table 3.1. The values in the table are the percentage of successful time series characteristics identified by each model. Column "CES" shows in how many instances the appropriate seasonal or non-seasonal model is chosen. We can observe that CES is very accurate and managed to select the appropriate model in the majority of cases. It is important to note that the fitted CES is not equivalent to the data generating process in each time series, but nevertheless it is able to approximate the time series structure.

The column "ETS, Trend" shows in how many cases the trend or its absence is identified correctly (not taking into account the type of trend). The "ETS, Seasonal" column shows similar accuracy in the identification of seasonal components. The lowest trends identification ETS accuracy is in the case of ETS(MNM) process with 65.9%, while the average accuracy in capturing trends correctly is 87.6%. The accu-

racy of seasonal components identification in ETS is much better, with the average accuracy for all the DGPs in this case being 99.8%. The column "Overall" shows in how many cases the exact DGP is identified by ETS, not taking parameters values into account, but considering type of estimated model. So for example, ETS(M,N,N) selected on data generated by ETS(A,N,N) would be considered as an error in this column. The average accuracy of ETS in this task is 62.1%. Note that contrary to CES, ETS should have identified the exact model in the majority of cases, since it was used to generate the series. The reason behind this misidentification may be due to the information criterion used and the sample size of the generated time series.

Table 3.1 also provides results for the ARIMA benchmark. The column "ARIMA, Trend" shows the number of cases where ARIMA identifies the trend or its absence correctly. Although trend in ARIMA is not defined directly, we use three criteria to indicate that ARIMA captured a trend in the time series: ARIMA has a drift, or ARIMA has second differences, or ARIMA has first differences with non-zero AR element. It can be noted that the lowest accuracy of ARIMA in trends identification is in the cases of level ETS models DGPs. The average accuracy of ARIMA in this task is 74.2%.

The column "ARIMA, Seasonal" shows the number of cases where ARIMA identifies the seasonality or its absence correctly. The cases where either seasonal AR or seasonal MA or seasonal differences contains non-zero values are identified as seasonal ARIMA models. The values in the Table 3.1 indicate that ARIMA makes a lot of mistakes in identifying the seasonal models in time series generated using non-seasonal ETS models, which leads to the average accuracy of 68.9% in this category. Even though the generating processes is different than the model used in this case, capturing seasonality in the pure non-seasonal data makes no sense. The column "ARIMA, Overall" shows the number of cases where both trend and seasonality are identified correctly by ARIMA using the criteria described above. Note in the "Overall" column

we only count the cases where ARIMA correctly identifies both the presence of trend and season, so as to make it comparable with the CES results. ARIMA identifies the correct form mainly in the cases of ETS(AAA) and ETS(MMM) models.

Obviously the ETS and ARIMA results are dependent on the identification methodology employed. But the comparison done here retains its significance since all three alternative models, CES, ETS and ARIMA use the same information criterion to identify the final model form.

We have also conducted similar experiment, generating data using several ARIMA models. In order for the results to be comparable with the ones obtained in table 3.1, we have decided to generate data using:

- ARIMA(0,1,1), because it underlies ETS(A,N,N) model,
- ARIMA(0,2,2) – underlies ETS(A,A,N),
- ARIMA(1,1,2) – underlies ETS(A,Ad,N),
- ARIMA(2,0,2) with constant – underlies CES,
- SARIMA$(0,1,1)(0,1,1)_{12}$ – well-known airline model from Box and Jenkins (1976),
- SARIMA$(0,2,2)(1,0,1)_{12}$ – has non-seasonal part corresponding to ETS(A,A,N) and a stationary seasonal part,
- SARIMA$(1,1,2)(1,0,1)_{12}$ – has non-seasonal part corresponding to ETS(A,Ad,N) and a stationary seasonal part,
- SARIMA$(2,0,2)(0,1,1)_{12}$ – has non-seasonal part corresponding to CES and a non-stationary seasonal part.

Obviously this is a restricted research as the variety of ARIMA models is wide and there can be many more other models that could potentially be included in the experiment. All the parameters for ARIMA models are generated randomly from station-

arity and invertibility regions using uniform distribution. Variance of the error is set to 50, initial values are randomly generated, making sure that there are no negative values in the data. For each of the 8 ARIMA process 1000 time series, containing 120 observations are generated.

Method	CES	ETS	ARIMA
ARIMA(0,1,1)	**100**	99.9	58.3
ARIMA(0,2,2)	**100**	99.8	47.1
ARIMA(1,1,2)	**98.4**	96.2	49.5
ARIMA(2,0,2)	**98.4**	94.5	40.6
SARIMA$(0,1,1)(0,1,1)_{12}$	93.5	**100**	98.5
SARIMA$(0,2,2)(1,0,1)_{12}$	11.9	18.5	**89.9**
SARIMA$(1,1,2)(1,0,1)_{12}$	14.1	16.5	**93.9**
SARIMA$(2,0,2)(0,1,1)_{12}$	98	**99.6**	95.9

Table 3.2: The percentage of forecasting models chosen correctly for each ARIMA DGP.

While it is possible to show that some ARIMA models imply that there is a trend in a time series, generated by ETS, it is not possible to do anything similar in the opposite direction. That is why we have only three rows in the table 3.2: "CES", "ETS" and "ARIMA". The rows show in how many cases each model managed to capture seasonality correctly. For CES it means that for seasonal data a seasonal model was selected, while for ETS it means that for a seasonal time series a model with any type of seasonal component was selected. "ARIMA accuracy" shows in how many cases the correct seasonal or non-seasonal ARIMA was selected (seasonal ARIMA in this case is the model with any non-zero seasonal part). We don't measure if orders of ARIMA were selected correctly, but focus on its ability to distinguish seasonal series from non-seasonal.

The results of this experiment are shown in the table 3.2. As it can be seen, CES with automatic model selection performed extremely good for non-seasonal data, beating both ETS and ARIMA and also performed very well for seasonal data. However, we need to point out that both CES and ETS failed to identify seasonality for

cases of SARIMA$(0,2,2)(1,0,1)_{12}$ and SARIMA$(1,1,2)(1,0,1)_{12}$. So, there are cases of data, when these models do not perform well and ARIMA should be used instead. However, analysing ARIMA performance, it can be concluded, that it overfitts the data, selecting seasonal components even in those cases, when they are not needed at all. While this could be fine in case with ETS DGP, there is no excuse for such a poor performance on data, generated using ARIMA itself.

Overall this simulation experiment shows that CES, having only two models to choose from, becomes more efficient in time series identification than ETS and ARIMA models, that both have to choose from multiple alternatives. The possible explanation of this is that CES is capable of capturing both level and trend series and its seasonal extension can produce both additive and multiplicative seasonality. This reduces the number of alternative models substantially. Furthermore any seasonal components are highly penalized with CES during model selection. Therefore the seasonal model is chosen only in cases when it fits data significantly better than its non-seasonal counterpart. This substantially simplifies the forecasting process as a whole.

3.3.2 Forecasting accuracy

To test the forecasting performance of CES and compare it against ETS and ARIMA we conduct an empirical evaluation on the monthly data from the M3 Competition (Makridakis and Hibon, 2000) that contains both trend and seasonal time series. The forecasting horizon (18 periods ahead) is retained the same as in the original competition, however a rolling origin evaluation scheme is used, with the last 24 observations withheld.

The Mean Absolute Scaled Error (MASE) is used to compare the performance of models for each forecast horizon from each origin (Hyndman and Koehler, 2006):

$$MASE = \frac{\sum_{j=1}^{h} |e_{T+j}|}{\frac{h}{n-1}\sum_{t=2}^{T} |y_t - y_{t-1}|} \tag{3.12}$$

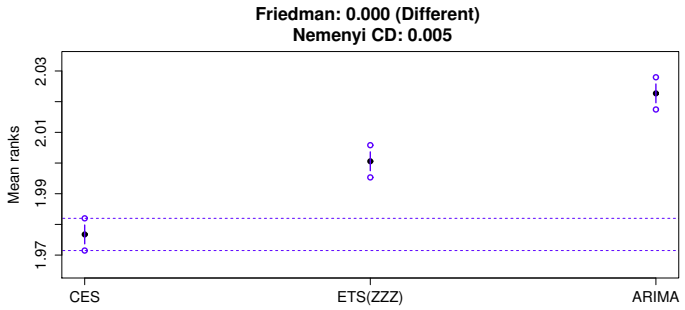


Figure 3.3: The results of MCB test on monthly data of M3. The dotted lines are the critical distances for the best model and we can see that both ETS and ARIMA are found to be significantly different.

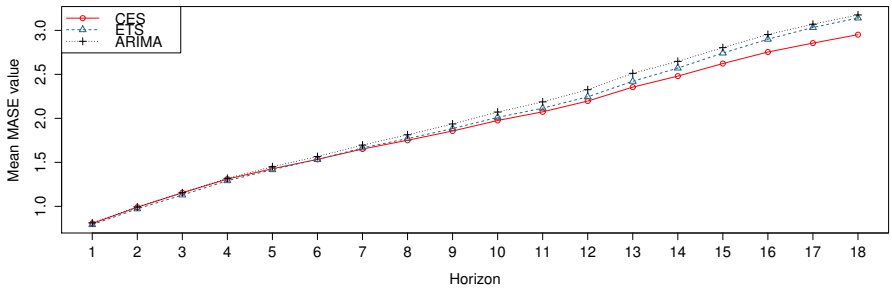
Using these errors we estimate quantiles of the distribution of MASE along with the mean values from each origin. Table 3.3 presents the results of this experiment.

	CES	ETS	ARIMA
Minimum	0.134	**0.084**	0.098
25% quantile	0.665	**0.664**	0.703
Median	**1.049**	1.058	1.093
75% quantile	**2.178**	2.318	2.224
Maximum	**28.440**	53.330	59.343
Mean	**1.922**	1.934	1.967

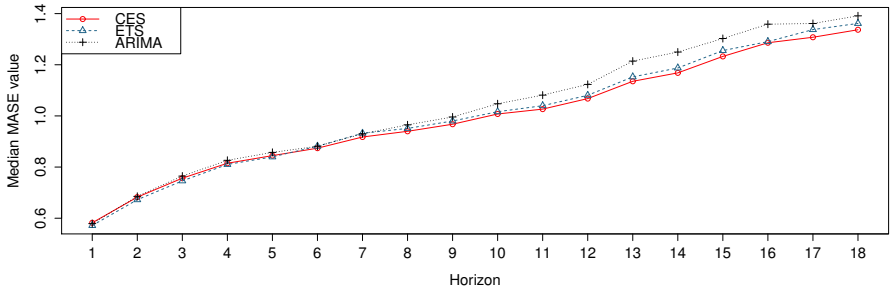
Table 3.3: MASE values of competing methods. The smallest values are in bold.

CES demonstrates the smallest mean and median MASE. It also has the smallest maximum error which indicates that when all the models failed to forecast some time series accurately, CES was still closer to the actual values. In the contrast ETS had the smallest minimum MASE, while CES had the highest respective value among all the competitors. This means that CES may not be as accurate as other models in the cases when the time series is relatively easy.

To see if the difference in the forecasting accuracy between CES and the other methods is significant, we use the Multiple Comparisons with Best (MCB) test (Koning et al., 2005). The results of this test indicate that CES is significantly more accurate than ETS and ARIMA (see Figure 3.3).



(a) Mean MASE



(b) Median MASE

Figure 3.4: Mean and median MASE value for CES, ETS and ARIMA on different horizons.

To investigate what could cause this difference in forecasting accuracy, the mean and median MASE are calculated for each forecast horizon, obtaining the one-step ahead, two-steps ahead etc. mean and median MASE values. These values are shown on the Figures 3.4a and 3.4b.

The analysis of the Figure 3.4 shows that while the errors of the methods are very similar for short horizons (with ETS being slightly better), the difference between them increases on longer horizons. The difference in mean values of MASE between methods starts increasing approximately after a year (12 observations) with CES being the most accurate. The same feature is present in median MASE values.

Concluding the results of this experiment, CES performs significantly better than ETS and ARIMA on the monthly M3 data. CES demonstrates that it can model various types of time series successfully and is particularly accurate in relation to the

benchmark models for longer horizons. We attribute this to the way that time series trends and seasonal components are modelled under CES.

3.4 Conclusions

We proposed a new type of seasonal exponential smoothing, based on Complex Exponential Smoothing and discussed the model selection procedure that can be used to identify whether the non-seasonal or seasonal CES should be selected for time series forecasting. While the non-seasonal CES can produce different non-linear trajectories and approximate additive and multiplicative trends, the seasonal model produces non-linear seasonality that can approximate both additive and multiplicative seasonal time series and even model new forms of seasonality that do not strictly lie under either of these cases. The latter can be achieved as a result of the independence of the seasonal and non-seasonal components and non-linearity of CES. This is one of the potential future research direction in the field of CES.

We also discussed the statistical properties of the general seasonal CES and showed that it has an underlying SARIMA$(2, 0, 2m + 2), (2, 0, 0)_m$ model. This model can be either stationary or not, depending on complex smoothing parameter values, which gives CES its flexibility and a clear advantage over the conventional exponential smoothing models which are strictly nonstationary.

The simulation conducted showed that the proposed model selection procedure allows choosing the appropriate CES effectively, making few mistakes in distinguishing seasonal from non-seasonal time series. In comparison ETS and ARIMA were not as effective as CES in this task. We argue that this is a particularly useful feature of the CES model family. Although the models initially may appear more complicated than conventional ETS, the fact that a forecaster needs to consider only two CES variants, while being able to model a wide variety of trend and seasonal shapes, can greatly simplify the forecasting process.

The forecast competition between CES, ETS and ARIMA conducted on the monthly data from M3 showed that CES achieved both the lowest mean and median forecasting errors. The MCB test showed that the differences in forecasting accuracy between these models were statistically significant. Further analysis showed that the superior performance of CES was mainly caused by the more accurate long-term forecasts. This is attributed to the ability of CES to capture the long-term dependencies in time series, due to its non-linear character that permits CES to obtain more flexible weight distributions in time than ETS.

Overall CES proved to be a good model that simplifies the model selection procedure substantially and increases the forecasting accuracy.

Chapter 4

Estimation of Complex Exponential Smoothing

This chapter is based on the materials of the paper "Trace forecast and shrinkage in time series models" submitted to Journal of American Statistical Association (currently under review).

4.1 Introduction

Estimation of time series models using methods based on multiple steps ahead errors has been widely researched by statisticians and econometricians. Many papers argue that using these estimators leads to increase in forecast accuracy (Weiss and Andersen, 1984; Chevillon and Hendry, 2005; Taylor, 2008; Chevillon, 2009; Franses and Legerstee, 2009; McElroy, 2015; Chevillon, 2016) especially when the model is misspecified (Proietti, 2011; Xia and Tong, 2011). However several researches demonstrate that this finding may not hold universally and depends mainly on data characteristics and the degree of misspecification of the model (Kang, 2003; Marcellino et al., 2006; Proietti, 2011; McElroy and Wildi, 2013). Discussing the statistical properties of the estimated parameters with such methods, it has been shown that multiple steps ahead estimators are asymptotically efficient (Haywood and Tunnicliffe-Wilson, 1997; Ing, 2003; Chevillon and Hendry, 2005; Chevillon, 2007), consistent and normal (Weiss, 1991; Haywood and Tunnicliffe-Wilson, 1997). But at the same time it

was shown by (Tiao and Xu, 1993) that parameters of models become less efficient on finite samples in comparison with the conventional one-step-ahead mean squared error estimator and that the efficiency may decrease even further when the forecast horizon increases.

There is also a general understanding that the usage of multiple steps ahead estimators leads to more conservative and robust models (Cox, 1961; Gersch and Kitagawa, 1983; Tiao and Xu, 1993; Marcellino et al., 2006; McElroy, 2015). Different experiments carried out over the years have demonstrated this statement. Still there is no plausible and detailed explanation why this happens, although many researchers imply the obviousness of such an explanation.

Furthermore, the literature has also proposed different model selection procedures derived especially for these estimators, and there are several papers discussing distribution of multiple steps ahead errors (Bhansali, 1996, 1997; Ing, 2004; Ing et al., 2009; Jordà and Marcellino, 2010; Pesaran et al., 2010). Unfortunately, these discussions are limited and have not proposed a simple and concise approach to distribution estimation and model selection using these estimators.

Finally, all the papers in the field discuss ARIMA models exclusively and there is no discussion of state-space models and the application of these estimators to them or other modelling frameworks.

The aim of this paper is to fill several gaps in the field of multiple steps ahead estimators mentioned above:

1. Give the simple and clear explanation of what happens with any time series model, when any multiple steps ahead estimator is used. This is shown using Single Source of Error (SSOE) State-space models, but an approach similar to the one discussed here can be applied to any other time series model;

2. Propose a simple likelihood function based on a multiple steps ahead estimator, which would guarantee that the estimates of parameters are less biased, more

efficient and consistent than the estimates produced by any other multiple steps ahead method. This likelihood function is called by the authors "trace forecast likelihood".

The paper is organised as follows: in section 4.2 the conventional multiple steps ahead estimators are discussed and the concise explanation of the parameter behaviour in case of their usage is given. The trace forecast likelihood is proposed and discussed in section 4.3. In the section 4.4 a simulation demonstrating the properties of discussed estimators is carried out. Finally section 4.5 presents examples on real data, followed by concluding remarks.

4.2 Conventional multiple steps estimators

One of the simplest and well-known estimators based on multiple steps ahead forecast error is mean squared h-steps ahead error:

$$\text{MSE}_h = \frac{1}{T}\sum_{t=1}^{T} e_{t+h|t}^2, \tag{4.1}$$

where $e_{t+h|t} = y_{t+h} - \mu_{t+h|t}$ is conditional h-steps ahead forecast error, y_{t+h} is the actual value, $\mu_{t+h|t}$ is the conditional expectation (point forecast) on the observation $t+h$ produced from the point in time t, while T is the number of observations in sample.

The estimator (4.1) is sometimes used in order to estimate a model several times for each horizon from 1 to h steps ahead (Kang, 2003; Chevillon and Hendry, 2005; Pesaran et al., 2010). The model is then optimised h times instead of one and has h different values of parameters. Such an estimator is sometimes called "direct multi-steps estimator" or "DMS" (Chevillon, 2007). The estimation procedure using DMS is obviously more complicated than the one based on (4.1) when used only once, but is reported to produce models with increased prediction accuracy. This is due to alignment of forecast with the objective function, i.e. predictions of h-steps ahead.

One of the other popular estimators implies that instead of taking only one error, on observation h, a forecaster takes all the errors for horizons from 1 to h, sums up the values of (4.1) and estimates the model (Weiss and Andersen, 1984; Xia and Tong, 2011). The resulting estimator can be called "Mean Squared Trace Forecast Error" and is calculated using:

$$\text{MSTFE} = \frac{1}{T}\sum_{j=1}^{h}\sum_{t=1}^{T} e_{t+j|t}^2. \tag{4.2}$$

The important fact about both estimators (4.1) and (4.2) is that they can be decomposed into bias and variance (which was discussed by Taieb and Atiya (2016) in detail). This means that they are proportional to variances of multiple steps ahead forecast errors: $\text{MSE}_h \propto \sigma_h^2$, MSTFE$\propto \sum_{j=1}^{h}\sigma_j^2$. This property in turn can be used in order to explain what happens with time series models when these estimators are used.

In order to give a simple and concise explanation, we use SSOE state-space model (Snyder, 1985):

$$\begin{cases} y_t = \mathbf{w}'\mathbf{v}_{t-1} + \epsilon_t \\ \mathbf{v}_t = \mathbf{F}\mathbf{v}_{t-1} + \mathbf{g}\epsilon_t \end{cases}, \tag{4.3}$$

where $\mathbf{v}_t$ is a state vector, $\mathbf{F}$ is a transition matrix, $\mathbf{g}$ is a persistence vector, $\mathbf{w}$ is a measurement vector and $\epsilon_t \sim N(0, \sigma^2)$ is a white noise process.

Using (4.3) it can be shown that the actual value for some observation $t+h$ (where $h \neq 1$) using the values of $\mathbf{v}_t, \mathbf{F}, \mathbf{w}$ and $\mathbf{g}$ can be calculated using:

$$y_{t+h} = \mathbf{w}'\mathbf{F}^{h-1}\mathbf{v}_t + \sum_{j=1}^{h-1} c_{j,h}\epsilon_{t+j} + \epsilon_{t+h} \tag{4.4}$$

where $c_{j,h} = \mathbf{w}'\mathbf{F}^{h-j-1}\mathbf{g}$.

The expected value at observation $t+h$ will then be equal to:

$$\mu_{t+h|t} = \mathbf{w}'\mathbf{F}^{h-1}\mathbf{v}_t \tag{4.5}$$

Substituting (4.5) in (4.4), we obtain the conditional h steps ahead forecast error:

$$e_{t+h|t} = y_{t+h} - \mu_{t+h|t} = \sum_{j=1}^{h-1} c_{j,h}\epsilon_{t+j} + \epsilon_{t+h}. \tag{4.6}$$

Note that when $h = 1$ the 1-step ahead error should be calculated using a different formula, it will simply be equal to the error term:

$$e_{t+1|t} = \epsilon_{t+1}. \tag{4.7}$$

The formulae (4.6) and (4.7) are essential for our analysis because they show how the forecast error is connected with the error term ϵ_t and what elements it contains.

Assuming that ϵ_t is not autocorrelated and is homoscedastic, the variance of the error term (4.6) can be calculated using (for a similar derivation see Hyndman et al. (2008b)):

$$\sigma_h^2 = \begin{cases} \sigma_1^2 \left(1 + \sum_{j=1}^{h-1} c_{j,h}^2\right) & \text{when } h > 1 \\ \sigma_1^2 & \text{when } h = 1 \end{cases}, \tag{4.8}$$

where σ_h^2 is variance of h-steps ahead forecast error.

The finding that the σ_h^2 can be decomposed into variance of one step ahead error and sum of squared parameters is not new; it has been briefly discussed in several papers (for examples, see Bhansali (1996, 1997); Haywood and Tunnicliffe-Wilson (1997)). But to our knowledge it has never been explored further. Here we would like to note that the minimisation of MSE_h in (4.1) implies the minimisation of the variance (4.8). This means that as both one-step ahead variance and the squared values of parameters decrease, transition matrix $\mathbf{F}$ and persistence vector $\mathbf{g}$ move towards zero. Efficiently this causes shrinkage of parameters and its intensity increases with increasing forecast horizon h. This finding helps explain results in the literature so far. For example, the results by Tiao and Xu (1993), who find using simulation that MA parameters tend towards one when MSE_h is used in the estimation (the detailed explanation of shrinkage mechanism for ARIMA models will be given later in this section). It also helps to explain the observed robustness of models estimated using MSE_h.

A similar shrinkage happens when using MSTFE:

$$\sum_{j=1}^{h} \sigma_j^2 = \sigma_1^2 \left(h + \sum_{j=2}^{h} \sum_{i=1}^{j-1} c_{i,j}^2 \right). \tag{4.9}$$

The parameters in (4.9) interact with the variance directly and the size of the sum in the right-hand side is at least h times greater than in the simpler case of (4.8). But at the same time the one step ahead variance in (4.9) is multiplied by h. So due to the presence of variances of short-term forecast errors in (4.9), the shrinkage effect in MSTFE is less intensive than in MSE_h, when the forecast horizon increases. But still the parameters will overshrink with the substantial increase of forecast horizon.

In order to see what exactly happens with different state-space models when different estimators are used, we examine several cases. In particular we discuss three model families: exponential smoothing, ARIMA and regression. Hereafter we skip the case of $h = 1$ as trivial, because no shrinkage is imposed on the parameters in this case.

4.2.1 Exponential smoothing

All of the exponential smoothing models (ETS) that we discuss here use Hyndman et al. (2008b) taxonomy. We examine additive ETS models only, keeping in mind that the main findings can be extended to multiplicative cases as well.

We first discuss ETS(A,N,N), which is also known as the local level model. It has $c_{j,h} = \alpha$, where α is smoothing parameter for level component of time series. This makes h-steps ahead variance equal to:

$$\sigma_h^2 = \sigma_1^2 \left(1 + \sum_{j=1}^{h-1} \alpha^2 \right). \tag{4.10}$$

Equation (4.10) shows that using any multiple steps ahead estimator leads to the shrinkage of smoothing parameter α in ETS(A,N,N) towards zero. Furthermore the speed of shrinkage increases with the increase of the forecast horizon because

the sum in the right hand side of (4.10) becomes greater. This also means that using multiple steps ahead estimators with long term forecasts leads to more uniform weights distribution in time for ETS(A,N,N). This corresponds to a model with the slower level changes that is less reactive to new information as α tends to zero. Asymptotically with the increase of h the model becomes a global level model.

Another commonly used model is ETS(A,A,N), also known as local trend model. In this case $c_{j,h} = \alpha + \beta j$:

$$\sigma_h^2 = \sigma_1^2 \left(1 + \sum_{j=1}^{h-1} (\alpha + \beta j)^2 \right), \tag{4.11}$$

where β is smoothing parameter for trend component of ETS.

The shrinkage effect in (4.11) is preserved but has a different form: both smoothing parameters shrink, but β shrinks faster than α with the increase of h. This leads to more stable estimates of trend and allows to capture long term tendencies in time series. Asymptotically, with the increase of horizon the model becomes a deterministic trend model.

The other model of interest is "damped-trend" ETS(A,Ad,N), which has been argued in the literature to be robust and performed well in several practical competitions (Makridakis and Hibon, 2000). It has a small difference in $c_{j,h}$ value in comparison to ETS(A,A,N): $c_{j,h} = \alpha + \beta \sum_{i=1}^{j} \phi^i$, which leads to:

$$\sigma_h^2 = \sigma_1^2 \left(1 + \sum_{j=1}^{h-1} (\alpha + \beta \sum_{i=1}^{j} \phi^i)^2 \right). \tag{4.12}$$

The damping parameter ϕ in (4.12) is usually constrained within bounds [0, 1]. This means that the second sum in the right hand side of (4.12) always satisfies the inequality:

$$0 \leq \sum_{i=1}^{j} \phi^i \leq j \tag{4.13}$$

As a result parameter ϕ slows down the shrinkage of β in comparison to ETS(A,A,N). In other words ϕ controls the speed of shrinkage of β. In cases when it is close to

zero, β may even not shrink at all. But when it is close to one, all the processes have the same features as ETS(A,A,N) model. This also gives a different interpretation of the observed good performance of ETS(A,Ad,N) over ETS(A,A,N) (Gardner and McKenzie, 2011). Asymptotically with the increase of the horizon, ETS(A,Ad,N) becomes a model with a deterministic non-linear trend.

It is also important to note that the shrinkage in ETS(A,Ad,N) is not controlled only by ϕ. It may happen by either β or ϕ. Also note that if one shrinks it is not necessary for the other one to shrink as well.

Finally we examine local level seasonal model ETS(A,N,A), which has $c_{j,h} = \alpha + \gamma j_m$, where $j_m = \lfloor \frac{j-1}{m} \rfloor$ and m is the frequency of the data. This is also a very commonly used ETS model and is interesting because it introduces a seasonal component. The h-steps ahead variance in this case is:

$$\sigma_h^2 = \sigma_1^2 \left(1 + \sum_{j=1}^{h-1} (\alpha + \gamma j_m)^2 \right), \tag{4.14}$$

where γ is smoothing parameter for seasonal component of ETS.

Due to the values of j_m, if forecast horizon is less than the frequency of the data (for example, 9 steps ahead in case of monthly data, where $m = 12$), then $j_m = 0$. As a result γ does not shrink at all. As the forecasting horizon increases, γ starts to shrink towards zero, but this happens in a stepwise manner, increasing as complete seasonal cycles increase. Parameter α shrinks faster than γ on smaller horizons, but as h increases, this is inverted and γ shrinks faster. This happens because γ has higher weight corresponding to j_m with higher horizons. The switch in the shrinkage speed happens when $h > 2m$, because γ starts prevailing in the sum in (4.14). Once again asymptotically the model becomes deterministic when $h \to \infty$.

Other additive state-space exponential smoothing models demonstrate behaviour similar to the ones discussed above. For example the shrinkage mechanism in ETS(A,Ad,A) has features of both ETS(A,Ad,N) and ETS(A,N,A).

It is important to note that models with the multiplicative errors have $c_{j,h}$ values similar to the ones already discussed. So the properties described for additive ETS models can be easily transferred to multiplicative ones.

Finally, it should be noted that the initial states of any ETS model influence the variance of the one-step ahead forecast error, so the shrinkage of the smoothing parameters is compensated by the change of the initial states. But the initial states do not shrink and their behaviour depends mainly on time series characteristics.

4.2.2 ARIMA

Due to Snyder (1985) any ARIMA model can be represented in state-space form. In order to preserve the same logic as with moving average components in ETS, we use the following polynomials for MA:

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \dots \tag{4.15}$$

This leads to the following components of the state-space model:

$$\mathbf{w} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \mathbf{F} = \begin{pmatrix} \eta_1 & \mathbf{I}_{k-1} \\ \vdots & \\ \eta_k & 0 \end{pmatrix}, \mathbf{g} = \begin{pmatrix} \eta_1 + \theta_1 \\ \vdots \\ \eta_k + \theta_k \end{pmatrix}, \tag{4.16}$$

here η_i is AR polynomial and θ_i is i^{th} moving average parameter. The variance of ARIMA can now be estimated in a manner similar to ETS. The analysis of (4.16) shows that in general AR parameters should shrink towards zero, as the transition matrix gets exponentiated with each step $j = 1, .., h$, while MA parameters should shrink towards minus AR parameters. Furthermore with the increase of forecasting horizon the shrinkage of AR parameters happens faster than the shrinkage of MA parameters, because of the exponentiation of transition matrix $\mathbf{F}$ in formula (4.8).

Analysing more specific models gives a better understanding of the shrinkage mechanism in ARIMA when any of multiple steps estimator is used.

For example ARIMA(1,1,1) can be represented in the state-space form (4.3) where:

$$\mathbf{w} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \mathbf{F} = \begin{pmatrix} 1+\phi_1 & 1 \\ -\phi_1 & 0 \end{pmatrix}, \mathbf{g} = \begin{pmatrix} 1+\phi_1+\theta_1 \\ -\phi_1 \end{pmatrix}. \tag{4.17}$$

The matrix $\mathbf{F}$ is exponentiated in the power of $j-1$ when h-steps ahead variance is calculated. The transition matrix decomposition shows (see Appendix H) that $c_{j,h}$ in this case is equal to:

$$c_{j,h} = 1 + (\phi_1 + \theta_1) \sum_{i=1}^{j} \phi_1^{i-1}. \tag{4.18}$$

Substituting the value (4.18) into the formula of h-steps ahead variance results in the following:

$$\sigma_h^2 = \sigma_1^2 \left(1 + \sum_{j=1}^{h-1} \left(1 + (\phi_1 + \theta_1) \sum_{i=1}^{j} \phi_1^{i-1} \right)^2 \right). \tag{4.19}$$

Now it can be concluded that the minimisation of (4.19) leads to the shrinkage of both AR and MA parameters: the sum of polynomials in the right hand side of (4.19) shrinks towards zero, implying the shrinkage of ϕ_1 towards zero as well. The speed of shrinkage of ϕ_1 increases with the increase of the forecasting horizon due to increase of number of elements in the sum of polynomials. At the same time the sum $(\phi_1 + \theta_1)$ also shrinks towards zero, meaning that θ_1 becomes closer to $-\phi_1$. This can have two different directions depending on the data: either θ_1 moves towards $-\phi_1$ or ϕ_1 moves towards $-\theta_1$.

Note that for $j \to \infty$ the second sum of right-hand side of (4.19) is equal to sum of infinite geometric progression, which converges when $|\phi_1| < 1$:

$$(\phi_1 + \theta_1) \sum_{i=1}^{\infty} \phi_1^{i-1} = \frac{\phi_1 + \theta_1}{1 - \phi_1}. \tag{4.20}$$

The elements in the first sum in (4.19) then asymptotically will be:

$$\left(1 + (\phi_1 + \theta_1) \sum_{i=1}^{\infty} \phi_1^{i-1} \right)^2 = \left(\frac{1+\theta_1}{1-\phi_1} \right)^2. \tag{4.21}$$

Therefore as the sum (4.21) tends towards zero, θ_1 shrinks towards -1. So asymptotically with $h \to \infty$ ARIMA(1,1,1) transforms into ARIMA(0,2,1) model with $\theta_1 = -1$.

In order to understand what this means in forecasting, we need to analyse the compact form of ARIMA(1,1,1):

$$(1-B)(1-\phi_1 B)y_t = (1+\theta_1 B)\epsilon_t. \tag{4.22}$$

Taking that asymptotically $\phi_1 = 1$ and $\theta_1 = -1$, the formula (4.22) transforms into:

$$(1-B)^2 y_t = (1-B)\epsilon_t. \tag{4.23}$$

Other ARIMA models can be analysed in a similar manner. Unfortunately the variety of ARIMA models is very wide and it is impossible to discuss all of them.

4.2.3 Regression models

Any regression model can also be represented in a state-space form. For example the multinomial regression:

$$y_t = a_0 + a_1 x_{1,t} + \cdots + a_{k-1} x_{k-1,t} + \epsilon_t, \tag{4.24}$$

where a_j is j^{th} coefficient of the model and $x_{j,t}$ is an exogenous variable.

The regression (4.24) can be split into measurement and transition equations (Hyndman et al., 2008b, p.138):

$$\begin{cases} y_t = \mathbf{w}_t' \mathbf{v}_{t-1} + \epsilon_t \\ \mathbf{v}_t = \mathbf{F}\mathbf{v}_{t-1} + \mathbf{g}\epsilon_t \end{cases}, \tag{4.25}$$

where

$$\mathbf{w}_t = \begin{pmatrix} 1 \\ x_{1,t} \\ \vdots \\ x_{k-1,t} \end{pmatrix}, \mathbf{v}_t = \begin{pmatrix} a_{0,t} \\ a_{1,t} \\ \vdots \\ a_{k-1,t} \end{pmatrix}, \mathbf{F} = I_k, \mathbf{g} = \mathbf{0}. \tag{4.26}$$

In the state-space form of regression the coefficients do not change in time. As a result all the values of the persistence vector are equal to zero, which in its turn means that $c_{j,h} = 0$ for any values of h. So applying any multiple steps estimator for regression model with exogenous variables only does not lead to shrinkage of parameters.

As we see in this paragraph the shrinkage that happens in multiple steps estimators means that models become more robust and persistent. Asymptotically, with the increase of h, any time series model estimated using a multiple steps estimator becomes deterministic. And if the forecast horizon is very high, then the parameters may become biased by overshrinking. They may take many more observations to converge to true values in comparison with the conventional MSE as estimator. The parameters are still asymptotically consistent (Weiss, 1991) but may revert to zeroes in cases of finite samples and high values of h. So in general it maybe not be appropriate to use the multiple steps estimators, especially when the sample size is small and the forecast horizon is high. This is especially important when the parameters of the model are the main interest.

4.2.4 Complex Exponential Smoothing

Complex Exponential Smoothing has the following underlying statistical state-space model:

$$\begin{cases} y_t = l_{t-1} + \epsilon_t \\ l_t = l_{t-1} - (1-\alpha_1)c_{t-1} - \alpha_1 p_t + \alpha_0 \epsilon_t \\ c_t = l_{t-1} + (1-\alpha_0)c_{t-1} + \alpha_0 p_t + \alpha_1 \epsilon_t \end{cases}, \tag{4.27}$$

which can be rewritten in the compact form:

$$\begin{cases} y_t = w' v_{t-1} + \epsilon_t \\ v_t = F v_{t-1} + q p_t + g \epsilon_t \end{cases}, \tag{4.28}$$

where $F = \begin{pmatrix} 1 & -(1-\alpha_1) \\ 1 & 1-\alpha_0 \end{pmatrix}$, $g = \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}$, $q = \begin{pmatrix} -\alpha_1 \\ \alpha_0 \end{pmatrix}$ and $w = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

The variance of this model depends on the selected information potential, but in general for h-steps ahead, where $h > 1$ can be written as:

$$\begin{aligned} \sigma_h^2 = \sigma_1^2 \left(1 + \sum_{j=1}^{h-1} (w' F^{h-1-j} g)^2\right) + V(p_t) \left(\sum_{j=1}^{h-1} (w' F^{h-1-j} q)^2\right) \\ + cov(p_t, \epsilon_t) \left(\sum_{i=1}^{h-1} \sum_{j=1}^{h-1} (w' F^{h-1-j} g)(w' F^{h-1-i} q)\right). \end{aligned} \tag{4.29}$$

Analysing formula (4.29), it can be concluded that by applying any multiple steps ahead estimation method, several things happen in CES:

1. Both variances of error term and information potential decrease;

2. Covariance between error term and information potential decreases as well;

3. Transition matrix and persistence vector shrink.

Now the transition matrix F has the following eigenvalues (2.22):

$$\lambda = \frac{2 - \alpha_0 \pm \sqrt{\alpha_0^2 + 4\alpha_1 - 4}}{2}. \tag{4.30}$$

These eigenvalues in a power of $h-1$ will have complicated interactions between real and imaginary parts of the complex smoothing parameter. For example the square of one of the eigenvalues (which corresponds to $h=3$) is equal to:

$$\lambda_1^2 = 0.5\alpha_0^2 - \alpha_0 + \alpha_1 + 2(1 - 0.5\alpha_0)\sqrt{0.25\alpha_0^2 + \alpha_1 - 1}, \tag{4.31}$$

while for the second it is:

$$\lambda_2^2 = 0.5\alpha_0^2 - \alpha_0 + \alpha_1 - 2(1 - 0.5\alpha_0)\sqrt{0.25\alpha_0^2 + \alpha_1 - 1}. \tag{4.32}$$

Minimising the values (4.31) and (4.32) even for such a short horizon means that although the parameters of CES shrink, this process happens in a non-linear fashion: parameters will move towards zero with the increase of h, but at the same time they will not overshrink because of the negative signs before the parameters in (4.31) and (4.32).

Still taking the power $h-1$ of each eigenvalue of the transition matrix F means that during the minimisation they will shrink towards zero. As a result CES with the multi-steps ahead cost functions will be forced towards stationarity. Unfortunately, there is no way to demonstrate this mechanism in a simpler and more clear way.

Summarising this subsection, CES can be efficiently estimated using simpler multi-steps ahead cost functions, such as MSTFE or MSE_h, because the parameters of CES will not overshrink. Using these functions will force CES to become stationary for higher horizons.

In order to address this issue we propose a different multiple steps estimator based on a likelihood function.

4.3 Trace forecast likelihood

The idea of trace forecast likelihood is based on the analysis of the joint conditional distribution of multiple steps ahead (1 to h) observed values, represented as a vector:

$$\mathbf{Y}_t = \begin{pmatrix} y_{t+1} \\ y_{t+2} \\ \vdots \\ y_{t+h} \end{pmatrix}. \tag{4.33}$$

This is a very universal approach because values of $\mathbf{Y}_t$ depend only on the underlying statistical model, which is not defined directly here. In the general case the likelihood that a model with some parameters produced such a vector of actual values can be estimated using the following conditional probability:

$$L(\theta, \mathbf{\Sigma}|\mathbf{Y}_t) = P(\mathbf{Y}_t|\theta, \mathbf{\Sigma}), \tag{4.34}$$

where θ is a vector of parameters of a statistical model and $\mathbf{\Sigma}$ is a conditional covariance matrix for variable $\mathbf{Y}_t$.

The same likelihood estimated on all the available observations T is equal to the product of the following joint probabilities:

$$L(\theta, \mathbf{\Sigma}|\mathbf{Y}) = \prod_{t=1}^{T} P(\mathbf{Y}_t|\theta, \mathbf{\Sigma}), \tag{4.35}$$

where $\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_T\}$.

Although the elements of vector $\mathbf{Y}_t$ are in general not independent from each other for different t, the maximisation of the likelihood (4.35) gives an estimate for θ, that is at least weakly consistent, first order efficient and asymptotically normally distributed (Heijmans and Magnus, 1986a,b). This means that using the likelihood (4.35) automatically guarantees that several important statistical properties of the estimates of θ hold.

The final value of the likelihood (4.35) depends mainly on the assumptions on the conditional distribution of $\mathbf{Y}_t$. Traditionally in time series literature the normal distribution is used, which in the multidimensional case leads to the following quasi-likelihood function:

$$L(\theta, \mathbf{\Sigma}|Y) = \prod_{t=1}^{T} \left[(2\pi)^{-\frac{h}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \mathbf{E_t'} \mathbf{\Sigma}^{-1} \mathbf{E_t} \right) \right] \tag{4.36}$$

where $\mathbf{E_t} = \begin{pmatrix} e_{t+1|t} \\ e_{t+2|t} \\ \vdots \\ e_{t+h|t} \end{pmatrix}$.

The covariance matrix $\mathbf{\Sigma}$ is unknown in (4.36) but can be estimated using maximum likelihood:

$$\hat{\mathbf{\Sigma}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{E_t} \mathbf{E_t'} \tag{4.37}$$

Linearising the function (4.36) and using the estimated covariance matrix (4.37) leads to the following concentrated log-likelihood:

$$\ell(\theta, \hat{\mathbf{\Sigma}}|\mathbf{Y}) = -\frac{T}{2} \left(h \log(2\pi) + \log |\hat{\mathbf{\Sigma}}| \right) - \frac{1}{2} \sum_{t=1}^{T} \left(\mathbf{E_t'} \hat{\mathbf{\Sigma}}^{-1} \mathbf{E_t} \right) \tag{4.38}$$

It can now be shown that (4.38) is equivalent to even simpler function (see Appendix I for the details):

$$\ell(\theta, \hat{\mathbf{\Sigma}}|\mathbf{Y}) = -\frac{T}{2} \left(h \log(2\pi e) + \log |\hat{\mathbf{\Sigma}}| \right) \tag{4.39}$$

By maximising function (4.39) we guarantee that the obtained estimates of θ will have all the desirable statistical properties of likelihood estimates, in contrast with the other multiple steps estimators. Furthermore function (4.39) can be used in model selection with information criteria. The rationale for this will be similar to the one used in Akaike (1974) with the following difference: the comparison between true and estimated models is done using multiple steps ahead forecasts instead of one-step ahead forecasts.

Analysing (4.39) shows that maximisation of concentrated log-likelihood is equivalent to minimisation of the determinant of the covariance matrix $\hat{\mathbf{\Sigma}}$. This determinant is called "Generalised Variance":

$$GV = |\hat{\mathbf{\Sigma}}| \tag{4.40}$$

For convenience of estimation, in cases of large h, the logarithm of (4.40) may be taken.

It is essential to understand what elements the covariance matrix $\mathbf{\Sigma}$ contains and what minimisation of its determinant means. The matrix has the following structure:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,h} \\ \sigma_{1,2} & \sigma_2^2 & \dots & \sigma_{2,h} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,h} & \sigma_{2,h} & \dots & \sigma_h^2 \end{pmatrix}, \tag{4.41}$$

where $\sigma_{i,j} = cov(e_{t+i|t}, e_{t+j|t})$ is the covariance between i-th and j-th steps ahead conditional errors. Note that these covariances in general differ from the conventional covariances between error terms ϵ_t and ϵ_{t+i} for any integer i because the former are errors conditional to some actual values on observation t, while the latter are unconditional errors.

Knowing the structure (4.41) it is easy to show that the function (4.40) can be represented in the following simpler way (see Appendix J for the derivation):

$$GV = |\mathbf{\Sigma}| = \prod_{j=1}^{h} \sigma_j^2 |\mathbf{R}|, \tag{4.42}$$

where $\mathbf{R} = \begin{pmatrix} 1 & r_{1,2} & \dots & r_{1,h} \\ r_{1,2} & 1 & \dots & r_{2,h} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1,h} & r_{2,h} & \dots & 1 \end{pmatrix}$ is the correlation matrix of conditional errors.

Linearising (4.42) leads to a function that is much easier to analyse:

$$\log GV = \sum_{j=1}^{h} \log \sigma_j^2 + \log |\mathbf{R}|. \tag{4.43}$$

We conclude that with the maximisation of likelihood (4.39) the logarithms of variances of 1 to h steps ahead errors are decreased and correlations between some of the elements of $\mathbf{R}$ are increased, ensuring that the determinant of matrix $\mathbf{R}$ becomes closer to zero.

The advantage of estimator (4.43) in comparison with (4.1) and (4.2) is that the variances in the former are scaled using the logarithm function. This means that the higher variances that are typical for long-term forecasts do not dominate the variances of short term during estimation. This effect was noted and studied by Taieb and Atiya (2016). Taking logarithms of these variances makes them closer to each other which in its turn allows minimising variances across all the horizons.

Note that the trace of matrix $\mathbf{\Sigma}$ is proportional to the value of MSTFE. This value is usually called "Total Variation" and does not take into account the covariances between variables in matrix $\mathbf{\Sigma}$. Furthermore values of MSE_h for any h are proportional to the h-th element of diagonal of the matrix $\mathbf{\Sigma}$, which makes this estimator a special case of trace forecast likelihood.

Overall it can be noted that the proposed likelihood approach is more general than the estimators discussed above.

In the case of state-space models, formula (4.43) can be rewritten using (4.8) in the following manner:

$$\log GV = h \log \sigma_1^2 + \sum_{j=2}^{h} \log \left(1 + \sum_{i=1}^{j-1} c_{i,j}^2 \right) + \log |\mathbf{R}|. \tag{4.44}$$

By minimising GV the variance of one-step ahead forecast error is decreased along with values of parameters. So this cost function also imposes shrinkage on parameters, but because of the logarithms in (4.44) the one-step ahead variance value is balanced out with the sum of parameters. This means that the shrinkage effect in GV is lower than in MSTFE.

Nevertheless function (4.44) has an important element that all the other functions lack; the determinant of the correlation matrix $\mathbf{R}$. In order to understand how this

matrix contributes to the final value, we need to estimate values of covariances $\sigma_{i,j}$ and calculate correlation coefficients $r_{i,j}$.

Using the same assumptions of no autocorrelation and homoscedasticity in error term, taking that $j > i$, the covariance $\sigma_{i,j} = cov(e_{t+j|t}, e_{t+i|t})$ can be derived using (4.6) and (4.7). The correlation coefficients then are equal to (see Appendix K for the details):

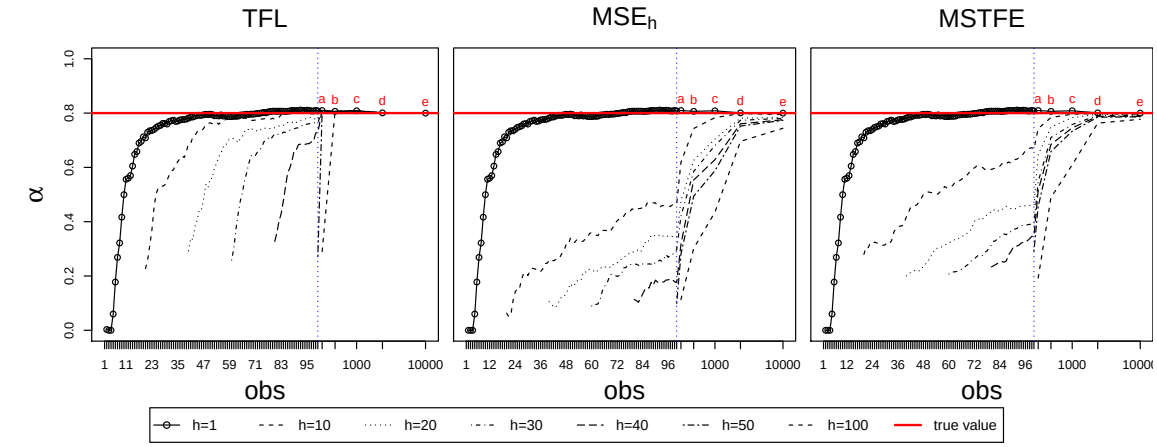
$$r_{i,j} = \begin{cases} \dfrac{\left(c_{i,j} + \sum_{l=1}^{i-1} c_{l,j}c_{l,i}\right)}{\sqrt{\left(1 + \sum_{l=1}^{i-1} c_{l,i}^2\right)\left(1 + \sum_{l=1}^{j-1} c_{l,j}^2\right)}} & \text{when } i, j \neq 1 \\ \dfrac{c_{1,j}}{\sqrt{\left(1 + \sum_{l=1}^{j-1} c_{l,j}^2\right)}} & \text{when } i = 1 \end{cases}. \tag{4.45}$$

The analysis of (4.45) shows that when i and j are close to each other then correlation between them increases with the increase of horizon, because the numerator in this case becomes close in value to the denominator. Secondly, the higher values of i and j in general also lead to increase of correlations. Asymptotically, when $i, j \to \infty$, correlation $r_{i,j} \to 1$, which leads to a singular matrix $\mathbf{R}$.

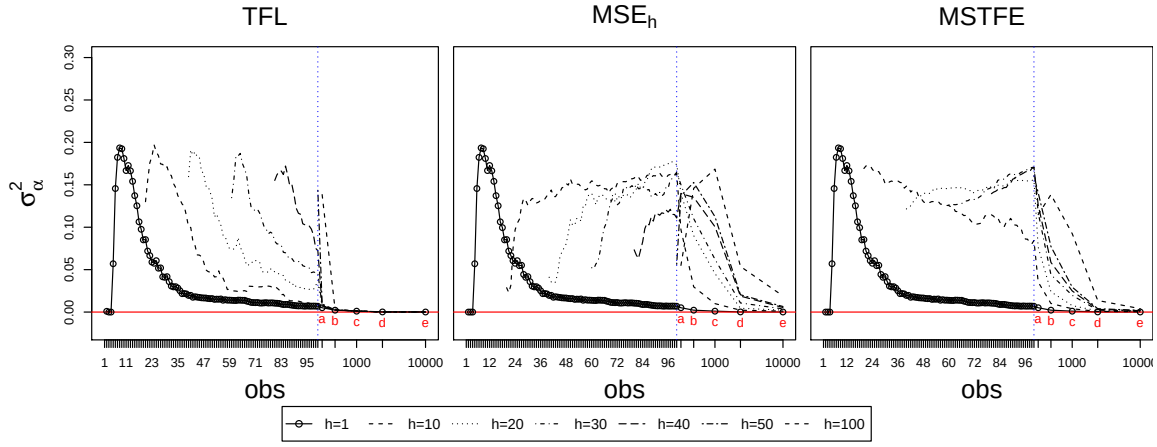
Vice versa the increase of correlation between some i and j steps ahead errors means that the numerator and the denominator in (4.45) are enforced to become closer to each other. This slows down the general shrinkage effect because the value $c_{i,j} = \mathbf{w}'\mathbf{F}^{j-i-1}\mathbf{g}$ in numerator tends towards 1, which is possible only when $\mathbf{F}$ and $\mathbf{g}$ do not shrink towards zero. Unfortunately, it is impossible to show the exact effect of $|\mathbf{R}|$ on GV in a general case because the determinant of $\mathbf{R}$ is hard to calculate analytically for $h > 3$.

Overall it can be concluded that the proposed estimator should be less biased and more efficient than the conventional multiple steps ahead estimators. The rate of convergence of parameters estimates to true values in this case is also higher than

for the other multiple steps ahead estimators due to the weakened shrinkage effect. In the next section we demonstrate these with a simulation experiment. We conduct simulations for both ETS and ARIMA models, the parameters of which are estimated using multiple steps functions.



(a) Mean values.



(b) Variance.

Figure 4.1: Correct model. Parameter α.

4.4 Simulation experiment

4.4.1 ETS

We conduct a simulation experiment in order to see the behaviour of different estimators depending on the sample size and the horizon. 100 time series with 10000

observations each are generated using ETS(A,N,N) model, with an arbitrary chosen smoothing parameter $\alpha = 0.8$, and then correct and wrong models are fit to the data using the estimators with forecasting horizon of 1, 10, 20, 30, 40, 50 and 100 on samples containing from 10 to 100 (with 1 observation step) and then 200, 500, 1000, 5000 and 10000 observations. The resulting estimated model parameters are recorded and analysed.

The data is simulated using "sim.es" function of package "smooth" for R (available from CRAN). The estimation of models is done using "es" function of the same package.

First we study the case when the true model is known. Figure 4.1a shows how parameters change when the sample size increases for different forecasting horizons used in the estimation with the different methods discussed in this paper: trace forecast likelihood (entitled as "TFL"), MSE_h and MSTFE. The case of $h = 1$ corresponds to the conventional method based on one-step-ahead forecast error. The vertical line splits samples with 10 to 100 observations from the samples with larger numbers of observations marked with the following letters: "a" – 200, "b" – 500, "c" – 1000, "d" – 5000 and "e" – 10000. The horizontal line represents the true parameter.

Figure 4.1a demonstrates that TFL gives consistent estimates although with a bias on smaller samples. This means that the method can be effectively used when dealing with large samples of data. On the contrary using both MSE_h and MSTFE leads to overshrinkage of parameters with high inefficiency and bias even on larger samples. For example, even when the sample size is 5000 observations MSE_h with $h = 100$ still produces biased estimates of the parameter. Note that MSTFE produces less biased estimates than MSE_h. Eventually the parameters estimated using both of these methods converge to the true value, but the speed of convergence is much lower than for TFL because of the uncontrolled shrinkage effect, the estimates reach true values only asymptotically. Furthermore the parameters converge to true value with

a speed proportional to $\frac{T}{h}$ with obvious advantage of TFL.

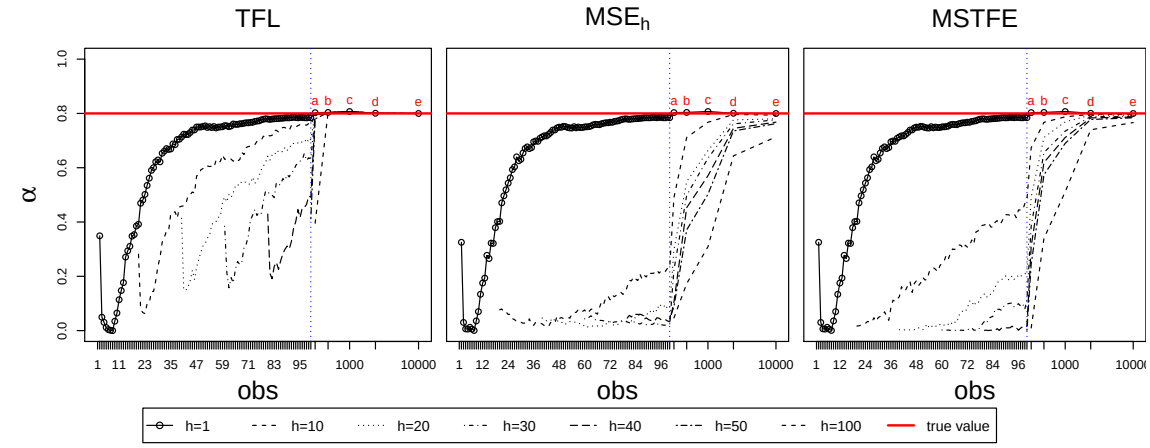
The Figure 4.1b demonstrates the convergence of variance of the same parameter. TFL estimator is inefficient for the sample sizes of approximately $T < 2h$, but it becomes efficient for larger samples. Both MSTFE and MSE_h demonstrate inefficiency even for larger samples. While the efficiency of MSTFE slowly increases (when T becomes approximately more than $10 \cdot h$), the MSE_h estimator is highly inefficient even on large samples (approximately $20 \cdot h$).

All of these results hold when the model is correctly specified. We proceed to analyse the behaviour of estimators when the wrong model is selected, which is a more realistic case in practice. In our case we use ETS(A,A,N).

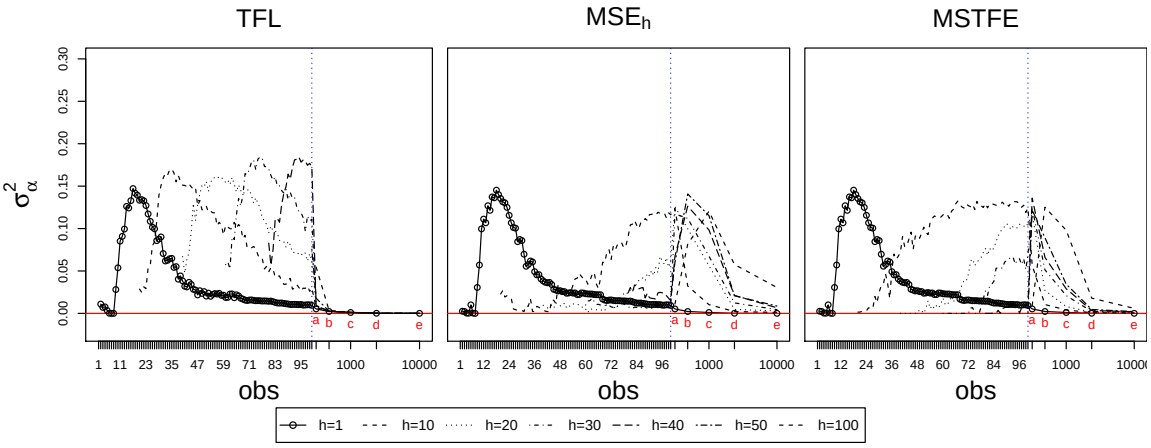
Figure 4.2a demonstrates the behaviour of parameter α of ETS(A,A,N) model in the same conditions discussed above. Note that the estimates are biased for all the methods but eventually converge to the true values. Still TFL has a faster rate of convergence than the other methods, although the behaviour of the smoothing parameter in this case is peculiar: it firstly diverges from, but then converges to $\alpha = 0.8$. Similar behaviour is demonstrated by the conventional method of estimation. This shows the connection between the two and also demonstrates that in order to have a precise estimates of the parameters using TFL, more observations are needed, than for the conventional method. This can be seen for lines with $h > 1$ - the estimates converge to true values slower than in case of MSE.

In the same situation MSE_h and MSTFE demonstrate behaviour similar to Figure 4.1a, when the correct model was selected. The bias of the parameter in this case is even greater, due to the same overshrinkage effect.

Furthermore the variance of α in this part of the experiment demonstrates that all the methods are inefficient each in its own way (Figure 4.2b). The inefficiency of TFL firstly grows with the growth of the sample size but then decreases for each of the horizons. MSE_h and MSTFE at the same time are more efficient on smaller samples,



(a) Mean values.

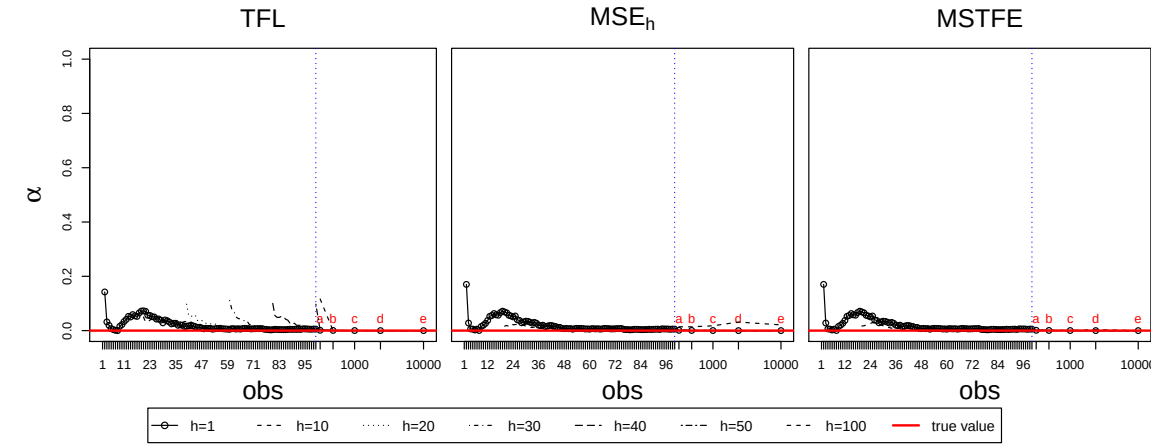


(b) Variance.

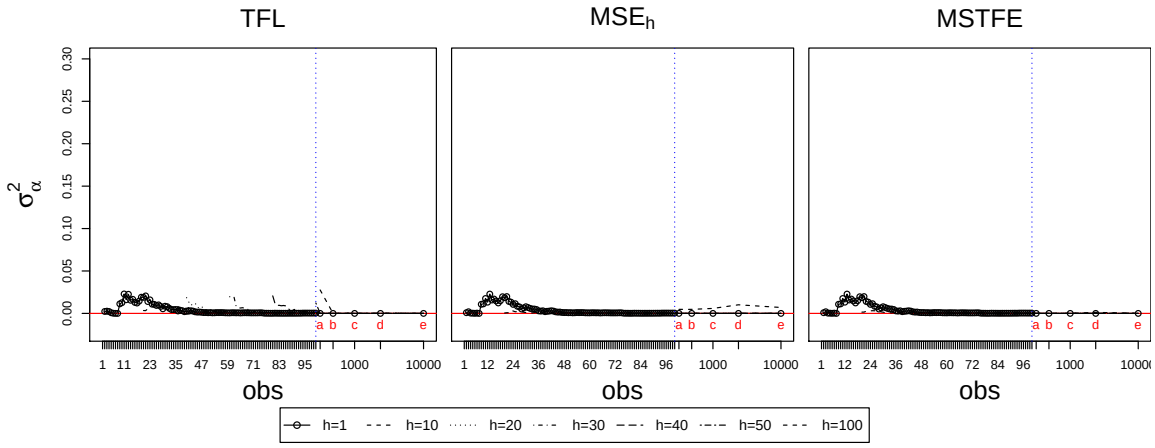
Figure 4.2: Wrong model. Parameter α.

than on larger ones. This interesting feature can probably be explained again by the shrinkage effect: on smaller samples it has a higher effect, than on bigger samples, which causes the parameter to become very close to zero.

Finally the mean values of parameter β are shown in Figure 4.3a. As we see due to the shrinkage the correct value of $\beta = 0$ is forced on the model for all estimation methods. MSE$_h$ and MSTFE are very efficient in this situation, although TFL demonstrates bias on samples with a small number of observations. The variances of the parameter β are shown in Figure 4.3b. Note that the conventional method is inefficient on smaller samples, TFL demonstrates small efficiency when the ratio $\frac{T}{h}$ is



(a) Mean values.



(b) Variance.

Figure 4.3: Wrong model. Parameter β.

small (about less than 2), but both MSE$_h$ and MSTFE give efficient estimators in this situation. This is once again because of the shrinkage effect that sets the redundant parameter to zero very fast.

Concluding the results of the experiment, it should be noted that in general TFL allows to obtain more efficient, less biased and consistent estimates, but the main advantage of the method appears only when a number of observations is sufficient (approximately $T > 2h$). MSE$_h$ and MSTFE are biased and inefficient due to the overshrinkage effect. They are still consistent, but the rate of convergence for these methods is very low. These conclusions hold for both cases of correctly and wrongly

specified models. Still when the model is wrong, due to the shrinkage effect of MSE_h and MSTFE, the redundant parameters become equal to zero much faster than using the conventional method or TFL. TFL shrinks the parameters as well, but it takes more observations to reach the same effect as other estimators.

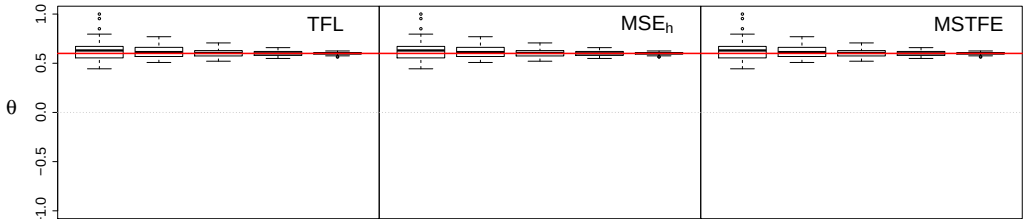
4.4.2 ARIMA

We conduct a similar experiment using ARIMA(0,1,1) as the DGP: we produce 100 series consisting of 5000 observations each with $\theta_1 = 0.6$. We then fit correct and wrong models to the simulated data using the three estimators and horizons of 1, 10, 20, 30, 40 and 50. The wrong model is ARIMA(1,1,1).

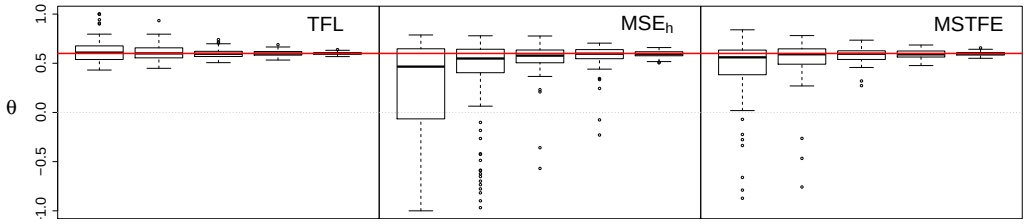
The results of this simulation in general correspond to the results above. In order to explore them in detail, we have produced several boxplots for each of the methods. Figure 4.4 demonstrates how the estimated parameters vary with different sample sizes and different forecast horizons. TFL gives consistent, efficient and unbiased estimates of the parameters. In contrast MSE_h and MSTFE are highly inefficient and biased. In many cases they even produce negative values for MA parameter, although the correct value is 0.6. This is due to the overshrinkage problem. As we have shown in the previous section, this is happening in conventional multiple steps estimators.

The case of the wrong model is also interesting. Figure 4.5 shows boxplots for MA(1) parameter and how its estimates deviate from the true value. Note that in this case all the multiple steps become less efficient than in the case with the correct model, but Trace Forecast Likelihood gives the most efficient and least biased estimates.

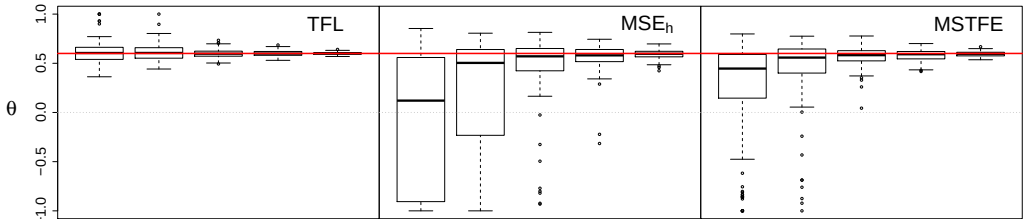
We provide a similar analysis of the estimated AR parameter in Appendix L. It confirms the results discussed in section 4.2.2. The AR(1) parameter in the simulation shrinks towards $-\theta$ when MSE_h and MSTFE are used. In the contrast TFL leads to unbiased estimates of AR(1). Note that MSTFE is more efficient and less biased than MSE_h which confirms the findings in the previous section.



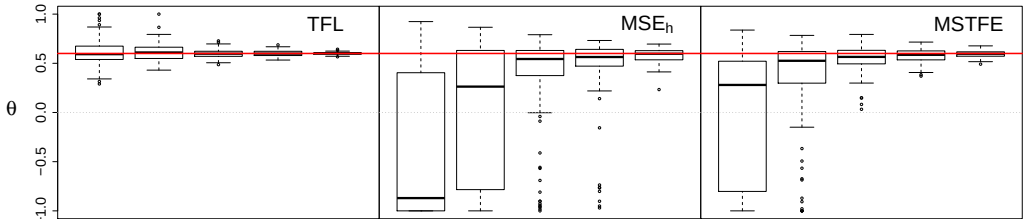
(a) h=1



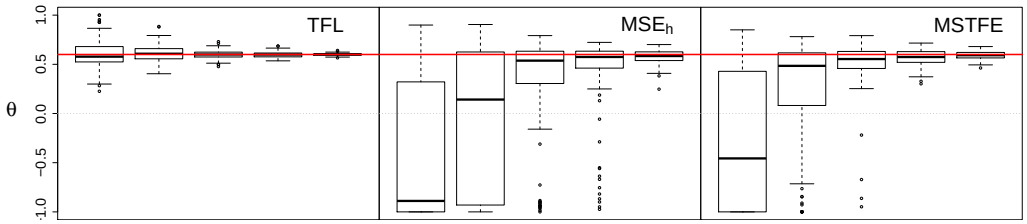
(b) h=10



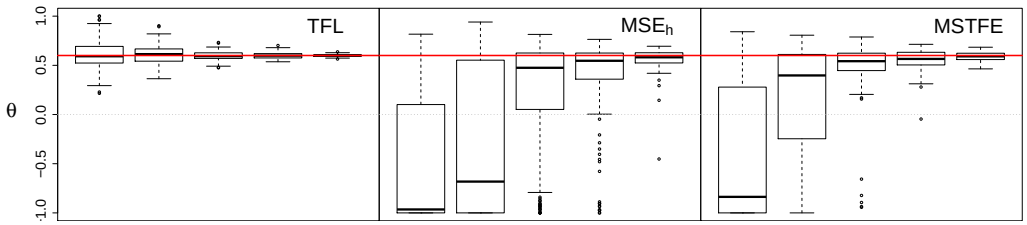
(c) h=20



(d) h=30

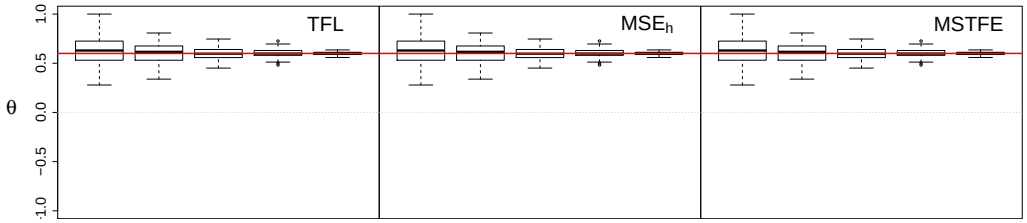


(e) h=40

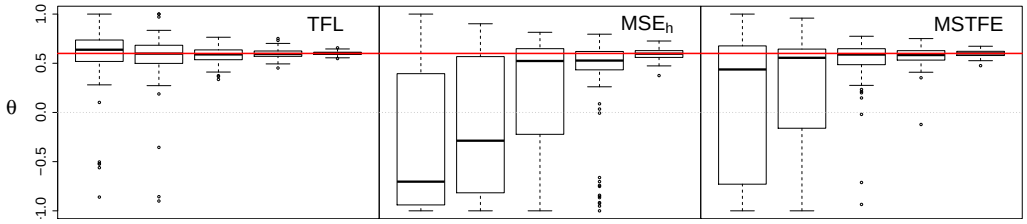


(f) h=50

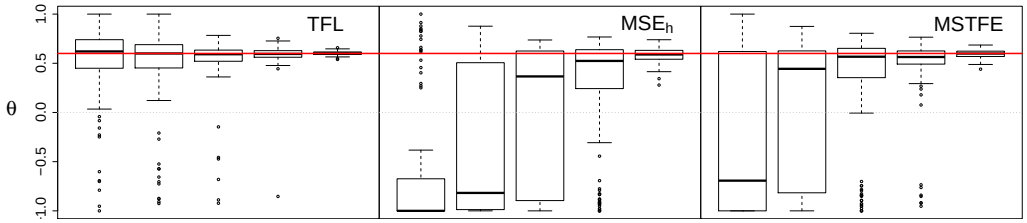
Figure 4.4: ARIMA estimated using different methods. Correct model. Boxplots of parameter θ.



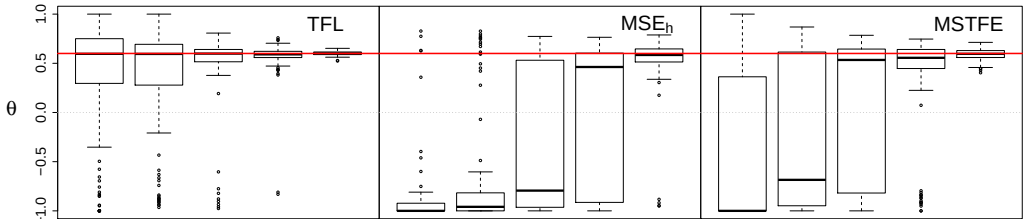
(a) h=1



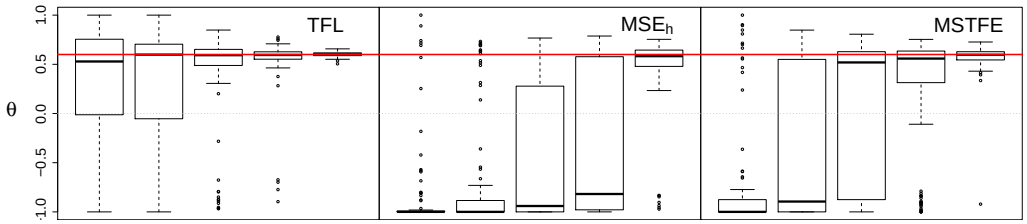
(b) h=10



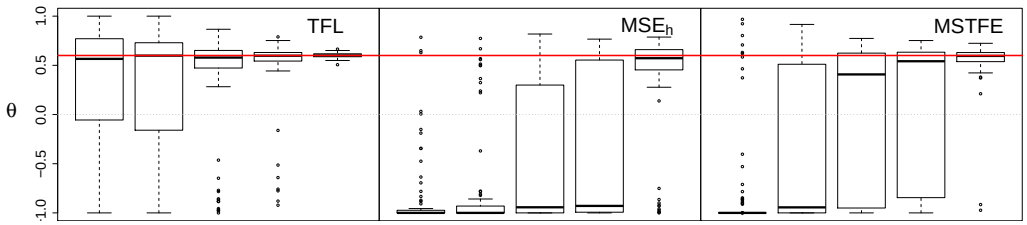
(c) h=20



(d) h=30



(e) h=40



(f) h=50

Figure 4.5: ARIMA estimated using different methods. Wrong model. Boxplots of parameter θ.

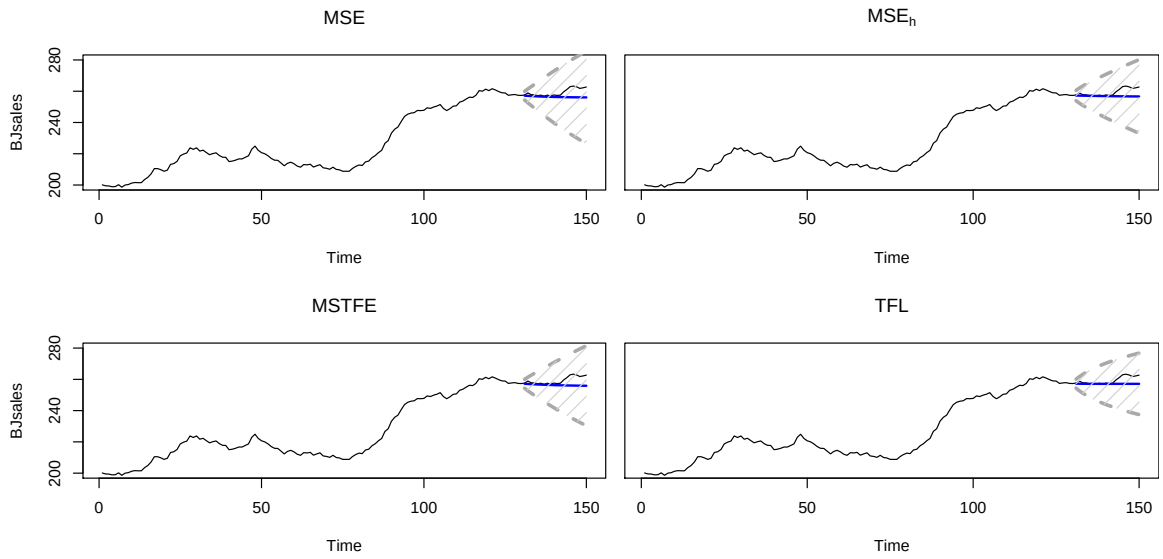


Figure 4.6: Sales series and forecasts using ARIMA(1,1,1).

4.5 Real time series example

In order to demonstrate what are the benefits of the TFL method for forecasting purpose, we fit ARIMA(1,1,1) model on Box and Jenkins sales data. The motivation here is to see how the model used for simulated data in previous section performs in real life with different estimators. Figure 4.6 demonstrates three cases of the model: (1) estimated using mean squared one-step-ahead error, (2) estimated using MSE$_h$, (3) estimated using MSTFE, and (4) estimated using TFL.

As we see from the Figure 4.6 the most accurate forecast is produced by ARIMA estimated using TFL, although the difference between the methods is slight – forecast errors for these three cases are respectively: (1) 1.13%, (2) 0.95%, (3) 1.15%, (4) 0.87%. TFL lead to the narrowest prediction intervals among the four estimators. Reason for that is values of parameters for AR and MA.

Table 4.1 demonstrates the values of parameters for the four estimators. Note that both MSE$_h$ and MSTFE once again overshrink the MA parameter, making it the lowest of the MAs of the four estimators. Note also that $\phi_1 + \theta_1$ for MSE$_h$ is the lowest (in comparison to the other estimation methods). This is because this

sum is included in the variance (4.19) and is minimised in the estimation process. If the forecast horizon would be higher, the shrinkage would be enforced, making MA to become closer to -1. Furthermore the variance of one-step-ahead forecast error is the highest for MSE_h, which indicates that due to the shrinkage of parameters, the variance was not minimised properly.

Although the one step ahead variance for MSE is the lowest among the three methods, the parameters values are still higher than they need to be. This results in the wider prediction intervals in Figure 4.6. Note also that $\phi_1 + \theta_1$ for MSE is the highest among the three methods for this case. This leads to a higher variance of h-steps ahead forecast error and as a result wide prediction intervals.

Finally TFL produced parameters closer to zero (in comparison with the other estimators) with $\phi_1 + \theta_1$ lower than in MSE. This is because of the shrinkage with the compensation that is happening during the estimation of the model. The variance of one-step-ahead error is the second lowest for TFL. These two factors balance each other out, leading to the lowest variance of h-steps ahead forecast error and as a result the prediction intervals for the model estimated with TFL appear to be the narrowest among the four estimation methods.

As we see TFL balances out the benefits of the other estimators, shrinking the parameters and producing the lower variance of one-step-ahead forecast error. This leads to more accurate forecasts and narrower prediction intervals, thus decreasing uncertainty.

Estimator	ϕ_1	θ_1	$\phi_1 + \theta_1$	σ_1^2
MSE	0.891	-0.649	0.242	1.90
MSE_h	0.938	-0.852	0.086	2.64
MSTFE	0.919	-0.763	0.156	1.99
TFL	0.632	-0.406	0.226	1.95

Table 4.1: AR and MA parameters values.

4.6 Conclusions

Estimation methods based on multiple steps ahead forecast error have been known for several decades. Their statistical properties have been studied in detail and it was found that using these estimators makes time series models more robust. Still there has never been any concise explanation of why this happens.

We have shown that the main reason for robustness is the implied shrinkage, that is automatically imposed on parameters when any of these estimators are used. We have also demonstrated that the shrinkage happens in ARIMA and ETS, but the coefficients of exogenous variables in regression do not shrink. Therefore this constitutes a univariate form of parameter shrinkage, in contrast to shrinkage methods such as Ridge or LASSO regression. Furthermore due to this effect with the increase of forecast horizon, time series models become deterministic. Thus there is a danger of overshrinking the parameters.

We propose a likelihood based on multiple steps ahead forecast errors which we called "Trace Forecast Likelihood". We showed that its maximisation implies the minimisation of generalised variance of multiple steps ahead forecast errors. Using the proposed likelihood obtains consistent, efficient and unbiased estimates of parameters and also allows the calculation of information criteria for model selection purpose. We have also demonstrated that Trace Forecast Likelihood imposes shrinkage on the parameters of models in a similar manner that other multiple steps estimators do, but it is done more gently and the parameters do not overshrink.

Using the simulation experiment we have demonstrated that the shrinkage effect in MSE_h and MSTFE worsens with the increase of the forecast horizon. Because of the "overshrinkage" effect the parameters of models become biased and inefficient, they eventually converge to the true values but with a very low rate. The speed of convergence also slows down with the increase of the forecast horizon. However the

shrinkage effect is weakened in the proposed Trace Forecast Likelihood which as it was observed, obtains more efficient, less biased and consistent estimates of parameters.

We then showed on the example of Box-Jenkins sales data how different estimation methods influence parameters of ARIMA(1,1,1) and obtain forecasts. TFL in this example resulted in more accurate forecasts with narrower prediction intervals.

Finally we conclude that if the parameters are the main interest, TFL should be preferred to MSE_h and MSTFE. However in this work we have not evaluated properly the effect of the different estimation methods for forecasting. It should be noted that the latter methods may be useful in forecasting due to their ability to decrease the influence of redundant parameters on final forecasts. Therefore future research should explore this question further.

Conclusions

Exponential smoothing is one of the well-known and effective forecasting methods that is widely used in practice. It has been shown in many papers that on average it produces more accurate forecasts than other extrapolative methods and at the same time is easier to use and explain than, for example, ARIMA model. It has been widely studied in many articles over the years from both theoretical and practical points of view and has seen substantial development since it was first introduced in 1956.

The literature review that we have conducted (chapter 1) has pointed towards interesting avenues for further research, extending the current state-of-the-art. We showed that the conventional idea of time series decomposition into "level", "trend", "seasonality" and "error" is ambiguous and that the true model (if such exists) can have more complicated structure than it is typically thought. This explains good results by various hybrid models and combinations of forecasts based on the conventional exponential smoothing model forms. Furthermore it is well understood that the true model is not known, which means that any time series model needs to be identified and estimated using appropriate methods. This motivated our research in multiple steps estimators.

Taking all of this into account we firstly proposed a new time series modelling approach based on the notion of "information potential". The idea behind it is that any time series contains the observed actual values and an unobservable information part. In chapter 2 we proposed to use the error term as an approximation of this unknown information and based on this introduced "Complex Exponential Smooth-

ing" (CES) model. CES is based on the theory of complex variables and captures the information potential into the imaginary part of a variable. This allows the CES model to produce both stationary and non-stationary forecasts depending on the data and smoothly transition between them instead of switching. This in turn obtains more accurate long-term forecasts, increasing overall accuracy in comparison to conventional ETS models. Therefore by introducing one model that approximates all possible types of trends in time series we avoid a model selection procedure and make forecasting process simpler.

We then modified the proposed model to take into account seasonality that many time series, in particular in business, exhibit. The resulting seasonal CES (chapter 3) can approximate both additive and multiplicative types of seasonality without the need to select between them. The modelling of the most appropriate type is based on the data characteristics and the value of smoothing parameters. Seasonal CES can also produce new types of seasonality, where the variance of seasonal profile may change even if the level is constant. Furthermore model selection between seasonal and non-seasonal CES was also introduced in the chapter 3. Because of the large difference in the number of parameters for these models, the model selection procedure appears to be more efficient than for ETS or ARIMA and the seasonal model is selected only when it approximates the time series significantly better than the non-seasonal counter part.

The key element of the proposed model is a complex smoothing parameter, which defines whether a stationary or non-stationary trajectory is produced and whether additive or multiplicative seasonality should be modelled. So the estimation of the parameters is essential for the model. In order to do that efficiently we propose in chapter 4 to use multiple steps ahead estimators, which produces more robust models, as it is argued in the literature. The properties of these estimators were analysed and it was shown that they impose shrinkage on the parameters of the CES models.

Furthermore, as the forecast horizon increases this leads to overshrinkage, making univariate models deterministic. We then proposed "Trace Forecast Likelihood", a function based on multiple steps ahead forecast errors and showed that the estimator is consistent, more efficient and less biased than the conventional multiple steps ahead estimators. Finally we demonstrated that the estimator can be efficiently applied to any univariate model resulting in better estimates of parameters, while retaining the useful property of shrinkage for univariate time series models.

As a result of this research, we proposed a new modelling principle which can easily be implemented in practice. CES can be efficiently used for demand forecasting and may substitute the existing exponential smoothing models used in these situations. Using only one model with a very simple mechanism of switching between seasonal and non-seasonal data would make producing of forecasts easy. Besides the good forecasting accuracy of CES also means that the uncertainty is decreased and that the forecasts can be used to support better decision making, supporting such functions as production planning, inventory management, budgeting and so on. Finally the proposed Trace Forecast Likelihood has desirable statistical properties which in practice lead to more consistent and stable forecasts with decreased sizes of the prediction intervals (due to shrinkage of smoothing parameters towards zero). This leads in turn to a further decrease of uncertainty.

There are some limitations of the proposed model. Due to the form of CES, precision of estimation is very important, but finding correct values of parameters may be a challenging task. Trace Forecast Likelihood solves this problem from fundamental perspective, but optimising the model in practice may sometimes be met with a mixed success. In some cases the surface of the objective function may become non-concave. The other limitation is in the fundamental idea behind CES – "information potential" is not a complicated notion, but it may be hard to explain to practitioners. This may cause problems in adoption of CES by companies.

Our research on the other hand opens several topics for the future works. First of all, we used only one form of information potential assuming that $p_t = \epsilon_t$. Other forms of information potential will lead to different forecasting results. For example, the approximation $p_t = \Delta y_t$ could lead to a model efficiently switching between long-term and short-term forecasts. Alternatively the information potential can be modelled using functions of exogenous variables which may lead to a further increase in forecasting accuracy.

Second, exogenous variables can be used in CES in different ways. For example, it can be included in a conventional way as in state-space ETS or it can be coded through the information potential. This can be especially useful for promotional modelling where the impact of promotions may vary in time in a non-linear fashion or to model shocks in the supply chain.

Third, it would be interesting to explore further the idea of Trace Forecast Likelihood and evaluate its impact on forecast accuracy. This will also lead to a better understanding of the effect of the univariate parameter shrinkage (that multiple step estimators impose) on the forecast accuracy of these models. Furthermore this likelihood allows the selection of a forecasting model among the set of models using information criteria, but the criteria themselves need to be updated to take the forecasting horizon into account. So additional studies of this likelihood need to be carried out.

Finally, the reseach opens up the topic of fuzzy state-space models, where the components are not defined strictly and do not correspond to some known in theory elements, such as "level", "trend" and "seasonality". It was shown that CES in state-space form has an important difference with ETS models: the transition matrix of CES includes smoothing parameters. This gives CES its main flexibility and can be exploited further. Defining any number of fuzzy components with any number of lags and estimating them along with smoothing parameters, measurement vector and

transition matrix may lead to a new class of models, which can be called "Generalised Exponential Smoothing". This class of models could underlie all the existing statistical univariate models, allowing to unite models and methods into one unified theory. Theoretical derivations, estimation and tests on the real data of this new class of models are possible directions of new research.

Appendices

A State-space form of CES

First of all any complex variable can be represented as a vector or as a matrix:

$$\begin{aligned} a + ib &= \begin{pmatrix} a \\ b \end{pmatrix}, \\ a + ib &= \begin{pmatrix} a & -b \\ b & a \end{pmatrix}. \end{aligned} \tag{46}$$

The general CES model (2.5) can be split into two parts: measurement and transition equations using (46):

$$\begin{cases} \begin{pmatrix} \hat{y}_t \\ \hat{x}_t \end{pmatrix} = \begin{pmatrix} l_{t-1} \\ c_{t-1} \end{pmatrix} \\ \begin{pmatrix} l_t \\ c_t \end{pmatrix} = \begin{pmatrix} \alpha_0 & -\alpha_1 \\ \alpha_1 & \alpha_0 \end{pmatrix} \begin{pmatrix} y_t \\ p_t \end{pmatrix} + \left(\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} - \begin{pmatrix} \alpha_0 & -\alpha_1 \\ \alpha_1 & \alpha_0 \end{pmatrix} \right) \begin{pmatrix} l_{t-1} \\ c_{t-1} \end{pmatrix} \end{cases}. \tag{47}$$

Regrouping the elements of transition equation in (47) the following equation can be obtained:

$$\begin{aligned} &\begin{pmatrix} l_t \\ c_t \end{pmatrix} = \begin{pmatrix} \alpha_0 & -\alpha_1 \\ \alpha_1 & \alpha_0 \end{pmatrix} \begin{pmatrix} 0 \\ p_t \end{pmatrix} + \begin{pmatrix} \alpha_0 & -\alpha_1 \\ \alpha_1 & \alpha_0 \end{pmatrix} \begin{pmatrix} y_t \\ 0 \end{pmatrix} + \\ &\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} l_{t-1} \\ c_{t-1} \end{pmatrix} - \begin{pmatrix} \alpha_0 & -\alpha_1 \\ \alpha_1 & \alpha_0 \end{pmatrix} \begin{pmatrix} 0 \\ c_{t-1} \end{pmatrix} - \begin{pmatrix} \alpha_0 & -\alpha_1 \\ \alpha_1 & \alpha_0 \end{pmatrix} \begin{pmatrix} l_{t-1} \\ 0 \end{pmatrix}. \end{aligned} \tag{48}$$

Grouping vectors of actual value and level component with complex smoothing parameter and then the level and information components leads to:

$$\begin{aligned} &\begin{pmatrix} l_t \\ c_t \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} l_{t-1} \\ c_{t-1} \end{pmatrix} - \begin{pmatrix} 0 & -\alpha_1 \\ 0 & \alpha_0 \end{pmatrix} \begin{pmatrix} l_{t-1} \\ c_{t-1} \end{pmatrix} + \\ &\begin{pmatrix} \alpha_0 & -\alpha_1 \\ \alpha_1 & \alpha_0 \end{pmatrix} \begin{pmatrix} 0 \\ p_t \end{pmatrix} + \begin{pmatrix} \alpha_0 & -\alpha_1 \\ \alpha_1 & \alpha_0 \end{pmatrix} \begin{pmatrix} y_t - l_{t-1} \\ 0 \end{pmatrix} \end{aligned}. \tag{49}$$

The difference between the actual value and the level in (49) is the error term: $y_t - l_{t-1} = \epsilon_t$. Using this and making several transformations gives the following state-space model:

$$\begin{cases} \begin{pmatrix} \hat{y}_t \\ \hat{x}_t \end{pmatrix} = \begin{pmatrix} l_{t-1} \\ c_{t-1} \end{pmatrix} \\ \begin{pmatrix} l_t \\ c_t \end{pmatrix} = \begin{pmatrix} 1 & -(1-\alpha_1) \\ 1 & (1-\alpha_0) \end{pmatrix} \begin{pmatrix} l_{t-1} \\ c_{t-1} \end{pmatrix} + \begin{pmatrix} -\alpha_1 \\ \alpha_0 \end{pmatrix} p_t + \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \epsilon_t \end{cases}. \tag{50}$$

Now if CES should be represented in the state-space form with the SSOE then the measurement equation should also contain the same error term as the transition equation. Since the imaginary part of the measurement equation in (50) is unobservable, it does not contain any useful information for forecasting and can be excluded from the final state-space model:

$$\begin{cases} y_t = l_{t-1} + \epsilon_t \\ \begin{pmatrix} l_t \\ c_t \end{pmatrix} = \begin{pmatrix} 1 & -(1-\alpha_1) \\ 1 & (1-\alpha_0) \end{pmatrix} \begin{pmatrix} l_{t-1} \\ c_{t-1} \end{pmatrix} + \begin{pmatrix} -\alpha_1 \\ \alpha_0 \end{pmatrix} p_t + \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \epsilon_t \end{cases}. \tag{51}$$

The other way of writing down this state-space model is by splitting the level and information components into two equations:

$$\begin{cases} y_t = l_{t-1} + \epsilon_t \\ l_t = l_{t-1} - (1-\alpha_1)c_{t-1} - \alpha_1 p_t + \alpha_0 \epsilon_t \\ c_t = l_{t-1} + (1-\alpha_0)c_{t-1} + \alpha_0 p_t + \alpha_1 \epsilon_t \end{cases}. \tag{52}$$

B Underlying ARIMA

The information component can be calculated using the second equation of (2.7) the following way:

$$c_{t-1} = -\frac{l_t - l_{t-1} + \alpha_1 p_t - \alpha_0 \epsilon_t}{1 - \alpha_1}. \tag{53}$$

Inserting (53) into the third equation of (2.7) leads to:

$$-\frac{l_{t+1}-l_t+\alpha_1 p_{t+1}-\alpha_0\epsilon_{t+1}}{1-\alpha_1}=l_{t-1}-(1-\alpha_0)\frac{l_t-l_{t-1}+\alpha_1 p_t-\alpha_0\epsilon_t}{1-\alpha_1}+\alpha_0 p_t+\alpha_1\epsilon_t. \tag{54}$$

Multiplying both parts of (54) by $-(1-\alpha_1)$ and taking one lag back results in:

$$\begin{gathered} l_t-l_{t-1}+\alpha_1 p_t-\alpha_0\epsilon_t= \\ -(1-\alpha_1)l_{t-2}+(1-\alpha_0)(l_{t-1}-l_{t-2}+\alpha_1 p_{t-1}-\alpha_0\epsilon_{t-1})-(1-\alpha_1)\alpha_0 p_{t-1}-(1-\alpha_1)\alpha_1\epsilon_{t-1}. \end{gathered} \tag{55}$$

Opening the brackets, transferring all the level components to the left hand side and all the error terms and information components to the right hand side and then regrouping the elements gives:

$$l_t-(2-\alpha_0)l_{t-1}-(\alpha_0+\alpha_1-2)l_{t-2}=\alpha_0\epsilon_t-(\alpha_0-\alpha_0^2+\alpha_1-\alpha_1^2)\epsilon_{t-1}-\alpha_1 p_t+(\alpha_1-\alpha_0)p_{t-1}. \tag{56}$$

Now making substitutions $l_t=y_{t+1}-\epsilon_{t+1}$ in (56), taking one more lag back and regrouping the error terms once again leads to:

$$\begin{gathered} y_t-(2-\alpha_0)y_{t-1}-(\alpha_0+\alpha_1-2)y_{t-2}= \\ \epsilon_t-(2-2\alpha_0)\epsilon_{t-1}-(2\alpha_0+2\alpha_1-2-\alpha_0^2-\alpha_1^2)\epsilon_{t-2}-\alpha_1 p_{t-1}-(\alpha_0-\alpha_1)p_{t-2}. \end{gathered} \tag{57}$$

The resulting model (57) is ARMAX(2,2) with lagged information potential:

$$(1-\phi_1 B-\phi_2 B^2)y_t=(1-\theta_1 B-\theta_2 B^2)\epsilon_t+(-\gamma_1 B-\gamma_2 B^2)p_t, \tag{58}$$

where $\phi_1=2-\alpha_0$, $\phi_2=\alpha_0+\alpha_1-2$, $\theta_1=2-2\alpha_0$, $\theta_2=2\alpha_0+2\alpha_1-2-\alpha_0^2-\alpha_1^2$, $\gamma_1=\alpha_1$ and $\gamma_2=\alpha_0-\alpha_1$.

If the information potential is equal to the error term then the model (57) transforms into ARMA(2,2):

$$(1-\phi_1 B-\phi_2 B^2)y_t=(1-\theta_1 B-\theta_2 B^2)\epsilon_t, \tag{59}$$

where $\phi_1=2-\alpha_0$, $\phi_2=\alpha_0+\alpha_1-2$, $\theta_1=2-2\alpha_0+\alpha_1$ and $\theta_2=3\alpha_0+\alpha_1-2-\alpha_0^2-\alpha_1^2$.

In a similar manner using (2.7) it can be shown that the imaginary part of the series has the following underlying model:

$$\begin{array}{l} c_t-(2-\alpha_0)c_{t-1}-(\alpha_0+\alpha_1-2)c_{t-2}= \\ \alpha_1\epsilon_t-(\alpha_1-\alpha_0)\epsilon_{t-1}+\alpha_0 p_t-(\alpha_0+\alpha_1)p_{t-1}, \end{array} \tag{60}$$

If $p_t-c_{t-1}=\xi_t$, where ξ_t is an information gap, then the equation (60) can be rewritten:

$$\begin{array}{l} p_{t+1}-(2-\alpha_0)p_t-(\alpha_0+\alpha_1-2)p_{t-1}-\alpha_0 p_t+(\alpha_0+\alpha_1)p_{t-1}= \\ \xi_{t+1}-(2-\alpha_0)\xi_t-(\alpha_0+\alpha_1-2)\xi_{t-1}+\alpha_1\epsilon_t-(\alpha_1-\alpha_0)\epsilon_{t-1} \end{array}, \tag{61}$$

which after taking one lag back leads to the following simpler ARMAX(2,2) model:

$$(1-\phi_1 B-\phi_2 B)\xi_t=(1-\theta_1 B-\theta_2 B^2)p_t+(-\gamma_1 B-\gamma_2 B^2)\epsilon_t, \tag{62}$$

where $\phi_1=2-\alpha_0$, $\phi_2=\alpha_0+\alpha_1-2$, $\theta_1=2$, $\theta_2=2$, $\gamma_1=\alpha_1$, and $\gamma_2=\alpha_0-\alpha_1$

In case of $p_t=\epsilon_t$ the model transforms into the following ARMA(2,2):

$$(1-\phi_1 B-\phi_2 B^2)\xi_t=(1-\theta_1 B-\theta_2 B^2)\epsilon_t, \tag{63}$$

where $\phi_1=2-\alpha_0$, $\phi_2=\alpha_0+\alpha_1-2$, $\theta_1=2+\alpha_1$ and $\theta_2=\alpha_0-\alpha_1-2$.

So CES has the following complex ARIMAX(2,0,2) model:

$$\begin{array}{l} (1-\phi_1 B-\phi_2 B^2)y_t=(1-\theta_{1,1}B-\theta_{1,2}B^2)\epsilon_t+(-\gamma_1 B-\gamma_2 B^2)p_t \\ (1-\phi_1 B-\phi_2 B^2)\xi_t=(1-\theta_{2,1}B-\theta_{2,2}B^2)p_t+(-\gamma_1 B-\gamma_2 B^2)\epsilon_t \end{array}, \tag{64}$$

where $\phi_1=2-\alpha_0$, $\phi_2=\alpha_0+\alpha_1-2$, $\theta_{1,1}=2-2\alpha_0$, $\theta_{1,2}=2\alpha_0+2\alpha_1-2-\alpha_0^2-\alpha_1^2$, $\theta_{2,1}=2$, $\theta_{2,2}=2$, $\gamma_1=\alpha_1$, $\gamma_2=\alpha_0-\alpha_1$.

C The connection of CES and ETS(A,N,N)

When $p_t = 0$ and $\alpha_1 = 1$ the following state-space model based on (2.7) can be obtained:

$$\begin{cases} y_t = l_{t-1} + \epsilon_t \\ l_t = l_{t-1} + \alpha_0 \epsilon_t \\ c_t = l_{t-1} + (1-\alpha_0) c_{t-1} + \epsilon_t \end{cases} . \tag{65}$$

The measurement equation in (65) implies that the information component c_t is constant over time, so the substitution $c_{t-1} = c_t$ can be made and after opening the brackets and making several simple substitutions the following model is obtained:

$$\begin{cases} y_t = l_{t-1} + \epsilon_t \\ l_t = l_{t-1} + \alpha_0 \epsilon_t \\ c_t = \frac{l_{t-1}}{\alpha_0} + \frac{\epsilon_t}{\alpha_0} \end{cases} . \tag{66}$$

D Stationarity condition for CES

The analysis of the equation (2.22) shows that the eigenvalues can be either real or complex. In the cases of the real eigenvalues they need to be less than one, so the corresponding forecasting trajectory can be stationary and exponentially decreasing. This means that the following condition must be satisfied:

$$\begin{cases} \left| \frac{2-\alpha_0 \pm \sqrt{\alpha_0^2 + 4\alpha_1 - 4}}{2} \right| < 1 \\ \alpha_0^2 + 4\alpha_1 - 4 > 0 \end{cases} \tag{67}$$

The first inequality in (67) leads to the following system of inequalities:

$$\begin{cases} \sqrt{\alpha_0^2 + 4\alpha_1 - 4} > \alpha_0 - 4 \\ \sqrt{\alpha_0^2 + 4\alpha_1 - 4} < \alpha_0 \\ -\sqrt{\alpha_0^2 + 4\alpha_1 - 4} > \alpha_0 - 4 \\ -\sqrt{\alpha_0^2 + 4\alpha_1 - 4} < \alpha_0 \end{cases} \tag{68}$$

The analysis of (68) shows that if $\alpha_0 > 4$, then the third inequality is violated and if $\alpha_0 < 0$, then the second inequality is violated. This means that the condition

$\alpha_0 \in (0, 4)$ is crucial for the stationarity of CES. This also means that the first and the forth inequalities in (68) are always satisfied. Furthermore the second inequality can be transformed into:

$$\alpha_0^2 + 4\alpha_1 - 4 < \alpha_0^2, \tag{69}$$

which after simple cancellations leads to:

$$\alpha_1 < 1. \tag{70}$$

The other important result follows from the third inequality in (68), which can be derived using the condition $\alpha_0 \in (0, 4)$:

$$\alpha_0^2 + 4\alpha_1 - 4 < (\alpha_0 - 4)^2, \tag{71}$$

which implies that:

$$\alpha_1 < 5 - 2\alpha_0, \tag{72}$$

Uniting all these condition and taking into account the second inequality in (67), CES will produce a stationary exponential trajectory when:

$$\begin{cases} 0 < \alpha_0 < 4 \\ \alpha_1 < 5 - 2\alpha_0 \\ \frac{4-\alpha_0^2}{4} < \alpha_1 < 1 \end{cases} \tag{73}$$

The other possible situation is When the first part of the inequality (67) is violated, which will lead to the complex eigenvalues, meaning that the harmonic forecasting trajectory is produced. CES still can be stationary if both eigenvalues in (2.22) have absolute values less than one, meaning that:

$$\sqrt{\Re(\lambda)^2 + \Im(\lambda)^2} < 1 \tag{74}$$

This means in its turn the satisfaction of the following condition:

$$\sqrt{\left(\frac{2-\alpha_0}{2}\right)^2+\left(i\frac{\sqrt{|\alpha_0^2+4\alpha_1-4|}}{2}\right)^2}<1 \tag{75}$$

or:

$$0\leq\frac{4+\alpha_0^2-4\alpha_0}{4}-\frac{\alpha_0^2+4\alpha_1-4}{2}<1 \tag{76}$$

which simplifying leads to:

$$1<\alpha_0+\alpha_1\leq 2$$

or:

$$\begin{cases}\alpha_1>1-\alpha_0\\ \alpha_1\leq 2-\alpha_0\end{cases}$$

The full condition that leads to the harmonic stationary trajectory of CES is:

$$\begin{cases}\alpha_1<\frac{4-\alpha_0^2}{4}\\ \alpha_1>1-\alpha_0\\ \alpha_1\leq 2-\alpha_0\end{cases} \tag{77}$$

Now it can be noted that the first inequality in (77) can be rewritten as a difference of squares:

$$\alpha_1<(2-\alpha_0)\frac{(2+\alpha_0)}{4} \tag{78}$$

Comparing the right hand part of (78) with the right hand side of the third inequality in (77) it can be shown that there is only one point, when both of these inequalities will lead to the same constraint: when $\alpha_0=2$. This is because the line $\alpha_1=2-\alpha_0$ is a tangent line for the function $\alpha_1=(2-\alpha_0)\frac{(2+\alpha_0)}{4}$ in this point. In all the other cases the right hand part of (78) will be less than the right hand side of

the third inequality in (77), which in its turn means that (78) can be substituted by stricter condition:

$$\begin{cases} \alpha_1 < \frac{4-\alpha_0^2}{4} \\ \alpha_1 > 1 - \alpha_0 \end{cases} \tag{79}$$

Uniting (79) with (73) leads to the following general stationarity condition:

$$\begin{cases} 0 < \alpha_0 < 4 \\ \alpha_1 < 5 - 2\alpha_0 \\ \frac{4-\alpha_0^2}{4} < \alpha_1 < 1 \\ \alpha_1 < \frac{4-\alpha_0^2}{4} \\ \alpha_1 > 1 - \alpha_0 \end{cases} \tag{80}$$

The third and forth conditions can be united while the second and fifth conditions are always satisfied when the first condition is met (because the corresponding lines of these inequalities have an intersection in the point $\alpha_0 = 4$). So the following simpler condition can be used instead of (80):

$$\begin{cases} \alpha_1 < 5 - 2\alpha_0 \\ \alpha_1 < 1 \\ \alpha_1 > 1 - \alpha_0 \end{cases} \tag{81}$$

E Stability condition for CES

The general ARMA(2,2) will be invertible when the following condition is satisfied:

$$\begin{cases} \theta_2 + \theta_1 < 1 \\ \theta_2 - \theta_1 < 1 \\ \theta_2 > -1 \\ \theta_2 < 1 \end{cases} \tag{82}$$

Following from subsection 2.4.5, inserting the parameters from the ARMA (2.30) underlying CES the following system of inequalities is obtained:

$$\begin{cases} 3\alpha_0 + \alpha_1 - 2 - \alpha_0^2 - \alpha_1^2 + 2 - 2\alpha_0 + \alpha_1 < 1 \\ 3\alpha_0 + \alpha_1 - 2 - \alpha_0^2 - \alpha_1^2 - 2 + 2\alpha_0 - \alpha_1 < 1 \\ 3\alpha_0 + \alpha_1 - 2 - \alpha_0^2 - \alpha_1^2 > -1 \\ 3\alpha_0 + \alpha_1 - 2 - \alpha_0^2 - \alpha_1^2 < 1 \end{cases} \tag{83}$$

After the cancellations and regrouping of elements the system (83) transforms into:

$$\begin{cases} -\alpha_0^2 + 5\alpha_0 - \alpha_1^2 - 5 < 0 \\ -\alpha_0^2 + \alpha_0 - \alpha_1^2 + 2\alpha_1 - 1 < 0 \\ -\alpha_0^2 + 3\alpha_0 - \alpha_1^2 + \alpha_1 - 1 > 0 \\ -\alpha_0^2 + 3\alpha_0 - \alpha_1^2 + \alpha_1 - 3 < 0 \end{cases} \tag{84}$$

The inequalities in (84) can be transformed into the inequalities, based on squares of differences:

$$\begin{cases} (\alpha_0 - 2.5)^2 + \alpha_1^2 > 1.25 \\ (\alpha_0 - 0.5)^2 + (\alpha_1 - 1)^2 > 0.25 \\ (\alpha_0 - 1.5)^2 + (\alpha_1 - 0.5)^2 < 1.5 \\ (\alpha_0 - 1.5)^2 + (\alpha_1 - 0.5)^2 > -0.5 \end{cases} \tag{85}$$

Note that any point on the plane of smoothing parameters satisfies the last inequality in (85), so it does not change the stability region and can be skipped.

F General seasonal CES and SARIMA

The model (3.7) can be written in the following state-space form:

$$\begin{matrix} y_t = w_0' x_{0,t-1} + w_1' x_{1,t-m} + \epsilon_t \\ x_{0,t} = F_0 x_{0,t-1} + g_0 \epsilon_t \\ x_{1,t} = F_1 x_{1,t-m} + g_1 \epsilon_t \end{matrix} , \tag{86}$$

where $x_{0,t} = \begin{pmatrix} l_{0,t} \\ c_{0,t} \end{pmatrix}$ is the state vector of the non-seasonal part of CES, $x_{1,t} = \begin{pmatrix} l_{1,t} \\ c_{1,t} \end{pmatrix}$ is the state vector of the seasonal part, $w_0 = w_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ are the measurement vectors, $F_0 = \begin{pmatrix} 1 & \alpha_1 - 1 \\ 1 & 1 - \alpha_0 \end{pmatrix}$, $F_1 = \begin{pmatrix} 1 & \beta_1 - 1 \\ 1 & 1 - \beta_0 \end{pmatrix}$ are transition matrices and $g_0 = \begin{pmatrix} \alpha_1 - \alpha_0 \\ \alpha_1 + \alpha_0 \end{pmatrix}$, $g_1 = \begin{pmatrix} \beta_1 - \beta_0 \\ \beta_1 + \beta_0 \end{pmatrix}$ are persistence vectors of non-seasonal and seasonal parts respectively.

Observe that the lags of the non-seasonal and seasonal parts in (86) differ, which leads to splitting the state-space model into two parts. But uniting these parts in a one bigger part will lead to the conventional state-space model:

$$\begin{aligned} y_t &= w' x_{t-l} + \epsilon_t \\ x_t &= F x_{t-l} + g\epsilon_t \end{aligned}, \tag{87}$$

where $x_t = \begin{pmatrix} x_{0,t} \\ x_{1,t} \end{pmatrix}$, $x_{t-l} = \begin{pmatrix} x_{0,t-1} \\ x_{1,t-m} \end{pmatrix}$, $w = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix}$, $F = \begin{pmatrix} F_0 & 0 \\ 0 & F_1 \end{pmatrix}$, $g = \begin{pmatrix} g_0 \\ g_1 \end{pmatrix}$. The state vector x_{t-l} can also be rewritten as $x_{t-l} = \begin{pmatrix} B & 0 \\ 0 & B^m \end{pmatrix} \begin{pmatrix} x_{0,t} \\ x_{1,t} \end{pmatrix}$, where B is a backshift operator. Making this substitution and taking $L = \begin{pmatrix} B & 0 \\ 0 & B^m \end{pmatrix}$ the state-space model (87) can be transformed into:

$$\begin{aligned} y_t &= w' L x_t + \epsilon_t \\ x_t &= F L x_t + g\epsilon_t \end{aligned} \tag{88}$$

The transition equation in (88) can also be rewritten as:

$$(I_2 - FL)x_t = g\epsilon_t, \tag{89}$$

which after a simple manipulation leads to:

$$x_t = (I_2 - FL)^{-1} g\epsilon_t, \tag{90}$$

Substituting (90) into measurement equation in (88) gives:

$$y_t = w' L (I_2 - FL)^{-1} g\epsilon_t + \epsilon_t. \tag{91}$$

Inserting the values of the vectors and multiplying the matrices leads to:

$$y_t = (1 + w_0'(I_2 - F_0 B)^{-1} g_0 B + w_1'(I_2 - F_1 B^m)^{-1} g_1 B^m)\epsilon_t. \tag{92}$$

Substituting the values by the matrices in (92) gives:

$$\begin{aligned} y_t = \Bigg(1 + w_0' \begin{pmatrix} 1 - B & (1-\alpha_1)B \\ -B & 1 - B + \alpha_0 B \end{pmatrix}^{-1} \begin{pmatrix} \alpha_1 - \alpha_0 \\ \alpha_1 + \alpha_0 \end{pmatrix} B + \\ w_1' \begin{pmatrix} 1 - B^m & (1-\beta_1)B^m \\ -B^m & 1 - B^m + \beta_0 B^m \end{pmatrix}^{-1} \begin{pmatrix} \beta_1 - \beta_0 \\ \beta_1 + \beta_0 \end{pmatrix} B^m \Bigg) \epsilon_t \end{aligned}. \tag{93}$$

The inverse of the first matrix in (93) is equal to:

$$(I_2 - F_0 B)^{-1} = \frac{1}{1 - 2B - (\alpha_0 + \alpha_1 - 2)B^2} \begin{pmatrix} 1 - (1 - \alpha_0)B & (\alpha_1 - 1)B \\ B & 1 - B \end{pmatrix}, \tag{94}$$

similarly the inverse of the second matrix is:

$$(I_2 - F_1 B^m)^{-1} = \frac{1}{1 - 2B^m - (\beta_0 + \beta_1 - 2)B^{2m}} \begin{pmatrix} 1 - (1 - \beta_0)B^m & (\beta_1 - 1)B^m \\ B^m & 1 - B^m \end{pmatrix}. \tag{95}$$

Inserting (94) and (95) into (93), after cancellations and regrouping of elements leads to:

$$\begin{aligned} &(1 - 2B - (\alpha_0 + \alpha_1 - 2)B^2)(1 - 2B^m - (\beta_0 + \beta_1 - 2)B^{2m})y_t = \\ &[(1 - 2B - (\alpha_0 + \alpha_1 - 2)B^2)(1 - 2B^m - (\beta_0 + \beta_1 - 2)B^{2m}) + \\ &(1 - 2B^m - (\beta_0 + \beta_1 - 2)B^{2m})(\alpha_1 - \alpha_0 - ((\alpha_0 - \alpha_1)^2 - 2\alpha_1)B) + \\ &(1 - 2B - (\alpha_0 + \alpha_1 - 2)B^2)(\beta_1 - \beta_0 - ((\beta_0 - \beta_1)^2 - 2\beta_1)B^m)]\,\epsilon_t \end{aligned} \tag{96}$$

Unfortunately, there is no way to simplify (96) to present it in a compact form, so the final model corresponds to SARIMA$(2, 0, 2m + 2)(2, 0, 0)_m$.

G Discount matrix of the general seasonal CES

Substituting the error term in the transition equation (87) by the value from the measurement equation leads to:

$$x_t = F x_{t-l} + g y_t - g w' x_{t-l} = (F - g w') x_{t-l} + g y_t, \tag{97}$$

which leads to the following discount matrix:

$$D = F - g w' = \begin{pmatrix} 1 - \alpha_0 + \alpha_1 & \alpha_1 - 1 & \alpha_1 - \alpha_0 & 0 \\ 1 - \alpha_0 - \alpha_1 & 1 - \alpha_0 & -\alpha_1 - \alpha_0 & 0 \\ \beta_1 - \beta_0 & 0 & 1 - \beta_0 + \beta_1 & \beta_1 - 1 \\ -\beta_1 - \beta_0 & 0 & 1 - \beta_0 - \beta_1 & 1 - \beta_0 \end{pmatrix} \tag{98}$$

H The calculation of $c_{j,h}$ for ARIMA(1,1,1)

The eigenvalues of the transition matrix $\mathbf{F} = \begin{pmatrix} 1+\phi_1 & 1 \\ -\phi_1 & 0 \end{pmatrix}$ are $\lambda_1 = 1, \lambda_2 = \phi_1$ while the corresponding eigenvectors can be $v_1 = \begin{pmatrix} 1 \\ -\phi_1 \end{pmatrix}$ and $v_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

Using matrix decomposition and inserting these values into $c_{j,h}$ gives:

$$c_{j,h} = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -\phi_1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \phi_1 \end{pmatrix}^{h-1-j} \begin{pmatrix} 1 & 1 \\ -\phi_1 & -1 \end{pmatrix}^{-1} \begin{pmatrix} 1+\phi_1+\theta_1 \\ -\phi_1 \end{pmatrix}. \tag{99}$$

The inverse of the matrix in the right hand side of (99) (when $\phi_1 \neq 1$) is equal to:

$$\begin{pmatrix} 1 & 1 \\ -\phi_1 & -1 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{1-\phi_1} & \frac{1}{1-\phi_1} \\ -\frac{\phi_1}{1-\phi_1} & -\frac{1}{1-\phi_1} \end{pmatrix}, \tag{100}$$

and inserting the value (100) into (99) and multiplying first two matrices leads to:

$$c_{j,h} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \phi_1^{j-1} \end{pmatrix} \begin{pmatrix} \frac{1}{1-\phi_1} & \frac{1}{1-\phi_1} \\ -\frac{\phi_1}{1-\phi_1} & -\frac{1}{1-\phi_1} \end{pmatrix} \begin{pmatrix} 1+\phi_1+\theta_1 \\ -\phi_1 \end{pmatrix}. \tag{101}$$

Multiplying the first two matrices one more time gives:

$$c_{j,h} = \begin{pmatrix} 1 & \phi_1^{j-1} \end{pmatrix} \begin{pmatrix} \frac{1}{1-\phi_1} & \frac{1}{1-\phi_1} \\ -\frac{\phi_1}{1-\phi_1} & -\frac{1}{1-\phi_1} \end{pmatrix} \begin{pmatrix} 1+\phi_1+\theta_1 \\ -\phi_1 \end{pmatrix}. \tag{102}$$

Opening the remaining brackets in (102) leads to:

$$c_{j,h} = \frac{1 + \theta_1 - \phi_1^{j+1} - \phi_1^j \theta_1}{1 - \phi_1} \tag{103}$$

After regrouping elements of (103), formula transforms into:

$$c_{j,h} = 1 + (\phi_1 + \theta_1)\frac{1 - \phi_1^j}{1 - \phi_1} \tag{104}$$

The quotient in the right hand side of the formula (104) is the sum of geometric series. Writing down the sum instead of the quotient gives the final formula for the calculation of $c_{j,h}$:

$$c_{j,h} = 1 + (\phi_1 + \theta_1)\sum_{i=1}^{j} \phi_1^{i-1} \tag{105}$$

I Simplifying the concentrated likelihood

In the function (4.38) the product $\mathbf{E}_\mathbf{t}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{E}_\mathbf{t}$ will always result in a scalar. This allows to use trace function in the sum in the right hand side of the equation in the following manner:

$$\sum_{t=1}^{T}\left(\mathbf{E}_\mathbf{t}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{E}_\mathbf{t}\right) = \sum_{t=1}^{T}\left(tr(\mathbf{E}_\mathbf{t}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{E}_\mathbf{t})\right). \tag{106}$$

Trace function allows to rearrange the right part of (106):

$$\sum_{t=1}^{T}\left(tr(\mathbf{E}_\mathbf{t}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{E}_\mathbf{t})\right) = \sum_{t=1}^{T}\left(tr(\hat{\mathbf{\Sigma}}^{-1}\mathbf{E}_\mathbf{t}\mathbf{E}_\mathbf{t}')\right) = tr\left(\hat{\mathbf{\Sigma}}^{-1}\sum_{t=1}^{T}\mathbf{E}_\mathbf{t}\mathbf{E}_\mathbf{t}'\right). \tag{107}$$

The sum of the matrices in the right hand side of (107) is equal to the estimated covariance (4.37) multiplied by T. The multiplication of $\hat{\mathbf{\Sigma}}^{-1}$ and $\hat{\mathbf{\Sigma}}$ gives the identity matrix with the trace equal to h. So finally the sum (106) can be rewritten as:

$$\sum_{t=1}^{T}\left(\mathbf{E}_\mathbf{t}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{E}_\mathbf{t}\right) = Th. \tag{108}$$

Inserting (108) in (4.38) leads to the following simplified concentrated log-likelihood:

$$l(\theta, \hat{\mathbf{\Sigma}}|\mathbf{Y}) = -\frac{T}{2}\left(h\log(2\pi) + \log|\hat{\mathbf{\Sigma}}|\right) - \frac{T}{2}h, \tag{109}$$

which after regrouping leads to:

$$l(\theta, \hat{\mathbf{\Sigma}}|\mathbf{Y}) = -\frac{T}{2}\left(h\log(2\pi e) + \log|\hat{\mathbf{\Sigma}}|\right). \tag{110}$$

J Simplifying the generalised variance

The covariances in the matrix Σ can be represented using correlations in the following manner:

$$\sigma_{i,j} = r_{i,j}\sigma_i\sigma_j, \tag{111}$$

where $r_{i,j}$ is the correlation coefficient between the error i-steps ahead and the error j-steps ahead.

Substituting (111) in the covariance matrix (4.41) leads to:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & r_{1,2}\sigma_1\sigma_2 & \dots & r_{1,h}\sigma_1\sigma_h \\ r_{1,2}\sigma_1\sigma_2 & \sigma_2^2 & \dots & r_{2,h}\sigma_2\sigma_h \\ \vdots & \vdots & \ddots & \vdots \\ r_{1,h}\sigma_1\sigma_h & r_{2,h}\sigma_2\sigma_h & \dots & \sigma_h^2 \end{pmatrix}. \tag{112}$$

Matrix (112) can now be represented as a multiplication of the following three matrices:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_h \end{pmatrix} \begin{pmatrix} 1 & r_{1,2} & \dots & r_{1,h} \\ r_{1,2} & 1 & \dots & r_{2,h} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1,h} & r_{2,h} & \dots & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_h \end{pmatrix}. \tag{113}$$

Taking the determinant of this matrix leads to:

$$|\mathbf{\Sigma}| = \prod_{j=1}^{h} \sigma_j^2 |\mathbf{R}|, \tag{114}$$

where $\mathbf{R} = \begin{pmatrix} 1 & r_{1,2} & \dots & r_{1,h} \\ r_{1,2} & 1 & \dots & r_{2,h} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1,h} & r_{2,h} & \dots & 1 \end{pmatrix}$ is the correlation matrix.

K Covariances and correlations of forecast errors

Using the formula (4.6) the covariance between i and j steps ahead errors when $j > i$ and $i, j \neq 1$ can be written the following way:

$$\sigma_{i,j} = cov\left(\sum_{l=1}^{i-1} c_{l,i}\epsilon_{t+l} + \epsilon_{t+i}, \sum_{m=1}^{j-1} c_{m,j}\epsilon_{t+m} + \epsilon_{t+j}\right). \tag{115}$$

Opening sums in (115) leads to the following:

$$\begin{aligned} \sigma_{i,j} = cov(\epsilon_{t+i}, \epsilon_{t+j}) + cov\left(\epsilon_{t+i}, \sum_{m=1}^{j-1} c_{m,j}\epsilon_{t+m}\right) + \\ cov\left(\sum_{l=1}^{i-1} c_{l,i}\epsilon_{t+l}, \epsilon_{t+j}\right) + cov\left(\sum_{l=1}^{i-1} c_{l,i}\epsilon_{t+l}, \sum_{m=1}^{j-1} c_{m,j}\epsilon_{t+m}\right). \end{aligned} \tag{116}$$

Assuming that the error term is not autocorrelated, the first and the third covariances in (116) will be equal to zero. In the second covariance all the elements will be equal to zero except for one: $cov(\epsilon_{t+i}, c_{i,j}\epsilon_{t+i}) = \sigma_1^2 c_{i,j}$. The last covariance in (116) will have $i-1$ non-zero elements that also become variances. All of that leads to:

$$\sigma_{i,j} = \sigma_1^2 c_{i,j} + \sigma_1^2 \sum_{l=1}^{i-1} c_{l,j} c_{l,i}. \tag{117}$$

In case when $i = 1$, the formula (115) will have a different form:

$$\sigma_{1,j} = cov\left(\epsilon_{t+1}, \sum_{m=1}^{j-1} c_{m,j}\epsilon_{t+m} + \epsilon_{t+j}\right), \tag{118}$$

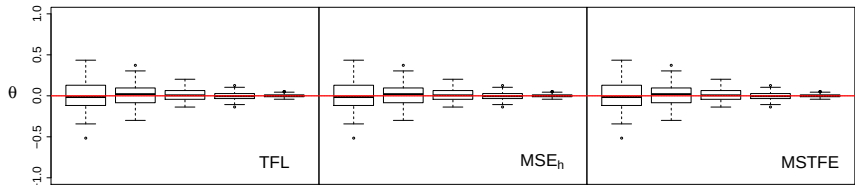
which after applying the similar approach leads to:

$$\sigma_{1,j} = \sigma_1^2 c_{1,j}. \tag{119}$$

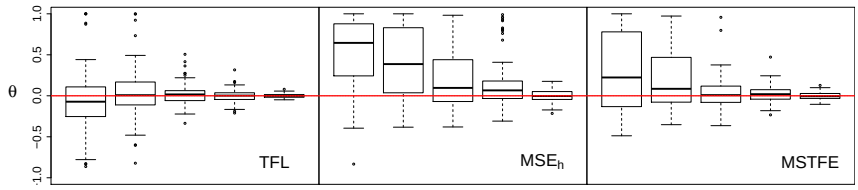
Dividing (117) and (119) by square roots of (4.8) for i and j the correlation coefficient can be estimated:

$$r_{i,j} = \begin{cases} \dfrac{\left(c_{i,j} + \sum_{l=1}^{i-1} c_{l,j} c_{l,i}\right)}{\sqrt{\left(1 + \sum_{l=1}^{i-1} c_{l,i}^2\right)\left(1 + \sum_{l=1}^{j-1} c_{l,j}^2\right)}} & \text{when } i, j \neq 1 \\ \dfrac{c_{1,j}}{\sqrt{\left(1 + \sum_{l=1}^{j-1} c_{l,j}^2\right)}} & \text{when } i = 1 \end{cases}. \tag{120}$$

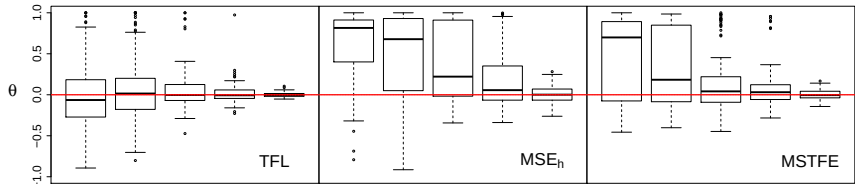
L Boxplots of AR(1) parameter



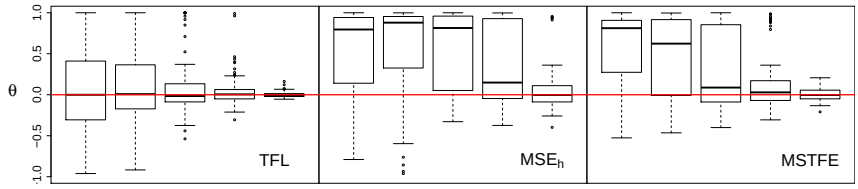
(a) h=1



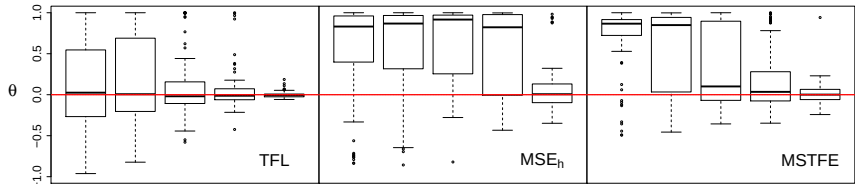
(b) h=10



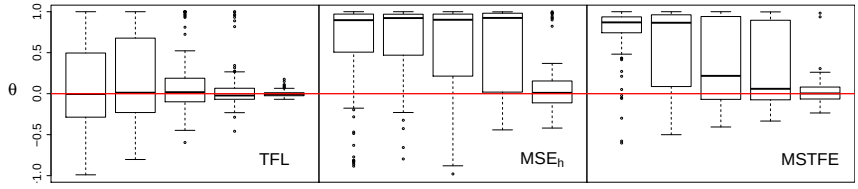
(c) h=20



(d) h=30



(e) h=40



(f) h=50

Figure 7: AR estimates using different methods. Wrong model. Boxplots of parameter ϕ.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactionson Automatic Control 19 (6), 716–723.

Andrawis, R. R., Atiya, A. F., El-Shishiny, H., 2011. Combination of long term and short term forecasts, with application to tourism demand forecasting. International Journal of Forecasting 27 (3), 870–886.

Assimakopoulos, V., Nikolopoulos, K., 2000. The theta model: a decomposition approach to forecasting. International Journal of Forecasting 16, 521–530.

Athanasopoulos, G., de Silva, A., 2012. Multivariate exponential smoothing for forecasting tourist arrivals. Journal of Travel Research 51 (5), 640–652.

Athanasopoulos, G., Hyndman, R. J., Song, H., Wu, D. C., 2011. The tourism forecasting competition. International Journal of Forecasting 27 (3), 822–844.

Bhansali, R., 1996. Asymptotically efficient autoregressive model selection for multistep prediction. Annals of the Institute of Statistical Mathematics 48 (3), 577–602.

Bhansali, R., 1997. Direct autoregressive predictors for multistep prediction: Order selection and performance relative to the plug in predictors. Statistica Sinica 7, 425–449.

Billah, B., King, M. L., Snyder, R. D., Koehler, A. B., 2006. Exponential smoothing model selection for forecasting. International Journal of Forecasting 22 (2), 239–247.

Box, G., Jenkins, G., 1976. Time series analysis: forecasting and control. Holden-day, Oakland, California.

Brenner, J. L., D'Esopo, D. a., Fowler, a. G., 1968. Difference equations in forecasting formulas. Management Science 15 (3), 141–159.

Brown, R. G., 1956. Exponential smoothing for predicting demand. Arthur D. Little, Inc.

Burnham, K. P., Anderson, D. R., 2004. Model selection and multimodel inference. Springer New York, New York, NY.

Chevillon, G., 2007. Direct multi-step estimation and forecasting. Journal of Economic Surveys 21 (4), 746–785.

Chevillon, G., 2009. Multi-step forecasting in emerging economies: An investigation of the South African GDP. International Journal of Forecasting 25 (3), 602–628.

Chevillon, G., 2016. Multistep forecasting in the presence of location shifts. International Journal of Forecasting 32 (1), 121–137.

Chevillon, G., Hendry, D. F., 2005. Non-parametric direct multi-step estimation for forecasting economic processes. International Journal of Forecasting 21 (2), 201–218.

Cox, D. R., 1961. Prediction by exponentially weighted moving averages and related methods. Journal of the Royal Statistical Society, Series B 23 (2), 414–422.

Davydenko, A., Fildes, R., 2012. A joint Bayesian forecasting model of judgment and observed data. Preprint. Lancaster University: The Department of Management Science.

De Livera, A. M., 2010. Exponentially weighted methods for multiple seasonal time series. International Journal of Forecasting 26 (4), 655–657.

De Livera, A. M., Hyndman, R. J., Snyder, R. D., 2011. Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American Statistical Association 106 (496), 1513–1527.

Fildes, R., Hibon, M., Makridakis, S., Meade, N., 1998. Generalising about univariate forecasting methods: further empirical evidence. International Journal of Forecasting 14 (3), 339–358.

Franses, P. H., Legerstee, R., 2009. A unifying view on multi-step forecasting using an autoregression. Journal of Economic Surveys 24 (3), 389–401.

Gardner, E. S., 1985. Exponential smoothing: The state of the art. Journal of Forecasting 4 (1), 1–28.

Gardner, E. S., 2006. Exponential smoothing: The state of the art-Part II. International Journal of Forecasting 22 (4), 637–666.

Gardner, E. S., Diaz-Saiz, J., 2008. Exponential smoothing in the telecommunications data. International Journal of Forecasting 24 (1), 170–174.

Gardner, E. S., McKenzie, E., 1985. Forecasting trends in time series. Management Science 31 (10), 1237–1246.

Gardner, E. S., McKenzie, E., 2011. Why the damped trend works. Journal of the Operational Research Society 62 (6), 1177–1180.

Gersch, W., Kitagawa, G., 1983. The prediction of time series with trends and seasonalities. Journal of Business & Economic Statistics 1 (3), 253–264.

Gould, P. G., Koehler, A. B., Ord, J. K., Snyder, R. D., Hyndman, R. J., Vahid-Araghi, F., Vahid-Araghi, F., 2008. Forecasting time series with multiple seasonal patterns. European Journal of Operational Research 191 (1), 205–220.

Haywood, J., Tunnicliffe-Wilson, G., 1997. Fitting time series models by minimizing multistep-ahead errors: a frequency domain approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59 (1), 237–254.

Heijmans, R. D., Magnus, J. R., 1986a. Asymptotic normality of maximum likelihood estimators obtained from normally distributed but dependent observations. Econometric Theory 2 (3), 374–412.

Heijmans, R. D., Magnus, J. R., 1986b. Consistent maximum-likelihood estimation with dependent observations. Journal of Econometrics 32 (2), 253–285.

Holt, C. C., 2004. Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting 20 (1), 5–10.

Hyndman, R. J., Akram, M., Archibald, B., 2008a. The admissible parameter space for exponential smoothing models. Annals of the Institute of Statistical Mathematics 60 (2), 407–426.

Hyndman, R. J., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. Journal of Statistical Software 27 (3), 1–22.

Hyndman, R. J., Koehler, A., Ord, K., Snyder, R., 2008b. Forecasting with exponential smoothing. Springer Berlin Heidelberg.

Hyndman, R. J., Koehler, A., Ord, K., Snyder, R. D., 2005. Prediction intervals for exponential smoothing using two new classes of state space models. Journal of Forecasting 24, 17–35.

Hyndman, R. J., Koehler, A., Snyder, R., Grose, S., 2002. A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting 13 (8), 439–454.

Hyndman, R. J., Koehler, A. B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22 (4), 679–688.

Ing, C.-K., 2003. Multistep prediction in autoregressive processes. Econometric Theory 19 (2), 254–279.

Ing, C.-K., 2004. Selecting optimal multistep predictors for autoregressive processes of unknown order. Annals of Statistics 32 (2), 693–722.

Ing, C.-K., Lin, J.-L., Yu, S.-H., 2009. Toward optimal multistep forecasts in non-stationary autoregressions. Bernoulli 15 (2), 402–437.

Jordà, Ò., Marcellino, M., 2010. Path forecast evaluation. Journal of Applied Econometrics 25 (4), 635–662.

Jose, V. R. R., Winkler, R. L., 2008. Simple robust averages of forecasts: Some empirical results. International Journal of Forecasting 24 (1), 163–169.

Kang, I.-B., 2003. Multi-period forecasting using different models for different horizons: an application to U.S. economic time series data. International Journal of Forecasting 19 (3), 387–400.

Koehler, A. B., Snyder, R. D., Ord, J. K., 2001. Forecasting models and prediction intervals for the multiplicative Holt-Winters method. International Journal of Forecasting 17 (2), 269–286.

Kolassa, S., 2011. Combining exponential smoothing forecasts using Akaike weights. International Journal of Forecasting 27 (2), 238–251.

Koning, A. J., Franses, P. H., Hibon, M., Stekler, H. O., 2005. The M3 competition: Statistical tests of the results. International Journal of Forecasting 21 (3), 397–409.

Kourentzes, N., Petropoulos, F., 2016. Forecasting with multivariate temporal aggregation: The case of promotional modelling. International Journal of Production Economics, 1–9.

Kourentzes, N., Petropoulos, F., Trapero, J. R., 2014. Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting 30 (2), 291–302.

Maia, A. L. S., de Carvalho, F. D. a. T., 2011. Holt's exponential smoothing and neural network models for forecasting interval-valued time series. International Journal of Forecasting 27 (3), 740–759.

Makridakis, S., Andersen, A. P., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., Winkler, R. L., 1982. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. Journal of Forecasting 1 (2), 111–153.

Makridakis, S., Hibon, M., 2000. The M3-Competition: results, conclusions and implications. International Journal of Forecasting 16, 451–476.

Marcellino, M., Stock, J. H., Watson, M. W., 2006. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. Journal of Econometrics 135 (1-2), 499–526.

McElroy, T., 2015. When are direct multi-step and iterative forecasts identical? Journal of Forecasting 34, 315–336.

McElroy, T., Wildi, M., 2013. Multi-step-ahead estimation of time series models. International Journal of Forecasting 29 (3), 378–394.

McKenzie, E., Gardner, E. S., 2010. Damped trend exponential smoothing: A modelling viewpoint. International Journal of Forecasting 26 (4), 661–665.

Ord, K., Fildes, R., 2012. Principles of business forecasting. Cengage Learning.

Ord, K., Koehler, A., Snyder, R., 1997. Estimation and prediction for a class of dynamic nonlinear statistical models. Journal of the American Statistical Association 92, 1621–1629.

Pegels, C. C., 1969. Exponential forecasting: some new variations. Management Science 15 (5), 311–315.

Pesaran, M. H., Pick, A., Timmermann, A., 2010. Variable selection, estimation and inference for multi-period forecasting problems.

Proietti, T., 2011. Direct and iterated multistep AR methods for difference stationary processes. International Journal of Forecasting 27 (2), 266–280.

Snyder, R. D., 1985. Recursive estimation of dynamic linear models. Journal of the Royal Statistical Society, Series B 47 (2), 272–276.

Snyder, R. D., Koehler, A. B., 2009. Incorporating a tracking signal into a state space model. International Journal of Forecasting 25 (3), 526–530.

Snyder, R. D., Ord, K., Koehler, A. B., 2001. Prediction intervals for ARIMA models. Journal of Business & Economic Statistics 19, 217–225.

Svetunkov, I., Kourentzes, N., 2015. Complex exponential smoothing. Lancaster University, Department of Management Science, 1 – 30.

Svetunkov, S., 2012. Complex-valued modeling in economics and finance. Springer New York.

Taieb, S. B., Atiya, A. F., 2016. A bias and variance analysis for multistep-ahead time series forecasting. IEEE Transactions on Neural Networks and Learning Systems 27 (1), 62–76.

Taylor, J. W., 2003. Exponential smoothing with a damped multiplicative trend. International Journal of Forecasting 19 (4), 715–725.

Taylor, J. W., 2008. An evaluation of methods for very short-term load forecasting using minute-by-minute British data. International Journal of Forecasting 24 (4), 645–658.

Taylor, J. W., 2010. Triple seasonal methods for short-term electricity demand forecasting. European Journal of Operational Research 204 (1), 139–152.

Taylor, J. W., Snyder, R., 2012. Forecasting intraday time series with multiple seasonal cycles using parsimonious seasonal exponential smoothing. Omega 40 (6), 748–757.

Tiao, G., Xu, D., 1993. Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. Biometrika 80 (3), 623–641.

Trapero, J. R., Kourentzes, N., Martin, A., 2015. Short-term solar irradiation forecasting based on dynamic harmonic regression. Energy 84, 289–295.

Wang, J.-J., Wang, J.-Z., Zhang, Z.-G., Guo, S.-P., 2012. Stock index forecasting based on a hybrid model. Omega 40 (6), 758–766.

Weiss, A., Andersen, A. P., 1984. Estimating time series models using the relevant forecast evaluation criterion. Journal of the Royal Statistical Society. Series A. 147 (3), 484.

Weiss, A. A., 1991. Multi-step estimation and forecasting in dynamic models. Journal of Econometrics 48, 135–149.

Xia, Y., Tong, H., 2011. Feature matching in time series modeling. Statistical Science 26 (1), 21–46.