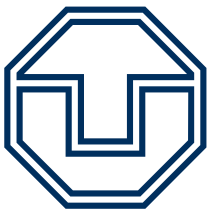

TECHNISCHE UNIVERSITÄT DRESDEN

MASTER THESIS

Forecasting univariate stock indices with parametric models and neural networks: A comparison

Author:
Niklas PAULIG

Supervisor:
Prof. Dr. Bernhard SCHIPP
M.Sc. Paul Felix REITER

A master thesis submitted in fulfillment of the requirements to aquire a master of science

at the

Chair of Quantitative Methods, esp. Econometrics

March 24, 2021

Abstract

Faculty of Business and Economics

Chair of Quantitative Methods, esp. Econometrics

Master thesis

Forecasting univariate stock indices with parametric models and neural networks: A comparison

by Niklas PAULIG

The purpose of this article is to conduct a Monte-Carlo simulation study that compares the efficiency, accuracy and empirical computation time of three different estimators in terms of selecting copulae and estimating their parameters. I will compare two moment-based estimators, the inverse of Kendall's Tau and Blomqvist's Beta, to the classic canonical maximum likelihood estimator. I find that, in terms of estimating parameters for a given copula, the moment-based estimators are inferior to maximum likelihood, especially for small sample sizes. For selecting copulae from a set of data, however, all three estimators are equally accurate while the moment-based ones are at least 16 times faster computationally.

Contents

List of Figures

List of Tables

List of Abbreviations

i.i.d	**i**ndependent and **i**dentically **d**istributed
CDF	**C**umulative **D**istribution **F**unction

1 Introduction

The aim of this thesis is to survey the accuracy and practicability of different univariate forecasting methods for different stock indices around the world. The methods used in this study are split into two groups based on their inner workings. The first group encompasses models with a well-established statistical foundation i.e. Exponential state-space and ARIMA(p,d,q) models, while the latter focuses on artificial neural networks (ANNs) which can be considered non-parametric as they do not need statistical assumptions before estimation but rather derive a structure directly from the underlying data.

The models are configured to provide a seven-day-ahead forecast which will be compared using scale invariant error metrics. The models with statistical foundation will also be compared among themselves using maximum entropy estimators or their derivatives.

The present thesis will be structured as follows. The first chapter reviews the history and literature of all here considered models. The second chapter gives an in depth explanation of those models and points out their utility in terms of univariate time-series forecasting. The third chapter presents the data to be analyzed and applies the models to the data. The fourth chapter discusses the results and concludes.

1.1 History of time series forecasting and literature review

The forecasting of serially structured data, i.e. data that has a temporal component ordering it from past to present, has challenged researchers ever since the first mathematical models for describing such data have evolved.

Exponential Smoothing models

More than 60 years ago, the family of exponential smoothing models emerged from the works of Brown (1962), Holt (1957) and Winters (1960), although they did not have profound statistical foundations but much rather were seen as ad-hoc techniques for extrapolating serially dependent data. Those models received not much attention among statisticians until the mid 1980s, when Gardner Jr (1985) (later revised as part two (see Gardner Jr (2006))) published a comprehensive review and classification of all known-to-date exponential smoothing models. In the same year Snyder (1985) laid the foundation for describing exponential smoothing models as innovation state space models and gave way for most of the recent developments in this field. Today, state-space models are the state of the art for modeling exponential smoothing, especially since Hyndman et al. (2002) and Taylor (2003)

designed automatic state space frameworks for nearly all smoothing models (following their own taxonomy), which are now also natively implemented into various statistical software packages.

Although not very complex, exponential smoothing models are widely adapted in industry and commerce for inventory control or sales forecasting. Until today surprisingly accurate forecasts can be generated with such models (Chatfield et al., 2001) and even out-perform more advanced models such as ARIMA (Hyndman, 2001).

ARIMA models

Inspired by the concept of a deterministic world, Yule (1927) was the first to develop the idea that time series can be seen as realizations of stochastic processes. This simple, yet novel approach laid the foundation for ample research efforts, developing many of the standard tools for time series analysis known today. After the subsequent description of autoregressive (AR) and moving average (MA) models, Box and Jenkins (1976) consolidated and integrated the existing knowledge and formulated a three stage approach for the identification, estimation and verification of time series. This approach, the *Box-Jenkins-approach*, is used until today in various applications.

It gave way for an abundance of studies successfully implementing their procedure (at least in parts) in different fields of investigation such as forecasting of the federal funds rate (Hein and Spudeck, 1988), monthly electricity consumption (Harris and Liu, 1993) or stock price prediction of the NYSE and NSE (Ariyo, Adewumi, and Ayo, 2014) which is closely related to this study's goal.

Artificial neural networks

Neural networks have a history that dates back nearly as long as that of the above mentioned methodologies, starting with Rosenblatt (1957) and his *Perceptron* as a first associative memory, based on the concept of neurons working in the human brain and Hu and Root (1964) using neural networks specifically for forecasting purposes. The practical utility, however, remained low in the early days as the number of solvable problems using such nets were limited due to lacking computational power and mathematical restraints.

In the subsequent decades, neural networks received only limited research attention until Rumelhart, Hinton, and Williams (1985) simplified training of networks by introducing the back-propagation algorithm to tackle complex learning problems with multi-layer-perceptrons (MLPs). This publication revived interest in the field and gave way for several advances, such as the proof that MLPs are able to approximate any measurable function arbitrarily well (Hornik, Stinchcombe, and White, 1989), the first commercial use of neural networks for handwritten zip-code recognition (LeCun et al., 1989), the emergence of convolutional neural networks (LeCun, Bengio, et al., 1995) or the invention of long-short-term memory cells (LSTM) by Hochreiter and Schmidhuber (1997) to circumvent learning problems of recurrent neural networks (RNNs).

Despite new research advancements, neural networks remained hard to train and work with, shifting away research efforts to related models such as Support Vector Machines or Random Forests (see Boser, Guyon, and Vapnik (1992) or Ho (1995)).

The era of deep learning began with the work of Hinton, Osindero, and Teh (2006), who showed that proper weight initialization made it possible to train networks with many hidden layers (deep networks), by pre-training every layer separately at first. These advancements lead to efficient weight intitializers that did not need the layers to be pre-trained (Glorot and Bengio, 2010). The same authors also highlighted the impact of activation functions on the capabilities of neural networks, leading to new research resulting in the well known rectified linear unit (ReLU) activation function and its derivatives (Jarrett et al. (2009) and Nair and Hinton (2010)).

The insights from the last decades, paired with an stark increase in available data and computational power, led to remarkable results in multiple fields of research, for example the forecasting of retail demand (Wen et al. (2017), Salinas et al. (2020)), traffic (Laptev et al. (2017), Li et al. (2017)) or energy (Dimoulkas, Mazidi, and Herre (2019)).

Despite the multitude of positive feedback, there are still doubts whether artificial neural networks are being oversold. There are several papers documenting ANNs being outperformed by basic random walks (see Conejo et al. (2005), or Tkacz (2001)).

2 Model description

2.1 Exponential smoothing

2.1.1 Model description

Developed independently by Robert G. Brown and Charles C. Holt in the early 1950s to develop a tracking model for fire-control information and forecasting spare parts, exponential smoothing methods quickly became a useful technique for extrapolating serial data.

In this thesis the notation and taxonomy implemented by Gardner Jr (2006), Hyndman et al. (2002) and Taylor (2003) is used. This thesis uses three different models presented below. The following notation is used:

Symbol	Definition
α	Smoothing parameter for the level of the series.
γ	Smoothing parameter for the trend.
m	Number of periods in the forecast.
ϕ	Autoregressive or damping parameter.
S_t	Smoothed level of the series, computed after X_t is observed.
X_t	Observed value of the time series in period t.
T_t	Smoothed additive trend at the end of period t.
$\hat{X}_t(m)$	Forecast for m periods ahead from origin t.
e_t	One-step-ahead forecast error, $e_t = X_t - \hat{X}_{t-1}$.
	$e_t(m)$ should be used for other forecast origins.

TABLE 2.1: Notation used to describe exponential smoothing models (same as in Gardner Jr (2006))

No trend, no seasonality

This type is the simple exponential smoothing method by Brown (1962). Following Gardner Jr (2006) there will be two separate equations for each model, one using the recursive form and the other being the error correction form.

Given a series $\{X_t\}$ with $t = \{1, \ldots, T\}$, the simple exponential smoothing model in recursive form is given by

$$\begin{aligned} S_t &= \alpha X_t + (1-\alpha) S_{t-1} \\ \hat{X}_t(m) &= S_t \end{aligned} \tag{2.1}$$

while the error correction form is given by

$$\begin{aligned} S_t &= S_{t-1} + \alpha e_t \\ \hat{X}_t(m) &= S_t. \end{aligned} \tag{2.2}$$

Additive trend, no seasonality

The model with additive trend is that of Holt (1957) (Holt's linear method), adding a trend term to the estimated parameter one time step prior in 2.1, resulting in a linear trend. The recursive form is given by

$$\begin{aligned} S_t &= \alpha X_t + (1-\alpha)\left(S_{t-1} + T_{t-1}\right) \\ T_t &= \gamma\left(S_t - S_{t-1}\right) + (1-\gamma) T_{t-1} \\ \hat{X}_t(m) &= S_t + m T_t \end{aligned} \tag{2.3}$$

and the error correction form by

$$\begin{aligned} S_t &= S_{t-1} + T_{t-1} + \alpha e_t \\ T_t &= T_{t-1} + \alpha\gamma e_t \\ \hat{X}_t(m) &= S_t + m T_t. \end{aligned} \tag{2.4}$$

Damped-additive trend, no seasonality

In order to allow the trend to decay over time, Gardner Jr and McKenzie (1989) used a dampening factor, reducing the trend influence over the course of the forecast horizon m. The recursive form is given by

$$\begin{aligned} S_t &= S_{t-1} + \phi T_{t-1} + \alpha e_t \\ T_t &= \phi T_{t-1} + \alpha\gamma e_t \\ \hat{X}_t(m) &= S_t + \textstyle\sum_{i=1}^{m} \phi^i T_t \end{aligned} \tag{2.5}$$

and the error correction form by

$$\begin{aligned} S_t &= S_{t-1} R_{t-1} + \alpha e_t \\ R_t &= R_{t-1} + \alpha\gamma e_t / S_{t-1} \\ \hat{X}_t(m) &= S_t R_t^m. \end{aligned} \tag{2.6}$$

For all above equations it is assumed for $\alpha, \gamma \in [0,1]$ otherwise observations would gain influence the further they are away from the forecast. This becomes evident once we look at the expanded form of 2.1. We substitute the expression of S_{t-1} back into itself and thus arrive at the geometric progression

$$
\begin{aligned}
S_t &= \alpha X_t + (1-\alpha) S_{t-1} \\
&= \alpha X_t + \alpha(1-\alpha) X_{t-1} + (1-\alpha)^2 S_{t-2} \\
&= \alpha \left[X_t + (1-\alpha) X_{t-1} + (1-\alpha)^2 X_{t-2} + \cdots + (1-\alpha)^{t-1} X_1 \right] + (1-\alpha)^t X_0.
\end{aligned}
$$

These expressions can be defined analogously for 2.3 and 2.5. For ϕ, different values can be used to give the trend convex, linear or even concave shape.

2.1.2 Model selection

Selecting an appropriate model can be done in several different ways, of which a few will be presented here. Depending on the number of time series to forecast either aggregate or individual model selection can be used. Fildes (2001) comes to the conclusion that in aggregate selection, the damped-additive trend model is hard to beat, although individual selection does yield better results sometimes. The "individual selection of exponential smoothing methods, [however], is best described as inconclusive." (Gardner Jr, 2006, p. 28).

On the one hand there are selection criteria based on time-series characteristics, as described for example in Shah (1997) or Meade (2000), which led to promising results when applied to time series that were generated using one of the processes to identify, but when applied to other series the results were less convincing. On the other hand there are expert-systems that generate rules based on experience made from earlier forecasting procedures (see for example Collopy and Armstrong (1992) or Flores and Pearce (2000)).

The model selection used in this study will be based on information criteria, as they are easy to derive and readily available. The results in selecting the appropriate model, however, are not convincing either. Comparing the studies of Gardner Jr and McKenzie (1985) and Hyndman et al. (2002), one can find that only for a forecast horizon of two and 15, the AIC as an information criterion selected more accurate models rather than the aggregate choice of a damped-additive trend model. For this reason this study also fits a damped-additive trend model to the data regardless of the decision based on information criteria. Nevertheless does the choice of an information criterion as a model selector provide an easily accessible procedure as this study estimates parameters via state-space maximum likelihood, from which arbitrary information criteria can be derived handily. The choice and description of information criteria used in this study will be discussed in more detail in section SECCCTIONN!!!!!

2.1.3 The state-space model

In order to estimate the parameters for the above described models, this study uses an "innovations", single-source of error (SSOE) state-space model. The model framework is that

of Ord, Koehler, and Snyder (1997), which was expanded by Hyndman et al. (2002). The basic state space framework can be described by the following equations:

$$y_t = w\left(\boldsymbol{X}_{t-1}\right) + r\left(\boldsymbol{X}_{t-1}\right)\varepsilon_t \tag{2.7a}$$

$$\boldsymbol{x}_t = f\left(\boldsymbol{X}_{t-1}\right) + g\left(\boldsymbol{X}_{t-1}\right)\varepsilon_t \tag{2.7b}$$

with y_t being the observation at time t, $\boldsymbol{X}_t$ the state vector containing unobserved components that describe the level, trend and seasonality of the series and w, r, f, g are continuous functions with $w, r : \mathbb{R}^p \to \mathbb{R}$ and $f, g : \mathbb{R} \to \mathbb{R}$. $\{\varepsilon_t\}$ is a Gaussian white noise process with variance σ^2. Equation 2.7a is called the *measurement equation* as it measures the relationship between the unobserved states $\boldsymbol{X}_{t-1}$ and the observation y_t. Equation 2.7b is called *transition equation*, describing the evolution of the states over time.

All of the equations from 2.1 to 2.5 can be translated into state-space terminology. To ensure that this thesis is self-contained I will present the equations in their state-space equations below, however they are exactly taken from Hyndman et al. (2002), which discusses them in great detail. The models in this study use additive trends, such that $r(\boldsymbol{X}_{t-1}) = 1$ and if we define the one-step forecast made in period $t-1$ as $\mu_t = F_{(t-1)+1} = w\left(\boldsymbol{X}_{t-1}\right)$. Further defining $e_t = r\left(X_{t-1}\right)\varepsilon_t$ we can rewrite $Y_t = \mu_t + e_t$ which is equal to $Y_t = \mu_t + \varepsilon_t$ for additive errors. Suppose that l_t is the level of our series at time t, b_t is the slope of the series at time t and α, γ, ϕ are constants, Brown's method from 2.1 becomes

$$\begin{aligned} \mu_t &= l_{t-1} \\ l_t &= l_{t-1} + \alpha\varepsilon_t. \end{aligned} \tag{2.8}$$

Holt's linear trend method becomes

$$\begin{aligned} \mu_t &= l_{t-1} + b_{t-1} \\ l_t &= l_{t-1} + b_{t-1} + \alpha\varepsilon_t \\ b_t &= b_{t-1} + \alpha\gamma\varepsilon_t, \end{aligned} \tag{2.9}$$

and Gardener's damped additive trend becomes

$$\begin{aligned} \mu_t &= l_{t-1} + b_{t-1} \\ l_t &= l_{t-1} + b_{t-1} + \alpha\varepsilon_t \\ b_t &= \phi b_{t-1} + \alpha\gamma\varepsilon_t. \end{aligned} \tag{2.10}$$

2.1.4 Parameter estimation and initial values

As we assumed the error term $\{\varepsilon_t\}$ to be Gaussian noise, the likelihood function will also be a Gaussian likelihood. In essence, the joint density of the series is the weighted product of the densities of the individual innovations:

$$p\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{X}_0, \sigma^2\right) = \prod_{t=1}^{n} p\left(\varepsilon_t\right) / \left|r\left(\boldsymbol{X}_{t-1}\right)\right|, \tag{2.11}$$

with $\boldsymbol{X}_0$ being the seed states or initial values. This given, the Gaussian likelihood can be described as

$$\mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{X}_0, \sigma^2 \mid \boldsymbol{y}\right) = \left(2\pi\sigma^2\right)^{-n/2} \left|\prod_{t=1}^{n} r\left(\boldsymbol{X}_{t-1}\right)\right|^{-1} \exp\left(-\frac{1}{2}\sum_{t=1}^{n} \varepsilon_t^2/\sigma^2\right), \tag{2.12}$$

and the log likelihood is given as

$$\log \mathcal{L} = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \sum_{t=1}^{n} \log\left|r\left(\boldsymbol{X}_{t-1}\right)\right| - \frac{1}{2}\sum_{t=1}^{n} \varepsilon_t^2/\sigma^2. \tag{2.13}$$

If we now set the partial derivative with respect to σ^2 to zero, we can calculate the maximum likelihood estimate of the variance as

$$\hat{\sigma}^2 = n^{-1}\sum_{t=1}^{n} \varepsilon_t^2. \tag{2.14}$$

This allows us to write the likelihood from 2.12 without dependency from the variance as

$$\mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{X}_0 \mid \boldsymbol{y}\right) = \left(2\pi\hat{\sigma}^2\right)^{-n/2} \left|\prod_{t=1}^{n} r\left(\boldsymbol{x}_{t-1}\right)\right|^{-1}. \tag{2.15}$$

Recapitulating from above we know that for additive models $r(\boldsymbol{X}_{t-1}) = 1$, and thus by taking the twice the negative logarithm we arrive at

$$\begin{aligned} \mathcal{L}^*\left(\boldsymbol{\theta}, \boldsymbol{X}_0\right) &= n\log\left(2\pi\right) + n\log\left(\hat{\sigma}^2\right) \\ &= n\log\left(2\pi\right) + n\log\left(\sum_{t=1}^{n} \varepsilon_t^2\right). \end{aligned} \tag{2.16}$$

To arrive at maximum likelihood estimates for the parameters this study minimizes 2.16.

One last step prior to fitting the model, the initial values for $\boldsymbol{X}_0$ must be set. Here we use $\alpha = \gamma = 0.5$ and $\phi = 0.9$, while for the level l_0 of the series we compute a linear trend of the first ten observations and use the intercept as a starting value. For the initial trend coefficient b_0 we use the slope of the before computed trend.

2.2 Autoregressive integrated moving average (ARIMA)

2.2.1 The linear filter

The basic framework for the stochastic models presented here, are based on an idea of Yule (1927) that an observable time series y_t in which past successive values are dependent on each other can be seen as a series of independent shocks ε_t. These shocks can be modeled as

random draws from a fixed distribution with mean $\mu = 0$ and time independent variance σ_ε^2. In most cases $\varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$. The thus generated sequence of random values $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2} \ldots$, is called a white noise process.

The process that transforms ε_t into y_t is then called a *linear filter*, that takes in a weighted sum of previous shocks, such that

$$\begin{aligned} z_t &= \mu + \varepsilon_t + \zeta_1 \varepsilon_{t-1} + \zeta_2 \varepsilon_{t-2} + \ldots \\ &= \mu + \zeta(\boldsymbol{B})\varepsilon_t. \end{aligned} \tag{2.17}$$

The second line of the above equation makes use of the *backwards shift operator*, which is is defined by $\boldsymbol{B}y_t = y_{t-1}$ and therefore $\boldsymbol{B}^m y_t = y_{t-m}$, and is called the *transfer function*, defined as

$$\zeta(\boldsymbol{B}) = 1 + \zeta_1 \boldsymbol{B} + \zeta_2 \boldsymbol{B}^2 + \ldots$$

The sequence of weights $\{\zeta_j\}$ can have arbitrary lengths and determines the processes' stationarity. If the weights are absolutely summable, i.e. $\sum_{j=0}^{\infty} \left|\zeta_j\right| < \infty$, the filter is called stable and the underlying process is stationary. Then, μ is the mean about which the process varies; otherwise μ has no specific interpretation, besides giving information about the process level.

2.2.2 The autoregressive model

The idea behind the autoregressive model is that the current value of a series can be expressed as the result of regressing past values of the series plus a random shock onto the current one, hence the name *autoregressive*. For the mathematical description we will stick with the above notation: Let $y_t, y_{t-1}, y_{t-2}, \ldots$ be a series with equidistant time steps and further let $\tilde{y}_t = y_t - \mu$ be the series deviation from its level or mean. Now,

$$\tilde{y}_t = \phi_1 \tilde{y}_{t-1} + \phi_2 \tilde{y}_{t-2} + \cdots + \phi_p \tilde{y}_{t-p} + \varepsilon_t \tag{2.18}$$

is an autoregressive process of order p (AR(p)). We can also make use of the backwards shift operator to define the transfer function of the process as

$$\phi(\boldsymbol{B}) = 1 - \phi_1 \boldsymbol{B} - \phi_2 \boldsymbol{B}^2 - \cdots - \phi_p \boldsymbol{B}^p \tag{2.19}$$

to arrive at the economical description of the process as

$$\phi(\boldsymbol{B})\tilde{y}_t = \varepsilon_t. \tag{2.20}$$

The number of estimated parameters of the model is $p + 2$, i.e. $\mu_y, \sigma_\varepsilon^2and\phi_j$ with $j = 1, \ldots, p$. An AR(p) process can be either stationary or non-stationary depending on the unit roots of

its characteristic polynomial $\phi(\boldsymbol{B})$. If the absolute value of all roots $\phi(\boldsymbol{B}) = 0$ lie outside the unit circle, the process is stationary.

2.2.3 The moving average model

As we have seen above, the autoregressive model regresses past terms of the series plus a random shock onto the current value. The same idea is adapted for the moving average model with the change that now the current value is regessed onto a linear combination of past shocks. The model can be described as

$$\tilde{y}_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}, \tag{2.21}$$

which is called a *moving average process* of order q (MA(q)). Again using the backwards shift operator we are able to define the moving average operator as

$$\theta(\boldsymbol{B}) = 1 - \theta_1 \boldsymbol{B} - \theta_2 \boldsymbol{B}^2 - \cdots - \theta_q \boldsymbol{B}^q,$$

which enables us to economically write the model as

$$\tilde{y}_t = \theta(\boldsymbol{B})\varepsilon_t. \tag{2.22}$$

The number of parameters to be estimated is again $q + 2$ following the same logic as above.

2.2.4 Autoregressive moving average models

Both of the above models can also be combined in order to achieve greater flexibility when modeling time series. The resulting model is called *autoregressive moving average*, or ARMA and is defined by

$$\tilde{y}_t = \phi_1 \tilde{z}_{t-1} + \cdots + \phi_p \tilde{z}_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}. \tag{2.23}$$

For easier identification the model order is usually written in parantheses after the model name, like ARMA(p,q). For a more economical writing stlye, often either the summation notation

$$\tilde{y}_t = \varepsilon_t + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{2.24}$$

or the operator notation

$$\phi(\boldsymbol{B})\tilde{y}_t = \theta(\boldsymbol{B})\varepsilon_t \tag{2.25}$$

is used. The combined model now has $p + q + 2$ parameters that need to be estimated from the data.

2.2.5 Achieving stationarity

The majority of time series encounterd in the real world are not stationary, meaning they do not fluctuate around a fixed mean but much rather exibit seasonal or global trends. However, it is possible for those series to still exhibit homogeneous behavior over time when we somehow account for those changes in level. Mathematically non-stationarity occurs if one ore more roots of the processes' characteristc polynomial lie exactly on the unit circle while all others lie outside. Following Box and Jenkins (1976), if such a process has exactly d roots on the unit circle, we can use the augmented autoregressive operator $\varphi(B) = \phi(B)(1-B)^d$ to transform the model back to stationarity via

$$\varphi(\boldsymbol{B})y_t = \phi(\boldsymbol{B})(1-\boldsymbol{B})^d y_t = \theta(\boldsymbol{B})\varepsilon_t. \tag{2.26}$$

So, a model that exhibits non-stationary, but homogeneous behavior can be brought back to stationarity by using the d-th differnece instead. We can now define some $b_t = (1-\boldsymbol{B})^d y_t$ to arrive at the well known ARIMA(p,d,q) model defined by

$$b_t = \phi_1 b_{t-1} + \ldots \phi_p b_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \tag{2.27}$$

or in summation notation

$$w_t = \mu + \varepsilon_t + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}. \tag{2.28}$$

2.2.6 Model identification and selection

The process of identification and selection can take place in two different manners. First, one can follow the *Box Jenkins approach*, that is inspecting the autocorrelation and partial autocorrelation functions to determine the order of the underlying process and second a grid-search algorithm in combination with some information criterion can be used to search though the models' parameter space to choose the model that minimizes the criterion

Autocovariance and Autocorrelation function

Once a process is stationary by means of the definition above, it features some important stochastic stabilities especially time invariant properties. In particular this means that its properties, such as the joint probability, mean or variance are unaffected by a change of time origin. Suppose a stationary stochastic process $\{y_t\}$ is observed for some period of time $y_{t_1}, y_{t_2}, \ldots, y_{t_m}$, then under stationarity conditions all describing moments and the joint probability remain equal if the series gets shiftet by some arbitrary integer s, say $y_{t_1+s}, y_{t_2+s}, \ldots, y_{t_m+s}$.

The time-invariant mean of a (continuous) stationarity process with a joint probability function $p(\cdot)$ is given by

$$\mu = E[y_t] = \int_{-\infty}^{\infty} y p(y) dy \tag{2.29}$$

and its variance by

$$\sigma_y^2 = E\left[(y_t - \mu)^2\right] = \int_{-\infty}^{\infty} (y - \mu)^2 p(y) dy. \tag{2.30}$$

As in real-world applications the series are always of finite length, the mean and variance must be estimated from the data. For a series with N observations the mean is estimated by

$$\bar{y} = \frac{1}{N} \sum_{t=1}^{N} y_t \tag{2.31}$$

and the variance via

$$\hat{\sigma}_y^2 = \frac{1}{N} \sum_{t=1}^{N} (y_t - \bar{y})^2 . \tag{2.32}$$

The stationarity from above implies that the joint probability distribution is the same for all times that are a constant value apart. It follows herefrom that also the covariance between two values y_t and y_{t+k} must be the same for every t. The separation integer k is also called the *lag value*, and therefore defines the autocovariance function at lag k as

$$\gamma_k = \operatorname{cov}[y_t, y_{t+k}] = E[(y_t - \mu)(y_{t+k} - \mu)] . \tag{2.33}$$

Equivalently, the autocorrelation function is defined as

$$\rho_k = \frac{E[(y_t - \mu)(y_{t+k} - \mu)]}{\sqrt{E\left[(y_t - \mu)^2\right] E\left[(y_{t+k} - \mu)^2\right]}} \tag{2.34}$$

Bibliography

Ariyo, Adebiyi A, Adewumi O Adewumi, and Charles K Ayo (2014). "Stock price prediction using the ARIMA model". In: *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*. IEEE, pp. 106–112.

Boser, Bernhard E, Isabelle M Guyon, and Vladimir N Vapnik (1992). "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152.

Box, George EP and Gwilym M Jenkins (1976). "Time series analysis. Forecasting and control". In: *Holden-Day Series in Time Series Analysis*.

Brown, Robert Goodell (1962). "Smoothing, forecasting and prediction of discrete time series". In:

Chatfield, Chris et al. (2001). "A new look at models for exponential smoothing". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 50.2, pp. 147–159.

Collopy, Fred and J Scott Armstrong (1992). "Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations". In: *Management science* 38.10, pp. 1394–1414.

Conejo, Antonio J et al. (2005). "Forecasting electricity prices for a day-ahead pool-based electric energy market". In: *International journal of forecasting* 21.3, pp. 435–462.

Dimoulkas, Ilias, Peyman Mazidi, and Lars Herre (2019). "Neural networks for GEFCom2017 probabilistic load forecasting". In: *International Journal of Forecasting* 35.4, pp. 1409–1423.

Fildes, RA (2001). "Beyond forecasting competitions". In: *International Journal of Forecasting* 17.4, pp. 556–560.

Flores, Benito E and Stephen L Pearce (2000). "The use of an expert system in the M3 competition". In: *International Journal of Forecasting* 16.4, pp. 485–496.

Gardner Jr, Everette S (1985). "Exponential smoothing: The state of the art". In: *Journal of forecasting* 4.1, pp. 1–28.

– (2006). "Exponential smoothing: The state of the art—Part II". In: *International journal of forecasting* 22.4, pp. 637–666.

Gardner Jr, Everette S and ED McKenzie (1985). "Forecasting trends in time series". In: *Management science* 31.10, pp. 1237–1246.

Gardner Jr, Everette S and Ed McKenzie (1989). "Note—Seasonal exponential smoothing with damped trends". In: *Management Science* 35.3, pp. 372–376.

Glorot, Xavier and Yoshua Bengio (2010). "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on*

artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, pp. 249–256.

Harris, John L and Lon-Mu Liu (1993). "Dynamic structural analysis and forecasting of residential electricity consumption". In: *International Journal of Forecasting* 9.4, pp. 437–455.

Hein, Scott E and Raymond E Spudeck (1988). "Forecasting the daily federal funds rate". In: *International Journal of Forecasting* 4.4, pp. 581–591.

Hinton, Geoffrey E, Simon Osindero, and Yee-Whye Teh (2006). "A fast learning algorithm for deep belief nets". In: *Neural computation* 18.7, pp. 1527–1554.

Ho, Tin Kam (1995). "Random decision forests". In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE, pp. 278–282.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Holt, Charles C (1957). "Forecasting trends and seasonals by exponentially weighted averages. carnegie institute of technology". In: *Pittsburgh ONR memorandum*.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5, pp. 359–366.

Hu, MJC and Halbert E Root (1964). "An adaptive data processing system for weather forecasting". In: *Journal of Applied Meteorology and Climatology* 3.5, pp. 513–523.

Hyndman, RJ (2001). "It's time to move from what to why'". In: *International Journal of Forecasting* 17.1, pp. 567–570.

Hyndman, Rob J et al. (2002). "A state space framework for automatic forecasting using exponential smoothing methods". In: *International Journal of forecasting* 18.3, pp. 439–454.

Jarrett, Kevin et al. (2009). "What is the best multi-stage architecture for object recognition?" In: *2009 IEEE 12th international conference on computer vision*. IEEE, pp. 2146–2153.

Laptev, Nikolay et al. (2017). "Time-series extreme event forecasting with neural networks at uber". In: *International conference on machine learning*. Vol. 34, pp. 1–5.

LeCun, Yann, Yoshua Bengio, et al. (1995). "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10, p. 1995.

LeCun, Yann et al. (1989). "Backpropagation applied to handwritten zip code recognition". In: *Neural computation* 1.4, pp. 541–551.

Li, Yaguang et al. (2017). "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting". In: *arXiv preprint arXiv:1707.01926*.

Meade, Nigel (2000). "Evidence for the selection of forecasting methods". In: *Journal of forecasting* 19.6, pp. 515–535.

Nair, Vinod and Geoffrey E Hinton (2010). "Rectified linear units improve restricted boltzmann machines". In: *Icml*.

Ord, John Keith, Anne B Koehler, and Ralph D Snyder (1997). "Estimation and prediction for a class of dynamic nonlinear statistical models". In: *Journal of the American Statistical Association* 92.440, pp. 1621–1629.

Rosenblatt, Frank (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1985). *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science.

Salinas, David et al. (2020). "DeepAR: Probabilistic forecasting with autoregressive recurrent networks". In: *International Journal of Forecasting* 36.3, pp. 1181–1191.

Shah, Chandra (1997). "Model selection in univariate time series forecasting using discriminant analysis". In: *International Journal of Forecasting* 13.4, pp. 489–500.

Snyder, RD (1985). "Recursive estimation of dynamic linear models". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 272–276.

Taylor, James W (2003). "Exponential smoothing with a damped multiplicative trend". In: *International journal of Forecasting* 19.4, pp. 715–725.

Tkacz, Greg (2001). "Neural network forecasting of Canadian GDP growth". In: *International Journal of Forecasting* 17.1, pp. 57–69.

Wen, Ruofeng et al. (2017). "A multi-horizon quantile recurrent forecaster". In: *arXiv preprint arXiv:1711.11053*.

Winters, Peter R (1960). "Forecasting sales by exponentially weighted moving averages". In: *Management science* 6.3, pp. 324–342.

Yule, George Udny (1927). "VII. On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 226.636-646, pp. 267–298.

Declaration of Authorship

I, Niklas PAULIG, declare that this thesis titled, "Forecasting univariate stock indices with parametric models and neural networks: A comparison" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:
