

Improving the accuracy of prediction using Singular Spectrum Analysis by incorporating internet activity

Dirk Jakobus Pretorius Badenhorst

Thesis presented in partial fulfilment of the requirements for the degree of Master of
Commerce in the department of Statistics and Actuarial Science in the Faculty of
Economic Management Sciences at Stellenbosch University



Supervisor: Prof. Sarel J. Steel
March 2013

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof, that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining a qualification.

Name: Dirk Jakobus Pretorius Badenhorst

Date: 30 November 2012

Abstract

Researchers and investors have been attempting to predict stock market activity for years. The possible financial gain that accurate predictions would offer lit a flame of greed and drive that would inspire all kinds of researchers. However, after many of these researchers have failed, they started to hypothesize that a goal such as this is not only improbable, but impossible.

Previous predictions were based on historical data of the stock market activity itself and would often incorporate different types of auxiliary data. This auxiliary data ranged as far as imagination allowed in an attempt to find some correlation and some insight into the future, that could in turn lead to the figurative pot of gold. More often than not, the auxiliary data would not prove helpful. However, with the birth of the internet, endless amounts of new sources of auxiliary data presented itself. In this thesis I propose that the near infinite amount of data available on the internet could provide us with information that would improve stock market predictions.

With this goal in mind, the different sources of information available on the internet are considered. Previous studies on similar topics presented possible ways in which we can measure internet activity, which might relate to stock market activity. These studies also gave some insights on the advantages and disadvantages of using some of these sources. These considerations are investigated in this thesis.

Since a lot of this work is therefore based on the prediction of a time series, it was necessary to choose a prediction algorithm. Previously used linear methods seemed too simple for prediction of stock market activity and a new non-linear method, called Singular Spectrum Analysis, is therefore considered. A detailed study of this algorithm is done to ensure that it is an appropriate prediction methodology to use. Furthermore, since we will be including auxiliary information, multivariate extensions of this algorithm are considered as well. Some of the inaccuracies and inadequacies of these current multivariate extensions are studied and an alternative multivariate technique is proposed and tested. This alternative approach addresses the inadequacies of existing methods.

With the appropriate methodology chosen and the appropriate sources of auxiliary information chosen, a concluding chapter is done on whether predictions that includes auxiliary information (obtained from the internet) improve on baseline predictions that are simply based on historical stock market data.

Abstrak

Navorsers en beleggers is vir jare al opsoek na maniere om aandeelpryse meer akkuraat te voorspel. Die moontlike finansiële implikasies wat akkurate vooruitskattings kan inhou het 'n vlam van geldgierigheid en dryf wakker gemaak binne navorsers regoor die wêreld. Nadat baie van hierdie navorsers onsuksesvol was, het hulle begin vermoed dat so 'n doel nie net onwaarskynlik is nie, maar onmoontlik.

Vorige vooruitskattings was bloot gebaseer op historiese aandeelprys data en sou soms verskillende tipes bykomende data inkorporeer. Die tipes data wat gebruik was het gestrek so ver soos wat die verbeelding toegelaat het, in 'n poging om korrelasie en inligting oor die toekoms te kry wat na die figuurlike pot goud sou lei. Navorsers het gereeld gevind dat hierdie verskillende tipes bykomende inligting nie van veel hulp was nie, maar met die geboorte van die internet het 'n oneindige hoeveelheid nuwe bronne van bykomende inligting bekombaar geraak. In hierdie tesis stel ek dus voor dat die data beskikbaar op die internet dalk vir ons kan inligting gee wat verwant is aan toekomstige aandeelpryse.

Met hierdie doel in die oog, is die verskillende bronne van inligting op die internet gebestudeer. Vorige studies op verwante werk het sekere spesifieke maniere voorgestel waarop ons internet aktiwiteit kan meet. Hierdie studies het ook insig gegee oor die voordele en die nadele wat sommige bronne inhou. Hierdie oorewegings word ook in hierdie tesis bespreek.

Aangesien 'n groot gedeelte van hierdie tesis dus gebaseer word op die vooruitskating van 'n tydreeks, is dit nodig om 'n toepaslike vooruitskattings algoritme te kies. Baie navorsers het verkies om eenvoudige lineêre metodes te gebruik. Hierdie metodes het egter te eenvoudig voorgekom en 'n relatiewe nuwe nie-lineêre metode (met die naam "Singular Spectrum Analysis") is oorweeg. 'n Deeglike studie van hierdie algoritme is gedoen om te verseker dat die metode van toepassing is op aandeelprys data. Verder, aangesien ons gebruik wou maak van bykomende inligting, is daar ook 'n studie gedoen op huidige multivariaat uitbreidings van hierdie algoritme en die probleme wat dit inhou. 'n Alternatiewe multivariaat metode is toe voorgestel en getoets wat hierdie probleme aanspreek.

Met 'n gekose vooruitskattingsmetode en gekose bronne van bykomende data is 'n gevolgtrekkende hoofstuk geskryf oor of vooruitskattings, wat die bykomende internet data inkorporeer, werklik in staat is om te verbeter op die eenvoudige vooruitskattings, wat slegs gebaseer is op die historiese aandeelprys data.

Acknowledgements

To my parents that passionately supported me throughout the entirety of my life, thank you.

To Prof S.J. Steel for very thorough and insightful guidance, thank you.

To MIH media lab for not only funding this thesis but also for providing a workspace unlike any other, thank you.

To old friends for keeping me sane and providing copious amounts of coffee, thank you.

To new friends that reminded me to remain balanced, thank you.

Contents

1	Introduction	1
2	Univariate SSA	7
2.1	Basic SSA algorithm	8
2.1.1	Filtering	8
2.1.2	Prediction	12
2.2	The parameters	12
2.2.1	Choice of the window length	13
2.2.2	Choice of rank	15
2.3	Simulation study	20
3	Baseline predictions	30
3.1	AR algorithm applied to stock market data	31
3.2	SSA algorithm applied to stock market data	34
3.3	Comparison and conclusion	40
4	Multivariate SSA	43
4.1	Current MSSA methods	43
4.1.1	Horizontal multivariate SSA algorithm	44
4.1.2	Vertical multivariate SSA algorithm	45
4.1.3	Disadvantages of HMSSA and VMSSA	46
4.2	Bayesian approach to combining time series	47
4.2.1	Multivariate normality of singular vectors	48
4.2.2	Bayesian MSSA	56
4.3	Simulation study	58
4.3.1	Effect of noise component	58
4.3.2	Conclusion on the effect of noise component	72
4.3.3	Effect of similar signal component	73
4.3.4	Conclusions on the effect of similar signal component	78
4.3.5	Effect of different signal components	79

4.3.6	Conclusions on the effect of different signal components	90
4.3.7	Effect of scale differences	92
4.3.8	Conclusions on the effect of scale differences	94
5	Incorporating search volumes	94
5.1	Predicting an automobiles index	95
6	Incorporating Twitter frequency counts	110
6.1	Predicting Google stock prices	110
6.2	Predicting Microsoft stock prices	116
6.3	Predicting Apple stock prices	121
7	Incorporating Twitter sentiment	126
7.1	Predicting Google stock prices	127
7.2	Predicting Microsoft stock prices	130
7.3	Predicting Apple stock prices	134
8	Conclusion	138
	Bibliography	141
	Appendix A	143

1 Introduction

Imagine for a moment travelling back to a time without the internet... A time where Google cannot provide quick answers to your queries. A time where Wikipedia cannot present quick peer reviewed information on both the present and the past. A time where News24 cannot update you on provincial, national and international news on an up-to-the-minute basis. A time where Amazon cannot assist you in buying products and having them delivered without leaving the comfort of your home. A time where PriceCompare cannot advise you where to find the best prices over a range of shops. A time where internet reviews cannot forewarn you against the cons or inform you on the pros of products that you are interested in. A time where forums cannot instruct you how to solve common technical problems. A time where Facebook cannot quickly reveal what your friends are up to. A time where Twitter cannot enlighten you with the opinions of like-minded people on all kinds of topics.

The internet is now a much more prominent part of our life in comparison with a decade ago. According to internetworldstats.com, there were about 360 million internet users world wide in the year 2000. This grew to a staggering 2.3 billion users by 2012 - A growth of 530%! Apart from the immense growth in internet users, the ease with which we can connect to the internet has also increased. To a large extent, we have become dependent on the real time connectivity the internet provides. According to a survey done by Pew internet in 2012, 88% of American adults own a cell phone. These Americans used their devices to access the internet and quickly solve everyday problems in the following manners:

- 35% used their cell phone to solve unexpected problems they had encountered.
- 30% used their cell phone to decide between different businesses or restaurants.
- 27% used their cell phone to obtain information to help settle an argument they were having.
- 25% used their cell phone to look up the price of a product on-line while they were in a store, to see if they could get a better price somewhere else.
- 24% used their cell phone to look up reviews of a product on-line while they were in a store.
- 23% used their cell phone to look up a score of a sporting event.
- 20% used their cell phone to get up-to-the-minute traffic information to find the fastest way to get somewhere.

In general, the survey showed that 51% used their cell phone at least once in 30 days to get information they needed right away. That means that roughly 140 million Americans are dependent on their cell phone for internet usage at least once a month.

The most obvious benefit of the internet must certainly be the ability to quickly obtain information on any topic one can think of. This is normally done by use of search engines. A Pew Internet's survey done in early 2012, showed that out of the 80% U.S. adults that use the internet, 59% uses a search engine on any given day. 83% of these said that they prefer the Google search engine. That amounts to roughly 125 million daily Google users!

The internet even changed the way we socialize. With the birth of social media websites such as Twitter, Facebook and Myspace, it became possible for us to tell our friends what we think, where we are, what we are doing, what we are planning on doing, what we are reading and so much more. The surveys done by Pew Internet also showed that 65% of adult internet users now say they use at least

one social networking site. 8% of these adults use twitter on a daily basis. This might not seem to be a significant proportion, but on the mere 6th birthday of Twitter, it announced that they have 140 million active users and post on average 340 million messages every day!

In this thesis, I propose that the internet not only changed our lives to such an extent that it is nearly unbearable to think back to a time without it, but that the enormous amount of information available on the internet also allows us to monitor exactly what is currently happening and perhaps also gain some insight into times to come. Herb Brody, senior editor at Technology Review, had the following to say about forecasting:

"Telling the future by looking at the past assumes that conditions remain constant. This is like driving a car by looking in the rear-view mirror."

The internet allows us to consider also what is happening in the present, i.e. current trends and breaking news. This would be analogous to looking out the side of your vehicle, hoping to obtain a peripheral glimpse of the road. This concept have been studied in several fields with some success.

One application that received attention from increasingly many researches, as the popularity of the internet grew, was that of early disease outbreak detection. Typically influenza pandemics were predicted based on indicators such as virological data and physician visits. The problem with indicators such as these was that they would typically be released every two weeks. Researchers then turned to internet data for a very important reason; real time information. If the internet presented real time information on current health prospects, flu outbreaks could be foreseen and addressed at an earlier stage, possibly resulting in a higher likelihood of effective treatment. Several internet-based applications were used to obtain information with the potential of predictive power. Some researchers hypothesized that internet search queries relating to health seeking advice could provide the necessary information; both Google and Yahoo seemed to provide helpful information (Ginsberg *et al*, 2009 and Polgreen *et al*, 2012). Blog data was also used as a possible alternative and it proved to be effective as a predictor variable (Corley *et al*, 2009). At a later stage, when Twitter became popular, Tweets were also shown to correlate with the number of people infected with influenza (Lampos and Cristianini, 2010). In an attempt to combine a wide range of useful internet data, several models were then researched that aggregated different web based data sources to form a system responding to collective internet information. An example of this has been researched by an organization called "HealthMap". This organization has a freely accessible web page (Brownstein *et al*, 2008), that combines over 20000 internet sources every hour, including news aggregators such as Google news, that monitors international influenza epidemics.

These studies showed firstly the beneficial properties of internet data; that is, availability, inexpensiveness and its real time nature. In this sense, using internet data as a predictor variable is extremely unique and superior to most other predictor variables. Secondly, these studies also give us an idea of the diverse number of web based applications providing easily obtainable data with predictive value. Search query volumes, blog and microblog data as well as news aggregators enabled these researchers to instantaneously obtain the information they needed. However, with the above referenced studies on disease outbreak detection, it is hypothesised that an increase in health conscious internet activity would be indicative of a disease pandemic. The statistical models used in these articles were therefore all based on detection of structural break models, as any deviation from some baseline internet activity would suggest more users being concerned about their health and therefore, possibly the start of a disease outbreak. Such models are therefore very simple and might not be able to generalize when applied to predictions in other fields.

More complex studies were also done in the entertainment industry, when researchers also quickly identified a possible use for the internet in product sales prediction; more specifically the prediction of box office revenues. Several papers were published studying the correlation between the number of blog posts containing a movie title and the revenue of said movie. For instance, Mishne and Glance (2006) considered the revenue of a movie during opening weekend together with relevant blog posts taken in a window starting one month prior to opening weekend and ending one month thereafter. Liu *et al* (2007) on the other hand correlated gross revenues with blog posts taken from one week before release date until 4 weeks thereafter. Both these studies produced significant correlations between number of blog mentions and movie revenues. With the growth of microblogging applications such as Twitter, correlations between Tweets and box office revenues could also be investigated (Asur and Huberman, 2010). In this paper, the authors not only showed that the rate at which people tweet about a certain movie is correlated with the movie's gross revenue, but also that using this tweet-rate as predictor variable outperformed traditionally used features such as the Hollywood stock exchange. All of the before mentioned researchers identified the blogosphere as a potential environment where the average Joe can give his opinion. Zhang and Skiena (2009) approached the problem differently and tried to analyse the opinion of the expert by considering the number of news articles obtained on the internet relating to a certain movie and correlating this with the movie's gross revenue .

In the above mentioned studies frequency counts of articles or posts that included the title of a certain movie were correlated with the revenue of the movie. However, this is based on the assumption that any news is good news, a similar assumption as with studies on disease outbreak detections. For movies such an assumption is certainly not a valid one. A piece of text could either express opinion on the quality of the film or the lack thereof. This is where studies in this area should be more carefully considered than those relevant to disease outbreak studies. Raw counts of posted text is not sufficient. Sentiment aware models should therefore also be considered. In most of the studies mentioned in the previous paragraph, a comparison was done between sentiment aware models and raw frequency counts. The sentiment aware models produced better correlations (or at least as good) than the models based on frequency counts. In most situations these sentiment aware models are necessary since it allows us to not only capture the hype around a certain topic, but also whether the opinion of the masses are positive or negative.

The concept of using the internet to improve predictions can be applied to nearly any imaginable field. Researchers in numerous fields saw the value of the opinions and the needs of the public, freely available on the internet. These researchers hypothesised that the internet data is not only correlated with certain current events, but that it also enables us to predict the nearby future with more accuracy. Focussing on search engines such as Google and Yahoo, these researchers proposed that search query volumes could be helpful in the prediction of unemployment figures (Varian and Choi, 2011 & Guzman, 2011 & Ettredge *et al*, 2005 & D'Amuri and Marcucci, 2009), house prices (Wu and Brynjolfsson, 2009) and even upcoming and trending news (Radinsky *et al*, 2008). Even though the proposed models in these papers were relatively simple ones, such as simple autoregressive models, they seemed to provide statistically significant improvements in prediction accuracy.

The papers referenced above illustrated several useful properties of internet data. That is:

- The property of real time information, freely available in one of numerous forms, as shown by studies on disease outbreak detection.
- The possibility of using sentiment aware models on text to extract the tone of any piece of text

available on the internet, as shown by studies on box office revenues.

- The predictive power of internet data on general economic activity.

In this thesis I plan to combine these properties of internet data and propose that the activity on the internet could contain information that would be beneficial to the prediction of stock market activity. We already saw that the internet had predictive power over some factors that indirectly influence economic activity such as house prices and unemployment figures. More directly, studies were also done on the predicting economic indicators such as private consumption (Schmidt and Vosen, 2009) and inflation (Guzman, 2011). These economic indicators are often used as predictor variables for prediction of stock market indices. Why therefore not try to apply similar methods directly to stock market prices? Through constant monitoring of internet activity, we will be able to observe unexpected changes in activity relevant to some of these economic indicators. And since the data is immediately available, we will be able to react to these unexpected changes immediately, where other investors might only react at a later stage. An increase in activity will therefore instruct us to react, while sentiment aware models might tell us how to react. This is a concept that has also received some attention by fellow researchers.

During the 90's, several papers were written on prediction of stock returns using financial forums such as Raging Bull and Yahoo! Finance. Tumarkin and Whitlaw (2001) presented one such paper. They considered 73 individual internet service companies and their adjusted stock returns together with data mined from Raging Bull. Since research such as this is predominantly relevant to short term investments, all posts that were tagged as potential long term investments could be disregarded. Furthermore, each post was also tagged with a score indicating how strongly its writer encourages buying or selling. Sentiment analysis was therefore not needed since the individual posts already contained a variable describing the opinion of the author. Initially Tumarkin and Whitlaw (2001) merely considered possible correlations between the individual stocks and the aggregate opinion expressed on the forum on that day. Results indicated only a few significant correlations at non-leading lags of the forum posts, i.e. today's forum was not significantly correlated with tomorrow's stock return. The postings were however significantly correlated with returns (as well as trading volumes) on the same day of forum posts. If the day had a positive indication in general, stock prices generally increases significantly on that same day. When subjecting the data to a simple vector autoregression prediction algorithm, similar results were found. This paper therefore resulted in inconclusive results. However we have to take several things into consideration. Firstly, since this paper was written in the early 90's, the amount of internet users were not as many then as they are today. Aggregate opinions of the people on the forum therefore did perhaps not represent the opinion of the population sufficiently. Secondly, even in the 90's, correlations were significant for the stock returns as well as trading values on the same day of the post. For all practical terms, we do not necessarily need to know whether the stock prices will change significantly by tomorrow, we merely need to know whether it will do so within the next hour (maybe even less). Taking both these aspects into consideration, the concept might achieve significantly better results now, nearly 20 years later.

Similar results were also obtained in an analysis done by Preis *et al* (2010). In this paper the authors attempt to find correlations between the returns of 500 of the companies that form part of the S&P500 index and information obtained from the web based application, Google Trends. In this paper, the exact company name or abbreviation thereof was used as search query and linear cross correlations were obtained between weekly search volume changes and corresponding weekly stock returns. Also, cross correlation were calculated between weekly search volume changes and stock volume changes as a proxy for stock market volatility. It seemed that the only significant cross correlation was that of search query volumes with trading volumes at zero lag, i.e. search query volumes coincided with stock market activity,

but had no predictive power over tomorrow's stock returns. This being said, the authors concluded by stating that:

"Increasing transaction volumes of stocks coincide with an increasing search volume and vice versa. Thus, one can conclude that search volume reflects the present attractiveness of trading a stock. But it seems that neither buying transactions nor selling transactions are preferred when one detects an increased search volume. Thus, the commonly accepted reasons for financial market movements (news and volume) are clearly linked together because news should be the most likely reason for searching company names in Internet search engines."

These might once again seem like unsatisfactory results. However, by concluding that stock market activity is affected by breaking news is a favourable result, since I propose that the internet allows us to be the first to respond to news. Furthermore, this paper completely disregards the benefit provided by the real time nature of the data available on the internet since it is based on data taken weekly. Once again, as mentioned in the discussion of the studies done by Tumarkin and Whitlaw (2001), predicting significant changes in today's stock market activity is helpful as predicting significant changes for tomorrow's (or next week's). We do not therefore necessarily need significant correlations at leading lags; correlations at zero lags might suffice. Another undesirable conclusion lies in the insignificant correlation of today's search volumes with today's stock returns. A possible reason for this is the fact that the paper is not based on sentiment aware models. As mentioned before, not all news is good news. If a company is highly spoken of, we do not yet know what people are saying. Over a large data set, positive correlations and negative correlations might have merely averaged out, resulting in insignificant correlations. Including the sentiment of the news in the model might address this problem.

Studies in both disease outbreak detection and the entertainment industry showed that blogs and microblogs are two valid web based applications that we can use to measure internet activity. Gilbert and Karahalios (2010) therefore considered the possibility of correlating blog activity with stock market activity using a dataset of 20 million blog posts during 2008 and correlating said dataset with the S&P500 index. Each day was measured on an Anxiety index. This value is measured as the normalized aggregate number of posts classified as Anxious, Worried, Nervous or Fearful. The authors then continued by showing that the Anxiety index "Granger-causes" S&P500 stock returns; this implies that online blogs provide information not already apparent from market history data. Even though no explicit predictions were performed in this paper, Gilbert and Karahalios (2010) concluded that a one standard deviation raise in Anxiety index corresponds to returns 0.4% lower than expected during the following day. In another paper, Zhang *et al* (2010) found similar results when using the popular microblog application Twitter to obtain information with predictive power. Instead of merely focussing on Anxiety as a sign of concerned investors, he classified each individual tweet into one of 7 possible categories (Hope, Happy, Fear, Worry, Nervous, Anxious and Upset). The percentage of tweets that fell into each of these categories daily were correlated with market indices like S&P500, NASDAQ, Dow Jones and the VIX. The results showed that Hope, Fear and Worry were significantly positively correlated with the VIX for up to lag of three days and significantly negatively correlated with the remaining three indices. These emotions therefore seem to cause the volatility of the stock market to increase and puts downward pressure on the actual stock returns.

The two papers mentioned above are based on the study of behavioural economics. Researchers in this field believe that emotions can profoundly affect individual behaviour and decision-making. In this context, the hypothesis is that blog and microblog data contain information on public mood which in turn affects the collective decision making of investors. Both studies showed that negative emotions of

the public - such as fear, anxiety, worry and such - resulted in negative stock returns during the next day/few days. In some ways, this collective public mood can be regarded as a version of the Consumer Confidence index, since it provides a broad, forward looking barometer of worry. Even though people primarily discuss everyday life on blogs and microblogs, the results showed that data mined from such applications could anticipate changes in a seemingly unrelated system, since both these studies showed that there are significant correlations between blog activity and stock market activity. The studies even went as far as to conclude that internet activity caused these market deviations to a certain extent. Whether this causality is valid or not is irrelevant. The more important question is whether the data mined from the internet contains actual predictive power.

Bollen *et al* (2011) addressed the question of the predictive power of blog posts on market indices. This paper was also considered from a behavioural economics point of view. Each mined tweet was classified into one of 6 emotions (Calm, Alert, Vital, Sure, Kind, Happy). After showing that the number of Calm Twitter posts were significantly correlated with the Dow Jones Industrial Average for up to a 7 day lag, the authors also used a self organizing fuzzy neural network to predict the index closing prices with out-of-sample predictions. The results showed that predictions improved significantly when allowing for this internet based predictor variable.

Moving away from behavioural economics, an alternative hypothesis could also be considered. Where behavioural economics is based on the idea that general market activity could be affected by the general public mood, we could speculate that individual company stocks could be affected by the public's opinion regarding said company. In a paper, presented by Choudury *et al* (2008), it was proposed that internet data does not only contain predictive value for stock market indices, but also for four individual company stock prices (Google, Microsoft, Apple and Nokia). The stock returns for these companies were linearly correlated with some contextual properties of blog posts containing the name of the company. It seemed that this linear model was able to capture the occurrences of big changes in stock returns; however a non-linear support vector regressor outperformed the linear model in predictive power since it was also able to capture the subtle changes in market fluctuations.

Since research is proposing that one can use the data available on the internet to more accurately predict stock market returns, one can't help to ask the question: "Can the model be used to beat the market and make money?". This question was addressed by Wuthrich *et al* (1998). Using web pages obtained from the Wall Street Journal website, the authors attempted to predict some well known market indices. This paper is different from previous studies on stock market prediction in the sense that it makes use of the real time nature of internet data to the full extent by implementing an automatic trading algorithm. The methodology used for prediction is only vaguely described; however results showed that during 60 days that this algorithm was tested on, it accumulated an effective profit of 30% capital appreciation in one year.

The majority of the papers referenced above use simple linear models to predict future stock market values. Linear methods such as these are obviously incapable of capturing the complex fluctuations in stock market behaviour. I will therefore consider an alternative prediction algorithm in this thesis called Singular Spectrum Analysis (SSA). My second chapter will thoroughly discuss this algorithm and related concepts such as model selection. In Chapter 3, I will continue by discussing why this algorithm is chosen above typically used stock market prediction algorithms. We will then also consider this univariate prediction methodology in a practical scenario and how it will be used in order to obtain our baseline predictions. Chapter 4 discusses some common derivations of this technique when predicting a time series

in the presence of an auxiliary time series. These derivations and possible inadequacies will be discussed thoroughly, after which I will propose an alternative derivation of the SSA algorithm that can be used to selectively include components of an auxiliary time series. In the fifth chapter we start to test our hypothesis that internet activity can be used to more accurately predict stock market activity. We do so by using the proposed multivariate methodology to predict stock market activity in the presence of Google search volume data. Chapter 6 continues to assess the validity of our hypothesis by including textual data in the form of frequency counts. This textual data will be extracted from the web-based applications called Twitter. Finally, Chapter 7 will continue to study the effect of the inclusion of text based information in prediction; however, in this chapter we will incorporate text based sentiment instead of simple frequency counts. A concluding chapter will thereafter be written to summarize the findings and results of this thesis.

2 Univariate SSA

In my opinion, any study focused on prediction of stock market values should mention the Efficient Market Hypothesis (EMH). This hypothesis in its crudest form states that stock market prices are effectively impossible to forecast. Often, stock prices are modelled as a random walk (Martingale model), as it is consistent with the Efficient Market Hypothesis. According to this model, the price of a stock tomorrow (Y_{T+1}) is merely the price of the stock today (Y_T), with some random disturbance term (ϵ_{T+1}) added. This model can be formulated mathematically by Equation 1.

$$Y_{T+1} = Y_T + \epsilon_{T+1} \tag{1}$$

In this model, ϵ_{T+1} is often modelled as Gaussian white noise. That is, $\{\epsilon_t | t = 1, \dots, T+1\}$ are independently distributed according to the normal distribution with zero mean and some constant standard deviation. Because of the symmetric nature of the Gaussian distribution, this model therefore states that we have a 50% chance of correctly predicting whether a stock will rise or fall. If this is the case, it is clear to see that there is no possibility for economic gain.

This Efficient Market Hypothesis is just that; a hypothesis. Many researchers have attempted to prove the EMH wrong. Early attempts focussed on modelling the disturbances, ϵ_i , through simple linear models such as the autoregressive (AR) models. By doing so the observed returns ($\{R_t = Y_t - Y_{t-1} | t = 2, \dots, T\}$) are modelled and one can linearly predict tomorrow's return based on the model together with previously observed returns. If the model and its predictions are accurate, one can therefore gain some knowledge on future stock prices and financial gain could follow.

Simple linear models such as AR models are however unlikely to capture the complexity of stock market activity. Non-linear models should also be considered. One such model that recently started receiving attention from increasingly many researchers is Singular Spectrum Analysis (SSA). Apart from the non-linearity of SSA, the methodology is also quite flexible in the sense that it does not require too many assumptions on the structure of the time series at hand. SSA has been compared to other popular time series analysis methods (Hassani, 2007). These studies showed that SSA proved to be significantly superior to some of the more classical approaches of time series analysis (such as SARIMA and AR as well as the seasonal Holt Winters model) in terms of prediction. In this chapter I will therefore first discuss the SSA algorithm and its parameters and then continue by comparing SSA with the AR model in terms of predictive power with regards to stock market prices.

2.1 Basic SSA algorithm

The basic SSA algorithm is discussed extensively by Golyandina *et al* (2001). Alternatively one can also consider De Klerk (2002) for in an depth discussion on the SSA algorithm. In this discussion we describe the basic SSA algorithm and rationale.

Consider a given time series of length T , $Y_T = \{y_1, \dots, y_T\}$. Let us assume this time series has a signal component ($S_T = \{s_1, \dots, s_T\}$) and noise component ($N_T = \{n_1, \dots, n_T\}$) allowing us to decompose the time series according to Equation 2.

$$Y_t = S_t + N_t \quad (2)$$

One advantage of this model is that we need not make any explicit assumptions on the nature of either the signal component or the noise component, making it a very adaptable and flexible approach.

The SSA methodology is a two phase procedure based on the decomposition in Equation 2. The first phase is a filtering methodology that estimates the decomposition of the time series into noise and signal and therefore filters the time series of unwanted noise. Once the noise has been extracted, the second phase of SSA can be implemented to calculate predictions. These predictions are based on filtered signal, ensuring that predictions are not impaired by noise.

2.1.1 Filtering

The filtering phase in the SSA methodology also consists of four steps:

- Embedding
- Singular value decomposition
- Grouping
- Diagonal averaging

The embedding step creates a multivariate scenario from the univariate time series, allowing us to perform a multivariate analysis on univariate data. This multivariate data is then subjected to a Singular Value Decomposition (SVD), a powerful dimension reduction technique used in multivariate analysis that will allow us to decompose the multivariate data into several elementary components. In the grouping step, we consider the individual elementary components found from the SVD and decide on the actual decomposition into signal and noise and grouping the components appropriately. The final step is then diagonal averaging. This step is to ensure that the multivariate data is now in an appropriate form for us to transform it back to univariate data.

Embedding

SSA can be seen as a multivariate analysis of univariate data. Since a single time series is essentially univariate data, we have to find a way to represent this univariate data in multivariate form. We do so by constructing the trajectory matrix in the embedding step. For this step, we have to decide on a parameter called the window length, L . This represents the dimension of the multivariate data and the

choice of this parameter is important for the success of the SSA technique. To define this multivariate data, we define the trajectory matrix for a given time series in Equation 3.

$$Y : [L \times (T - L + 1)] = \begin{bmatrix} y_1 & y_2 & \cdots & y_{T-L+1} \\ y_2 & y_3 & \cdots & y_{T-L+2} \\ \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & \cdots & y_T \end{bmatrix} \quad (3)$$

Some important characteristics of this trajectory matrix should be noticed. The first thing we should take note of is that the choice of the window length directly controls the form of this trajectory matrix. We will later discuss the fact that we can interpret this trajectory matrix as $N - L + 1$ replicates of a vector of size L since the matrix contains $N - L + 1$ segments of L consecutive time series observations. A final very important feature of this trajectory matrix is that it is of a very definite structure. That is, the reverse diagonals of the matrix all contain the same value. A matrix with this structure is called a Hankel matrix.

Singular Value Decomposition

After the trajectory matrix has been constructed, we analyse the multivariate data by using SVD decomposition. This SVD step allows us to decompose our trajectory matrices into several basic components. This technique decomposes our trajectory matrix according to Equation 4.

$$Y = UDV' \quad (4)$$

In this decomposition, D is a diagonal matrix with elements $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_L$. The matrices U and V on the other hand are orthogonal matrices containing the left and right singular vectors of the Y matrix respectively ($U = [\underline{u}_1, \cdots, \underline{u}_L]$ and $V = [\underline{v}_1, \cdots, \underline{v}_L]$). We can therefore alternatively rewrite the decomposition according to equation 5.

$$\begin{aligned} Y &= [\underline{u}_1, \cdots, \underline{u}_L] \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_L \end{bmatrix} \begin{bmatrix} \underline{v}'_1 \\ \vdots \\ \underline{v}'_L \end{bmatrix} \\ &= \sum_{i=1}^L \lambda_i \underline{u}_i \underline{v}'_i \\ &= Y^1 + \cdots + Y^L \quad \text{where } Y^i = \lambda_i \underline{u}_i \underline{v}'_i \end{aligned} \quad (5)$$

The rationale behind this decomposition of Y into a summation of L single rank matrices can be described at the hand of the properties of the elementary matrix components as proved by the Eckhart-Young theorem (interested reader can see Eckhart Young, 1936). According to this theorem, the matrix of rank j that best approximates a given trajectory matrix (according to the Frobenius matrix norm as defined by Equation 6) can be calculated by $\sum_{i=1}^j Y^i$.

$$\|A\|_F = \sqrt{\text{trace}(AA^T)} \quad (6)$$

In other words, the j th single rank matrix, Y^j , attempts to capture the maximum amount of variation in Y left unexplained by the first $(j - 1)$ matrices, in the form of a single rank matrix. Since each of these individual components, Y^j , has column space spanned by \underline{u}_j , the SVD process can alternatively be described as an algorithm that finds a direction, \underline{u}_j , orthogonal to previous $(j - 1)$ vectors, that captures most of the deviation in the $N - L + 1$ observations in L -dimensional space. These directional vectors are the unit vectors with the same direction as the principal components of our trajectory matrix. A

practical implication of explaining the SVD in terms of directional vectors as above is the fact that we can interpret the obtained \underline{u}_j as the first step in the direction of answering the following question: what would a typical L consecutive values in a given time series look like?

Grouping

In the grouping step, we focus on the actual decomposition of a time series into noise and signal as per our original assumption. Once our time series has been converted into a trajectory matrix, this decomposition assumption can be reformulated by stating that the trajectory matrix of the given time series (Y) can be decomposed into a trajectory matrix describing the noise (N) and one describing the signal (S).

$$Y = S + N \quad (7)$$

Approximating the signal time series, S_T , is therefore equivalent to approximating the trajectory matrix of the signal time series, S .

Using the single rank matrices obtained in the SVD step, we can find a decomposition in the form of Equation 7 by associating some of the single rank matrices obtained from the SVD with noise and some of them with signal. Let \mathcal{I}_N be the set of indices of matrices associated with noise and \mathcal{I}_S the set of indices of matrices associated with signal. We can then find a decomposition into noise and signal according to Equation 8.

$$\begin{aligned} Y &= \sum_{i \in \mathcal{I}_S} Y^i + \sum_{i \in \mathcal{I}_N} Y^i \\ &= \hat{S} + \hat{Y} \end{aligned} \quad (8)$$

Once again by interpreting Equation 8 in terms of the individual left singular vectors, \underline{u}_j , allows us to gain further insight into this decomposition. Consider for the moment the projection of a given trajectory matrix on the column space spanned by the left singular vectors $\{\underline{u}_i | i \in \mathcal{I}_S\}$. If we were to let U_S denote the matrix containing the set of singular vectors associated with signal as columns, the projection matrix can be written as in Equation 9, since the left singular vectors are unit vectors that are orthogonal to each other.

$$P_{span(\{\underline{u}_i | i \in \mathcal{I}_S\})} = U_S U_S^T \quad (9)$$

An expression for the projection of a trajectory matrix, Y , onto this space therefore follows according to Equation 10. In this equation, U_S denotes the matrix containing the set of left singular vectors associated with signal as columns, U_N denotes matrix containing the remaining left singular vectors associated with noise as columns and V_S, V_N are defined similarly, but with appropriate right singular vectors as columns. Furthermore, D_S and D_N are diagonal matrices, the former containing the set of singular values $\{\lambda_i | i \in \mathcal{I}_S\}$ and the latter contains $\{\lambda_i | i \in \mathcal{I}_N\}$ on its diagonals.

$$\begin{aligned} P_{span(\{\underline{u}_i | i \in \mathcal{I}_S\})} Y &= U_S U_S^T Y \\ &= U_S U_S^T \begin{bmatrix} U_S & U_N \end{bmatrix} D V^T \\ &= U_S \begin{bmatrix} I_d & 0 \end{bmatrix} \begin{bmatrix} D_S & 0 \\ 0 & D_N \end{bmatrix} \begin{bmatrix} V_S & V_N \end{bmatrix}^T \\ &= U_S D_S V_S \\ &= \sum_{i \in \mathcal{I}_S} \lambda_i \underline{u}_i \underline{v}'_i \\ &= \hat{S} \end{aligned} \quad (10)$$

We can therefore interpret the filtered signal matrix as the projection of the trajectory matrix Y onto the column space defined by the chosen left singular values. This is how the SVD is used to answer the question of what a typical consecutive L time series values would look like. Several of the obtained singular vectors are used to define some L -dimensional space in which L consecutive time series values are likely to be in. The filtering process then projects each column onto this space.

Since the individual singular vectors $\underline{u}_1, \dots, \underline{u}_L$ are arranged in terms of decreasing significance, with \underline{u}_1 giving the direction capturing the most variation and \underline{u}_L giving the direction explaining the least variation, it is intuitive to choose $I_S = \{1, \dots, d\}$ and $I_N = \{d+1, \dots, L\}$ (where d is some parameter to be determined). With the two sets of indices chosen as such we are able to associate the single rank matrices capturing the predominant amount of variation with signal, thereby constructing a signal matrix that adequately approximates Y . Those single rank matrices explaining an insignificant amount of variation can then be associated with noise. This choice of I_S and I_N is not a prerequisite, however, it is seldom chosen differently.

After the grouping step we have the multivariate decomposition of the time series trajectory matrix into a signal matrix and a noise matrix. To convert this multivariate decomposition into a univariate decomposition, we consider the final step in the filtering process, namely diagonal averaging.

Diagonal averaging

Now that we have the multivariate decomposition of Y into a signal matrix and a noise matrix, we need to find the corresponding signal and noise time series. We had a very simple procedure to transform our univariate data into a matrix representing multivariate data. This procedure can of course be reversed. However, it is necessary that the given matrix is a Hankel matrix. The signal matrix obtained by the first three steps of the SSA approach most likely is not a Hankel matrix. We can therefore not obtain a filtered time series before we have transformed the estimated signal matrix into a Hankel matrix. We do so by subjecting the matrices to the Hankelization operation, denoted by \mathcal{H} . This operation can be formally defined element-wise for $\hat{S} : N \times (T - L + 1)$ by Equation 11:

$$[\mathcal{H}(\hat{S})]_{i,j} = \begin{cases} \frac{1}{i+j-1} \sum_{l=1}^{i+j-1} [\hat{S}]_{l,i+j-l} & \text{for } 2 \leq i+j \leq L-1 \\ \frac{1}{L} \sum_{l=1}^L [\hat{S}]_{l,i+j-l} & \text{for } L \leq i+j \leq T-L+2 \\ \frac{1}{T-i-j+2} \sum_{l=i+j-T+L-1}^L [\hat{S}]_{l,i+j-l} & \text{for } T-L+3 \leq i+j \leq T \end{cases} \quad (11)$$

Effectively, this operator merely implements diagonal averaging; forcing our final approximated signal matrix into the form of a Hankel matrix by setting all the reverse diagonal elements equal to the average of the reverse diagonal elements.

Obtaining a Hankel matrix representing the signal in this fashion (through diagonal averaging) is optimal in the sense that it finds the Hankel matrix ($\mathcal{H}(\hat{S})$) that is closest to \hat{S} , according to the Frobenius norm. However, there are disadvantages of the Hankelization operator. The first disadvantage is the fact that after Hankelization, the $\mathcal{H}(\hat{S})$ matrix is not necessarily in an L dimensional subspace since the elements are affected by all elements in the reverse diagonal. The methodology therefore loses some of its interpretability. Secondly, for most of the observations in the centre of the observed time series, we find an average over L values; however, at the end and at the beginning of the time series the reverse diagonals become shorter and an average is taken over fewer values. In general, an average over fewer values causes a loss in certainty. This implies that the values at the end of the time series could possibly

be less accurate and since predictions are based on these values, this could be problematic. What should be said on these topics is firstly the fact that the effect of the diagonal averaging is often very small. To such an extent that the matrix could still be considered to be approximately in an L -dimensional space and also to such an extent that the end values can be considered relatively accurate. Also, since the SVD approximates the given trajectory matrix according to the Frobenius norm, it would be unwise to alter our matrix distance metric.

After this final step of the filtering method, we now have a matrix that represents a filtered time series, $\mathcal{H}(S)$, from which we can easily obtain our signal time series, $S_T = \{s_1, \dots, s_T\}$, upon which we base our predictions.

2.1.2 Prediction

Golyandina *et al* (2001) propose an algorithm for prediction using this filtered time series. However, it is based on the assumption that the given time series is approximately in the form of a linear recurrent formula. That is, the time series can be expressed by Equation 12:

$$y_t = \sum_{k=1}^q \alpha_k(t) e^{\mu_k t} \sin(2\pi t / \omega_k + \varphi_k) \quad (12)$$

In this equation, $\alpha_k(t)$ are polynomials in t and μ , ω and φ_k are arbitrary parameters. This allows for a very large variety of time series with different structural forms. Moreover, the methodology only requires that this property holds approximately.

The prediction algorithm then uses the singular vectors obtained in the SVD step as well as the signal obtained during the filtering process. It can be shown that the last component of any sequence of L consecutive values in this reconstructed time series can be expressed as a linear combination of the other values as shown in Equation 13 (interested reader can refer to Golyandina *et al*, 2001).

$$s_{L+i} = a_1 s_{L+i-1} + \dots + a_{L-1} s_{1+i} \quad \text{where } i = 0, \dots, T - L \quad (13)$$

Here the a_i values can be expressed in terms of the elements in the U_S matrix. Let \underline{u}_j^∇ denote a vector containing the first $L - 1$ components of the j th chosen singular vector. Further, let also $v^2 = \sum_j \pi_j$ where π_j represents the L th element in the j th singular vector. We then find that the \underline{a} vector can be written as follows:

$$\underline{a} = \frac{1}{1 - v^2} \sum \pi_j \underline{u}_j^\nabla \quad (14)$$

The one step ahead prediction can then be calculated by taking the inner product of \underline{a} with the vector containing the final $L - 1$ values of the time series. This step can be repeated iteratively k times on the predicted values to form k step ahead predictions.

2.2 The parameters

As with most methodologies, we now have the problem of choosing the appropriate parameters. De Klerk (2002) thoroughly explains possible methods of parameter selection. A couple of these methods will be examined here.

Model selection is crucial if we want the algorithm to provide accurate results. The parameters we have to assign values to are the window length L and the rank of the signal matrix d . There are some general guidelines as to how these parameters should be chosen which will be discussed here. After a brief discussion of these parameters, we will use some simulation examples in order to assess the effect that different choices of our parameters have on the methodology.

2.2.1 Choice of the window length

The choice of window length controls to a large extent the SSA algorithm, since it defines the dimension of the multivariate scenario that we are considering. By specifying this parameter, we are specifying how many consecutive time series values we want to investigate. The question arising is whether there comes a point where the $(L + 1)$ th lagged value do not provide significant contribution to the description of the signal in the time series.

The possible values for our window length parameter is restricted according to Equation 15. Of course we need to choose a window length of at least two in order to capture some trend in the L consecutive time series values. Furthermore, because of symmetry properties, it is not meaningful to choose a window length larger than $\frac{N+1}{2}$ (interested reader can refer to Golyandina *et al*, 2001).

$$2 \leq L \leq \text{integer part of } \left(\frac{N+1}{2} \right) \quad \text{where } L \in \mathcal{N} \quad (15)$$

Based on these possible values for our window length parameter, there are several concepts generally used to decide on a value for L . These concepts include basing parameter selection on w-correlation. A study of the SSA methodology would be incomplete without referring to w-correlation and the concept of separability. However, in this thesis we will more often use prediction error to decide on a value for L .

W-correlation

The concept of w-correlation goes hand in hand with studying separability in Singular Spectrum Analysis. Hassani (2007) mentions that separability in the SSA context characterizes how well the SSA algorithm is able to effectively separate the individual elementary components in some given time series. This can either refer to the separation between noise and signal components or, more specifically, the effective separation between individual signal components.

This procedure where we attempt to separate the individual components in a given time series takes place in the SVD step of the SSA algorithm. The trajectory matrix is decomposed into several single rank matrices. Each of these matrices are in different column spaces (each defined by their corresponding singular vectors) and these column spaces are orthogonal to (and therefore also uncorrelated with) each other. However, if we were to subject each of these elementary single rank matrix components to diagonal averaging and transformation to univariate data, the obtained univariate time series components, $Y_T^{(i)}$, are not necessarily uncorrelated any more. The w-correlation is a weighted measure of orthogonality between the individual single rank matrices after diagonal averaging. If the obtained time series are not significantly correlated with each other, according to the calculated w-correlation, we call the time series separable.

The w-correlation measure can be defined and calculated according to Equation 16, where $Y_T^{(i)} =$

$\{y_1^{(i)}, \dots, y_T^{(i)}\}$ and $Y_T^{(j)} = \{y_1^{(j)}, \dots, y_T^{(j)}\}$ represent the time series obtained from the single rank matrices Y^i and Y^j respectively after Hankelization and transformation.

$$\rho_{ij}^w = (Y_T^{(i)}, Y_T^{(j)})_w / \sqrt{(Y_T^{(j)}, Y_T^{(j)})_w (Y_T^{(i)}, Y_T^{(i)})_w} \quad (16)$$

where $(Y^{(i)}, Y^{(j)})_w = \sum_{k=1}^T w_k y_k^i y_k^j$ and $w_k = \min\{k, L, T - k\}$

Notice in this equation how the weights w_k simply represent the number of values over which an average is calculated during the diagonal averaging step. The measure therefore corrects for drastic changes that might have occurred near the end and the beginning of the given time series.

Since i and j take on values in the set $\{1, \dots, L\}$, we can obtain an $L \times L$ symmetric matrix containing the w-correlations between the different pairs of individual elementary components, $Y^{(i)}$ and $Y^{(j)}$. This symmetric matrix is often diagrammatically illustrated in the form of an $L \times L$ matrix with each cell coloured according to the magnitude of the w-correlation, with black cells corresponding to a w-correlation of magnitude equal to 1, white cells corresponding to a w-correlation of magnitude equal to 0 and shades of grey representing the possible values in between.

As mentioned before, separability can either refer to the ability of the SVD to effectively separate the individual signal components, or more importantly (but less strict) the ability to separate noise from signal. If the individual signal components are not separated well, prediction will not be affected as badly since the signal space could still be adequately approximated. However, if the signal is correlated with the noise, problems will arise in terms of prediction accuracy.

Even though separability is widely researched, this thesis will not study the concept in detail since separability is not our main concern. Our main concern is prediction accuracy.

Prediction error

When choosing the value of a parameter in a methodology, we have to keep our goal in mind. What do we consider to be favourable results? Often some loss function is calculated and parameters are chosen to optimize said loss function. One possibility would be to choose the window length that maximizes separability by minimizing the w-correlations. In this thesis on the other hand, since our main goal is one step ahead prediction accuracy, we rather wish to find the window length minimizing the error of one step ahead predictions.

Prediction accuracy of one step ahead predictions can be evaluated according to several measures. Common measures include Mean Squared Error (MSE), Mean Absolute Error (MAE) and Direction of Change (DOC). For a given time series Y_T of length T , the accuracy of k one step ahead predictions, denoted by $\{\hat{y}_{T+i} | i = 1, \dots, k\}$ can be evaluated according to the before mentioned measures of accuracy as defined by Equation 17, Equation 18 and Equation 19 respectively, once the true y_{T+i} values have been observed.

$$MSE = \frac{1}{N} \sum_{i=1}^k (y_{T+i} - \hat{y}_{T+i})^2 \quad (17)$$

$$MAE = \frac{1}{N} \sum_{i=1}^k |y_{T+i} - \hat{y}_{T+i}| \quad (18)$$

$$DOC = \frac{1}{N} \sum_{i=1}^k \delta(\text{sign}(y_{T+i} - y_{T+i-1}), \text{sign}(\hat{y}_{T+i} - \hat{y}_{T+i-1})) \quad (19)$$

In Equation 19, notice that $\delta(a, b)$ is the Kronecker delta function, returning the value one if $a = b$ and returning value zero if $a \neq b$. The DOC measure therefore simply finds the proportion of times the algorithm correctly predicted either an increase or a decrease in the time series. The MAE and MSE on the other hand measures of the magnitude of the error of the predicted values. These measures will be used throughout this entire thesis as an indication of the accuracy of prediction.

There is some connection between prediction accuracy and separability of a time series. If the SSA algorithm is unable to adequately separate the noise from the signal, it would either imply that some part of the signal component is omitted from or some part of the noise component is included in the estimated signal component or both. Predictions would then struggle to be accurate, since it is based on this inadequately filtered time series. However, as we will see in a simulation study at the end of this chapter, more often than not, the algorithm is able to adequately separate noise from signal even under heavily contaminated situations.

Literature

The choice of window length has received some attention in the literature. These papers based the choice of the value of L on different rationales, keeping different objectives in mind. Golyandina (2001) reminds us that, when choosing a value for L , it is important to take into consideration that if we expect the time series to include a periodic component, it is essential for us to choose L larger than or equal to the expected periodicity. This ensures that the periodicity, if it exists, is captured in the SVD. She then continues by recommending that the window length is chosen as a multiple of the periodicity of the time series. Other authors (Golyandina *et al*, 2001) suggest using $L = \frac{N+1}{2}$ or alternatively, as large as possible. Hassani (2011) showed that choosing the maximum window length is optimal in the sense that it maximizes separability by minimizing W-correlations.

It seems evident from most literature studies that larger choices of window length are safer choices. However, as we will see in a simulation study, there are disadvantages when choosing too large a window length.

2.2.2 Choice of rank

The choice of the rank of the inherent signal matrix is significantly more complicated than the choice of the window length, but at the same time also much more important. There are several things we need to take into consideration when choosing this parameter. During the grouping stage, we need to choose d single rank matrices to be associated with signal. As we mentioned before, since the single rank matrices are already ordered in terms of decreasing significance, we therefore merely choose the first d single rank matrices. The choice of single rank matrices therefore reduces to a mere choice of the rank, since the rank of the obtained signal matrix is equal to the number of single rank matrices that form the estimated signal matrix. This discussion elaborates on a couple of methods we can use to choose the appropriate rank. Popular methods include the following:

- Scree plots
- Study of singular vectors
- Forward cross-validation

Though these methods are widely used, they can often be ineffective. I believe that further research can be done on more effective ways in which to determine the correct value for the parameter d .

Scree plots

It is common practice in algorithms such as Principal Component Analysis to use the singular values obtained from the SVD, λ_i , to distinguish between significant variation and variation that should or can be associated with noise. Or in the framework of SSA, this implies choosing d large enough to ensure that the trajectory matrix is significantly represented, but also small enough to ensure that the error is completely removed. These singular values are an indication of the accuracy with which the Y matrix is estimated by the lower rank S matrix. It can be shown that:

$$\text{trace}[(Y - \sum_{i=1}^d Y^i)(Y - \sum_{i=1}^d Y^i)'] = \sum_{i=d+1}^L \lambda_i^2 \quad (20)$$

Therefore, λ_i is proportional to the amount of variation of Y included in the Y^i matrices, since inclusion of Y^i into the signal matrix would result in the Frobenius norm decreasing by λ_i^2 . These values can therefore be used to decide when the signal matrix is sufficiently represented. If the λ_i values become insignificantly small, the contribution of the corresponding Y^i to the variation in Y becomes negligible as well and should therefore rather be associated with noise.

A scree plot is a way in which we can consider these singular values, by merely graphically illustrating the ordered λ_i values. In other words, a scree plot is a plot of the paired values $(\lambda_i, i) | i = 1, \dots, L$. Theoretically, when this decreasing curve forms an elbow followed by a less steep decline, it would indicate that the singular values up to this point had significant contribution in the variation of the given matrix. Values after (and including) this point on the other hand should be associated with noise since the singular values (and therefore the contribution to the estimation) have become insignificant and should not be included in estimation of the signal matrix. Apart from the fact that using scree plots as guideline is extremely subjective, it has some other inadequacies in the context of SSA. Since there is a lot of repeated values in the trajectory matrix, the singular values are often such that there are large decreases in subsequent λ_i . The first singular value is often much greater than the singular values thereafter. Therefore it is often difficult to interpret where the curve forms an elbow, thereby exacerbating the problem of subjectivity when interpreting a scree plot. This problem is often addressed by considering monotone transformations of the singular values in order to better identify this turning point in the transformed singular values.

Notice that choosing the rank parameter based on a scree plot is effectively a process where we allow the parameter d to vary and then basing the model selection decision on the in-sample model fit for each d . This is evident since the inclusion of each individual single rank matrix monotonically increases the accuracy of the estimation of Y . Therefore, by including single rank matrices that insignificantly contribute to the estimation process, we are subjecting the algorithm to over-fitting, even though the estimation of Y becomes more accurate. Basing our parameter selection procedure on an in-sample model fit such as this is not sufficient.

Plot of singular vectors

Another method that can be used to assess the necessity of adding a singular vector in the filtering and prediction process, is a study of the singular vectors. More often than not, this study is of a graphical nature, based on plotting the elements of the singular vectors against their index values and assessing the nature of the singular vector. Plots such as these can be of help in identifying the nature of each individual component obtained in the SVD.

Upon inspection of the assumption that a given time series can be approximated by a linear recurrent formula, we would be able to derive that our given time series consists of a combination of polynomials, exponential functions and wave functions. Each of these components would have different effects on our individual signal components.

Consider a time series of a quadratic polynomial that is uncontaminated with noise. If we were to create some trajectory matrix of this time series, we would be able to show that this trajectory matrix would lie in a three dimensional space. One dimension would be attributed to a constant, another would be necessary to describe the linear component of the polynomial and the final dimension would be associated with the quadratic component of the time series. Corresponding singular vectors needed to describe nature such as this could therefore be graphically represented as follows:

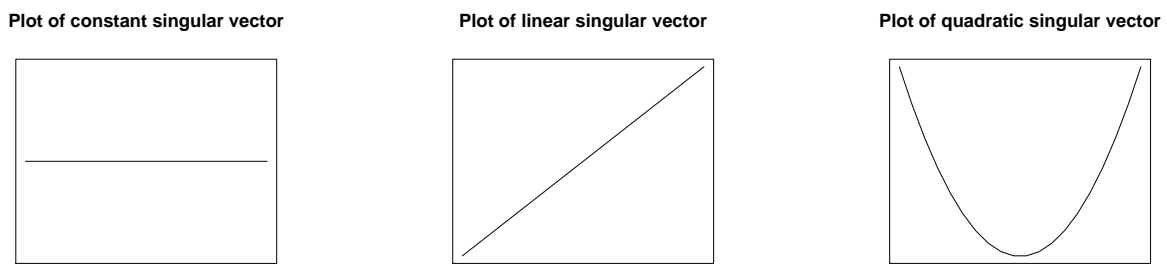


FIGURE 1: These three figures represent the nature of the 3 singular vectors associated with the signal present in a polynomial time series of degree 2. The left figure represents the singular vector associated with the constant in said polynomial, the middle figure is associated with the linear trend and the right figure is associated with the quadratic component of the quadratic polynomial.

This concept can be extended to describe any time series in the form of a polynomial of degree p . Such a time series, if uncontaminated, would necessarily lie in a $(p + 1)$ dimensional space, with each singular vector describing either a constant or one of the terms of the polynomial.

For a time series in the form of an exponential function, without noise component, we can follow a similar rationale. By transforming such a time series into a trajectory matrix, we will see that the matrix will be of dimension 2. One component will describe some constant deviation and one component will describe the exponential nature of the time series. Corresponding singular vectors would therefore resemble those in Figure 2.

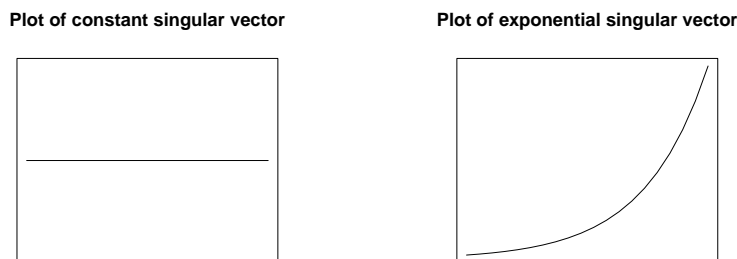


FIGURE 2: These two figures represent the nature of the 2 singular vectors associated with the signal in a time series of exponential nature. The left figure represents the singular vector associated with the constant in the time series and the right figure is associated with the exponential component of the time series.

However, if we had a time series with periodic wave component, also uncontaminated with noise, we have to be slightly more careful. Say for instance we had a sine function with periodicity of 10 ($s_t = \sin\left(\frac{2\pi t}{10}\right)$). If we transform this time series into a trajectory matrix with $L = 5$, the SVD analysis would not be able to consider a full cycle and 5 consecutive time series values would better resemble the behaviour of a quadratic (or maybe cubic) polynomial rather than a sine function. However, if a full 10 (or more) consecutive values are considered, a full cycle would be observed. We will then see that the trajectory matrix would lie in a 2 dimensional space, since any sine function with phase deviation can be expressed in the form of a linear combination of a sine and cosine function. The corresponding singular vectors could therefore graphically be represented as in Figure 3.

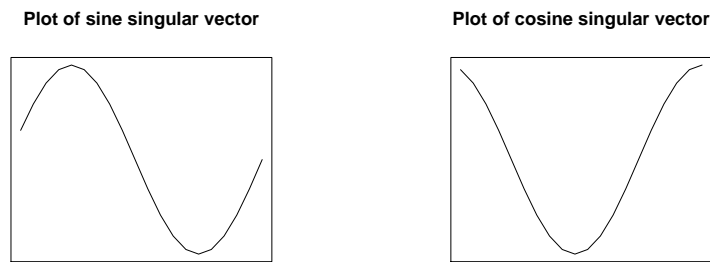


FIGURE 3: These two figures represent the nature of the 2 singular vectors associated with the signal of a time series with harmonic component. The left figure represents the singular vector associated with the sine component and the right figure is associated with the cosine component. Any 12 consecutive values in a harmonic component can then be described by a linear combination of these two singular vectors.

Furthermore, we can consider cyclic behaviour not only by plotting the elements of the singular vector against index values, but also when plotting the elements of one singular vector against the elements of another singular vector. These points are then linked with a line connecting the pairs as the index increases. These plots are called scatterplots of the paired singular vectors. These figures take interesting forms when cyclic behaviour is present. For instance, if we were to plot the sine and cosine singular vectors of wave functions with periodicity 12,6 and 4 we would obtain the following scatterplots:

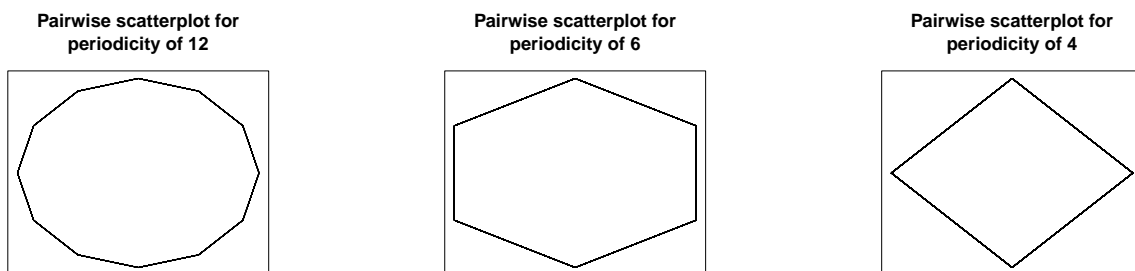


FIGURE 4: These three figures represent the nature of the scatterplots of paired singular vectors of a time series with harmonic component. The left figure is a paired scatterplot of two singular vectors associated with cyclic component of period 12, the middle figure is a paired scatterplot of vectors describing wave-like behaviour with period of 6 and the final figure is similar to the other two figures, but for cyclic behaviour with periodicity of 4.

The above scatterplots clearly show that for a time series with periodicity ω , a paired scatterplot would resemble a regular polygon (either convex or complex) with ω sides. It is of course entirely impractical to plot all possible pairs of singular vectors since we often work with dimension of 10 or more. However, with cyclic behaviour we have that two singular vectors attempt to capture a single type of behaviour. Therefore, the amount of variation described by both components will be similar, resulting in the corresponding singular values being similar. We therefore merely need to plot the scatterplots of paired singular vectors with similar singular values in order to assess this periodic nature.

Notice that in the first two situations (polynomials and exponential functions), we had a singular vector associated with some constant deviation. This was not the case for the third scenario. This is merely because the class of polynomials and the class of exponential functions are unlikely to have mean of zero. If a situation arises where the time series has a mean of zero (as is the case with a time series with only a periodic wave component), the constant singular vector will simply be omitted. It is here where the rationale for the centred version of SSA originated. This addition to SSA simply centres the trajectory matrix before SVD and then reverses the centring procedure after the algorithm is completed. However, since a constant singular vector would do something very similar, I believe that this is unnecessary and therefore focus my studies on the non-centred version of SSA.

In conclusion, I would like to mention that it is unlikely to have a time series of this form, uncontaminated by noise, in any practical situation. Even if we did, the singular vectors would merely resemble the figures presented here. We are therefore unlikely to see singular vectors exactly like the ones shown above. However, it is still an interesting study to consider these individual components and their characteristics. Such a study often gives valuable information on which singular vectors to include and the nature of the individual components.

Forward cross-validation

Cross-validation is often used for model selection or parameter estimation. Forward cross-validation techniques are often associated with models for time series analysis, since we are less interested in the accurate estimation when interpolating and more interested in accuracy when extrapolating. With forward cross-validation, our time series is partitioned into three parts. The first part (y_1, \dots, y_{train}) forms the data set used to train the model. The second part ($y_{train+1}, \dots, y_{valid}$) is used in the validation step to test the efficiency of different parameters. This cross-validation step is then used to find the most apt parameter. The final data set ($y_{valid+1}, \dots, y_T$) is the data set we wish to predict with the optimum amount of accuracy. The fitted model is therefore tested on this data set.

Graphical illustration of data partitioning for forward cross-validation

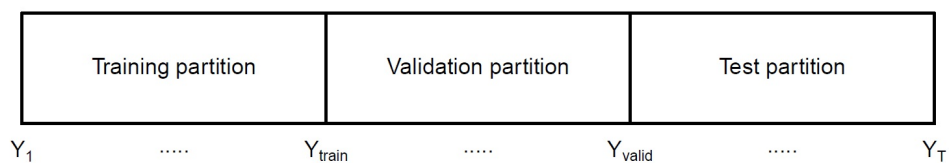


FIGURE 5: This figure illustrates the way in which a time series will be partitioned in order for us to subject said time series to forward cross-validation methods.

In this thesis, I will be using forward cross-validation for estimation of the parameter d . A portion of the time series will be set aside for model training. In the cross-validation step, a value for d will be suggested and rolling one step ahead predictions will iteratively be calculated for the validation set. This will

be repeated for each possible value $d = \{1, \dots, L\}$ and the optimal d will be chosen based on minimization of validation error. After the dimension has been chosen, the test data set will be predicted and analysed.

In order to gain some intuition on how the dimension affects predictions, we consider now a simulation study. This simulation study will also answer questions left unanswered in previous discussions.

2.3 Simulation study

In this simulation study we will discuss some simulation examples in order to gain an intuitive feeling for how the different parameters influence the SSA algorithm. In each example, we will give some formula according to which the given time series is constructed. The time series will be of length $T = 130$ and this time series will be split into a training set (of length 100) and a validation set (from the remaining 30). The effects of different choices of parameters will then be considered.

Example 1

For the first example, we consider an uncontaminated time series, as given by Equation 21.

$$y_t = t + 10\sin(2\pi t/10), \quad \text{where } t = 1, \dots, 130 \quad (21)$$

We then continue the study by considering the effect different window lengths have on the SSA methodology. Of course it is clear to us that in this scenario the signal matrix of this time series is of dimension 4: one singular vector for a constant and one for a linear trend, as we normally have with a polynomial of degree 2. Finally, two singular vectors are attributed to the cyclic component. However, in practice, we will most likely not know the true value of d . We therefore consider the effect of the choice of window length for 6 different choices of rank; $d = 1, \dots, 6$. For each of these dimensions, we vary the window length from minimum ($L = d$) to maximum ($L = 50$) and consider the mean squared error of the rolling one step ahead predictions of the test partition.

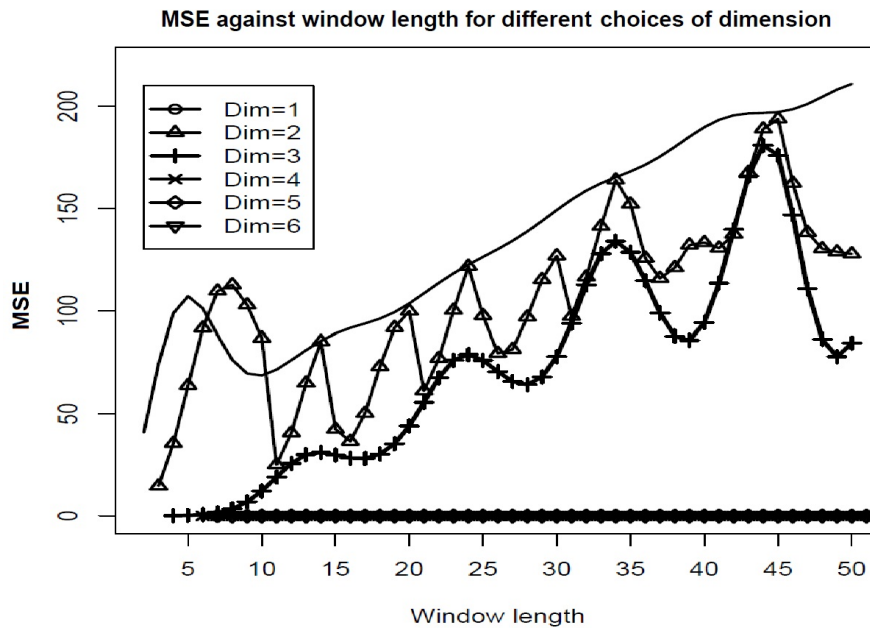


FIGURE 6: This figure shows how the Mean Squared Error for one step ahead predictions changes as the window length is varied in Example 1. The plot contains 6 lines; each one representing a different choice of signal space dimension, d , ranging from 1 to 6.

Upon inspection of this figure, we see that for $d \in \{1, 2, 3\}$, an increase in window length is destructive to prediction accuracy. Consider for example for $d = 1$; this first singular vector in this scenario will be the singular vector resembling a constant (this will be the case for any choice of window length since the time series clearly does not have zero mean). If we were to predict the one step ahead value according to this singular vector, we would project to the space spanned by a constant and it would imply that the prediction is roughly an average of the previous $L - 1$ values. For L between 2 and 10, the prediction accuracy shows interesting results. Up until $L = 6$, the prediction error increases drastically. The reason for this can be explained intuitively. If we have currently just passed a dip in the sine function, we are predicting tomorrow's peak by the average of the values in the dip observed in the previous 6 days; alternatively, if we have just passed a peak, we are predicting an upcoming dip by the previously observed peak. On the other hand, if we consider an entire cycle ($L = 10$) and average over such a cycle, the dips and peaks are irrelevant. In general however, by increasing L , we are looking further into the past which will therefore decrease the prediction accuracy as a result of the increasing trend.

The behaviour of the MSE (for $d = 2/3$) for changes in choice of window length can be explained in a similar fashion as above (for $d = 1$). We do so by considering the nature of the individual singular vectors in the form of scatterplots. Figure 7 shows these scatterplots for a window length of $L = 20$.

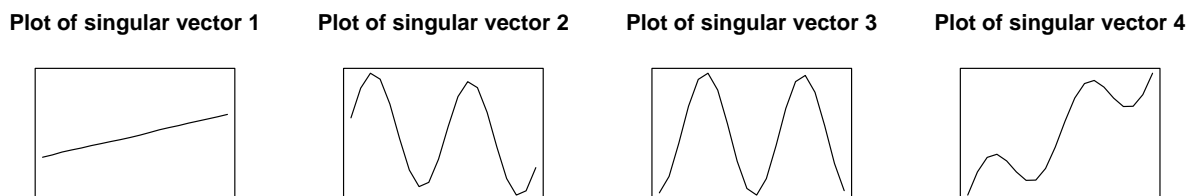


FIGURE 7: These 4 figures each represent a scatterplot of one of the 4 significant singular vectors in the given deterministic time series of Example 1, for a window length of 20.

As mentioned before, the first singular vector is merely a constant (with slight indication of linearity) over the index values. The second significant vector describes some of the cyclic behaviour as well as a slight trend, while singular vector three simply describes the cosine component of the periodic behaviour and the final singular vector describes the remaining part of the periodic behaviour together with the trend. Therefore, by only including the second singular vector, the singular vectors will not be able to capture the periodic component adequately. Notice how the prediction accuracy seems to contain some periodic component as L increases. The same holds when we choose rank of 3. In both cases, this happens since some of the periodic nature is left unexplained in the filtering process and therefore also in the prediction process.

However, once the fourth singular vector is included, the signal is adequately captured and prediction error drops to zero for all L . Even if the window length is smaller than the periodicity of 10, approximation of the periodic component (as polynomial) is close enough to the true signal. As a result of this there is little or no error in prediction. This is also true for dimension of 5 and 6. By choosing these dimensions, we are overestimating the dimension and effectively also incorporating noise into our signal; however, since there is no noise, the prediction is unaffected. These figures will of course be heavily influenced if noise were added to the time series.

Example 2

In this second example, we consider time series with the same deterministic component as we had in Example 1; however noise is added as given by Equation 22.

$$y_T = t + 10\sin(2\pi t/10) + \epsilon_t \quad \text{where } t = 1, \dots, 130 \quad (22)$$

where $\epsilon_t \sim \text{Uniform}(-1, 1)$

Notice in Equation 22 that the added noise time series is simulated according to the Uniform distribution. This is perhaps an unintuitive choice for the distribution of the noise component; the Gaussian distribution is a more common choice. The main reason for this Uniformly distributed noise component will be clear at a later point in this thesis. A secondary reason that we can mention at this point is the fact that it is easier to describe the amplitude of a Uniformly distributed random variable and therefore easier to refer to the signal to noise ratio. For the sceptical reader I would like to mention that the Gaussian distribution delivered similar results in this (and every other) simulation study.

Once again we are interested in the prediction power of the SSA algorithm and how it is affected by the choice of our parameters. In order to measure the prediction power of the algorithm, we consider the ability of the algorithm to correctly predict the signal or deterministic part of the time series. The mean squared difference between the prediction and signal component is measured for each choice of the dimension, $d = 1, \dots, 6$, while varying the window length as before. By measuring deviation from signal rather than deviation from observed time series values, we are disregarding the variation that the noise component would add to the MSE, since this variation is merely noise and impossible to predict. In order to measure the MSE accurately, we therefore repeat this process 500 times and obtain the average of the MSE for each situation.

For this simulation example, we find that the figures for choices $d \in \{1, 2, 3\}$ are nearly exactly the same as those obtained with Example 1. However, for $d \in \{4, 5, 6\}$, we find the Figure 8 describing the calculated MSE over different window lengths.

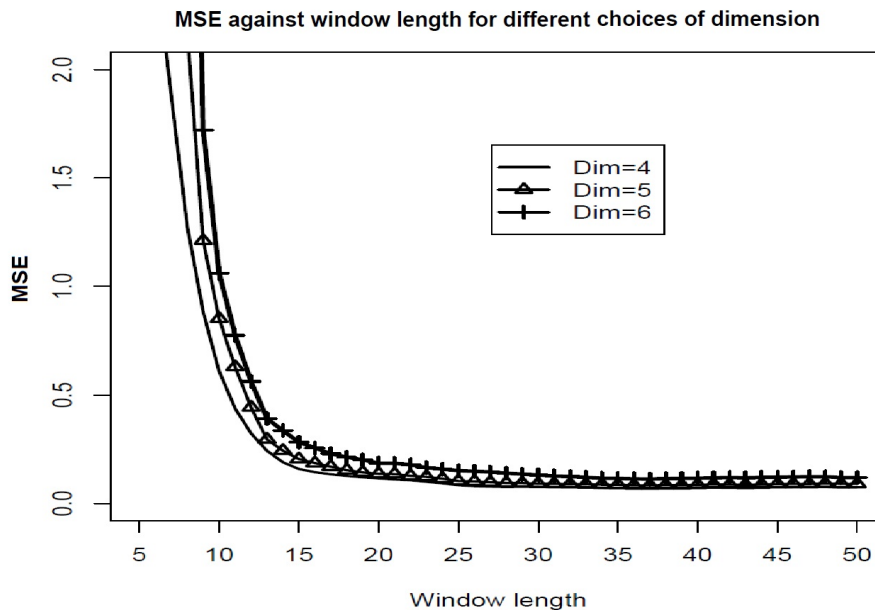


FIGURE 8: This figure shows how the Mean Squared Error for one step ahead predictions changes as the window length is varied in Example 2. The plot contains 3 lines; each one representing a different choice of signal space dimension, d , ranging from 4 to 6.

Here we can clearly start seeing the benefit of choosing L large. When we had a situation where signal was unaffected by noise, as in Example 1, the SSA methodology had negligible error terms for $d \in \{4, 5, 6\}$, while varying across all window lengths. However, now that a noise component is additively added to the signal component, we see that window lengths smaller than the periodicity is completely ineffective with prediction. With the addition of the noise the methodology is completely unable to capture the cyclic behaviour and therefore unable to estimate and predict the true underlying signal.

We should also notice how the curves for dimension equal to 5 and 6 are slightly above that of $d = 4$. As we mentioned before, we know that the inherent signal matrix is of rank 4. Therefore by including a fifth and sixth dimension, we are including some of the noise into the estimates of the signal. Consequently prediction based on this inaccurately filtered time series is therefore influenced by slight unpredictable deviations and prediction accuracy decreases.

In this example, by choosing a too large value for d , we do not decrease prediction accuracy as drastically. However, choosing window length smaller than the periodicity does. We therefore come to conclusions similar to those mentioned in the literature; that is, the window length in the SSA algorithm should be larger than or equal to the periodicity of the cyclic behaviour in the time series. Furthermore, for window lengths larger than the periodicity of 10, we do still see some slight decrease in the MSE. However, after $L = 20$, this decrease seems to cease. Literature (Golyandina *et al*, 2001) mentioned that choosing L as large as possible has the advantage that it minimizes w-correlations. In this scenario, we also see how larger values of L slightly increases prediction accuracy. To explain this, we consider two w-correlation plots at different window lengths, for a single time series, randomly generated according to Equation 22.

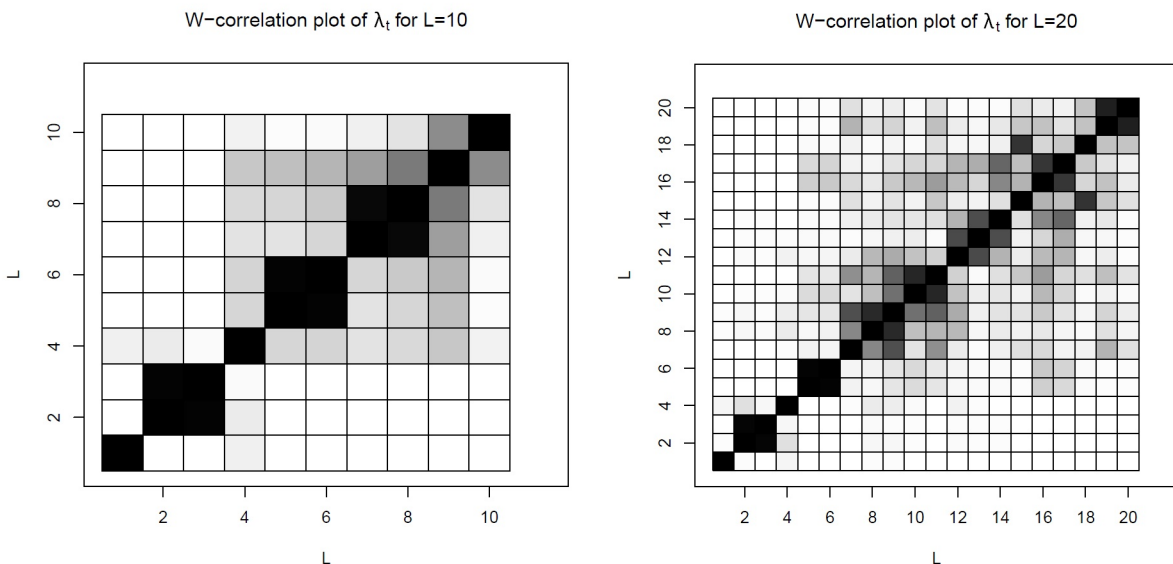


FIGURE 9: These two figures represent w-correlation plots associated with a single time series simulated according to Equation 22. The left figure graphically illustrates the w-correlation matrix for a window length $L=10$ and the right describes the w-correlation matrix for window length $L=20$.

Even though prediction is our main priority, w-correlation plots here give some insight as to why prediction at window length of 10 is still slightly worse than we have with window length of 20. In both the left and right figure, we see that the second and third components are highly correlated with each

other. This is of course since both these components are related to the periodic behaviour of the time series. However, as mentioned before, prediction accuracy is not as much dependent on the separability of individual signal components, it is more important for the algorithm to successfully separate signal from noise. The figures also show that with window length of 20, the first four components seem to be uncorrelated with the noise components. However, in the left figure, with window length of 10, we see that the fourth signal component seems to have slight correlation with the components associated with noise. Therefore, even by including the correct number of single rank matrices for prediction, we are still in fact including a very small amount of noise into estimation of the signal. Consequently, prediction accuracy decreases.

In this scenario, we had a time series contaminated with noise. However, the signal to noise ratio was still very high. The methodology was still therefore able to distinguish between noise and signal, resulting in accurate predictions. In the next example we consider a scenario where the signal is much less prominent.

Example 3

We were able to see in the previous two examples how the SSA methodology successfully separated signal from noise and predicting accordingly. However, this was in the presence of little or no noise. This example will consider the effect that noise with large variation has on prediction accuracy and how that changes for different parameter values. We therefore consider time series generated according to Equation 23.

$$y_T = t + 10\sin(2\pi t/10) + \epsilon_t \quad \text{where } t = 1, \dots, 130 \quad (23)$$

where $\epsilon_t \sim \text{Uniform}(-10, 10)$

This example is similar to Example 2. The only difference is the fact that the noise component now has an amplitude of 20 instead of 2. This is the same as that of the periodic component of the signal. Furthermore, it is more than the variation as a cause of the linear trend in the signal for window length up to $L = 20$. To consider how this affects the prediction accuracy of the methodology, let us once again consider the effect that a varying window length has on the MSE for rank of the signal matrix ranging from 4 to 6. We only consider these possible values for d , since the MSE curves for d ranging from 1 to 3 once again were similar to that of Examples 1 and 2, only with more erratic behaviour at very small window lengths. To ensure accurate estimation of the MSE for rolling one step ahead predictions, we repeat the process 500 times and consider the average observed MSE for varying window lengths.

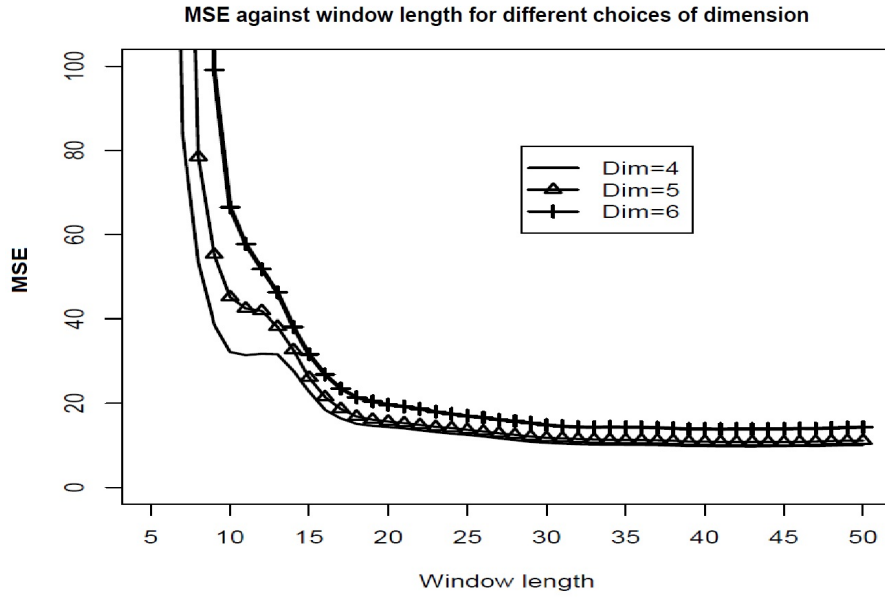


FIGURE 10: This figure shows how the Mean Squared Error for one step ahead predictions changes as the window length is varied in Example 3. The plot contains 3 lines, each one representing a different choice of signal space dimension, d , ranging from 4 to 6.

This figure resembles Figure 8 of Example 2. There are however two very significant differences. The first is the fact that there seem to be a more harsh penalization when choosing the value of d too large. With Example 2, the MSE for the three possible values of d differed slightly. In this simulation example however, if d is chosen as 5 or 6, there is a much more significant increase in MSE for all window lengths. Another noticeable difference is the behaviour between window length of 10 and 20, especially for $d = 4$. The MSE decreases sharply until it reaches $L = 10$, stabilizes up to $L = 15$ and then it starts decreasing again until it reaches some plateau around $L = 20$, where further decreases in MSE seem to become insignificant. In a similar fashion as in Example 2, we consider the w-correlation plots when $L = 10$ and $L = 20$, for a single time series generated according to Equation 23 in order to explain why the MSE behaves strangely between window lengths of 10 and 15.

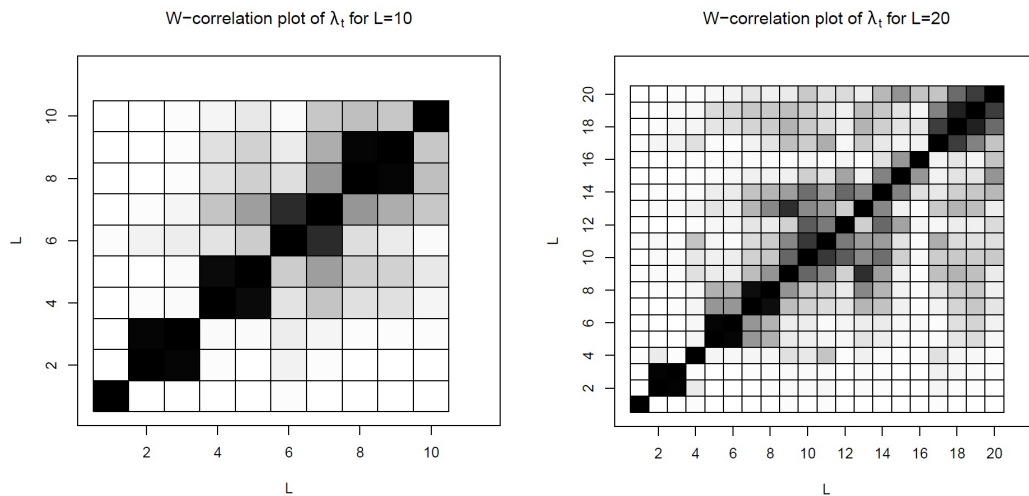


FIGURE 11: These two figures represent w-correlation plots associated with a single time series simulated according to Equation 23. The left figure graphically illustrates the w-correlation matrix for a window length $L=10$ and the right describes the w-correlation matrix for window length $L=20$.

In the above w-correlation plots, we see that the fourth component is the one proving to be problematic. With window length of 10, this component is significantly correlated with the fifth component. However, for window length of 20, this correlation effectively disappears and only slight correlations between signal and noise components remain. Because of the bad separability between signal and noise when $L = 10$, we find that the filtering process is unable to filter the contaminated time series from noise. Consequently, prediction is still effectively based on noise contaminated time series. This problem also seems noticeable if we consider the scree plots for window lengths of 10 and 20 in Figure 12.

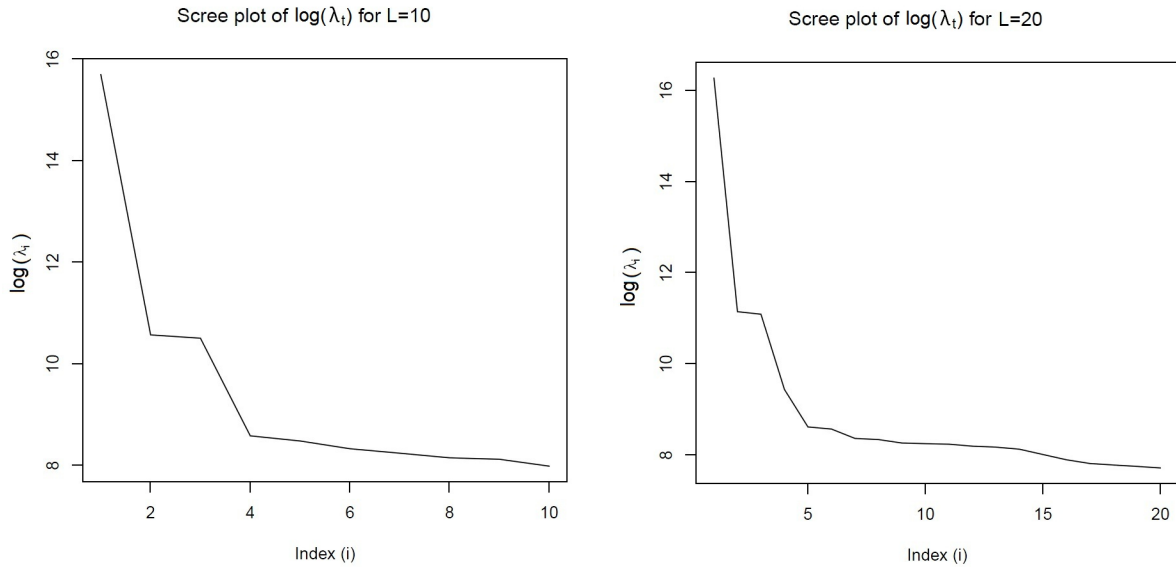


FIGURE 12: These two figures illustrate the scree plots associated with a single time series simulated according to Equation 23; however, because of the nature of the singular values, we plot the log of the decreasing singular values. The left figure graphically illustrates the transformed scree plot for a window length $L=10$ and the right illustrates the transformed scree plot for window length $L=20$.

In this figure, the log of the λ_i values are drawn. Because of the nature of the magnitudes of the singular values, as discussed before, plots of the logged singular values are more interpretable than normal scree plots. Notice the significant drop in singular value between index 4 and 5 for window length of 20. This is much less clear for window length of 10. This would imply that for a window length of 10, the fourth singular vector is describing a less significant part of the variation in the trajectory matrix, in comparison with the fifth component, than in the second scenario with window length of 20. Both the w-correlation plots and the scree plots therefore indicate that there is some relationship between the fourth and the fifth singular vectors at window length of 10 that is not present at a window length of 20. To try and explain this, we consider the plots of the singular vectors for both situations.

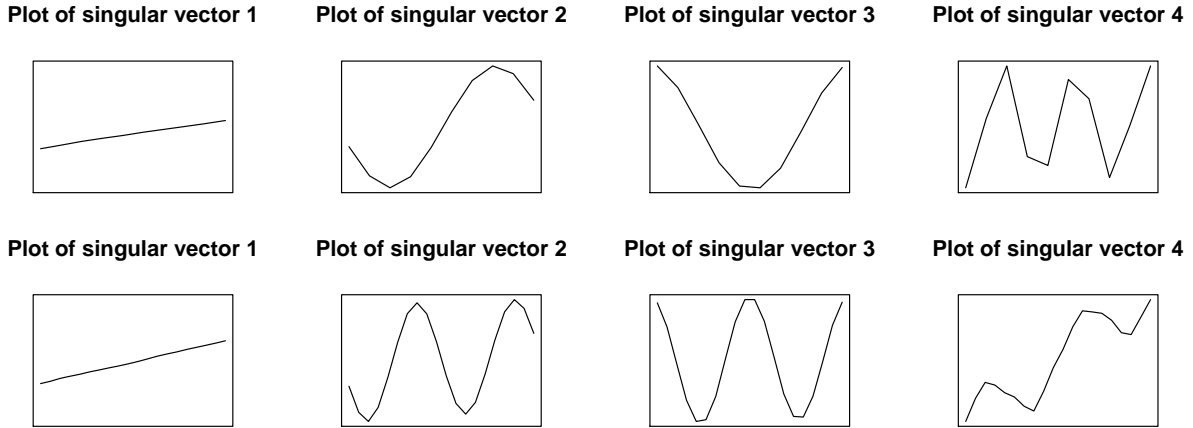


FIGURE 13: These two sets of figures illustrate the differences between the 4 signal singular vectors when different window lengths are chosen. The top four figures are graphical representations of the singular vectors obtained from a single time series when $L=10$ and the bottom four figures are graphical representations of the singular vectors obtained from a single time series when $L=20$.

This clearly explains the inadequacy of the algorithm at a window length of 10. Singular vectors 1,2 and 3 are similar for both situations, however, as we mentioned before, the problem lies with the fourth singular vector. For smaller window lengths, the algorithm is overwhelmed by the amount of noise and is therefore unable to capture the remainder of both the linear and the periodic component. Larger window lengths on the other hand adequately captures this final component. The reason for this is the fact that for a window length of 10, the amplitude of the linear component over this bandwidth is a mere 10, while the amplitude of the noise is double as much. With $L = 20$, the amplitude associated with the linear component is also 20 and since this linear component is more consistent than random deviations, the algorithm adequately captures this final component. The same rationale holds for the inadequate acquisition of the harmonic component in the final singular vector.

From these three simulation examples we can clearly see some benefits of choosing large window lengths and why it is recommended. However, under some circumstances, larger window lengths can be destructive to prediction.

Example 4

In our final example we consider a time series undergoing some structural change and we suspect that such a time series might be disadvantaged by choosing L too large. Consider the time series with structural break at $t = 100$ described by Equation 24.

$$y_t = \begin{cases} t + 10\sin(2\pi t/10) + \epsilon_t & t \leq 100 \\ 80 - t + 10\sin(2\pi t/10) + \epsilon_t & t > 100 \end{cases} \quad \text{for } t = 1, \dots, 110 \quad (24)$$

where $\epsilon_t \sim \text{Uniform}(-10, 10)$

Notice once again that this time series is similar to that of Example 3 up until the point of structural break at $t = 100$, where the linear trend merely changes direction. Suppose we were asked to predict a time series, simulated according to Equation 24, from $t = 100$ to $t = 110$ (only ten validation values are considered here so as to put emphasis on the change in structural break). Based on all observations up to the point of structural break ($t = 100$), all evidence show that the appropriate choice for d is 4. If the

estimation of the d parameter is therefore accurate, prediction will follow based on a signal space spanned by 4 singular vectors. To investigate whether some choices of window length provide more accurate predictions in the presence of a structural break such as this, we once again consider the MSE for 10 rolling one step ahead predictions. Even though the more accurate choice for d (based on historical data) would be 4, we do consider choices $d = 4, 5, 6$ and to accurately estimate the true MSE from signal component, we repeat this simulation 500 times and calculate the average value of the MSE. These average MSE values are shown in Figure 14.

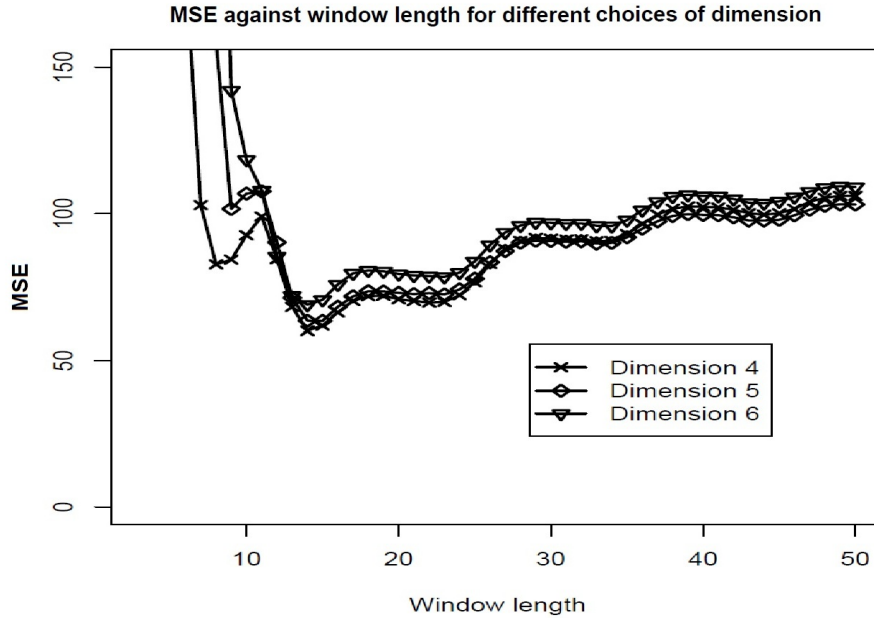


FIGURE 14: This figure shows how the Mean Squared Error for one step ahead predictions changes as the window length is varied in Example 4. The plot contains 3 lines, each one representing a different choice of signal space dimension, d , ranging from 4 to 6.

From Figure 14, we see that regardless of the choice of d , the MSE is minimized by choosing L slightly larger than 10. In Example 3, the large amount of noise made it necessary for us to choose a larger window length so as to include more observations over which we try and capture some signal component. We even saw a very significant decrease in MSE between window length values $L = 10$ and $L = 20$ in Figure 10. Here however, even under the presence of noise with large amplitude, a larger window length penalizes us in terms of prediction accuracy, while for L chosen small, the prediction is also inaccurate. The reason for this is that we need to at least consider 10 consecutive time series values in order to effectively capture the periodic component. However, the algorithm needs to be able to disregard observations prior to the structural break as soon as possible and rather base predictions on newer, more recent observations. This can only be done if a smaller window length is chosen.

It is interesting to see in Figure 14 that, even though this time series is highly contaminated with noise, the choice of d does not seem to influence prediction accuracy greatly for window lengths larger than 10. This is especially the case between dimension of 4 and 5. The reason for this is the uncertainty regarding the inherent structure of the time series during this structural break. Consequently we do not know the true value for the dimension. The question therefore arises: How would the methodology perform if we were to consider smaller choices of d and how would prediction accuracy compare with that of the model with $d = 4$? To answer this question we consider Figure 15 representing the behaviour of MSE over varying window lengths for choices of d ranging from 1 to 6.

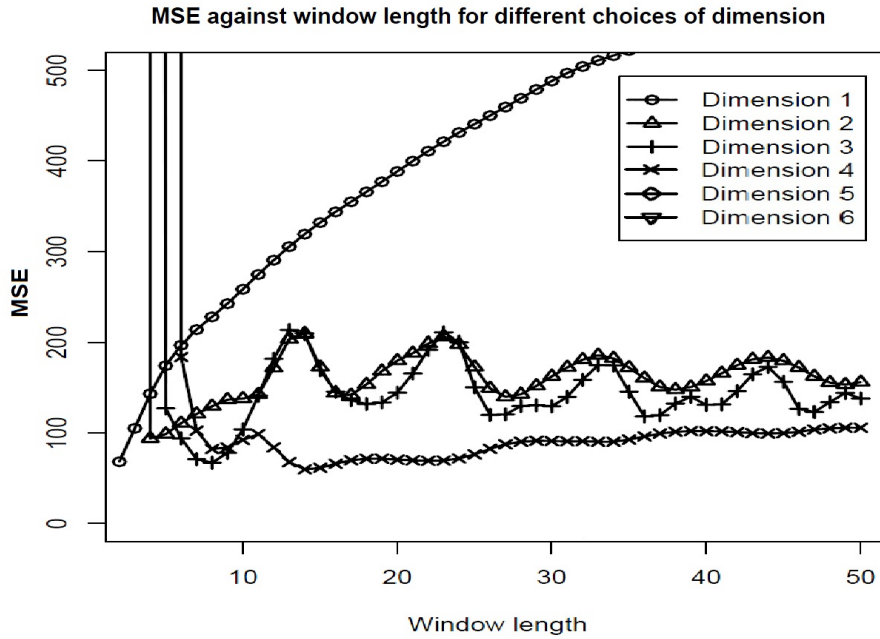


FIGURE 15: This figure shows how the Mean Squared Error for one step ahead predictions changes as the window length is varied in Example 4. The plot contains 6 lines, each one representing a different choice of signal space dimension, d , ranging from 1 to 6.

The 3 curves representing the MSE, associated with predictions where lower dimensions for the signal space are considered, are significantly different from that of Examples 1, 2 and 3. However, there are similarities. For instance, the curve for $d = 1$ still increases as the window length increases, even though the initial behaviour is different. This can be explained once again by referring to the fact that the first singular vector once again roughly represents a constant deviation and therefore roughly predicts according to an average of the previous $L - 1$ observations together with some slight increasing trend. Also, we see that the curves for $d = 2$ and $d = 3$ still have a periodic nature which can be explained by once again referring to the fact that the second and third singular vectors are associated with some periodic component, which is only fully captured once we include the fourth singular vector. What is most important to notice here is the fact that choosing $d = 4$ is still optimal in terms of prediction accuracy in this scenario, even though the true dimension of the signal is unknown as a result of the change in structure.

Concluding this simulation study, we see that the window length should be chosen as large as possible. However, if it is possible that the time series could be undergoing some structural change, it is necessary that the window length be chosen so as to incorporate any cyclic behaviour, but not too much larger than that. Some evidence (both in the simulation studies and in the literature) also show that the window length should be chosen as a multiple of the periodicity.

Choosing the rank of the signal matrix is our next obstacle. Even though I believe the choice of this parameter can be researched in much more detail, in this thesis I will use cross-validation methods, such as the ones discussed above, to decide on the appropriate value of d .

3 Baseline predictions

Now that we have some general intuition regarding the SSA methodology and how it is affected by parameter choice, we consider practical data on stock market activity in order to compare traditionally used AR models with SSA models. In this chapter, we consider the daily closing prices of three general stock market indices in America, namely the Dow Jones Industrial Average (DJIA), the NASDAQ and the S&P500. These daily closing prices are obtained from <http://finance.yahoo.com>. Figure 16 below shows these three stock market indices over time and how the time series is divided into three segments to incorporate the forward cross-validation procedure estimating the relevant parameters.

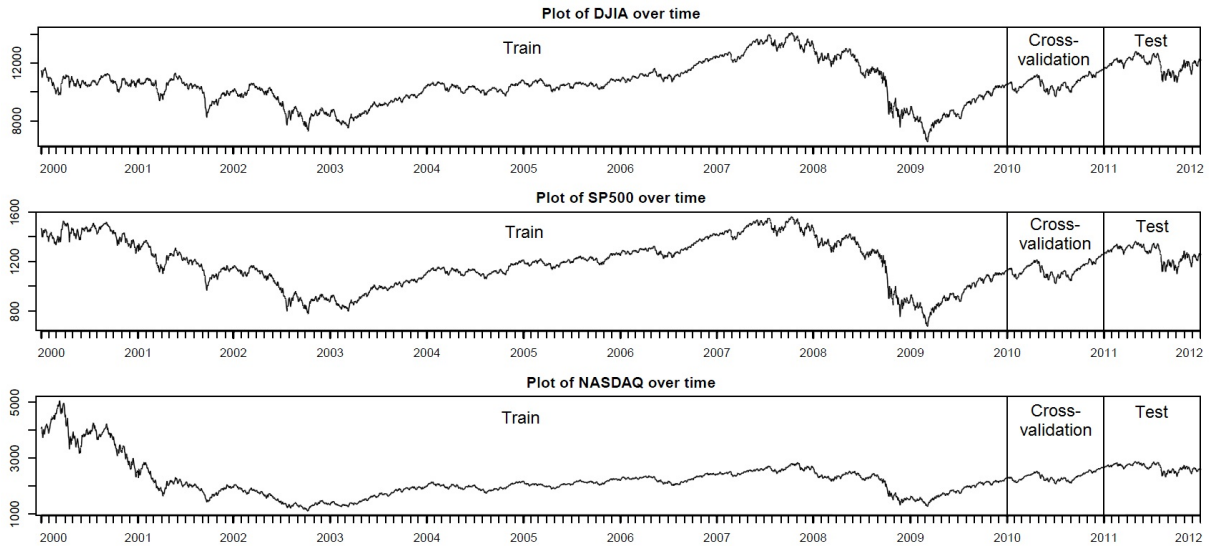


FIGURE 16: These three figures illustrate the three stock market time series to be predicted. The top figure contains the observed daily DJIA closing prices, the middle contains the observed daily S&P500 closing prices and the final contains that of the NASDAQ index. In addition to this, these three figures then also graphically illustrate the partition of each time series into training data, cross-validation data and test data.

The three indices shown here are considered to be measures of the value of the stock market in general, i.e. they measure the general economic activity in America. Since this is the case, the three time series should be relatively related. Upon inspection, we see that this is indeed the case; the general trend for the three time series seem quite similar. This is especially true for the DJIA and the S&P500. The NASDAQ also resembles the first two indices, however it seems to have less variation in some scenarios where the DJIA and the S&P500 undergoes frequent and unexpected changes.

In this study on practical data, we will be considering each of these time series individually. Each time series will undergo the same process. First, observations between 2000 and 2010 will be used in order to train both our SSA and AR models. Thereafter rolling one step ahead predictions will be calculated for observations during 2010 while varying the parameters involved in the model in order to find values for these parameters in a forward cross-validation step. Finally these parameters will be used in order to predict the final year's observations and to compare the optimum models for both methodologies.

Firstly, the AR model is considered and parameter estimation is briefly discussed and performed. Thereafter we discuss the parameter selection process for these stock market time series in terms of the SSA algorithm. Finally, we then compare the accuracy of the predictions of the test data for the selected

models.

3.1 AR algorithm applied to stock market data

Autoregressive models are an entire study on its own, and we will therefore not describe the algorithm in detail (Interested reader can consider Brockwell and Davis, 2002). It is however important to take note of the fact that AR models require stationary data as input. A stationary time series, $\{Y_t | t = 1, \dots, T\}$, is one whose joint probability distribution, mean and variance do not depend on the point in time. Time series describing stock market data are clearly not stationary, since (amongst other things) they often include some increasing or decreasing trends. That is why, as we mentioned before, AR models are usually applied to the returns of a stock ($\{R_T = Y_t - Y_{t-1} | t = 2, \dots, T\}$) and not the stock values themselves. This differencing procedure is a common method we apply to some time series in order to eliminate some trend in the data thereby attempting to transform the non-stationary time series into a stationary time series. Many other transformations can also be used, but we will only consider differencing.

The AR algorithm then attempts to model the return of a given stock linearly in terms of previous returns, thereby stating that the returns produced for the relevant stock depends on the success of the stock in the past. It also incorporates additively for the possibility of some random Gaussian white noise error. This model can be formulated mathematically according to Equation 25.

$$\begin{aligned} R_T &= \alpha_1 R_{t-1} + \alpha_2 R_{t-2} + \dots + \alpha_p R_{t-p} + \epsilon_t \\ &= \sum_{i=1}^p \alpha_i R_{t-i} + \epsilon_t \end{aligned} \quad (25)$$

In this equation the only parameter to be specified is p . This parameter is similar to the window length parameter in the SSA methodology since it determines how far into the past one should peer in order to gain useful and significant information on the observation made today. For a given value of p , the α_i coefficients can easily be determined by minimizing the sum of squared ϵ_t values. Once these coefficients are determined, prediction can follow according to the model by applying Equation 26.

$$\hat{R}_{T+1} = \hat{\alpha}_1 R_T + \hat{\alpha}_2 R_{T-1} + \dots + \hat{\alpha}_p R_{T-p+1} \quad (26)$$

One thing to take note of in Equation 26 is the fact that predictions are based on noise contaminated observations and not observations filtered of noise like we had with the SSA methodology. It is true that the estimates of the coefficients are calculated in order to minimize the effect of the error component; however, since stock market activity is so volatile, this might prove to be a downfall of the AR algorithm in this context.

Since the α_i coefficients are numerically calculated by minimization of the sum of errors squared, we simply need methods to find the appropriate value for p . Traditionally one of two methods is used. The first method I will discuss entails the use of forward cross-validation to determine the parameter. In this process rolling one step ahead predictions are made, while varying the value of p . The choice of p resulting in the most accurate predictions of the cross validation segment should then be used on the test data segment. Alternatively, plots of the partial autocorrelations are often considered to determine the lags at which observations are still significantly correlated. This alternative approach can be thought of as a method using an in-sample approach to determine the appropriate value of p .

Returning to our practical scenario, we can clearly see that our three indices are not stationary. We

subject the three time series to the differencing operator in order to attempt to transform the data into stationary data. After said transformation, we have the following three time series representing the returns of the indices.

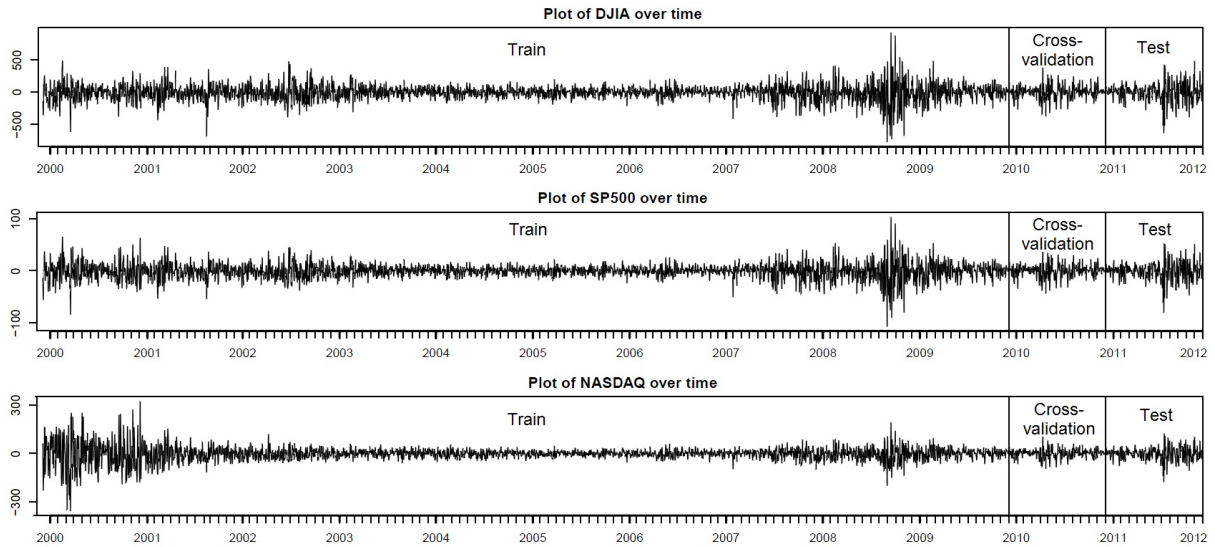


FIGURE 17: These three figures graphically illustrate the returns of the 3 stock market time series of interest. The first figure shows the returns for the DJIA index, the second illustrates that of the S&P500 index and the third figure is a plot of the returns of the NASDAQ index.

In these three figures, we clearly still see that the data is not quite stationary yet. Even though the expected value of the time series seems to be relatively constant, this is certainly not true for the variance. Clearly some segments has a higher level of variation. The requirements necessary for the AR model are therefore not strictly met; however, since we are merely considering a simple comparison between AR models and SSA models we disregard this fact.

In order to find the appropriate parameter p that we need to use for prediction, we turn to our cross-validation data. Rolling one step ahead predictions are found for all the values during 2010, for values of p ranging from 1 to 50. Mean Squared Errors were then calculated for each possible value of p in order to assess the accuracy of predictions and therefore the efficiency of the model. The following plots describe these MSE values.

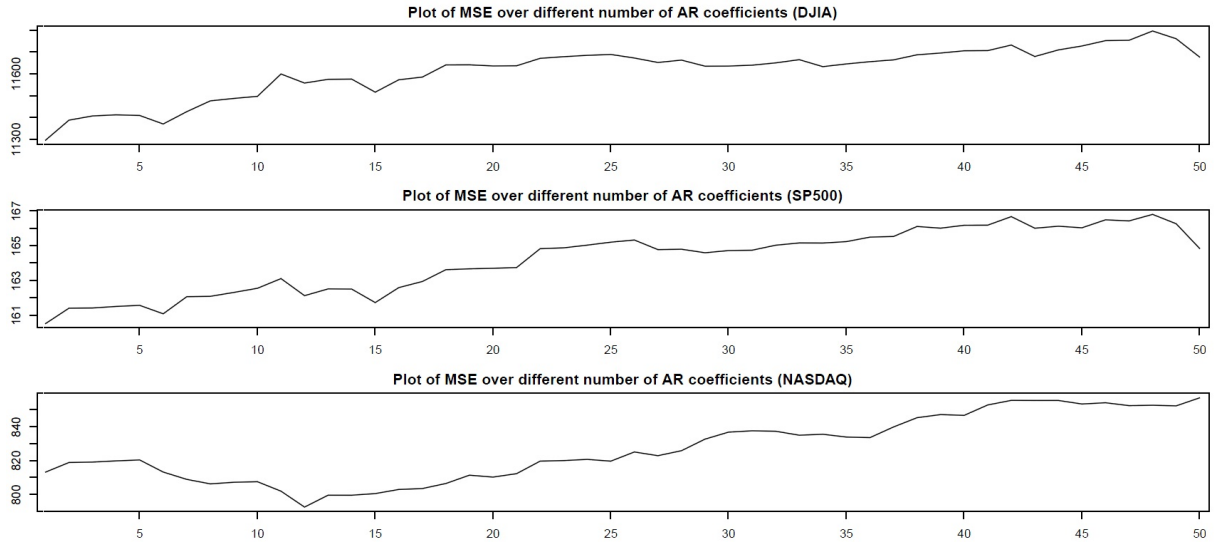


FIGURE 18: The three figures here illustrate how the Mean Squared Error of rolling one step ahead predictions of our cross-validation segment change as we vary the parameter p . The top figure shows the behaviour of the MSE of predictions on the DJIA returns, the middle figure represents that of the S&P500 index and the bottom figure illustrates the behaviour of the MSE of the predictions on the NASDAQ returns.

The top and bottom figures are clearly very similar. This is of course a result of the fact that the underlying index time series themselves are comparable. According to these figures it is clear that a value of $p = 1$ seems to provide quite good results in predictions of the DJIA and S&P500 indices. Some local minimum also seems to be evident for $p = 6$. For the NASDAQ index on the other hand, the bottom figure in Figure 18 shows that $p = 12$ results in the global minimum MSE. A possible explanation for the NASDAQ figure being different might be attributed to the fact that the variation in the NASDAQ index during the cross-validation period is significantly lower than that for the DJIA and the S&P500 and consequently "more predictable". Notice however that for this figure we also seem to have local minima at $p = 1$. According to this forward cross-validation process, models that should be considered are those where $p = 1/6/12$.

As we mentioned, one could also consider partial autocorrelations to determine the lags where significant correlations still reside. Consider the data given up to 2011; in-sample partial autocorrelations can be calculated with the lagged time series values for lags up to 30. A plot of these in-sample partial autocorrelations follows.

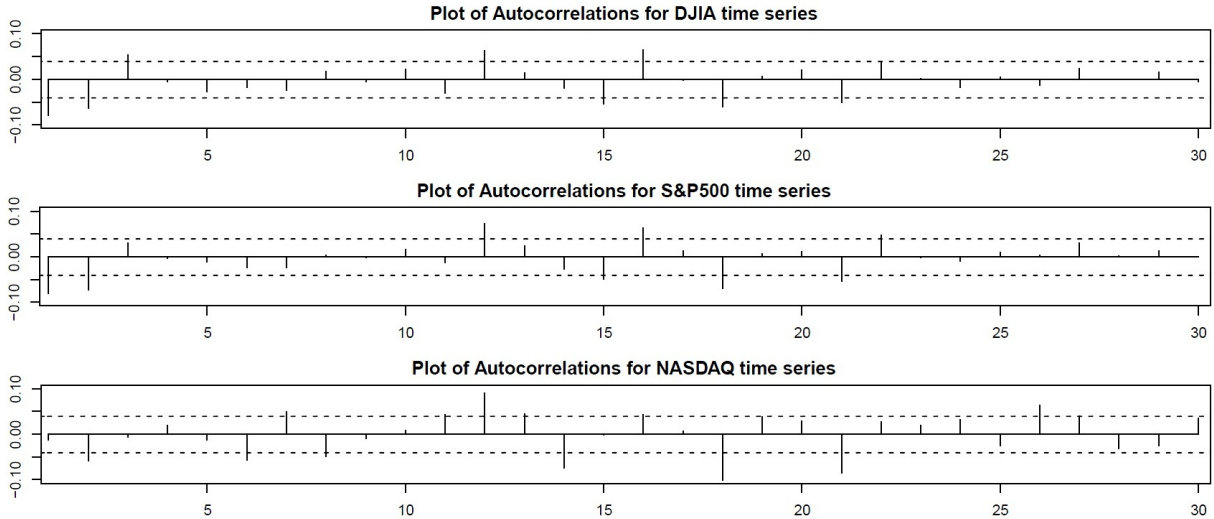


FIGURE 19: These 3 figures illustrate the in-sample partial autocorrelations calculated for each of the three given stock market index returns. The top figure can be associated with the DJIA, the middle figure can be associated with the S&P500 index and the final figure can be associated with the NASDAQ index.

It is interesting to see here that the partial autocorrelations are significant at lag of one, for the DJIA and the S&P500. This is not the case for the NASDAQ time series, where there is however significant autocorrelation at a lag of 6. This confirms the conclusions made during forward cross-validation, since predictions were more accurate when we had $p = 6$. Apart from this, the three autocorrelation plots seem comparable with significant correlations at lags of 2, 12, 16, 18 and 21.

In conclusion, considering all models proposed by forward cross-validation and in-sample partial autocorrelations, we have that possible values for models to be used include $p \in \{1, 2, 6, 12, 16, 18, 21\}$. This being said, I believe that cross-validation methods will prove to be more reliable, since they directly address the question of accurate prediction and not just in-sample autocorrelation. Furthermore, stock market returns are unlikely to be correlated at large lags. In general, stock market returns are modelled to be correlated only with the returns of the previous day ($p = 1$) or at most with the returns of the previous week ($p = 5$). I therefore believe that p should be chosen either as 1 or 6. The high variance noise in stock market data generally causes an overfitted model if p is chosen too large. We will however consider all of the values proposed above for completeness' sake.

3.2 SSA algorithm applied to stock market data

Now that we have found possible models that we can use from the family of AR models, we can continue to select the appropriate SSA models. After we have found the appropriate parameters for the SSA methodology, we can calculate predictions for the test data set using both SSA and AR models so a comparison can be made between the accuracy of the predictions of the individual prediction models. Starting the discussion on parameter selection for the SSA algorithm, it is important to notice that the AR model insisted on stationary data. This is not the case for a SSA. We can therefore apply the methodology directly to the stock index values and any transformation is optional.

To find the appropriate values for the window length and the dimension of the signal space, we once again turn to the cross-validation data. Rolling one step ahead predictions are calculated for this cross-validation segment while varying window length over the first 20 possible values ($L \in \{2, \dots, 21\}$) while dimension varied between all possible values $d = 1, \dots, L - 1$, for each of these choices of window

length. We therefore considered each possible combination of window length and dimension, for all window lengths up to 21. Larger values for window lengths can also be considered; however, as we will see in the results, larger window length values did not provide more accurate predictions. The curves of MSE over different choices of dimension were drawn for each window length on the same set of axes in Figure 20.

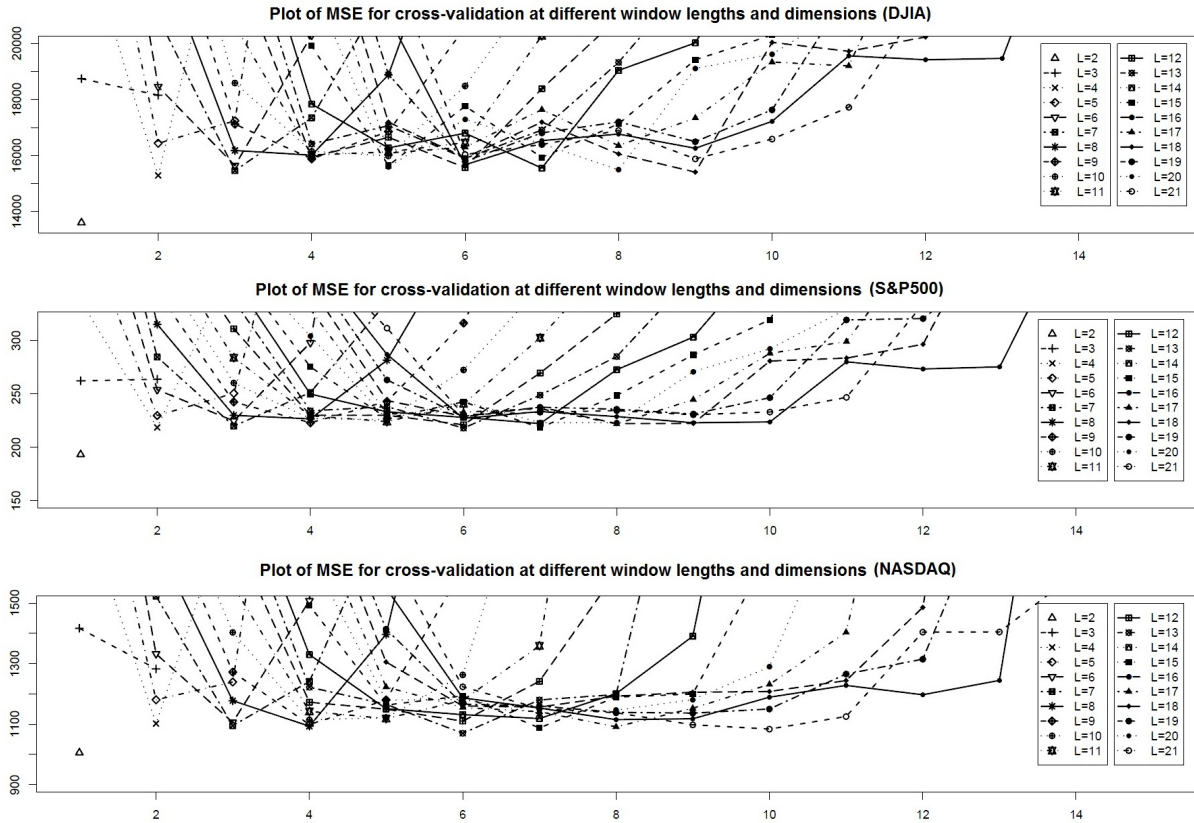


FIGURE 20: These three figures represent the prediction accuracy (MSE) associated with each of the models considered when applying these models to the cross validation segment of the time series to be predicted. In a given figure, a single curve represents how the MSE changes across different choices of dimension while keeping the window length fixed. In each figure there are therefore 20 lines, each representing one of the window length choices considered. The top figure plots these curves for the DJIA index, the second figure does the same for the S&P500 index and the final figure graphically illustrates these curves for the NASDAQ index.

Figure 20 is quite difficult to interpret since the plot is densely populated with points and lines. We therefore calculate the minimum MSE value for each window length and simply plot these points on the same set of axes.

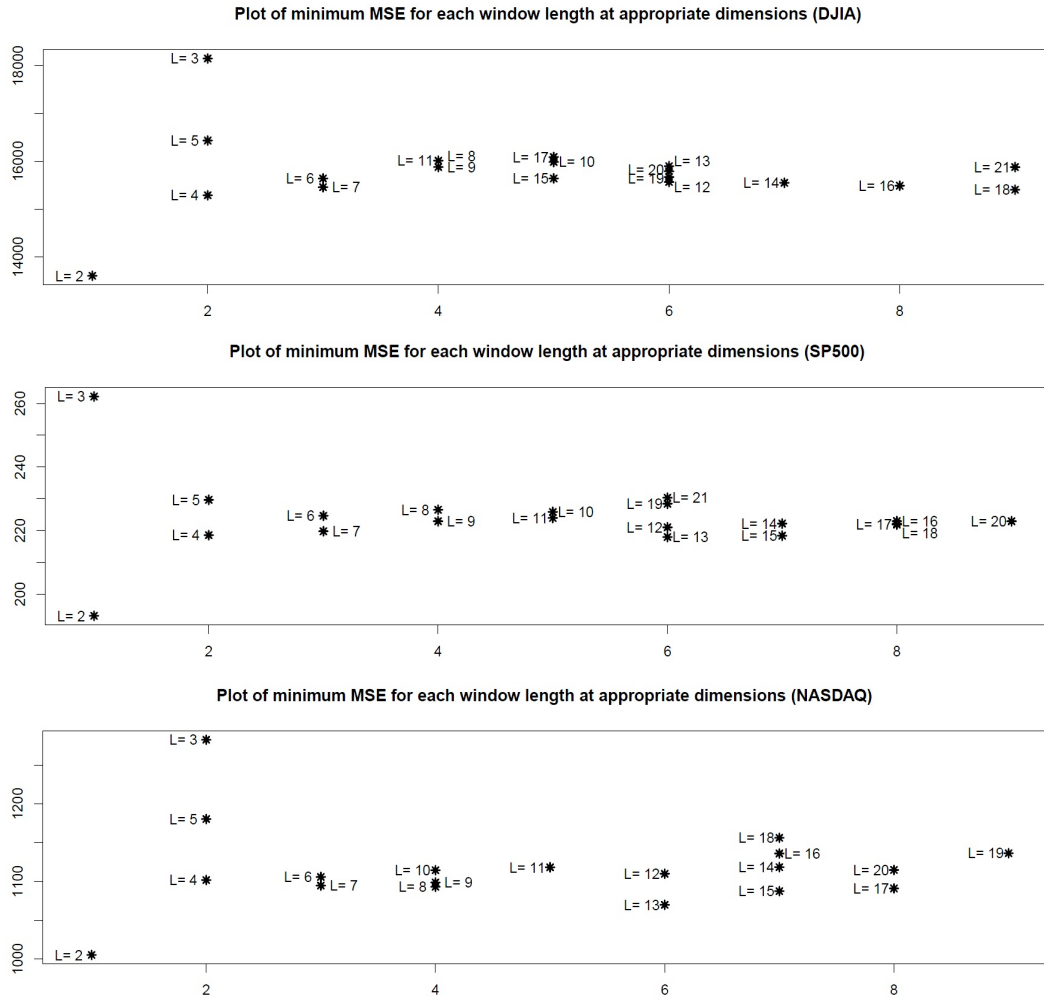


FIGURE 21: This plot is similar to Figure 20; however, only the optimum point (in terms of minimum MSE) for each choice of possible window length is drawn on the graph. These figures represent, from top to bottom, the minimum MSE obtained when choosing the appropriate window length and dimension for the DJIA index, the S&P500 index and the NASDAQ index.

It is clear that there exists a global minimum MSE at window length of 2 (with dimension 1) for each of these three time series. An increase in the window length clearly does not improve the prediction accuracy (which is why only the first 20 possible window lengths were considered). The fact that the SSA algorithm seem to prefer a window length of 2 implies that it also (similar to the AR algorithm) believes that these stock market indices are best predicted if we only consider the current value of the time series and accordingly predict that of tomorrow. This is similar to what we found with the AR model when predicting with $p = 1$. There is however cause for concern for two reasons. The first is the fact that research and our simulation studies showed that larger choices of window length is beneficial. Even when we simulated a structural break, choosing the window length too small resulted in very inaccurate predictions. The second cause for concern, and the more important one, comparing the MSE of these three scenarios with that of the cross-validation procedure performed for the AR model, we see that the minimum MSE is significantly larger here than we had with the AR algorithm. To therefore include a larger diversity (and perhaps a more competitive set) of models to compare with the AR algorithm, we also consider the possibility of applying the SSA algorithm with the differenced time series as we had with the AR model.

Once again, a similar cross-validation procedure is applied to the time series of returns. We let the relevant parameters vary over similar regions. This time however, we allow for all window lengths up to 35 ($L = \{2, \dots, 35\}$). The reason for this, as we will see now, is that when applying the SSA algorithm to the observed returns, larger values of L seem to provide accurate predictions. Thus, we consider a larger variety of window lengths in order to better study the effect of large window lengths. The dimension parameter then once again varies over all possible values for given window length ($d = \{1, \dots, L - 1\}$). For each of these combinations, we find rolling one step ahead predictions for all observations in the cross-validation segment.

When measuring the prediction accuracy of these three indices, results showed that for each of these possible window lengths, the optimum chosen dimension was $d = 1$. It also showed that any other choice of rank was not comparable in terms of prediction accuracy and can therefore be disregarded. Since this is the case, the prediction accuracy for different window lengths could easily be illustrated according to Figure 22.

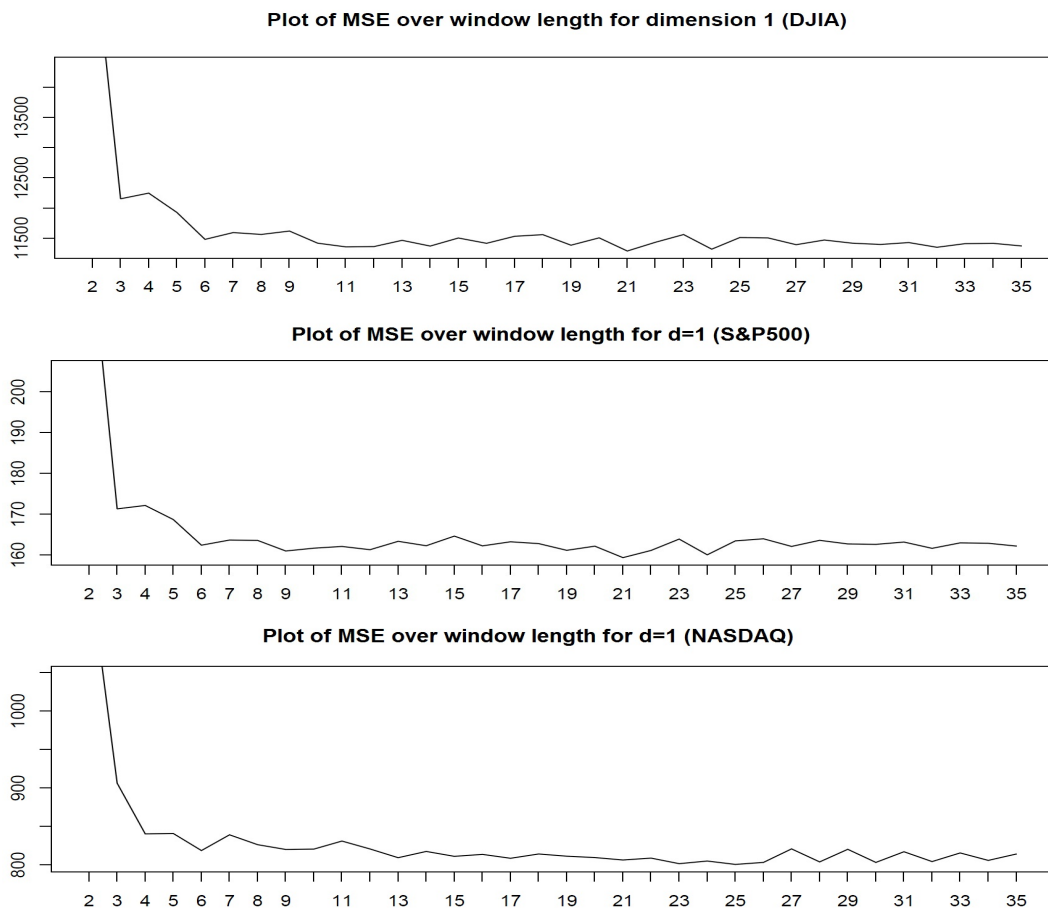


FIGURE 22: These three figures illustrate the prediction accuracy for each of the models considered. Since it was clear that the cross-validation predictions indicated that d should be equal to one, these figures show how the MSE (for the predictions on the cross-validation set) changes as the window length increases. Once again, the figures represent, from top to bottom, the results for the DJIA index, the S&P500 index and the NASDAQ index.

The results obtained from this cross-validation procedure is more what we would expect it to be according to the knowledge we gained during simulation examples. Small window lengths result in inaccurate predictions, but as the window length grows, the accuracy of the predictions increases and

stabilises around $L = 6$. According to these three figures, we can therefore propose any SSA model with $L \in \{6, \dots, 35\}$. If we were simply to consider the minimum MSE we would choose $L = 24$ with chosen dimension of 1 for the DJIA, $L = 21$ for the S&P500 index and $L = 25$ for the NASDAQ index. However, since the MSE values are so close together, we could perhaps consider alternative measures of prediction accuracy to better distinguish between the individual models. Table 1 tabulates all three mentioned measures of accuracy we are interested in, for window length values $L \in \{6, \dots, 35\}$ and choice of dimension $d = 1$.

The table confirms that the MSE stabilizes when the window length parameter reaches the value of 6. For values of window length larger than 6, the MSE fluctuates with not more than 5% of the minimum obtained value. These global minimum MSE's are reached at $L = 24$ for the DJIA index, at $L = 21$ for the S&P500 index and at $L = 25$ for the NASDAQ index. The MAE fluctuates even slightly less after $L = 6$ with variation of no more than 4% from the minimum obtained MAE. These minimum MAE values can be found in Table 1. Using MAE as function to be optimized might however not be appropriate for stock market data, since we are more interested in capturing large deviations and would therefore like to penalize more harshly for large errors. Furthermore the differences in the MAE are so small that one can hardly use this measure to effectively distinguish between different models. Very little information can therefore be obtained from the MAE.

The final measure of accuracy is the DOC. This is quite an important measure when considering stock market activity, since we want to accurately predict whether the value of the stock goes up or down. The DOC measures the proportion of instances where we correctly predict a positive or a negative return. This measure changes much more drastically than the MSE and MAE with values as low as 42% and reaching as high as 58%. One thing to notice with the DOC is the fact that the DOC values tend to have quite a few instances where it reaches proportions lower than the desired 50%. This is especially true for window lengths smaller than 20 and larger than 30. The minimum values for DOC when considering each of the three indices are obtained at $L = 23$ for the DJIA index, at $L = 13$ for the S&P500 index and at $L = 24$ for the NASDAQ index. However, DOC can be good when prediction accuracy is bad. For instance overfitting the model causes the algorithm to drastically change direction based on insignificant information. Consider the obtained minimum for the S&P500 index; at this choice of window length ($L = 13$), we can see from the obtained MSE and MAE that in general the algorithm does not predict accurately even though the DOC shows that it is optimum. We therefore need a good balance between DOC and the MSE.

In the comparison, we will consider each of the models that is said to be optimum with respect to given measure of accuracy for the sake of completeness. In other words for $L \in \{13, 17, 21, 23, 24, 26, 32\}$. However, as mentioned before, the MAE and DOC measures can often be misleading and one therefore needs to find a balance between these three measures so as to find a model that accurately predicts errors, while still harshly penalizing for large discrepancies and correctly predicting whether a stock increases or decreases. I will therefore focus on models between window length of 20 and 30, since at this point, the DOC seem to be quite stable and consistently above 50%. I propose that the model to be used for the DJIA index should be at $L = 24$, $L = 27$ or $L = 30$, for the S&P500 we should choose $L = 21$ or $L = 27$ and for the final index, NASDAQ, I suspect that the algorithm will perform well when $L = 25$ or $L = 26$. I propose these models since they produce near optimum measures of accuracy across all three measures.

	MSE			MAE			DOC		
	DJIA	SP500	NASDAQ	DJIA	SP500	NASDAQ	DJIA	SP500	NASDAQ
SSA(L=6)	11552.66	163.03	819.28	75.93	9.19	20.87	53.17%	51.19%	48.81%
SSA(L=7)	11662.87	164.18	839.33	75.64	9.11	21.09	51.59%	49.60%	53.97%
SSA(L=8)	11620.45	164.00	827.54	74.77	9.02	21.20	48.41%	49.20%	46.43%
SSA(L=9)	11680.28	161.59	819.78	75.49	9.09	20.62	44.44%	46.43%	50.79%
SSA(L=10)	11482.07	162.16	821.99	74.29	9.02	21.15	51.59%	50.79%	46.43%
SSA(L=11)	11416.05	162.58	831.90	73.85	9.06	21.18	51.59%	46.43%	48.86%
SSA(L=12)	11428.52	161.82	822.06	74.51	9.05	21.19	48.02%	48.81%	48.41%
SSA(L=13)	11525.20	163.78	809.49	74.41	9.04	20.59	55.16%	54.37%*	53.97%
SSA(L=14)	11429.25	162.75	818.31	74.02	9.03	20.99	50.00%	53.17%	48.02%
SSA(L=15)	11563.46	165.07	810.81	75.01	9.17	20.46	54.76%	52.38%	54.76%
SSA(L=16)	11473.68	162.69	813.94	74.39	9.08	20.68	50.00%	51.59%	48.41%
SSA(L=17)	11592.98	163.75	808.25	75.54	9.11	20.42*	53.57%	51.59%	57.14%
SSA(L=18)	11610.88	163.16	814.46	75.29	9.09	20.61	51.98%	50.00%	47.22%
SSA(L=19)	11453.35	161.77	810.99	75.04	9.07	20.66	49.60%	52.38%	55.16%
SSA(L=20)	11559.39	162.51	809.00	74.77	9.03	20.68	51.98%	53.57%	51.98%
SSA(L=21)	11358.03	159.97*	805.96	74.20	9.04	20.56	51.98%	52.78%	55.96%
SSA(L=22)	11434.87	161.18	808.84	74.26	8.99	20.66	50.40%	50.79%	54.37%
SSA(L=23)	11562.90	163.96	801.63	74.41	9.04	20.46	56.75%*	53.97%	55.95%
SSA(L=24)	11320.72*	160.11	805.04	73.97	8.96*	20.57	53.57%	50.00%	57.94%*
SSA(L=25)	11513.32	163.51	800.67*	74.26	9.02	20.47	53.97%	52.38%	55.95%
SSA(L=26)	11507.05	164.03	803.25	74.40	9.06	20.56	50.79%	50.00%	55.95%
SSA(L=27)	11394.54	162.16	820.85	74.04	9.00	20.73	53.57%	53.17%	51.98%
SSA(L=28)	11471.86	163.65	803.90	74.20	9.05	20.56	53.57%	51.19%	56.35%
SSA(L=29)	11419.14	162.77	820.31	74.22	9.05	20.76	51.19%	49.60%	50.79%
SSA(L=30)	11398.50	162.66	803.25	73.99	9.02	20.53	55.95%	52.38%	54.76%
SSA(L=31)	11429.86	163.23	817.14	74.16	9.05	20.75	51.98%	51.19%	51.98%
SSA(L=32)	11351.72	161.69	804.34	73.97*	9.03	20.53	49.60%	50.79%	55.56%
SSA(L=33)	11411.87	163.03	815.59	74.04	9.03	20.73	55.56%	51.19%	51.19%
SSA(L=34)	11417.05	162.93	805.86	74.18	9.05	20.53	50.00%	48.41%	53.97%
SSA(L=35)	11373.76	162.25	814.20	74.06	9.03	20.67	55.95%	50.79%	52.38%

TABLE 1: This table summarizes the three measures of prediction accuracy when predicting over the cross-validation segment of each of the three time series, while varying the window length between possible values ranging from 6 to 35. Values marked with an asterisk indicate the optimum window length for a given column.

3.3 Comparison and conclusion

Now that the appropriate parameters have been decided on for both the AR algorithm and the SSA algorithm, we can continue by comparing the two methodologies. Briefly revising the theory behind the two algorithms, with the AR algorithm, we modelled the returns of tomorrow linearly in terms of previous returns, through minimization of squared error. Furthermore, prediction followed according to unfiltered returns. Based on these results, we propose that the SSA methodology is superior to the AR model, not only since we model the returns non-linearly in some space spanned by singular vectors, but also since prediction follows from filtered time series observations. If we consider the underlying rationale of the methodology, we notice that the AR concept does not provide intuitive rationale when applied to stock market data, since we have little evidence to believe that stock market data can be modelled linearly. It can merely be interpreted as a method that models the returns according to the best first order approximation. When considering the SSA methodology on the other hand, we know that the algorithm considers segments of L consecutive observations, based on these observations the algorithm then decides what happens typically in such a segment and predicts accordingly. Analysts in the industry do exactly this. They believe (as all statisticians) that history will repeat itself. They therefore consider segments in the past, try to relate it with what is currently happening and predict accordingly. The SSA methodology is therefore effectively a technical analysis. A further benefit of the SSA algorithm is the fact that we do not require stationary data as we did with the AR model. We can therefore apply the methodology either to the time series representing the index values directly or alternatively take differences and consider these. Even though this is the case, we found during cross-validation techniques that when we apply the SSA algorithm to the closing prices directly, it resulted in inadequate predictions. We also believe that this will resonate when predictions on the test data are calculated.

To compare the two approaches in terms of prediction accuracy, we subject the test segment of the three time series (all observations during 2011) to all the proposed prediction models. Once again we will measure prediction accuracy in terms of our three before mentioned measures as seen in Table 2. In this table we tabulate the accuracy of the proposed models. These models include the AR algorithm with parameter $p \in \{1, 2, 6, 12, 16, 18, 22\}$, the SSA algorithm applied directly to the observed stock market index values (with $d = 1$ and $L = 2$) and finally the SSA models applied to the stock market index returns with $d = 1$, while varying $L \in \{13, 17, 21, 23, 24, 25, 26, 27, 30, 32\}$.

Even though we consider all of these possible models, we have certain models that are of primary interest. We mentioned that if we were to analyse stock market activity with an AR algorithm, we should not choose p too large so as not to overfit the model to the data. Our models of primary interest are therefore the AR(1) and AR(6) models, since these models are both parsimonious and gave most accurate predictions on the cross-validation set. For the SSA algorithm, we specified for each stock market index, the choice of L that minimizes the MSE for cross-validation ($L = 24$ for the DJIA, $L = 21$ for the S&P500 and $L = 25$ for the NASDAQ). We then also mentioned some alternative models for the SSA algorithm: $L = 27$ or $L = 30$ for the DJIA, $L = 27$ for the S&P500 and $L = 25$ for the NASDAQ. These models were not chosen to minimise the MSE for prediction, but rather to have near optimum measures of accuracy for all three measures. The before mentioned 9 models are tabulated in the first 9 lines of Table 2. Models chosen on alternative considerations are only included for completeness' sake and are therefore listed thereafter as models of secondary interest.

Upon first inspection, the differences might seem slight. For instance when comparing the optimally chosen SSA models of primary interest with the optimally chosen AR models, we see that the SSA algo-

	DJIA			SP500			NASDAQ		
	MSE	MAE	DOC	MSE	MAE	DOC	MSE	MAE	DOC
Models of primary interest chosen by minimization of the MSE for cross validation									
AR(1)	23300.43	110.94	51.19%	313.72	12.90	47.62%	1676.00	30.31	46.82%
AR(6)	23712.54	111.20	54.76%	319.75	12.93	49.60%	1686.73	30.37	50.40%
SSA (on R_t) (L=21)	23740.39	111.89	49.60%	310.28	12.79	53.17%	1674.48	30.48	50.00%
SSA (on R_t) (L=24)	23990.90	111.78	56.74%	310.14	12.80	56.74%	1663.27	30.22	58.73%
SSA (on R_t) (L=25)	23745.14	112.47	53.97%	319.48	12.99	50.40%	1667.89	30.26	52.38%
SSA models of primary interest chosen to have near optimal MSE, MAE and DOC for cross validation									
SSA (on R_t) (L=26)	24200.23	112.79	46.43%	324.04	13.04	46.03%	1661.68	30.17	56.35%
SSA (on R_t) (L=27)	23291.14	111.49	51.59%	314.86	12.91	50.00%	1680.82	30.38	51.98%
SSA (on R_t) (L=30)	23645.96	112.31	51.98%	318.73	13.01	50.40%	1666.91	30.13	53.17%
SSA models of secondary interest chosen on the basis of alternative considerations									
AR(2)	23650.24	111.94	50.79%	319.82	12.98	49.60%	1696.90	30.49	48.41%
AR(12)	24036.74	112.15	52.38%	322.46	12.95	48.02%	1728.27	30.57	51.59%
AR(16)	24109.00	112.53	55.16%	322.74	12.97	51.98%	1723.10	30.61	50.00%
AR(18)	23802.09	112.06	54.37%	317.84	12.89	51.98%	1701.30	30.46	48.02%
AR(21)	23725.35	111.67	55.95%	317.30	12.87	54.76%	1696.15	30.43	50.79%
SSA (on Y_t)	27147.46	124.07	50.40%	362.49	14.34	47.22%	1976.86	34.32	46.83%
SSA (on R_t) (L=13)	24496.21	112.78	52.78%	340.26	13.11	51.59%	1711.68	30.61	48.81%
SSA (on R_t) (L=17)	24148.92	111.94	55.16%	323.74	12.87	50.40%	1675.61	30.40	48.81%
SSA (on R_t) (L=23)	24239.56	112.92	53.97%	324.29	13.03	47.62%	1667.80	30.26	47.22%
SSA (on R_t) (L=32)	23655.71	112.00	50.00%	320.16	12.99	46.03%	1671.24	30.18	53.17%

TABLE 2: This table summarizes the three measures of prediction accuracy for the predictions on the test segment of each of the three stock market indices, as calculated by each of the models proposed during the cross-validation step. The bold cells indicate for which of the three time series the specific model was chosen.

rithm can only reduce the MSE with 1% and 0.5% for the S&P500 and the NASDAQ indices respectively and the SSA algorithm only slightly outperforms the AR algorithm when considering the DJIA. However, upon closer inspection we can see the additional benefit of the SSA model; that is the improvement it can provide in terms of DOC. The proposed SSA algorithms can consistently predict direction of movement for an index correctly an additional 5% more than the proposed AR algorithm. Furthermore, it does so not only without losing the accuracy of the prediction in terms of MAE and MSE, but improving (albeit little) on these measurements of prediction accuracy. This is of great use to us when considering stock market activity.

When considering all the AR models in general, we see that as the parameter p increases, the accuracy of prediction reduces in terms of the MSE and MAE. However, the algorithm correctly predicts the direction of change more frequently. This is because the algorithm is effectively overfitting the model to the data, causing sporadic predictions that are not accurate in terms of predicting magnitude, but coincidentally correct in terms of direction. This is of no use to us, since we not only want to correctly predict a positive or negative return, but also the magnitude of the change. This results in only one AR model being effective in terms of prediction, that is the AR model with $p = 1$, even though this optimum AR model is less effective in terms of DOC. This trade-off between correctly predicting magnitude of change and predicting DOC is not present with SSA models.

The stability of the SSA models can be seen when considering the proposed SSA models implemented on the index returns, both the models of primary and secondary interest. In the cross-validation step we noticed that the models with window lengths between 20 and 30 had predictions that were not only accurate in terms of magnitude of the error (MSE and MAE) but also in terms of the direction of change (DOC). This behaviour seems to reproduce here when testing the proposed models on the test data segment. By varying the window length in this interval, we are therefore not overfitting the model to the data (as we did with the AR model) but rather seeking some model that is optimal in terms of all three measurements of prediction accuracy. Notice also from the proposed SSA models that when the SSA algorithm was applied directly to the index closing prices, we found predictions that were completely ineffective in comparison with the other models considered here - as we expected. This supports our previous statement that the SSA methodology should therefore not be applied directly to stock market data. The reason for this is unknown and can be researched; however, intuitively I believe that it is a consequence of the volatility of stock market data. Because of this seemingly random up and down fluctuations in the time series, the SSA methodology is unable to extract some constant signal trend from the closing prices. Differencing methods can then be used in order to transform the time series into a nearly chaotic series of observations from which non-linear methods, such as the SSA methodology, can more easily extract some signal trend.

Another benefit of the SSA algorithm is the diverse number of models that perform adequately. With the AR algorithm, we quickly realize that only the AR(1) model can compete in terms of prediction accuracy. The SSA algorithm had models of secondary interest that produced relatively accurate predictions. Furthermore, the proposed models of primary interest, chosen based on optimization considerations applied to the cross-validation procedure, showed to have near optimum results in the test partition of our data as well. The chosen SSA models therefore seem to be relatively consistent. This is also very reassuring. In conclusion, this practical scenario considering stock market data seem to show that the SSA models not only outperform the AR models in terms of prediction accuracy measurements, but they are also more diverse, consistent and reliable.

4 Multivariate SSA

From the previous we now have the knowledge to evaluate a univariate stock market time series. However, this thesis attempts to improve accuracy of univariate predictions by including additional information obtained from the internet. The rationale of this hypothesis can be described when we more formally consider the Efficient Market Hypothesis. In the second chapter, we briefly referred to this hypothesis and said that it effectively implies that no financial gain can be obtained by predicting stock market activity. More formally Jensen (1978) provides the following definition for the Efficient Market Hypothesis:

A market is efficient with respect to information set, Θ_t , if it is impossible to make economic profits by trading on the basis of information set, Θ_t .

This thesis attempts to contradict this hypothesis by proposing that the internet is one of very few sources that actually allows us to react on information set Θ_t . If we are able to act on Θ_t while other investors are still reacting to information set $\Theta_{t-\delta t}$ (where δt is some lag at which information is obtained) we could perhaps gain useful information over other investors, enabling us to improve on prediction accuracy.

The SSA methodology is a univariate time series analysis methodology. Since we now need to incorporate additional information in our predictions, we have to formulate how we would do this in the SSA context. This chapter will therefore briefly describe current Multivariate Singular Spectrum Analysis (MSSA) methodologies and their disadvantages. Thereafter, an alternative multivariate methodology will be proposed and simulation examples will compare these MSSA techniques with each other and with the univariate SSA methodology. Finally, these methodologies will also be compared by evaluating their effectiveness in predicting stock market data with internet data used as auxiliary information in a practical scenario.

4.1 Current MSSA methods

Now that univariate SSA methodology have been described in detail, we can continue to the multiple time series analysis extension of the SSA methodology, i.e. the prediction of a time series in the presence of one or many auxiliary time series. There are two current methods worth mentioning here in the SSA context; namely Horizontal Multivariate Singular Spectrum Analysis (HMSSA) and Vertical Multivariate Singular Spectrum Analysis (VMSSA). Both of these methods are very intuitive extensions of SSA. For us to be able to discuss these multivariate extensions, we need to set the multivariate framework. Suppose therefore that we have a multiple time series consisting of K univariate series, denoted as given by Equation 27.

$$Y_{T,k} = (y_{1,k}, \dots, y_{T,k}) \quad \text{for } k = 1, \dots, K \quad (27)$$

In the SSA framework, each of these time series can be transformed into a trajectory matrix as denoted by Equation 28.

$$X_k = \begin{bmatrix} y_{1,k} & y_{2,k} & \cdots & y_{T-L_k+1,k} \\ y_{2,k} & y_{3,k} & \cdots & y_{T-L_k+2,k} \\ \vdots & \vdots & \ddots & \vdots \\ y_{L_k,k} & y_{L_k+1,k} & \cdots & y_{T,k} \end{bmatrix} \quad \text{for } k = 1, \dots, K \quad (28)$$

These K matrices can then be used to form a single trajectory matrix, which can then be subjected to the SSA algorithm. This concatenation of multiple trajectory matrices into a single trajectory matrix can be done in two ways as described by Golyandina (2001); vertical concatenation or horizontal concatenation. These two methods result in two different multivariate SSA techniques.

4.1.1 Horizontal multivariate SSA algorithm

Firstly, we discuss the more commonly used MSSA technique; that is the Horizontal Multivariate Singular Spectrum Analysis technique. In this approach we construct a single trajectory matrix by concatenating the K individual trajectory matrices in a horizontal manner next to each other as illustrated in Equation 29.

$$X^* = \begin{bmatrix} X_1 & \cdots & X_K \end{bmatrix} \quad (29)$$

Notice that this concatenation is only possible if we have that all K trajectory matrices have the same number of rows and therefore the same window length.

After the single trajectory matrix has been constructed, the HMSSA procedure follows much like the univariate SSA procedure. The trajectory matrix is subjected to an SVD, after which the individual single rank matrices are classified as either signal or noise. The final step in the filtering procedure on the other hand has to be modified slightly to take into consideration the fact that we have more than one time series in our trajectory matrix. Before subjecting our obtained signal matrix to diagonal averaging, we have to decompose the signal matrix obtained from X^* back to its original K matrices, so as to distinguish between the individual time series. Each individual trajectory matrix representing the signal matrix of a single time series is then transformed to a univariate time series by diagonal averaging and transformation. The prediction step is also identical to the univariate SSA approach only using the singular vectors obtained from the SVD of X^* . These singular vectors can then be used to predict any of the K time series. For example, to predict the k th time series we simply use Equation 30.

$$s_{T+1,k} = a_1 s_{T,k} + \cdots + a_L s_{T-L+1,k} \quad (30)$$

In Equation 30, $s_{i,k}$ is the i th signal component obtained for the k th time series after the filtering process have been completed and the coefficients for prediction can be determined from Equation 31.

$$\underline{a} = \frac{1}{1-v^2} \sum \pi_j \underline{u}_j^\nabla \quad (31)$$

Refer to Chapter 2.1.2 for definitions of π_j , v and \underline{u}_j^∇ .

The rationale behind this multivariate extension of the SSA methodology is quite clear. If we were to assume that the primary and the auxiliary time series share some mutual component, by including the auxiliary time series, we are including more observations (column wise) of length L . With this additional data, we might be able to more accurately estimate the singular vectors, which in turn improves on both the filtering and prediction stages of the SSA algorithm. Therefore, inclusion of the auxiliary time series emphasizes the mutual signal component, making it easier to extract in the SVD step.

4.1.2 Vertical multivariate SSA algorithm

A lesser known multivariate extension of the SSA approach is VMSSA. In the previous approach, numerous trajectory matrices were concatenated in a horizontal manner. In Vertical Multivariate Singular Spectrum Analysis (VMSSA), on the other hand, our K trajectory matrices are merely concatenated in a vertical manner on top of each other as illustrated in Equation 32.

$$X^* = \begin{bmatrix} X_1 \\ \vdots \\ X_K \end{bmatrix} \quad (32)$$

Notice in this equation that we are not restricted by matrices that must have equal window lengths, the concatenation is possible for different window lengths. However, the number of columns in all the matrices have to be the same. Typically, however, we have that all K trajectory matrices have the same dimensions.

Once again, the process then follows in a similar fashion to the HMSSA algorithm and the SSA algorithm. The X^* matrix is subjected to an SVD, after which the individual single rank matrices are grouped. After the signal matrix is obtained from X^* , we once again have to decompose the estimated signal into its individual time series matrices representing the signal of each time series before diagonal averaging is performed. Prediction is then also performed in a similar fashion to normal SSA, based on the singular vectors obtained from the SVD step. However, since the singular vectors were obtained from a trajectory matrix with $\sum_{k=1}^K L_k$ rows, the singular vectors are also of length $\sum_{k=1}^K L_k$. Subtle changes are therefore needed with VMSSA prediction. In order to adequately discuss the prediction process for VMSSA, we introduce the following notation of the singular vector obtained from SVD of the trajectory matrix in Equation 33.

$$\underline{u}_i = \begin{bmatrix} \underline{u}_i^1 \\ \vdots \\ \underline{u}_i^K \end{bmatrix} \quad \text{for } i = 1, \dots, d \quad (33)$$

In this equation, each \underline{u}_i^k is of length L_k . By making this decomposition on each singular vector, we are able to associate each segment of the singular vector with a specific time series.

Suppose we now want to predict the k th time series. In the univariate SSA method, prediction was based on the omission of the final element of the singular vectors when defining \underline{u}_i^∇ . However, since the final element in these singular vectors as obtained from the VMSSA methodology is associated with the K th time series, we need to adjust the algorithm appropriately. In order to make sure the final element in the singular vector refers to the k th time series, we define \underline{u}_i^* as in Equation 34.

$$\underline{u}_i^* = \begin{bmatrix} \underline{u}_i^1 \\ \vdots \\ \underline{u}_i^{k-1} \\ \underline{u}_i^{k+1} \\ \vdots \\ \underline{u}_i^K \\ \underline{u}_i^k \end{bmatrix} \quad \text{for } i = 1, \dots, d \quad (34)$$

Equation 34 therefore simply reorganizes the segments of the given singular vector by removing the k th segment and concatenating it to the end of \underline{u}_i^* .

By then defining $\underline{u}_j^{\nabla*}$ similar as we did in the univariate scenario in Chapter 2.2.2, as the vector containing the first $\sum_{k=1}^K L_k - 1$ elements in \underline{u}_j^* and $(v^*)^2 = \sum_j \pi_j^*$ where π_j^* represents the final element in \underline{u}_j^* , we can predict the next value of the time series according to Equation 35.

$$s_{L+1,k} = \underline{s}_k^{*'} \underline{a}^* \quad (35)$$

In the above Equation 35, \underline{s}_k^* consists of the final L_i signal values of time series $\{y_{T,i} | i = 1, \dots, k - 1, k + 1, \dots, K\}$ as well as the final $L_k - 1$ signal values of the k th time series (so that it matches the constructed $\underline{u}_j^{\nabla*}$) and \underline{a}^* can be calculated according to Equation 36.

$$\underline{a}^* = \frac{1}{1 - (v^*)^2} \sum \pi_j^* \underline{u}_j^{\nabla*} \quad (36)$$

Notice that if we want to calculate predictions several steps ahead iteratively, we need to use future values of the auxiliary time series as well. For k step ahead predictions, we therefore need to first calculate the predictions for the auxiliary time series in order to use the future values of the auxiliary time series in order to predict the future values of the primary time series.

The idea of the VMSSA procedure is slightly different to that of the HMSSA approach. It is not based on the assumption that the primary and auxiliary time series share some inherent signal component. Since prediction is based on the actual observed values in the auxiliary time series, we realize that the VMSSA methodology attempts to predict activity in the primary time series by also considering the current activity in the auxiliary time series. Rather than using auxiliary information to better estimate the inherent signal of the time series, the VMSSA methodology believes that if the auxiliary time series is behaving in some fashion, it is likely to cause some change in the primary time series as well. In statistical terminology, the HMSSA approach increases the number of observations (column wise), while the VMSSA approach increases the number of features (row wise) according to which prediction takes place.

4.1.3 Disadvantages of HMSSA and VMSSA

After a brief discussion of the two current multivariate extensions of SSA and the rationale behind these methodologies, we realise that these methodologies have possible flaws.

Considering first the VMSSA methodology, we should realize that by concatenating the individual trajectory matrices vertically, we are increasing the number of features and therefore the dimension of the observations. With an increase in dimensionality, we find ourselves more vulnerable to the curse of dimensionality. In brief, the idea behind the curse of dimensionality is that as dimension increases, the distance measures between points become less interpretable and less effective. Since Singular Value Decomposition is based on distances, the effectiveness of the decomposition decreases as well. Intuitively speaking, we can say that by including more features into our trajectory matrix, the estimation of the inherent signal becomes more complicated. Furthermore, if the included features are unnecessary, we are including additional unavoidable noise. This in turn affects prediction.

Another restriction of the VMSSA prediction method is the fact that the methodology requires tra-

jectory matrices with equal numbers of columns, $T - L + 1$. This might not seem like a very restrictive requirement, but since the number of columns depends directly on T , we are forced to choose a window length based on the length of the auxiliary time series. This often results in choices for window length that are not intuitive.

The preferred HMSSA methodology on the other hand is not as vulnerable to the curse of dimensionality, since with the concatenation, the dimension of the problem is not increased. Also, the restriction on the window lengths of the individual trajectory matrices having to be equal is less of a problem, since it produces an intuitive way to analyse the time series problem. However, there are other inadequacies of the HMSSA methodology.

Since the methodology is based on the assumption of some common inherent signal component present in both the primary and auxiliary time series, a possible disadvantage of this idea can stem from situations where this assumption might not be valid. There are numerous such scenarios and some of these will be considered in the simulation example so as to evaluate what results will hold and why. The point here however is that inclusion of some auxiliary time series can be severely destructive to prediction accuracy when this assumption is invalid. Together with this uncertainty in the inherent signal component as a result of inclusion of the auxiliary time series, we are also losing some of the interpretability on the choice of the dimension. Some of the signal components could belong to the primary time series and some to the auxiliary time series. If this is the case, how should the grouping step be approached?

Another disadvantage of the HMSSA methodology is evident when we consider two time series on different scales. If our primary time series assumes values with large variation, while the auxiliary information has a small variation, the auxiliary time series will be of very little help. During the SVD the algorithm will attempt to find the singular vectors describing most of the variation in the observed L consecutive time series. Since the auxiliary time series is regarded as observations out of the same population, their relatively small values will not be regarded as significant and singular vectors will remain unaffected. This problem could of course be addressed by simply rescaling our auxiliary time series, but this could simply lead to additional problems such as the optimal method of scaling.

These are merely theoretical disadvantages of both multivariate SSA methodologies. We will consider the problem areas in a simulation study. However, based on these possible inadequacies of the HMSSA and VMSSA algorithms I propose an alternative manner in which we can include auxiliary information in the SSA context.

4.2 Bayesian approach to combining time series

In general, with any Bayesian approach, we consider the estimate of some parameter to simply be an observation out of some distribution of possible parameter values. In the SSA context, since both signal estimation and prediction are based on the obtained singular vectors, we are interested in estimation of the true signal singular vectors. If we can therefore assign some distribution to these vectors, we could perhaps obtain some knowledge as to the true inherent signal. The singular vectors can then be estimated by considering the auxiliary time series as possible prior information while using the primary time series as additional information resulting in some posterior estimate of the column space.

We will first show that we can assume normality of the singular vectors obtained from the SVD step.

Based on this assumption, we will formulate an approach where we can then include auxiliary information in the SSA context, by using a Bayesian framework.

4.2.1 Multivariate normality of singular vectors

In this section, we propose that each of the d singular vectors associated with signal could perhaps independently be distributed according to the multivariate normal distribution. If we were to test multivariate normality of the singular vectors obtained in the SVD decomposition, we would have the following hypotheses:

$$H_{0i} : \underline{u}_i \sim N_L(\underline{\theta}_i, \Sigma_i) \quad \text{for } i = 1, \dots, d \quad (37)$$

These d hypotheses therefore test whether each of the obtained singular vectors could be distributed according to the multivariate normal distribution. Under this assumption we can multiply by the square root matrix of the inverse of the covariance matrix to find the hypothesis as given by Equation 38.

$$H_{0i} : \Sigma_i^{-1/2} \underline{u}_i \sim N_L(\Sigma_i^{-1/2} \underline{\theta}_i, I) \quad \text{for } i = 1, \dots, d \quad (38)$$

From this, we can see that under the original hypothesis, we have that each of the L elements in this vector ($\Sigma_i^{-1/2} \underline{u}_i$) should have univariate normal distribution, independently of the other elements in the vector. In order to test our initial hypotheses as given by Equation 37, we therefore effectively have $d \times L$ univariate normal hypothesis tests, since each of the d singular vectors has L elements. Many authors (Gnanadesikan, 1977 and Johnson & Wichern, 1992) suggest that an essential first step in assessing multivariate normality is to test univariate normality as above. Notice that this test of normality of the marginal distributions is compulsory for multivariate normality, but not necessarily sufficient. It would however give us some confidence in the assumption of normality.

In order to test hypotheses such as these, we need a number of observations of each of the singular vectors. These replicates are necessary firstly to estimate both θ and Σ and secondly to actually perform each of the hypothesis tests since one observation is not sufficient to estimate some distribution of the stochastic variable.

Singular vector replicates

To obtain a number of replicates of the singular vectors, we consider once again the decomposition of our given time series as described by Equation 39.

$$Y_T = S_T + N_T \quad (39)$$

In this equation the individual noise observations are stochastic and therefore follow some random distribution, not restricted to the normal distribution, while the signal component is fixed. Suppose for the moment that the true signal of the time series as well as the distribution of the noise are known (this will of course not be the case in practice). With this knowledge, bootstrapping methods can be used to construct several replicates of this time series by simulating from the distribution of the noise. For each of these time series, we can perform a Singular Spectrum Analysis on said time series and obtain a set of d singular vectors for each time series replicate. By doing so, we can obtain several singular vector replicates that can be used for our hypothesis test.

When we obtain these singular vector replicates, we have to keep two things in mind. The first thing to take into consideration is the fact that any singular vector is only unique up to multiplication with negative one. Secondly, the order of the singular vectors corresponds to the magnitude of the λ values and these singular values are often quite close to each other and even interchangeable (especially in the case of harmonic components). It is therefore not unlikely that for two consecutive time series we might find that some of the singular vectors are reflected about the x axis or that the order of some of the singular vectors might have changed. This has to be taken into consideration during this hypothesis test. When replicates of the singular vectors are obtained, we must ensure that the set of obtained singular vectors are all in similar form, representing the same signal component. The singular vectors therefore undergo a procedure where they are subjected to the applicable transformations necessary to ensure that the correct singular vectors, of the same sign, are grouped in a set.

We might for instance find, for two consecutive time series, the following singular vectors:

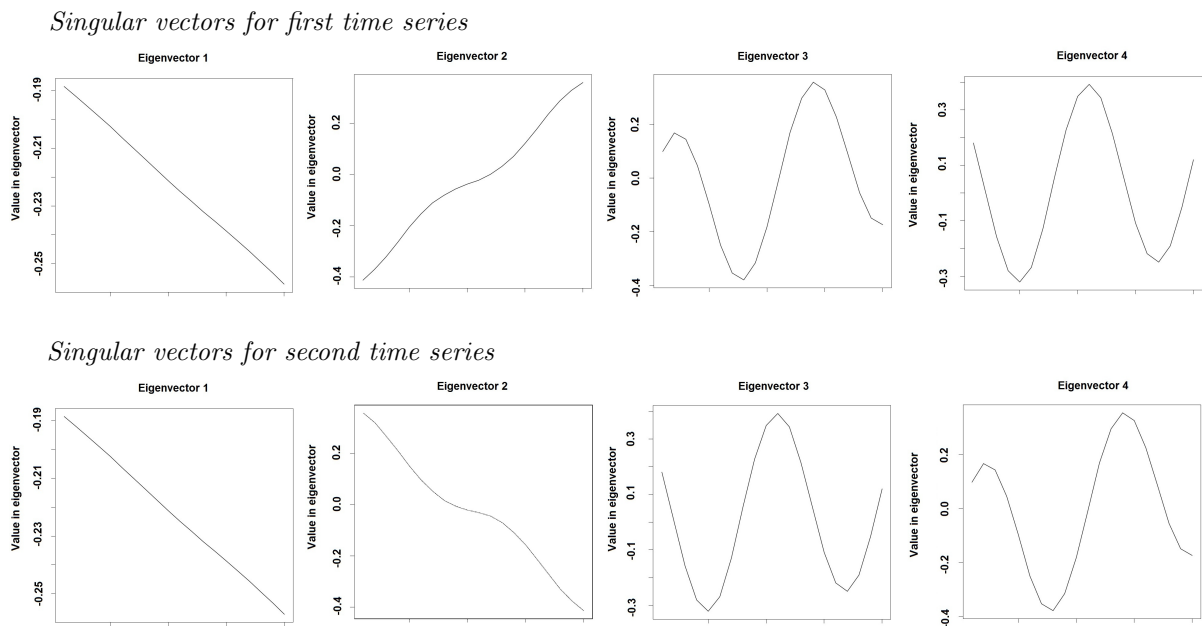


FIGURE 23: These figures illustrate how two similar time series could produce singular vectors that are slightly different in terms of order and multiplication with negative one

We therefore need to find from the second set of singular vectors, those that correspond to the singular vectors in the first set. We do this by considering the Euclidean distance between the singular vectors as well as that of the singular vectors multiplied by -1. The optimum arrangement of the second set of singular vectors is chosen to minimize these Euclidean distances between the singular vectors and the first set of singular vectors. With this algorithm, the second set of singular vectors is transformed to singular vectors as illustrated in Figure 24.

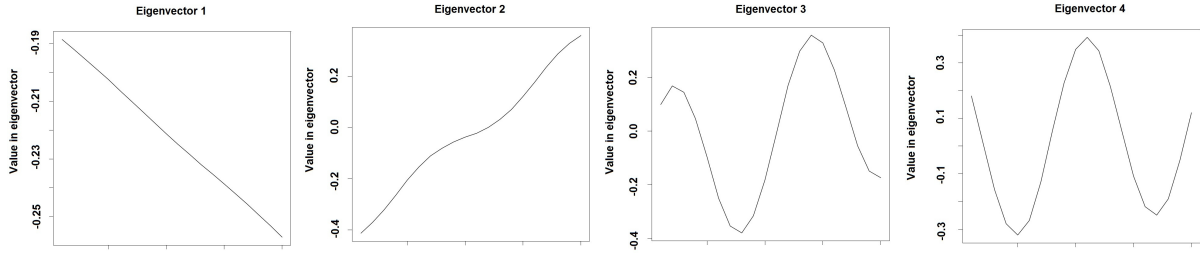


FIGURE 24: These figures illustrate how we can transform given singular vectors to ensure that two sets of singular vectors are similar

This set then corresponds to the first set of singular vectors. We can then consider the set of singular vectors corresponding to each individual component so that these changes in the sign of the singular vectors and the arrangement of the singular vectors do not influence the marginal distributions of each singular vectors.

With the obtained singular vectors, we can continue to the next step in the hypothesis test; that is testing normality of the individual elements in the transformed singular vectors as given by Equation 38.

Univariate test for normality

With replicates of the singular vectors obtained, we can estimate both θ and Σ and calculate the several replicates of the vectors $\hat{\Sigma}_i^{-1/2}\underline{u}_i$. Once these replicates are obtained, we simply need to individually test normality of the elements of these vectors by using univariate tests for normality. Notice of course that using $\hat{\Sigma}_i^{-1/2}\underline{u}_i$ instead of $\Sigma_i^{-1/2}\underline{u}_i$ affects the distribution of these vectors; however, with a large enough number of replicates, the estimate of Σ will be accurate enough and the effect will be negligible. Univariate tests for normality start with null and alternative hypotheses as given by Equation 40.

$$\begin{aligned} H_0 : x &\sim N_L(\mu, \sigma) \\ H_a : x &\text{ is not normally distributed} \end{aligned} \tag{40}$$

Normally, a hypothesis test is constructed in such a manner that the null hypothesis would describe the status quo, while the researcher believes that his data sufficiently contradicts the null hypothesis. By doing so, he can directly control the level of certainty in our decision. Unfortunately, this is not an option when testing for normality and we therefore have here that the null hypothesis describes the assumption the researcher wants to accept. However, we only want to do so if the hypothesis is correct. We are therefore much more interested in type two errors than type one errors, i.e. we are more concerned with rejecting the hypothesis when it is required to do so than we are with incorrectly rejecting the hypothesis. The power of the hypothesis test is therefore of great importance.

There are a number of tests for univariate normality. According to Pearson *et al.* (1977) these tests mainly consists of two types; directional tests and omnibus tests. Directional tests focus on a specific type of departure from normality. These tests are more effective if we have some prior knowledge on the distribution and would like the hypothesis to be sensitive to certain types of departure from normality. Omnibus tests however are generally based on two test statistics measuring skewness and kurtosis from an observed data set. These two test statistics are then combined to form a single omnibus test statistic which is then more efficient in testing any kind of departure from normality. Since we do not have any prior knowledge on the distribution of these elements we are concerned with here, we will be using omnibus tests rather than directional tests.

Several papers were studied to consider different possible tests for normality (Pearson *et al.*, 1977; Oztuna *et al.*, 2006; Razali and Wah, 2010). Some of the popular omnibus tests used in these papers include the Shapiro-Wilk test for departure from normality and the D'Agostino test for normality (interested reader can see Shapiro and Wilk, 1965 as well as D'Agostino and Pearson, 1973). There is speculation as to which of the broad spectrum of possible hypothesis tests performs better, but no conclusion has been drawn with certainty. Pearson *et al.* (1977) mentioned in their very broad study on these tests, that it is not yet possible to conclude on a uniform most powerful test for departure from normality. In Razali and Wah's comparison (2010), it was however found that in general the Shapiro-Wilk hypothesis test has high power properties. Oztuna *et al.* (2006) came to the same conclusion, viz. that the high power properties of the Shapiro-Wilk test are desirable and prescribe using this test. The writers then continue by mentioning that this test does however have one major disadvantage; that is the Shapiro-Wilk test does not perform well when there are ties in the data. We therefore use both the Shapiro-Wilk's test as well as D'Agostino's univariate test for departure from normality.

These univariate tests for normality can therefore be used to individually test each element of the transformed singular vectors. However, since we need to find some general conclusion as to the distribution of the vector as a whole, we need to calculate some combined p-value over all L hypothesis tests in a single singular vector. We do so by using theory from multiple hypothesis tests.

Multiple hypothesis test

When we are working with multiple hypothesis tests, we should remember that if the number of hypotheses increases, the probability of rejecting a few of the true hypotheses incorrectly increases. We therefore also have to take the number of hypothesis tests into consideration. There are several methods to control for the type 1 error rate mentioned above. Well known methods include the Bonferoni principle and Bonferoni related techniques. These maintain the Family Wise Error Rate (FWER); however, it significantly decreases the power of our hypothesis test.

Newer methods create a potential gain of power by maintaining the false discovery rate instead of the error rate. The False Discovery Rate (FDR) is a concept introduced by Benjamini and Hochberg (1995) and is defined to be the expected proportion of errors committed by falsely rejecting the null hypothesis. Notice that, if all tested hypotheses are true, controlling the FDR will effectively be the same as controlling the FWER. We will be using Benjamini and Hochberg's procedure (1995) of adjusting our p-values of the multiple hypothesis test accordingly to maintain this FDR. We can therefore test these hypotheses at the same error rate with more power.

Simulation example

Now that each of these components of our hypothesis tests have been discussed, we can use a simple simulation example to test this proposition of normality for a given situation.

We started the above theory by assuming that we know the signal of a given time series as well as the distribution of the noise. In order to remain consistent, we continue our simulation study on a time series similar to those discussed in Chapter 2. Equation 41 illustrates both the deterministic trend as well as the noise component of the time series.

$$\begin{aligned} S_t &= t + 10\sin(2\pi t/12) \\ N_t &\sim \text{Uniform}(-2, 2) \end{aligned} \quad \text{for } t = 1, \dots, 100 \quad (41)$$

Notice in Equation 41 that the uniform distribution was chosen for the error component. Since we propose that the singular vectors are normally distributed regardless of the underlying noise distribution, we want to ensure that the underlying distribution of the error component does not influence our results and we therefore did not choose the commonly used normal distribution. It is also the same distribution of error that we used in our previous simulation study in Chapter 2.

Using this signal component as given by Equation 41, we can clearly see that the time series has signal space of dimension 4. Simulation studies in Chapter 2 showed that for a time series such as this, it is sufficient to choose a window length of 24. We therefore now want to show that if we were to analyse this time series using the SSA methodology, with window length of 24 ($2 \times L$), we would find that each of the 4 singular vectors associated with signal can be regarded as observations from a multivariate normal distribution. Our proposed hypotheses then follow according to Equation 42.

$$\begin{aligned} H_{0i} : \underline{u}_i &\sim N_L(\underline{\theta}_i, \Sigma_i) && \text{for } i = 1, \dots, d \\ H_{ai} : &\text{The singular vector does not have a multivariate normal distribution} \end{aligned} \quad (42)$$

In order to test these hypotheses, we use the theory described above. Firstly we need to find several singular vector replicates through bootstrapping methods. Thereafter, the singular vectors are subjected to the appropriate transformation and the individual singular vectors are tested for univariate normality using both the Shapiro-Wilk and D'Agostino tests for departure from normality. Finally, the p-values are adjusted to control for the False Discovery Rate. We chose to find 250 replicates of the singular vectors, by simulating 250 time series according to Equation 43.

$$\begin{aligned} Y_{j,t} &= S_t + N_{j,t} \\ N_{j,t} &\sim \text{Uniform}(-2, 2) && \text{for } t = 1, \dots, 100 \text{ and } j = 1, \dots, 250 \end{aligned} \quad (43)$$

Notice once again that this will of course not be possible in practice. In a practical scenario we will be able to empirically estimate both the noise and signal components. The signal component will be estimated from the SSA analysis and the noise component can then be estimated by using empirical bootstrap methods. This entails simulating (with replacement) from the observed residuals obtained from the SSA analysis. However, since this simulation is simply intended to give some evidence supporting our assumption of normality, we use the true inherent signal and noise components to simulate the bootstrap time series.

From each of these 250 time series, we obtain 4 singular vectors. The appropriate transformations are made to these singular vectors as discussed.

Before we can consider the normality of each of the individual elements of the singular vectors, we first have to multiply the singular vectors with the appropriate matrix in order to ensure that we can assume independence of the elements of the singular vector and therefore consider these elements separately. Let $\underline{u}_{k,j}$ represent the k th singular vector of the j th time series, each of length 24. We can now use the 250 replicates of each of the 4 singular vectors to assess the distribution of each singular vector. We calculate the sample covariance matrices S_k for these singular vectors and pre-multiply the set of singular vectors by the inverse of the matrix square root of the sample covariance matrices to obtain the required $S_k^{-1/2} \underline{u}_{k,j}$ vector.

In a first attempt to assess the distribution of the singular vectors, it might give us some insight to consider the quantile-quantile normality plots for each of the elements in the 4 vectors. See Appendix A (Figures A.1 to A.4) for these plots. We can see clearly that these quantile-quantile plots for normality behave really well near the mean, it does however deviate from the hypothesised red line at the far ends in some figures. We can use the Shapiro-Wilk and D’Agostino hypothesis tests to see whether these deviations are significant. We derived in previously that our hypothesis test of normality of each singular vector is equivalent to testing the following 96 hypotheses:

$$\begin{aligned}
 H_{0i} : \underline{\epsilon}_j \Sigma_i^{-1/2} \underline{u}_i &\sim N_L(\Sigma_i^{-1/2} \underline{\theta}_i, I) && \text{for } i = 1, \dots, 4 \text{ and } j = 1, \dots, 24 \\
 H_{ai} : &\text{For some } j \text{ we do not have normality}
 \end{aligned}
 \tag{44}$$

For these 96 tests we can find the 96 p-values for both the Shapiro-Wilk and the D’Agostino univariate normality tests. These p-values are given in Table 3.

Comp	Singular vector 1		Singular vector 2		Singular vector 3		Singular vector 4	
	p-value (S-W)	p-value (D’Ago)	p-value (S-W)	p-value (D’Ago)	p-value (S-W)	p-value (D’Ago)	p-value (S-W)	p-value (D’Ago)
1	<0.001*	0.108	0.103	0.084	0.292	0.214	0.182	0.095
2	<0.001*	0.155	0.044*	0.030*	0.920	0.705	0.571	0.458
3	<0.001*	0.550	0.003*	0.006*	0.627	0.468	0.295	0.402
4	<0.001*	0.263	0.092*	0.203	0.034*	0.036	0.417	0.735
5	<0.001*	0.175	0.232	0.635	0.456	0.354	0.501	0.434
6	<0.001*	0.016*	0.317	0.996	0.077	0.239	0.920	0.820
7	<0.001*	0.290	0.948	0.938	0.073	0.553	0.748	0.452
8	<0.001*	0.234	0.125	0.030*	0.423	0.152	0.275	0.575
9	<0.001*	0.313	0.741	0.828	0.969	0.929	0.167	0.226
10	<0.001*	0.121	0.301	0.592	0.175	0.193	0.435	0.534
11	<0.001*	0.840	0.263	0.052	0.993	0.921	0.736	0.628
12	<0.001*	0.504	0.820	0.501	0.842	0.569	0.512	0.144
13	<0.001*	0.877	0.464	0.588	0.240	0.098	0.504	0.334
14	<0.001*	0.028*	0.626	0.913	0.806	0.621	0.121	0.079
15	<0.001*	0.158	0.440	0.395	0.067	0.209	0.654	0.956
16	<0.001*	0.305	0.919	0.605	0.539	0.501	0.695	0.447
17	<0.001*	0.203	0.965	0.937	0.334	0.705	0.131	0.302
18	<0.001*	0.314	0.277	0.111	0.442	0.565	0.639	0.469
19	<0.001*	0.018*	0.768	0.819	0.315	0.075	0.374	0.748
20	<0.001*	0.307	0.604	0.822	0.247	0.111	0.028*	0.017*
21	<0.001*	0.140	0.926	0.864	0.806	0.483	0.407	0.517
22	<0.001*	0.672	0.877	0.754	0.792	0.526	0.243	0.279
23	<0.001*	0.021*	0.120	0.805	0.399	0.475	0.681	0.593
24	<0.001*	0.125	0.942	0.929	0.572	0.455	0.564	0.516

TABLE 3: This table consists of p-values testing normality for each of the components in the transformed singular vectors. There are therefore 4 columns, one for each of the significant singular vectors and 24 rows for each of the elements in these vectors. The p-values for the univariate tests of normality are tabulated above.

The first thing we notice from Table 3 is the column of very small p-values associated with the Shapiro-Wilk test applied on the first singular vector. According to these p-values, we have sufficient reason to believe that none of these elements are normally distributed. We have to however keep in mind that the Shapiro-Wilk test is proven to be inaccurate when observations are close together or tied. Since there is very little variation in the first singular vector, we find that the Shapiro-Wilk test concludes that the elements are not normally distributed. D’Agostino’s test on the other hand shows some evidence as to the normality of the first singular vector. For the other singular vectors, we see that some of the p-values are significantly small; small enough for us to reject a single hypothesis. However, as we mentioned, with an increase in the number of hypotheses, it becomes more likely for us to reject a couple of hypotheses incorrectly. We therefore have to adjust our p-values in order to reduce this probability of incorrectly rejecting our hypotheses. Our p-values are therefore adjusted according to the Benjamini and Hochberg (1995) procedure and tabulated in Table 4.

Comp	Singular vector 1		Singular vector 2		Singular vector 3		Singular vector 4	
	<i>p-value</i> (<i>S-W</i>)	<i>p-value</i> (<i>D’Ago</i>)	<i>p-value</i> (<i>S-W</i>)	<i>p-value</i> (<i>D’Ago</i>)	<i>p-value</i> (<i>S-W</i>)	<i>p-value</i> (<i>D’Ago</i>)	<i>p-value</i> (<i>S-W</i>)	<i>p-value</i> (<i>D’Ago</i>)
1	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
2	<0.001*	0.877	0.965	0.665	0.993	0.929	0.920	0.956
3	<0.001*	0.877	0.082	0.146	0.993	0.929	0.920	0.956
4	<0.001*	0.877	0.965	0.996	0.993	0.856	0.920	0.956
5	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
6	<0.001*	0.391	0.965	0.996	0.993	0.929	0.920	0.956
7	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
8	<0.001*	0.877	0.965	0.665	0.993	0.929	0.920	0.956
9	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
10	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
11	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
12	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
13	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
14	<0.001*	0.587	0.965	0.996	0.993	0.929	0.920	0.956
15	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
16	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
17	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
18	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
19	<0.001*	0.420	0.965	0.996	0.993	0.929	0.920	0.956
20	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.415
21	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
22	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956
23	<0.001*	0.465	0.965	0.996	0.993	0.929	0.920	0.956
24	<0.001*	0.877	0.965	0.996	0.993	0.929	0.920	0.956

TABLE 4: This table consists of adjusted p-values testing normality for each of the components in the transformed singular vectors. There are therefore 4 columns, one for each of the significant singular vectors and 24 rows for each of the elements in these vectors. The adjusted p-values for the univariate tests of normality are tabulated above.

Considering Table 4, we see that for the D'Agostino test for normality, all our adjusted p-values are not significant. Most of these adjusted p-values are above 90%, giving very persuasive evidence to show that the elements are indeed normally distributed. A single adjusted p-value of 8.2% would be a cause for concern if it was not for the fact that the remaining adjusted p-values all deliver very persuasive evidence.

For the Shapiro-Wilk test for normality, we see that for the second, third and fourth singular vectors, we also have adjusted p-values that give very compelling evidence for normality of the individual elements. The p-values associated with the first singular vector, however, still show that this singular vector is not normally distributed. This can once again be attributed to the fact that there is so little deviation in the elements of the first singular vector that the Shapiro-Wilk test is simply ineffective in testing for normality.

Conclusions on normality

A suspicious reader might not be persuaded by the positive results of this single simulation example. Criticism could include the number of observations (250) over which these hypothesis tests are performed, the structure assigned to the error component, and whether the choice of window length could possibly affect this assumption. In order to study these effects, further simulation studies were considered.

The first step in these further simulation studies was to consider the effect that different number of bootstrap replicates had on this assumption of normality. Considering the adjusted p-values for tests based on varying numbers of simulated time series, we found that for smaller numbers of simulated time series these results are even more persuasive. For larger sample sizes however, we found that adjusted p-values reject the assumption of normality more often. Occasionally results are still conclusive, but as the sample size increases this happens less often. This is to be expected. Normality tests in general are less effective when applied to large data sets. In fact, for the articles cited in the univariate tests for departure from normality, the normality tests are seldom evaluated for sample sizes larger than 100. When tests for normality are applied to very small samples, it is unlikely that the non-normality will be detected, making the type two error probability very large and the power consequently very low. When the test is applied to very large samples on the other hand, the assumption of normality is rejected more easily, since normality is rejected even when slight imperfections occur. Sample size therefore does have a very important role in testing for normality here; however, even when taking this into consideration, the adjusted p-values still deliver compelling reason to believe in normality of the singular vectors.

The second possible reason to be sceptical could be the choice of uniformly distributed error component as in the simulation study above. This distribution for our error is of course not a practical one since our error component is seldom restricted to some interval. We therefore experimented with some other well known distributions. For the symmetric Gaussian distribution and Student's t-distribution we found similar results. With our error component following one of these distributions we could with confidence assume normality of the underlying singular vectors. When we considered less intuitive and non-symmetric distributions such as the exponential distribution, the gamma distribution and the chi-squared distribution, we found that if our error component followed any of these distributions (with varying parameters) that the adjusted p-values would more often reject the assumption of normality. Decreasing our sample size to 50 (as the tests were intended for) did improve our results to give more convincing evidence on normality, but not convincingly so. One thing to remember is that these non-symmetric distributions are not intuitive ones in the context of time series analysis.

Another possible point of criticism might be allegations that the structure of the signal component might have had an affect on our result of normality in the simulation example. Several different struc-

tures were tested using the same procedure and we found that this structure had little to no effect on our conclusions. Instead of the signal component being a cause for concern, we found that the signal to noise ratio had a much larger effect on our hypothesis tests. If the amplitude associated with the noise was low in comparison with the amplitude of the signal component, our results were in favour of the assumption of normality. For instance, when we considered the possibility of our error component having a non-symmetric beta distribution (with varying parameters), we still found in favour of normality for our singular vectors, since our noise component was restricted to a small interval. As the error component grew in magnitude however, we found less persuasive results, often rejecting the hypothesis of normality.

In this more extensive simulation study, we found that some evidence found in favour of our proposed distribution of the singular vectors and some evidence found results against the assumption. However, this is when considering the adjusted p-values for our hypothesis test. When considering quantile-quantile normal plots, we always saw that the dominant amount of observed quantile points were similar to the theoretical normal distribution quantile points, with some exceptions at far ends which is (as mentioned previously) typical when considering normality tests for large samples.

In practice it is often suggested that when we are testing for normality, instead of applying (often inaccurate) tests for normality, we should rather consider the nature of the problem and consider the practical validity of the assumption. One should see these hypothesis tests on the marginal distributions as reason to believe that this assumption of normality is not an unlikely one rather than a formal proof of normality. Additional to these hypothesis tests, quantile-quantile plots were used to answer the question "Is our data normal enough?". Often we find that the mere assumption of normality (without any evidence thereof) allows us to understand the problem at hand better and to elaborate on the scenario. In this case, we might not have formally proven normality, but we might still be able to use this assumption to improve some of the methods in SSA.

One final inadequacy that I feel needs to be mentioned here is the assumption of independence between the individual singular vectors. This assumption is inherently implied by constructing the hypothesis test given by Equation 37. This is of course not a valid assumption at all in the SSA context. However, with the application of this assumption of normality we will see that this assumption of independence produces an additional benefit in the SSA context. Furthermore, we would be able to test whether the entire set of singular vectors might have multivariate normal distributions and so we could incorporate any possible correlations between the singular vectors; however, this significantly complicates simulation studies and (as we will see later) would require significantly more computational power. A similar assumption is made in the naive Bayes classifier, for similar reasons.

4.2.2 Bayesian MSSA

Based on the above theory, stating that the obtained singular vectors representing the signal space of a given time series is distributed according to the multivariate normal distribution, we are now able to consider SSA and the estimation of the singular vectors in a Bayesian framework.

Suppose we have primary time series $Y_{1,T} = \{y_{1,1}, \dots, y_{1,T}\}$ which can be decomposed into a signal component, $S_{1,T} = \{s_{1,1}, \dots, s_{1,T}\}$, and a noise component, $N_{1,T} = \{n_{1,1}, \dots, n_{1,T}\}$. Suppose further we have that the signal component is of dimension d_1 with singular vectors representing the orthogonal basis for the column space, $\underline{u}_{1,i}$ for $i = 1, \dots, d_1$. We can now assign some distribution to these singular

vectors according to Equation 45.

$$\underline{u}_{1,i}|\underline{\theta}_i, \Sigma_{1,i} \sim N_L(\underline{\theta}_i, \Sigma_{1,i}) \quad \text{for } i = 1, \dots, d_1 \quad (45)$$

In order for us to be able to use this distribution, we have to estimate the mean vector $\underline{\theta}_i$ as well as the covariance matrix $\Sigma_{1,i}$. Estimates of these parameters can be found by obtaining bootstrap replicates of these singular vectors in a non-parametric fashion. With a given dimension d_1 , we can estimate $S_{1,T}$ and $N_{1,T}$. Even though we do not know the distribution of the noise component, we can simulate B bootstrap error components, $N_{1,t}^b$, by empirically drawing (with replacement) from the estimated $N_{1,t}$. With these bootstrap error components, we can create bootstrap replicates of the primary time series $Y_{1,t}^b$ and consequently bootstrap replicates $\underline{u}_{1,i}^b$ of the singular vectors for $i = 1, \dots, d_1$. With these bootstrap replicates we can therefore estimate $\underline{\theta}_i$ and $\Sigma_{1,i}$ by the sample mean vector ($\bar{\underline{u}}_{1,i}$ as given by Equation 46) and the sample covariance matrix ($W_{1,i}$ as given by Equation 46) of the singular vectors.

$$\begin{aligned} \bar{\underline{u}}_{1,i} &= \frac{1}{B} \sum_{b=1}^B \underline{u}_{1,i}^b \\ W_{1,i} &= \frac{1}{B-1} \sum_{b=1}^B (\underline{u}_{1,i}^b - \bar{\underline{u}}_{1,i})(\underline{u}_{1,i}^b - \bar{\underline{u}}_{1,i})' \quad \text{for } i = 1, \dots, d_1 \end{aligned} \quad (46)$$

Analogous to this we also have some auxiliary time series $Y_{2,T} = \{y_{2,1}, \dots, y_{2,T}\}$ which can also be decomposed into a noise component, $N_{2,T} = \{n_{2,1}, \dots, n_{2,T}\}$, and a signal component, $S_{2,T} = \{s_{2,1}, \dots, s_{2,T}\}$, of dimension d_2 with singular vectors representing signal denoted by $\underline{u}_{2,i}$ for $i = 1, \dots, d_2$. Suppose we suspect that one of these singular vectors contains some information about our true signal vector in the primary time series, $\underline{\theta}_i$. Under the assumption of normality, this information can be incorporated into our original estimation of the unknown signal parameter in the form of a prior distribution as shown in Equation 47.

$$\theta_i \sim N_L(\bar{\underline{u}}_{2,j}, W_{2,j}) \quad \text{for some } j \quad (47)$$

Combining our prior distribution with the likelihood derived from Equation 45, we find the posterior distribution of $\underline{\theta}_i$ as given by Equation 48.

$$\begin{aligned} \theta_i|\bar{\underline{u}}_{1,j} &\sim N_L(\underline{u}_{1,i}^{Bayes}, (W_{1,j}^{-1} + W_{2,j}^{-1})^{-1}) \\ \text{where } \underline{u}_{1,i}^{Bayes} &= (W_{1,j}^{-1} + W_{2,j}^{-1})^{-1}(W_{1,i}^{-1}\bar{\underline{u}}_{1,i} + W_{2,j}^{-1}\bar{\underline{u}}_{2,j}) \end{aligned} \quad (48)$$

After these transformations of our singular vectors, we have d_1 vectors describing the signal space of the given primary time series. Unfortunately these vectors are not necessarily orthonormal vectors. Since the prediction algorithm is based on orthonormal vectors, the obtained $\underline{u}_{1,i}^{Bayes}$ must first be subjected to the Gram-Schmidt procedure in order to obtain the set of orthonormal vectors ($\{\underline{u}_{1,i}^*|i = 1, \dots, d_1\}$) describing the same signal space as that described by the $\underline{u}_{1,i}^{Bayes}$ vectors.

The idea now is that the Bayesian estimate ($\underline{u}_{1,i}^*$) of the true signal singular vector ($\underline{\theta}_i$) is now a more accurate estimate than $\underline{u}_{1,i}$. Since estimation and prediction depend on the calculated singular vectors, the Bayesian estimators of the signal space will result in less accurate filtering, but it might be more accurate in terms of prediction.

Notice that this entire methodology, as described above, relies on the assumption that some auxiliary singular vector contains some information on one of the primary singular vectors. For the time being, we simply assume that this assumption is valid. However, in the fifth scenario in the simulation study, we

will propose a way to test whether this assumption is a valid one.

4.3 Simulation study

In order to test the validity of this proposed Bayesian methodology, we start an extensive simulation study where the Bayesian methodology is compared to current MSSA methods in the context of several different scenarios. By doing so, we can assess the strengths and weaknesses of the methodology and compare it with those of HMSSA and VMSSA.

We will consider the effect that noise components of different magnitude have on prediction accuracy as well as the effect that an auxiliary time series with slightly different signal component has on prediction accuracy. After that we will consider two areas where the existing multivariate SSA techniques are famous for being inadequate; that is when the auxiliary time series either has signal with little deviation (relative to the primary time series) or when both time series have harmonic components but with different periodicity. For each scenario, we will change only one aspect of our time series so as to investigate only the effect of the change. Our baseline scenario will be identical to those investigated in the simulation study in Chapter 2.

4.3.1 Effect of noise component

The first concept we wish to study in this simulation example is the effect that different noise components can have on prediction accuracy. Several scenarios will be discussed here. Primarily, we will investigate high noise and low noise scenarios. We will then also discuss a situation where the two time series have identical structure, but different noise components. Finally, we also consider the effect of having correlated noise components. Notice that throughout this simulation study, we will use uniformly distributed noise components. The reason for this is to ensure that the Bayesian approach does not unfairly benefit from normally distributed noise. Also, with uniformly distributed noise, we can more easily control the amplitude of the noise component.

Scenario 1

In the first of many scenarios in our simulation study, we consider a primary and auxiliary time series both having identical underlying signal component similar to the signal component studied in Chapter 2. The noise components included in the time series are also identical in this scenario, both noise components being simulated from a uniform distribution with low variance. Mathematically formulated, our primary and auxiliary time series can be described by Equation 49.

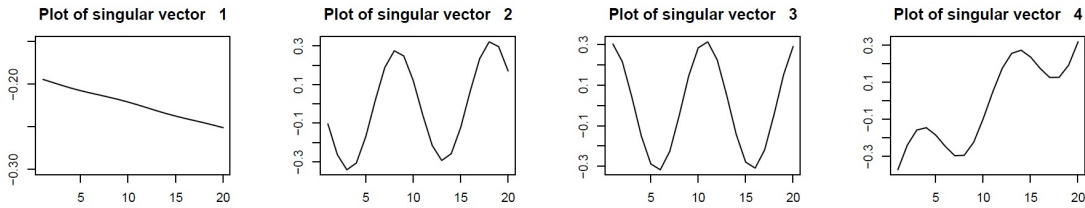
$$\begin{aligned}
 y_{1,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{1,t} \\
 y_{2,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{2,t} \\
 \varepsilon_{1,t} &\sim \text{Uniform}(-1, 1) \\
 \varepsilon_{2,t} &\sim \text{Uniform}(-1, 1)
 \end{aligned}
 \quad \text{for } t = 1, \dots, 120
 \tag{49}$$

Both time series here are of length 120. The first 100 observations from both time series are used in order to train our model and rolling one step ahead predictions can then be made for the remaining 20 observations. No cross-validation segment is necessary, since we know from Chapter 2 how our parameters should be chosen. The predictions on the final 20 observations can be made by using each of our multivariate prediction methods as well as the univariate SSA approach. The 20 one-step-ahead predictions can then be used to compare the accuracy of predictions for the three different MSSA procedures

as well as with the predictions of the univariate SSA methodology.

Consider a single given time series simulated according to Equation 49. We know by now that these time series both lie in a 4 dimensional signal space. Furthermore, simulation studies in Chapter 2 showed that using a window length of 20 provided near optimal prediction accuracy. Using these values as our parameters we can perform an SSA analysis on the first 100 observations of both the primary and the auxiliary time series. We would typically find the following singular vectors representing signal space for the primary and auxiliary time series.

For the primary time series:



For the auxiliary time series:

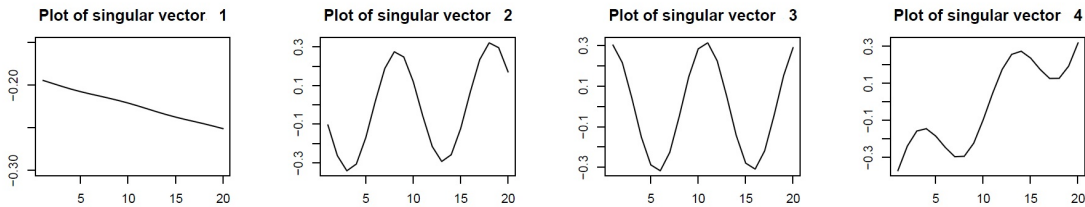


FIGURE 25: These figures represent plots of the estimated singular vectors associated with the signal space of the primary time series (top) and auxiliary time series (bottom).

These two sets of plots do not seem to differ at all. This similarity in the singular vectors is of course merely a consequence of the fact that there is little variation in the underlying noise component. Signal components can therefore be extracted very accurately. According to these plots, we can clearly see that the singular vectors obtained from the auxiliary time series can indeed give us some prior knowledge on the vectors associated with the true signal space, since the two sets of singular vectors are so similar. The assumption necessary for the Bayesian extension of the SSA algorithm is therefore met and we can apply the proposed algorithm. This would result in singular vectors representing signal space, as shown in Figure 26 (when taking 100 bootstrap replicates to estimate the appropriate parameters).

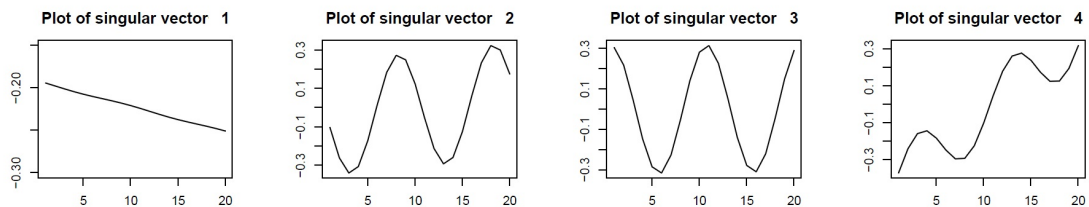


FIGURE 26: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the BMSSA approach. Notice that these singular vectors have already been subjected to the Gram Schmidt procedure in order to ensure that the vectors are orthonormal.

Once again, this set of singular vectors do not even seem to have changed. This is to be expected

since the low variance noise component ensures already accurate estimates of the vectors corresponding with signal, and little transformation is therefore needed.

The question is now, how do these Bayesian singular vectors perform in terms of prediction? Does it improve firstly on the univariate predictions and secondly can it provide more accurate predictions than the HMSSA and VMSSA methodologies? In order to answer these questions, we calculate 20 one-step-ahead predictions and determine the accuracy of these predictions by comparing the predicted values with the true deterministic signal component (as we did in Chapter 2). Table 5 tabulates the MSE for each of these methodologies.

<i>Method</i>	<i>MSE</i>
SSA	0.1405
HMSSA	0.1137
VMSSA	6.7950
BMSSA	0.1183

TABLE 5: This table represents the average MSE values for the four prediction methodologies. These MSE values are obtained from a single simulation.

For this single simulation we see that the Bayesian methodology indeed improves on the univariate SSA predictions. However, the HMSSA predictions are slightly better than that of BMSSA. The VMSSA predictions on the other hand are completely inaccurate. Unfortunately, one simulation is not sufficient evidence to conclude significant differences. We therefore perform 200 such simulations for the scenario described above and calculate the MSE for the 4 prediction methods for each of the simulated time series.

In order to compare these 200 MSE values, we first consider a plot of the Empirical Cumulative Distribution Function (ECDF) as given by Figure 27. This plot can give us information on the distribution of the obtained MSE values.

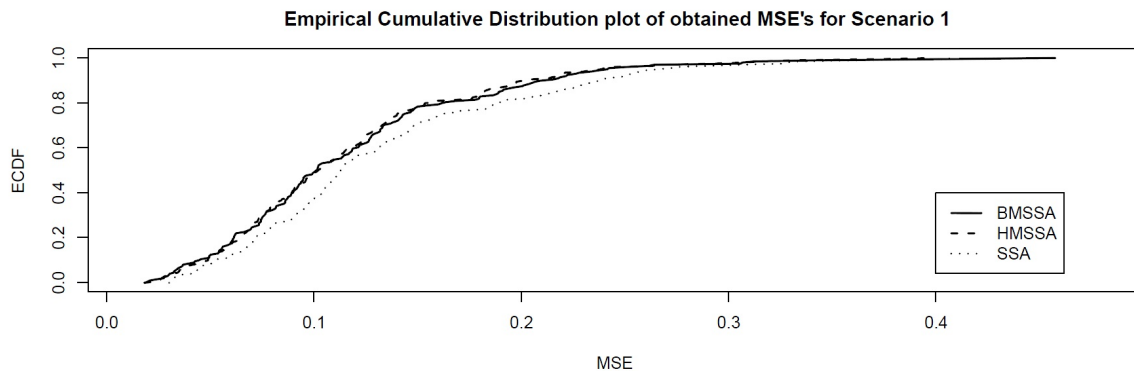


FIGURE 27: This figure represents the Empirical Cumulative Distribution Function plots of the 200 MSE values obtained from three of the possible prediction methods.

Figure 27 omits the ECDF of the VMSSA prediction MSE values. This omission is simply a consequence of the fact that the VMSSA algorithm continued to produce very inaccurate predictions that were not comparable to any of the other prediction methods. Figure 27 does however give some insightful information on the BMSSA and HMSSA prediction methods. Clearly, both these multivariate techniques improve on the univariate prediction methods. However, there are also very slight differences between

the HMSSA and BMSSA predictions in terms of MSE values.

In order to conclude on the significance of the differences between the four prediction methods, we calculate the average value of the 200 MSE's. Based on these average MSE values, we can test the hypothesis that there is a statistically significant difference between the average MSE obtained from the Bayesian approach and that of the other prediction methods by performing a Student's t-test on the average MSE values. The test statistics and associated p-values are tabulated in Table 6.

<i>Method</i>	<i>Average MSE</i>	<i>T-statistic</i>	<i>p-value</i>
SSA	0.1326	2.0559	0.0405 *
HMSSA	0.1158	-0.2750	0.7835
VMSSA	6.8700	180.6689	<0.001 *
BMSSA	0.1177		

TABLE 6: This table represents the average MSE values for the four prediction methodologies. These MSE values are obtained from 200 simulations. Students' t-test statistics and p-values are also calculated to test whether there are significant differences between the BMSSA predictions and that of other prediction methods.

According to Table 6 the Bayesian approach significantly improves on predictions obtained from both the VMSSA approach as well as that of the SSA approach. We can also conclude from this table that even though the average MSE associated with the Bayesian approach is slightly higher than the HMSSA average MSE, the difference is not significant. Notice also that the VMSSA approach is completely inaccurate, even in comparison with the univariate SSA predictions.

Considering each of the 200 simulations individually, we can also find the prediction method that produced the smallest average MSE for each specific simulation. Table 7 tabulates the proportion of times the BMSSA methodology outperformed another approach in terms of average MSE. The table also includes p-values that test the hypothesis that these proportions equal 50%.

	<i>Proportion</i>	<i>p-value</i>
BMSSA outperformed HMSSA	42.5%	0.0339 *
BMSSA outperformed VMSSA	100%	<0.001 *
BMSSA outperformed SSA	82.5%	<0.001 *

TABLE 7: This table gives the proportion of the 200 simulations where the Bayesian approach produced a lower average MSE than the other prediction methods. The significance of these differences are then also tested.

According to these tests on the proportions, we can conclude that the HMSSA methodology outperforms the BMSSA methodology a significant proportion of times. The BMSSA methodology did however consistently improve on univariate SSA predictions and always on VMSSA predictions.

In conclusion, it seems that the HMSSA methodology is significantly superior to the proposed Bayesian methodology in this scenario, but does not reduce the MSE by significant amounts. Also, since the BMSSA predictions outperform the SSA predictions, we can clearly see that the BMSSA methodology is able to extract valuable information from the auxiliary time series. This is not the case with the Vertical

Multivariate SSA methodology.

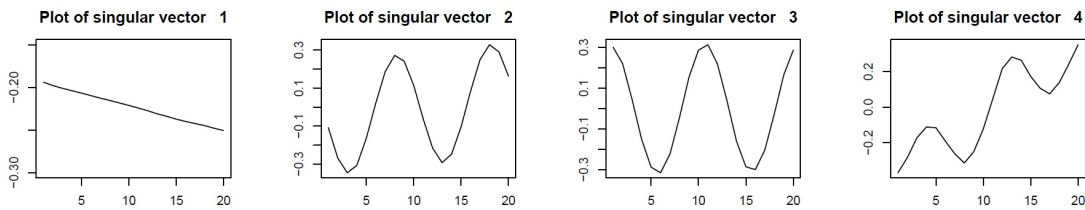
Scenario 2:

In Scenario 1 we found that when we have little noise or low amplitude noise, the HMSSA methodology is slightly superior to the BMSSA methodology. In this scenario, we continue our simulation study by inspecting the consequences of high amplitude noise on prediction accuracy of each of our methodologies. We do so by simulating time series according to Equation 50.

$$\begin{aligned}
 y_{1,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{1,t} \\
 y_{2,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{2,t} \\
 \varepsilon_{1,t} &\sim \text{Uniform}(-10, 10) \\
 \varepsilon_{2,t} &\sim \text{Uniform}(-10, 10)
 \end{aligned}
 \quad \text{for } t = 1, \dots, 120
 \tag{50}$$

A time series simulated according to Equation 50 therefore has a noise component with 10 times the amplitude of that of Scenario 1. Upon inspection of the singular vectors obtained for each of the individual time series, we see that the noise has started to affect estimation of the significant singular vectors. According to Figure 28, the final singular vector clearly deviates from the true signal component, since the harmonic component seems to be slightly contaminated with noise.

For the primary time series:



For the auxiliary time series:

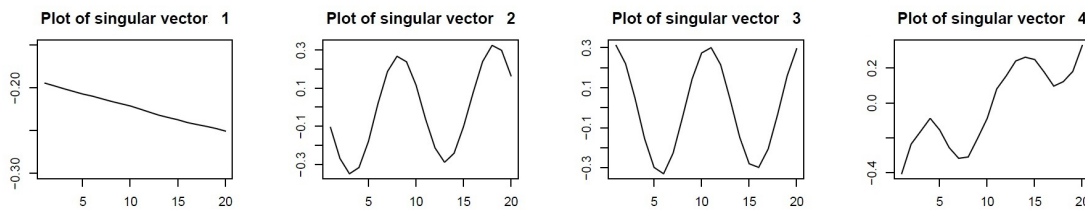


FIGURE 28: These figures represent plots of the estimated singular vectors associated with the signal space of the primary time series (top) and auxiliary time series (bottom).

Even though the individual components obtained from each time series are slightly inaccurate, we can still clearly see that they are attempting to estimate the same signal components. The required assumption for the proposed methodology, that the auxiliary time series provide prior information on our primary time series, is therefore a valid one and we can therefore combine them in a Bayesian context to perhaps obtain a more accurate estimate of the inherent signal component. Figure 29 shows the singular vectors obtained from both the Bayesian approach as well as that of the HMSSA methodology.

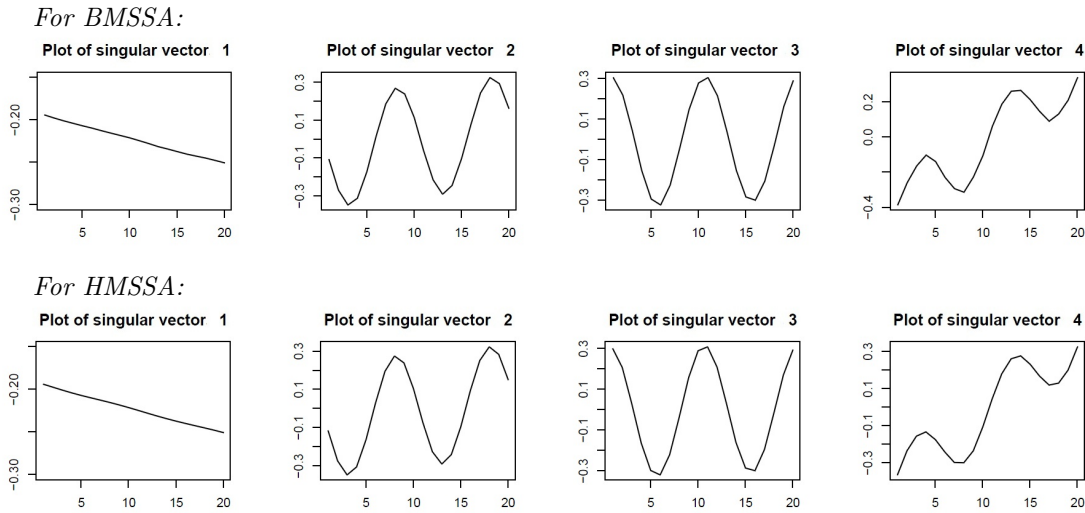


FIGURE 29: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the BMSSA approach (top) and the HMSSA approach (bottom).

Notice that the VMSSA singular vectors are not given here, since they are of different dimension and therefore not comparable to the singular vectors as obtained by the BMSSA and HMSSA approaches. Upon inspection of the singular vectors in Figure 29, we see that there are no obvious changes in the first three singular vectors. The fourth singular vector on the other hand seems to represent a smoother harmonic component than the initial SSA singular vectors and therefore seems to be a more accurate estimate of the true signal space in comparison with what we had with Figure 28. This holds for both the HMSSA algorithm and the BMSSA algorithm. The question is now, which of these singular vectors can provide more accurate predictions. To answer this question, we once again simulate 200 time series according to Equation 50 and calculate rolling one-one-step-ahead predictions for the final 20 time series values using all multivariate SSA prediction methods as well as the univariate SSA algorithm. For each of these simulations, we calculate the MSE of the prediction. Comparing these 200 simulated MSE values, we consider a plot of the ECDF for each of our prediction methods. This plot is given in Figure 30.

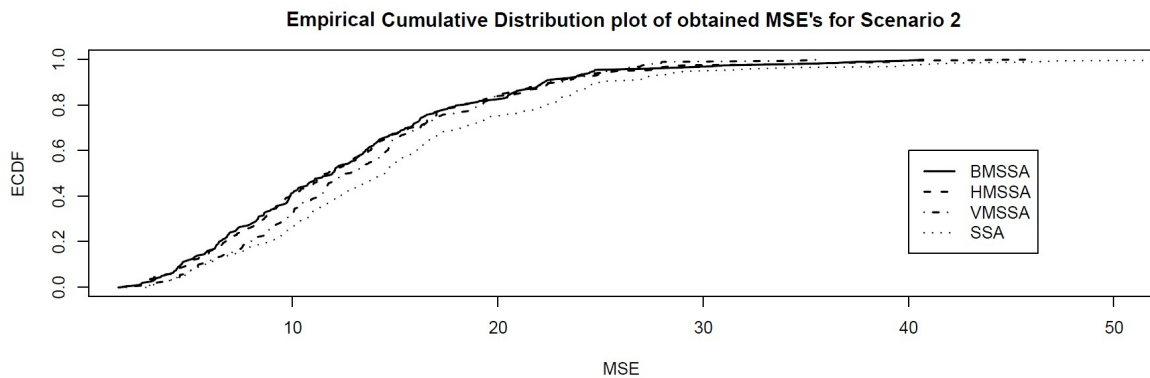


FIGURE 30: This figure represents the Empirical Cumulative Distribution Function plots of the 200 MSE values obtained from the four prediction methods.

With this high noise scenario, we find that the VMSSA methodology now actually competes with the other 2 multivariate approaches. According to Figure 30, each of the multivariate techniques improves on the univariate prediction. However, differences seem slight. To evaluate these differences, we calculate

the average MSE value for each prediction methods and test whether there are significant differences between the BMSSA approach and any other approach based on these average MSE values, using the Student's T test statistic, as tabulated in Table 8.

<i>Method</i>	<i>Average MSE</i>	<i>T-statistic</i>	<i>p-value</i>
SSA	15.6117	3.3275	0.0010*
HMSSA	13.0498	0.2135	0.8311
VMSSA	13.5289	0.9172	0.3596
BMSSA	12.8900		

TABLE 8: This table represents the average MSE values for the four prediction methodologies. These MSE values are obtained from 200 simulations. Student's t-test statistics and p-values are also calculated to test whether there are significant differences between the BMSSA predictions and that of the other prediction methods.

Table 8 now shows us that in this high noise scenario, the Bayesian approach produces the best average MSE value. This average value is significantly lower than that of the univariate predictions, but does not seem to be significantly different from the other multivariate approaches. Even though the differences in the average MSE values are not significant, considering the proportion of times the proposed Bayesian methodology outperformed its competitors (as given by Table 9) we find significant differences.

	<i>Proportion</i>	<i>p-value</i>
BMSSA outperformed HMSSA	55%	0.1573
BMSSA outperformed VMSSA	59%	0.0109*
BMSSA outperformed SSA	79%	<0.001*

TABLE 9: This table gives the proportion of the 200 simulations where the Bayesian approach produced a lower average MSE than the other prediction methods. The significance of these differences are then also tested.

According to this table, we see that the proposed Bayesian algorithm outperforms its competitors consistently. These p-values show that the Bayesian approach should therefore be the preferred approach since it improves on other multivariate techniques, even though it might not reduce the MSE by significant amounts. Notice that even though the multivariate approaches do not differ by significant amounts (in terms of average MSE) all three multivariate approaches provide significantly reduced average MSE values in comparison with the univariate SSA predictions.

Scenario 3:

In Scenarios 1 and 2 we had a simulation scenario where both the primary and auxiliary time series had exactly the same structure, both in terms of signal and noise. This is of course a very improbable scenario in practice. Two time series might have similar components, but will rarely represent exactly the same signal with exactly the same noise structure. This similarity in structure benefits both the HMSSA and BMSSA methodologies even though one approach might be benefited more than the other in certain cases. For the HMSSA methodology, by adding the auxiliary information into the trajectory matrix we are simply adding more observations whereby we can more accurately find the inherent signal space. For the BMSSA methodology, since both primary and auxiliary time series have the same structure, we can assume that the auxiliary time series can give us prior information on the primary time series and we

can therefore apply the Bayesian algorithm to the two time series. The VMSSA procedure on the other hand is yet to provide convincing results.

To deviate from time series with identical structures, we consider a situation where our primary and auxiliary time series have different noise components. In order to study the effect of this on the predictions as calculated using these methodologies, we simulate primary and auxiliary time series according to Equation 51.

$$\begin{aligned}
 y_{1,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{1,t} \\
 y_{2,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{2,t} \\
 \varepsilon_{1,t} &\sim \text{Uniform}(-10, 10) \\
 \varepsilon_{2,t} &\sim \text{Uniform}(-1, 1)
 \end{aligned}
 \quad \text{for } t = 1, \dots, 120
 \tag{51}$$

The first time series here has a much larger noise component than the second. Therefore, when predicting $Y_{1,t}$, including the information in the second time series should be of great help since both time series represent the same signal component, the first only more contaminated by noise. Inclusion of the second time series should allow us to more accurately capture the inherent signal component and improve on univariate predictions.

One can reconsider Figures 25 and 28 to inspect typical singular vectors obtained from the individual time series. Both the HMSSA approach and the BMSSA combine the two sets of singular vectors. The vectors representing signal space, as obtained by these multivariate methodologies, are given by Figure 31.

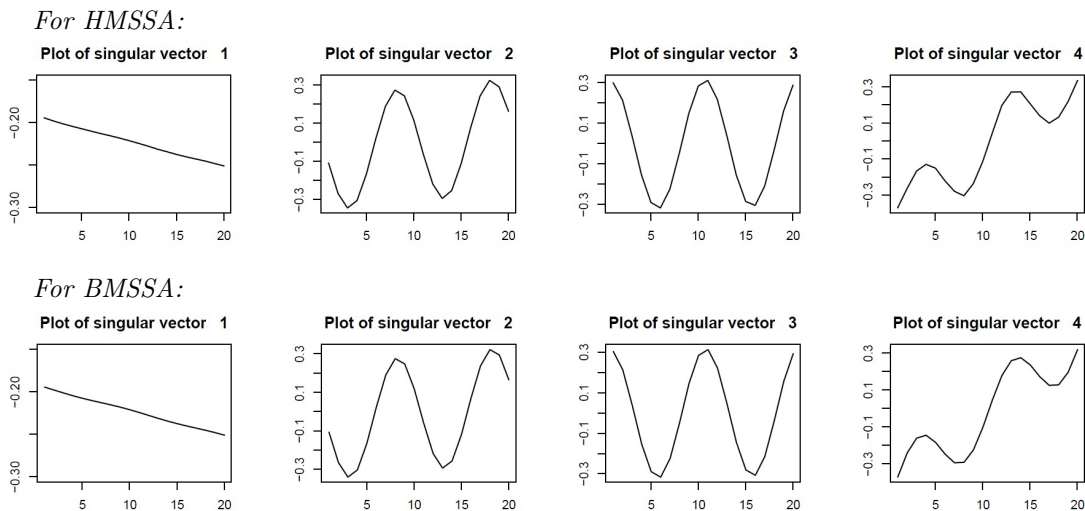


FIGURE 31: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the BMSSA approach (top) and the HMSSA approach (bottom).

The two sets of singular vectors as obtained by these two multivariate approaches once again look quite similar. However, when it comes to prediction, there are slight differences. 200 of these time series were then once again simulated and predictions were made for the final 20 observations in each replicate. The MSE of the predictions in each time series was calculated. The plot of the ECDF of these MSE values for each of the four prediction methods, drawn in Figure 32, allows us to inspect the differences between the different prediction methods.

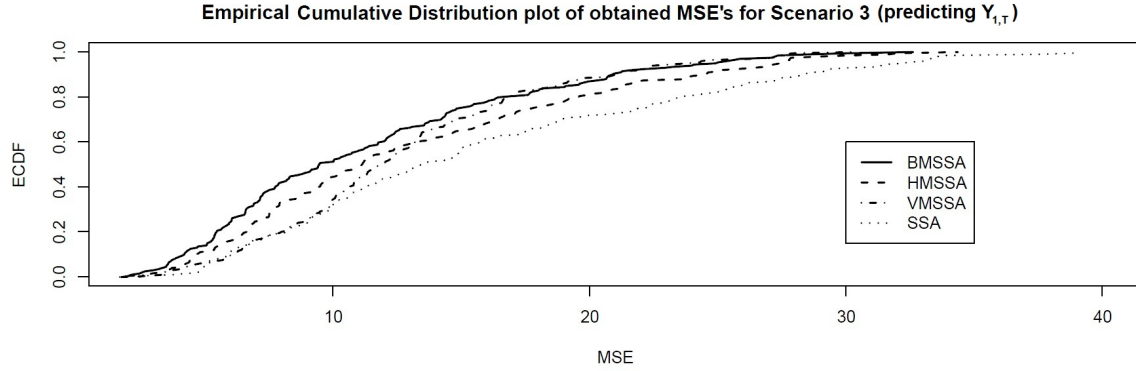


FIGURE 32: This figure represents the Empirical Cumulative Distribution Function plots of the 200 MSE values obtained from the four possible prediction methods.

This figure clearly distinguishes between the four prediction methods. All three multivariate techniques benefited from inclusion of the low noise auxiliary information. According to the ECDF, it seems that the Bayesian approach produced the smallest average MSE. Testing for statistical significance between the Bayesian approach and the other three methodologies can be done by calculating the average value for the 200 MSE values as done in Table 10.

Method	Average MSE	T-statistic	p-value
SSA	15.5808	5.6194	<0.0001 *
HMSSA	12.9555	2.4054	0.0166 *
VMSSA	12.6754	2.2624	0.0242 *
BMSSA	11.2704		

TABLE 10: This table represents the average MSE values for the four prediction methodologies. These MSE values are obtained from 200 simulations. Student's t-test statistics and p-values are also calculated to test whether there are significant differences between the BMSSA predictions and that of the other prediction methods.

Once again, we see that the proposed algorithm produces the lowest average MSE value. Surprisingly, the VMSSA procedure produces the second best average MSE. An intuitive explanation for this can be researched, but the inconsistency of the VMSSA makes it an unpredictable prediction methodology and we did therefore not seek such an explanation. According to the T-statistic, the BMSSA average MSE is the lowest and significantly so. Calculating also the proportion of simulation replicates where the Bayesian methodology produces more accurate predictions, we find Table 11.

	Proportion	p-value
BMSSA outperformed HMSSA	89%	<0.001 *
BMSSA outperformed VMSSA	57%	0.0477 *
BMSSA outperformed SSA	92.5%	<0.001 *

TABLE 11: This table gives the proportion of the 200 simulations where the Bayesian approach produced a lower average MSE than the other prediction methods. The significance of these differences are then also tested.

According to the above table, we find that the Bayesian approach consistently produces predicted values with less error than the univariate approach as well as both the other multivariate techniques. It also seems evident that only the VMSSA predictions seem to compete with the BMSSA predictions. Even though the HMSSA multivariate approach improves on univariate predictions, both the VMSSA and BMSSA approaches seem superior.

In an alternative situation, we could be interested in the prediction of the second time series. In that case, including the noise contaminated information in $Y_{1,t}$ in the prediction of $Y_{2,t}$ could perhaps be destructive to prediction accuracy. We could therefore have that the multivariate methodologies are even less accurate than the univariate predictions. However, since both time series still contain information on the same inherent signal, we still have reason to believe that the auxiliary time series may contain predictive information on the primary time series and can therefore apply the proposed Bayesian algorithm. Repeating this simulation 200 times and finding predictions for the final 20 observations using singular vectors as obtained through each of the prediction techniques, we can compare the accuracy of the predictions for our 4 different methodologies. Empirical Cumulative Distribution Functions for the 200 obtained MSE values are given in Figure 33.

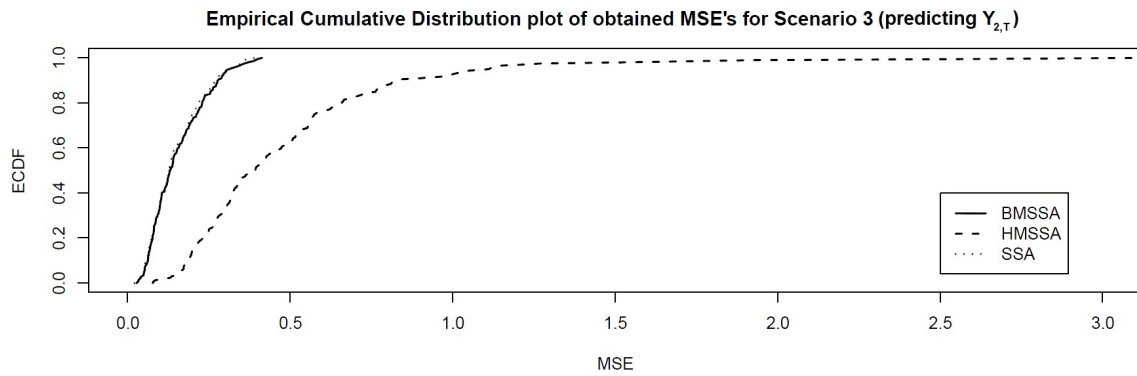


FIGURE 33: This figure represents the Empirical Cumulative Distribution Function plots of the 200 MSE values obtained from three of the possible prediction methods.

The first clear benefit of the Bayesian approach is evident in this figure. The HMSSA multivariate technique produces predictions that are clearly inadequate and inaccurate, while, once again, the VMSSA methodology is not even included in the figure. Therefore, according to these multivariate techniques, including the noise contaminated auxiliary information is actually destructive to prediction accuracy. The Bayesian approach however is relatively unaffected. The reason for this can be explained when considering calculation of the singular vector with the Bayesian approach (see Equation 48). With this formula, we can clearly see that the variation in the singular vectors obtained from both time series is incorporated into calculation of the final singular vector. Therefore, since the variation in the singular vectors obtained from the auxiliary time series is large, said singular vector carries less weight in estimation of the signal space and the estimated signal space is therefore less affected by this variation.

According to these plots, we can clearly see that the Bayesian approach is significantly better than other multivariate techniques. The question remains whether the proposed approach improves on the average MSE of univariate predictions. Calculating the average MSE value for the 200 obtained time series gives Table 12.

<i>Method</i>	<i>Average MSSE</i>	<i>T-statistic</i>	<i>p-value</i>
SSA	0.1472	-0.5804	0.562
HMSSA	0.4862	12.1053	<0.0001 *
VMSSA	9.2335	41.5289	<0.0001 *
BMSSA	0.1521		

TABLE 12: This table represents the average MSE values for the four prediction methodologies. These MSE values are obtained from 200 simulations. Student's *t*-test statistics and *p*-values are also calculated to test whether there are significant differences between the BMSSA predictions and that of the other prediction methods.

As we mentioned, we can clearly see that the Bayesian approach is superior to the other multivariate techniques. However, according to this table, the BMSSA produces a higher average MSE than the SSA approach. Even though this is the case, the *p*-value suggests that this difference between the univariate methodology and the Bayesian approach is very small and not significant. Continuing the comparison between SSA and BMSSA in this scenario, we found that the Bayesian approach only improved on univariate predictions 35% of the time. According to this, the Bayesian predictions are consistently less accurate than the univariate predictions, however the change in MSE is not significant. Other multivariate techniques on the other hand never improved on the univariate predictions with immense increases in average MSE values. The advantage of the Bayesian approach is therefore clear in this scenario.

Scenario 4:

In the previous 3 scenarios we have seen some of the strengths and weaknesses of the different prediction techniques when varying the noise component. In Scenario 3, when noise components were different, the Bayesian approach seemed to be superior to the HMSSA prediction algorithm in both cases. For Scenarios 1 and 2 on the other hand, the BMSSA approach was preferred in the presence of high noise and HMSSA predictions seemed superior in the presence of low amplitude noise. In order for us to study the effect of other possible changes, we need to find some compromise between high and low amplitude noise where the HMSSA and BMSSA predictions are equivalent. Thereby effects resulting from any changes to either the structure or the noise, are merely attributed to the changes and not to the amplitude of the noise component. Several different noise components were considered and we found that at an amplitude of 14, the two methodologies seemed equivalent. When considering the same signal component (as in previous scenarios) we found (over 200 replicates) that the proposed Bayesian methodology resulted in an average MSE of 6.07 while the HMSSA methodology resulted in an average MSE of 6.06. These were both improvements on the univariate predictions which had average MSE of 7.05, while the VMSSA methodology still produced entirely inaccurate predictions with average MSE of 9.66. The HMSSA and BMSSA procedures were also relatively similar in terms of the proportion of times the one methodology provided better results than the other since the Bayesian approach provided more accurate predictions than the HMSSA methodology 46% of the times. Clearly at this amplitude noise structure, no one method seems to be significantly more accurate than the other. We can then use this as a baseline from which the effects of other factors can be investigated.

Before we continue to the effects of time series with different structural components, we consider a final possible change in error component. This scenario attempts to answer the question: When can auxiliary information be helpful? Of course, it should be intuitive that if two time series explain some common signal component, the additional observations obtained from the auxiliary time series produce more accurate estimates of this component which in turn improves signal estimation and prediction. Ad-

ditional to this, some specific characteristics of the structure of the noise could be helpful. For example, would we prefer negatively correlated noise or positively correlated noise?

To investigate the effect that correlated noise components have on prediction accuracy of multivariate techniques, we consider two simulation studies. The first study is based on noise that is negatively correlated. In practice, it is possible (and sometimes even likely) that deviations from some expected signal component could be correlated. We can consider the effects of such correlations in the noise component by simulating primary and auxiliary time series according to Equation 52.

$$\begin{aligned}
 y_{1,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{1,t} \\
 y_{2,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{2,t} \\
 \varepsilon_{1,t} &\sim \text{Uniform}(-7,7) \\
 \varepsilon_{2,t} &\sim \text{Uniform}(-7,7) \\
 \text{cor}(\varepsilon_{1,t}, \varepsilon_{2,t}) &= -0.75
 \end{aligned}
 \quad \text{for } t = 1, \dots, 120
 \tag{52}$$

In this first simulation example of Scenario 4, we consider two time series with similar structure as in previous scenarios. The difference here however is the fact that the noise components included in the two time series are negatively correlated. In a practical sense, we would have that some factor would cause the one time series to increase while the second decreases. In order to determine whether this would be of benefit to our multivariate SSA methodologies we simulate 200 time series according to Equation 52. For each of these time series we predict the final 20 observations and determine the deviation from the true signal component. These 200 MSE values are then considered.

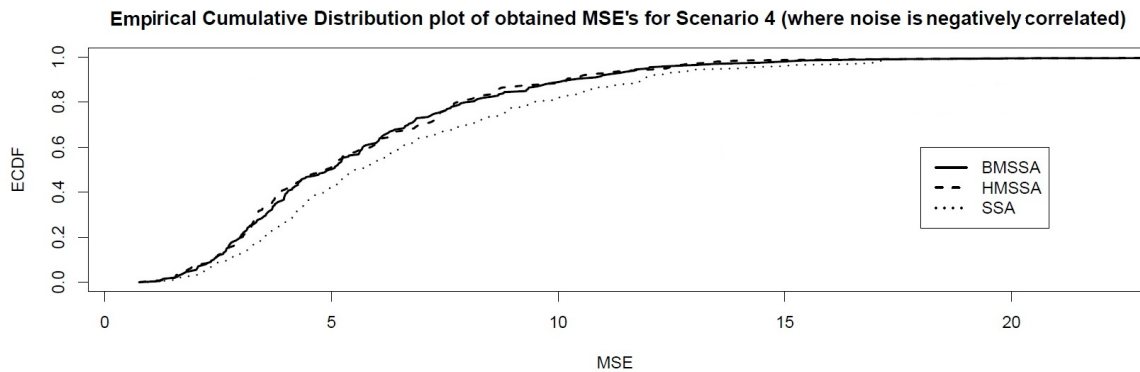


FIGURE 34: This figure represents the Empirical Cumulative Distribution Function plots of the 200 MSE values obtained from the three prediction methods under inspection.

The above figure illustrates the Empirical Cumulative Distribution Functions of the MSE values obtained from each of the prediction methods. Once again, when considering the MSE values, we notice that the VMSSA approach provides completely ineffective predictions and the ECDF of these MSE values is therefore not included in Figure 34. The HMSSA and BMSSA methodologies on the other hand seem to provide significant improvements on the univariate SSA predictions. However, there does not seem to be significant differences between the HMSSA and the BMSSA predictions. Table 13 confirms this by testing for statistically significant differences between the average MSE values of the 4 prediction methods.

<i>Method</i>	<i>Average MSE</i>	<i>T-statistic</i>	<i>p-value</i>
SSA	7.0200	2.3504	0.0193 *
HMSSA	6.0040	-0.3342	0.7384
VMSSA	10.0389	9.9346	<0.0001 *
BMSSA	6.1228		

TABLE 13: This table represents the average MSE values for the four prediction methodologies. These MSE values are obtained from 200 simulations. Student's *t*-test statistics and *p*-values are also calculated to test whether there are significant differences between the BMSSA predictions and that of the other prediction methods.

Clearly the only significant differences we can conclude from Table 13 are that the BMSSA methodology significantly improves on both the VMSSA predictions as well as the univariate SSA predictions. Upon further inspection of Table 13, we find that the test on significant differences between the HMSSA and BMSSA methodologies was inconclusive. We also notice that the Bayesian approach only outperformed the HMSSA approach 45% of the times, once again showing that the two methodologies provide similar results.

In comparison with our baseline measures of accuracy, we see that the correlation of the noise slightly increased the average MSE from 6.07 to 6.12, for the BMSSA approach and the average MSE decreased from 6.06 to 6 for the HMSSA approach. Clearly these differences are insignificant and we find (against intuition) that the additional structure in the noise could not be used to increase prediction accuracy. This was the case for both the BMSSA and the HMSSA prediction methods.

In a second situation, we consider the possibility of positively correlated noise. In a practical sense, some factor could cause unexpected increases in both time series. Our simulation example is therefore very similar to the previous scenario, but only with positively correlated noise. The two time series are simulated according to Equation 53.

$$\begin{aligned}
 y_{1,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{1,t} \\
 y_{2,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{2,t} \\
 \varepsilon_{1,t} &\sim \text{Uniform}(-7, 7) \\
 \varepsilon_{2,t} &\sim \text{Uniform}(-7, 7) \\
 \text{cor}(\varepsilon_{1,t}, \varepsilon_{2,t}) &= 0.75
 \end{aligned}
 \quad \text{for } t = 1, \dots, 120 \tag{53}$$

To assess whether this positive correlation benefits multivariate techniques, we repeat our simulation process whereby we simulate 200 time series according to Equation 53 and predict the final 20 observations in the time series. From the 200 simulations, we first plot the ECDF of the MSE values as given by Figure 35.

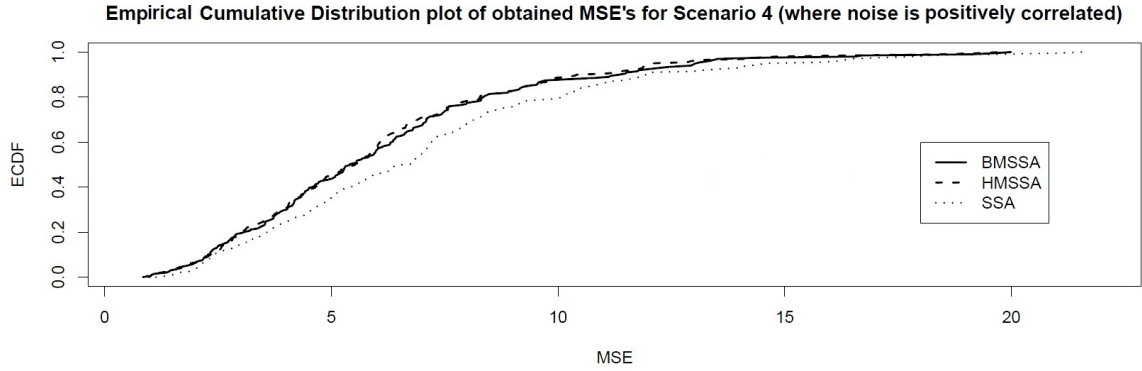


FIGURE 35: This figure represents the Empirical Cumulative Distribution Function plots of the 200 MSE values obtained from the three possible prediction methods.

Once again the VMSSA method produces completely ineffective predictions with MSE values even less accurate than the univariate predictions. The other two multivariate methods on the other hand improve on the SSA predictions. Testing the significance of this differences, we can consider Table 14.

<i>Method</i>	<i>Average MSE</i>	<i>T-statistic</i>	<i>p-value</i>
SSA	7.0359	2.3698	0.0183 *
HMSSA	5.94315	-0.1903	0.8492
VMSSA	9.5522	8.1857	<0.0001 *
BMSSA	6.0185		

TABLE 14: This table represents the average MSE values for the four prediction methodologies. These MSE values are obtained from 200 simulations. Students' *t*-test statistics and *p*-values are also calculated to test whether there are significant differences between the BMSSA predictions and that of other prediction methods.

According to the average of the MSE values in Table 14, it seems that the HMSSA methodology is slightly more accurate (but not significantly so) than the BMSSA methodology. In terms of the proportion of times that the BMSSA approach produces predictions with lower MSE, we find that this happens 48% of the time. The two prediction methods are therefore very close in terms of prediction accuracy. Furthermore, it seems that neither of these approaches were able to effectively use the additional noise structure to improve significantly on previous baseline predictions. The important thing however is that both these multivariate predictions are still more accurate than the univariate SSA predictions, showing that inclusion of the secondary time series was helpful to prediction. This is once again not the case for the VMSSA approach.

In an attempt to come to some conclusion as to whether these multivariate techniques can effectively use positively or negatively correlated noise, we see that the average MSE values for the multivariate predictions do not change by significant amounts. If anything, we can conclude that the HMSSA methodology was slightly better than the BMSSA approach, but it is most likely just coincidence and not because of the correlation structure.

4.3.2 Conclusion on the effect of noise component

During this segment of our simulation study, we considered the effects that different noise components have on prediction accuracy. We saw that by varying the magnitude of the noise component, not one single approach was uniformly superior in all situations. In high noise scenarios our proposed Bayesian approach was preferred, in low noise scenarios on the other hand the HMSSA methodology seemed to provide superior results. Even though these differences did not always seem significant when considering the reduction in MSE, the preferred methodology did consistently provide more accurate predictions in most cases. It might be unsatisfactory that the Bayesian approach did not produce the most accurate predictions in the low amplitude noise scenario, however it should be kept in mind that one seldom finds such a low noise to signal ratio in a time series in practice. Furthermore, one could speculate that this difference could be a cause of the limited number of bootstrap replicates (100) taken in the BMSSA approach. By increasing this number of replicates, we could perhaps increase the accuracy of the BMSSA approach, however computational time would increase significantly.

Considering time series with different noise components, we saw that the Bayesian approach proved to be superior to other multivariate methods both when the auxiliary time series had the smaller noise component and when it had the larger noise component. This can be attributed to the manner in which the Bayesian approach calculates the singular vectors representing the signal space. Where the HMSSA methodology simply considers additional information to be additional observations in a trajectory matrix, the BMSSA approach assigns a weight to the auxiliary information according to the variation in said auxiliary information. Therefore, when including information with large variation, less weight is assigned to the auxiliary information. On the other hand, when including information with little variation, the approach assigns more weight to the additional information. A very important result here was the fact that the Bayesian approach was the only multivariate approach that was at all able to extract valuable information with predictive power (albeit little) from an auxiliary time series with larger variation.

After a compromise between high variance and low variance noise was found where the BMSSA and HMSSA provide similar results, we were able to consider the effect a correlated noise component has on the prediction accuracy of our two multivariate prediction methodologies. Even when we had very high correlation (both positive and negative) between the noise components in our two time series, we did not seem to find statistically significant differences in the prediction accuracy with this additional structure in our noise component. One would expect that these two methodologies would be better equipped to effectively use this additional structure to their advantage. It seems however that the multivariate SSA techniques prefer data that shares a common signal component rather than data with correlated noise.

Regarding the VMSSA multivariate approach, we found that this approach was completely unstable and never really produced near optimum results. This is of course because of the fact that we are introducing extra features to the prediction algorithm that is explaining the same signal. Correlations between the features are therefore prominent, thereby unnecessarily increasing the complexity of the prediction algorithm. There are scenarios where this approach provides accurate predictions, but the lack of consistency in the performance of the algorithm is very unappealing. We will therefore not consider the success of this algorithm in any more detail during this thesis.

In this discussion on the effect the noise component has on prediction accuracy, we varied the magnitude of the noise component and we added additional structure to the noise time series in terms of correlation. We did not however change the distribution of the noise component. The previous 4 sce-

narios were all based on noise components with uniform distribution. The reason for this, was primarily to ensure that the Bayesian approach (and its assumption of normality) is not benefited by Gaussian noise. When similar simulations were performed with Gaussian noise components the Bayesian approach benefited and the results showed, more convincingly, that the Bayesian approach was superior. For future simulations, in order to expose the proposed methodology to unfavourable circumstances, we will continue with uniformly distributed noise components. Keep in mind however that (the more likely) Gaussian noise component produced similar results throughout, only more in favour of the proposed algorithm.

4.3.3 Effect of similar signal component

During the first segment of our simulation study, we came to the conclusion that the noise in a time series influences the accuracy of our multivariate prediction techniques. However, we found that the primary reason for the success of these techniques is the common signal component present in both time series. In this segment, we will study this concept in more detail. With two time series identical in structure, the multivariate techniques were very successful. To inspect the success of the multivariate approaches when the two time series had similar structure rather than identical structure, we consider the fifth scenario.

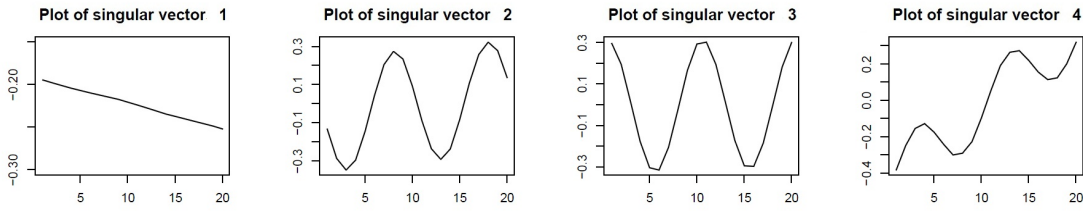
Scenario 5:

In this scenario, we commence our study on time series with similar structure. What we imply with similar structure is when two time series have identical signal components, only differing in magnitude. For instance both time series might have the same harmonic component and linear component, but one time series might have harmonic component with greater magnitude and linear component with less steep gradient. We therefore consider two time series simulated according to Equation 54.

$$\begin{aligned}
 y_{1,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{1,t} \\
 y_{2,t} &= 2t + 5\sin(2\pi t/10) + \varepsilon_{2,t} \\
 \varepsilon_{1,t} &\sim \text{Uniform}(-7, 7) \\
 \varepsilon_{2,t} &\sim \text{Uniform}(-7, 7)
 \end{aligned}
 \quad \text{for } t = 1, \dots, 120
 \tag{54}$$

Notice that by changing the amplitude of a sine wave or the gradient of the linear component, we are not effectively changing the structure of the signal as a whole, we are merely changing both the signal to noise ratio and the significance of the individual signal components. Consider for example the plot of singular vectors as given by Figure 36.

For the primary time series:



For the auxiliary time series:

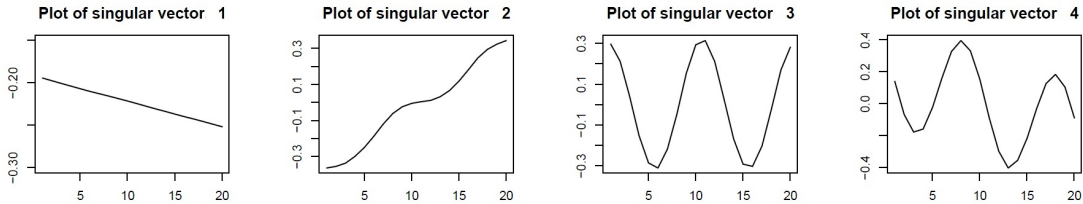


FIGURE 36: These figures represent plots of the estimated singular vectors associated with the signal space of the primary time series (top) and auxiliary time series (bottom).

We can clearly see that there are still little differences in the first and third singular vectors for the two time series. The second and fourth singular vectors however differ quite significantly. For both time series, we have that the the second and fourth singular vectors combined represent part linear signal component and part harmonic signal component. However, in the first time series the harmonic component is more significant because of the higher amplitude and the second singular vector therefore includes more of a harmonic component rather than a linear trend. The exact opposite holds for the second time series. The more prominent linear component dominates the harmonic component of smaller amplitude and is therefore included to a larger extent in the second singular vector. If we were to subject these two time series to a HMSSA procedure, the trajectory matrix would consider the signal space of the two individual time series as a whole. Extracting the four significant singular vectors from this trajectory matrix would result in Figure 37.

For HMSSA:

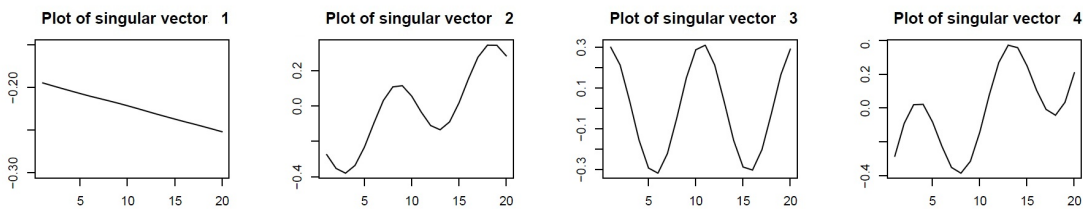


FIGURE 37: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the HMSSA approach.

These four singular vectors are clearly still describing the same signal component with linear trend, constant and harmonic component. The only difference is now the fact that the significance of the individual components have changed thereby changing the extent to which each singular vector is describing some specific component.

Unlike Scenarios 1 to 4, clearly the singular vectors are significantly different and we now doubt the assumption required for the Bayesian methodology. Is it a valid assumption to use the singular values

obtained in the second time series as prior information for those obtained in the first time series in the Bayesian context? In the HMSSA context, the two individual signal spaces are combined in their entirety. For the BMSSA methodology on the other hand, the individual components are independently combined. If we were to blindly accept the assumption, by inspection of Figure 36, we would attempt to combine the first and third singular vectors of the second time series with that of the first time series. However, according to the proposed Bayesian algorithm, we would prefer to combine the second vector in time series one with the fourth in the auxiliary time series and the fourth signal vector in the primary time series with the second in the auxiliary time series. This combination groups the individual vectors in such a manner that they are most similar to each other, enabling us to combine the two singular vectors in an independent fashion. Figure 38 shows what would happen if we were to do so.

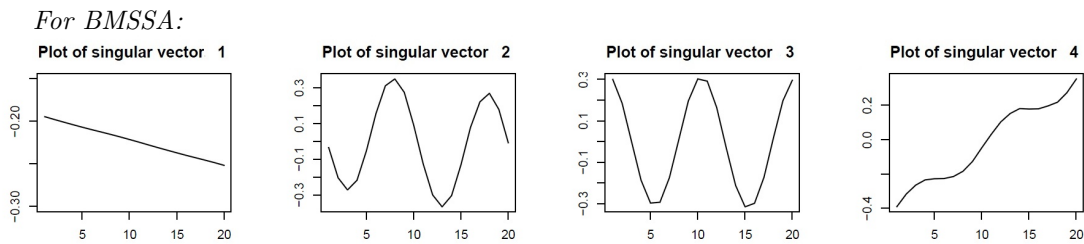


FIGURE 38: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the BMSSA approach. Here we choose to include the auxiliary information regardless of the hypothesis test testing the validity of the inclusion.

The first three singular vectors seem quite similar to that obtained by the univariate SSA algorithm. However, the final singular vector seems to have lost some of the harmonic component because of the inclusion of the additional information in the auxiliary time series. This is a result of the manner in which these singular vectors are combined. Referring back to Equation 48, we realise that the inclusion is weighted based on the covariance matrix of the included singular vector. Since the signal vector obtained from the auxiliary time series has less variation, this component is weighted more and the harmonic component obtained from the primary time series is weighted less. By inspection of Figure 38, we therefore have reason to be suspicious of this prior information. To test the validity of this suspicion, we simulate 200 time series according to Equation 54 and predict the final 20 values by rolling one-step-ahead predictions of the two multivariate techniques as well as the univariate approach. The MSE for these prediction methods produces Empirical Cumulative Distribution Functions as given by Figure 39.

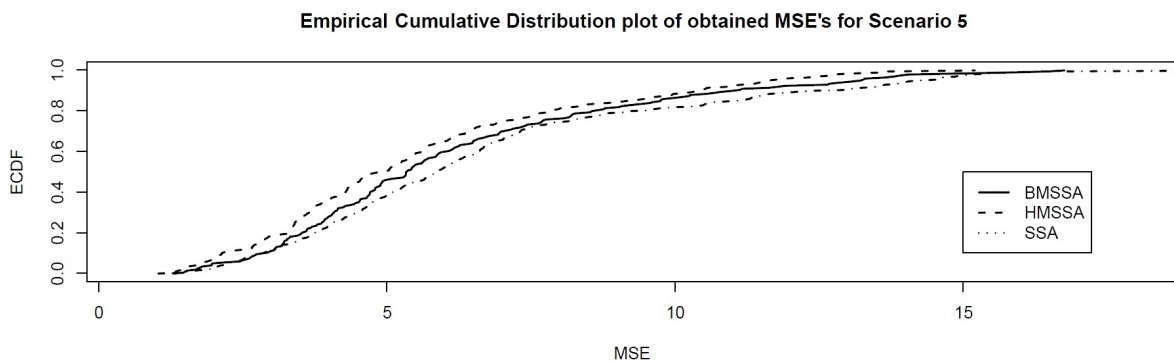


FIGURE 39: This figure represents the Empirical Cumulative Distribution Function plots of the 200 MSE values obtained from the three prediction methods under inspection.

Clearly we see here that there is a hierarchy of prediction accuracy. The HMSSA predictions improve on the BMSSA predictions which in turn improve on the SSA predictions. To evaluate the significance of these differences, we consider Table 15.

<i>Method</i>	<i>Average MSE</i>	<i>T-statistic</i>	<i>p-value</i>
SSA	6.5743	0.9659	0.3348
HMSSA	5.5582	-1.7941	0.0738
BMSSA	6.2004		

TABLE 15: This table represents the average MSE values for the three prediction methodologies. These MSE values are obtained from 200 simulations. Student's *t*-test statistics and *p*-values are also calculated to test whether there are significant differences between the BMSSA predictions and that of the other prediction methods.

These average MSE values show that both the BMSSA and the HMSSA improve on the univariate SSA predictions. Including the auxiliary time series was therefore helpful in terms of prediction accuracy. However, the HMSSA methodology yielded predictions with average MSE (nearly significantly) lower than that of the BMSSA methodology. The HMSSA algorithm therefore uses the auxiliary information more effectively than the proposed Bayesian algorithm.

This inadequacy of the Bayesian approach needs to be addressed. The reason for this flaw is a result of our assumption that the obtained components can be used as prior information. This assumption is not valid in this scenario. Suppose we had, for each singular vector, a hypothesis test to test the validity of using the vector obtained from the auxiliary time series as prior information for the signal component of our primary time series. This could be formulated as in Equation 55.

$$\begin{aligned} H_0 : E\{\underline{\theta}_i\} &= \underline{u}_i^* \\ H_a : E\{\underline{\theta}_i\} &\neq \underline{u}_i^* \end{aligned} \quad \text{for some } i \quad (55)$$

This hypothesis is a simple test on the expected value of a multivariate vector and addresses the validity of using the vector obtained from the Bayesian approach to describe the signal associated with the primary time series. The bootstrap sample obtained for the singular vectors of the primary time series can be used to calculate the test statistic for this hypothesis test (as given by Equation 56) and to test the hypothesis.

$$\begin{aligned} \text{Hotelling } T^2 &= (\bar{\underline{u}}_{1,i} - \underline{u}_{1,i}^*)' W_{1,i} (\bar{\underline{u}}_{1,i} - \underline{u}_{1,i}^*) \\ \text{With } \frac{2B - L - 1}{2LB - 2L} (\text{Hotelling } T^2) &\sim F(df_1 = L, df_2 = 2B - 1) \end{aligned} \quad (56)$$

In this equation, $\bar{\underline{u}}_i$ is the sample mean of the singular vector replicates of the first time series and $W_{1,i}$ is the sample covariance matrix of these singular vector replicates. Both of these are defined by Equation 46. Furthermore, B is the number of bootstrap replicates used in the Bayesian algorithm and L is the chosen window length.

Based on this calculated test statistic, one can then reject or accept the hypothesis at some level of significance α , by comparing the calculated *p*-value of the above calculated test statistic with the chosen α . Notice that this hypothesis test is equivalent to creating an L -dimensional confidence ellipsoid around

$\underline{u}_{1,i}$ and accepting the prior information if the obtained $\underline{u}_{1,i}^*$ is included in this ellipsoid.

The level of significance at which this test must be performed can vary based on the situation. We propose testing the hypothesis at 5% level of significance. However, this parameter can be altered by the user, thereby incorporating his own certainty as to whether the auxiliary information should be included or not from the practical context.

If the hypothesis is not rejected, we can use $\underline{u}_{1,i}^*$ as an estimate of the true signal singular vector, θ_i . If on the other hand, the hypothesis test is rejected, we use $\underline{u}_{1,i}$ as obtained from the univariate SSA. This process is followed for each $i = 1 \dots, d$ and after all singular vectors have been calculated individually, we once again have a set of vectors describing the signal space of our time series. Since these vectors are not necessarily orthonormal and since prediction is based on orthonormal vectors, these calculated singular vectors must once again undergo a Gram-Schmidt procedure to ensure that they can be used for both signal estimation and prediction.

Reconsidering Figure 36, we said that singular vectors 1 and 3 should be able to incorporate the information obtained from the auxiliary time series. Testing the hypothesis according to the above discussed theory, we find p-values for the hypothesis test as tabulated in Table 16.

Singular vector	1	2	3	4
p-value	0.5221	< 0.001	0.4537	0.0769

TABLE 16: This table gives the p-values associated with testing whether inclusion of auxiliary information is valid.

These p-values confirm what we suspected. That is that the difference between the second pair of singular vectors and the fourth pair of singular vectors are too significant for us to believe that these auxiliary signal vectors can be used as auxiliary information. If we were to omit the second and fourth signal vector pairs and were only to use the first and third vector pairs, we would then obtain the set of singular vectors given by Figure 40.

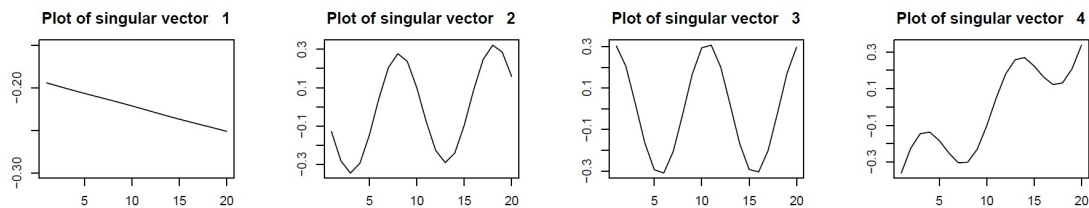


FIGURE 40: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the BMSSA approach. Here we choose to include the auxiliary information only if the hypothesis test, testing the validity of the inclusion, delivers a p-value larger than 0.05.

These signal vectors are quite similar to those in Figure 36. Only slight changes were therefore made to the description of the signal since we only included certain components of the auxiliary time series. If we were to perform this hypothesis test before each prediction of the same 200 time series simulated before, we would never include the second and fourth signal vectors from the auxiliary time series while including the first singular vector 88% of the time and the second singular vector 41% of the simulation

replicates. Including the MSE values, of the predictions obtained using this selectively inclusive Bayesian approach, in Figure 39 produces the updated Figure 41.

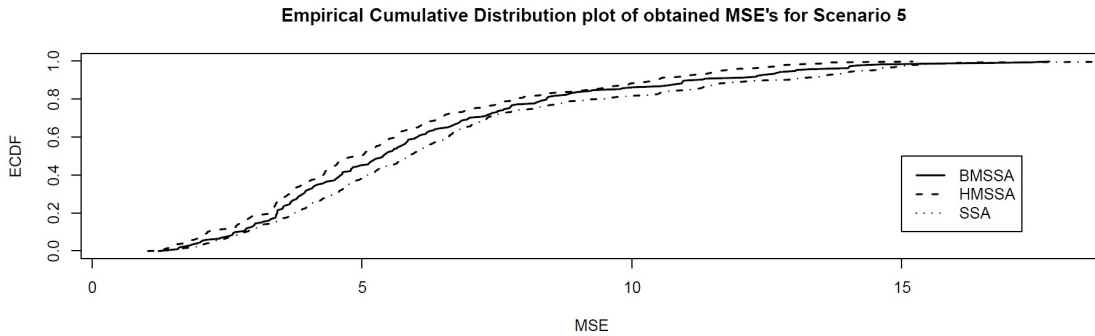


FIGURE 41: This figure represents the Empirical Cumulative Distribution Function plots of the 200 MSE values obtained from the three prediction methods under inspection.

Selectively including singular vectors as described above reduces the average MSE to 6.1039. This reduction might not be significant here, but the concept will prove helpful in future simulation scenarios. Even though the HMSSA procedure still produces more accurate predictions than this selective Bayesian approach, it seems that these hypothesis tests improve the prediction accuracy of the Bayesian approach. The hypothesis tests therefore prevent us from unnecessarily including auxiliary information when it might be destructive to prediction accuracy, which will be very helpful in later studies. This enables us to include components of an auxiliary time series instead of all of the information included in the auxiliary time series.

4.3.4 Conclusions on the effect of similar signal component

Even though we only considered a single scenario where time series have different but similar signal components, it is clear to see how the results can be generalized. As we increase or decrease the emphasis on the individual components, it is reflected in the singular vector components. In certain cases, this then in turn causes the individual signal vectors of the two time series to represent different signal spaces, even though the set of signal vectors as a whole still span the same space. Since the Bayesian approach is based on including the vectors individually and independently of each other and the HMSSA approach considers the signal space in its entirety, the Bayesian approach can not incorporate the auxiliary information while the HMSSA approach can.

The main problem here is therefore the fact that the SSA analysis is not able to effectively distinguish between the individual signal components. Because this is the case, the signal vectors cannot be included individually. Recent work on SSA (Golyandina, 2012) proposed using Independent Component Analysis (ICA) to attempt to better separate the individual signal components. By using an ICA approach, we can first find the signal space in its entirety and then describe it in terms of vectors that are optimum in the sense that they maximise some measure of independence. This might solve the inadequacy of the Bayesian approach since it would allow us to include the auxiliary information in the form of the individual signal components with more ease. However, since the Bayesian approach is already a computationally expensive approach, the possibility of using the independent principal component analysis was not included in this thesis.

An important aspect studied in this simulation study was the discussion of the selective including process. In previous scenarios, we simply assumed that the assumption on which the Bayesian approach is based was valid. After this simulation study, we have a simple hypothesis test to test the validity of the assumption. By using the hypothesis tests to test the validity of using the singular vectors obtained through the Bayesian approach, we have several advantages. Primarily, the fact that we are able to statistically determine whether some information could possibly provide additional predictive power is in itself a great advantage. When a prediction is made, we can determine whether the additional information was included or not. In a practical scenario, if the additional time series cannot be used for inclusion, we might prefer to simply continue looking for auxiliary information that can be used. If no such information is found, the algorithm automatically simplifies to the univariate SSA predictions which makes it highly unlikely that we will find predictions that are less accurate than that of the SSA algorithm. Furthermore, with this approach we can include components of a time series rather than the entire time series. We will see in later studies that this is an enormous advantage. Another advantage of this process is the fact that the user can incorporate his or her own knowledge on the validity of including certain components by choosing the α parameter appropriately. It is however important to notice that the introduction of this selective including process had no effect on the previous simulation study scenarios. This is true since with these previous scenarios, this hypothesis test always produced p-values larger than 0.9 and it was therefore clear that the vectors obtained from the Bayesian approach could and should be used for prediction.

From this simulation study we might have reason to believe that there are unresolved inadequacies of the BMSSA algorithm. Clearly, the HMSSA algorithm can (in some instances) use auxiliary information more effectively than the Bayesian approach. Of course no prediction algorithm is uniformly optimum and it is therefore necessary to know when to use certain approaches and when not to. However, by including the hypothesis tests we have the possibility of either including auxiliary information with predictive power or returning to the default univariate predictions, making it unlikely to include information that could be destructive to prediction accuracy. The Bayesian approach can therefore be regarded as a safer prediction method than the HMSSA approach.

4.3.5 Effect of different signal components

In the previous scenario we considered the consequences of having auxiliary time series with signal components that are similar yet not identical to that of the primary time series. Effectively we had that the two sets of singular vectors obtained from the two time series were different when compared individually, but they still spanned the same signal space when considered as a set rather than individually. We found that the HMSSA approach could more effectively use the auxiliary information in situations such as these. The Bayesian approach was still able to obtain some useful information from the auxiliary information, but had to do so selectively and some information with predictive power was therefore lost. Even though this is an inadequacy of the Bayesian approach, the option to selectively include different components of a time series is an appealing one. This segment of the simulation study focuses on using this option to increase prediction accuracy significantly.

On occasion, the fact that the HMSSA algorithm considers the signal space in its entirety is advantageous for prediction accuracy. This we saw in Scenario 5. However, it can also be a disadvantage. Suppose the signal spaces of the two time series have common components, but the time series are not similar. The HMSSA approach is famous for being inadequate in such scenarios (Golyandina, 2012).

One specific such case is the very specific type of structural difference where we have two time series with similar signal components but seasonal effects of different periodicity. The Bayesian approach here will not consider the signal space as a whole, but rather the individual components. Theoretically, the selective inclusion of the BMSSA approach should therefore be a point of importance here.

We will therefore be considering three different scenarios in this segment of the simulation study. Our primary time series will remain constant with some periodicity. We will use the same baseline primary time series we have been using in Scenarios 4 and 5. An auxiliary time series will also be simulated. This time series will however have a harmonic component of different periodicity. In comparison with the primary time series, the amplitude of the harmonic component in the secondary time series will either be smaller, larger or similar to that of the primary time series. These three different situations have different effects on the multivariate SSA techniques.

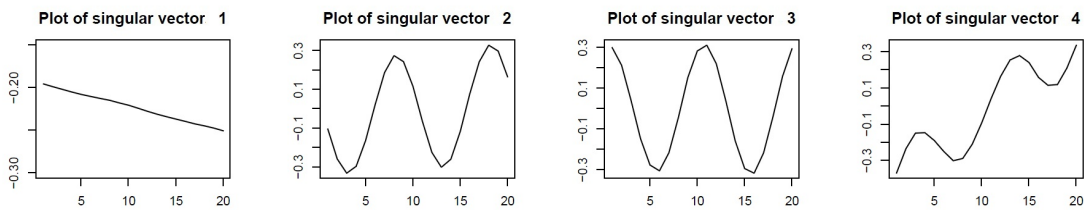
Scenario 6

Starting this simulation study on the effects of time series with different periodicities on prediction accuracy, we consider primary and secondary time series as simulated by Equation 57.

$$\begin{aligned}
 y_{1,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{1,t} \\
 y_{2,t} &= t + 10\sin(2\pi t/5) + \varepsilon_{2,t} \\
 \varepsilon_{1,t} &\sim \text{Uniform}(-7, 7) \\
 \varepsilon_{2,t} &\sim \text{Uniform}(-7, 7)
 \end{aligned}
 \quad \text{for } t = 1, \dots, 120
 \tag{57}$$

By simulating our time series according to this equation, we have that our secondary time series have the same linear component as the primary time series. However, the harmonic component in the auxiliary time series is different from that in the primary time series only in terms of periodicity. Individually these time series would typically produce singular vector plots as given by Figure 42.

For the primary time series:



For the auxiliary time series:

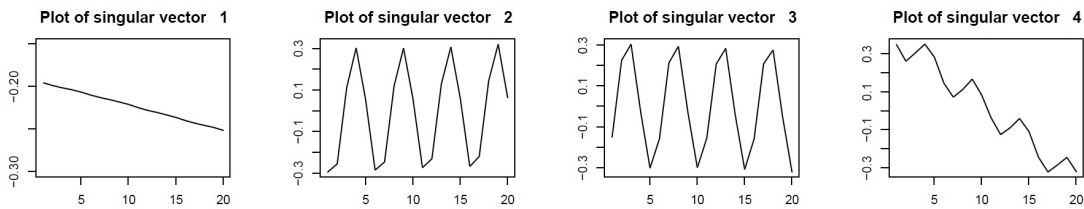


FIGURE 42: These figures represent plots of the estimated singular vectors associated with the signal space of the primary time series (top) and auxiliary time series (bottom).

The individual components seem similar to each other apart from the clear difference in periodicity. The first component is clearly associated with a constant. The second and third components contain information on the harmonic component. The final component is associated with some linear trend and the remainder of the harmonic component. In order to examine whether the HMSSA methodology can

use the auxiliary information to improve prediction accuracy on the primary time series, we first consider the singular vectors obtained from the HMSSA as given by Figure 43.

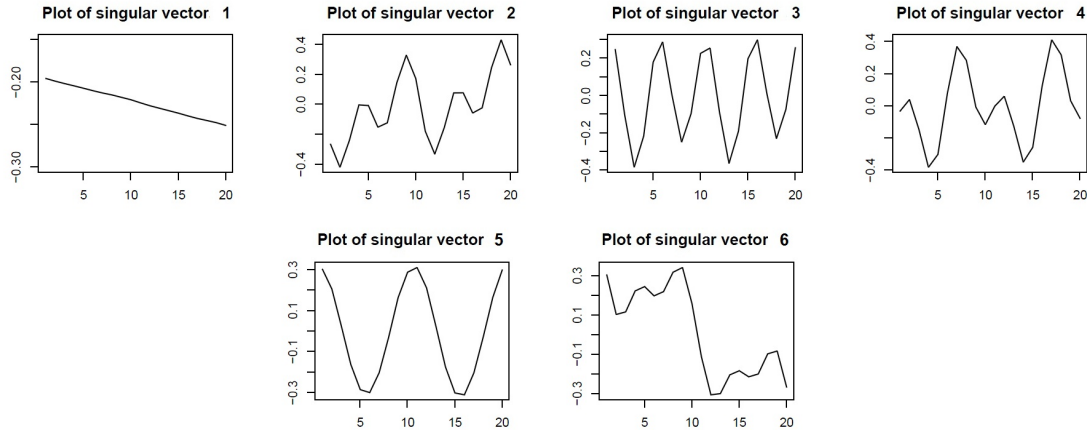


FIGURE 43: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the HMSSA approach.

Notice that the first six singular vectors are given here. This is since the trajectory matrix, as obtained during the HMSSA approach, has six structural components (based on Equation 57). Individually the primary and secondary time series both have vectors representing a constant and a linear trend. In the HMSSA trajectory matrix, these two components can be combined and represented by two vectors. The harmonic components included in the two individual time series on the other hand cannot be evaluated as a single harmonic component since they have different periodicities. Therefore, these two harmonic components need to be described by 4 signal vectors. This type of analysis on the structure component is only possible since the inherent signal of the time series is known and will not be possible in practice. However, since this simulation example is intended to study the differences between the HMSSA and BMSSA methodology when the dimension is correctly chosen, we choose the value for this parameter based on our knowledge of the signal component. In a practical scenario, this parameter would have to be chosen based on cross-validation methods.

From the 6 singular vectors illustrated in Figure 43, we can clearly see that the first vector represents some constant deviation. The remaining linear trend and harmonic components are represented by the remaining 5 signal vectors. Considering this set of vectors, the question is now how many of these vectors should be used for prediction. The individual time series are both of dimension 4. However if we were to combine the two time series there are 6 significant signal components, only four of them related to the primary time series.

For the Bayesian approach on the other hand, the choice of the dimension is unaffected. The auxiliary information is merely included component by component. The Bayesian approach however assumes that the singular vectors obtained from the auxiliary time series can be used as prior information to the primary time series. According to Figure 42, this is clearly not a viable assumption. If we were to merely accept this assumption, we would obtain singular vectors as given by Figure 44.

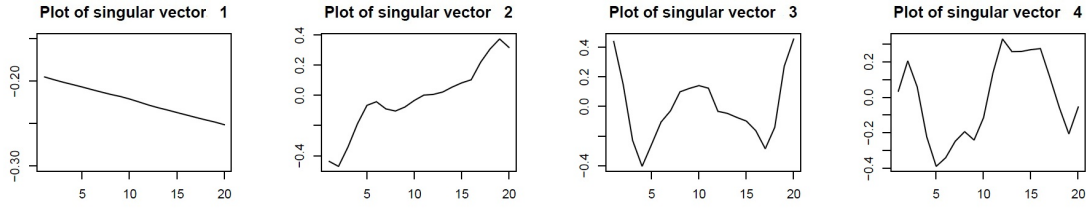


FIGURE 44: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the BMSSA approach. Here we choose to include the auxiliary information regardless of the hypothesis test testing the validity of the inclusion.

Clearly these estimates of the signal space are inaccurate since they do not seem to describe a linear and harmonic signal space; our assumption is therefore not feasible. However, if we were to perform the hypothesis test as discussed in the previous scenario, we could find p-values to test the validity of using these singular vectors for prediction. The p-values in this case indicate that only the first singular vector can be used (p-value of 0.1454). The p-values for the other vectors were all negligibly small if we were to compare them with the proposed α of 0.05 (notice of course that any other level of significance could also be used depending on the user). Rejecting the hypothesis that signal vectors two to four can be used and therefore only including the first singular vector, results in singular vectors as given by Figure 45.

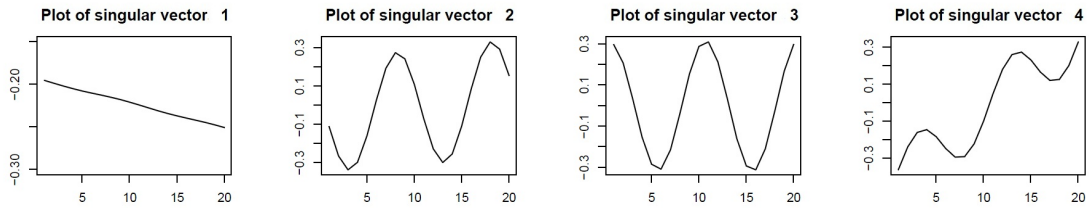


FIGURE 45: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the BMSSA approach. Here we choose to include the auxiliary information only if the hypothesis test, testing the validity of the inclusion, delivers a p-value larger than 0.05.

Since we are more interested in the prediction power of the different methodologies, we continue the simulation study by simulating 200 time series according to Equation 57. Rolling one-step-ahead predictions are calculated for the final 20 observations by using the singular vectors given by Figures 42 to 45. We are therefore interested in the predictive power of the Bayesian approach both when assuming that auxiliary information can be included and when inclusion only occurs when tested at 5% level of significance. Also, for the HMSSA approach, we consider the possibility of predicting using the first four singular vectors (since our primary time series is of dimension four) and using the first 6 singular vectors (since the dimension of the trajectory matrix is effectively 6). The MSE values for these multivariate prediction methods as well as for the univariate SSA predictions as calculated for these 200 replicates resulted in average MSE values as tabulated in Table 17.

<i>Method</i>	<i>MSE</i>
SSA	7.2997
HMSSA 4	102.8091
HMSSA 6	8.6460
BMSSA 0	56.5054
BMSSA 0.05	7.3216

TABLE 17: This table represents the average MSE values for five prediction methodologies. These MSE values are obtained from 200 simulations.

According to this table, we can clearly see that the Bayesian methodology should test the inclusion of the auxiliary time series in order to protect against inclusion of information that will be destructive to prediction, as we expected. The HMSSA methodology should also clearly use a dimension of 6 and not 4. In order to compare these two multivariate prediction methods as well as the univariate SSA prediction technique, we therefore construct the ECDF as given by Figure 46.

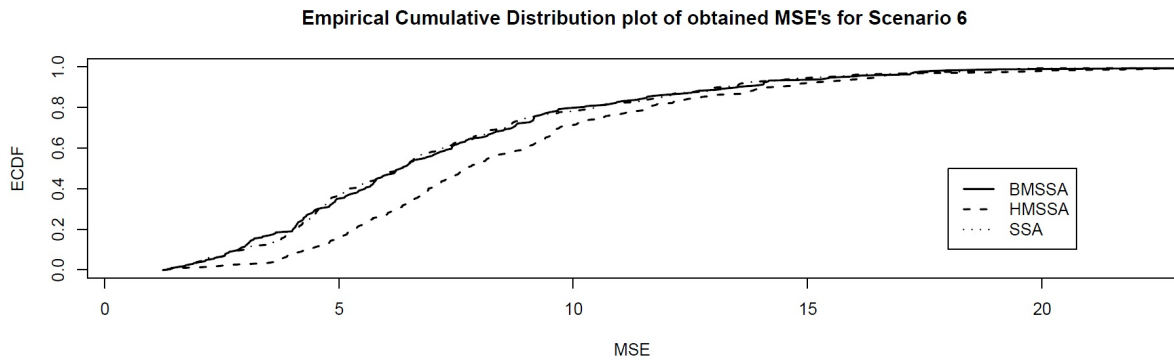


FIGURE 46: This figure represents the Empirical Cumulative Distribution Function plots of the 200 MSE values obtained from the three prediction methods under inspection.

Both the above figure, as well as Table 17 show that the HMSSA methodology provides us with less accurate predictions than both the univariate predictions as well as the predictions calculated from the Bayesian approach. The reason for this decrease in accuracy can be explained by considering Figure 43. Clearly we see that the sixth signal component still contains some linear trend. Prediction of the linear component would therefore be inaccurate when only considering the first four singular vectors (as we saw in Table 17). We are therefore forced to include all singular vectors up to the sixth signal vector, but by doing so we are also including the harmonic component that is not relevant to our primary time series, thereby effectively including unnecessary noise into our signal estimate. This reduces the accuracy of the filtering process and consequently also prediction accuracy. The difference between the Bayesian approach and the SSA approach on the other hand seem insignificant.

Therefore, according to these results, we see that the inclusion of the auxiliary information proved to significantly decrease the accuracy of prediction for the HMSSA methodology. However, the hypothesis test included in the Bayesian approach ensures that we do not include information that is destructive to prediction, resulting in predictions very similar to the univariate SSA predictions, which are the best for this scenario.

Scenario 7

Continuing this study on time series with different harmonic components, we simulate our next scenario. In this scenario, we consider the effects of including an auxiliary time series that has different periodic component, but one of less significance. We reduce the significance of this harmonic component by reducing the amplitude of said harmonic component. The simulation study here is based on two time series simulated according to Equation 58.

$$\begin{aligned}
 y_{1,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{1,t} \\
 y_{2,t} &= t + 5\sin(2\pi t/5) + \varepsilon_{2,t} \\
 \varepsilon_{1,t} &\sim \text{Uniform}(-7, 7) \\
 \varepsilon_{2,t} &\sim \text{Uniform}(-7, 7)
 \end{aligned}
 \quad \text{for } t = 1, \dots, 120
 \tag{58}$$

The rationale here is that if our auxiliary time series has a less significant periodic component, it might affect our signal estimation and prediction to the same extent it did when the signal components of the two time series were of equal magnitude. The difference in the significance is evident when considering the scatterplot of the singular vectors for the second time series.

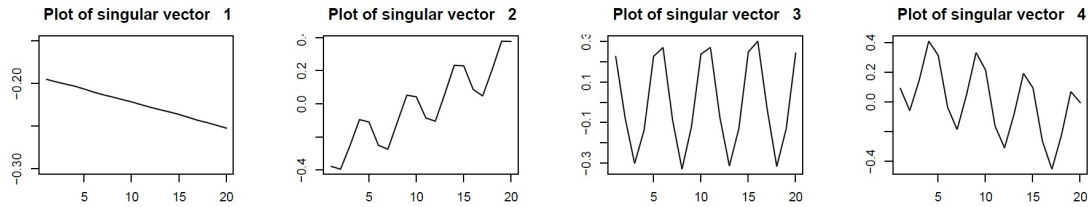


FIGURE 47: These figures represent plots of the estimated singular vectors associated with the signal space of auxiliary time series.

The singular vectors obtained for our first time series will be identical to those in Figure 42. However the second set of singular vectors changed slightly in comparison with Figure 42. Clearly, the signal space these singular vectors represent seem similar to that of Scenario 6. The change in the amplitude of the harmonic component merely caused the harmonic component to become less significant and therefore a larger part of the harmonic component is explained in the fourth and final singular vector, where previously the majority of the component was explained by the second and third vectors. To answer the question whether some of these characteristics might carry over to a HMSSA analysis, we first consider the singular vectors obtained from the HMSSA procedure for a typically simulated scenario.

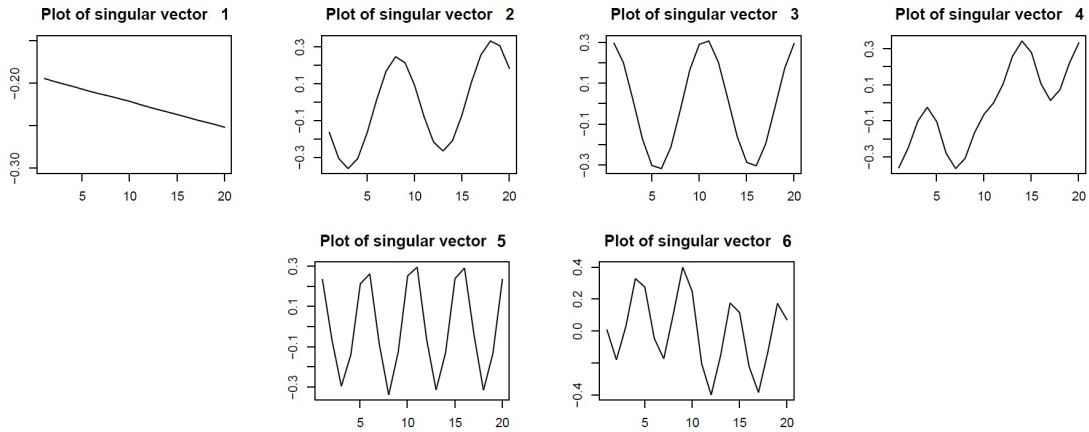


FIGURE 48: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the HMSSA approach.

In Figure 48 we consider the 6 signal vectors obtained from the HMSSA procedure. By reducing the significance of the periodic component in the auxiliary time series, we were able to find the first four singular vectors that are clearly associated with the primary time series. The fifth singular vector is clearly only associated with the auxiliary time series and the sixth singular vector represents a combination of the auxiliary time series' harmonic component as well as some other component (perhaps linear). Possibly, in cases where the amplitude of the auxiliary time series' harmonic component is smaller than that of the primary time series, we are able to isolate the singular vectors that should be associated with the auxiliary time series, thereby enabling us to predict the primary time series only using the first four singular vectors.

For the Bayesian approach, on the other hand, we once again first consider the singular vectors obtained if we were to blindly assume that the auxiliary singular vectors can be used as prior information for the vectors obtained from the primary time series. These vectors are illustrated in Figure 49.

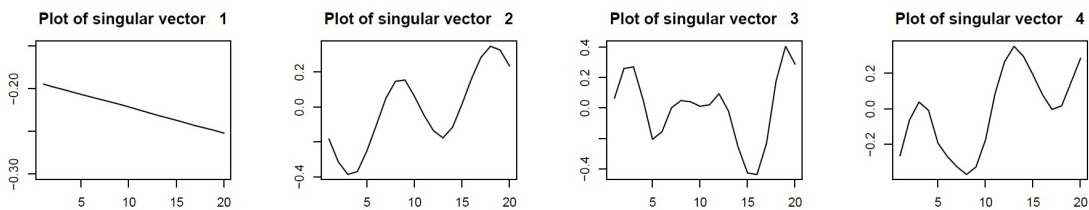


FIGURE 49: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the BMSSA approach. Here we choose to include the auxiliary information regardless of the hypothesis test testing the validity of the inclusion

We once again find singular vectors that do not seem to be accurate descriptions of our true signal space. Merely upon inspection of the individual singular vectors as obtained from the individual time series (Figure 47 and Figure 42), it would seem intuitive to believe that they describe different components and should not be combined in a Bayesian context. The result of such a combination clearly confirms the intuitive statement. If we were to test the validity of including the auxiliary information and using these Bayesian vectors, we would find that only the first singular vector obtained from the auxiliary time series should be included at a 5% level of significance (with a p-value of 0.0946). This would result in

singular vectors in Figure 50.

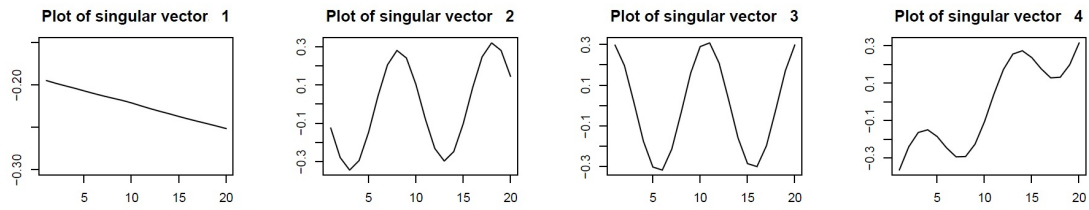


FIGURE 50: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the BMSSA approach. Here we choose to include the auxiliary information only if the hypothesis test, testing the validity of the inclusion, delivers a p -value larger than 0.05.

Clearly these singular vectors seem to represent a much more accurate estimate of our signal space. In fact they are nearly identical to the signal vectors obtained from the univariate SSA approach. However, the mere estimation of these singular vectors are not what we are interested in. We are interested in the predictive power of these singular vectors. Once again we simulate 200 time series according to Equation 58. The final 20 observations are predicted using the same SSA methodologies as in Scenario 6. The average values of the MSE's produced by these prediction methods are tabulated in Table 18.

<i>Method</i>	<i>MSE</i>
SSA	6.9373
HMSSA 4	6.9235
HMSSA 6	8.3043
BMSSA 0	42.1822
BMSSA 0.05	6.7006

TABLE 18: This table represents the average MSE values for five prediction methodologies. These MSE values are obtained from 200 simulations.

These average MSE values clearly show us that our proposed Bayesian estimate cannot simply assume that the auxiliary information is helpful. This assumption can be tested at 5% level of significance (notice once again, that this level of significance can be changed to adapt to the will and knowledge of the user). Interestingly, we see here that the HMSSA methodology now performs more accurately when using a 4 dimensional approach in comparison to the 6 dimensional approach that produced more accurate predictions in the previous scenario. In a more detailed comparison between the relevant predictions techniques, the ECDF of the MSE values are drawn in Figure 51.

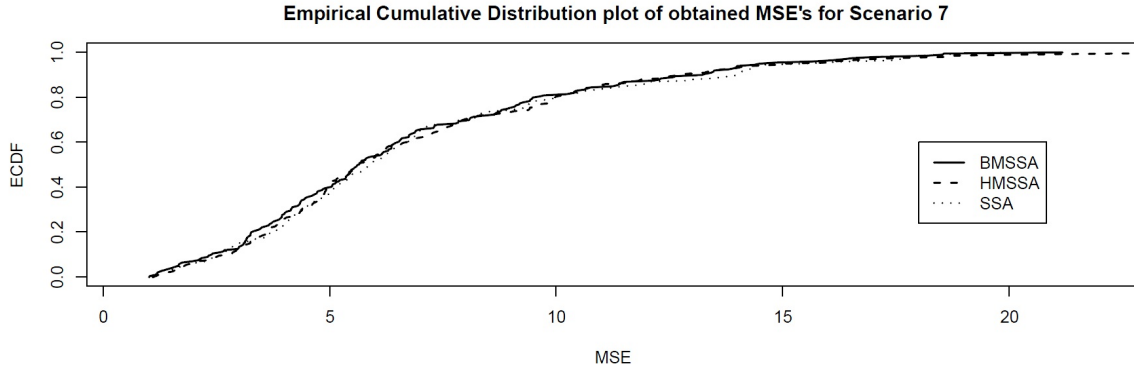


FIGURE 51: This figure represents the Empirical Cumulative Distribution Function plots of the 200 MSE values obtained from the three prediction methods under inspection.

Both the ECDF as well as the average MSE values in Table 18 show us that the HMSSA method produces predictions that are very similar to the univariate predictions. The Bayesian approach is the only multivariate technique that is able to actually extract valuable information from the auxiliary time series. Even though we found that the four dimensional HMSSA approach is more effective than the six dimensional approach, the predictions are still less accurate than those obtained by the Bayesian approach. The reason for this is that there are still signal components in the fifth and sixth singular vectors that are relevant to our primary time series. However, these singular vectors still describe the auxiliary time series to a much larger extent. Therefore by including these singular vectors we are including irrelevant information more than we are including signal relevant to our primary time series. This unnecessary inclusion decreases our prediction accuracy.

For the Bayesian approach on the other hand we see that not only were we able to disregard this irrelevant auxiliary information, but we were even able to extract some valuable information from the auxiliary time series, thereby slightly improving on the prediction accuracy of the univariate SSA predictions. We were able to do so since we did not include the auxiliary time series as a whole, but only individual components. We merely used the first singular vector (when it was allowed by the hypothesis test) associated with a constant and slight linear trend.

Scenario 8

In a penultimate scenario we consider here a secondary time series with a harmonic component of different periodicity, but this component is much more prominent than in our primary time series. In order to study such a scenario, we consider two time series simulated according to Equation 59.

$$\begin{aligned}
 y_{1,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{1,t} \\
 y_{2,t} &= t + 20\sin(2\pi t/5) + \varepsilon_{2,t} \\
 \varepsilon_{1,t} &\sim \text{Uniform}(-7, 7) \\
 \varepsilon_{2,t} &\sim \text{Uniform}(-7, 7)
 \end{aligned}
 \quad \text{for } t = 1, \dots, 120
 \tag{59}$$

Before continuing to prediction accuracy, we can already speculate what the effect of this inclusion would be for the HMSSA methodology. We will once again have to include all 6 singular vectors into our signal estimation and prediction process, because of the increase in the significance of the periodic component in the auxiliary time series. This additional emphasis on the harmonic component is evident when considering Figure 52.

For the auxiliary time series:

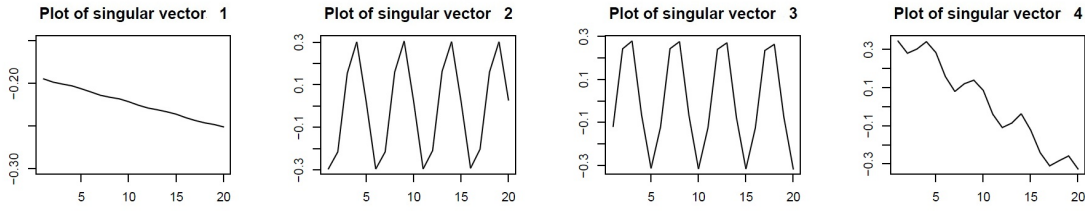


FIGURE 52: These figures represent plots of the estimated singular vectors associated with the signal space of the auxiliary time series.

Clearly the second and third singular vectors are solely associated with the harmonic component in the auxiliary time series and very little of the harmonic component is explained by the final signal vector. This significance of the harmonic component resonates in the singular vectors obtained through the HMSSA analysis, as we can clearly see in Figure 53.

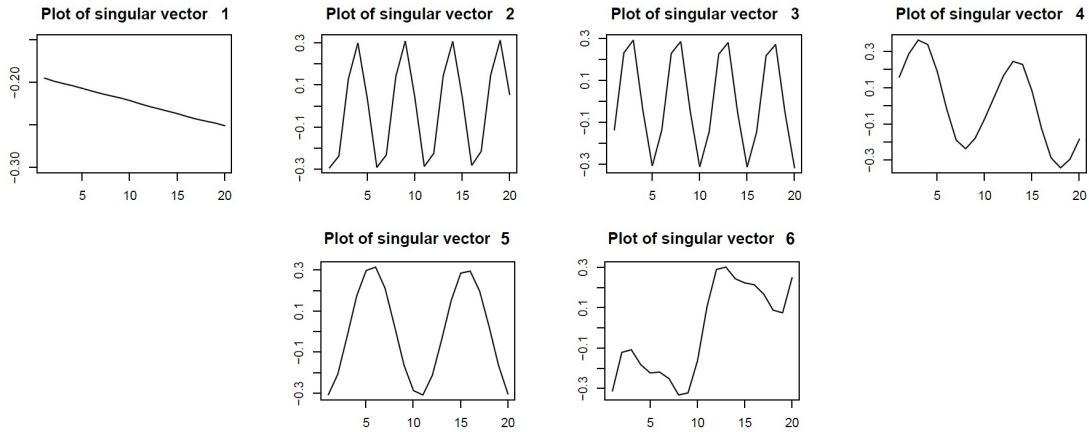


FIGURE 53: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the HMSSA approach.

According to Figure 53, the harmonic component present in the auxiliary time series is represented in the second and third singular vectors. It is also noticeable in the final vector. The components in the primary time series are present in singular vectors 1, 4, 5 and 6. Inclusion of all 6 singular vectors will therefore most likely be necessary in the HMSSA context. We were able to anticipate from previous simulation examples that this would be the effect on the HMSSA methodology. However, we are uncertain as to the effect this auxiliary information will have on the proposed Bayesian approach. If we were to blindly (without any testing) assume that the auxiliary time series delivers valuable prior information on all our singular vectors, we would obtain singular vectors as given by Figure 54.

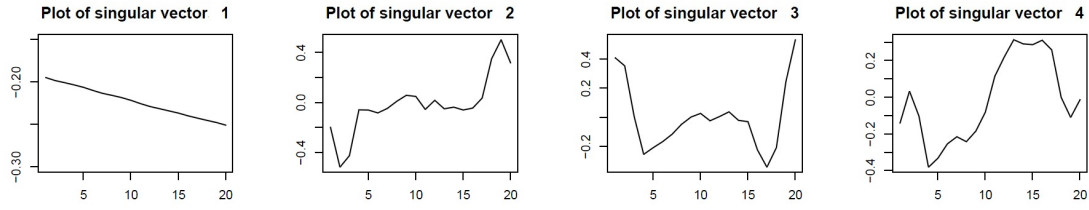


FIGURE 54: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the BMSSA approach. Here we choose to include the auxiliary information regardless of the hypothesis test testing the validity of the inclusion.

It is clear that these singular vectors are not accurate estimates of the signal space. Testing the validity of using these singular vectors results in a p-value of 0.18 for the first singular vector and negligibly small p-values for the remaining 3 components. Selectively including only the first singular vector as suggested by the obtained p-values and the proposed level of significance, the end result of our proposed Bayesian approach therefore produces singular vectors as given by Figure 55.

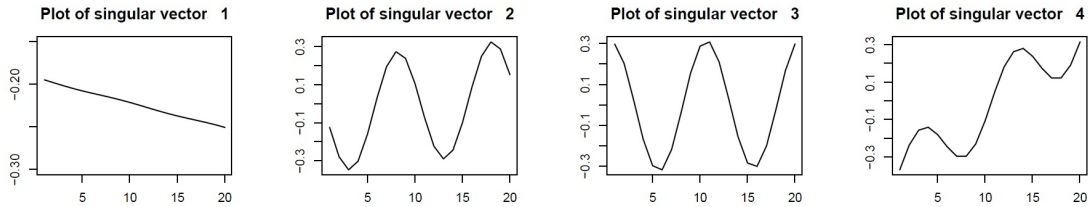


FIGURE 55: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the BMSSA approach. Here we choose to include the auxiliary information only if the hypothesis test, testing the validity of the inclusion, delivers a p-value larger than 0.05.

These singular vectors seem to be more accurate estimates of our signal space than those in Figure 54, since these clearly represent a linear, constant and harmonic signal component. To test whether this is indeed true, we consider 200 simulations of time series simulated according to Equation 59. Predicting the final 20 observations, according to each of the prediction methodologies we had in Scenarios 6 and 7, produces average MSE values as given in Table 19.

Method	MSE
SSA	7.0188
HMSSA 4	107.3739
HMSSA 6	8.2575
BMSSA 0	55.0828
BMSSA 0.05	7.0696

TABLE 19: This table represents the average MSE values for five prediction methodologies. These MSE values are obtained from 200 simulations.

As we anticipated, the hypothesis test included in the Bayesian approach is indeed necessary. Without it, predictions would be completely unstable. Also, the HMSSA methodology indeed produces more accurate predictions when including the first 6 signal vectors, resulting in average MSE comparable to

those in Scenario 6. Assessing the plot of the ECDF of the MSE values for the prediction methods would allow us to make more definite conclusions on the differences between the relevant prediction methods.

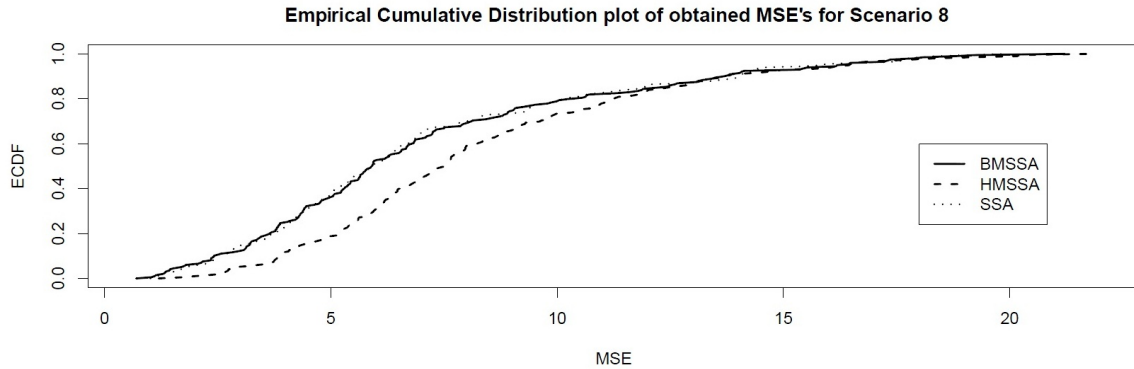


FIGURE 56: This figure represents the Empirical Cumulative Distribution Function plots of the 200 MSE values obtained from the three prediction methods under inspection.

Once again we see that the HMSSA methodology is not able to match either the univariate predictions or the BMSSA predictions for similar reasons as before. Upon inspection of Figure 56, one might consider basing HMSSA predictions on the first singular vector together with the fourth to sixth singular vectors, thereby omitting the second and third singular vector. This could possibly result in more accurate predictions, however still not in comparison to that of the Bayesian approach since the sixth HMSSA singular vector still contains some harmonic component of the auxiliary time series. Furthermore, this illustrates another disadvantage of the Horizontal Multivariate SSA approach; the fact that the choice of dimension then becomes less interpretable.

The Bayesian approach on the other hand still produces predictions that are comparable to those of the univariate SSA approach in terms of MSE. Table 19 shows that the Bayesian approach does have average MSE value higher than the SSA approach (though not significantly so). This could possibly be because of the fact that the harmonic component in the auxiliary time series is more dominant and the first vector might contain some slight unnoticeable harmonic component.

4.3.6 Conclusions on the effect of different signal components

It should become clear that the problem in scenarios such as these is the fact that the SVD step in the SSA analysis is unable to distinguish between individual signal components. For the HMSSA this forces us to include some signal components that are effectively only valid for our auxiliary time series. For the Bayesian approach this also has some slight effect as we saw in the final simulation. However, the Bayesian approach is much more robust against this inadequacy.

If we were therefore better able to separate the individual components more definitely, both multivariate techniques could benefit from it. As mentioned before, the possibility of using ICA methods in order to more effectively distinguish between the individual components is currently being studied. The use of Independent Components Analysis has been proposed and seems to increase separability in the univariate approach. However, research is yet to be done on the effect of this in multivariate approaches such as the HMSSA and BMSSA methodologies.

One problem remains for the HMSSA algorithm. Even if we were able to distinguish between the different signal components, the choice of dimension becomes significantly less interpretable. This was already evident in the previous simulation Scenarios 6 and 7; we saw that we needed to choose our dimension differently for the three scenarios. The final study (Scenario 8) then showed us that the optimum choice of the signal space would not necessarily be the first d singular vectors, but some non-intuitive combination of the singular vectors obtained from the HMSSA algorithm. Something similar to this final scenario is what we could expect if we were to use an Independent Component Analysis (ICA) instead of the SVD step in the decomposition step since the ICA will simply more effectively separate the signal components associated with the primary time series from those associated with the auxiliary time series; the first d singular vectors will not necessarily be associated with the primary time series. This problem of a non-intuitive optimum choice of dimension would of course be exacerbated as the complexity of the two time series increases, causing more confusion as to which singular vectors should be associated with which time series.

Based on these simulation scenarios, where the possibility of an ICA was not considered, we saw that the auxiliary information were only destructive to prediction accuracy in the HMSSA context. The Bayesian approach on the other hand was still able to extract relevant (albeit little) information with predictive power from the auxiliary time series component wise and if it was unable to do so, the Bayesian approach was much more robust against auxiliary time series with different signal components. If we therefore believe a specific singular vector in the decomposition of the auxiliary time series will be destructive to the prediction of the primary time series, we simply do not include this singular vector. The problem therefore changes to deciding when the inclusion of an singular vector is destructive to prediction and when it is not. The proposed hypothesis test is a possible solution to this problem, however the user is also able to incorporate his own predetermined knowledge by changing the level of significance at which the hypothesis test is to be performed.

We have already mentioned that the HMSSA prediction technique is known to be very ineffective when the two time series are of different periodicities (Golyandina, 2012). This was confirmed in this simulation scenario here. Under these unfavourable circumstances, the Bayesian approach was able to provide predictions better than or at the very least comparable to the univariate SSA predictions. The results found here can be generalized by considering any situation where our auxiliary and primary time series are spanned by entirely different signal components. In other words if the individual signal components are inherently different and cannot be combined. This does not necessarily have to be based on harmonic components with different periodicities, but it can also include polynomials of different degree. Or one time series with some signal component and the other without. Identical problems will arise and only the Bayesian approach will be able to withstand these problems. The incapability of the HMSSA algorithm to consider the components individually will cause unnecessary inclusion of some signal that is present in the auxiliary time series but not the primary time series. The Bayesian approach will be able to independently include or omit this component.

Notice here that when testing whether the singular vectors obtained from the SVD step were normally distributed, we chose to test the singular vectors independently. We mentioned that this is an unrealistic assumption. However, in this simulation study, we see that this assumption allows us to individually and independently incorporate signal components obtained from an auxiliary time series. Clearly, this assumption therefore not only decreases the required computational power, but it also allows us to consider our time series component-wise, which is a great advantage in cases such as the ones mentioned here.

4.3.7 Effect of scale differences

In the previous scenario we considered one of the very common situations where the Bayesian multivariate approach has a very significant advantage over the existing HMSSA approach; that is when the two time series are inherently different in terms of structural components. Another scenario where the Bayesian approach addresses inadequacies of the HMSSA methodology will be considered here.

Often we will have scenarios where we have two time series perhaps describing some identical signal (or signal components), but the two time series are on different scales. This discussion is included in this study since it is specifically relevant in the rest of this thesis. Stock market data in general is predominantly much larger in magnitude in comparison to the values we will be using, describing internet activity.

Scenario 9

In order to assess the effect that an auxiliary time series has on the prediction accuracy of our primary time series when the auxiliary time series is on a significantly smaller scale than the primary time series, we consider time series simulated by Equation 60.

$$\begin{aligned}
 y_{1,t} &= t + 10\sin(2\pi t/10) + \varepsilon_{1,t} \\
 y_{2,t} &= (t + 10\sin(2\pi t/10) + \varepsilon_{2,t})/5 \\
 \varepsilon_{1,t} &\sim \text{Uniform}(-7, 7) \\
 \varepsilon_{2,t} &\sim \text{Uniform}(-7, 7)
 \end{aligned}
 \quad \text{for } t = 1, \dots, 120
 \tag{60}$$

Notice the very specific manner in which we reduce the scale of the auxiliary time series. By simulating according to Equation 60, we ensure that the signal to noise ratio is maintained. The two time series are therefore identical, but merely on different scales. Since this is the case, there is of course no change in the singular vectors as calculated from each individual time series. Previous figures can be considered to evaluate these singular vectors. According to these, one can clearly conclude that the assumption of the Bayesian approach is a valid one and it is indeed confirmed by the hypothesis tests. Singular vectors as obtained from our two multivariate techniques are however included in this discussion.

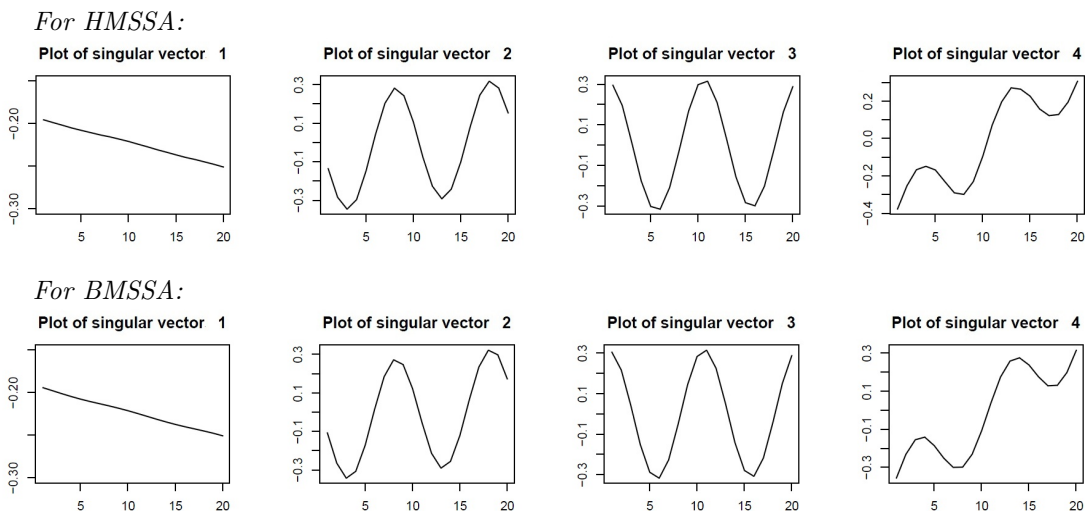


FIGURE 57: These figures are plots of the estimated singular vectors associated with the signal space of the primary time series calculated by combining both the primary and secondary time series using the HMSSA approach (top) and the BMSSA approach (bottom).

The difference between these two sets of singular vectors seems negligible. However, we must remember that slight differences in these singular vectors could result in significant differences in the corresponding predictions. We therefore repeat this simulation 200 times and calculate the MSE for each of the relevant multivariate prediction methods as well as the univariate approach. The MSE values for these prediction methods produce an ECDF as given in Figure 58.

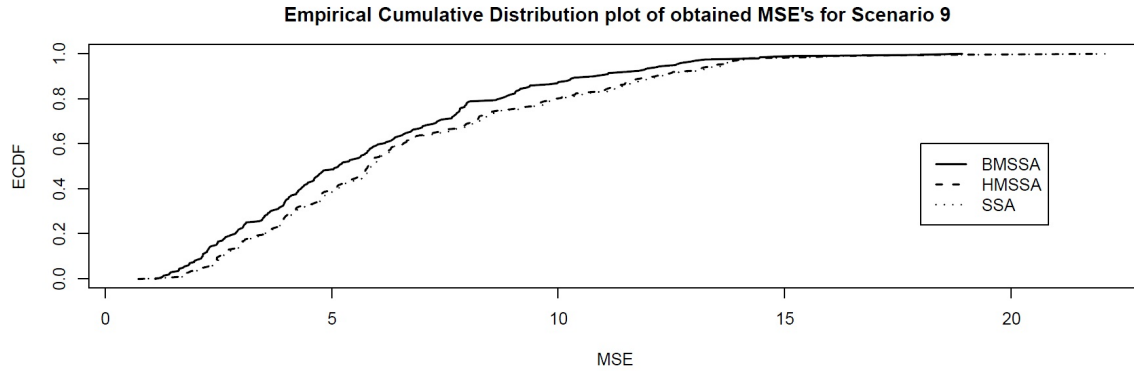


FIGURE 58: This figure represents the Empirical Cumulative Distribution Function plots of the 200 MSE values obtained from the three prediction methods under inspection.

Upon inspection of this figure, we find that both multivariate approaches are able to slightly improve on the univariate predictions. However, the Bayesian approach is able to improve considerably more than the HMSSA approach. To test the significance of the improvement, we consider Table 20.

Method	Average MSE	T-statistic	p-value
SSA	6.7307	2.3697	0.0183 *
HMSSA	6.6438	1.4706	0.0332 *
BMSSA	5.8778		

TABLE 20: This table represents the average MSE values for the three prediction methodologies. These MSE values are obtained from 200 simulations. Student's *t*-test statistics and *p*-values are also calculated to test whether there are significant differences between the BMSSA predictions and that of the other prediction methods.

Table 20 shows that the proposed Bayesian approach significantly improves on both the univariate SSA predictions as well as the HMSSA predictions in terms of average MSE. Considering the proportion of times the BMSSA improve on the other prediction methods, we find 77.5% improvements on HMSSA predictions and 78.5% improvements on SSA predictions. This clearly illustrates another inadequacy of the HMSSA methodology. Since the two time series are combined into a single trajectory matrix, the auxiliary time series does not significantly contribute to the variation in the trajectory matrix. The SVD is much more concerned with the larger values and therefore finds singular vectors predominantly based on the primary time series. The Bayesian approach on the other hand considers the two time series separately in order to find the corresponding signal space. Since the covariance matrices of the singular vectors then determines the weights assigned to each of the time series, we have an inherent scaling procedure included in the Bayesian approach not present in the HMSSA technique.

4.3.8 Conclusions on the effect of scale differences

In this final segment of the simulation study, we only included a study where the auxiliary time series is on a smaller scale in comparison with the primary times series. In the situation where an auxiliary time series is on a larger scale, we can anticipate results based on the explanations given above. Since the HMSSA method simply emphasizes the time series with observations the highest in magnitude, the predictions will mainly be based on the auxiliary time series and less so on the primary time series. However, because these two time series are indeed describing the same signal space with the same signal to noise ratio, this will not have a big effect on the HMSSA methodology but we might be able to improve on the accuracy of the univariate SSA predictions. The Bayesian approach will still weigh the individual signal components by their inherent variation and the weights will effectively scale the information accordingly and the Bayesian approach will therefore also be able to improve on the accuracy of the SSA predictions. Comparing the HMSSA and BMSSA predictions, I would suspect that they would provide similar results. Such a scenario could be included to confirm these assumptions, however since this is not directly relevant to the thesis it is omitted. It should be noted, however, that the scenario was briefly investigated and the conclusions discussed above seemed accurate.

According to this rationale, the only reason why the BMSSA procedure is benefited in such scenarios is the fact that the information is rescaled before the individual components are combined. This is not the case for the HMSSA technique. One can of course subject the auxiliary trajectory matrix to a scaling procedure before the combined trajectory matrix is constructed. The problem however then merely reduces to finding the optimum (or at the very least an appropriate) scaling procedure.

This segment therefore concludes by saying that the proposed Bayesian approach is superior to the HMSSA approach in scenarios where there is a significant difference in the scale of the two time series. There are manners in which one can elaborate on the HMSSA approach in order to solve its inadequacies, however such methods are inherently included in the Bayesian approach.

5 Incorporating search volumes

Now that our multivariate methods have been discussed thoroughly, we can apply these methods to a practical scenario in order to test our hypothesis that internet activity can improve prediction accuracy. Essentially, what we have in this practical scenario is a primary time series regarding some stock market activity and a secondary time series measuring some form of internet activity over time. In general, stock market activity is very volatile. Since our simulation studies showed that the Bayesian approach is preferred in scenarios where our primary time series is highly contaminated with noise (regardless of the amount of noise in the secondary time series) we believe that the Bayesian approach will deliver more accurate predictions in comparison with the HMSSA prediction method. We will also see in this specific practical example, that our secondary time series will be significantly lower in scale when compared to the primary time series. We have concluded from the simulation study that the HMSSA methodology is not helpful when we attempt to predict some time series in the presence of an auxiliary time series that is significantly lower in scale. The Bayesian approach on the other hand can overcome this scaling problem. We believe that these characteristics of the practical example at hand will benefit the BMSSA methodology and allow the BMSSA methodology to produce more accurate predictions than the HMSSA technique.

5.1 Predicting an automobiles index

The primary time series we will be considering in this chapter is the Dow Jones Automobiles and Parts Titans 30 Index (DJTATO). This index represents the value of leading companies in the global automobiles and parts sector. It is calculated by using a weighted combination of companies in the auto-mobile sector, such as Toyota, BMW, Bridgestone etc. Figure 59 graphically illustrates all available observations of this index over time (obtained from Yahoo Finance). We saw in Chapter 3 that by analysing the returns of some stock market time series resulted in more accurate predictions and also simpler parameter estimation in the context of the SSA methodology. Figure 59 therefore also shows the returns of the DJTATO index.

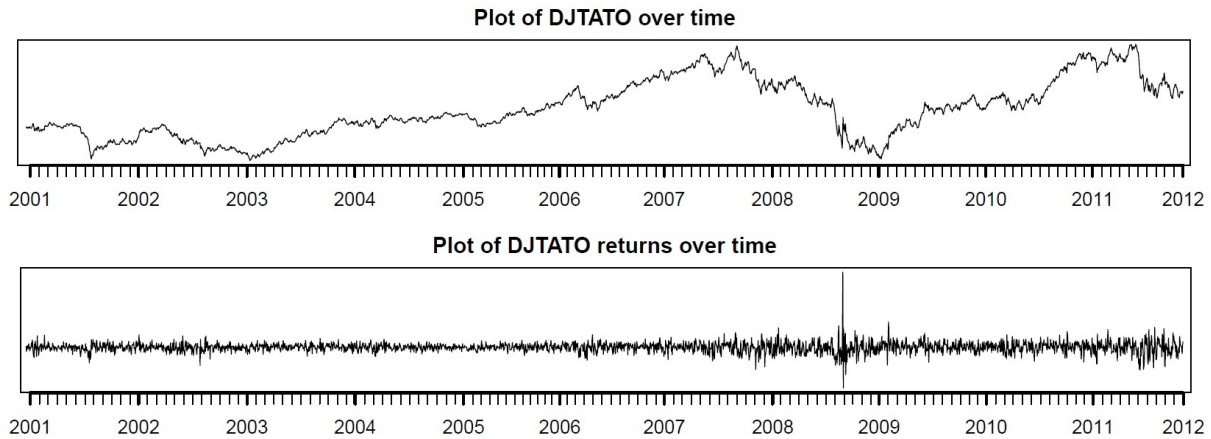


FIGURE 59: These figures illustrate the primary stock market time series to be predicted. The top figure contains the observed daily DJTATO index closing prices and the bottom figure contains the returns of the observed daily DJTATO index closing prices.

The first noticeable observation we can make from the primary DJTATO time series is the very significant decline apparent during 2008. This is of course a direct consequence of the recession that started late in the year 2007. Furthermore, both figures clearly illustrate the amount of noise present in the primary time series. When considering the return time series, we see that the amount of noise become even more evident after the recession during 2008. This characteristic makes it quite difficult for us to analyze (let alone predict) the primary time series.

The first necessary step, in order to predict the returns of this stock market index for the year 2011, is to once again divide this time series into a model training set, a cross-validation set and a test set. This will allow us to perform cross-validation methods on the time series, which will in turn allow us to choose the appropriate values for our parameters. The partition of the primary time series is identical to what we had in Chapter 3. The observations until the end of 2009 are used for model training. Based on this subset of the data, we continue our cross-validation process by calculating rolling one-step-ahead predictions for all 2010 returns, while varying our parameters across some domain. By doing so, we are investigating the efficiency of the chosen parameters and we can decide on the appropriate choice of the parameters (L and d) to be used in the prediction of the test partition.

In Chapter 3 we found that by choosing a dimension parameter other than $d = 1$ we would significantly decrease prediction accuracy. In this specific example, cross-validation produced similar results. Based on the predictions found for the cross-validation segment, while varying our window length parameter over values $L \in \{1, \dots, 35\}$ and our dimension parameters in the set $\{1, \dots, L - 1\}$ (as in Chapter 3),

we found that $d = 1$ produced the most accurate predictions, regardless of the choice of window length. Figure 60 therefore illustrates the prediction accuracy of the cross-validation predictions for different chosen window lengths when our choice of the dimension parameter is one.

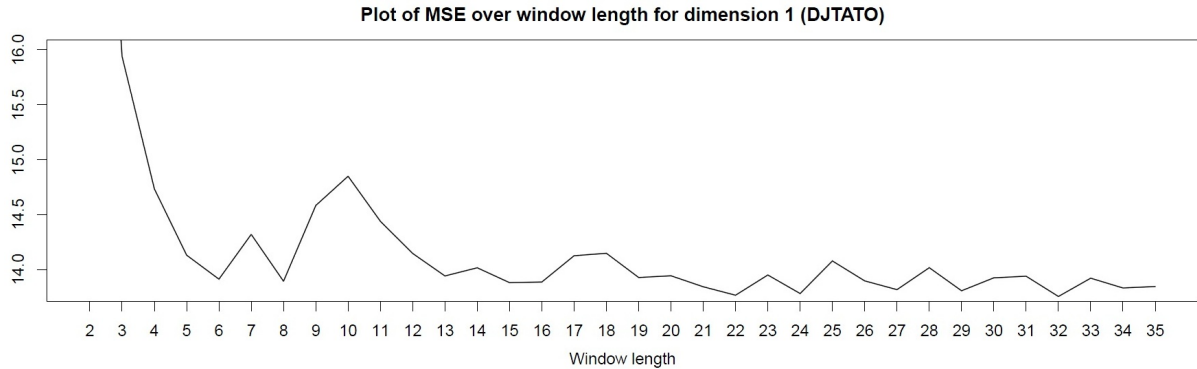


FIGURE 60: This figure illustrates the prediction accuracy (of the predictions of the DJTATO return time series) as window length is varied. Since it was clear that the cross-validation predictions indicated that d should be equal to one, this figure shows how the MSE (for the predictions on the cross-validation set) changes as the window length increases.

This figure once again shows that window lengths smaller than 20 are unstable and ineffective when using MSE to measure prediction accuracy. Window lengths larger than 35 could also have been considered, but to make predictions according to the SSA algorithm while using large window lengths, requires large amounts of computational power (especially in the BMSSA approach). However, we can clearly see that prediction accuracy stabilises around $L = 20$, and we do therefore not need to consider larger window lengths. The minimum MSE value is reached at window length of 22. This is also the window length that maximises the DOC measure, with DOC of 50.39%. Clearly, according to these measurements of prediction accuracy, we should use $L = 22$ and $d = 1$ to obtain predictions for the final test set (the DJTATO values observed during the year 2011).

Based on these values for our parameters, we can obtain baseline predictions, i.e. predictions that do not incorporate any additional information. These baseline predictions can then be compared to the predictions that do incorporate internet activity measurements. This comparison can answer the question proposed in this thesis: Can internet activity be used to improve the accuracy of prediction? The internet data we wish to incorporate can be extracted from numerous sources. In this specific chapter, we consider the possibility that data obtained from Google Domestic Trends (http://www.google.com/finance/domestic_trends) could perhaps have predictive power for the DJTATO time series. This web-based application is related to both Google Insights for Search and Google Finance. It measures Google search volumes in one of several economic sectors. There are however some inadequacies of the data obtained via this application. The first possible problem is the fact that the data given by Google Domestic Trends do not represent actual search volumes related to a given sector, it merely measures the amount of search queries relative to the activity on that day. Furthermore, this time series is also normalized according to other variables (such as the geographical size of a region) to eliminate the effect of such variables. After this normalization, the individual time series values are expressed as a value out of 100, truncated to two decimals. This truncation also causes us to lose some of the accuracy in these values. Even though this is the case, we continue our attempt to use this information in the prediction of the DJTATO index.

The Google Domestic Trends application has three categories relevant to the automotive industry, that is the "Auto Buyers Index" (ABI), the "Auto Financing Index" (AFI) and the "Automotive Index" (AI). Each of these three categories will be used in our practical scenario. The ABI tracks queries related to buying used cars and websites that sell cars. The AFI tracks queries related to certain financial aspects involved when purchasing a vehicle; this includes loans, vehicle payments etc. Finally the AI tracks search queries relating to specific motor companies such as Ford, Toyota etc. There are very slight differences in the categorization of these three indices, but Figure 61 shows that there are significant differences in the actual search volumes.

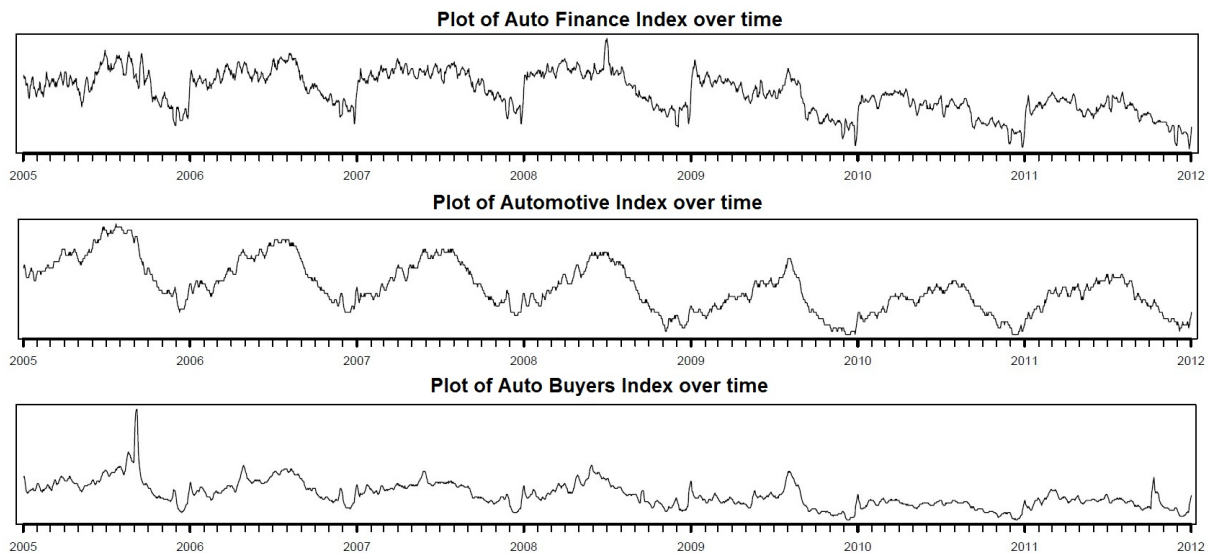


FIGURE 61: These three figures illustrate the auxiliary time series that we are attempting to include in our prediction of the DJTATO index returns. The top figure contains the observed daily search volumes related to the AFI category, the middle contains the observed daily search volumes related to AI category and the final contains that of the ABI category.

A significant obstacle in these time series is the fact that they clearly have seasonal effects. One could use SSA to eliminate this yearly harmonic component in each time series; however, the computational time involved when applying the SSA algorithm to over 2000 observations with minimum window length of 365 observations is quite impractical. If we were to calculate the partial autocorrelations at a lag of 365, we would see that there is valid reason to believe that there are significant correlations between observations at a lag of exactly one year. This gives us some reason to believe that seasonal differences might be an appropriate (and simple) way to remove seasonality. Additionally, a Dickey-Fuller augmented unit root test (at a lag of 365), on each of these time series at a 5% level of significance, shows that we can indeed take seasonal differences in order to crudely remove the seasonal effects present. Doing so produces time series as given by Figure 62.

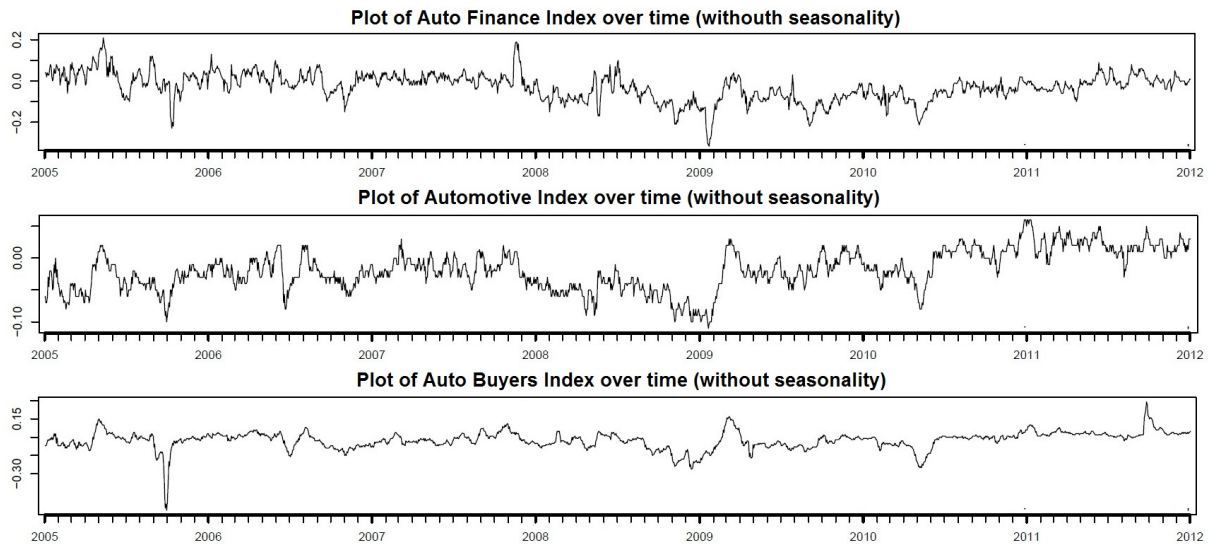


FIGURE 62: These three figures illustrate the auxiliary time series (with seasonality removed) that we are attempting to include in our prediction of the DJTATO index returns. The top figure contains the differenced daily search volumes related to the AFI category, the middle contains the differenced daily search volumes related to AI category and the final contains that of the ABI category.

There are numerous disadvantages involved when taking differences in order to remove seasonality. The most important of which is that it is not considered to be a very accurate way to remove seasonality. Apart from the fact that it can be seen as an inaccurate seasonal filtering method, another one of these inadequacies is of course that we are effectively losing 365 observations when taking differences at a lag of 365. Even though this is the case, we found that the remaining 6 years' observations were sufficient to produce results in favour of our hypothesis.

With seasonality removed (albeit crudely), our next step is to consider the differences between the inherent characteristics of the primary and auxiliary time series. The primary time series measures the changes in the DJTATO index. Therefore, in order to incorporate the auxiliary information, we must consider the changes in internet activity rather than the actual (seasonally filtered) search volumes. Consequently, we need to take differences in our secondary time series at a lag of one as well, resulting in secondary time series as given by Figure 63.

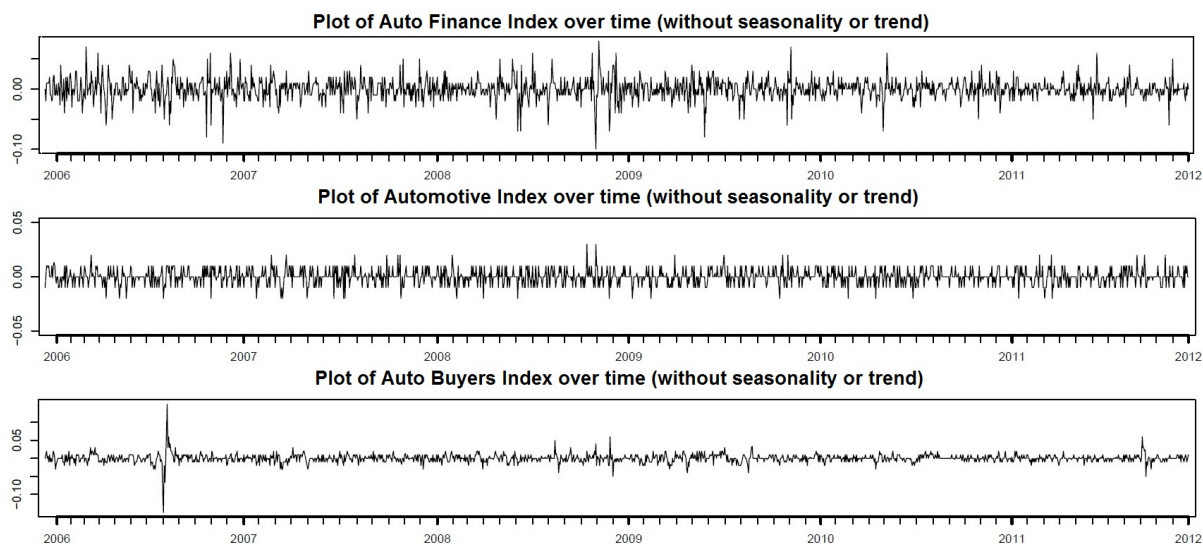


FIGURE 63: These three figures illustrate the internet auxiliary time series (with seasonality as well as trend removed) that we are attempting to include in our prediction of the DJTATO index returns. The top figure contains the auxiliary information related to the AFI category, the middle contains the auxiliary information related to AI category and the final contains that of the ABI category.

Finally, before we can continue to the actual prediction of our primary time series, with this additional information obtained from Google Domestic Trends, we must notice that the observations in the auxiliary time series are taken every day, while the market activity is only measured on weekdays (excluding public holidays). The above Figure 63 represents the auxiliary data after data cleansing which ensures that the two time series are of identical nature and taken on the same dates. During this process of data cleaning, the observations in the auxiliary time series (on days where there are no corresponding stock market observations) are combined with the following observations (on dates at which stock market data is available) by simply taking the average of all the consecutive internet measurements for which no stock market data exist. For instance, during a weekend the auxiliary information is measured on the Saturday, Sunday and Monday, where the DJTATO index is only taken on the Monday. The three observations taken during the weekend are therefore combined to represent a single observation taken on the Monday. After this data cleansing process, both time series are effectively the same in terms of dates on which the measurements are taken as well as the inherent characteristics of the measurements.

These three auxiliary time series have quite distinctive properties. The AFI has more variation and seems to behave quite erratically. The AI index on the other hand has very little deviation from zero, rarely reaching values larger in magnitude than 0.01. Finally the ABI is also quite stable, generally with values smaller than 0.02 in magnitude, but with several prominent spikes, especially during mid 2006. Reconsidering Figure 61, we see that this spike is not because of a significant change during 2006, but rather because of a significant change relative to the observed search volume of the previous year. This highly significant outlying observation in 2005 is caused by consequences of hurricane Katrina during 2005. Damage caused by this hurricane resulted in fuel leakages which in turn caused dramatic increases in fuel prices. Since seasonal differencing is used to eliminate seasonality, this outlying observation is resonated in 2006 even though it is only relevant in 2005.

The question this Chapter attempts to answer is whether it is valid to include the information extracted from the Google Domestic Trends web-based application when attempting to predict our primary DJTATO returns. In previous articles on associated work (as cited in Chapter 1), predominantly linear

methods were used to include auxiliary internet information. In order to see whether these linear methods would be applicable here, we consider the linear correlation between the two time series at different lags. For both primary and auxiliary time series, the observations between January 2006 and December 2009 are considered and in-sample cross-correlations are calculated between these two time series. Figure 64 illustrates these calculated cross-correlations and their significance.

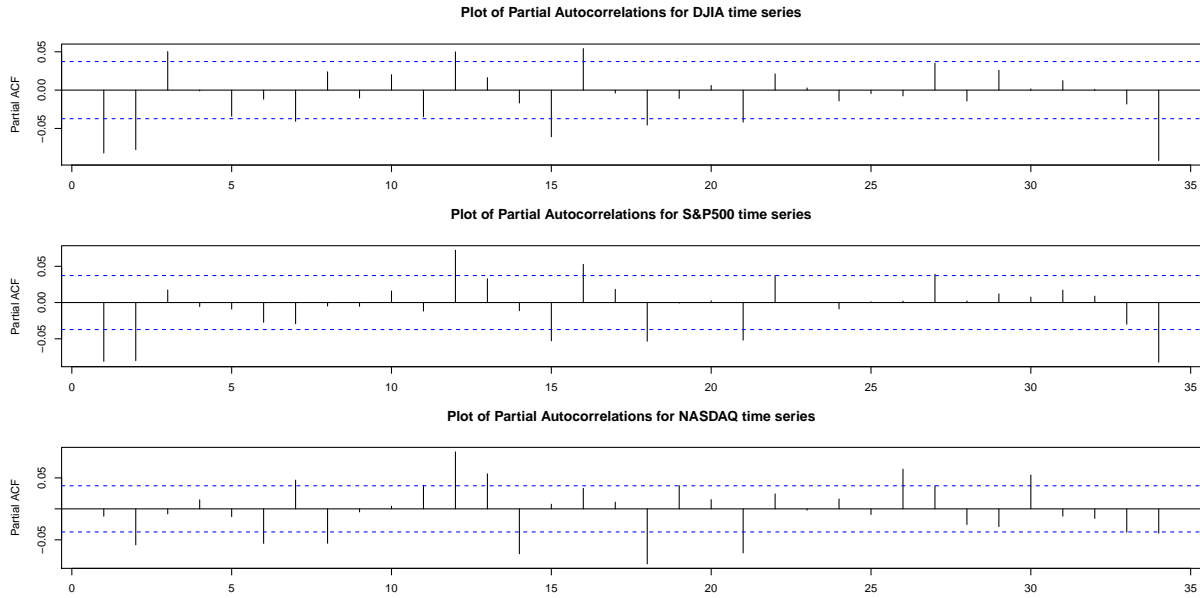


FIGURE 64: These three figures illustrate in-sample cross-correlations between the time series obtained from Google Domestic Trends and the DJTATO returns. The top figure contains the cross-correlations between the DJTATO returns and the time series obtained from the AFI category, the middle contains the cross-correlations between the DJTATO returns and the time series obtained from the AI category and the final contains the cross-correlations between the DJTATO returns and the time series obtained from the ABI category.

Notice in the above figures that the in-sample cross-correlations, $Cor(y_{1,t+lag}, y_{2,t})$ (where $y_{1,t}$ represents the t th DJTATO return and $y_{2,t}$ represents the t th auxiliary time series value obtained from Google Domestic Trends), are drawn against possible lag values. Therefore, if correlations are significant for positive lag values, it can be said that internet data leads stock market data. In other words, historical data from the internet is significantly correlated with today's DJTATO returns. Upon inspection of Figure 64 we conclude that there are no significant linear cross-correlations either at positive or negative lags. Based on this figure, including the auxiliary information in a linear fashion would therefore most likely not improve prediction accuracy. However, perhaps by using non-linear, multivariate SSA techniques we could still possibly use this auxiliary information to gain predictive power over the primary time series.

In order to investigate whether the Google Domestic Trends application can be useful in prediction of the DJTATO time series, we first need to use the univariate SSA prediction methodology to obtain baseline predictions for the DJTATO returns observed during 2011. Predicting these 259 observations with the univariate SSA approach resulted in a MSE of 33.2635, a MAE of 4.3626 and a DOC of 52.8957%. We attempt to improve on these predictions using data obtained from the internet and incorporating this data by using one of the multivariate SSA approaches.

Firstly we consider the possibility of using the HMSSA approach to include the auxiliary information

in prediction. It should be intuitive (after the thorough simulation study in Chapter 2) that in this practical scenario, even when including the auxiliary information, we should still be using a dimension of 1 in the HMSSA context. This was confirmed in a cross-validation study. By predicting the DJTATO returns observed during 2011 (using the HMSSA approach with window length of 22 and dimension of 1) we found that none of the 3 auxiliary time series had any effect on the predictions at all. This is of course a result of the difference in scale between the primary and secondary time series. The range associated with the primary time series is nearly 90, whereas the range associated with the secondary time series is never more than 0.3. These deviations in the secondary time series are therefore deemed irrelevant in the SVD of the HMSSA trajectory matrix and predictions are therefore still merely based on the historical data of the primary time series.

In the simulation study we saw that we can overcome such scale differences by using the proposed Bayesian approach. This gives us reason to believe that even though the HMSSA approach was not able to incorporate the auxiliary information, the Bayesian approach might still be able to do so. Since this proposed Bayesian approach requires us to use the same window length for the individual analyses of both time series and since we are only interested in the accurate prediction of the primary time series, we continue using a window length of 22 as suggested by the cross-validation study of the primary time series. By performing a cross-validation study on the secondary time series at a window length of 22, we find that all three the secondary time series should also be considered to be in a signal space with dimension of 1.

With all the required parameters estimated, we can continue to the actual prediction of the DJTATO returns during 2011 in the presence of the Google Domestic Trend data. Papers cited in Chapter 1 attempted to predict tomorrow's stock market activity using today's internet activity. This is one possible approach. Another is to use the current internet activity to attempt to predict today's closing price. The difference is, with the first method we have a lag of one day, while with the second method we have zero lag. Proposing that current internet activity could perhaps contain predictive power for today's closing price, we therefore use the Bayesian approach to predict the 259 values of 2011. The Bayesian approach however requires two additional parameters. The first parameter is the number of bootstrap replicates we should base the approach on (B) and more importantly, the second parameter is the level of significance (α) at which the hypothesis test, testing the validity of inclusion of the auxiliary time series, should be evaluated. The first parameter is simply a trade-off between computational complexity and robustness of the Bayesian predictions. As B increases, the bootstrapping approach is more accurate, but the computational time increases as well and vice versa for a decreasing B . The α parameter on the other hand is significantly more important. It reflects the user's confidence in whether the auxiliary information should be included in estimation of the signal space. Since this thesis tests the hypothesis that we can indeed include this information, we first attempt to base predictions on an α value of zero (thereby always allowing for inclusion of the auxiliary information) and $B = 250$. This results in predictions as given by Figure 65.

Upon inspection of Figure 65, we notice some interesting things. The first, most obvious thing is the fact that none of the predictions are particularly good; the predictions merely show slight deviations around zero. Even though this additional information influenced our predictions significantly, the magnitude of our residuals are still large. However, keeping in mind the nature of the primary data to be predicted, we must acknowledge the fact that stock market activity is considered to be nearly unpredictable. We are not interested in precise predictions, since this would be unrealistic. What we are interested in is whether inclusion of additional web based information increases the accuracy of prediction

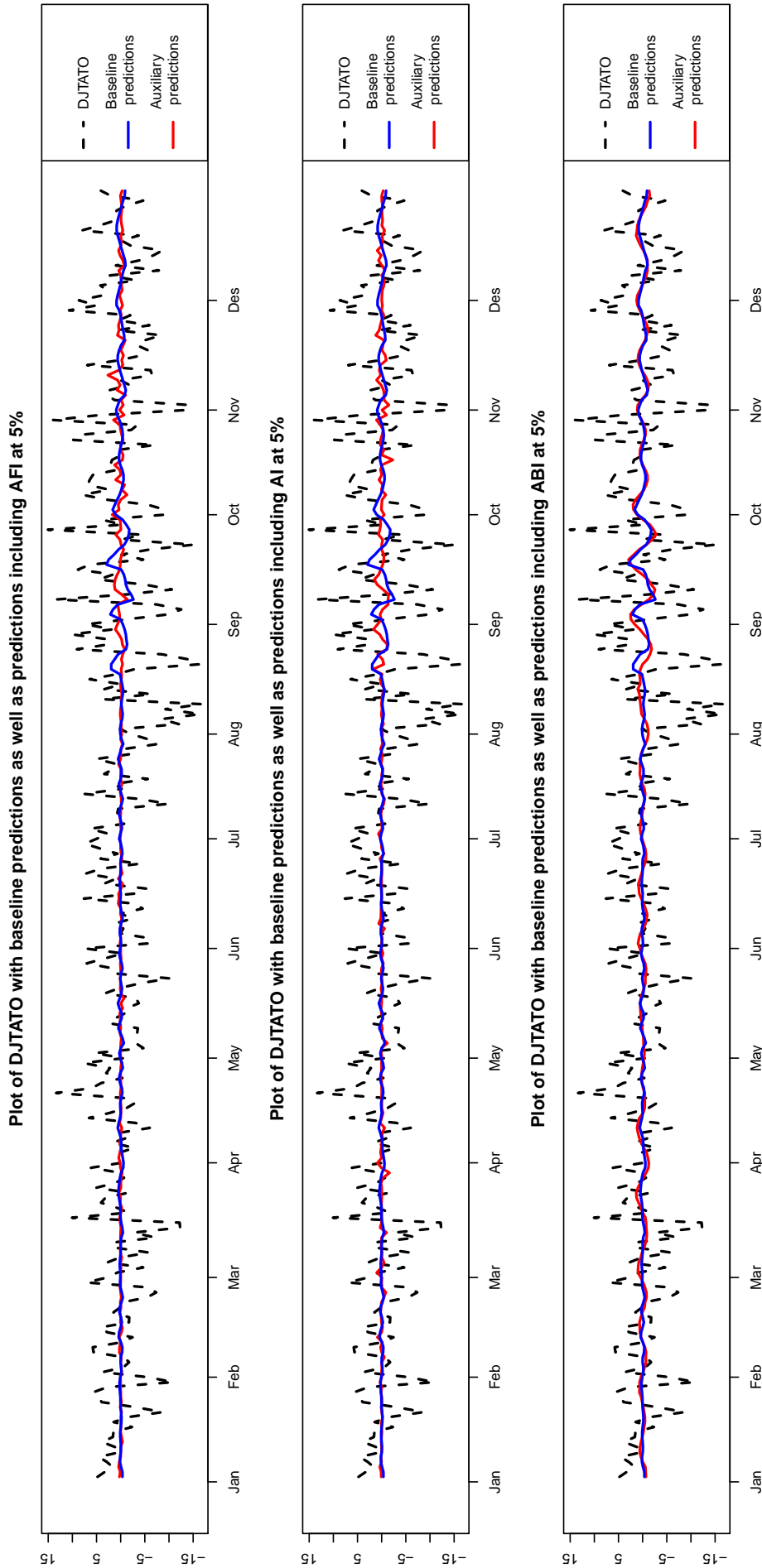


FIGURE 65: These figures represent the predictions obtained through both the univariate and Bayesian multivariate prediction algorithms and compares them with the actual observed DJTATO returns. The top figure illustrates how inclusion of the AFI time series affects prediction, the middle figure illustrates how the inclusion of the AI time series affects prediction and the bottom figure illustrates how the inclusion of the ABI time series affects prediction.

significantly. On a secondary note, we are also interested in whether we can keep our predictions stable by not increasing the measures quantifying the magnitude of the residual, while also improving our DOC measure. To address these objectives, we consider the measures of accuracy for the baseline predictions as well as the predictions incorporating internet activity as given in Table 21.

Univariate predictions		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
33.2635	4.3626	52.8957%
Including AFI		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
31.83871	4.2949	49.8069%
Including AI		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
31.77602	4.3065	52.1236%
Including ABI		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
32.4907	4.3199	54.0541%

TABLE 21: This table shows both the measures of accuracy for the baseline DJTATO predictions during 2011 as well as that of the predictions when including data obtained from Google Domestic Trends.

According to these figures we see that the Bayesian approach consistently provided predictions with lower MSE and MAE. The DOC measure of the predictions is unfortunately not as satisfying. When including the AFI or the AI auxiliary information, we see that the DOC decreases. This means that we correctly predict the direction of DJTATO returns less often. In an attempt to explain this phenomena, we consider the nature of the singular vectors and the validity of the inclusion of the internet time series.

During 2011 we had 259 predictions and for each of these predictions we had a single singular vector associated with the baseline predictions. It could be interesting to consider how this singular vector changed over time. This is illustrated in Figure 66. Each vertical entry in the bottom figure illustrates a single singular vector observed at a given time and each of these vertical entries consists of L coloured blocks representing the L entries in the singular vector. The colour of each of these blocks represent the value of the relevant entry in the relevant singular vector.

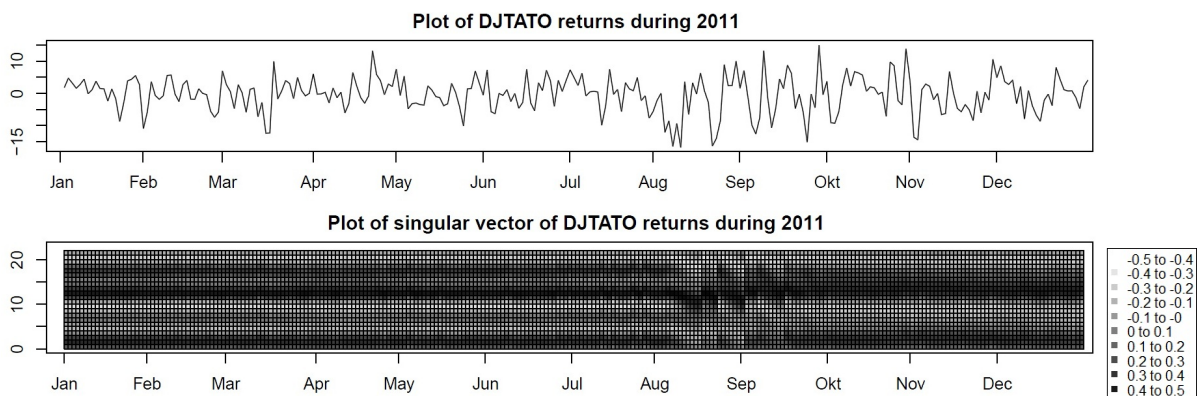


FIGURE 66: These figures represent how the primary time series as well as the associated singular vector changed during 2011. The top figure shows the test segment of the DJTATO returns and the bot-

tom figure illustrates how the singular vector (which determines the predicted value) changed during 2011.

The above figure shows us that the SSA algorithm notices some signal resembling a harmonic trend in the historical DJTATO returns. However, during August the nature of the singular vector undergoes some changes and the singular vector becomes more unstable. The reason for this instability is clear when we consider the top figure. During August, the returns seem to be generally lower and more volatile than previously, causing the SSA algorithm to believe that the time series might be undergoing some structural break. During September, both the DJTATO returns as well as the singular vector associated with prediction seem to stabilize once again. The behaviour of the singular vector obtained from the auxiliary information is however slightly different.

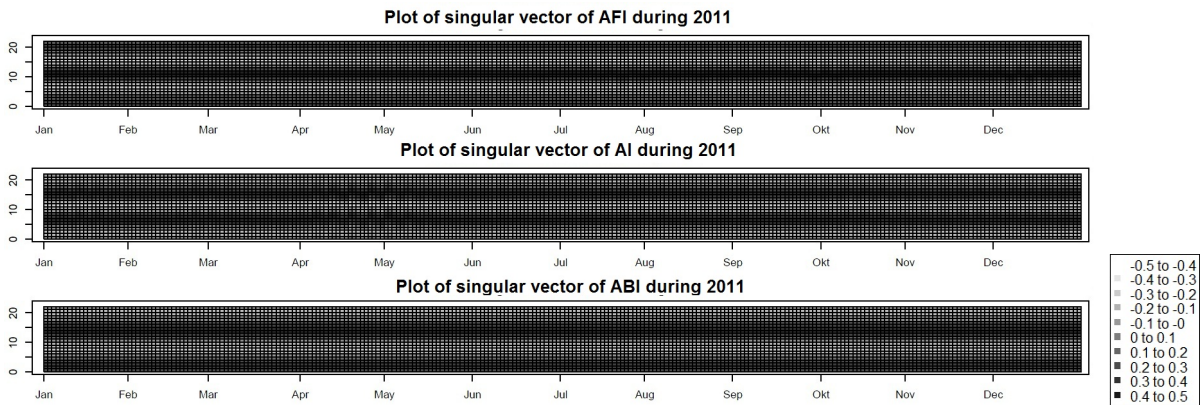


FIGURE 67: These figures represent how the singular vector associated with the secondary time series changed during 2011. The top figure illustrates the significant singular vector for the AFI time series, the middle figure illustrates the significant singular vector for the AI time series and the bottom figure illustrates the significant singular vector for the ABI time series.

In Figure 67, we can see slight differences among the singular vectors of the three auxiliary time series. Even though the plot of these three time series (Figure 63) seem to be relatively random to the naked eye, according to this analysis there seems to be some inherent signal resembling a harmonic component in each of the time series. It is also quite interesting to see how little variation there is in these singular vectors. Comparing these singular vectors with those obtained from the primary time series, we see that they all share some harmonic nature. However, only the singular vector obtained from the ABI time series really seems to resemble the singular vector associated with the primary time series. In the above predictions, we simply assumed that we can incorporate the auxiliary information, by letting $\alpha = 0$. However, these figures (Figures 66 and 67) cast some doubt on this assumption. By assuming that inclusion is valid (by letting $\alpha = 0$), we would combine the vector associated with the primary time series with the vector associated with the secondary time series at each prediction step. In order to investigate how this prediction vector changes over time, we can consider Figure 68.

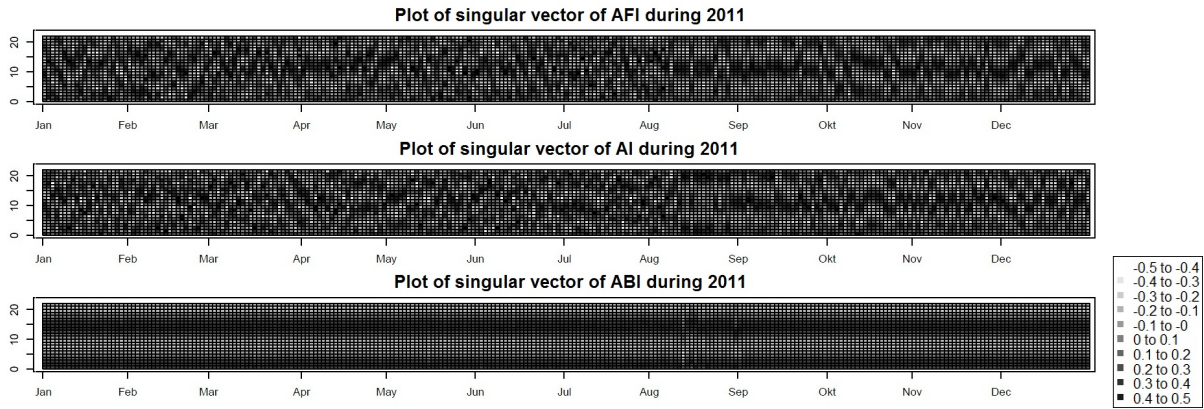


FIGURE 68: These figures represent how the singular vector associated with prediction (as calculated with the Bayesian algorithm) changed during 2011. The top figure illustrates the prediction vector when including the AFI time series at $\alpha = 0$, the middle figure illustrates the prediction vector when including the AI time series at $\alpha = 0$ and the bottom figure illustrates the prediction vector when including the ABI time series at $\alpha = 0$.

The singular vectors in the top and middle plots of Figure 68 are clearly not stable. Even though they produce predictions that result in lower MSE and MAE, we tend not to have confidence in these predictions because of this instability in the associated prediction vectors. Also, the predictions obtained from the vectors in the top and middle figures resulted in a lower DOC (see Table 21). We might therefore have reason to believe that, at some of these time points, inclusion of the auxiliary information in the AFI and AI time series is not valid. Considering the ABI auxiliary time series on the other hand, we see that the prediction vectors obtained through the BMSSA algorithm seem much more stable and also resembles the singular vectors in the primary time series, only with less variation. Inclusion of the ABI time series therefore seems valid at most (perhaps all) time points.

Figure 68 reminds us that we should perhaps not merely assume that we can include the auxiliary information since inclusion might produce unstable singular vectors and therefore unstable predictions. The safer approach would be to apply the BMSSA algorithm with $\alpha = 0.05$, thereby testing whether inclusion of auxiliary time series is appropriate beforehand. With this approach, our vectors associated with prediction would change slightly. This is illustrated in Figure 69. Notice that the time points where inclusion is allowed is marked with a tick above the produced prediction vector.

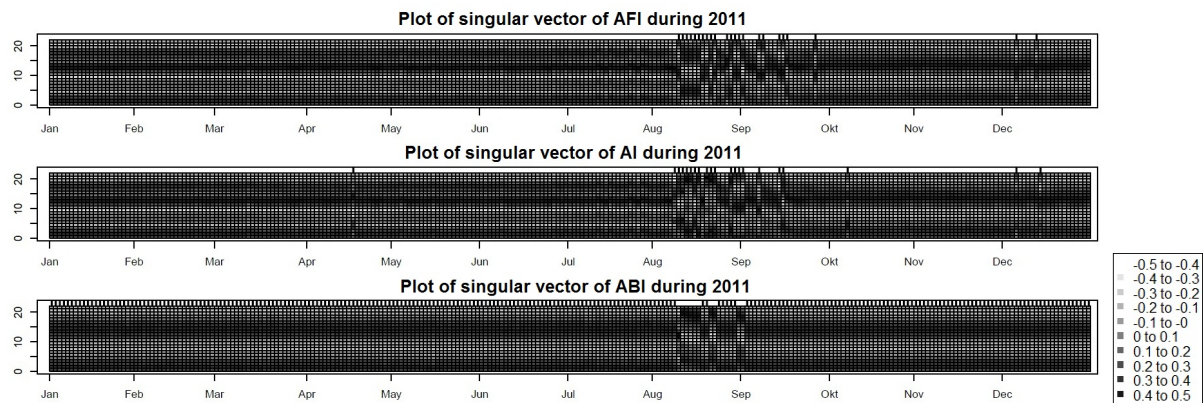


FIGURE 69: These figures represent how the singular vector associated with prediction (as calculated with the Bayesian algorithm) changed during 2011. The top figure illustrates the prediction vector when including the AFI time series at $\alpha = 0.05$, the middle figure illustrates the prediction vector when includ-

ing the AI time series at $\alpha = 0.05$ and the bottom figure illustrates the prediction vector when including the ABI time series at $\alpha = 0.05$.

This figure raises some interesting points. We already mentioned that the AFI and AI time series should perhaps not be incorporated into our predictions. When testing this inclusion at 5% level of significance, we find that the inclusion of these time series were only allowed 21 and 23 times (out of the 259 possible time points) for the AFI and the AI time series respectively. Coincidentally, inclusion was only allowed specifically at time points where there seemed to be instability in the primary time series (during August and early September). When incorporating the ABI time series on the other hand, the algorithm allowed for inclusion at 249 time points. This inclusion was only allowed when the primary time series seemed stable. To investigate the effect of these inclusions on prediction, we can consider Table 22. This table gives us the measures of accuracy when predicting according to the prediction vectors in Figure 69.

Univariate predictions		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
33.2635	4.3626	52.8957%
Including AFI		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
32.2507	4.3127	55.5985%
Including AI		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
32.6372	4.3240	55.2124%
Including ABI		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
32.9039	4.3385	52.5097%

TABLE 22: This table shows the measures of accuracy for the baseline DJTATO predictions during 2011 as well as that of the predictions when including data obtained from Google Domestic Trends at a 5% level of significance.

Comparing the accuracy of these adjusted predictions to that of our baseline predictions, we see that even though the AFI and AI time series were only included 21 and 23 times respectively, there were clear increases in the DOC measures, and slight decreases in MSE and MAE. This is also true when including the ABI time series (at 245 time points). These are of course favourable results, but are they statistically significant? A single measure of the accuracy of prediction is clearly not enough to base our conclusion on and we therefore continue to investigate the differences between these predictions.

In the simulation examples, we used a Student's t-test to determine whether there are statistically significant differences between the different prediction methods in a certain scenario. In this practical scenario, we are more interested to test whether there are statistically significant differences in the residuals of the prediction methodologies at a certain time point. Suppose for instance that the univariate SSA algorithm produces a residual $e_{t,1}$, at time point t and the Bayesian approach produces a residual $e_{t,2}$, at the same time point. To test whether $e_{t,2}$ is statistically smaller in magnitude than $e_{t,1}$, we need to compare the individual differences between the absolute residuals (or squared residuals) and not the differences between the average of the absolute residuals (or squared residuals). Figure 70 illustrates our first attempt to evaluate this hypothesis.

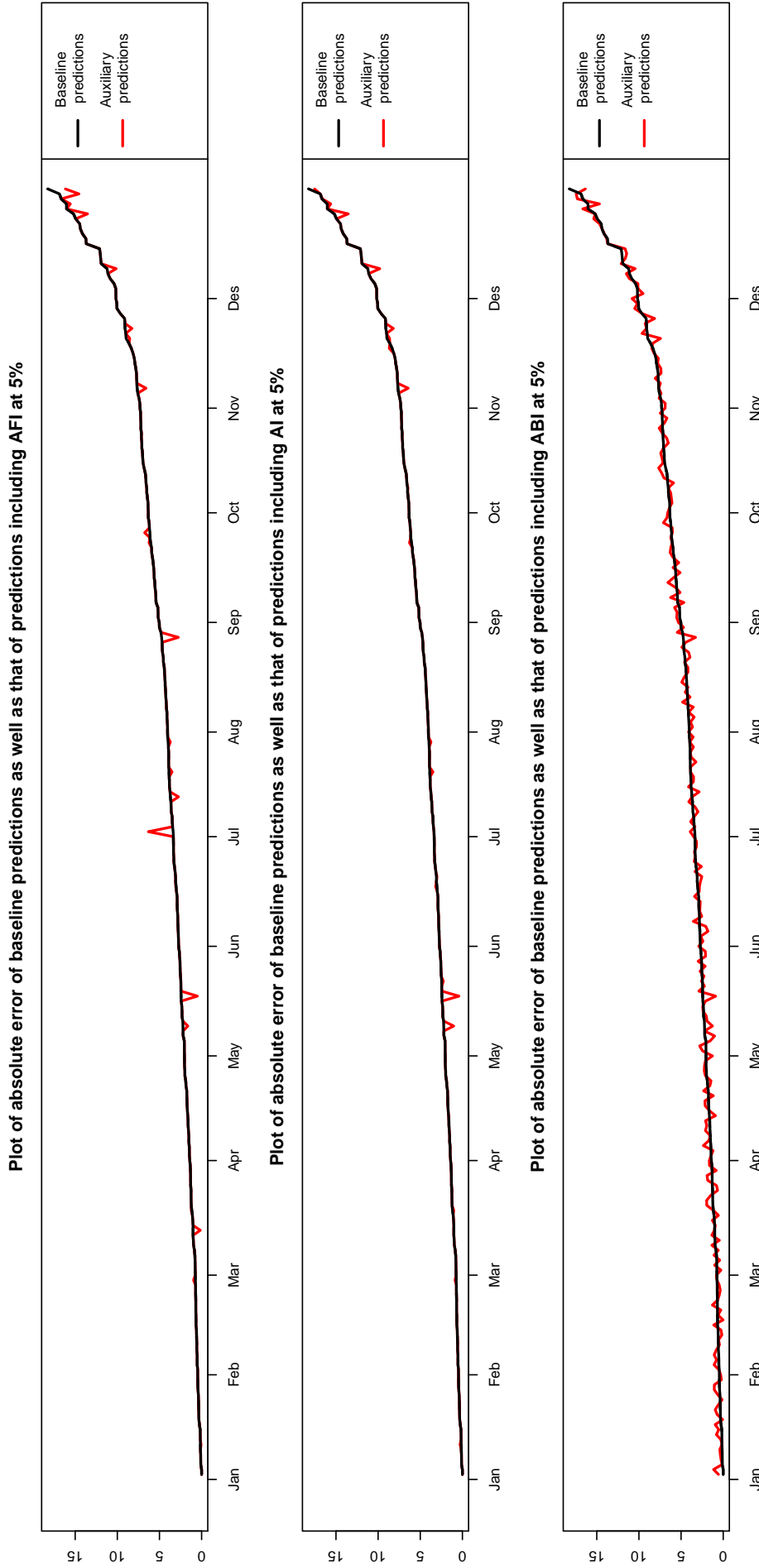


FIGURE 70: This plot illustrates the increasing absolute errors when predicting according to the univariate approach and how the Bayesian predictions (when including at 5% level of significance) affected these absolute errors.

Figure 70 clearly illustrates the reason why inclusion of the AFI and AI time series resulted in better measures of prediction accuracy even though inclusion was only allowed 21 and 23 times for the AFI and AI time series respectively. At the few time points - where we did in fact include the AFI and AI time series - the Bayesian predictions often decreased the absolute error when compared with Baseline predictions. Inclusion of the ABI time series on the other hand resulted in subtler changes. The change in absolute error varies substantially; however, by reinspecting Table 22 we can see that these subtle changes were mostly beneficial to prediction accuracy.

Since these changes in our individual absolute error terms might seem slight, we are therefore interested in testing whether they are significant. To test the significance of these changes, we use a paired Student's t-test, since we want to compare the residuals individually. This entails obtaining the differences between the individual absolute error pairs $(e_{t,1} - e_{t,2})$, or squared error pairs $(e_{t,1}^2 - e_{t,2}^2)$, and testing whether the mean of these differences is significantly greater than zero. In order for us to perform such a paired Student's t-test, we must assume that the sample mean is normally distributed. This assumption is justified by the Central Limit Theorem which states that this is asymptotically true. Furthermore, we must also assume that the individual differences are independent and homoscedastic. This is of course not necessarily true, but the assumption is made here in order to have some way in which to test for statistical differences. Table 23 tabulates the observed mean differences and the p-values associated with testing whether the mean differences are significantly greater than zero. These differences are tested both when we assume that inclusion is appropriate ($\alpha = 0$) and also when we apply the Bayesian algorithm at 5% level of significance.

<i>Paired t-test on absolute errors</i>			
Including AFI at $\alpha=0.05$		Always including AFI	
<i>Mean difference</i>	<i>p-value</i>	<i>Mean difference</i>	<i>p-value</i>
0.0677	0.05163	0.0499	0.01437
Including AI at $\alpha=0.05$		Always including AI	
<i>Mean difference</i>	<i>p-value</i>	<i>Mean difference</i>	<i>p-value</i>
0.0561	0.1241	0.0385	0.0036
Including ABI at $\alpha=0.05$		Always including ABI	
<i>Mean difference</i>	<i>p-value</i>	<i>Mean difference</i>	<i>p-value</i>
0.0426	0.0996	0.0241	0.2062
<i>Paired t-test on squared errors</i>			
Including AFI at $\alpha=0.05$		Always including AFI	
<i>Mean difference</i>	<i>p-value</i>	<i>Mean difference</i>	<i>p-value</i>
1.0127	0.0164	1.4248	0.0153
Including AI at $\alpha=0.05$		Always including AI	
<i>Mean difference</i>	<i>p-value</i>	<i>Mean difference</i>	<i>p-value</i>
0.6262	0.0085	1.4875	0.0242
Including ABI at $\alpha=0.05$		Always including ABI	
<i>Mean difference</i>	<i>p-value</i>	<i>Mean difference</i>	<i>p-value</i>
0.3596	0.202	0.7728	0.0627

TABLE 23: This table gives us the difference between the baseline predictions and the Bayesian predictions in terms of difference in the MAE and MSE measurements. The significance of these differences are then also tabulated in the form of p-values obtained from paired a Student's t-test.

Even though not all the p-values here are significant, in their entirety they indicate that the internet auxiliary information significantly influenced the predictions. This is already significantly better than the HMSSA prediction algorithm, which was not even able to change the predictions at all.

The right side of Table 23 illustrates that when we let $\alpha = 0$, inclusion of either the AFI or the AI time series significantly decreases both the squared error and the absolute error. However, we mentioned before that the instability of the singular vectors associated with these predictions made us doubt the confidence we had in said predictions. For example, reconsidering Table 21, we see that even though this inclusion significantly decreased the squared error as well as absolute error, it also decreased the DOC and because it is of practical importance to keep the DOC above 50%, we conclude that choosing $\alpha = 0$ would be a high risk option. This resonated in further practical scenarios and this approach was therefore omitted in following practical examples.

Considering on the other hand the left hand side in Table 23, we find predictions that are uniformly better, instead of merely decreasing the absolute errors and the squared errors. Inclusion of the AFI and AI time series still produce predictions that have significantly lower squared errors and even though the absolute errors associated with these predictions are not significant at 5% level of significance, the p-values are still noticeably low. Apart from these p-values, the inclusion of the auxiliary time series increased the DOC measure to over 55% with both inclusions. Inclusion of the ABI time series (with $\alpha = 0.05$) on the other hand had slightly different implications. The predictions obtained when including the ABI time series in this manner decreased both the absolute errors as well as the squared errors. Unfortunately, Table 23 indicates that these changes are not significant at a 5% level of significance; however, the p-values are still noticeably low. Furthermore, the inclusion of this auxiliary information resulted in slightly lower DOC relative to that of the baseline predictions, even though the DOC is still slightly higher than 50%. This once again shows that when testing for inclusion, we might not find predictions that are immensely more accurate, but we will obtain safer predictions than when we choose $\alpha = 0$.

In conclusion, we might ask the question, why did inclusion of the AFI and AI time series significantly improve on the baseline predictions, even though they were only included in 21 and 23 predictions? The reason for this is quite simply because of the time points at which these 21 and 23 inclusions took place. We saw in Figure 69 that inclusion was allowed predominantly during August and September. During this time, the primary time series was slightly more volatile and the singular vectors associated with the primary time series were unstable. Since the information regarding internet activity was more stable during this time, it proved helpful when attempting to predict the primary time series. In practical terms we can explain this phenomena by saying that the primary time series had a structural break during this time and the internet activity (as measured by the AFI and AI time series) explained this structural break and was able to provide information on future values of the DJTATO index when the primary time series was unable to do so.

Similar reasons hold for why inclusion (at 5% level of significance) of the ABI time series did not provide significantly improved results. When the univariate predictions needed auxiliary information during August and September, the ABI time series was unable to provide such information according to the hypothesis test. The inclusion was therefore unable to help during this period where it was essential. Notice that if we were to merely include the ABI during this period as well, we would have found more apparent improvements (see Table 21). However, since the singular vector associated with the auxiliary

information was too different from the singular vector associated with the primary time series at the time, the hypothesis test did not allow for inclusion and we were left with the safer predictions that were merely based on the univariate data.

Even though these results might not provide undeniable evidence to support the hypothesis in this thesis, we would just like to briefly revisit the fact that when we used the Bayesian approach (with $\alpha = 0.05$) to include auxiliary information, we consistently improved on the absolute errors and the squared errors, while at the same time keeping the DOC stable and above 50%, or sometimes even improving on it. This was when we were predicting a volatile time series in the presence of an auxiliary time series which has lost a lot of information due to normalization, truncation and differencing.

6 Incorporating Twitter frequency counts

In the previous chapter, we considered a single practical scenario where we attempted to predict stock market activity by using search volume information obtained from Google Domestic Trends. Some favourable results were found; however, we discussed some inadequacies of the auxiliary information. The clear seasonal trend forced us to use seasonal differencing in order to obtain auxiliary information that did not include a seasonal component. Taking seasonal differences is a very crude way of removing cyclical effects. Apart from seasonal differencing, we also took differences at a lag of one. This together with Google Domestic Trends' normalisation procedures inevitably caused the auxiliary time series to lose some information that could have been useful in prediction.

Alternative web-based applications can also be used to obtain some measurement of internet activity. In an attempt to consider alternative web-based auxiliary information sources, we discuss in this chapter the possibility of using the microblog website, Twitter, to obtain information with predictive power. The information we use in this chapter is obtained from a website called Topsy (<http://www.topsy.com>). This website is essentially a database of all Tweets since 2008. Perhaps the number of Tweets obtained from this web-based application can be indicative of future stock market activity.

It seems intuitive to believe that people posting messages on Twitter are technologically aware and therefore we believe that some of the information available on Twitter can perhaps give us insight into the present desirability to buy shares in certain technological companies such as Microsoft, Google and Apple. This chapter proposes that the number of Tweets regarding a company can be used to improve on predictions of the returns of said company.

6.1 Predicting Google stock prices

The first company we consider is Google. Google has grown significantly and has received quite some attention in the blogosphere. Internet users have grown fond of giving their opinion on Google and related products on web-blogs such as Twitter. This practical example therefore focuses on determining whether the number of times people mention Google on the internet could be used to predict Google stock prices.

Our primary time series in this practical example is therefore Google's daily closing prices. The closing prices of Google shares, since late 2004, are obtained from Yahoo! Finance and illustrated in Figure 71. Previous practical examples (see Chapter 3) on prediction of stock market activity showed that the SSA

methodology produces more accurate predictions when applied to returns of shares rather than when it is applied directly to stock market activity. Parameter estimation was also significantly easier when we applied the SSA algorithm to stock returns. Figure 71 therefore also displays the returns corresponding with daily Google closing prices (since August 2004).

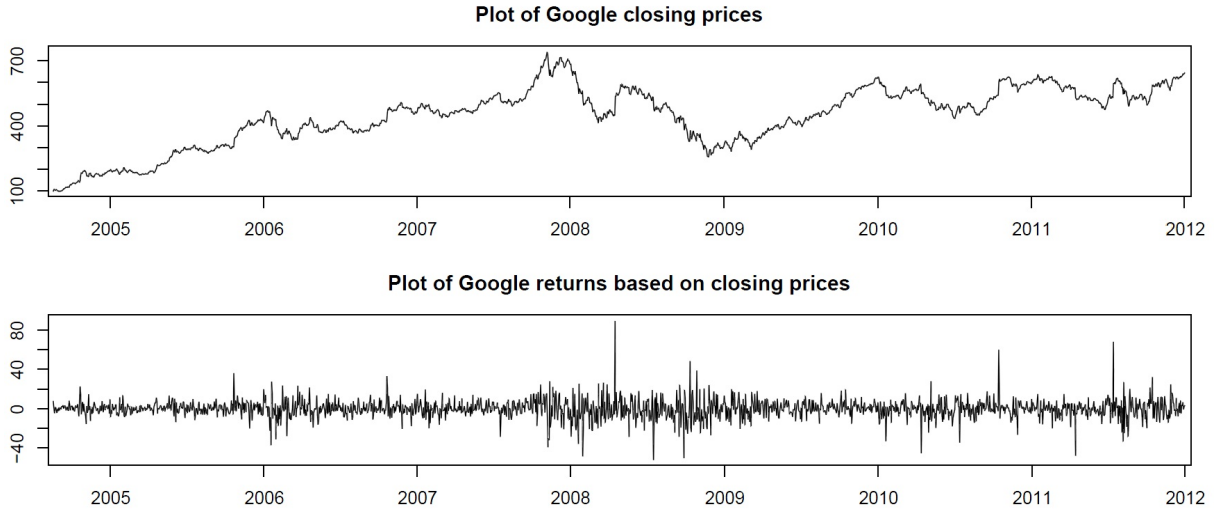


FIGURE 71: These two figures illustrate the primary stock market time series to be predicted. The top figure contains the observed daily Google closing prices and the bottom figure contains the returns of the observed daily Google closing prices.

In this figure we see, once again, that Google was not unaffected by the recession that started near the end of 2007. However, once the recession was over, the company value showed steady increases with some volatile behaviour during mid 2011. Similar conclusions can be made by considering the bottom figure. The returns seem to show erratic deviations during 2008 (as a consequence of the recession) and also some outlying observations during 2011. Once again, we partition our primary time series (the returns of the Google closing prices) in order to subject it to cross-validation methods. Historical data up until the last day of 2009 is used for model fitting. This model fitting segment is then used to obtain cross-validation predictions for all the observed returns during the year 2010. The final segment of the primary time series can then be used to test the accuracy of our final predictions.

These cross-validation predictions produced results very similar to that of previous chapters (Chapters 3 and 5). When predicting the cross-validation segment with rolling one-step-ahead predictions and parameters varying over the same domain as before ($L \in \{2, \dots, 35\}$ and $d \in \{1, \dots, L - 1\}$), we saw that by choosing any value other than $d = 1$ significantly decreased the accuracy of prediction for the cross-validation segment, regardless of the choice of window length. Also, we found that the MSE stabilizes for L greater than 20 (or thereabout). Based on these cross-validation predictions, we found that the optimum choice of window length was 22 since this minimized the MSE of cross-validation predictions. After values for the window length and dimension parameters have been decided on ($L = 22$ and $d = 1$), we were able to calculate rolling one-step-ahead predictions for the test segment. These represent our baseline predictions. These predictions produced measures of accuracy as later given in Table 24.

In order to improve on the baseline predictions already calculated, we consider our auxiliary time series. This information is obtained by searching the web-based database, Topsy, for Tweets containing the word "Google". All Tweets containing the term "Google" (since January 2009) were scraped from

Topsy and Figure 72 illustrates these daily frequency counts. Apart from these frequency counts, Figure 72 also graphically illustrates the differenced frequency counts since this is what we will use as auxiliary time series.

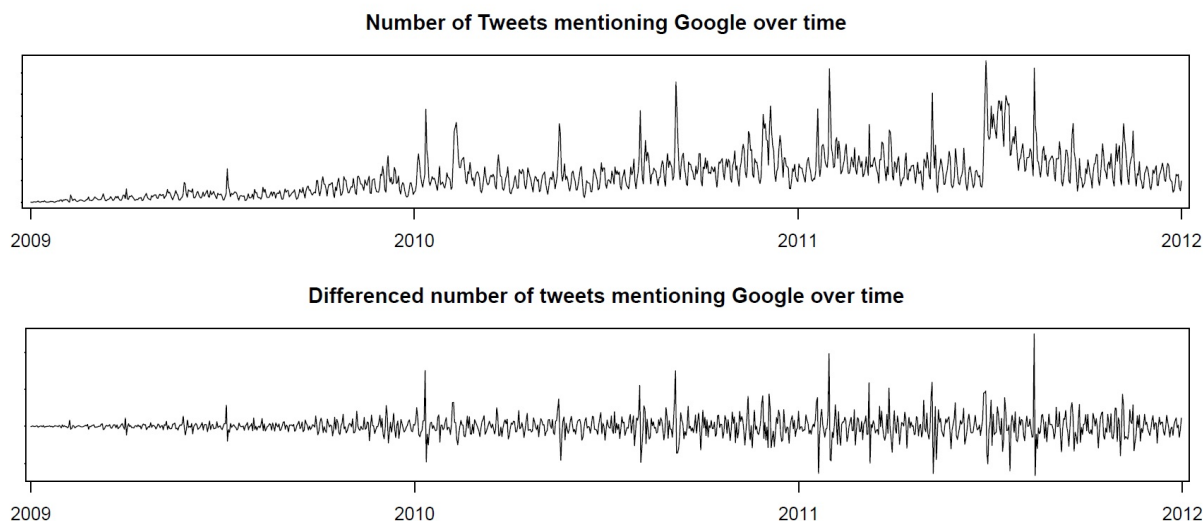


FIGURE 72: These two figures illustrate the auxiliary information obtained from Topsy. The top figure contains the frequency counts of Tweets with keyword "Google" and the bottom figure simply contains the differenced frequency counts.

Notice once again that the bottom figure represents the frequency counts after it was subjected to a data cleaning process. The data cleaning process here is similar to that of Chapter 5, where we remove values in the auxiliary time series that was taken on dates for which no stock market activity was measured, by merely combining them with future auxiliary time series values that do correspond with stock market measurements.

It is interesting to take note of the way these two time series change over time. The top time series clearly starts at a value close to zero and steadily increases both in terms of magnitude and volatility. The bottom figure also shows how the the time series became more volatile; initially the time series showed little fluctuation, but during 2011 the time series seems to show much more prominent variation. These properties illustrate how the popularity of Twitter grew since 2009. Similar characteristics will be noticed in other Twitter time series as well. Apart from these characteristics describing general Twitter behaviour, we should take note of other important points as well. For instance, consider the behaviour of the top time series during 2011. There is a slight decreasing trend present and in the middle of the year a big spike, after which the time series returns to some baseline behaviour. This resembles to some extent the behaviour in the primary time series as well (see Figure 71). The reason why Google was mentioned often during this time was the fact that on 29 June, Google had launched a series called Project Glass, which essentially entails a cell-phone like device built into a pair of glasses. This then caused investors to become optimistic on future prospects of Google and it reflected in both Tweet frequency counts and Google stock prices. Also notice the very significant spike in the bottom time series. On 14 August, Google purchased a company called Motorola Mobility in an attempt to combine Google's software with Motorola's cellphone hardware. This time point will also be of significance when considering the results.

In previous practical examples, cross-validation methods showed that the auxiliary time series (as obtained from the internet) should also be considered to be of low dimension ($d = 1$). By subjecting

the above auxiliary time series to cross-validation methods, we quickly realise that this is also the case here. This is of course because of the fact that we are using the differenced time series and not the actual frequency counts. Deviations from zero therefore seem random and there seems to be little signal present. With all the necessary parameters estimated, we calculate the rolling one-step-ahead predictions with the BMSSA approach (including at 5% level of significance with 250 bootstrap replicates) for the test segment and compare these with baseline predictions. Figure 73 illustrates the increasing absolute values of the baseline predictions and how the auxiliary information influenced these absolute errors.

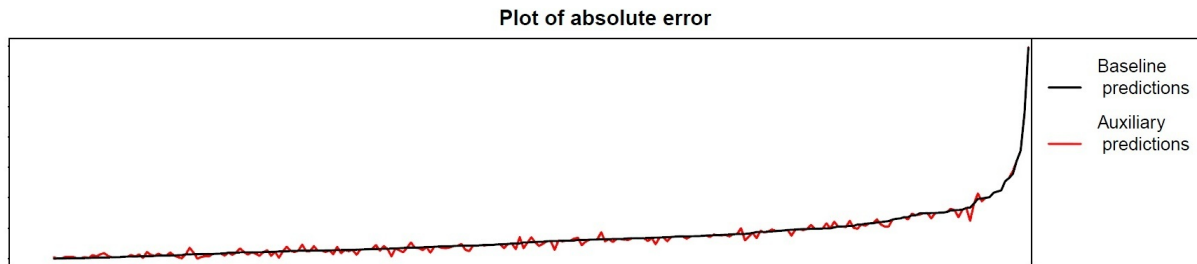


FIGURE 73: This plot illustrates the increasing absolute errors when predicting according to the univariate approach and how the Bayesian predictions (when including at 5% level of significance) affected these absolute errors.

Figure 73 clearly shows that the predictions were only slightly influenced by the inclusion of the auxiliary time series. At some points, the inclusion caused slight increases in absolute errors; however, it seems that the inclusion predominantly caused subtle decreases in absolute errors. To further study the effect of the inclusion, we consider first the singular vectors associated with the primary time series and those associated with the auxiliary time series.

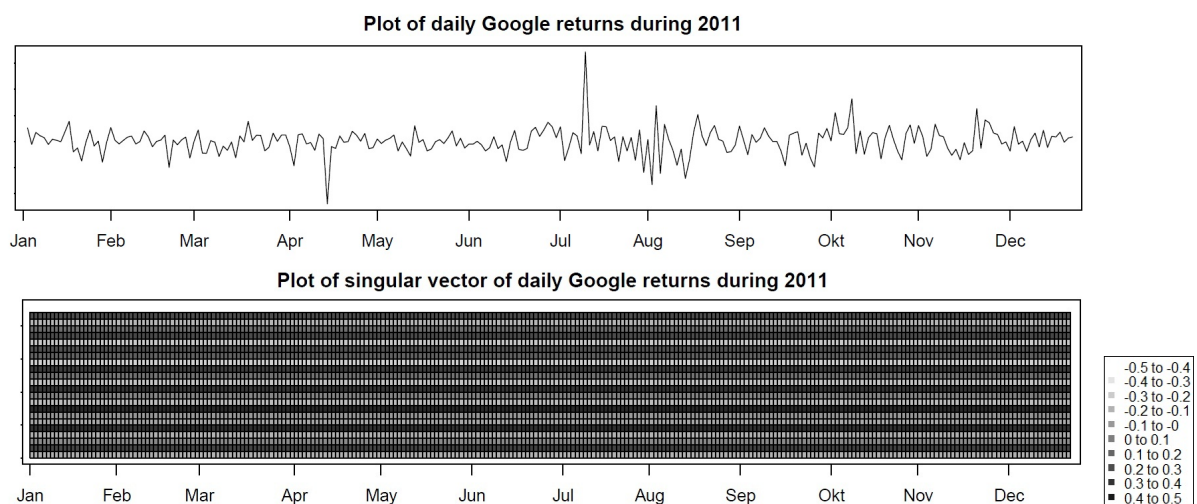


FIGURE 74: These figures represent how the primary time series as well as the associated singular vector changed during 2011. The top figure shows the test segment of the Google stock price returns and the bottom figure illustrates how the singular vector (which determines the predicted value) changed during 2011.

Figure 74 shows that there is little variation visible in the singular vectors associated with the primary time series. This is unexpected since in general we have that stock market activity is highly contaminated with noise and even slightly random. This variation should then resonate in the singular vectors. However, we must remember that variation in these singular vectors is much less visible. This is because of the fact that slight changes in the singular vector could have drastic effects in both filtering and pre-

diction. This visual representation might therefore not show significant changes in the singular vector, but this does not mean they are not there. Consider for instance the period between July and August. The primary time series is clearly more volatile here and this would necessarily affect the singular vectors during this time interval as well; however, the change is not evident in the bottom figure. Unfortunately, little can be said about the signal associated with these singular vectors. It seems that they represent some sort of high frequency harmonic component.

To consider the effect the auxiliary time series had on prediction, we can consider the manner in which the singular vectors associated with the secondary time series changed over time. This is given in Figure 75.

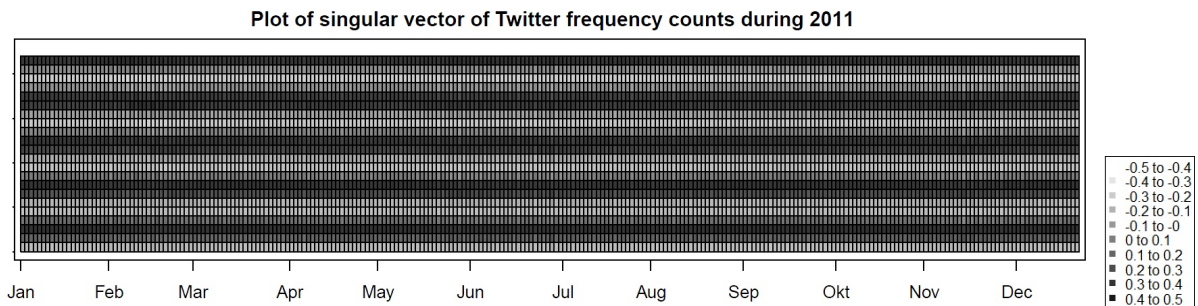


FIGURE 75: This figure represents how the singular vector associated with the secondary time series changed during 2011.

There seem to be some similarities between these singular vectors and those associated with the primary time series. The singular vectors in the above figure appear to be relatively stable and they also seem to represent a high frequency harmonic trend. Notice however that these singular vectors seem to have 5 peaks, where the singular vectors in Figure 74 have 8 peaks. We can therefore see that these two sets of singular vectors are similar but slightly different. To investigate whether the Bayesian approach considered these singular vectors to be similar enough to be combined (at 5% level of significance), we consider Figure 76.

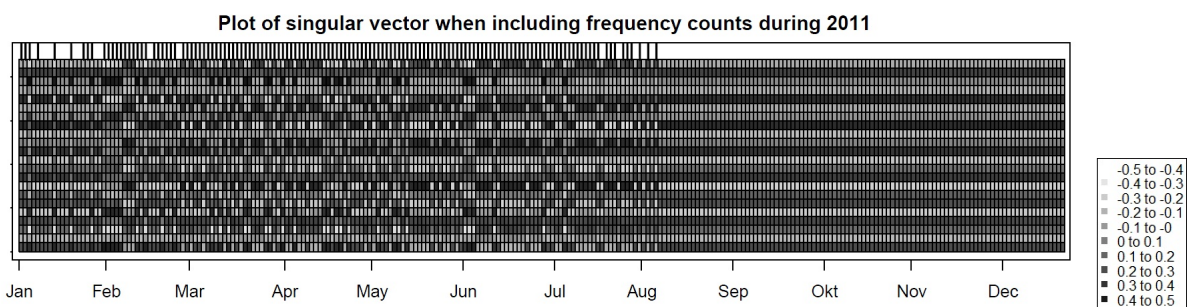


FIGURE 76: This figure represents how the singular vector associated prediction (as calculated with the Bayesian algorithm, while including the auxiliary information at 5% level of significance) changed during 2011.

The tick marks in the above figure shows us that inclusion was only allowed up until late August. There are two possible reasons for this. If we were to reinspect Figure 74, we realize that the primary time series seems to be at some structural break point. The inherent variation in the primary time series seems to increase significantly and singular vectors inevitably changed (even though it is not apparent from Figure 74). Also, at roughly the same time the secondary time series has very significant outlying

observations; this could explain why the auxiliary time series was not included after August. The primary time series showed changes in structure that was not supported by the frequency counts. Notice on the other hand the time points where the auxiliary time series was included. During the beginning of 2011, both time series seemed volatile. Hopefully the inclusion brought about the same effect we had in Chapter 5, where the primary time series became volatile and the auxiliary time series was able to provide information on why the stock prices behave more erratically and on how to react on the erratic behaviour. To investigate if this was the case, we inspect Table 24.

Univariate predictions		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
108.1566	7.1607	48.4127%
HMSSA predictions		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
108.1054	7.0627	53.1746%
BMSSA predictions		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
106.7065	7.0364	55.5556%

TABLE 24: This table shows the measures of accuracy for the baseline Google predictions during 2011 as well as that of the predictions when including data obtained from Topsy using either the HMSSA methodology or the BMSSA methodology at a 5% level of significance.

We can clearly see here that, based on all three measures of accuracy, the Bayesian approach produced predictions more accurate than the baseline predictions. We also included in this table the measures of accuracy when predicting according to the HMSSA methodology (with $L = 22$ and $d = 1$). The HMSSA methodology was able to improve slightly on the univariate predictions; however, the Bayesian approach seems to be considerably superior. Not only did the Bayesian approach decrease the MSE by nearly 1.4% and the MAE by nearly 2.1% but it increased the DOC by 7%. It seems that this inclusion did indeed influence the prediction accuracy and we can conclude that the auxiliary time series was able to provide information with predictive power at times where the primary time series was unstable and incapable of doing so. To investigate whether these changes were significantly superior to our baseline predictions, we can consider Table 25.

<i>Paired t-test on absolute errors</i>	
Including frequency counts at 0.05	
<i>Mean difference</i>	<i>p-value</i>
0.1243	0.0123
<i>Paired t-test on squared errors</i>	
Including frequency counts at 0.05	
<i>Mean difference</i>	<i>p-value</i>
1.4501	0.0664

TABLE 25: This table gives us the difference between the baseline predictions and the Bayesian predictions in terms of difference in the MAE and MSE measurements. The significance of these differences are then also tabulated in the form of p -values obtained from paired Student's t -test.

The p-values in this table are associated with the paired Student's t-tests that tested the significance between the pairs of absolute errors and the pairs of squared errors. According to these p-values, we find that the absolute errors did significantly decrease with a p-value of 0.0123. The test on the difference between the squared errors produced a p-value of 0.0664. This might not be significant at a 5% level of significance, but it is clearly not merely coincidental. Apart from these changes, we must remember that the inclusion of the auxiliary information also increased the DOC with 7%.

In conclusion it seems that the number of Twitter mentions gave us additional knowledge on up and coming news regarding Google. This information was then successfully used in order to better predict Google's closing prices, not only in magnitude but also in direction.

6.2 Predicting Microsoft stock prices

Another company we chose to consider here is Microsoft. This company is a hot topic for all computer fanatics and such fanatics are fond of giving their opinion on Microsoft's newest products or on Bill Gates' personal success. Often these opinions are given on the internet in blog form. Twitter is one such platform on which people can make their opinion known, but it is also a valid form of advertising. Microsoft therefore often advertises new products or new news in the blogosphere. There is therefore reason to believe that by incorporating this information available on Twitter we can improve our predictions.

Our primary time series here is therefore the daily closing prices of Microsoft. The SSA methodology allows us to directly investigate this time series, since it does not require a stationary time series; however, previous studies have shown us that the SSA algorithm produces more accurate predictions when applied to the returns of stock market activity. Cross-validation methods confirmed this. Figure 77 therefore gives us both the actual stock market data of Microsoft, as well as the returns for the time period starting late 2004 and ending 2012 (this data is once again obtained from Yahoo! Finance).

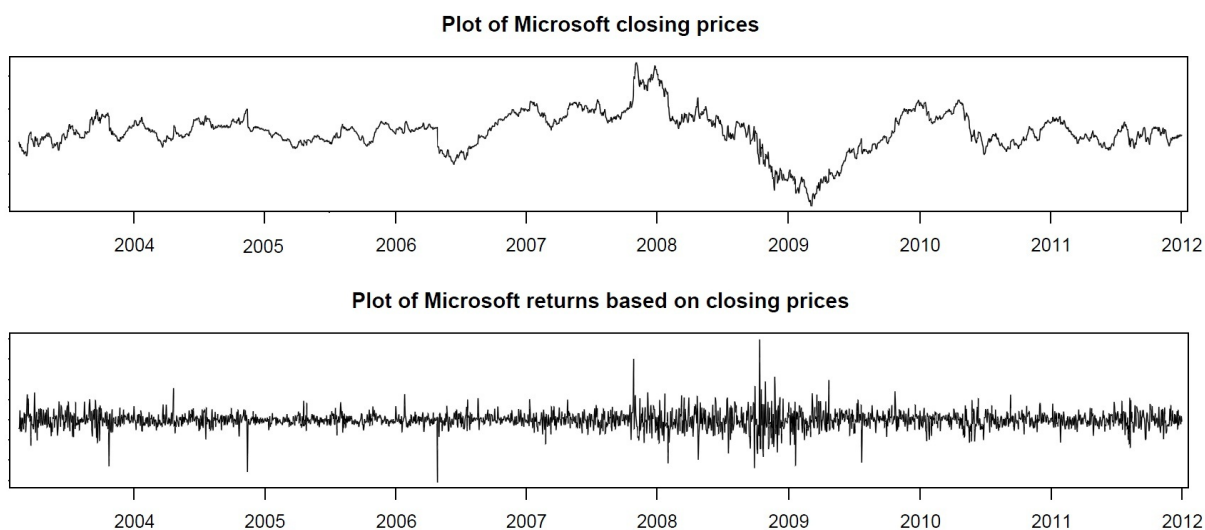


FIGURE 77: These two figures illustrate the primary stock market time series to be predicted. The top figure contains the observed daily Microsoft closing prices and the bottom figure contains the returns of the observed daily Microsoft closing prices.

Once again, we see from the top figure that the 2008 recession took its toll on Microsoft. Thereafter,

steady growth was prominent up until 2010, after which the stock prices seemed to stabilize. When considering the bottom figure, we see that the year 2008 as well as the year 2011 was associated with volatile returns. This volatility in 2011 will be quite significant, since the 2011 observations form our test partition when we apply cross-validation methods to this time series, as before. The cross-validation process has been discussed before. Essentially we partition the time series into a model training set (all observations up until the end of 2009), a cross-validation set (all observations during 2010) and a test set (all observations during 2011). Rolling one-step-ahead predictions are calculated for the cross-validation segment while varying our parameters over the same domain as before ($L \in \{2, \dots, 35\}$ and $d \in \{1, \dots, L - 1\}$). The window length and dimension parameters are then chosen as the pair of parameters that minimizes the MSE of the cross-validation predictions. Not only was the cross-validation method identical to those applied in previous discussions, but the results were also nearly identical. The parameters chosen based on these cross-validation procedures were $d = 1$ and $L = 20$. The test segment was then predicted with rolling one-step-ahead predictions and these predictions formed our baseline predictions without inclusion of the auxiliary internet information. The measures of accuracy associated with these predictions will later be given in Table 26.

In this chapter, we only consider auxiliary information in the form of the number of Tweets posted in a day. Since this example discusses the prediction of the Microsoft stock prices, we propose that the number of times "Microsoft" is mentioned on Twitter, in a single day, could contain information on the current success of Microsoft and therefore possibly also information that could have predictive power on future returns. All Tweets mentioning Microsoft (since 2009) were once again scraped from the Topsy (web-based) database and the number of Tweets on each day was counted. Figure 78 shows this time series. Since our primary time series is a differenced one, Figure 78 also illustrates the differenced number of mentions on each day; this time series will then form our secondary time series.

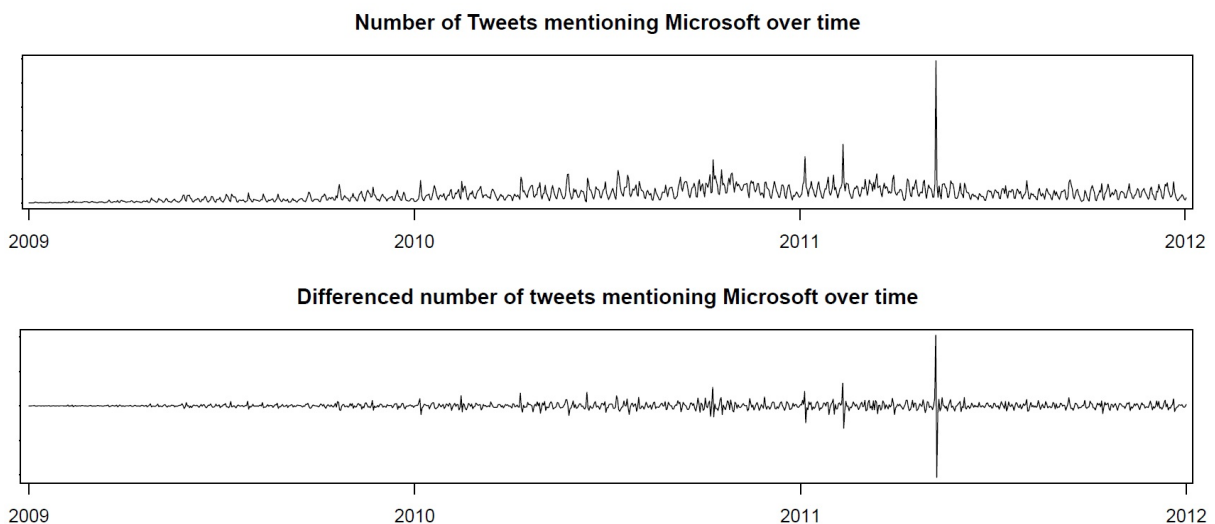


FIGURE 78: These two figures illustrate the auxiliary information obtained from Topsy. The top figure contains the frequency counts of Tweets with keyword "Microsoft" and the bottom figure simply contains the differenced frequency counts after data cleansing as discussed in Chapter 5.

According to Figure 78, the secondary time series seems to show very little variation. This is only a result of the fact that there were more than 10 000 Tweets mentioning Microsoft on the 9th of May, 2011. On this date, Microsoft decided to buy Skype. All other observations seem small in comparison with this outlying observation, even though they often reach values larger than 1000. The variation in

this time series is therefore significantly larger than it seems, but only three outlying observations are prominent, especially when considering the bottom figure. The two observations early in 2011 as well as the very significant observation on 9 May might allow us to identify points where something of note happened regarding Microsoft. These points could represent points where the primary time series might also deviate significantly from what is to be expected.

First it is necessary for us to identify the dimension of the secondary time series in the SSA context. Cross-validation methods quickly showed (as previously) that our auxiliary time series should also be considered to be of low dimension ($d = 1$). After this parameter has also been identified, we can include this auxiliary information by using the Bayesian approach with 250 bootstrap replicates and allowing for inclusion at level of significance of 5%. Figure 79 illustrates the increasing absolute errors of the baseline prediction and how they are affected by the inclusion of the auxiliary information.

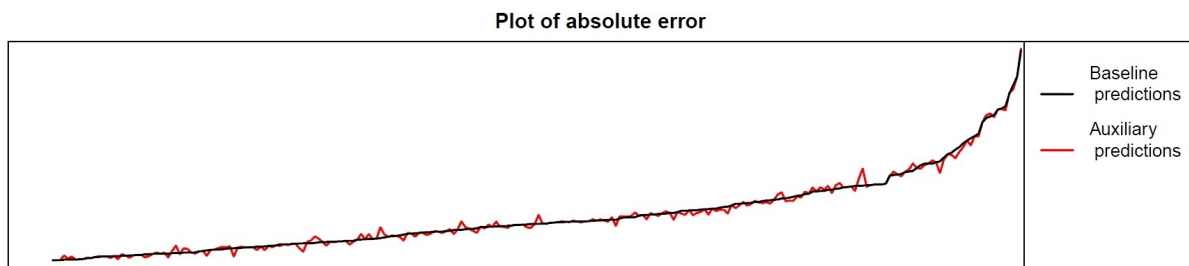


FIGURE 79: This plot illustrates the increasing absolute errors when predicting according to the univariate approach and how the Bayesian predictions (when including at 5% level of significance) affected these absolute errors.

Once again, these changes in absolute error seem slight. It does however seem that some significant changes were made to predictions where the univariate predictions produced residuals very high in magnitude. Before we consider whether these changes were predominantly helpful or predominantly destructive, let us first consider the singular vectors associated with each of these two time series (during the test segment of the data) and how the Bayesian approach combined them.

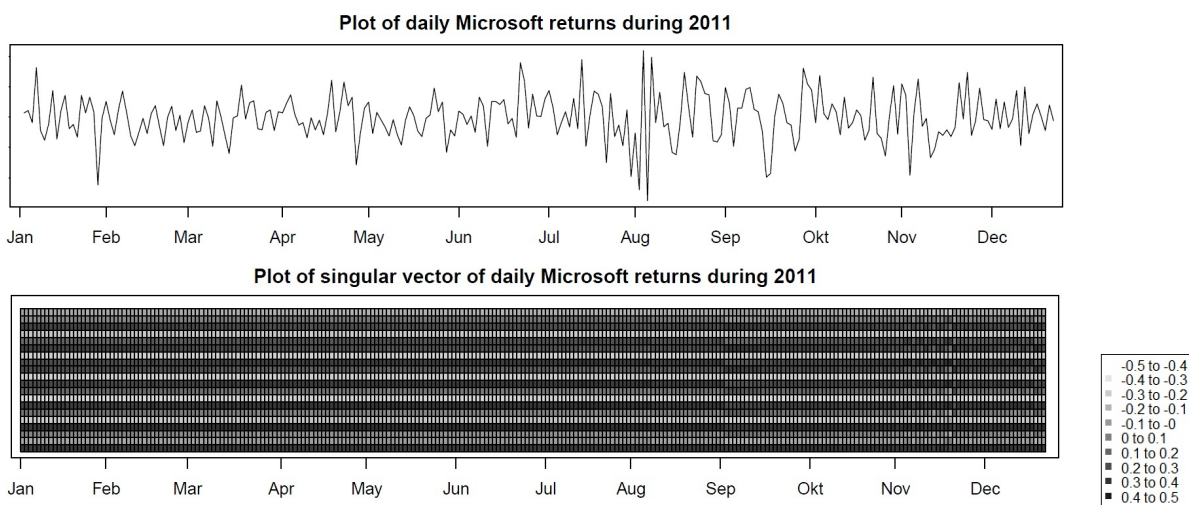


FIGURE 80: These figures represent how the primary time series as well as the associated singular vector changed during 2011. The top figure shows the test segment of the Microsoft stock price returns and the bottom figure illustrates how the singular vector (which determines the predicted value) changed during 2011.

Figure 80 shows that, according to our stock returns, we have very volatile data. There does not seem to be specific structural break points, just general unpredictable behaviour. The same cannot be said for the singular vectors during this time period. There seems to be very little variation associated with these singular vectors, even though the variation associated with the time series it originated from is very prominent, but as we said before, small changes in the singular vector will have much more significant effects on filtering and prediction. The signal associated with these singular vectors on the other hand once again seems to represent some high frequency harmonic component. This high frequency component might be a result of the fact that the primary time series shows erratic changes and very little inherent signal. To see whether the secondary time series perhaps produces similar singular vectors as time goes by, we consider Figure 81.

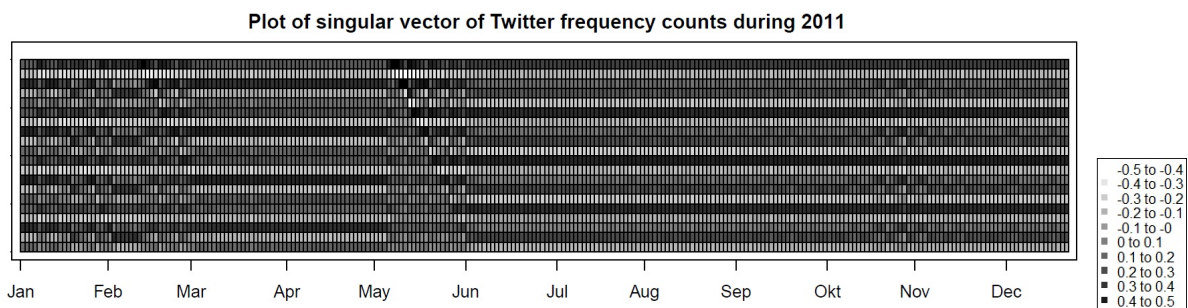


FIGURE 81: This figure represents how the singular vector associated with the secondary time series changed during 2011.

Figure 81 shows that the singular vectors associated with the secondary time series also seem to represent some high frequency harmonic component, similar to that of the primary time series. One significant difference is however the variation in the singular vectors. Between January and March, the singular vectors seem to be a bit unstable. The same is true during the month of May. The reason behind this instability has already been mentioned when we were considering Figure 78. We mentioned 3 significant outlying observations in the auxiliary time series; 2 at the beginning of 2011 and one on 9 May. The effects of these outlying observations on the singular vectors associated with the auxiliary time series are clearly visible in this figure. The question is how these outlying observations (and the effects it had on the singular vector) would be combined with the singular vectors associated with the primary time series. Also, is this combination valid? Figure 82 answers these questions.

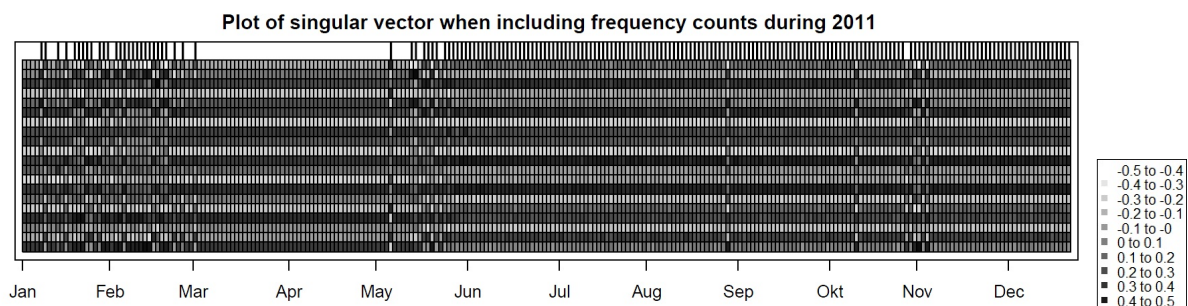


FIGURE 82: This figure represents how the singular vector associated prediction (as calculated with the Bayesian algorithm, while including the auxiliary information at 5% level of significance) changed during 2011.

Upon inspection of Figure 82, the first thing we should realize is the fact that inclusion of the aux-

iliary time series was not always considered to be valid according to the hypothesis test included in the Bayesian algorithm. However inclusion was deemed valid during most of January to March, as well as during most of May. The inclusion of the auxiliary time series during January, February and May is of great importance to us. These were times where the secondary time series produced observations that were significantly different from what we would expect. Hopefully, the Bayesian algorithm not only found a way to make the inclusion of the auxiliary time series valid during these times of uncertainty, but was also able to use this information to its advantage. After June, where both time series seemed to have become stable, inclusion was also allowed. Table 26 tabulates the accuracy measures for the predictions obtained with the univariate prediction method, the HMSSA methodology and the BMSSA methodology. This table allows us to consider whether the inclusion produced improvements in our measures of accuracy.

Univariate predictions		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
0.1473	0.2922	52.3809%
HMSSA predictions		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
0.1467	0.2930	48.8095%
BMSSA predictions		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
0.1462	0.2910	57.5397%

TABLE 26: This table shows the measures of accuracy for the baseline Microsoft predictions during 2011 as well as that of the predictions when including data obtained from Topsy using either the HMSSA methodology or the BMSSA methodology at a 5% level of significance.

Table 26 once again shows that the HMSSA methodology was not of much help. These predictions did improve on the MSE of the baseline predictions, but produced weaker MAE and DOC than the univariate predictions. The Bayesian methodology on the other hand produced predictions that decreased the MAE by 0.4% and also decreased the MSE by 0.7%. These decreases might seem small, but when we consider the fact that the predictions obtained from the Bayesian approach also produced a DOC value 5% higher than that of the baseline predictions, we realize that the inclusion of the internet data produced predictions that were still stable and also slightly better than the univariate predictions (in terms of the magnitude of error) while they increased the DOC by a considerable amount. Even though the decreases in MSE and MAE seem small, we still consider Table 27 to test the significance of these changes by using a paired Student's t-test as before.

<i>Paired t-test on absolute errors</i>	
Including frequency counts at 0.05	
<i>Mean difference</i>	<i>p-value</i>
0.0012	0.1751
<i>Paired t-test on squared errors</i>	
Including frequency counts at 0.05	
<i>Mean difference</i>	<i>p-value</i>
0.0011	0.1613

TABLE 27: This table gives us the difference between the baseline predictions and the Bayesian predictions in terms of difference in the MAE and MSE measurements. The significance of these differences

are then also tabulated in the form of p -values obtained from paired Student's t -test.

This table confirms what we suspected. Even though we did see clear improvements in prediction accuracy, the paired Student's t -test concluded that these differences are insignificant at 5% level of significance. The p -values are indeed larger than 0.05; however, we can still clearly see that they are relatively small. This shows that even though we cannot conclude that we are 95% sure that the frequency counts improve on the prediction accuracy of the baseline predictions, we can still see that the improvements in this example is fairly large. This shows that by including the auxiliary time series, at time points where the secondary time series seemed to have significant information, we did indeed improve on the baseline predictions; not only in terms of the magnitude of the residuals, but also in terms of correctly predicting a positive or negative return.

6.3 Predicting Apple stock prices

The final company we consider is Apple. This company is also highly spoken of in technologically aware circles. With every release of a new Apple product, people jump to the opportunity to post their opinions and reviews on the internet. This relationship between Apple as a company and Apple as interpreted by the internet user, makes Apple the ideal company to consider when attempting to answer the question: Can Twitter frequency counts be used to improve the predictions of a company's share value? Since our primary time series in this discussion will clearly be the time series representing the daily closing prices of Apple, we present Figure 83, illustrating this primary time series, observed since 2006.

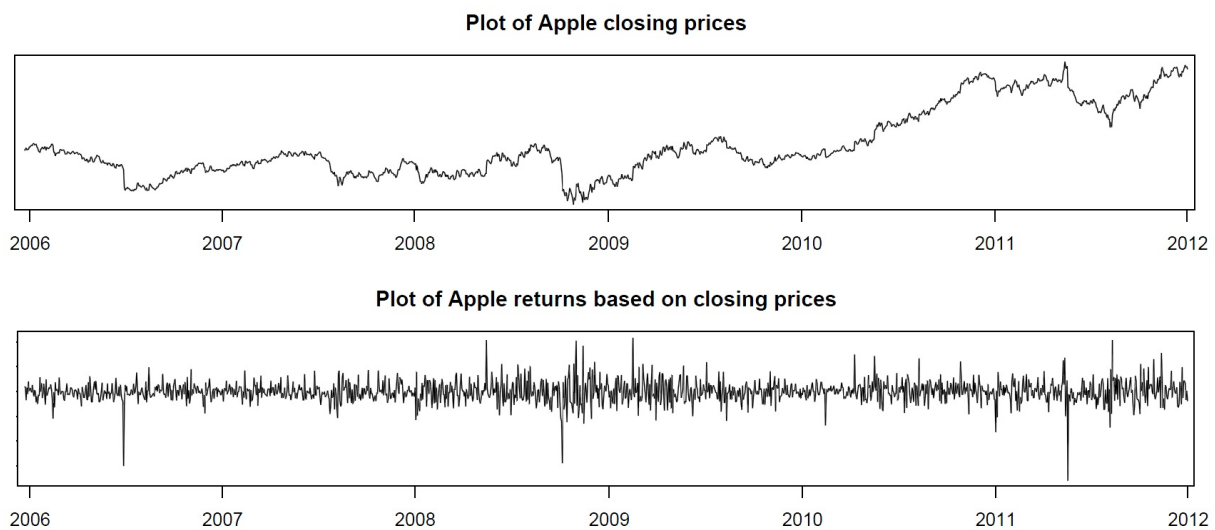


FIGURE 83: These two figures illustrate the primary stock market time series to be predicted. The top figure contains the observed daily Apple closing prices and the bottom figure contains the returns of the observed daily Apple closing prices.

Figure 83 not only gives the actual Apple closing prices, but it also gives the returns of these closing prices. This is because previous examples on stock market activity showed that the SSA algorithm provides more accurate predictions, as well as simpler parameter estimation, when the algorithm is applied to the returns of stock market activity instead of directly to the actual stock values. Apple's closing prices show behaviour slightly different to that of Google and Microsoft. The prices seem less affected by the 2008 recession and quickly thereafter the company once again shows constant growth. In 2011, however, the time series behaves slightly different to that what we would expect. This is clear when con-

sidering either the top or bottom figure. Apple shares seem to drop after several years of constant growth.

Using the time series given in Figure 83, we can consider the next step in our prediction process: the estimation of the window length and dimension parameters to be used through cross-validation methods identical to previous examples in this thesis. Rolling one-step-ahead predictions on the observed returns during 2010 (cross-validation set) are calculated and show that predictions are most accurate when we choose $L = 26$ and $d = 1$. With parameters as estimated through cross-validation, we can find the baseline predictions of the returns during 2011 (test segment). Results for these baseline predictions will be shown later in Table 28. The auxiliary information we wish to include in this discussion is once again the number of times Apple was mentioned in Tweets. All Tweets (since 2009) referring to Apple was once again scraped from Topsy. The daily frequency counts were observed. This time series of daily frequency counts as well as the differenced time series are illustrated in Figure 84.

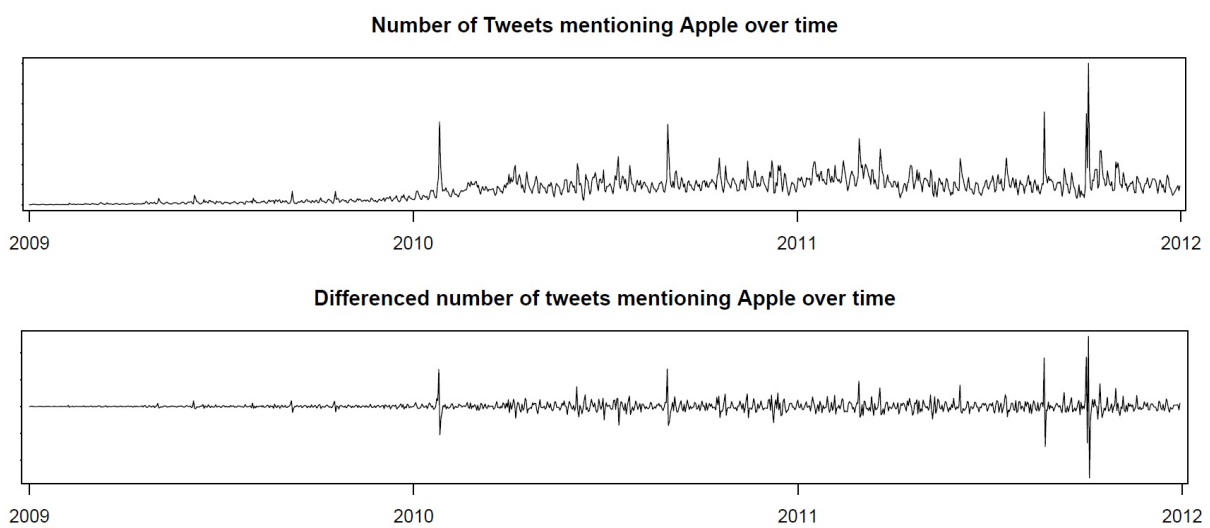


FIGURE 84: These two figures illustrate the auxiliary information obtained from Topsy. The top figure contains the frequency counts of Tweets with keyword "Apple" and the bottom figure simply contains the differenced frequency counts after data cleansing as discussed in Chapter 5.

The above time series clearly have two very significant outlying observations during the test segment of the data. Upon closer inspection of the Tweets, we see that the first significant observation was taken during August 2011. On this day, Apple won a preliminary trial against HTC wherein Apple sued HTC for patent infringements. Shortly thereafter (November 2011), Apple announced that they will also sue Samsung based on similar patent infringements. We already saw that the primary time series behaved slightly differently to what we would expect of it during 2011. Perhaps this additional information, as provided by these Tweets will allow us to more accurately predict the returns of Apple shares. Figure 85 allows us to graphically study what the effect of the auxiliary information would be if we were to include it by using the Bayesian prediction algorithm.

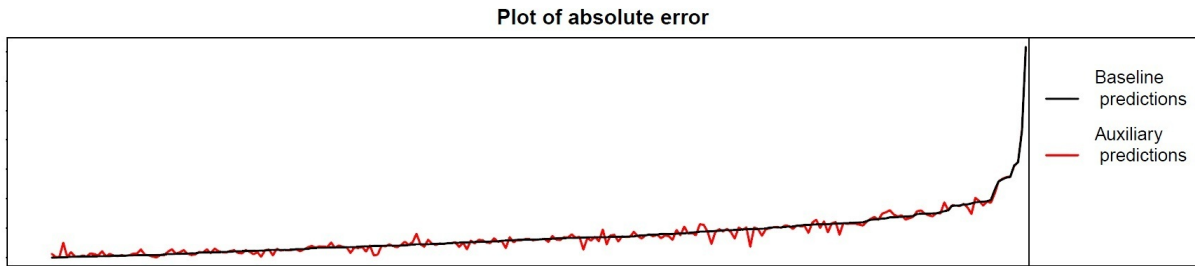


FIGURE 85: This plot illustrates the increasing absolute errors when predicting according to the univariate approach and how the Bayesian predictions (when including at 5% level of significance) affected these absolute errors.

Clearly the inclusion of the auxiliary time series did not seem to change the baseline predictions by significant amounts. It seems that most of the adjustments provided very slight decreases in absolute error. We will discuss the significance of the adjustments later in this example, but let us first consider the nature of the singular vectors and how they were affected by the inclusion of the auxiliary information.

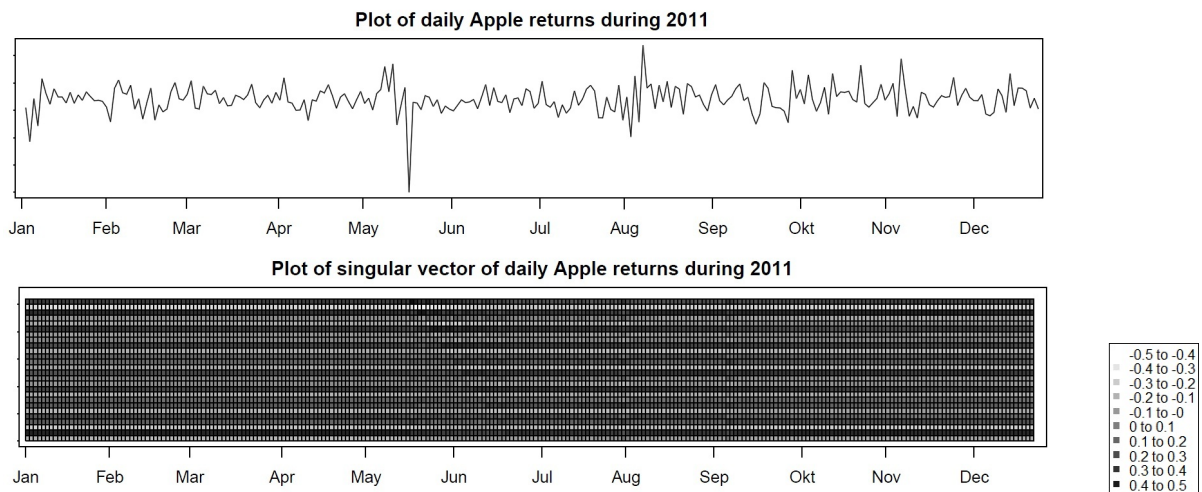


FIGURE 86: These figures represent how the primary time series as well as the associated singular vector changed during 2011. The top figure shows the test segment of the Apple stock price returns and the bottom figure illustrates how the singular vector (which determines the predicted value) changed during 2011.

Once again, we see in this figure that the singular vectors associated with the primary time series are of such a nature that they seem to describe a high frequency harmonic component (possibly because of the lack of significant signal in the time series). Furthermore, there seems to be very little variation in the singular vector during 2011, even though the primary time series itself is very volatile. The same can not be said for the significant singular vectors obtained from the auxiliary time series (once again cross-validation methods showed that only one singular vector was significant).

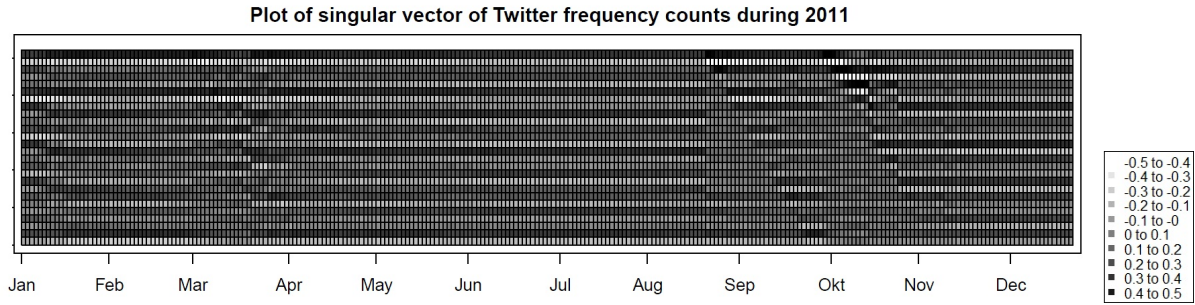


FIGURE 87: This figure shows how the singular vector associated with the secondary time series changed during 2011.

Figure 87 clearly shows that there are some variation associated with the singular vectors during March and between August and November. We already mentioned that Apple was involved in lawsuits during August to November; this is clearly the reason for the variation in the singular vectors during that time interval. Upon inspection of the scraped Tweets, we see that the reason for the structural break in early March was because of the fact that Apple launched the Ipad 2 on 2 March. This then clearly caused some hype in the blogosphere, which is now visible in singular vectors illustrated in Figure 87. Combining these singular vectors with those in Figure 86, using the Bayesian approach, results in prediction vectors as given in Figure 88.

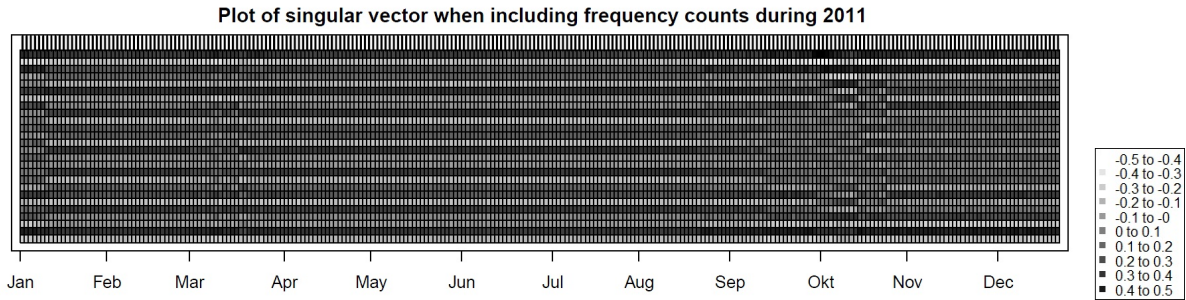


FIGURE 88: This figure shows how the singular vector associated prediction (as calculated with the Bayesian algorithm, while including the auxiliary information at 5% level of significance) changed during 2011.

The first thing we notice from this figure is the fact that the Bayesian algorithm allowed for inclusion at each step. This is of course a result of the similarity between the two sets of singular vectors. Notice also that some of the variation that was present in the singular vector of the secondary time series is also now present in these prediction vectors. Hopefully these structural changes in the singular vectors, during times where internet users felt the need to mention Apple, can resemble current news on the company to such an extent that it will improve significantly on the baseline predictions. Figure 85 already graphically illustrated the effect this inclusion had on the predictions to some extent. In Table 28, we tabulate the differences between the two sets of predictions with numerical measures of accuracy.

Univariate predictions		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
1.1607	0.7733	50.3969%
Including AFI		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
1.1591	0.7725	49.2064%
Including AI		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
1.1519	0.7671	49.2064%

TABLE 28: This table shows the measures of accuracy for the baseline Apple predictions during 2011 as well as that of the predictions when including data obtained from Topsy using either the HMSSA methodology or the BMSSA methodology at a 5% level of significance.

This table, for the first time, shows less than desirable results. Even though the Bayesian approach did improve on both the univariate predictions as well as the HMSSA predictions, it is worrisome that the DOC associated with the BMSSA predictions are now less than the desired 50%. The magnitude of the error therefore slightly decreased, but at the expense of correctly classifying the direction of the movement less often. To see whether the decrease in the absolute errors and the squared errors is significant, we consider Table 29.

<i>Paired t-test on absolute errors</i>	
Including frequency counts at 0.05	
<i>Mean difference</i>	<i>p-value</i>
0.0062	0.2053
<i>Paired t-test on squared errors</i>	
Including frequency counts at 0.05	
<i>Mean difference</i>	<i>p-value</i>
0.0089	0.2560

TABLE 29: This table gives us the difference between the baseline predictions and the Bayesian predictions in terms of difference in the MAE and MSE measurements. The significance of these differences are then also tabulated in the form of *p*-values obtained from paired Student's *t*-test.

This table once again shows that the decreases in MSE and MAE are not significant at 95% level of certainty. However, we can once again take note of the fact that the *p*-values are relatively small, implying that even though these decreases are not significant, they do seem to be considerable. The more disappointing result in this example is the fact that the adjusted predictions produced a DOC lower than 50%. Taking the nature of the auxiliary time series into consideration, we can perhaps gain insight into why the predictions were unable to correctly predict the direction of change here. The auxiliary time series merely contains frequency counts. By using these frequency counts as auxiliary information, we are effectively assuming that all news is good news, i.e. we believe that an unexpectedly high observation in the auxiliary time series would cause a positive return in the primary time series to be more likely. However, we found here that large frequency counts were often associated with law suites. This is certainly not good news for investors. The following chapter attempts to address this inadequacy related to using frequency counts as auxiliary information.

7 Incorporating Twitter sentiment

The previous chapter was based on the assumption that any news is good news. This assumption seemed valid when considering the stock prices of Google and Microsoft. However, since Apple went through some unpleasant legal processes, the large amounts of Tweets regarding Apple was not necessarily good news. This chapter therefore tries to rectify this assumption by not merely counting the number of Tweets relating to a certain company, but trying to incorporate the sentiment of the text posted on Twitter. By doing so, we would perhaps be able to distinguish between positive opinions (implying positive future prospects) and negative opinions (implying negative future prospects).

Sentiment classification in text is an entire study on its own. Numerous approaches can be used to either classify text into positive or negative sentiment or into one of several different emotional classes. There are several different problems with sentiment classifiers. Firstly, the accuracy of simple classifiers are often weak; more complex classifiers on the other hand can be relatively accurate, but often increase computational time to an impractical extent. Also, these sentiment classifiers are often based on a training data set. Such training data sets are available; however, for the classifier to be accurate, it is recommended that the training data should contain text relating to that of the test data (in terms of topic and field) and also written in similar style. Such topical data sets are hard to find and timely to create. Another problem with sentiment classifiers is that they often provide weak classifications when words are misspelt or when slang words are used. Tweets often contain both spelling errors as well as slang words.

Considering all of these problems associated with text classifiers, it might seem that it would be impractical to subject these Tweets to sentiment classifiers. However, we do have one significant advantage when using Tweets; that is the fact that we have hundreds and thousands of Tweets in a single day. By using a simple sentiment classifier (that is computationally inexpensive) we can classify text into one of several categories. The accuracy of the individual classifications is less important, since by combining large amounts of these classifications, we are bound to obtain a statistic that accurately measures the opinion of the internet user on a given day.

In this section we will therefore use the simple sentiment classifier included in the R package called *sentiment* (<http://cran.r-project.org/web/package/sentiment/index.html>). This package can classify text as either positive or negative by subjecting the text to a naive Bayes classifier that has been trained on the Janyce Wiebe's subjectivity lexicon. The classifier gives a score for each of the polarity classes and then classifies to the class with the highest score (interested reader can see the above url for more information on this classification method). This is definitely not the most accurate or most effective classifier. However, the fact that it can classify relatively quickly and that it has already been trained, made it the obvious choice. Furthermore, the accuracy of the classifier was less important because of the large number of daily Tweets that it was applied to and for the same reason, the time it took to classify was much more important since it would need to classify more than 3 million Tweets.

This chosen polarity classifier is then applied to the same Tweets we used in Chapter 6. For each day, we then calculated the sum total of all the positive scores and subtracted the sum total of all the negative scores. The calculated number then showed us the general sentiment of that day's Tweets, with a positive number implying positive sentiment for that day and a negative number implying negative sentiment for that day. This chapter proposes that this sentiment aware auxiliary information could produce even more accurate predictions than the simple frequency counts used in Chapter 6. Notice that in all three

of these examples, we saw that the HMSSA algorithm was never competitive with the proposed Bayesian algorithm. The HMSSA approach is therefore omitted in these examples.

7.1 Predicting Google stock prices

In this first example, we try to show the validity of incorporating sentiment associated with Tweets instead of amount of Tweets. We therefore revisit the Google example studied in Chapter 6.1. We therefore use here the same primary time series of daily closing prices of Google. We saw that this time series was very volatile and cross-validation methods showed that we should analyze the time series with $L = 22$ and $d = 1$, in the SSA context. Singular vectors associated with this time series however seemed stable and showed little variation. When including frequency counts, we saw that the auxiliary information was also quite stable. We mentioned how the auxiliary data was able to identify important time points and use these time points to improve adequately on baseline predictions in terms of decreasing all three prediction accuracy measurements.

The auxiliary time series in this example on the other hand is different from that of Chapter 6.1. Instead of merely counting the number of Tweets mentioning Google, we now try to obtain additional information from these Tweets by classifying them as either positive or negative. Each Tweet receives a negative score and a positive score (calculated by the naive Bayes classifier in the *sentiment* package in R). For each day we obtain the total positive score and subtract the total negative score. This value then represents the overall sentiment regarding Google on that day. Figure 89 shows this time series of daily sentiment scores.

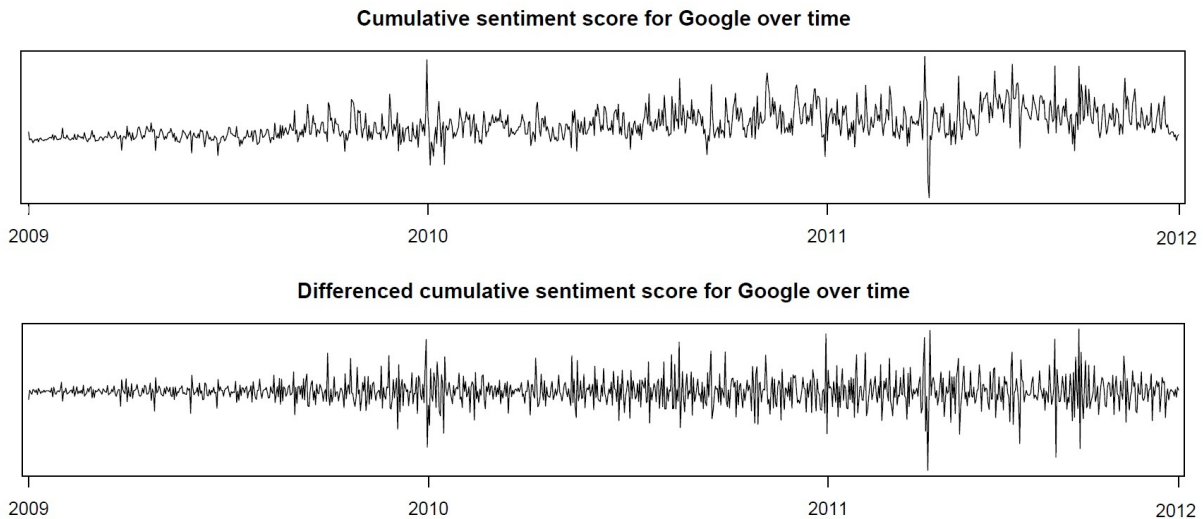


FIGURE 89: These two figures illustrate the auxiliary information obtained from Topsy. The top figure contains the daily sentiment scores for Tweets with the keyword "Google" and the bottom figure simply contains the differenced sentiment scores after data cleaning.

Figure 89 not only shows the daily sentiment scores, but the bottom figure also shows the differenced sentiment score observations (after ensuring that both primary and secondary time series were taken on the same dates). This time series (illustrated in the bottom figure) will be used as auxiliary information when attempting to predict the primary time series. Parts of this auxiliary time series resemble the auxiliary time series in Chapter 6.1. For instance there seems to be an increase just before Google launches the Project Glass and when Google bought Motorola Mobility. However, this time series clearly

has significantly more noise than the auxiliary time series in Chapter 6.1. Cross-validation studies once again showed that we have little signal in this time series and that we should assume that this time series also only has one significant singular vector. This significant singular vector (and how it changes during the test segment of our time series) can be considered in Figure 90.

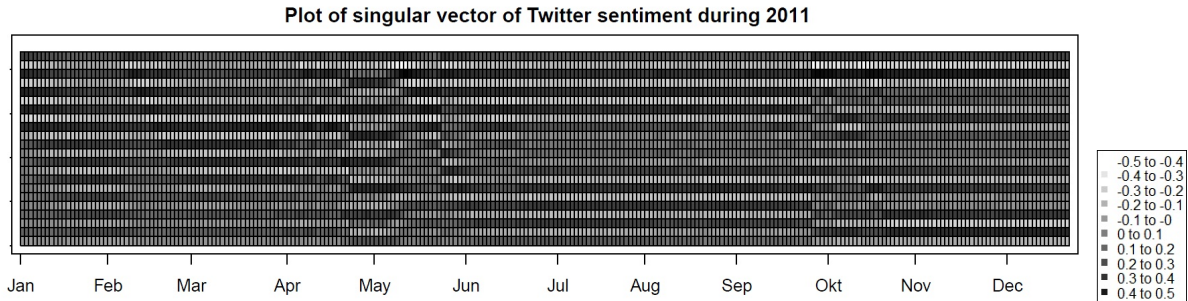


FIGURE 90: This figure represents how the singular vector associated with the secondary time series changed during 2011.

The increase in noise visible in the primary time series is clearly also visible in Figure 90. The singular vectors show much more variation than it did when considering frequency counts. This additional variation could merely be a consequence of inaccurate sentiment classifications and it might then not be helpful in prediction. However, if this variation is indicative of the public’s general feeling towards Google, it could contain information with predictive power. If we were to combine, at each prediction step, these additional singular vectors with the singular vectors associated with the primary time series (see Figure 74), by using the BMSSA approach while including information at 5% level of significance, we would obtain prediction vectors as illustrated in Figure 91.

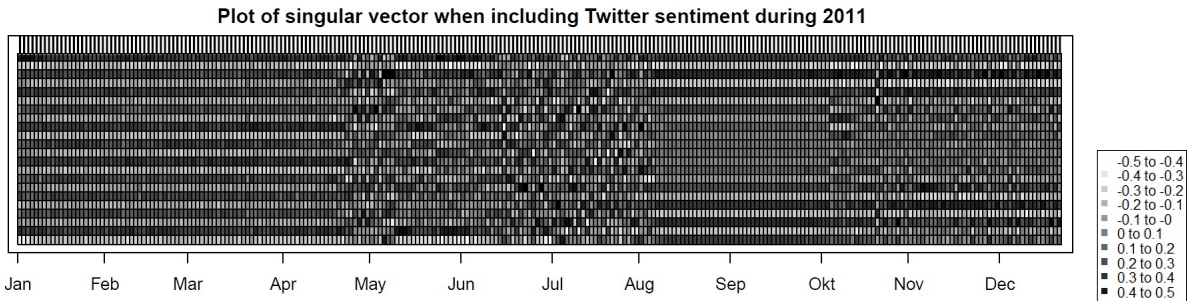


FIGURE 91: This figure represents how the singular vector associated prediction (as calculated with the Bayesian algorithm, while including the auxiliary information at 5% level of significance) changed during 2011.

The tick marks at the top of the figure once again show that the algorithm allowed for inclusion at each prediction point. As a result of this, we see that the prediction vectors during May, June and July were quite unstable. The reason for this instability was the inclusion of the auxiliary information during said time periods. This auxiliary time series itself had some significant variation as a result of slight, unexpected changes in the sentiment regarding Google. The inclusion of this information was deemed valid according to the hypothesis test included in the Bayesian algorithm, but to assess whether the inclusion of the information was actually helpful to prediction, we consider Figure 92.

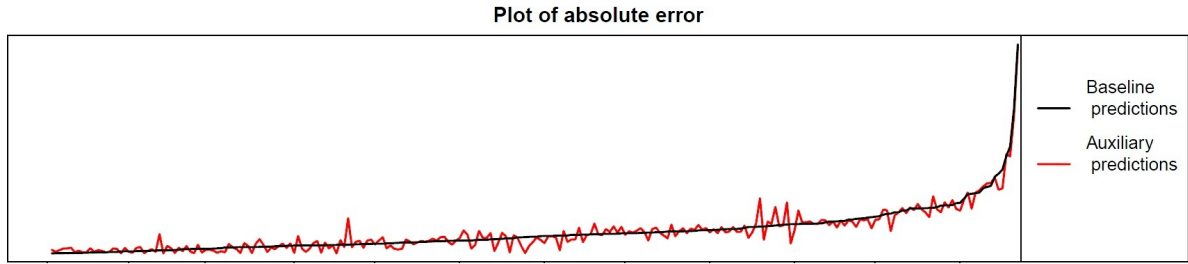


FIGURE 92: This plot illustrates the increasing absolute errors when predicting according to the univariate approach and how the Bayesian predictions (when including at 5% level of significance) affected these absolute errors.

As before, it is difficult to see the effect inclusion had on the prediction errors when considering this figure. It seems that the inclusion, in general, had a positive effect on the prediction errors; however, this graphical representation of the difference between the two sets of predictions is not clear enough. We therefore also consider Table 30.

Univariate predictions		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
108.1566	7.0731	48.4127%
Predictions including frequency counts		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
106.7065	7.036	55.5556%
Predictions including sentiment scores		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
105.3710	7.0621	57.5397%

TABLE 30: This table shows the measures of accuracy for the baseline Apple predictions during 2011 as well as that of the predictions when including data obtained from Topsy using either the frequency counts or sentiment scores.

Table 30 clearly shows that the predictions when including the sentiment scores not only improve uniformly on the univariate predictions, but also on the MSE of predictions when including frequency counts. More importantly, when considering DOC, we see that these predictions produce an increase of 10% relative to the univariate predictions. We can consider Table 31 to verify whether these predictions are significantly better, in terms of magnitude of error, than the univariate predictions.

<i>Paired t-test on absolute errors</i>	
Including frequency counts at 0.05	
<i>Mean difference</i>	<i>p-value</i>
0.0986	0.2210
<i>Paired t-test on squared errors</i>	
Including frequency counts at 0.05	
<i>Mean difference</i>	<i>p-value</i>
2.7863	0.1854

TABLE 31: This table gives us the difference between the baseline predictions and the Bayesian predictions in terms of difference in the MAE and MSE measurements. The significance of these differences are then also tabulated in the form of p-values obtained from paired Student's t-test.

Once again unfortunately, the p-values indicate that these improvements are insignificant. Even though the improvements are large in magnitude, when we reconsider Figure 92, we see that these differences also vary much more. This variation is the reason for these p-values being higher than we expected. This being said, once again, we have to keep in mind that the inclusion of the auxiliary information considerably decreased MSE and MAE while increasing the DOC by a very significant 10%. We also saw in Table 30, that by adding sentiment to our auxiliary information, we increased on most measures of prediction accuracy. Therefore, in conclusion, even though the effects of including sentiment scores are not significant they did provide us with predictions that are superior to the other prediction methods discussed.

7.2 Predicting Microsoft stock prices

The Microsoft example can also be reconsidered. Chapter 6.2 showed us that when incorporating frequency counts of Tweets mentioning Microsoft, we obtained predictions that were uniformly more accurate than the univariate predictions. The frequency counts identified some significant time points relating to Microsoft and used these time points to gain insight into current and perhaps future prospects of the company, but this auxiliary information did not also include the sentiment regarding the events on these significant days. This chapter proposes that if we were to incorporate this sentiment in our prediction, that we might obtain even more accurate predictions.

Our primary time series here is once again the Microsoft closing prices and returns (Figure 77). Cross-validation studies on this time series showed that we should subject it to the SSA algorithm with parameters $L = 20$ and $d = 1$. In order to obtain our auxiliary time series, we use the same Tweets mentioning Microsoft that we obtained from Topsy. However, this time, each Tweet is subjected to the sentiment classifier and given a score proportional to the likelihood that the text is positive and a score proportional to the likelihood that it is negative. On each day, we add all the positive scores and subtract all the negative scores in order to obtain a figure representing the sentiment for that specific day. This time series is given in Figure 93.

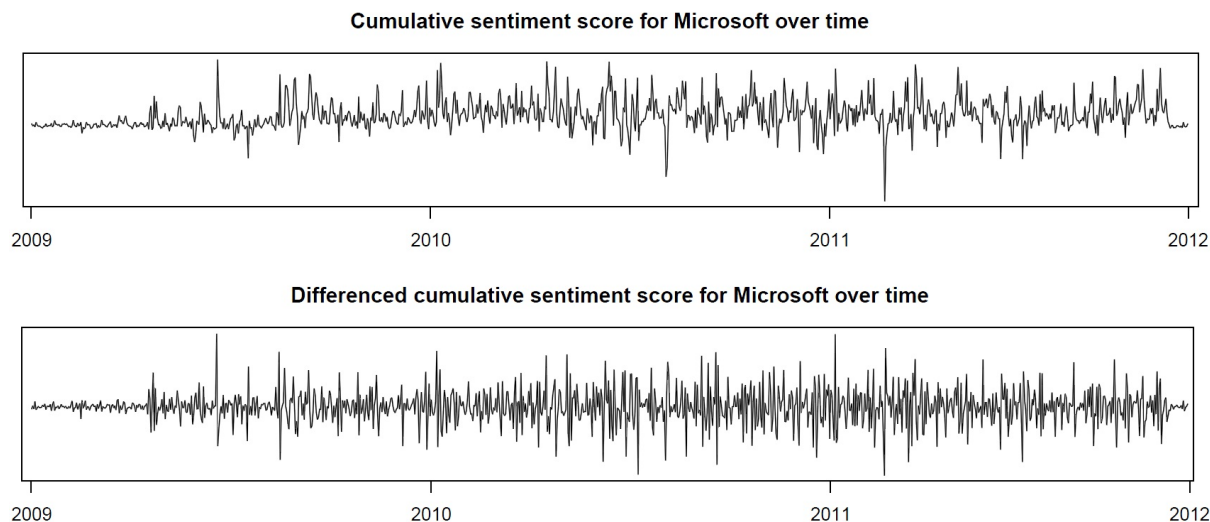


FIGURE 93: These two figures illustrate the auxiliary information obtained from Topsy. The top figure contains the daily sentiment scores for Tweets with the keyword "Microsoft" and the bottom figure simply contains the differenced sentiment scores after data cleaning.

Notice in the top figure that there is much more variation in the auxiliary time series than we had in Chapter 6.2. This was beneficial for prediction in the previous example (Chapter 7.1). However, here we also notice something else. When we merely considered frequency counts, there was a very significant outlying observation in May 2011, on the day Microsoft bought Skype. In this time series, this significant observation disappeared. This is reason for concern. Clearly such a big investment would affect the value of Microsoft shares; however, according to this time series of sentiment scores, the date of this purchase was insignificant. This phenomenon is difficult to explain. An intuitive reason for this could simply be inaccurate classification of sentiment as a result of using a simple classifier. However, I believe that the reason could also be inaccurate classification due to people using negative words to describe Skype, rather than positive words describing Microsoft. For example, suppose we have the following Tweet: "Skype has always been giving me problems. Maybe Microsoft will bring them out of the gutter". Such a Tweet will be interpreted very negatively and will be associated with Microsoft even though it should not. When manually considering the Tweets, numerous such examples were observed. Another possible explanation of course is large amounts of neutral posts, perhaps merely containing a link to some article about Microsoft.

Despite the fact that this auxiliary information seems to be inaccurate, we continue the study to see whether the information might not still be able to improve on baseline predictions. The singular vectors associated with this secondary time series (during the test partition of the data set) are illustrated in Figure 94.

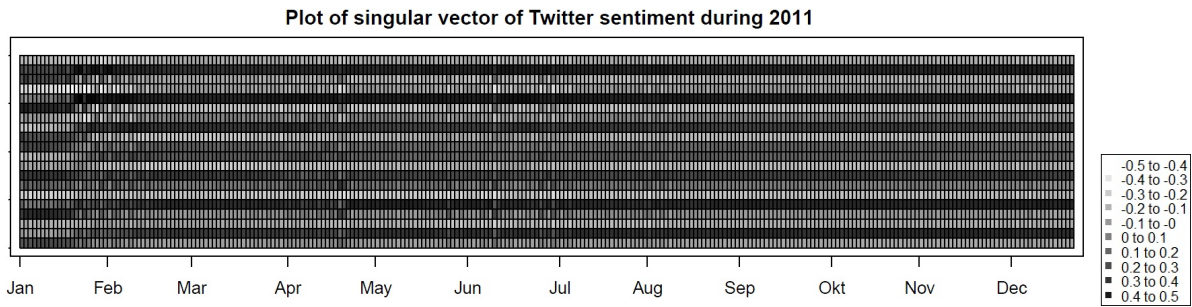


FIGURE 94: This figure represents how the singular vector associated with the secondary time series changed during 2011.

The singular vectors in the above figure seem to have some variation. Based on the primary time series we might expect to see more variation, but as we mentioned before, slight variation in the singular vector would have significant effects in both filtering and prediction. Another thing we can conclude from Figure 94 is the fact that there does not seem to be any clear points indicating some form of structural break. The general form of the singular vector seems to be quite constant even though there is a reasonable amount of variation around this relatively constant signal. Combining these singular vectors with those given in Figure 77 would provide prediction vectors as given by Figure 95.

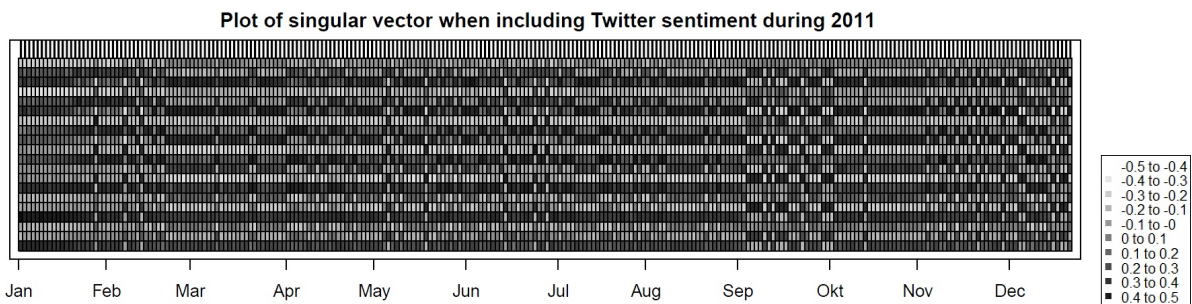


FIGURE 95: This figure represents how the singular vector associated prediction (as calculated with the Bayesian algorithm, while including the auxiliary information at 5% level of significance) changed during 2011.

The signal associated with these singular vectors seems to change frequently. It is of course possible that these changes accurately reflect future prospects of the Microsoft returns. However, these changes occur quite often and it would seem that the inclusion of the high variation auxiliary singular vectors merely caused variation in the produced prediction vectors. One would hope that the hypothesis test would reject inclusion in order to protect the estimated prediction vectors from unwanted deviations such as these; however, inclusion was allowed at each prediction step. This might simply be a consequence of the fact that both the primary and secondary time series have significant amounts of noise. To investigate whether these predictions are accurate, we consider a graphical representation of the absolute errors and how they were affected by the inclusion of the sentiment scores.

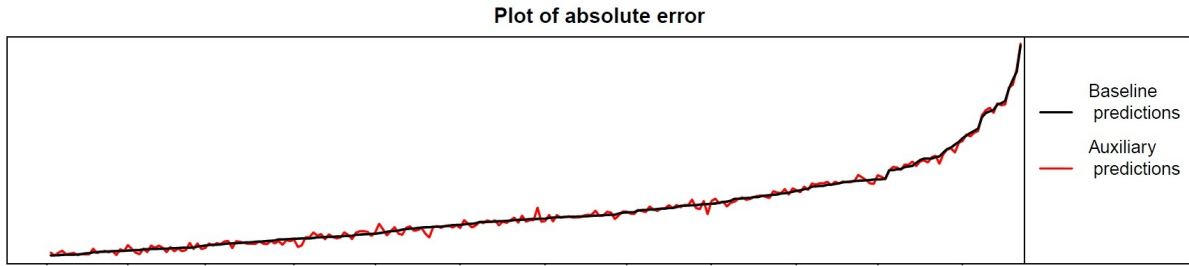


FIGURE 96: This plot illustrates the increasing absolute errors when predicting according to the univariate approach and how the Bayesian predictions (when including at 5% level of significance) affected these absolute errors.

We see here that even though the singular vector changed dramatically quite often, the effect of these changes on predictions were small. Slight deviations from the baseline absolute errors are visible, but nothing drastic. Based on this figure we are unable to state with confidence which of these prediction methods are better and we therefore consider Table 32.

Univariate predictions		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
0.1473	0.2922	52.3810%
Predictions including frequency counts		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
0.1462	0.2910	57.5397%
Predictions including sentiment scores		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
0.1471	0.2917	55.1587%

TABLE 32: This table shows the measures of accuracy for the baseline Microsoft predictions during 2011 as well as that of the predictions when including data obtained from Topsy using either the frequency counts or sentiment scores.

This table clearly shows us that the predictions produced when including sentiment scores did indeed improve on the baseline predictions. Both the MSE and MAE measurements decreased by small amounts and also, the DOC increased to 55%. Clearly, as Table 33 illustrates, these improvements are small and not significant. Another unfortunate result is the fact that inclusion of frequency counts proved to be even more helpful than inclusion of sentiment.

<i>Paired t-test on absolute errors</i>	
Including frequency counts at 0.05	
<i>Mean difference</i>	<i>p-value</i>
0.0005	0.3272
<i>Paired t-test on squared errors</i>	
Including frequency counts at 0.05	
<i>Mean difference</i>	<i>p-value</i>
0.0002	0.4366

TABLE 33: This table gives us the difference between the baseline predictions and the Bayesian predictions in terms of difference in the MAE and MSE measurements. The significance of these differences are then also tabulated in the form of p -values obtained from paired Student's t -test.

These results are disappointing since the predictions based on sentiment scores were not able to significantly improve on the baseline predictions nor were they able to improve at all on the predictions including mere frequency counts. I believe that the reason for this is the weak sentiment classifier. The impact of this classifier was not as drastic in other examples, but here it seemed to have a larger effect on the predictions of the Microsoft returns. Consolation can be found in the fact that internet activity (in the form of frequency counts) were able to produce favourable results in Chapter 6.2. Perhaps if we were to subject these Tweets to better sentiment classifiers, the secondary time series will be less erratic and consequently able to improve on predictions. Furthermore, we should once again remember that even though we are unsatisfied with the results here, we once again see that with the inclusion of sentiment scores, we were able to decrease the magnitude of the error of prediction while pushing the DOC to above 55%. Even though the results are not significant at a 5% level of significance, the results are still good.

7.3 Predicting Apple stock prices

In this final practical example, we revisit the prediction of Apple's daily closing prices we discussed in Chapter 6.3. In our previous discussion on these closing prices, we obtained the only set of predictions that delivered a DOC of less than 50%, even though this measure was only slightly higher in the univariate predictions and the MSE and MAE measurements showed improvements. These were unsatisfying results. Nevertheless, we contemplated that the possible reason for these weak results could be the inherent assumption associated with using frequency counts as auxiliary information; that is, the assumption that all news is good news. This entire chapter specifically addresses this assumption by trying to incorporate sentiment of text into predictions.

The primary time series in this example is therefore identical to the primary time series used in Chapter 6.3 and illustrated in Figure 83. The secondary time series is obtained from the same set of Tweets scraped in the previous discussion; however, instead of merely counting the number of times Apple was mentioned, we assigned a positive sentiment score and a negative sentiment score to each of these Tweets. The daily secondary time series observations was then obtained by taking the total positive score of that day and subtracting the total negative score of that day. These values (illustrated in Figure 97) gives an indication of the public's general feeling towards Apple.

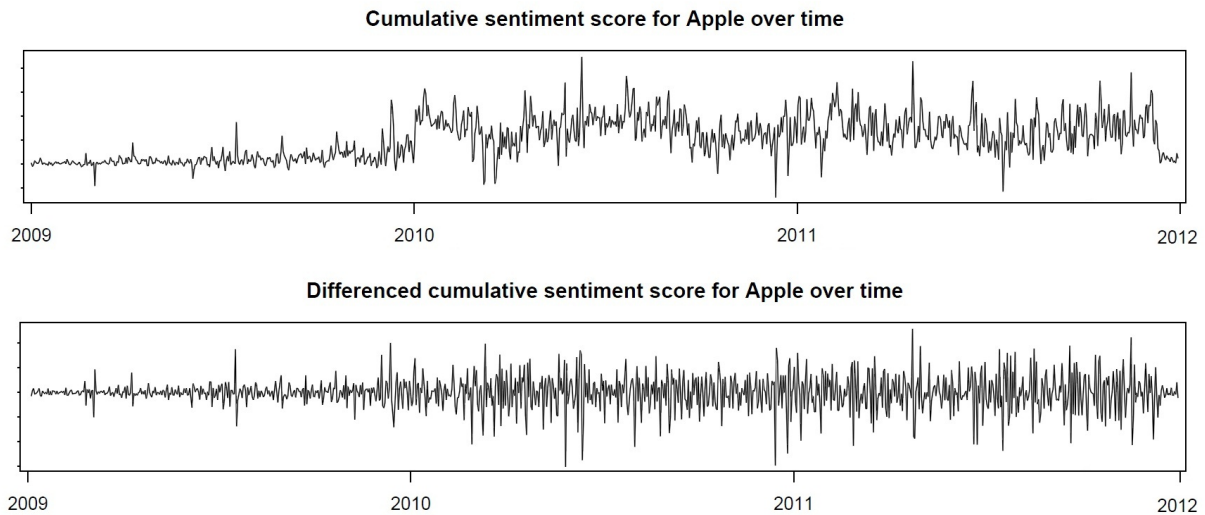


FIGURE 97: These two figures illustrate the auxiliary information obtained from Topsy. The top figure contains the daily sentiment scores for Tweets with the keyword "Apple" and the bottom figure simply contains the differenced sentiment scores after data cleaning.

Notice once again that the bottom figure represents the differenced time series (after data cleaning procedures) that we will actually be using as secondary time series. In this time series there still seems to be a considerably large (positive) observation in August (where Apple won the law suite against HTC). This significant date is therefore not only apparent in the frequency counts time series, but the sentiment classifier was accurate enough to have a similar observation when considering the sentiment scores. The other significant date in the frequency counts time series (November) is not as apparent in this sentiment time series. This is simply because typical Tweets regarding the start of a law suit against Samsung were factual rather than positive or negative. The aggregate sentiment for the day was therefore not as evident as the number of Tweets. In order to see the effect of sentiment classifier on the singular vector associated with the internet activity, we consider Figure 98.

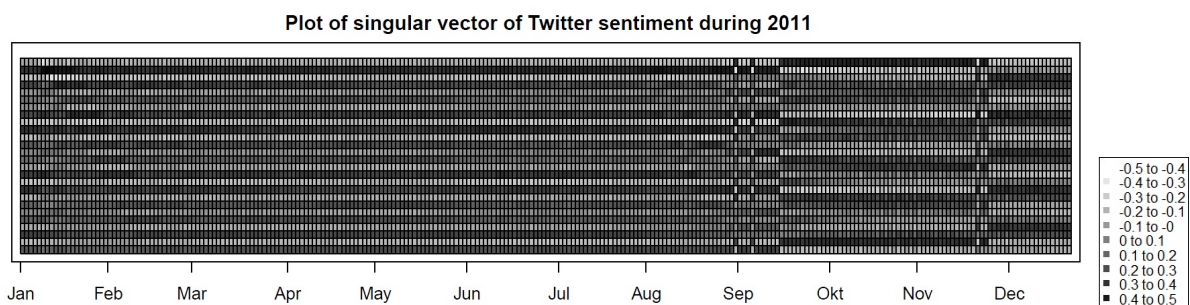


FIGURE 98: This figure represents how the singular vector associated with the secondary time series changed during 2011.

These singular vectors seem to change slightly quite often over time. During September there seems to be quite a prominent structural break. However, we must remember that a singular vector is only unique up to multiplication of negative one. If we were to multiply the singular vectors from middle September to middle November with negative one, we would see that there is no structural break. We must also take note of the fact that these singular vectors do not show signs of prominent structural breaks as we had in Chapter 6.3. This is because the sentiment analysis incorporated the concept that not all Tweets are in praise of Apple. One can therefore consider these singular vectors to merely notice

slight changes in the public’s feeling towards Apple. These slight changes in the singular vectors could result in significant changes in the predictions. If we were to combine these singular vectors with that obtained from the primary time series with the proposed Bayesian approach (at a 5% level of significance and 250 bootstrap replicates), we would obtain prediction vectors as given in Figure 99.

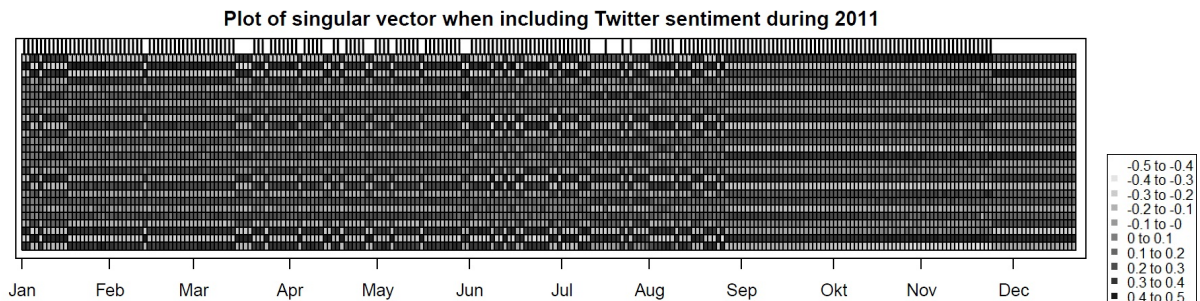


FIGURE 99: This figure represents how the singular vector associated prediction (as calculated with the Bayesian algorithm, while including the auxiliary information at 5% level of significance) changed during 2011.

We see by the ticks in the above figure that inclusion was not always allowed. At certain time points, the hypothesis test rejected the possibility of including the auxiliary information. This is mainly a result of inherent variation. If the variation in the primary time series at a certain point is very large, we would include the auxiliary information more often. If the variation is relatively small on the other hand, we would have reason to believe that the univariate predictions are sufficient and would then not include the auxiliary information. With regards to the actual signal and whether it shows signs of structural breaks, Figure 99 might give us reason to believe that the singular vectors change slightly at some time points. However, this is once again perhaps merely a result of the non-uniqueness of a singular vector. Multiplying some of these vectors with negative one will make the changes less apparent. These prediction vectors therefore do not seem to be indicative of structural breaks. They now merely incorporate (to a slight extent) the slight changes in the public’s opinion. Figure 100 allows us to also consider the error associated with these prediction vectors.

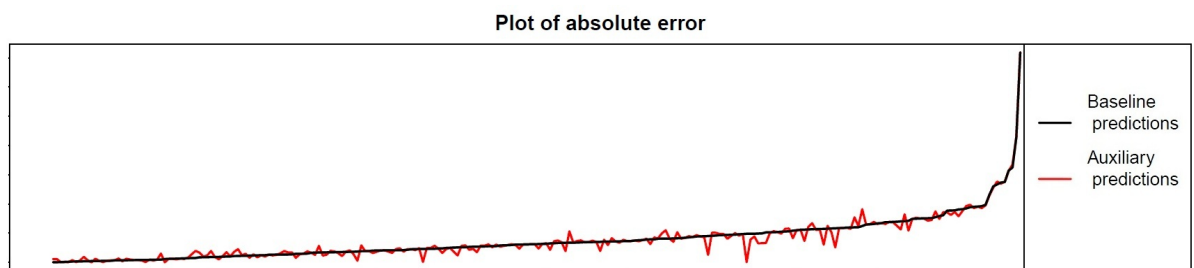


FIGURE 100: This plot illustrates the increasing absolute errors when predicting according to the univariate approach and how the Bayesian predictions (when including at 5% level of significance) affected these absolute errors.

This figure seems to show very significant drops in the absolute error when including the auxiliary information. Furthermore, there are only a few points where inclusion increased the absolute error. These are promising results, but to investigate them in more detail, we also consider Table 34.

Univariate predictions		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
1.1607	0.7733	50.3968%
Predictions including frequency counts		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
1.1519	0.7671	49.2064%
Predictions including sentiment scores		
<i>MSE</i>	<i>MAE</i>	<i>DOC</i>
1.1451	0.7586	55.1587%

TABLE 34: This table shows the measures of accuracy for the baseline Apple predictions during 2011 as well as that of the predictions when including data obtained from Topsy using either the frequency counts or sentiment scores.

This table shows very promising results. When we considered using frequency counts as auxiliary information, we were able to decrease the MSE and MAE slightly, but we were unsatisfied with the fact that the DOC dropped below 50%. This inadequacy is addressed when considering the sentiment scores instead of the frequency counts. With these new predictions, we were not only able to further decrease the MSA and MAE, but we were also able to increase the DOC by 5%, putting us considerably above the 50% mark. To investigate whether these predictions are significantly superior to our baseline predictions, we consider Table 35.

<i>Paired t-test on absolute errors</i>	
Including frequency counts at 0.05	
<i>Mean difference</i>	<i>p-value</i>
0.0147	0.0596
<i>Paired t-test on squared errors</i>	
Including frequency counts at 0.05	
<i>Mean difference</i>	<i>p-value</i>
0.0156	0.1709

TABLE 35: This table gives us the difference between the baseline predictions and the Bayesian predictions in terms of difference in the MAE and MSE measurements. The significance of these differences are then also tabulated in the form of p-values obtained from paired Student's t-test.

Once again, unfortunately, we find that these decreases are not significant at a 95% level of certainty. The p-value associated with testing the difference between the paired absolute errors is very close to 0.05, but not quite there. The p-value testing whether the squared errors associated with the Bayesian approach are significantly smaller than those produced by the univariate approach is also not 0.05, but it is relatively low showing once again that we were able to increase prediction accuracy considerably while maintaining a promising DOC.

Concluding this chapter, we found that the results in these practical studies were both disappointing and encouraging. Unfortunately we were unable to show that by including Tweets we could significantly improve on baseline predictions. However, we did see that this inclusion consistently produced predictions that are uniformly better than the univariate baseline predictions, albeit not by significant amounts. The inclusion of the frequency counts always produced predictions that had lower MSE and MAE.

Furthermore, these predictions only produced a DOC lower than that of the univariate predictions in one example. We realized that by using the frequency counts, we often have a time series with very significant outliers. This could have much too big of an effect on the singular vectors associated with this auxiliary time series. When we considered the possibility of using sentiment scores, we obtained prediction vectors that had much more subtle variation. These prediction vectors then also produced encouraging results. The inclusion of these sentiment scores did not provide significantly more accurate predictions, but in each of the three scenarios the predictions were more accurate than the baseline predictions in terms of all three measures of accuracy. Furthermore, the DOC for this inclusion was never smaller than 55%. We could not help to ask ourselves the question whether statistically significant results are really of concern here. If we were to consider these results from a practical viewpoint, we must remember that the difference between the MAE of the univariate predictions and that of the Bayesian predictions literally represent the average amount of money we would gain when buying and selling a single share according to this selective algorithm instead of the Bayesian algorithm. Also the practical implication of a DOC consistently (even slightly) above 50% is of the utmost importance, since it contradicts the Efficient Market Hypothesis. This chapter might therefore not have proven that the inclusion produces significantly more accurate results, at a 95% level of confidence. It did however consistently show that this inclusion is significant at an 80% level of confidence. Once again, if we were to consider the inherent nature of the data and the practical implication, these are very promising results.

8 Conclusion

Looking back on this thesis, we can partition it into two segments. The first segment is a theoretical discussion on SSA and its extensions. The second segment of this thesis discusses how SSA can be applied to stock market activity in a practical sense and how internet activity can be incorporated into prediction.

In the theoretical discussion on SSA, we did several simulation examples in order to establish how the parameters of the univariate approach affects prediction accuracy. Several authors suggest that the window length parameter should be chosen as large as possible in order to maximize prediction accuracy. However, the simulation studies contradicted this. In the presence of structural breaks, smaller window lengths are sometimes preferred. In other scenarios, we realized that it is often not necessary to choose the window length as large as possible, but that it suffices to choose the parameter large enough. Very large window lengths merely increases computational time and the decrease in prediction accuracy is often negligible.

After a thorough investigation of the univariate approach, existing multivariate extensions of the SSA algorithm were also considered. In the theoretical discussion we already noticed some possible flaws in the rationale behind these existing approaches. These inadequacies quickly also became evident in simulation studies. An alternative multivariate extension of the SSA algorithm was therefore proposed.

The formulation of this alternative multivariate approach was based on the assumption that singular vectors associated with the trajectory matrix approximately follows a multivariate normal distribution. This assumption was necessary in order for us to consider the estimation of the signal singular vector from a Bayesian viewpoint. Even though this assumption was not formally proved, we did a thorough simulation study which showed that the assumption is not a completely unlikely one. Further studies could go towards investigating whether this assumption is indeed valid.

With the assumed distribution, we were able to construct a Bayesian approach to the estimation of

the signal singular vectors associated with the primary time series. This procedure was able to selectively include components of an auxiliary time series instead of an entire time series. Even though the assumption of independence of singular vectors (as we had in the formulation of the Bayesian approach) is unlikely to be entirely accurate, this assumption proved to have numerous advantages in simulation studies. In fact, it was this assumption that allowed us to include components of the auxiliary information independently. The simulation studies clearly showed that this multivariate approach consistently provided predictions that were at least as accurate as the univariate predictions. Furthermore, it seemed to be superior to the existing multivariate techniques in most simulation examples. Even though this multivariate technique seemed to be based on assumptions that we could be pessimistic about, the results spoke for themselves. Unfortunately, the Bayesian approach of course did not produce uniformly superior results, but the simulation study gave enough insight into the rationale behind the approach that we were able to at least identify when the approach should be used. The inadequacies of the Bayesian approach were (more often than not) a result of the assumption of independence between the individual singular vectors. Studies currently being done on using Independent Component Analysis instead of Singular Value Decomposition in the SSA algorithm could perhaps solve some of the inadequacies of the Bayesian approach. In general, these studies seem to hold a promising future for the SSA approach. The effect it would have on the Bayesian approach is definitely one that could be investigated further.

In the practical segment of this thesis, we first considered whether SSA is indeed an appropriate choice of methodology when predicting stock market activity. We compared this approach with the AR model, which is often used in stock market prediction. We quickly saw that the SSA approach is only competitive with the AR model if we apply the algorithm to the returns rather than actual closing prices. Applying the algorithm to the returns also allowed us to consider a broader variety of models and simpler parameter estimation procedures. Choosing from these models, we saw that the SSA algorithm was able to produce predictions that were not only accurate in terms of magnitude of error, but also accurate in terms of correctly predicting positive and negative returns. The same could be said for the AR algorithm.

In the final few chapters, my thesis finally addressed the hypothesized question; that is whether internet activity can actually be useful for prediction of stock market activity. We attempted to answer this question based on theory discussed in previous chapters. In a first study, we attempted to incorporate Google search volumes. Predictions showed significant improvements both in terms of magnitude of error as well as correctly predicting positive and negative returns. Thereafter, we considered the possibility of using the amount of Tweets regarding some company as auxiliary information. This chapter also gave promising but not significant improvements. We speculated that a possible reason for some of the slightly disappointing results could be a consequence of using frequency counts. By using these frequency counts we are assuming that all news is good news. We attempted to address this problem by adding sentiment to each of the Tweets and rather incorporating daily sentiment scores instead of frequency counts. These results were once again promising, but often not significant.

In this final chapter of my thesis, I would once again like to emphasize the fact that we must constantly take into consideration the inherent nature of the primary data. Stock market activity is claimed to be unpredictable and completely random. Slight improvements in prediction accuracy might therefore be all that we can hope for. Most of the practical studies in this thesis showed that internet activity could not significantly increase prediction accuracy. However, in every single one of the seven practical examples considered, we were always able to obtain predictions that were more accurate in terms of magnitude of error and only once did inclusion of internet activity cause us to correctly classify the direction of change less often. These differences might not have been significant, but they were definitely consistent. The

practical implications of consistently increasing the accuracy of stock market predictions, while at the same time correctly classifying positive and negative returns more often, merely by including internet activity are indeed significant.

References

- [1] Antweiler, W., Frank M. 2004. Is all that talk just noise? The information content of internet stock message boards. *Journal of finance*, 59(3): 1259-1294.
- [2] Asur, S., Huberman, B. 2010. Predicting the future with social media. *arXiv preprint arXiv:1003.5699*.
- [3] Benjamini, Y., Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*:289-300.
- [4] Bollen, J., Mao, H., Zeng, X. 2011. Twitter mood predicts the stock market. *Journal of computational sciences*, 2(1): 1-8.
- [5] Brockwell, P., Davis, R. 2002. *Introduction to time series and forecasting*. Springer Verlag.
- [6] Brownstein, J., Freifeld, C., Reis, B., Mandl, K. 2008. Internet based emerging infectious disease intelligence and the HealthMap project. *PLoS medicine*, 5(7), e151.
- [7] De Choudhury, M., Sundaram, H., John, A., Seligmann, D. 2008. Can blog communication dynamics be correlated with stock market activity? *Proc of 19th ACM conference on hypertext and hypermedia*, 55-60.
- [8] Corley, C., Mikler, A., Singh, K., Cook, D. 2009. Monitoring influenza trends through mining social media. *International conference on bioinformatics & computational biology*, 340-346.
- [9] D'Agostino, R., Pearson, E. 1973. Tests for departure from normality: Empirical results for the distribution of b_2 and $\sqrt{b_1}$. *Biometrika*, 60(3), 613-622.
- [10] D'Amuri, F., Marcucci, J. 2009. "Google it!" Forecasting the US unemployment rate with a Google job search. *Social science research network*.
Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1594132, Accessed in June 2011.
- [11] De Klerk, J. 2002. *Time series forecasting and model selection in singular spectrum analysis*. (Doctoral dissertation, Stellenbosch: Stellenbosch University).
- [12] Doshi, L., Krauss, J., Nann, S., Gloor, P. 2009. Predicting movie prices through dynamic social network analysis. *Procedia-social and behavioral sciences*, 2(4), 6423-6433.
- [13] Eckhart, C., Young, G. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), 211-218.
- [14] Ettredge, M., Gerdes, J., Karuga, G. 2005. Using web based search data to predict macroeconomic statistics. *Communications of the ACM*, 48(11), 87-92.
- [15] Gilbert, E., Karahalios, K. 2010. Widespread worry and the stock market. *Proc of international conference on weblogs and social media* (Vol. 2 No. 1, 229-247).
- [16] Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., Brilliant, L. 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012-1014.
- [17] Gnanadesikan, R. (1997). *Methods for statistical data analysis of multivariate observations*. New York: John Wiley.

- [18] Golyandina, N., Nekrutkin, V., Zhigljavsky, A. 2001. *Analysis of time series structure: SSA and related techniques*. Chapman & Hall/CRC.
- [19] Golyandina, N. 2012. New insights into SSA separability. *Proc of 3rd conference on Singular Spectrum Analysis and its applications*.
- [20] Guzman, G. 2011. Internet search behaviour as an economic forecasting tool: The case of inflation expectation. *Journal of economic and social measurement*, 36(3), 119-167.
- [21] Hassani, H. 2007. Singular spectrum analysis: Methodology and comparison. *Journal of data sciences*, 5(2), 239-257.
- [22] Hassani, H., Mahmoudvand, R., Zokaei, M. 2011. Separability and window length in Singular Spectrum Analysis. *Comptes rendus mathematique*, 349 (17), 987-990.
- [23] Jensen, M. 1978. Some anomalous evidence regarding market efficiency. *Journal of financial economics*, 6(2), 95-101.
- [24] Johnson, R., Wichern, D. (1992) *Applied multivariate statistical analysis*. Englewood Cliffs, NJ: Prentice-Hall.
- [25] Lampos, V., Cristianini, N. 2010. Tracking the flu pandemic by monitoring the social web. Available at: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5604088>, Accessed in October 2012.
- [26] Liu, Y., Huang, X., An, A., Yu, X. 2007. A sentiment aware model for predicting sales performance using blogs. *Proc of 30th ACM SIGIR conference on Research and development in information retrieval*, 607-617.
- [27] Mishne, G., Glance, N. 2006. Predicting movie sales from blogger sentiment. *AAAI 2006 spring symposium on computational approaches to analyzing weblogs*.
- [28] Pearson, E., D'Agostino, R., Bowman, K. 1977. Tests for departure from normality: Comparison of powers. *Biometrika*, 64(2), 231-246.
- [29] Polgreen, P., Chen, Y., Pennock, D., Nelson, F. 2012. Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11), 1443-1448.
- [30] Preis, T., Reith, D., Stanley, E. 2010. Complex dynamics of our economic life on different scales: Insights from search engine query data. *Philosophical transactions of the royal society A: Mathematical, physical and engineering sciences*, 368(1933), 5707-5719.
- [31] Radinsky, K., Davidovich, S., Markovitch, S. 2008. Predicting the news of tomorrow using patterns in web search queries. *IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, 363-367.
- [32] Razali, N., Wah, Y. 2010. Power comparison of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of statistical modeling and analytics* 2(1), 21-33.
- [33] Schmidt, T., Vosen, S. 2009. Forecasting private consumption: Survey-based indicators vs. Google Trends. *Ruhr Economic paper*, (155)
- [34] Shapiro, S., Wilk, M. 1965. An analysis of variance test for normality. *Biometrika*, 52(3), 591-611.
- [35] Tumarkin, R., Whitlaw, R. 2001. News or noise? Internet postings and stock prices. *Financial analyst journal*, 57(3), 41-51.

- [36] Varian, H., Choi, H.. 2011. Predicting the present with Google Trends. *Economic record*, 88(s1), 2-9.
- [37] Wu, L., Brynjolfsson, E. 2009. The future of prediction: How Google searches foreshadow housing prices and quantities. *Social science research network*.
Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2022293, Accessed in April 2012 .
- [38] Wuthrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J., Lam, W. 1998. Daily prediction of major stock indices from textual www data. *Proc. of 4th conference on knowledge discovery and data mining*, KDD-98.
- [39] Wysocki, P. 1998. Cheap talk on the web: The determinants of postings on stock message boards. *University of Michigan Business School Working Paper*, 98025.
- [40] Zhang, X., Fuehres, H., Gloor, P. 2010. Predicting stock market indicators through Twitter "I hope its is not as bad as I feared". *1st international workshop on mining social media*.
- [41] Zhang, W., Skiena, S. 2009. Improving movie gross prediction through news analysis. *Proc of IEEE/WIC/ACM international joint conference on web intelligence and intelligent agent technology*, Vol. 1, 301-304.

Appendix A

Quantile-quantile plots testing normality for components in first singular vector

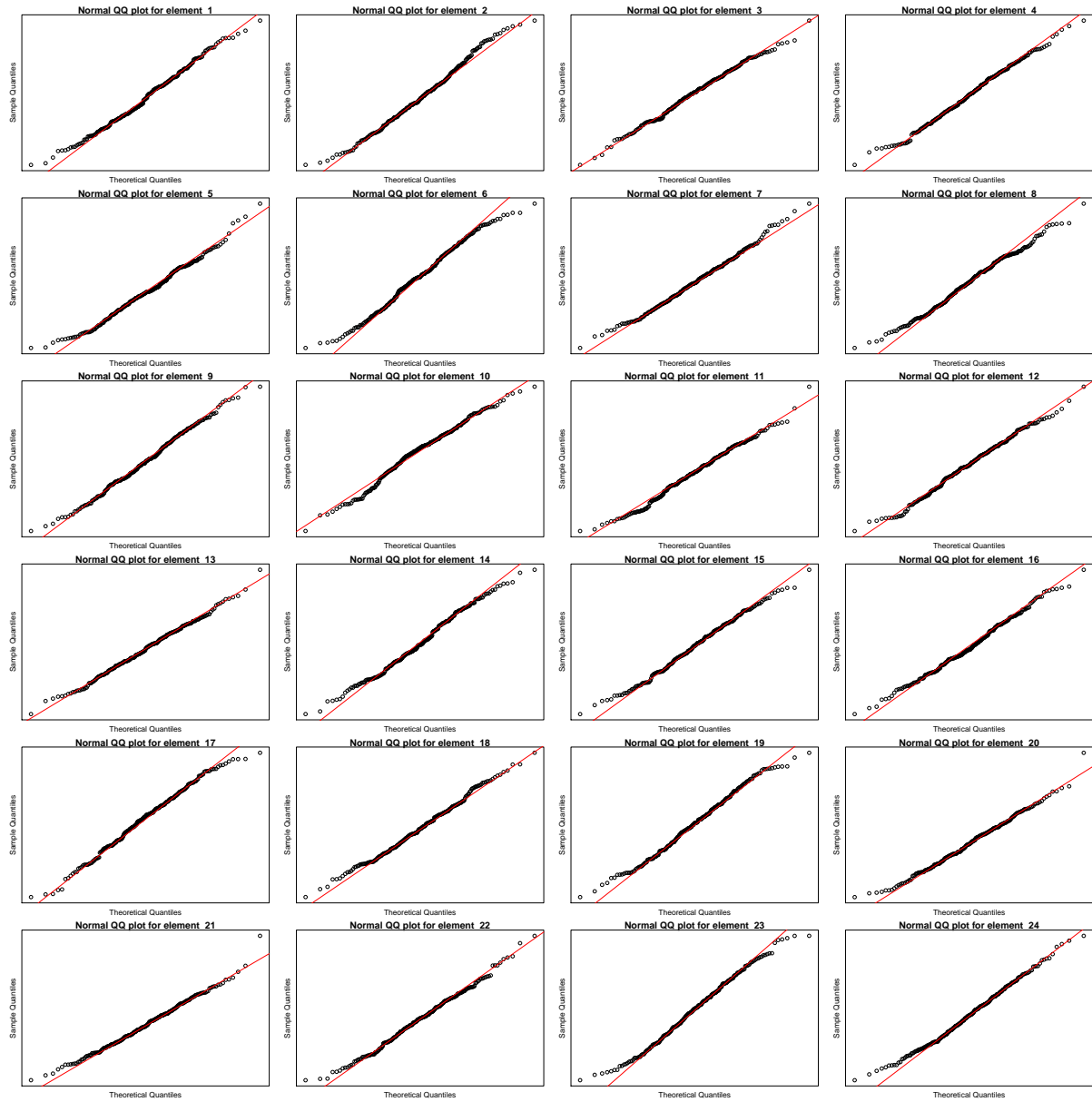


FIGURE A.1: These figures represent the quantile-quantile normality plots for each of the elements in the transformed first singular vector of the simulation example in Chapter 4

Quantile-quantile plots testing normality for components in second singular vector

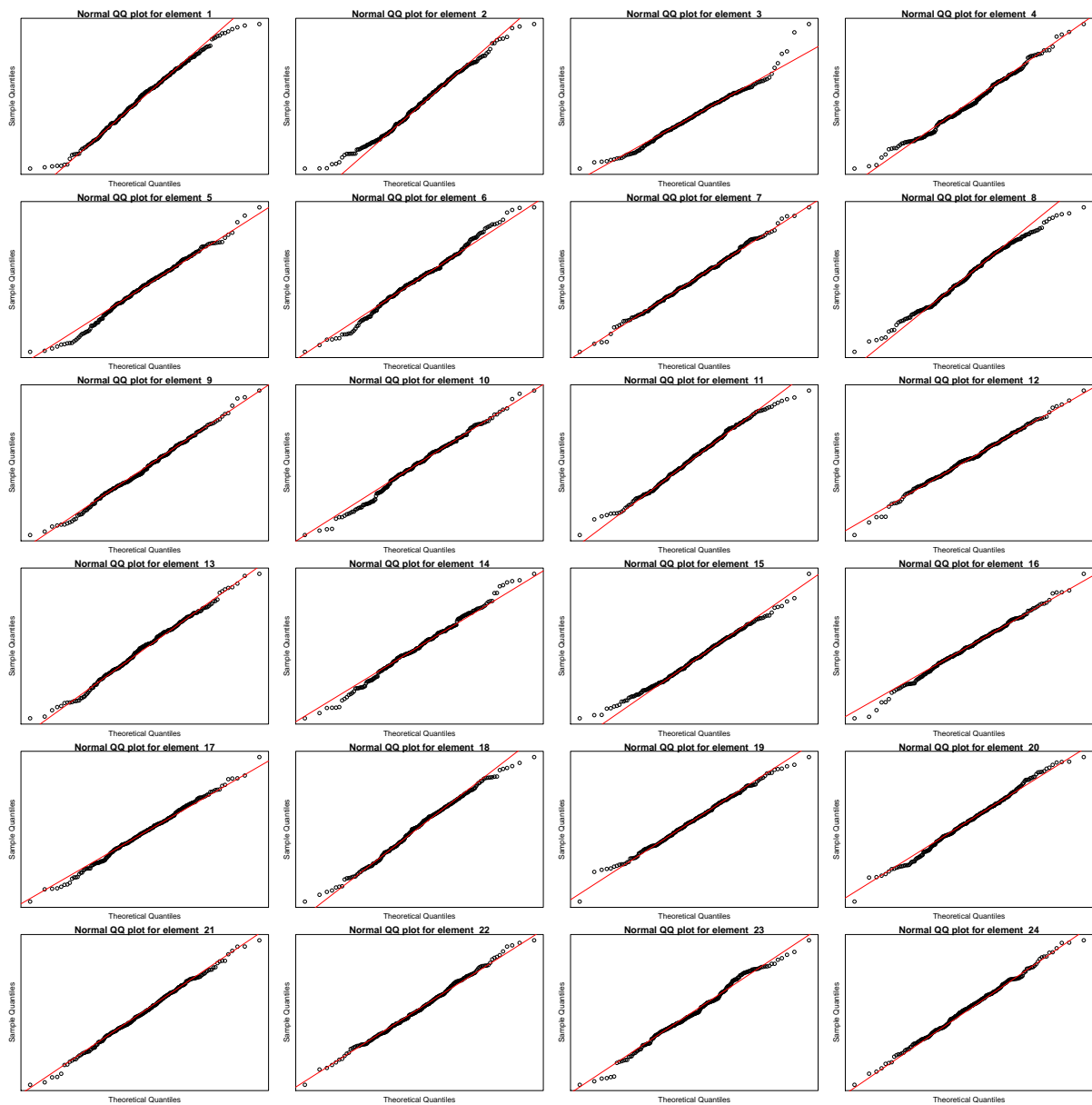


FIGURE A.2: These figures represent the quantile-quantile normality plots for each of the elements in the transformed second singular vector of the simulation example in Chapter 4

Quantile-quantile plots testing normality for components in third singular vector

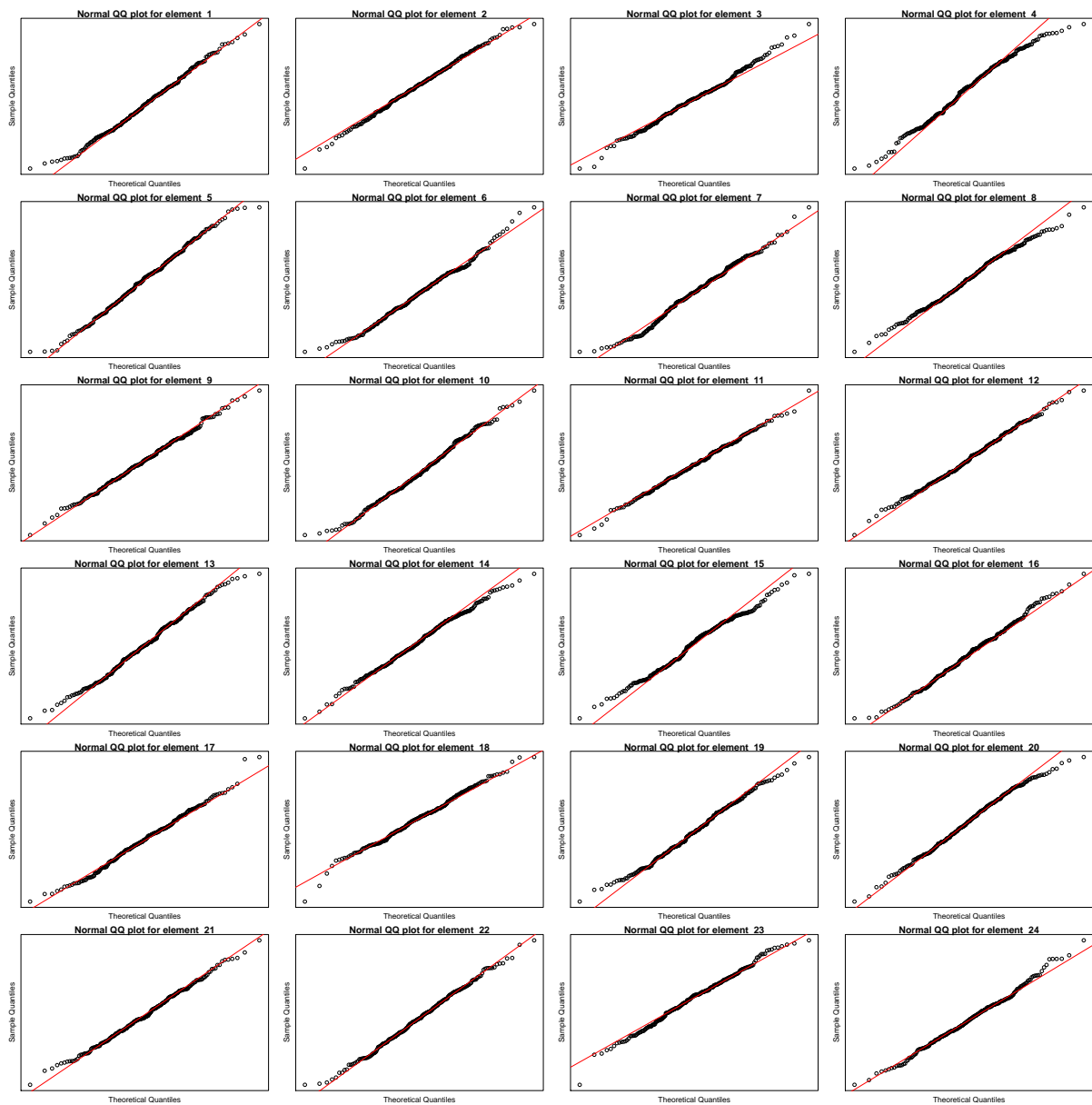


FIGURE A.3: These figures represent the quantile-quantile normality plots for each of the elements in the transformed third singular vector of the simulation example in Chapter 4

Quantile-quantile plots testing normality for components in fourth singular vector

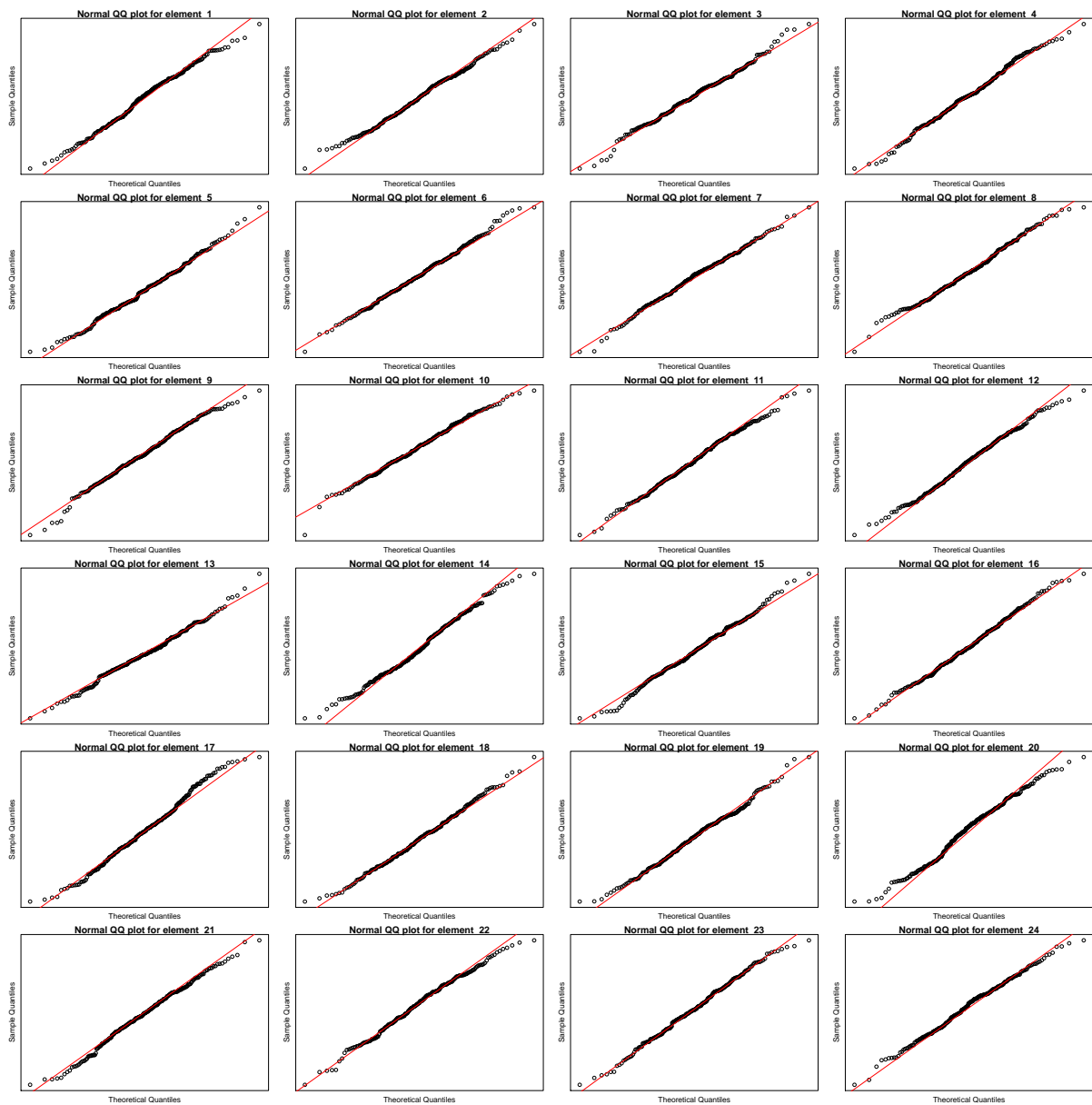


FIGURE A.4: These figures represent the quantile-quantile normality plots for each of the elements in the transformed fourth singular vector of the simulation example in Chapter 4