Stability of Singular Spectrum Analysis and Causality in Time Series

Maria Vronskaya

School of Mathematics
Cardiff University

A thesis submitted for the degree of
Doctor of Philosophy

October 18, 2013

ACKNOWLEDGMENTS

At first I would like to thank my first supervisor Anatoly Zhigljavsky for granting me an opportunity to do the current thesis in School of Mathematics in Cardiff University. Especially, I would like to thank my second supervisor Karl Michael Schmidt not only for sharing his knowledge and wisdom with me, but also for his moral support and enormous patience with me throughout the thesis. I am incredibly grateful for the time he spent educating me and improving my analytical and logical thinking.

I want to thank my parents, my brother, my sister and Katia who were incredibly supportive and kept me on a positive wave. Well, I am lucky to have you all and love you very much.

I also want to thank my friends on both sides of the world, who kept in me entertained and put up with me during grumpiest times. Thank you, Katia, Olia and Valia. Special thanks to my dear friend Svetlana, whose strong character helped to go through rough periods.

And thank you, Flo, for believing in me.

ABSTRACT

The concept of causality has been widely studied in econometrics and statistics since 1969, when C. J. Granger published his paper "Investigating causal relations by econometric models and cross-spectral methods". The intuitive basis for his definition of causality is the following: time series Y is causing time series X if the use of the additional information provided by Y improves the forecast of series X.

In the present thesis we focus on combining Granger's causality concept with the Singular Spectrum Analysis (SSA) technique. SSA is founded on the idea of transforming the time series into a multidimensional trajectory form (Hankel matrices), Singular Value Decomposition with subsequent projection to a lower-dimensional subspace and diagonal averaging.

The main aim of the present thesis is to study the causality concept through SSA prism in details and suggest a novel causality measure, which can be used outside the stationary autoregressive class, which is the framework for Granger's original causality concept.

We first apply standard statistical tests directly to simulated data to assess the improvement of forecast quality of bivariate multidimensional SSA (MSSA) of time series X and Y compared with SSA of time series X only. Although the results of performance of these tests are reasonably conclusive, the simulation method is time consuming and, thus, more theoretical understanding is desirable.

We solve a fundamental scaling problem of the MSSA approach by introducing so-called linearized MSSA. The linearized MSSA approach shows a way towards a causality measure, calculated from the forecast linear recurrence formula (LRF) coefficients.

We finally analyze SSA and (non-linear) bivariate MSSA approach in terms of first order stability analysis under perturbations leading to the construction of a valid suitable measure of causality. The construction of the measure requires some simplifying assumptions in the stability analysis whose validity we verify for both simulated and real data.

CONTENTS

LIST OF FIGURES

INTRODUCTION

The concept of causality has been widely studied in econometrics and statistics since 1969, when C. Granger published his paper "Investigating causal relations by econometric models and cross-spectral methods" [15]. In the present thesis Granger's definition of causality is taken as a conceptual basis, although we give it a different mathematical expression. Intuitively we say that a time series Y causes a time series X if the forecast of the time series X is improved when Y is taken into account. C. Granger defined a measure of causality, based on *causality coherence*, for the particular class of time series which are realizations of *stationary* autoregressive stochastic processes. Although the stationarity condition is widely used in time series analysis, it often remains unclear whether and how real data meet this condition.

The focus of the thesis is to study causality appearing between two time series using the Singular Spectrum Analysis (SSA) technique. SSA is a relatively new method which is now often used in time series analysis. The first reference about SSA can be found in papers written by Broomhead and King in 1986 [7, 6]. Main advantages of this technique are the absence of the stochastic model for the noise (residuals), except the idea that noise should have low self-correlation compared to the signal, and existence of a general and flexible model for the signal, which can be described using linear recurrence. In other words, SSA can be considered a nonparametric method and can be used for arbitrary time series.

The procedure of SSA is based on embedding the time series into multidimensional trajectory space (building Hankel matrices) and Singular Value Decomposition (SVD) with subsequent projection to a lower-dimensional subspace and averaging (Hankelization). Typical tasks for SSA based algorithms include investigating the structure of the time series, in particular identifying trends and periodicities, signal extraction, reducing noise, forecasting time series. The monograph "Analysis of Time Series Structure" [11] contains a detailed description of the method and its usage.

To be more precise, the aim of the present work is to combine Singular Spectrum Analysis technique with Granger's causality concept in such a way that it is possible to find a suitable measure for causality appearing between two time series, without recourse to a stationary autoregressive model. According to Granger's definition of causality, the concept of causality has an immediate connection with the ability to forecast. SSA is an ideal candidate for a new causality test as it provides flexible tools for forecasting.

The thesis has the following structure.

Chapter 1 starts with a description of the basic SSA procedure for reconstruction and forecasting of univariate and multivariate time series. In this chapter we introduce SSA by representing the procedure in two different ways as, first, the stage-to-stage algorithm and as, second, a convolution (linear filter) representation for both univariate and bivariate

cases.

To prepare our connection of Granger's idea to SSA, we study his causality concept with underlying model in details in Chapter 2. This chapter contains the main definitions and ideas introduced by C. Granger in 1969 and also describes the concept Granger suggested as a measure of causality, so-called *coherence*, which is based on the cross spectrum of two stationary time series. In this chapter we also discuss the most common and widely used Granger's causality tests and their implementation. These statistical tests are ultimately based on regression done on two time series and therefore they do not need the assumptions of autoregressive model. We perform numerical study for the time series within the class of stationary autoregressive processes and also for the time series outside this class. Performing Granger's causality test on different time series reveals its sensitivity to changes in model parameters.

However, these tests are not the only tools which can be used to detect a causal relationship. Based on Granger's idea that causality means an improved forecast when the second series is taken into account, we investigate, in Chapter 3, several different statistical tests combined with SSA/MSSA to see if causality takes place.

Chapter 3 contains a detailed description of statistical tests, which we use for causality detection. The chapter is constructed in such a manner that each section consists of two parts, theoretical and practical. Theoretical part contains information about the statistical tests and the practical part contains several examples, which give a general overview on the performance of statistical tests in different cases. In addition to F-test, Kolmogorov-Smirnov and Anderson-Darling tests, Sign test and its variations, we also study *dominance* tests, which involve direct comparison of empirical distributions. However, distribution comparison is a visual tool, which may be time consuming, if one looks at a large number of distributions. Thus, in this chapter we introduce indicator measures, which are used to help to decide which case of dominance occurs in a given data. At the end of this chapter we summarize the observations and explain certain regular patterns that were revealed during simulation studies.

Chapter 4 is dedicated to derivation of a novel approach to causality. Here using Granger's concept we are trying to build a causality measure modeled more closely on Granger's concept, but using SSA/MSSA which is suitable for time series outside the class of stationary autoregressive time series. We show the fundamental difference between the SSA based model and the autoregressive model. Our discussion highlights a general difficulty one faces when dealing with bivariate SSA (MSSA). The main difficulty of the MSSA approach is the lack of homogeneity, which means that if one changes the scaling of one of the time series involved in MSSA the forecast is changing non-linearly. This is problematic if two time series do not have a natural common unit. We propose a measure of causality based on linearized MSSA using parts of the recurrence vectors, however it is not quite clear how it links to either Granger's tests of Chapter 2 or those described in Chapter 3. Therefore, we return to the non-linear MSSA in the next chapter.

Chapter 5 concentrates on the relationship between the noise component of the original time series and the resultant output of SSA/MSSA analysis, i.e. the time series reconstruction and its forecast. This relationship is closely connected to the F-test, which lies at the core of common Granger's causality tests, used for the stationary autoregressive model.

F-test considers the resultant variances in forecast, which in the case of Chapter 3 were obtained with SSA/MSSA linear recurrence formulae. These variances in their turn arise from randomness in the initial time series. Hence, we establish a connection between the input and output noise by considering how the forecast of SSA/MSSA changes as we add a random perturbation to the initial time series, which is the key idea of the stability analysis in this chapter. In particular, we study the propagation of noise through the SSA and MSSA procedures and look for a connection between the initial perturbation in time series and the obtained forecast error.

Note that throughout the thesis we study the forecast as an intermediate tool to obtain a causality measure. Specifically in Chapter 5, we use the recurrence vectors and convolution kernels to construct such a measure. Hence, we are not interested in the forecast itself, but only in linear recurrence formulae the forecast is obtained with and therefore we consider SSA/MSSA forecast for one point only.

In order to obtain a managing formula for the variance,we need to make simplifying assumptions, whose validity we discuss with regard to simulated and real data. On this basis we derive a theoretical model for the output variance in terms of the recurrence vectors and convolution kernels.

1. SINGULAR SPECTRUM ANALYSIS

In this chapter we describe in detail the Singular Spectrum Analysis (SSA) algorithms used for series analysis, reconstruction and forecast. First two sections of this chapter contain the description of the SSA reconstruction algorithms for univariate (SSA) and multivariate cases (MSSA). The basic idea is to take a time series and present it in a multidimensional form, specifically trajectory matrix, which gives the opportunity to study the structure of the time series through spectral analysis, based on singular value decomposition of this matrix. By choosing the parts of the spectral decomposition of the matrix we then are able to reconstruct a partial one-dimensional time series. This approach is useful for noise reduction, signal extraction, structure analysis and many other purposes. Also, we introduce an alternative way of deriving the SSA and MSSA reconstructions.

We introduce the MSSA reconstruction procedure for general multivariate case. We then narrow it down to the bivariate case throughout this thesis, since the subject of interest is the interaction, i.e. causal relationship, between two time series.

The description of reconstruction algorithms is followed by the SSA forecast algorithms for univariate and bivariate cases. The main idea of the SSA forecast algorithm is to find a forecast, such that the distance between the projection subspace and the forecast trajectory is minimal.

1.1 SSA reconstruction procedure

Consider a non-zero time series $F_N = (f_1, ..., f_N)$. The Basic SSA algorithm starts from decomposition stage and is followed by reconstruction stage.

First stage: decomposition

Embedding. Let us choose some window length $L : 1 < L < N$ (which will be the only parameter of the embedding) and define $K = N - L + 1$. At this step we form K lagged vectors $\{X_i\} : X_i = (f_i, ..., f_{L+i-1})^T$, $i \in \{1, ..., K\}$ of dimension L, defining our (L-)trajectory matrix (Hankel matrix) of the time series F_N:

$$\mathbb{X} = \begin{bmatrix} f_1 & f_2 & f_3 \dots & f_K \\ f_2 & f_3 & \cdots & f_{K+1} \\ f_3 & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ f_L & f_{L+1} & \cdots & f_{L+K-1} \end{bmatrix} \tag{1.1}$$

In other words, embedding can be treated as a mapping which transfers a one-dimensional time series $F_N = (f_1, ..., f_N)$ to the multidimensional series $X_1, ..., X_K$.

The choice of window length L could be a task on its own. When choosing a window length, one should consider all the information on the analyzed time series as well as decide on the purpose of the analysis. The purpose could be time series trend extraction, smoothening the series, looking for periodic or oscillation and many more.

Singular Value Decomposition (SVD) is the main tool in the Basic SSA algorithm. Let us denote $S = \mathbb{X}\mathbb{X}^T$ and let $\lambda_1, ..., \lambda_L$ be eigenvalues of S: $\lambda_1 \geq ... \geq \lambda_L \geq 0$. By $\eta_1, ..., \eta_L$ we define orthonormal system of eigenvectors of S corresponding to eigenvalues λ_i.

Let $d = max\{i : \lambda_i > 0\}$. If we denote $V_i = \mathbb{X}^T \eta_i / \sqrt{\lambda_i}$ $(i = 1, ..., d)$ then we can rewrite our matrix $\mathbb{X}$:

$$\mathbb{X} = \mathbf{X}_1 + ... + \mathbf{X}_d, \tag{1.2}$$

where $\mathbf{X}_i = \sqrt{\lambda_i}\eta_i V_i^T$ have rank 1 and therefore are called elementary matrices. The collection $(\sqrt{\lambda_i}, \eta_i, V_i)$ is called the i^{th} eigentriple of the SVD of $\mathbb{X}$.

Second stage: reconstruction

Grouping starts with separation of eigentriple into several groups. In other words if we denote I as a set of indices $\{1, ..., d\}$ and separate it into m disjoint sets:

$$I = \bigcup_{i=1}^{m} I_i, \quad \text{such that} \quad I_i \cap I_j = \emptyset \qquad (i \neq j) \tag{1.3}$$

Then we choose some sets of our indices (I_i for $i = 1, ..., m$). For simplicity let's take $m = 2$. Then $I_1 = \{i_1, ..., i_p\}$ and $I_2 = I \backslash I_1$. Computing a corresponding matrix X_{I_1} we get:

$$\mathbf{X}_{I_1} = \mathbf{X}_{i_1} + ... + \mathbf{X}_{i_p} \tag{1.4}$$

The choice of indices is called eigentriple grouping. We choose I_1 so that $\sum_{i \in I_1} \mathbf{X}_i$ is the approximation of Hankel matrix X

$$\mathbf{X}_{I_1} = \sum_{i \in I_1} \sqrt{\lambda_i}\eta_i V_i^T = P_{I_1}\mathbb{X} \tag{1.5}$$

where P_{I_1} is the orthogonal projector in $\mathbb{R}^L$ onto the subspace spanned by $\{\eta_{i_1}, ..., \eta_{i_p}\}$:

$$P_{I_1} = \sum_{k=i_1}^{i_p} \eta_k \eta_k^T \tag{1.6}$$

Diagonal averaging. After all the decomposition and grouping steps we need to come back to the Hankel matrix form to obtain the reconstructed series.

First of all, we set $L^\star = min(L, K)$ and $K^\star = max(L, K)$ and denote a matrix Y with elements y_{ij}(resultant matrix of grouping step) and $Y^\star$ be a matrix, such that $y^\star_{ij} = y_{ij}$ if $L < K$ and $y^\star_{ij} = y_{ji}$ if $L > K$.

Diagonal averaging transfers arbitrary matrix $Y^\star$ into a Hankel matrix $\widetilde{X}$:

$$\widetilde{X} = HY^\star, \tag{1.7}$$

where H is Hankelization operator and $\widetilde{X}$ is a trajectory matrix of some series $\widetilde{F}_N$. Therefore, we get decomposition of the initial series F_N which is $\widetilde{F}_N$, where $\widetilde{x}_{ij} = \widetilde{f}_{i+j-1} : \{i \in (1, ..., L),\ j \in (1, ..., K)\}$

$$\widetilde{f}_k = \begin{cases} \dfrac{1}{k}\displaystyle\sum_{m=1}^{k} y^\star_{m,k+1-m}, & \text{for} \quad 1 \le k \le L^\star - 1, \\ \dfrac{1}{L^\star}\displaystyle\sum_{m=1}^{L^\star} y^\star_{m,k+1-m}, & \text{for} \quad L^\star \le k \le K^\star, \\ \dfrac{1}{N-k+1}\displaystyle\sum_{m=k-K^\star+1}^{N-K^\star+1} y^\star_{m,k+1-m}, & \text{for} \quad K^\star + 1 \le k \le K^\star + L^\star - 1 \end{cases} \tag{1.8}$$

where $k \in (1, ..., N)$.

$$\widetilde{X} = \begin{bmatrix} \widetilde{f}_1 & \widetilde{f}_2 & \widetilde{f}_3 \dots & \widetilde{f}_K \\ \widetilde{f}_2 & \widetilde{f}_3 & \dots & \widetilde{f}_{K+1} \\ \widetilde{f}_3 & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{f}_L & \widetilde{f}_{L+1} & \dots & \widetilde{f}_N \end{bmatrix} \tag{1.9}$$

or $\widetilde{X}$ can be represented in terms of K-lagged vectors $\widetilde{X} = [\widetilde{X}_1 : ... : \widetilde{X}_K]$.

This notation is a classical SSA algorithm, which could be also found in various sources. For example, see [29].

There is also an alternative notation for the SSA procedure which combines all stages of SSA into one expression. See Proposition 1.2.1.

1.2 Alternative representation of SSA reconstruction

Alternatively, SSA reconstruction procedure could be presented in the convolution form, also known as linear filter [14, p.113], [33].

Proposition 1.2.1. *Let $f = (f_1, ..., f_N)$ be a time series and $\widetilde{f} = (\widetilde{f}_1, ..., \widetilde{f}_N)$ its SSA reconstruction (for a suitable choice of window length L and I_1 set of eigentriples). Then there exists a convolution representation*

$$\widetilde{f_n} = (q \star f)_n = \sum_{m=-L+1}^{L-1} q_m f_{n+m}, \qquad n \in \{L+1, ..., N-L\} \tag{1.10}$$

where the reconstruction kernel $q : \{-L+1, ..., L-1\} \to \mathbb{R}$ *depends on the choice of window length and eigentriples and has the symmetry property*

$$q_{-m} = q_m, \qquad m \in \{-L+1, ..., L-1\} \tag{1.11}$$

Remark As shown in the above Proposition, the convolution formula gives the exact same result as the standard SSA reconstruction, except in the first and last L terms. In fact, the convolution concept is more natural to doubly-infinite time series $f = (f_j)_{j\in\mathbb{Z}}$. Our approach for the proof is to derive the convolution formula for the doubly-infinite case, and make it possible to apply it to finite series f by extending f to the doubly-infinite time series by zeros.

Proof. Consider a double infinite time series $f = (f_j)_{j\in\mathbb{Z}}$, where $f_j \in \mathbb{R}(j \in \mathbb{Z})$ and a right shift operator

$$S : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}, \text{ such that } Sf_t = f_{t+1} \ (t \in \mathbb{Z}) \tag{1.12}$$

so that the trajectory matrix of infinite time series f is

$$\mathbb{X} = \begin{pmatrix} f \\ Sf \\ \vdots \\ S^{L-1}f \end{pmatrix} \tag{1.13}$$

and the lag-covariance matrix

$$\mathbb{X}\mathbb{X}^T = \begin{pmatrix} ff^T & f(Sf)^T & \dots & f(S^{(L-1)}f)^T \\ (Sf)f^T & Sf(Sf)^T & \dots & Sf(S^{(L-1)}f)^T \\ \vdots & & & \\ (S^{(L-1)}f)f^T & S^{(L-1)}f(Sf)^T & \dots & S^{(L-1)}f(S^{(L-1)}f)^T \end{pmatrix}, \tag{1.14}$$

assuming that inner products are finite.
In fact, the matrix $\mathbb{X}\mathbb{X}^T$ is a Toeplitz matrix as

$$S^j f(S^k f)^T = S^{(j+k)} ff^T \qquad (i, j \in \mathbb{Z}) \tag{1.15}$$

The SSA projector P is a spectral projector of rank r

$$P = \begin{pmatrix} p_{0,0} & p_{0,1} & p_{0,2} & \cdots & p_{0,L-1} \\ p_{1,0} & p_{1,1} & p_{1,2} & \cdots & p_{1,L-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{L-1,0} & p_{L-1,1} & p_{L-1,2} & \cdots & p_{L-1,L-1} \end{pmatrix}, \tag{1.16}$$

obtained for first r eigentriples from SVD of the Hankel matrix $\mathbb{X}$, where

$$p_{k,j} = \sum_{i=1}^{r} \eta_{i,k}\eta_{i,j}, \tag{1.17}$$

where $\eta_{i,j}$ corresponds to the j^{th} term of i^{th} eigenvector η_i. Note that

$$p_{k,j} = \sum_{i=1}^{r} \eta_{i,k}\eta_{i,j} = \sum_{i=1}^{r} \eta_{i,j}\eta_{i,k} = p_{j,k} \tag{1.18}$$

Therefore, applying the SSA projector to the Hankel matrix we get

$$P\mathbb{X} = \begin{pmatrix} \sum_{j=0}^{L-1} p_{0,j} S^j f \\ \sum_{j=0}^{L-1} p_{1,j} S^j f \\ \vdots \\ \sum_{j=0}^{L-1} p_{L-1,j} S^j f \end{pmatrix}, \tag{1.19}$$

Define an operator $\mathbb{H}$

$$\mathbb{H} : \mathbb{R}^n \to \mathbb{R} \qquad \text{such that } \mathbb{H}(a_0, ..., a_{n-1}) = \frac{1}{n} \sum_{k=0}^{n-1} S^{-k} a_k. \tag{1.20}$$

Then applying (1.20) to (1.19) we get

$$\begin{aligned} \widetilde{f} = \mathbb{H}P\mathbb{X} &= \frac{1}{L} \sum_{k=0}^{L-1} P\mathbb{X} \\ &= \frac{1}{L} \sum_{k=0}^{L-1} \sum_{j=0}^{L-1} p_{k,j} S^{j-k} f \\ &= \sum_{m=-L}^{L} q_m S^m f \end{aligned} \tag{1.21}$$

or

$$\widetilde{f}_n = \sum_{m=-L+1}^{L-1} q_m f_{n+m} \qquad (n \in \mathbb{Z})$$

which is the SSA reconstruction of the series written as a convolution with convolution coefficients q_m

$$q_m = \begin{cases} \frac{1}{L} \sum_{i=0}^{L-1} p_{i,i}, \, m = 0 \\ \frac{1}{L} \sum_{i=m}^{L-1} p_{i-m,i}, \, m \in (1, ..., L-1) \\ \frac{1}{L} \sum_{i=-m}^{L-1} p_{i,m+i}, \, m \in (-L+1, ..., -1) \end{cases},$$

As the calculation of convolution coefficients q_m is based on summing projector elements (1.17), which have the symmetry (1.18), the q_m are symmetric as well,

$$q_m = q_{-m} (m \in \mathbb{Z}). \tag{1.22}$$

The expression (1.21) can be rewritten as

$$\boxed{\widetilde{f}_n = \sum_{m \in \mathbb{Z}} q_{n-m} f_m = (q \star f)_n} \tag{1.23}$$

where we set

$$q_m = 0 \text{ if } m \notin (-L+1, ..., L-1) \tag{1.24}$$

If f is a finite time series, convolution (1.23) will extend f by zeros to double-infinite time series and produce the correct reconstruction for $\widetilde{f}_n$ for $n \in (L+1, ..., N-L)$. The first and last L points will give reconstruction, which differs from the standard SSA one, due to the the fact, that the series was extended by zeros, on both sides. □

The SSA procedure presented in terms of convolution has an advantage in stability analysis of SSA reconstruction, recurrence and forecasts. In particular, noise propagation through SSA becomes more transparent using the convolution. These points are discussed in more details in Chapter 5.

1.3 Multivariate SSA (MSSA) reconstruction procedure

MSSA is an extended version of the SSA algorithm applied to more than one series. Instead of having one F_N, we now have several series $F_N^{(s)} = (f_0^{(s)}, ..., f_{N-1}^{(s)})$, where s is the number of series and N is the length of $F_N^{(s)}$. Define trajectory matrices of each time series $F_N^{(s)}$ as $\mathbb{X}^{(s)}$ as in (1.1). The combined trajectory matrix for multivariate case is

$$\mathbb{X} = \begin{pmatrix} \mathbb{X}^{(1)} \\ \mathbb{X}^{(2)} \\ \vdots \\ \mathbb{X}^{(s)} \end{pmatrix} \tag{1.25}$$

If the length of time series $F^{(i)}$ vary, K_i should be adjusted for trajectory matrix so that $L = const$, i.e. $K_i = N_i - L + 1$.

Decomposition (SVD) and grouping steps are similar to the Basic SSA procedure. For diagonal averaging first let us define K_i $(i = 1, ..., s)$ of the trajectory matrix $X^{(i)}$ and an operator Δ_i:

$$\Delta_i = \begin{cases} 0 & \text{for} \quad i = 1 \\ \sum_{s=0}^{i-1} K_s & \text{for} \quad i > 1 \end{cases} \tag{1.26}$$

Then the diagonal averaging has a following form:

$$\widetilde{f}_k^{(i)} = \begin{cases} \dfrac{1}{k}\displaystyle\sum_{m=1}^{k} y^{\star}_{m,k+1-m+\Delta_i}, & \text{for} \quad 1 \le k \le L-1, \\ \dfrac{1}{L}\displaystyle\sum_{m=1}^{L} y^{\star}_{m,k+1-m+\Delta_i}, & \text{for} \quad L \le k \le K_i, \\ \dfrac{1}{N_i-k+1}\displaystyle\sum_{m=k-K_i+1}^{N_i-K_i+1} y^{\star}_{m,k+1-m+\Delta_i}, & \text{for} \quad K_i+1 \le k \le N_i \end{cases} \tag{1.27}$$

where $y^{\star}_{ij}$ are elements of matrix $Y^{\star}$, result of grouping step.

Analogously to the univariate SSA procedure, the MSSA reconstruction can be represented by a convolution formula as well. We only consider bivariate case.

1.4 Alternative representation of bivariate SSA reconstruction

The possibility of representing the reconstruction as a convolution for the bivariate case expand opportunities of stability analysis, in particular, studying noise propagation in the bivariate case.

Proposition 1.4.1. *Let $f = (f_1, ..., f_N)$, $g = (g_1, ..., g_N)$ be time series and $\widetilde{f} = (\widetilde{g}_1, ..., \widetilde{g}_N)$ $\widetilde{f} = (\widetilde{f}_1, ..., \widetilde{f}_N)$ their MSSA reconstruction respectively (for a suitable choice of window length L and r eigentriples). Then there exists a convolution representation*

$$\begin{pmatrix} \widetilde{f} \\ \widetilde{g} \end{pmatrix} = \begin{pmatrix} q^1 \star f + q^2 \star g \\ q^3 \star f + q^4 \star g \end{pmatrix} \tag{1.28}$$

where the reconstruction kernel $q^i : \{-L+1, ..., L-1\} \to \mathbb{R}$ for $i \in \{1, 2, 3, 4\}$ depends on the choice of window length and eigentriples

Proof. Suppose we now have two doubly-infinite time series $(f_n)_{n\in\mathbb{Z}}, (g_n)_{n\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$, i.e. $f_j, g_j \in \mathbb{R} (j \in \mathbb{Z})$. Here $(f_n)_{n\in\mathbb{Z}}$ is the main series we are interested in and the $(g_n)_{n\in\mathbb{Z}}$ is the support one.

The Hankel matrix for the two dimensional case is a block matrix consisting of two blocks: one is the Hankel matrix $\mathbb{X}_1$ of f_n and the other one is the Hankel matrix $\mathbb{X}_2$ of g_n

$$\mathbb{X} = \begin{pmatrix} \mathbb{X}_1 \\ \mathbb{X}_2 \end{pmatrix} = \begin{pmatrix} f \\ Sf \\ \vdots \\ S^{L-1}f \\ g \\ Sg \\ \vdots \\ S^{L-1}g \end{pmatrix} \tag{1.29}$$

and the lag-covariance matrix

$$\mathbb{X}\mathbb{X}^T = \begin{pmatrix} ff^T & \dots & f(S^{(L-1)}f)^T & fg^T & \dots & f(S^{(L-1)}g)^T \\ (Sf)f^T & & \dots & \dots & \dots & \\ \vdots & & \vdots & \vdots & \vdots & \\ (S^{(L-1)}f)f^T & \dots & S^{(L-1)}f(S^{(L-1)}f)^T & & \dots & \\ gf^T & \dots & g(S^{(L-1)}f)^T & gg^T & \dots & g(S^{(L-1)}g)^T \\ (Sg)f^T & \dots & Sf(S^{(L-1)}f)^T & (Sg)g^T & \dots & \\ \vdots & & \vdots & & \vdots & \\ (S^{(L-1)}g)f^T & \dots & \dots & S^{(L-1)}g(S^{(L-1)}g)^T & \dots & \dots \end{pmatrix} \tag{1.30}$$

The MSSA projector P is obtained for first r eigentriples from SVD of Hankel matrix $\mathbb{X}$.

$$P = \begin{pmatrix} p^1_{0,0} & p^i_{0,1} & \cdots & p^1_{0,L-1} & p^2_{0,L} & p^2_{0,L+1} & \cdots & p^2_{0,2L-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p^1_{L-1,0} & p^1_{L-1,1} & \cdots & p^1_{L-1,L-1} & p^2_{L-1,L} & p^2_{L-1,L+1} & \cdots & p^2_{L-1,2L-1} \\ p^3_{L,0} & p^3_{L,1} & \cdots & p^3_{L,L-1} & p^4_{L,L} & p^4_{L,L+1} & \cdots & p^4_{L,2L-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p^3_{2L-1,0} & p^3_{2L-1,1} & \cdots & p^3_{2L-1,L-1} & p^4_{2L-1,L} & p^4_{2L-1,L+1} & \cdots & p^4_{2L-1,2L-1} \end{pmatrix}.$$

The projector P consist of 4 blocks

$$P = \begin{pmatrix} p^1 & p^2 \\ p^3 & p^4 \end{pmatrix}$$

where

$$\begin{aligned} p^1_{k,j} &= \sum_{i=1}^{r} \eta_{i,k}\eta_{i,j} \text{ for } k,j \in \{0,...,L-1\} \\ p^2_{k,j} &= \sum_{i=1}^{r} \eta_{i,k}\eta_{i,j} \text{ for } k \in \{0,...,L-1\}, j \in \{L,...,2L-1\} \\ p^3_{k,j} &= \sum_{i=1}^{r} \eta_{i,k}\eta_{i,j} \text{ for } k \in \{L,...,2L-2\}, j \in \{0,...,L-1\} \\ p^4_{k,j} &= \sum_{i=1}^{r} \eta_{i,k}\eta_{i,j} \text{ for } k,j \in \{L,...,2L-1\} \end{aligned} \tag{1.31}$$

where η_i are eigenvectors, obtained from SVD of the matrix $\mathbb{X}\mathbb{X}^\intercal$

$$P\mathbb{X} = \begin{pmatrix} \sum_{j=0}^{L-1} p^1_{0,j} S^j f + \sum_{j=0}^{L-1} p^2_{0,j} S^j g \\ \sum_{j=0}^{L-1} p^1_{1,j} S^j f + \sum_{j=0}^{L-1} p^2_{1,j} S^j g \\ \vdots \\ \sum_{j=0}^{L-1} p^1_{L-1,j} S^j f + \sum_{j=0}^{L-1} p^2_{L-1,j} S^j g \\ \sum_{j=0}^{L-1} p^3_{0,j} S^j f + \sum_{j=0}^{L-1} p^4_{0,j} S^j g \\ \sum_{j=0}^{L-1} p^3_{1,j} S^j f + \sum_{j=0}^{L-1} p^4_{1,j} S^j g \\ \vdots \\ \sum_{j=0}^{L-1} p^3_{L-1,j} S^j f + \sum_{j=0}^{L-1} p^4_{L-1,j} S^j g \end{pmatrix}, \tag{1.32}$$

Similarly to the univariate case, we now can write down the expression for the reconstruction of f_n and g_n as a convolution

$$\begin{pmatrix} \widetilde{f} \\ \widetilde{g} \end{pmatrix} = \begin{pmatrix} \mathbb{H}\left(\sum_{k=0}^{L-1} \sum_{k=0}^{L-1} p^1_{k,j} S^{j-k} f + \sum_{k=0}^{L-1} \sum_{k=0}^{L-1} p^2_{k,j} S^{j-k} g \right) \\ \mathbb{H}\left(\sum_{k=0}^{L-1} \sum_{k=0}^{L-1} p^3_{k,j} S^{j-k} f + \sum_{k=0}^{L-1} \sum_{k=0}^{L-1} p^4_{k,j} S^{j-k} g \right) \end{pmatrix} \tag{1.33}$$

$$= \begin{pmatrix} \sum_{m=-L}^{L} q^1_m S^m f + \sum_{m=-L}^{L} q^2_m S^m g \\ \sum_{m=-L}^{L} q^3_m S^m f + \sum_{m=-L}^{L} q^4_m S^m g \end{pmatrix} \tag{1.34}$$

as $S^m f = f_m$ and $S^m g = g_m$ we get two reconstructions

$$\begin{pmatrix} \widetilde{f_n} \\ \widetilde{g_n} \end{pmatrix} = \begin{pmatrix} \sum_{m\in\mathbb{Z}} q^1_{n-m} f_m + \sum_{m\in\mathbb{Z}} q^2_{n-m} g_m \\ \sum_{m\in\mathbb{Z}} q^3_{n-m} f_m + \sum_{m\in\mathbb{Z}} q^4_{n-m} g_m \end{pmatrix} = \begin{pmatrix} (q^1 \star f)_n + (q^2 \star g)_n \\ (q^3 \star f)_n + (q^4 \star g)_n \end{pmatrix} \ (n \in \mathbb{Z}) \tag{1.35}$$

or briefly

$$\boxed{\begin{pmatrix}\widetilde{f}\\ \widetilde{g}\end{pmatrix} = \begin{pmatrix}q^1 \star f + q^2 \star g\\ q^3 \star f + q^4 \star g\end{pmatrix}} \tag{1.36}$$

Here $q^j_m (j \in \{1,2,3,4\})$ are convolution coefficients

$$q^j_m = \begin{cases} \frac{1}{L}\sum\limits_{i=0}^{L-1} p^j_{i,i},\ m = 0 \\ \frac{1}{L}\sum\limits_{i=m}^{L-1} p^j_{i-m,i},\ m \in (1,...,L-1) \\ \frac{1}{L}\sum\limits_{i=-m}^{L-1} p^j_{i,m+i},\ m \in (-L+1,...,-1) \end{cases}, \tag{1.37}$$

The expression (1.36) gives the reconstructions $\widetilde{f}_n$, $\widetilde{g}_n$ of the main series f_n and support series g_n for $n \in \{L+1,...,N-L\}$. This expression is identical to the MSSA reconstructions, provided that the parameters of both procedures are the same. □

It is worth mentioning that convolution vectors q^1 and q^4 have the same symmetry as in 1.22 as in the univariate case but q^2, q^3 are not. In fact, $q^2_m = q^3_{-m} (m \in \mathbb{Z})$.

1.5 The standard SSA forecast

To show the continuation of the series we need a forecast base. The forecast base is represented by the reconstructed series, which are described below. We are using the following notation in this section, $f = (\widetilde{f}_{N-L+2},...,\widetilde{f}_N)^T$, $f_+ = (\widetilde{f}_{N-L+2},...\widetilde{f}_{N+1})^T$ and $\widetilde{f}_m$ is an obtained forecast point for $i \in \{N+1,...,N+M\}$.

In terms of K-lagged vectors one needs to find $\widetilde{X}_{K+M}$ vectors:

$$\widetilde{X}_{K+M} = \begin{pmatrix} \hat{f}_{K+M} \\ \hat{f}_{K+M+1} \\ \vdots \\ \hat{f}_{N+M} \end{pmatrix} \tag{1.38}$$

where

$$\hat{f}_i = \begin{cases} \widetilde{f}_i, & \text{if } i = K+1,...,N \\ \hat{f}_i, & \text{if } i = N+1,...,N+M \end{cases}$$

In other words, $\hat{f}_i$ corresponds to the reconstructed time series for $i \in \{K+1,...,N\}$ and forecast points for $i \in (N+1,...,N+M)$.

M is the number of points to forecast. For simplicity we take $M = 1$. To find vector $\widetilde{X}_{K+1}$ we choose $\hat{f}_{N+1}$ to minimize the distance from the span $\{\eta_{i_1}...\eta_{i_p}\}$ to the vector $\widetilde{X}_{K+1}$, using that $P_{I_1}f_+$ is the closest point to the point f_+ in span $\{\eta_{i_1}...\eta_{i_p}\}$:

$$z(\hat{f}_{N+1}) = \| f_+ - P_{I_1}f_+ \|^2 = \| (I - P_{I_1})f_+ \|^2 \longrightarrow min. \tag{1.39}$$

Taking a derivative with respect to $\hat{f}_{N+1}$ we get:

$$z'(\hat{f}_{N+1}) = 2f_+^T(I - P_{I_1})e_L,$$

where $e_L = (0, ..., 0, 1)^T$ and $e \in \mathbb{R}^L$.
To minimize the distance from the vector $(f_{N-L+2}, ..., f_{N+1})^T$ onto space spanned by $\{\eta_{i_1}...\eta_{i_p}\}$, we obtain

$$z'(\hat{f}_{N+1}) = 2f_+^T(I - P_{I_1})e_L = 0,$$

which can be rewritten as

$$f_+^T(I - P_{I_1})_L = 0, \tag{1.40}$$

where

$$(I - P_{I_1})_L^T = e_L - \sum_{k=i_1}^{i_p} \eta_k \eta_k^T e_L.$$

Then we have

$$0 = (I - P_{I_1})_L f_+^T = \hat{f}_{N+1} - \sum_{k=i_1}^{i_p} e_L \eta_k^T \eta_k f_+,$$

where

$$\sum_{k=i_1}^{i_p} e_L \eta_k^T = \eta_{k,L},$$

$$\eta_k f_+ = \eta_k^{\nabla T} f + \eta_{k,L}\hat{f}_{N+1}.$$

The solution of (1.40) is

$$\hat{f}_{N+1} = f^T \frac{\sum_{k=i_1}^{i_p} \eta_k^{\nabla} U_{k,L}}{1 - \sum_{k=i_1}^{i_p} \eta_{k,L}^2},$$

where $U_{k,L}$ is the L^{th} component of eigenvector η_k and $\eta_k^{\nabla} \in \mathbb{R}^{L-1}$ (represents first ($L - 1$) components of eigenvector η_k). In other words $\hat{f}_{N+1}$ can be expressed in terms of the following linear recurrence formula:

$$\hat{f}_{N+1} = a_1\hat{f}_N + a_2\hat{f}_{N-1} + ... + a_{L-1}\hat{f}_{N-L+1}, \tag{1.41}$$

where a_i are components of the recurrent vector $R = (a_{L-1}, a_{L-2}, ..., a_1)^T$:

$$R = \frac{\sum_{k=i_1}^{i_p} \eta_{k,L}\eta_k^{\nabla}}{1 - \sum_{k=i_1}^{i_p} \eta_{k,L}^2}. \tag{1.42}$$

1.6 MSSA forecast

$MSSA_+$ forecast

Here in addition to the notation in the previous section we use $g = (\widetilde{g}_{N-L+2}, ..., \widetilde{g}_N)^T$, $g_+ = (\widetilde{g}_{N-L+2}, ..., \widetilde{g}_{N+1})^T$, where $\widetilde{g}_{N+1}$ is known or unknown value depending on the method used.

For the MSSA forecast we consider two possible scenarios. First scenario takes place if the "future" value at $(N+1)^{th}$ point of the support series is known. The second scenario is if there is equal amount of information available on time series used for the forecast.

To distinguish between these two scenarios we use following notation: the procedure with extra information on support series is defined as MSSA_+ and with equal amount of information - MSSA.

For MSSA_+ forecast there is a Hankel matrix of the following form:

$$\mathbb{X} = \begin{pmatrix} \mathbb{X}_1 \\ \mathbb{X}_2 \end{pmatrix},$$

$\mathbb{X}_1$ is a Hankel matrix of the main series F_N(we are interested in) with window length L_1 and $\mathbb{X}_2$ is a Hankel matrix of a supportive series G_N with window length L_2. For the simplicity we suppose that window lengths are equal $L_1 = L_2 = L$. Here we deal with the scenario when extra information about supportive series G_N, i.e. g_{N+1} is known. As in previous case with SSA forecast algorithm we take an orthogonal projector P (a $2L \times 2L$ matrix) of the matrix $\widetilde{X}$:

$$P = P_{I_1} = \sum_{k=i_1}^{i_p} \eta_k \eta_k^T$$

and as in previously described SSA case we choose f_{N+1} to minimize the distance from the span $\{\eta_{i_1} ... \eta_{i_p}\}$ to the vector

$$\widetilde{X}_{K+1} = \begin{pmatrix} \widetilde{X_1} \\ \widetilde{X_2} \end{pmatrix}_{K+1}.$$

Then we take the following norm:

$$z(f_{N+1}) = \| \begin{pmatrix} f_+ \\ g_+ \end{pmatrix} - P \begin{pmatrix} f_+ \\ g_+ \end{pmatrix} \|^2 \longrightarrow min, \tag{1.43}$$

where f_+ represents terms of the series F_N and terms g_+ represents terms of the series G_N and g_{N+1} is the known value in g_+. We can rewrite the expression as

$$z(f_{N+1}) = \| (I - P) \begin{pmatrix} f_+ \\ g_+ \end{pmatrix} \|^2 \longrightarrow min. \tag{1.44}$$

When there is a known value of supportive series G_N, and the only unknown value of the given vector $(f_+, g_+)^T$ is f_{N+1} the problem is similar to SSA forecast case. The only

change is the size of the vector $\widetilde{X}_{K+1}$.
Taking the derivative of (1.44) with respect to f_{N+1}, we get

$$z'(f_{N+1}) = 2\begin{pmatrix} f_+ \\ g_+ \end{pmatrix}^T (I-P)e_L = 2\begin{pmatrix} f_+ \\ g_+ \end{pmatrix}^T (I-P)_L,$$

where

$$(I-P)_L = e_L - \sum_{k=i_1}^{i_p} \eta_k \eta_k^T e_L = e_L - \sum_{k=i_1}^{i_p} \eta_k \eta_{k,L}$$

or we can rewrite it as

$$e_L - \sum_{k=i_1}^{i_p} \begin{pmatrix} \eta_k^1 \\ \eta_{k,L} \\ \eta_k^2 \end{pmatrix} \eta_{k,L}.$$

Thus the derivative can be rewritten as

$$z'(f_{N+1}) = \begin{pmatrix} f_+ \\ g_+ \end{pmatrix}^T (e_L - \sum_{k=i_1}^{i_p} \begin{pmatrix} \eta_k^{(1)} \\ \eta_{k,L} \\ \eta_k^{(2)} \end{pmatrix} \eta_{k,L}),$$

where $e_L = (0,...,0,1,0,...,0)^T \in \mathbb{R}^{2L}$, $\eta_k^{(1)} = (\eta_1^{(k)},...,\eta_{L-1}^{(k)})^T \in \mathbb{R}^{L-1}$, $\eta_{k,L} = (\eta_L^{(k)})$ and $\eta_k^{(2)} = (\eta_{L+1}^{(k)},...,\eta_{2L}^{(k)})^T \in \mathbb{R}^L$. $(I-P)_L$ is the L-th row of the matrix $(I-P)$.
To find the the f_{N+1} value, which is optimal we need to find the solution to $z'(f_{N+1}) = 0$, where

$$z'(f_{N+1}) = f_{N+1} - (f^T \sum_{k=i_1}^{i_p} \eta_k^{(1)} + \sum_{k=i_1}^{i_p} \eta_{k,L}^2 f_{N+1} + g_+^T \sum_{k=i_1}^{i_p} \eta_k^{(2)}).$$

Thus, we have

$$f_{N+1} = \frac{1}{1 - \sum_{k=i_1}^{i_p} \eta_{k,L}^2} \left(f^T \sum_{k=i_1}^{i_p} \eta_{k,L} U_k^{(1)} + g_+^T \sum_{k=i_1}^{i_p} \eta_{k,L} \eta_k^{(2)} \right)$$

or

$$f_{N+1} = f \frac{\sum_{k=i_1}^{i_p} \eta_{k,L} \eta_k^{(1)}}{1 - \sum_{k=i_1}^{i_p} \eta_{k,L}^2} + g_+ \frac{\sum_{k=i_1}^{i_p} \eta_{k,L} \eta_k^{(2)}}{1 - \sum_{k=i_1}^{i_p} \eta_{k,L}^2}$$

We can express f_{N+1} using the following recurrence formula

$$f_{N+1} = a_1 f_N + a_2 f_{N-1} + ... + a_{L-1} f_{N-L+1} + b_1 g_{N+1} + ... + b_L g_{N-L+2}, \tag{1.45}$$

where $R_1 = (a_{L-1}, a_{L-2}, ..., a_1)^T$, $R_2 = (b_L, b_{L-1}, ..., b_1)^T$

$$R_1 = \frac{\sum_{k=i_1}^{i_p} \eta_{k,L}\eta_k^{(1)}}{1 - \sum_{k=i_1}^{i_p} \eta_{k,L}^2} \tag{1.46}$$

$$R_2 = \frac{\sum_{k=i_1}^{i_p} \eta_{k,L}\eta_k^{(2)}}{1 - \sum_{k=i_1}^{i_p} \eta_{k,L}^2} \tag{1.47}$$

In this case we get two recurrent vectors of different sizes - $R_1 \in \mathbb{R}^{L-1}$ and $R_2 \in \mathbb{R}^L$.

MSSA forecast

For MSSA case there appears the second unknown value g_{N+1}. So here we have 2 series of the same length N. As in MSSA$_+$ we take the same norm z (1.43), but here we need to minimize it with respect to 2 unknown values f_{N+1} and g_{N+1}. Differentiating will give the following equations:

$$\frac{\partial z}{\partial \hat{f}_{N+1}} = 2 \begin{pmatrix} f_+ \\ g_+ \end{pmatrix}^T (I - P)_L \tag{1.48}$$

and

$$\frac{\partial z}{\partial \hat{g}_{N+1}} = 2 \begin{pmatrix} f_+ \\ g_+ \end{pmatrix}^T (I - P)_{2L} \tag{1.49}$$

now we need to solve the system of equations:

$$\begin{cases} \frac{\partial z}{\partial \hat{f}_{N+1}} = 0 \\ \frac{\partial z}{\partial \hat{g}_{N+1}} = 0 \end{cases} \tag{1.50}$$

using the fact that $P =$

$$\frac{\partial z}{\partial \hat{f}_{N+1}} = \begin{pmatrix} f_+ \\ g_+ \end{pmatrix}^T (e_L - \sum_{k=i_1}^{i_p} \begin{pmatrix} \eta_k^{(1)} \\ \eta_{k,L} \\ \eta_k^{(2)} \\ \eta_{k,2L} \end{pmatrix} \eta_{k,L}) \tag{1.51}$$

and

$$\frac{\partial z}{\partial \hat{g}_{N+1}} = \begin{pmatrix} f_+ \\ g_+ \end{pmatrix}^T (e_{2L} - \sum_{k=i_1}^{i_p} \begin{pmatrix} \eta_k^{(1)} \\ \eta_{k,L} \\ \eta_k^{(2)} \\ \eta_{k,2L} \end{pmatrix} \eta_{k,2L}) \tag{1.52}$$

where $e_{2L} = (0, ..., 0, 1)^T \in \mathbb{R}^{2L}$. Rearranging the system of equations will give us next system in matrix form

$$A \begin{pmatrix} \hat{f}_{N+1} \\ \hat{g}_{N+1} \end{pmatrix} = b \tag{1.53}$$

where A is a 2×2 matrix:

$$A = \begin{pmatrix} 1 - \sum_{k=i_1}^{i_p} \eta_{k,L}^2 & -\sum_{k=i_1}^{i_p} \eta_{k,2L}\eta_{k,L} \\ -\sum_{k=i_1}^{i_p} \eta_{k,L}\eta_{k,2L} & 1 - \sum_{k=i_1}^{i_p} \eta_{k,2L}^2 \end{pmatrix} \tag{1.54}$$

and b is a vector:

$$b = \begin{pmatrix} f\sum_{k=i_1}^{i_p} \eta_k^{(1)}\eta_{k,L} + g\sum_{k=i_1}^{i_p} \eta_k^{(2)}\eta_{k,L} \\ g\sum_{k=i_1}^{i_p} \eta_k^{(2)}\eta_{k,2L} + f\sum_{k=i_1}^{i_p} \eta_k^{(1)}\eta_{k,2L} \end{pmatrix} \tag{1.55}$$

So the system (1.53) can be rewritten:

$$\begin{pmatrix} \hat{f}_{N+1} \\ \hat{g}_{N+1} \end{pmatrix} = A^{-1}b$$

where A^{-1} is an inverse matrix.

The solution to the system is vector $(f_{N+1}, g_{N+1})^T$ so that:

$$\begin{aligned} \hat{f}_{N+1} = {} & \frac{1 - \sum_{k=i_1}^{i_p} \eta_{k,2L}^2}{\det A}(f^T \sum_{k=i_1}^{i_p} \eta_k^{(1)}\eta_{k,L} + g^T \sum_{k=i_1}^{i_p} \eta_k^{(2)}\eta_{k,L}) \\ & + \frac{\sum_{k=i_1}^{i_p} \eta_{k,2L}\eta_{k,L}}{\det A}(g^T \sum_{k=i_1}^{i_p} \eta_k^{(2)}\eta_{k,2L} + f^T \sum_{k=i_1}^{i_p} \eta_k^{(1)}\eta_{k,2L}) \end{aligned} \tag{1.56}$$

$$\begin{aligned} \hat{g}_{N+1} = {} & \frac{1 - \sum_{k=i_1}^{i_p} \eta_{k,L}^2}{\det A}(g^T \sum_{k=i_1}^{i_p} \eta_k^{(2)}\eta_{k,2L} + f^T \sum_{k=i_1}^{i_p} \eta_k^{(1)}\eta_{k,2L}) \\ & + \frac{\sum_{k=i_1}^{i_p} \eta_{k,L}\eta_{k,2L}}{\det A}(f^T \sum_{k=i_1}^{i_p} \eta_k^{(1)}\eta_{k,L} + g^T \sum_{k=i_1}^{i_p} \eta_k^{(2)}\eta_{k,L}) \end{aligned} \tag{1.57}$$

As can be seen from the formulae above, the forecast recurrence formula in this case will consist of two parts, including values from two series F_N and G_N and the coefficients in this recurrent formula are represented as vectors $\in \mathbb{R}^{L-1}$.

We can express $\hat{f}_{N+1}$ and $\hat{g}_{N+1}$ using following linear recurrence formulae

$$\hat{f}_{N+1} = a_{1,1}\widetilde{f}_N + ... + a_{1,L-1}\widetilde{f}_{N-L+1} + b_{1,1}\widetilde{g}_N + ... + b_{1,L-1}\widetilde{g}_{N-L+1} \tag{1.58}$$

$$\hat{g}_{N+1} = a_{2,1}\widetilde{f}_N + ... + a_{2,L-1}\widetilde{f}_{N-L+1} + b_{2,1}\widetilde{g}_N + ... + b_{2,L-1}\widetilde{g}_{N-L+1} \tag{1.59}$$

where $R_{11} = (a_{1,L-1}, a_{1,L-2}, ..., a_{1,1})^T$ is given by

$$R_{11} = \frac{1}{\det A}\big((1 - \sum_{k=i_1}^{i_p} \eta_{k,2L}^2) \sum_{k=i_1}^{i_p} \eta_k^{(1)} \eta_{k,L} + \sum_{k=i_1}^{i_p} \eta_{k,2L}\eta_{k,L} \sum_{k=i_1}^{i_p} \eta_k^{(1)} \eta_{k,2L}\big), \tag{1.60}$$

$R_{12} = (b_{1,L-1}, b_{1,L-2}, ..., b_{1,1})^T$ is given by

$$R_{12} = \frac{1}{\det A}\big((1 - \sum_{k=i_1}^{i_p} \eta_{k,2L}^2) \sum_{k=i_1}^{i_p} \eta_k^{(2)} \eta_{k,L} + \sum_{k=i_1}^{i_p} \eta_{k,2L}\eta_{k,L} \sum_{k=i_1}^{i_p} \eta_k^{(2)} \eta_{k,2L}\big), \tag{1.61}$$

$R_{21} = (a_{2,L-1}, a_{2,L-2}, ..., a_{2,1})^T$ is given by

$$R_{21} = \frac{1}{\det A}\big((1 - \sum_{k=i_1}^{i_p} \eta_{k,L}^2) \sum_{k=i_1}^{i_p} \eta_k^{(1)} \eta_{k,2L} + \sum_{k=i_1}^{i_p} \eta_{k,L}\eta_{k,2L} \sum_{k=i_1}^{i_p} \eta_k^{(1)} \eta_{k,L}\big), \tag{1.62}$$

and $R_{22} = (b_{2,L-1}, b_{2,L-2}, ..., b_{2,1})^T$ is given by

$$R_{22} = \frac{1}{\det A}\big((1 - \sum_{k=i_1}^{i_p} \eta_{k,L}^2) \sum_{k=i_1}^{i_p} \eta_k^{(2)} \eta_{k,2L} + \sum_{k=i_1}^{i_p} \eta_{k,L}\eta_{k,2L} \sum_{k=i_1}^{i_p} \eta_k^{(2)} \eta_{k,L}\big) \tag{1.63}$$

Throughout this thesis we are using the MSSA forecast algorithm.

To summarize this chapter, we discussed the basic stages of the SSA technique in details. Theoretically there is no model restrictions for the current analysis. The technique is flexible and may be applied to general time series. We looked at the reconstruction and forecast algorithms for two cases, univariate and multivariate.

The main idea of SSA is to employ a time series in a multidimensional form as a trajectory matrix, which makes it possible to treat the self-covariance in the structure of the time series through spectral analysis. Specifically, by choosing parts of the spectral decomposition of the matrix we then reconstruct a partial one-dimensional time series.

We have introduced the SSA and MSSA reconstructions as linear convolution filter, which later are required for the stability analysis we discuss in Chapter 5.

We have also described forecast algorithms for univariate and bivariate cases. The key idea of the forecast in terms of SSA/MSSA is to minimize the distance between the forecast value and the subspace, spanned by eigenvectors obtained from the original trajectory matrix of one or two time series, which are involved in reconstruction. For the bivariate forecast MSSA algorithm we considered two cases: MSSA_+ forecast LRF if future values of the support series are know and standard MSSA LRF otherwise.

2. GRANGER CAUSALITY COHERENCE

One of the most known and widely used concepts of causality in time series analysis and econometrics is the concept which was introduced by C.W.J. Granger [15]. However, this concept was defined and developed for time series corresponding to the properties of stationary autoregressive model.

Before introducing the concept itself we first need to describe the class of autoregressive processes and study its main properties. Also, looking ahead, we need the formal description of autoregressive processes to be able to compare them with the time series models obtained using (M)SSA.

2.1 The autoregressive model

In this section we consider discrete-time stochastic processes which are stationary in the wide sense.

The stochastic process X_t is called *strictly stationary* if its cumulative distribution function (F_X) for $(k+1)$ consecutive times is not dependent on shift in time, for all $k \in \mathbb{N}_0$:

$$F_X(X_t, ...X_{t+k}) = F_X(X_{t+\tau}, ...X_{t+k+\tau}), \tau \in \mathbb{Z}. \tag{2.1}$$

It is called *weakly stationary* or *stationary in the wide sense* if its expectation and variance are constant and finite, and the covariance $cov(X_t, X_{t+\tau})$ does not depend on t, for any $\tau \in \mathbb{N}$, i.e.

1. $EX_t < \infty$,
2. $\sigma^2_{X_t} < \infty$,
3. $cov[X_t X_{t+\tau}]$ does not depend on time t for any τ.

The process X is denoted autoregressive moving average (ARMA(p,q)) process with mean zero if (X_t) is stationary with mean $EX_t = 0$ and can be written as

$$X_t - \alpha_1 X_{t-1} - ... - \alpha_p X_{t-p} = \varepsilon_t + \beta_1 \varepsilon_{t-1} + ... + \beta_q \varepsilon_{t-q}, \tag{2.2}$$

where α_i, β_i are the parameters of the model and ε_t is i.i.d. white noise with mean zero and common variance σ^2_ε. More generally, X_t is an ARMA(p,q) process with mean $EX_t = m$ if $(X_t - m)$ is an ARMA(p,q) process with mean zero.

The autoregressive (AR) process is a special case of an ARMA(p,q) process describing a stochastic linear dependence between present and past values of the same time series.

The autoregressive process of order p, AR(p), is ARMA(p,0), i.e. (2.2) simplifies to

$$X_t - \alpha_1 X_{t-1} - ... - \alpha_p X_{t-p} = \varepsilon_t. \tag{2.3}$$

The moving average process of order q, MA(q), is ARMA(0,q) process, i.e. (2.2) simplifies to

$$X_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + ... + \beta_q \varepsilon_{t-q}. \tag{2.4}$$

In particular, the infinite moving average process, MA(∞) is defined as

$$X_t = \sum_{k=0}^{\infty} \beta_t \varepsilon_{t-k}. \tag{2.5}$$

Lemma 2.1.1. *The MA(∞) process (2.5) is stationary if its coefficients are absolutely summable*

$$\sum_{i=1}^{\infty} \mid \beta_i \mid < \infty$$

[19, p.69-70]

Proof. The process $X_t = \sum_{j=1}^{\infty} \beta_j \varepsilon_{t-j}$ is stationary in the wide sense if:

1. The expectation is finite

$$EX_t = \sum_{j=1}^{\infty} E\beta_j \varepsilon_{t-j} = \sum_{j=1}^{\infty} \beta_j \underbrace{E\varepsilon_{t-j}}_{=0} \Rightarrow EX_t = 0, \tag{2.6}$$

2. since $\sum_{i=1}^{\infty} \mid \beta_i \mid < \infty$, there $\exists N :\mid \beta_j \mid < 1$ for $j \geq N \Rightarrow \beta_j^2 <\mid \beta^j \mid$, therefore, by comparison test, $\sum_{i=1}^{\infty} \beta_i^2 < \infty$.

   Hence, the variance is finite

$$\sigma_X^2 = \sum_{j=1}^{\infty} \beta_j^2 < \infty. \tag{2.7}$$

3. The autocovariance function of X_t is independent of t

$$\begin{aligned} E[X_t X_{t-\tau}] &= E[\textstyle\sum_{j=1}^{\infty} \beta_j \varepsilon_{t-j} \sum_{k=1}^{\infty} \beta_k \varepsilon_{t-\tau-k}] \\ &= E[\textstyle\sum_{l=1}^{\infty} \sum_{m=1}^{l} \beta_m \varepsilon_{t-m} \beta_{l+1-m} \varepsilon_{t-\tau-l-1+m}] \\ &= \textstyle\sum_{l=1}^{\infty} \sum_{m=1}^{l} \beta_m \beta_{l+1-m} E[\varepsilon_{t-m} \varepsilon_{t-\tau-l-1+m}] \\ &= \textstyle\sum_{l=1}^{\infty} \sum_{m=1}^{l} \beta_m \beta_{l+1-m}\ \sigma_\varepsilon^2\ \delta_{2m,\tau+l+1}, \end{aligned} \tag{2.8}$$

   where $E[\varepsilon_{t-m}\ \varepsilon_{t-\tau-l-1+m}] = 0$, unless $t-m = t-\tau-l-1+m \Rightarrow$ the autocovariance function is independent on t.

From above we conclude that MA(∞) is stationary. $\square$

Using the shift operator S defined in (1.12), we can formally rewrite (2.3) as

$$X_t - \alpha_1 S^{-1} X_t - ... - \alpha_p S^{-p} X_t = \varepsilon_t \tag{2.9}$$

or equivalently

$$(1 - \alpha_1 S^{-1} - ... - \alpha_p S^{-p}) X_t = \varepsilon_t \tag{2.10}$$

For the purposes of this section only, we work in the Hilbert space $\ell^2(\mathbb{Z}) = \{X_{t(t\in\mathbb{Z})} : \sum_{t\in\mathbb{Z}} |X_t|^2 < \infty\}$. The operator S is a bounded operator on this space. Let $D = \sum_{j=1}^{p} \alpha_j S^{-j}$, s.t.

$$(1 - D) X_t = \varepsilon_t \tag{2.11}$$

The operator D is a linear operator on Hilbert space with the characteristic polynomial

$$1 - \alpha_1 z - \alpha_2 z^2 - ... - \alpha_p z^p. \tag{2.12}$$

The roots of this characteristic polynomial are the eigenvalues $\pi_i (i \in \{1, ..., p\})$ of the operator D.

If $(1 - D)^{-1}$ exists, it can be formally expressed as Neumann series

$$\begin{aligned} &(1 - D)^{-1} = \sum_{j=0}^{\infty} D^j = \sum_{j=0}^{\infty} (\alpha_1 S^{-1} + \alpha_2 S^{-2} + ... + \alpha_p S^{-p}) = \\ &= 1 + \alpha_1 S^{-1} + \alpha_2 S^{-2} + ... + \alpha_p S^{-p} + (\alpha_1 S^{-1} + \alpha_2 S^{-2} + ... + \alpha_p S^{-p})^2 + ... \end{aligned} \tag{2.13}$$

This gives rise to a power series

$$(1 - D)^{-1} = 1 + \hat{\beta}_1 S^{-1} + \hat{\beta}_2 S^{-2} + \hat{\beta}_3 S^{-3} + ... + \hat{\beta}_p S^{-p} + ... = \sum_{j=0}^{\infty} \hat{\beta}_j S^{-j} \tag{2.14}$$

where we define $\hat{\beta}_0 = 1$, and (2.11) can be formally rewritten as

$$X_t = (1 - D)^{-1} \varepsilon_t. \tag{2.15}$$

From (2.15) and (2.14), X_t is a MA(∞) process,

$$X_t = \sum_{j=0}^{\infty} \hat{\beta}_j \varepsilon_{t-j}. \tag{2.16}$$

Assume that the hypothesis of the Lemma 2.1.1 is satisfied, i.e.

$$\sum_{i=0}^{\infty} | \hat{\beta}_i | < \infty. \tag{2.17}$$

Then we can conclude that the process (2.16) is stationary.

Now under the assumption that the process (2.16) is stationary and assumption that we are working in the Hilbert space described above, using the triangle inequality and knowing that the operator norm of the shift operator is $\| S^{-i} \| = 1$, we obtain

$$\| \sum_{i=k}^{m} \hat{\beta}_i S^{-i} \| \leq \sum_{i=k}^{m} | \hat{\beta}_i | \| S^{-i} \| = \sum_{i=k}^{m} | \hat{\beta}_i | \underset{k,m\to\infty}{\longrightarrow} 0 \tag{2.18}$$

which shows that the series (2.14) is norm convergent.

As the process (2.16) is stationary, which is equivalent to the stationarity of the AR process (2.9), then $cov[X_t X_{t-\tau}]$ does not depend on t $\forall\tau$. Multiplying (2.3) by $X_{t-\tau}$ and taking the expectation, we get

$$E[X_t X_{t-\tau}] - E[\alpha_1 X_{t-1} X_{t-\tau}] - ... - E[\alpha_p X_{t-p} X_{t-\tau}] = E[\varepsilon_t X_{t-\tau}] \tag{2.19}$$

where

$$E[\varepsilon_t X_{t-\tau}] = \sum_{j=1}^{\infty} E\hat{\beta}_j \varepsilon_{t-\tau-j} \varepsilon_t = 0, \text{ for all } k \in \mathbb{Z}_0 \tag{2.20}$$

so

$$E[X_t X_{t-\tau}] - \alpha_1 E[X_{t-1} X_{t-\tau}] - ... - E[X_{t-p} X_{t-\tau}] = 0. \tag{2.21}$$

Now assuming that X_t is stationary, we get

$$E[X_t X_{t-\tau}] - \alpha_1 E[X_t X_{t-\tau+1}] - ... - E[X_t X_{t-\tau+p}] = 0, \tag{2.22}$$

or equivalently

$$R(\tau) = \alpha_1 R(\tau - 1) + ... + \alpha_p R(\tau - p), \tag{2.23}$$

where $R(\tau) = E[X_t X_{t-\tau}]$ is the autocovariance function.
Because $\sigma^2_{X_t}$ is constant, it follows that the autocorrelation function $\rho(\tau) = \frac{R(\tau)}{\sigma^2_X}$ satisfies the recurrence.

$$\rho(\tau) = \alpha_1 \rho(\tau - 1) + ... + \alpha_p \rho(\tau - p) \tag{2.24}$$

is the autocorrelation function. The difference equation (2.24) for the autocorrelation function has the general solution

$$\rho(\tau) = A_1 \pi_1^{|\tau|} + ... + A_p \pi_p^{|\tau|} \ (\tau \in \mathbb{Z}), \tag{2.25}$$

where π_i are the roots of the characteristic polynomial

$$z^p - \alpha_1 z^{p-1} - ... - \alpha_{p-1} z - \alpha_p = 0, \tag{2.26}$$

see [4, A4.1]. The coefficients A_i in (2.26) are arbitrary constants for the autocorrelation function of a particular solution of (2.26), such that $\sum_i A_i = 1$, which provides $\rho(0) = 1$.

Proposition 2.1.2. *The roots of characteristic polynomial of (2.24) are all in the unit circle if and only if* $\| D \| < 1$.

Proof. The linear normal operator D has the same characteristic polynomial as (2.26), π_i are its eigenvalues.

To show that the operator D is normal one needs to show that $DD^* = D^*D$, where D^* is an adjoint of D.

$$D^* = (\alpha_1 S^{-1} + \alpha_2 S^{-2} + ... + \alpha_p S^{-p})^* = \alpha_1 (S^{-1})^* + \alpha_2 (S^{-2})^* + ... + \alpha_p (S^{-p})^* \quad (2.27)$$

and $(S^{-j})^* = S^j$, because S^{-j} is a unitary operator. The equation (2.27) can be rewritten as

$$D^* = \alpha_1 S + \alpha_2 S^2 + ... + \alpha_p S^p, \quad (2.28)$$

and then

$$DD^* = \sum_{j=1}^{p} \alpha_j S^{-j} \sum_{k=1}^{p} \alpha_k S^k = \sum_{k=1}^{p} \sum_{j=1}^{p} \alpha_j \alpha_k S^{-j} S^k = \sum_{k=1}^{p} \sum_{j=1}^{p} \alpha_j \alpha_k S^k S^{-j} = D^*D. \quad (2.29)$$

Since the D operator is normal, its norm is equal to its spectral radius [22, I(6.44)]

$$\| D \| = \max_{i \in \{1,...,p\}} | \pi_i | . \quad (2.30)$$

If all π_i are inside the unit circle, then the norm $\| D \| = max | \pi_i | < 1$. □

Same results were obtained in a different form in [4, p.51-52].

Remark If roots of the characteristic polynomial (2.12) are less than one, then $\| D \| < 1$, which provides the existence of the inverse $(1 - D)^{-1}$ and $\lim_{k \to \infty} \rho(k) = 0$. Additionally, $\| D \| < 1$ implies the convergence of Neumann series (2.13)

To sum up the above discussion, the AR(p) process can be rewritten as an MA(∞) process with absolutely summable coefficients, which ensures its stationarity.

2.2 Granger causality

The causality concept in general is very intuitive and small changes in the definition may cause bigger changes in the way we understand it and approach it. It is necessary to refer to the origins of the concept: the definition of causality, its proposed measures and tests to verify such measures.

As it was mentioned in the introduction to this chapter, Granger's causality concept was more or less fully described in Granger's paper [15]. There he gave the main definitions of causal relations for stationary stochastic processes, which are now widely used in statistics and econometrics. In this paper, he distinguishes several causal relationships: a simple causal model, instantaneous causality and a feedback model. And also proposed ways of detecting and measuring causality, if one occurs.

Definition The intuitive definition of *Simple Causality* states the following: suppose we have two stationary stochastic processes F_N and G_N. It is said that G_N is *causing* F_N ($G_N \rightsquigarrow F_N$) if F_N is better predictable with all available past information, rather than with all available past information excluding the past and present values of G_N [15].

Feedback occurs if simple causality appears in both directions ($F_N \leftrightsquigarrow G_N$).

Definition *Instantaneous Causality* occurs when the time series F_N is better predicted if the **present** value of G_N is included in the forecasting process [15].

The general model used for instantaneous causality is a *stationary* bi-variate (vector) autoregressive (VAR) model of order m:

$$\begin{aligned} X_t + b_0 Y_t &= \sum_{j=1}^{m} a_j X_{t-j} + \sum_{j=1}^{m} b_j Y_{t-j} + \varepsilon_t \\ Y_t + c_0 X_t &= \sum_{j=1}^{m} c_j X_{t-j} + \sum_{j=1}^{m} d_j Y_{t-j} + \xi_t \end{aligned} \tag{2.31}$$

which can be written as

$$\begin{aligned} X_t &= \sum_{j=1}^{m} a_j X_{t-j} + \sum_{\mathbf{j=0}}^{m} b_j Y_{t-j} + \varepsilon_t \\ Y_t &= \sum_{\mathbf{j=0}}^{m} c_j X_{t-j} + \sum_{j=1}^{m} d_j Y_{t-j} + \xi_t \end{aligned} \tag{2.32}$$

Here X_t and Y_t are some multidimensional stationary time series, both with mean zero, and ε_t, ξ_t are some uncorrelated white noise.

Definition White noise ε_t is a strictly stationary stochastic process with mean zero and variance σ_ε^2, and with autocovariance function

$$E[\varepsilon_k \varepsilon_{k+\tau}] = \begin{cases} \sigma_\varepsilon^2 & \text{if } \tau = 0 \\ 0, & \text{otherwise} \end{cases} \tag{2.33}$$

where ε_i are independent and identically distributed (i.i.d.) and uncorrelated random variables.

For the simple causality model, the instantaneous terms are removed: $b_0 = c_0 = 0$.

Lemma 2.2.1. *Let (2.32) be a stationary VAR model with $b_0 = c_0 = 0$. Then, without loss of generality, one can subtracts mean to get a stationary process with mean zero*

Proof. Calculating the expectation for the simple causality model, we find

$$\begin{aligned} EX_t &= E\sum_{j=1}^{m} a_j X_{t-j} + E\sum_{j=1}^{m} b_j Y_{t-j} + \underbrace{E\varepsilon_t}_{=0} \\ EY_t &= E\sum_{j=1}^{m} c_j X_{t-j} + E\sum_{j=1}^{m} d_j Y_{t-j} + \underbrace{E\xi_t}_{=0}. \end{aligned} \tag{2.34}$$

Thus

$$\begin{aligned} EX_t &= \sum_{j=1}^{m} a_j EX_{t-j} + \sum_{j=1}^{m} b_j EY_{t-j} \\ EY_t &= \sum_{j=1}^{m} c_j EX_{t-j} + \sum_{j=1}^{m} d_j EY_{t-j} \end{aligned} \tag{2.35}$$

and defining $\widetilde{X_t} = EX_t$ and $\widetilde{Y_t} = EY_t$, we get two non-random deterministic

$$\begin{aligned} \widetilde{X_t} &= \sum_{j=1}^{m} a_j \widetilde{X}_{t-j} + \sum_{j=1}^{m} b_j \widetilde{Y}_{t-j} \\ \widetilde{Y_t} &= \sum_{j=1}^{m} c_j \widetilde{X}_{t-j} + \sum_{j=1}^{m} d_j \widetilde{Y}_{t-j} \end{aligned} \tag{2.36}$$

If we consider the differences $F = X - \widetilde{X}$, $G = Y - \widetilde{Y}$, then their expectations are zero, and we can show that they also satisfy the model (2.31):

$$\begin{aligned} X_t - EX_t &= \sum_{j=1}^{m} a_j X_{t-j} + \sum_{j=1}^{m} b_j Y_{t-j} + \varepsilon_t - \sum_{j=1}^{m} a_j EX_{t-j} + \sum_{j=1}^{m} b_j EY_{t-j} \\ Y_t - EY_t &= \sum_{j=1}^{m} c_j X_{t-j} + \sum_{j=1}^{m} d_j Y_{t-j} + \xi_t - \sum_{j=1}^{m} c_j EX_{t-j} + \sum_{j=1}^{m} d_j EY_{t-j} \end{aligned} \tag{2.37}$$

or

$$\begin{aligned} X_t - EX_t &= \sum_{j=1}^{m} a_j (X_{t-j} - EX_{t-j}) + \sum_{j=1}^{m} b_j (Y_{t-j} - EY_{t-j}) + \varepsilon_t \\ Y_t - EY_t &= \sum_{j=1}^{m} c_j (X_{t-j} - EX_{t-j}) + \sum_{j=1}^{m} d_j (Y_{t-j} - EY_{t-j}) + \xi_t \end{aligned} \tag{2.38}$$

which is equivalent to

$$\begin{aligned} F_t &= \sum_{j=1}^{m} a_j F_{t-j} + \sum_{j=1}^{m} b_j G_{t-j} + \varepsilon_t \\ G_t &= \sum_{j=1}^{m} c_j F_{t-j} + \sum_{j=1}^{m} d_j G_{t-j} + \xi_t. \end{aligned} \tag{2.39}$$

Obtained model (2.39) is identical to VAR model (2.31) with $b_0 = c_0 = 0$ (simple causality case). This shows that when one uses the VAR model, without loss of generality one can subtract mean and assume that the processes have mean zero. □

Stationarity and zero mean imply that we can use the Cramer representation for F_t (and analogously for G_t) [15]:

$$F_t = \int_{-\pi}^{\pi} e^{it\omega} dZ_F(\omega) \tag{2.40}$$

where the integrator Z_F is a complex stochastic process with uncorrelated increments [15], [30, p.558].

Proposition 2.2.2. *Let (2.39) be a stationary VAR process with mean zero. Then the power spectra of processes F and G are*

$$\begin{aligned} f_F(\omega) &= \tfrac{1}{2\pi|\delta|^2}\left(\mid 1-d \mid^2 \sigma_\varepsilon^2 + \mid b \mid^2 \sigma_\xi^2\right) \\ f_G(\omega) &= \tfrac{1}{2\pi|\delta|^2}\left(\mid c \mid^2 \sigma_\varepsilon^2 + \mid 1-a \mid^2 \sigma_\xi^2\right) \end{aligned} \tag{2.41}$$

and the cross spectrum is

$$f_{FG}(\omega) = \frac{1}{2\pi \mid \delta \mid^2}(1-d(\omega))\overline{c(\omega)}\sigma_\varepsilon^2 + \frac{1}{2\pi \mid \delta \mid^2}\overline{(1-a(\omega))}b(\omega)\sigma_\xi^2 \tag{2.42}$$

Remark *This proposition is fully based on the Granger's paper [15].*

Proof. Knowing that processes F_t and G_t have Cramer representation, using the (2.40) we have

$$F_{t-j} = \int_{-\pi}^{\pi} e^{i(t-j)\omega} dZ_F(\omega) \tag{2.43}$$

so that the sum $\sum_{j=1}^m a_j F_{t-j}$ can be rewritten as

$$\begin{aligned} \textstyle\sum_{j=1}^m a_j F_{t-j} &= \textstyle\sum_{j=1}^m a_j \int_{-\pi}^{\pi} e^{i(t-j)\omega} dZ_F(\omega) \\ &= \textstyle\int_{-\pi}^{\pi} e^{it\omega} \sum_{j=1}^m a_j (e^{-i\omega})^j dZ_F(\omega) \\ &= \textstyle\int_{-\pi}^{\pi} e^{it\omega} a(e^{i\omega}) dZ_F(\omega) \end{aligned} \tag{2.44}$$

where we define $a(z) = \sum_{j=1}^m a_j z^j \ (z \in \mathbb{C})$.
Thus, the simple causal model (2.38) can be expressed in the following manner

$$\begin{aligned} \textstyle\int_{-\pi}^{\pi} e^{it\omega} dZ_F(\omega) &= \textstyle\int_{-\pi}^{\pi} a e^{it\omega}(e^{-i\omega}) dZ_F(\omega) + \int_{-\pi}^{\pi} b e^{it\omega}(e^{-i\omega}) dZ_G(\omega) + \int_{-\pi}^{\pi} e^{it\omega} dZ_\varepsilon(\omega) \\ \textstyle\int_{-\pi}^{\pi} e^{it\omega} dZ_G(\omega) &= \textstyle\int_{-\pi}^{\pi} c e^{it\omega}(e^{-i\omega}) dZ_F(\omega) + \int_{-\pi}^{\pi} d e^{it\omega}(e^{-i\omega}) dZ_G(\omega) + \int_{-\pi}^{\pi} e^{it\omega} dZ_\xi(\omega) \end{aligned} \tag{2.45}$$

where $Z_G, Z_\varepsilon, Z_\xi$ are stochastic integrators analogous to Z_F.

Now, bringing all except the noise components to the left hand side and rewriting the set of equations in a matrix form, we get

$$\int_{-\pi}^{\pi} e^{it\omega} A \begin{bmatrix} dZ_F(\omega) \\ dZ_G(\omega) \end{bmatrix} = \int_{-\pi}^{\pi} e^{it\omega} \begin{bmatrix} dZ_\varepsilon(\omega) \\ dZ_\xi(\omega) \end{bmatrix} \text{ for all } t \in \mathbb{Z} \tag{2.46}$$

where

$$A = \begin{bmatrix} (1-a(e^{-i\omega})) & -b(e^{-i\omega}) \\ -c(e^{-i\omega}) & (1-d(e^{-i\omega})) \end{bmatrix}. \tag{2.47}$$

If A is invertible, then we conclude

$$\begin{bmatrix} dZ_F(\omega) \\ dZ_G(\omega) \end{bmatrix} = A^{-1} \begin{bmatrix} dZ_\varepsilon(\omega) \\ dZ_\xi(\omega) \end{bmatrix} \tag{2.48}$$

for almost all $\omega \in [-\pi, \pi]$.

The power spectral density functions f_F and f_G of the series F and G respectively can be obtained from the expectation of (2.48).

On one hand, the expectation is

$$E \begin{bmatrix} dZ_F(\omega) \\ dZ_G(\omega) \end{bmatrix} \begin{bmatrix} \overline{dZ_F(\lambda)} & \overline{dZ_G(\lambda)} \end{bmatrix} = \begin{bmatrix} f_F(\omega)d\omega & f_{FG}(\omega)d\omega \\ f_{GF}(\omega) & f_G(\omega)d\omega \end{bmatrix} \tag{2.49}$$

where f_{FG} and f_{GF} correspond to the cross-spectrum.

And on the other hand, using the right hand side of the equation (2.48), the expectation is

$$\begin{aligned} & E \begin{bmatrix} dZ_F(\omega) \\ dZ_G(\omega) \end{bmatrix} \begin{bmatrix} \overline{dZ_F(\lambda)} & \overline{dZ_G(\lambda)} \end{bmatrix} \\ & = A^{-1} E \begin{bmatrix} dZ_\varepsilon(\omega) \\ dZ_\xi(\omega) \end{bmatrix} \begin{bmatrix} \overline{dZ_\varepsilon(\lambda)} & \overline{dZ_\xi(\lambda)} \end{bmatrix} \overline{(A^{-1})^T} \\ & = A^{-1} \begin{bmatrix} E[dZ_\varepsilon(\omega)\overline{dZ_\varepsilon(\lambda)}] & E[dZ_\varepsilon(\omega)\overline{dZ_\xi(\lambda)}] \\ E[dZ_\xi(\omega)\overline{dZ_\varepsilon(\lambda)}] & E[dZ_\xi(\omega)\overline{dZ_\xi(\lambda)}] \end{bmatrix} \overline{(A^{-1})^T} \end{aligned} \tag{2.50}$$

For $\omega = \lambda$ we express spectral density functions f_ξ and f_ε of ξ and ε respectively, using the fact that ξ and ε are uncorrelated

$$\begin{aligned} E[dZ_\varepsilon(\omega)\overline{dZ_\varepsilon(\omega)}] &= f_\varepsilon(\omega)d\omega = \frac{\sigma_\varepsilon^2}{2\pi}d\omega \\ E[dZ_\xi(\omega)\overline{dZ_\xi(\omega)}] &= f_\xi(\omega)d\omega = \frac{\sigma_\xi^2}{2\pi}d\omega \\ E[dZ_\varepsilon(\omega)\overline{dZ_\xi(\omega)}] &= \overline{E[dZ_\varepsilon(\omega)\overline{dZ_\xi(\omega)}]} = 0 \end{aligned} \tag{2.51}$$

and for $\omega \neq \lambda$ all terms vanish:

$$E[dZ_\varepsilon\overline{dZ_\varepsilon}] = E[dZ_\xi\overline{dZ_\xi}] = E[dZ_\varepsilon\overline{dZ_\xi}] = \overline{E[dZ_\varepsilon\overline{dZ_\xi}]} = 0 \tag{2.52}$$

The inverse of the matrix A is

$$A^{-1} = \frac{1}{\delta} \begin{pmatrix} (1 - d(e^{-i\omega})) & b(e^{-i\omega}) \\ c(e^{-i\omega}) & (1 - a(e^{-i\omega})) \end{pmatrix} \tag{2.53}$$

where

$$\delta = \det A = (1 - d(e^{-i\omega}))(1 - a(e^{-i\omega})) - b(e^{-i\omega})c(e^{-i\omega}). \tag{2.54}$$

Then the (2.50) for $\omega = \lambda$ becomes

$$\begin{aligned} &A^{-1} \begin{bmatrix} \frac{\sigma_\varepsilon^2}{2\pi} d\omega & 0 \\ 0 & \frac{\sigma_\xi^2}{2\pi} d\omega \end{bmatrix} \overline{(A^{-1})^T} = \\ &\frac{1}{|\delta|^2} \begin{bmatrix} \left(\mid 1 - d \mid^2 \frac{\sigma_\varepsilon^2}{2\pi} d\omega + \mid b \mid^2 \frac{\sigma_\xi^2}{2\pi} d\omega \right) & \left((1-d)\bar{c}\frac{\sigma_\varepsilon^2}{2\pi} d\omega + \overline{(1-a)} b \frac{\sigma_\xi^2}{2\pi} d\omega \right) \\ \left(c\overline{(1-d)} \frac{\sigma_\varepsilon^2}{2\pi} d\omega + (1-a)\bar{b}\frac{\sigma_\xi^2}{2\pi} d\omega \right) & \left(\mid c \mid^2 \frac{\sigma_\varepsilon^2}{2\pi} d\omega + \mid 1 - a \mid^2 \frac{\sigma_\xi^2}{2\pi} d\omega \right) \end{bmatrix} \end{aligned} \tag{2.55}$$

Therefore the power spectra of F and G are

$$\boxed{\begin{aligned} f_F(\omega) &= \frac{1}{2\pi|\delta|^2} \left(\mid 1 - d \mid^2 \sigma_\varepsilon^2 + \mid b \mid^2 \sigma_\xi^2 \right) \\ f_G(\omega) &= \frac{1}{2\pi|\delta|^2} \left(\mid c \mid^2 \sigma_\varepsilon^2 + \mid 1 - a \mid^2 \sigma_\xi^2 \right) \end{aligned}} \tag{2.56}$$

respectively. As it was mentioned before, $f_{FG}(\omega)$ corresponds to the cross-spectrum, which is of interest for the question of causality,

$$\boxed{\begin{aligned} f_{FG}(\omega) &= \frac{1}{2\pi|\delta|^2} \left((1 - d(\omega))\overline{c(\omega)}\sigma_\varepsilon^2 + \overline{(1 - a(\omega))} b(\omega) \sigma_\xi^2 \right) \\ &= \frac{1}{2\pi|\delta|^2} (1 - d(\omega))\overline{c(\omega)}\sigma_\varepsilon^2 + \frac{1}{2\pi|\delta|^2} \overline{(1 - a(\omega))} b(\omega) \sigma_\xi^2 \end{aligned}} \tag{2.57}$$

□

The power spectrum $f_{FG}(\omega)$ can be expressed as a sum of two components, one associated with F via ε and another one with G via ξ:

$$f_{FG} = C_1(\omega) + C_2(\omega) \tag{2.58}$$

with

$$C_1(\omega) = \frac{1}{2\pi \mid \delta \mid^2} \left((1 - d(\omega))\overline{c(\omega)} \right) \sigma_\varepsilon^2 \tag{2.59}$$

and

$$C_2(\omega) = \frac{1}{2\pi \mid \delta \mid^2} \left(\overline{(1 - a(\omega))} b(\omega) \right) \sigma_\xi^2 \tag{2.60}$$

If the coefficients b_i in model (2.31) are zero, then $b(\omega) = 0$ consequently and $C_2(\omega)$ vanishes, therefore $f_{FG} = C_1(\omega)$. This is interpreted as time series G not causing F. Similarly, if c_i in (2.31) are zero, then $f_{FG} = C_2(\omega)$, and therefore G is not causing F.

In view of this, Granger suggested the following measure for causality, which he defined as *causality coherence* [15]

$$C_{FG}(\omega) = \frac{\mid C_1(\omega) \mid^2}{f_F(\omega) f_G(\omega)} \tag{2.61}$$

$$C_{GF}(\omega) = \frac{\mid C_2(\omega) \mid^2}{f_F(\omega) f_G(\omega)} \tag{2.62}$$

where C_{FG} measures the strength of causality of F upon G ($F \rightsquigarrow G$) and C_{GF} measures the strength of causality of $G \rightsquigarrow F$.

Denominators in (2.61) and (2.62) are added for normalization. Indeed, as

$$f_F(\omega) f_G(\omega) = \frac{1}{4\pi^2 \mid \lambda \mid^4} \left(\mid 1 - d(\omega) \mid^2 \sigma_\varepsilon^2 + \mid b(\omega) \mid^2 \sigma_\xi^2 \right) \left(\mid c(\omega) \mid^2 \sigma_\varepsilon^2 + \mid 1 - a(\omega) \mid^2 \sigma_\xi^2 \right), \tag{2.63}$$

we see that

$$f_F(\omega) f_G(\omega) \geq \frac{\mid 1 - d(\omega) \mid^2 \mid c(\omega) \mid^2 \sigma_\varepsilon^4}{4\pi^2 \mid \lambda \mid^4} \tag{2.64}$$

so

$$f_F(\omega) f_G(\omega) > 0 \tag{2.65}$$

the inequality is strict, as $f_F(\omega) f_G(\omega) = 0$ only if $\sigma_\varepsilon^2 = \sigma_\xi^2 = 0$ and in terms of the autoregressive model, that means that there is no noise to develop the model and therefore, the model becomes deterministic.

Knowing that $\mid C_1(\omega) \mid^2 \geq 0$, we find

$$0 \leq C_{FG} \leq \frac{|1-d|^2 |c|^2 \sigma_\varepsilon^4}{|1-d|^2 |c|^2 \sigma_\varepsilon^4} = 1 \tag{2.66}$$

Thus, the measure is normalized s.t. $0 \leq C_{FG} \leq 1$.

By looking at (2.31) one can see that noise in this type of model plays the crucial role in its development, as the future values contain the randomness of the past.

The expectation value of the process can be subtracted as in equations (2.37) and (2.38); for a VAR model, the structure of the model is encoded in the noise pattern. This explains why the noise variances σ_ε^2 and σ_ξ^2 enter the spectra (2.56) and (2.57) and measures for causality (2.61) and (2.62).

2.3 Granger causality tests

There are several statistical tests for Granger causality estimation, which are widely used in econometrics. The most common of them are: the Granger-Wald test (1969), the Sims test (1972) and its modifications (Geweke, Meese, Dent, 1982) and the Granger-Sargent test (1976) [18].

Consider the following models:

$$X_t = \sum_{j=1}^{L} a_j X_{t-j} + \sum_{j=0}^{L} b_j Y_{t-j} + \varepsilon_t \tag{2.67}$$

and

$$X_t = \sum_{j=1}^{L} \widetilde{a}_j X_{t-j} + \widetilde{\varepsilon}_t \tag{2.68}$$

Here (2.67) is a AR model of the time series X_t based on the past values of time series X_t and Y_t, and (2.68) is a AR model based only on the past values of X_t. Here L is a time lag of last L values from the time series that being used for the regression. To obtain coefficients a_j, b_j and $\widetilde{a}_j$, one would need to fit the real time series X to the above ARMA model. For this purpose there exist several methods, such as least-square and maximum likelihood methods.

For all mentioned tests and their variations in this section, we assume under the null hypothesis that the main series (X_t) is not Granger-caused by the support series (Y_t). In other words, $H_0 : b_j = 0 \ \forall j \in \{1, ..., L\}$. To distinguish between the two models, we call (2.68), the *constrained* when $b_i = 0$ and (2.67) the *unconstrained*.

All tests are based on comparing the variance $\sigma^2\varepsilon$ of the constrained residual error ε_t with the variance $\sigma^2_{\widetilde{\varepsilon}}$ of the unconstrained residual error $\widetilde{\varepsilon}_t$. For example, the Granger test statistic or **Granger-Sargent** statistic [18] is

$$GS = \frac{\sum_{i=1}^{L} \widetilde{\varepsilon}_i^2 - \sum_{i=1}^{L} \varepsilon_i^2}{L} \cdot \frac{(N - (2L + 2))}{\sum_{i=1}^{L} \varepsilon_i^2} \sim \frac{\chi^2_L}{L} \text{ under } H_0 \tag{2.69}$$

here χ^2_L stands for chi squared distribution with L degrees of freedom.

There are other statistics, which are calculated in a similar manner, but normalized differently (for example, Sims test [32] and its modifications [18]).

The **Granger-Wald** test is the adaptation of the Wald test to estimating Granger causality. The Granger-Wald statistic is defined as:

$$GW = N \frac{\hat{\sigma}^2_{\widetilde{\varepsilon}} - \hat{\sigma}^2_{\varepsilon}}{\hat{\sigma}^2_{\varepsilon}} \sim \chi^2_L \text{ under } H_0 \tag{2.70}$$

where again $\hat{\sigma}_\varepsilon$ and $\hat{\sigma}_{\widetilde{\varepsilon}}$ are the estimators of variances of the noise ε_t and $\widetilde{\varepsilon}_t$ respectively. The equation for GW (2.70) can be rewritten as

$$GW = N \left(\frac{\hat{\sigma}^2_{\widetilde{\varepsilon}}}{\hat{\sigma}^2_{\varepsilon}} - 1 \right) \tag{2.71}$$

which points to the main similarity with the F-test, since $\frac{\hat{\sigma}^2_{\widetilde{\varepsilon}}}{\hat{\sigma}^2_{\varepsilon}}$ is of the form of an F-statistic, as in the F-test, the statistics is based on the ratio between error variances (see the description of F-test in section 3.3).

Tab. 2.1: Granger test p-values with null hypothesis s.s. $\not\rightsquigarrow$ m.s.

time lag L	case (1)	case (2)	case (3)
2	0	0	0.014
3	0	0	0.006
7	0	0	0.03
10	0	0	0.12
50	$5.5 \cdot 10^{-15}$	$7.2 \cdot 10^{-5}$	0.43

Now recalling the Granger's measure for causality, which is calculated from the cross spectrum of analyzed process (2.62), one would not see obvious connection between the F-test and the calculation of causality coherence. However, Granger's causality tests are all derived from the F-test, based on the idea of error comparison.

The R project for statistical computing (package MSBVAR) provides us with the common Granger causality test (granger.test). This test is used for "bivariate Granger causality testing for multiple time series". *The null hypothesis is that the past values of some time series Y do not improve the prediction of the time series X.* As a result of the test, we get two regressions of X_t based on lag time L: one is based on the L past values of X_t, and another one is based on L past values of X_t and L past values Y_t. By means of F-test, it is determined whether the coefficients of the past values of Y are jointly zero [5]. If we reject the null hypothesis, we support the presence of Granger causality between time series.

The description of Granger causality test provided in R documentation [5], suggests that as a test they have chosen straightforward F-test, described here in section 3.3.

To try out Granger causality test described above we generated several different sets of time series. Note that generated series are not all stationary AR processes. The reason to choose some time series, which differ from AR processes, is to see how the test is implemented if the data is out of the common class of usage. In fact, the Granger test is often used for the time series which cannot be considered AR processes, moreover, stationary AR processes.

However, we start with AR processes. Firstly, using MATLAB command randn, we generated normally distributed i.i.d. pseudo numbers, i.e. white noise $\varepsilon_n \sim N(0,1)$ and $\xi_n \sim N(0,1)$, which correspond to the very common case of stationary AR processes. Then we take two cases, which should pick up causality and one which should not. Let $n \in \{1, ..., 200\}$ and $m = n - 1$.

1. main series (m.s.) $sin(\varepsilon_m) + 10^{-10}\xi_m$ (is zero for $n = 1$), support series (s.s.) ε_n;
2. m.s. $(\varepsilon_m + \xi_m)$ (is zero for $n = 1$), s.s. ξ_n;
3. m.s. ε_n, s.s. ξ_n.

We would expect causality in (1) and (2). Although processes in (3) come from the same distribution, we still do not expect any causality, as the processes are still different.

All p-values in Table (2.1) for the time lag $L < 10$ reject the null hypothesis that the support time series do not cause main series (p-value$\leq 0.05 \Rightarrow$ s.s. $\rightsquigarrow$ m.s.). Looking at

Tab. 2.2: Granger test p-values with null hypothesis m.s. $\not\rightarrow$ s.s.

time lag L	case (1)	case (2)	case (3)
2	0.46	0.98	0.99
3	0.15	0.96	0.99
7	0.44	0.87	0.9
10	0.59	0.91	0.95
50	0.77	0.81	0.92

Tab. 2.3: Granger test p-values with null hypothesis $y^{(1)} \not\rightarrow x^{(1)}$

time lag L	$x^{(1)}(0.1), y^{(1)}(1)$	$x^{(1)}(0.01), y^{(1)}(1)$	$x^{(1)}(0.1), y^{(1)}(2)$
10	$3.03 \cdot 10^{-10}$	$2.62 \cdot 10^{-10}$	$7.82 \cdot 10^{-10}$
25	0.417	0.228	0.488
50	0.76	0.478	0.664

the causality in different direction, the null hypothesis is supported by p-values in Table 2.2 for all three cases. However, one would expect absence of causality between two processes described in case (3). And it seems that the smaller is the chosen time lag, the more difficult it is for the test to pick up non-causal relationship. But if the time lag L is large enough, then p-values support null hypothesis both ways (see Tables 2.1,2.2, p-values for $L = 10$ and 50). The random variables described in case three were generated randomly and independently one from another and, therefore, there should be no connection whatsoever.

The next set of examples are represented by series which are not in the class of stationary AR processes. Here are the examples of main series:

$$\begin{aligned} &x_n^{(1)}(\sigma^2) = sin(\pi n\omega) + \sigma\varepsilon_n \qquad \varepsilon_n \sim WN(0,1), \\ &x_n^{(2)} = sin(3\pi n\omega), \\ &x_n^{(3)}(\sigma^2) = sin(\pi n\omega) + sin(3\pi n\omega) + \sigma\varepsilon_n \qquad \varepsilon_n \sim WN(0,1) \end{aligned} \tag{2.72}$$

where $WN(\mu, \sigma^2)$ stands for normally distributed white noise with mean μ and variance σ^2, and support series

$$\begin{aligned} &y_n^{(1)}(\lambda) = \lambda sin(\pi n\omega), \\ &y_n^{(2)} = e^{\pi n\omega}, \end{aligned} \tag{2.73}$$

where $n \in \{1, ..., 200\}$, $\omega = 0.05$, $\sigma^2 \in \{0.1, 0.01\}$ and $\lambda \in \{0.1, 1, 2, 10\}$.

The three tables related to this subsection, Tables 2.3 - 2.5, contain p-values of the Granger test (obtained with R) with the null hypothesis that the past L values of y do not help in predicting the value of x ($y^{(i)} \not\rightarrow x^{(j)}$).

All three Tables 2.3 - 2.5 have the same consistent pattern: the bigger the time lag L, the bigger is the p-value. The values in Table 2.3 for $L = 10$ are very close to zero. Therefore, based on these values, we are rejecting the null hypothesis and conclude that

Tab. 2.4: Granger test p-values with null hypothesis $y \not\rightsquigarrow x$

L	$x^{(2)}, y^{(1)}(1)$	$x^{(1)}(0), y^{(2)}$
10	0.661	0.649
25	0.371	0.118
50	0.52	0.954

Tab. 2.5: Granger test p-values with null hypothesis $y^{(1)} \not\rightsquigarrow x^{(3)}$

L	$x^{(3)}(0.1), y^{(1)}(1)$	$x^{(3)}(0.1), y^{(1)}(0.1)$	$x^{(3)}(0.1), y^{(1)}(10)$
10	$3.33 \cdot 10^{-16}$	$3.33 \cdot 10^{-16}$	$3.33 \cdot 10^{-16}$
25	0.273	0.273	0.273
50	0.482	0.482	0.482

$y^{(1)} \rightsquigarrow x^{(1)}$. However, if we are taking bigger lag L, based on the results in Table 2.3, we cannot reject the null hypothesis. Thus, no convincing conclusions can be achieved.

Nevertheless, the Granger test shows a convincing result for the counter example (see Table 2.4, last column), where we do not expect any causality. All p-values here support the null hypothesis. And in this case there is no monotonic dependence between p-values and the time lag: p-value for time lag $L = 25$ (p_{25}) is smaller then p-value for $L = 10$ (p_{10}), p-value for time lag 10 is smaller then p-value for $L = 50$ (p_{50}) (see Table 2.4, second column).

In Table 2.5 p-values are different, because they were obtained from different runs. As for p-values in Table 2.6, they were obtained from the same run (from identical generated time series) and the only parameter which changes is the constant λ in the support series.

In the next set of examples, the main series x_n differs from the support series y_n only by a fixed shift $\nu = 1, 2, 3$ and the addition of noise.

$$\begin{aligned} &x_n(0.1) = sin(\pi(n-\tau)\omega) + \sigma\varepsilon_n \qquad \tau \in \{1,2,3\}, \\ &y_n(1) = sin(\pi n\omega), \end{aligned} \tag{2.74}$$

in other words, we know that signal y_n is causing signal x_n. Table 2.6 contains Granger test p-values for (2.74) generated time series.

Example (2.74), similarly to previous examples, shows the same pattern for p-values, which seem to be dependent on the chosen time lag: the bigger the lag, the bigger is the

Tab. 2.6: Granger test p-values (chi square p-values) with null hypothesis $y(1) \not\rightsquigarrow x(0.1)$ for the example (2.74) with three different shifts, first trial

L \ τ	1	2	3
10	$1.94 \cdot 10^{-09}$ ($1.73 \cdot 10^{-11}$)	$9.49 \cdot 10^{-09}$ ($1.62 \cdot 10^{-10}$)	$9.46 \cdot 10^{-11}$ ($2.14 \cdot 10^{-13}$)
25	$0.001(9.54 \cdot 10^{-05})$	0.44 (0.427)	0.049 (0.030)
50	0.736 (0.790)	0.558 (0.554)	0.599 (0.612)

Tab. 2.7: Granger test p-values with null hypothesis $y(1) \not\rightarrow x(0.1)$ for the example (2.74) with three different shifts, check trial

L \ τ	1	2	3
10	$1.66 \cdot 10^{-10}$	$1.64 \cdot 10^{-11}$	$2.53 \cdot 10^{-09}$
25	0.057	0.071	0.0005
50	0.282	0.801	0.199

Tab. 2.8: Granger test p-values with null hypothesis $z(0.1) \not\rightarrow x^{(1)}$

L	$x^{(1)}(0), z(0.1)$	$x^{(1)}(0.1), z(0.1)$	$x^{(1)}(0.01), z(0.1)$
10	0.633	0.0002	0.583
25	0.471	0.774	0.815
50	0.893	0.842	0.626

p-value. Also, judging by the values in Table 2.7, one can notice that p-values are sensitive not only to the lag time L, but to the size of the shift τ as well.

Recalling two measures C_1 (2.59) and C_2 (2.60), one can see that if the noise ε in main series is absent, then $C_1 = 0$ and if the noise ξ is absent in the support series, then $C_2 = 0$ and hence, causality coherences $C_{FG} = 0$ (2.61, $F \not\rightarrow G$) and $C_{GF} = 0$ (2.62, $G \not\rightarrow F$). Suppose we have

$$z_n(\sigma^2) = y_n^{(1)}(1) + \sigma\xi,\ \xi \sim WN(0,1) \tag{2.75}$$

with $\sigma^2 = 0.1$

Table 2.8 illustrates that if we have two identical signal components in the main and support series and change the noise level in the main series, then by Granger test there is no clear indication of causal relationship, as the null hypothesis is rejected just for $x^{(1)}(0.1), z(0.1)$ with time lag $L = 10$. In fact, this example shows that the identity of the series does not mean causality in Granger's sense. From Granger's point of view, the causality may occur only between different time series with no redundant information. For example, if there we have some data measured in kilograms and the same data, measured in tonnes, there is no point to include both time series in one analysis [16, p. 330, Axiom B]. However, it is not always clear what should be considered identical, and what redundant information actually means in this case. For example, if two time series have exactly the same periodic but different noise component, it is not clear if the periodic component should be treated as an identical part and be excluded from the analysis or not.

To sum up all the observations, we see that Granger test in general is a very sensitive tool. It seems to be very dependent on the chosen time lag and is sensitive to the shift in the series. Even in the case of obvious causal relation between series, the test might mislead the tester. However, looking at the p-values in Tables 2.5 and 2.3, the Granger test behaves quite consistently under the change of multiplier of the causing series.

3. STATISTICAL TESTS FOR SSA/MSSA CAUSALITY

In the first two chapters we described two essential subjects, on which this work is based. As was discussed in Chapter 2, Granger's causality test by means of the F-test and some slight variations have been used to estimate causality for stationary autoregressive models. There also have been attempts to estimate causality for SSA based model (see [9], [20], [25]).

In this chapter we are applying some basic tools, statistical tests, which can be used for the purpose of causality detection. Here we use a wider definition of causality: time series g causes time series f if the additional information that we get from series g improves the forecast of the series f.

This chapter contains several section. Each section gives a description of the statistical test, that could be used for forecast estimation and hence be useful to determine causality. Each statistical test is validated by several generated data examples. We first study how the test performs under the null hypothesis and/or if the null hypothesis is likely to be rejected. Afterwards we look at several cases when the support series replicates the signal of the main series fully or partially and also we study the case when the signal in the support series has a delay in comparison to the main series.

Note that we are validating following tests only for one forecast point at a time as the aim of this work is to build a causality measure analogous to Granger's measure, which consider one forecast point at a time [15].

3.1 Data and basic test setup

To find and validate a suitable criterion, we need to analyze some simple predictable data. For these purposes we choose a set of time series, consisting of data modeled as combinations of a signal and white noise. As a signal, we choose a sine or a combination of sines. This gives a signal series of periodic (or quasi periodic) structure, which can be well represented by only few SSA eigentriples. In this case we know what to expect from the forecast and we know its structure, so an improvement of the forecast using some additional information will be easily recognizable.

We consider two cases of the main series (F):

$$f_n = \sin(\pi n\omega) + \sigma\varepsilon_n \qquad n = 1, ..., N, \tag{3.1}$$

$$f_n = \sin(\pi n\omega) + \sin(3\pi n\omega) + \sigma\varepsilon_n \qquad n = 1, ..., N, \tag{3.2}$$

and two cases of the support series (G):

$$g_n = \sin(\pi n\omega + \tau) + \sigma\zeta_n \qquad n = 1, ..., N, \tag{3.3}$$

$$g_n = \exp(\pi n\omega) \qquad n = 1, ..., N, \tag{3.4}$$

where $\epsilon_n, \zeta_n \sim N(0, \sigma^2)$ are i.i.d. noise. We are going to look at the different combinations of main and support series described above.

The explanation of the choice of the support series is following. The time series (3.3) has identical signal to the series (3.1) and corresponds to one part of the signal of the series (3.2). Hence, one would expect feedback relationship (3.1)$\leftrightsquigarrow$(3.3) and causal relationship respectively (3.3)$\rightsquigarrow$(3.2). However, the time series (3.4) has completely different structure with (3.1)and (3.2) and therefore, we are expecting no causality at all.

The next essential moment of the test setup is to construct it in such a manner, that it is possible to compare forecasts not only between each other, but also with some actual values of the series. Of course, in real life it is not possible to do so, but for generated series we are able to construct such a setup.

For simplicity we are going to use the notation $\hat{f}$, $\hat{g}$ for both the reconstructed series and the forecast of the series. This is reasonable as the reconstructed part of the series presents the base for the linear recurrence formula.

We are going to look at the forecast of time series with different types of signal as was described above (see (3.1) - (3.4)). Noise for the time series F and G was generated within MATLAB using the built-in pseudo-random *randn* command.

First we split both series into two parts. In the main series $F = (f_1, ..., f_{N+M})$ we distinguish two parts $F_R = (f_1, ..., f_N)$ and $F_F = (f_{N+1}, ..., f_{N+M})$; similarly, the support series $G = (g_1, ..., g_{N+M})$ is split into $G_R = (g_1, ..., g_N)$ and $G_F = (g_{N+1}, ..., g_{N+M})$, where M is the number of forecast points. F_F and G_F are those parts of two given time series F and G, which are used for comparing with a forecast (not forecast series themselves). The first N points of both series are used for the reconstruction procedures: SSA, MSSA$_+$, MSSA. In fact, SSA procedure uses just F_R (univariate case); MSSA$_+$ uses F_R, G_R and information known from G_F part; whereas MSSA uses an equal amount of information from both series, F_R and G_R only.

The SSA procedure starts from the series F_R and produces a reconstructed series $\widetilde{F}_R$. To forecast M points of the series, we use the linear recurrence formula obtained from the SSA forecast algorithm (1.41), calculating point $\hat{f}_{N+1}$ from the $(\hat{f}_{N-L+2}, ..., \hat{f}_N)$ part of the series $\widetilde{F}_R$. Generally, to forecast point $\hat{f}_{N+m}$ we use $(\hat{f}_{N-L+m+1}, ..., \hat{f}_{N+m-1})$ in formula (1.41), where $m \in \{1, ..., M\}$; here $\hat{f}_i$ is the part of the reconstructed series for $i \in \{1, ..., N\}$ and previously forecasted values for $i \in \{N+1, ..., N+M\}$.

For MSSA$_+$ we take series F_R and G_R, applying the MSSA algorithm to obtain two reconstructed series $\widetilde{F}_R = (\hat{f}_1, ..., \hat{f}_N)$ and $\widetilde{G}_R = (\hat{g}_1, ..., \hat{g}_N)$. Here $\hat{f}_{N+m}$ is obtained by the MSSA$_+$ linear recurrence formula (1.45). We take first $(M-1)$ points of the G_F part $(g_1, ..., g_{M-1})$ to get a valid comparison between three SSA-based procedures. For the forecast we use $\hat{f}_i$ from the reconstructed series for $i \in \{1, ..., N\}$ and previously forecasted

values for $i \in \{N+1, ..., N+M\}$, $\hat{g}_i$ from the reconstructed series for $i \in \{1, ..., N\}$ and actual g_i values for $i \in \{N+1, ..., N+M-1\}$, so the expression for the $\hat{f}_{N+m}$ is as described in (1.45).

For MSSA we don't use the additional information in G_F, i.e. by the obtaining reconstructed series $\widetilde{F}_R$, $\widetilde{G}_R$ by the MSSA algorithm, we use the MSSA forecast algorithm to forecast values $\hat{f}_i$ and $\hat{g}_i$ for $i \in \{N+1, ..., N+M-1\}$. The values $\hat{f}_{N+m}$ are calculated using the linear recurrence formula (1.56), where $m \in \{1, ..., M\}$.

To see wider picture of the forecast behavior one can study its changing dynamic in terms of different parameters. These parameters can be divided in following groups: time series parameters and procedure parameters. Time series parameters refer to properties of the input series themselves, and cannot be controlled in case of real data. Procedure parameters refer to choice made in the analysis and can always be controlled.

We study the following SSA/MSSA procedure parameters:

- window length L
- number of eigentriples I_r
- length of the forecast M

For our constructed test series, we also have control of the time series parameters:

- noise level σ^2
- time delay τ in the support series

The choice of procedure parameters is shown in the order of their appearance in the procedure. First, we choose the window length L, and run the first stage of the SSA algorithm. On the second stage (grouping) we are choosing eigentriples I_r, which are going to be used for the reconstruction. As we take them in order of decreasing singular values, the number of first eigentriples, related to the biggest eigenvalues, is the usual choice for the eigentriple parameter. After reconstruction, we can find the linear recurrence formula (LRF) for the forecast. At this stage we are choosing the length M of the forecast we want to obtain. We vary the sample size N_s to see whether and how quickly the applied tests converge.

By changing noise level σ^2 and adding some time delay τ to the support series G one can see the effect of changing these parameters.

The set of the scenarios was constructed so that it is possible to see the effect of all parameters in different combinations described above on the result of the forecast.

We define a scenario as a procedure, repeated N_s times, performing SSA, MSSA_+ and MSSA forecast algorithms with chosen parameters. For each individual trial of the scenario the noise of the main (and the support) series is regenerated. Thus, each scenario is based on the principle of bootstrapping: as the noise is regenerated N_s times, the observations we get for each forecast point can be considered independent.

For each scenario we obtain three sets of M distributions. Each set corresponds to the forecast algorithm that was used; each distribution of the set corresponds to an m^{th} forecast point distribution ($m \in \{1, ..., M\}$).

3.2 Simple causality criteria

The most straightforward comparison of forecasts, which can be used as a causality criterion, is to look at the ratios of loss functions.

Theory

Loss functions in this case are functions assessing the forecast error. If these errors have a random element arising from the noise in the main and support series, we run a number N_s of trials to estimate the error distribution or some of its characteristics, especially the variance.

We denote the forecast of j^{th} point by

$$\hat{f}^{(i)}_{N+j} \qquad (i \in \{1, ..., N_s\}, j \in \{1, ..., M\}),$$

the actual value, signal with noise, of the $(N+j)^{th}$ by

$$f^{(i)}_{N+j} \qquad (i \in \{1, ..., N_s\}, j \in \{1, ..., M\})$$

and the signal at the $(N+j)^{th}$ point by

$$f_{N+j,sig}, \qquad (i \in \{1, ..., N_s\}, j \in \{1, ..., M\}) | \qquad \forall \quad i,k \in N_s : f^{(i)}_{N+j,sig} = f^{(k)}_{N+j,sig}$$

Furthermore,

$$\mu_{N+j} = \frac{1}{N_s} \sum_{i=1}^{N_s} f^{(i)}_{N+j}$$

is the empirical mean of the $(N+j)^{th}$ point sample, which we take as an estimate for the mean of its distribution. We use the general notation $f^{(i)}_{N+j,comp}$ for the value used for comparison at the i^{th} trial, where *comp* designates its origin as signal value, actual value, empirical mean or median.

For each trial the noise ϵ is regenerated as pseudo-random values, thus the trials can be considered statistically independent. Therefore, values obtained for each forecast point are i.i.d. random values. We consider the following loss functions

$$S^{(1)}_{N+j} = \frac{\sum_{i=1}^{N_s} (\hat{f}^{(i)}_{N+j} - f^{(i)}_{N+j,sig})^2}{N_s} \ , \tag{3.5}$$

$$S^{(2)}_{N+j} = \frac{\sum_{i=1}^{N_s} (\hat{f}^{(i)}_{N+j} - f^{(i)}_{N+j})^2}{N_s} \ , \text{ and} \tag{3.6}$$

$$S^{(3)}_{N+j} = \frac{\sum_{i=1}^{N_s} (\hat{f}^{(i)}_{N+j} - \mu_{N+j})^2}{N_s} \tag{3.7}$$

as measures of the forecast quality.

$S^{(1)}$ is the measure for the *accuracy* of the forecast, where we test how close the forecast value is to the signal at that point. $S^{(2)}$ measures the accuracy as well, but with respect to the actual value (signal+noise) of the series at the point. $S^{(3)}$ is a measure of the *stability* of the forecast.

With the help of these functions, the forecast can be analyzed from several points of view: with respect to the actual (signal+noise) value of the time series, its signal and the empirical mean of the distribution of the forecast point. In practice, comparing the forecast with real data is not possible, because we don't know the actual future value of the analyzed time series. Comparing the forecast with a signal value can be a complicated task. In this case we know that the signal is just a sine wave, but for real data the extraction of a signal is a problem. There is the residual correlation between noise and a signal, and that is why it is difficult to separate them. The most convenient comparison is with the mean of the point distribution.

Clearly, a smaller value of $S^{(i)}_{N+j}$ indicates a more accurate or stable forecast, respectively.

As it was mentioned at the beginning of this section, the simple causality criteria is the ratio of the loss functions. If $S^{(k)}_{SSA} > S^{(k)}_{MSSA}$ $(k = \{1, 2, 3\})$ then MSSA is better, otherwise worse or the same:

$$\begin{cases} \frac{S^{(k)}_{MSSA}}{S^{(k)}_{SSA}} < 1 & \text{MSSA improves upon SSA} \\ \frac{S^{(k)}_{MSSA}}{S^{(k)}_{SSA}} \geq 1 & \text{MSSA does not improve upon SSA} \end{cases} \tag{3.8}$$

In fact, the ratio (3.8) with $k = 3$ corresponds to the F statistics of the *F-test*, which is described in Section 3.3.

Practice

We start of with validating if loss functions ratios could be considered good criteria of detecting the causality. And firstly we look at their performance in case of null hypothesis is true.

H_0 hypothesis testing. We take time series F_R (3.2) with the signal frequency $\omega = 0.046$ to be of the length $N = 200$ and run the SSA procedure twice to obtain $(N+1)^{th}$ forecast point N_s times for each different choice of time series and procedure parameters. Then we choose time series parameters: the noise level to be $\sigma^2 = \{0.01, 0.5, 1, 2\}$; due to the absence of the support series, there is no time delay τ. Moving on to procedure parameters, we choose: window length to be $L = \{100, 50, 25\}$; first four eigentriples for F_R reconstruction; forecast only one point $M = 1$; the number of trials $N_s = \{100, 200, ..., 500, 1000, 1500, ...3000\}$. Thus, for each sample size N_s we now have 12 scenarios (all possible combinations of the chosen window length L and the noise level σ^2).

The reconstruction of the time series is based on first four eigentriples, which correspond to the signal $\sin(\pi n\omega) + \sin(3\pi n\omega)$. We do the SSA analysis of series (3.2) using Caterpillar software. The weight of first eigentriple varies from 13% to 15.6%

Tab. 3.1: The loss function $S^{(1)}$ for the SSA procedure for sample sizes $N_s = 100(3000)$

σ^2 \ L	100	50	25
1	0.10(0.11)	0.12 (0.11)	0.16 (0.16)
2	0.23 (0.24)	0.19 (0.22)	0.40 (0.32)
0.5	0.07 (0.05)	0.06 (0.06)	0.06 (0.09)
0.01	0.002 (0.003)	0.011 (0.009)	0.03 (0.03)
0	0.002 (0.002)	0.009 (0.009)	0.03 (0.03)

(depending on the choice of window length) and contains mostly information about the first part of the signal $\sin(\pi n\omega)$; the weight of the second eigentriple is $10.5 - 11\%$ and contains information of both parts of the signal; the weight of the third eigentriple is $10.2 - 10.8\%$ and the fourth eigentriple is $9.4 - 10.3\%$ containing information about second part of the signal $\sin(3\pi n\omega)$. The weight of eigentriples vary depending on the window length: the smaller is the window length, the bigger weight leading eigentriples get.

Looking at the values of loss functions, we first study the relation between this loss functions and their dependence on the noise level σ^2 and procedure parameters. The three Tables of loss functions $S^{(k)}$ 3.1-3.3, contain values for the smallest sample size ($N_s = 100$) and contain the values in brackets corresponding to the largest sample size ($N_s = 3000$). As we see from the values in Tables 3.1-3.3, the loss function $S^{(k)}$ gives a consistent result under the sample size N_s variation. We also can notice that the values obtained for the loss functions $S^{(1)}$ and $S^{(3)}$ are approximately the same, i.e. the ratio of corresponding loss functions $\frac{S^{(1)}}{S^{(3)}} \approx 1$.

Now moving on to the loss functions ratios (see Tables 3.4-3.6), we see that on average mostly all values obtained are close to one, although there are several ratios, which are slightly higher. However, we expect to see 5 rejections on null hypothesis due to the 5% chance for error.

Under the null hypothesis testing it does not make sense to look at the ratios of $S^{(3)}$ values for noise level $\sigma = 0$, as the values themselves are practically zeros and obtained ratios are dominated by random fluctuation.

All loss functions ratios $\frac{S^{(i)}_{SSA1}}{S^{(i)}_{SSA2}}$ are ≈ 1 $(i = 1, 2, 3)$, where SSA1 is the first run of SSA, and SSA2 is the second one. Therefore both times we run SSA, we obtain very similar results as expected. Also, if the signal of the time series was extracted correctly, i.e. eigentriples were chosen correctly, then the mean of the fixed forecast point distribution gives a value close to the real signal value, i.e. $S^{(1)} \approx S^{(3)}$.

The loss function, where we compare the forecast value with the actual value seem to reflect the variance of the noise directly $S^{(2)} \approx S^{(1)} + \sigma^2$.

The following example is to show the goodness of loss function based measures, i.e. their ratios, when we compare SSA and MSSA procedures between each other.

Tab. 3.2: The loss function $S^{(3)}$ for the SSA procedure for sample sizes $N_s = 100(3000)$

σ^2 \ L	100	50	25
1	0.10 (0.10)	0.10 (0.09)	0.12 (0.13)
2	0.23 (0.24)	0.18 (0.22)	0.36 (0.30)
0.5	0.07 (0.05)	0.05 (0.05)	0.04 (0.07)
0.01	0.001 (0.001)	0.0011 (0.0009)	0.001 (0.0012)
0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

Tab. 3.3: The loss function $S^{(2)}$ for the SSA procedure for sample sizes $N_s = 100(3000)$

σ^2 \ L	100	50	25
1	1.09 (1.11)	1.12 (1.12)	1.04 (1.16)
2	2.48 (2.16)	1.74 (2.27)	1.88 (2.21)
0.5	0.47 (0.54)	0.66 (0.56)	0.56 (0.57)
0.01	0.012 (0.013)	0.018 (0.021)	0.04 (0.04)
0	0.002 (0.002)	0.009 (0.009)	0.03 (0.03)

Tab. 3.4: The loss functions $S^{(1)}$ ratio $\frac{S^{(1)}_{SSA1}}{S^{(1)}_{SSA2}}$ for sample sizes $N_s = 100(3000)$

σ^2 \ L	100	50	25
1	0.82 (1.05)	0.97 (0.96)	1.12 (0.96)
2	0.83 (0.93)	1.36 (0.96)	0.83 (0.97)
0.5	0.68 (0.99)	0.98 (0.97)	**1.40** (1.01)
0.01	1.36 (0.98)	0.93 (1.00)	1.05 (1.00)
0	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)

Tab. 3.5: The loss functions $S^{(2)}$ ratio $\frac{S^{(2)}_{SSA1}}{S^{(2)}_{SSA2}}$ for sample sizes $N_s = 100(3000)$

σ^2 \ L	100	50	25
1	0.999 (0.99)	1.04 (0.99)	1.11 (1.00)
2	0.87 (1.00)	1.03 (0.98)	1.04 (1.01)
0.5	0.96 (1.00)	0.91 (0.99)	1.03 (1.02)
0.01	1.01 (0.99)	0.93 (1.00)	1.03 (1.00)
0	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)

Tab. 3.6: The loss functions $S^{(3)}$ ratio $\frac{S^{(3)}_{SSA1}}{S^{(3)}_{SSA2}}$ for sample sizes $N_s = 100(3000)$

σ^2 \ L	100	50	25
1	0.87 (1.05)	0.90 (0.98)	1.26 (0.96)
2	0.82 (0.93)	1.38 (0.96)	0.80 (0.96)
0.5	0.70 (0.99)	1.01 (0.99)	**1.48** (0.99)
0.01	**1.50** (1.02)	0.94 (1.02)	1.30 (1.01)

Simple signal. We look at the case when the signal of the main f_n and the support g_n series are the same. However, the main series is perturbed by the white noise ε_n with given variance σ^2_ε. For this example we take the time series f_n (3.1) as main series and g_n (3.3) as a support time series with $\sigma^2_\zeta = 0$. Here there is no delay in time series $\tau = 0$ and the frequency $\omega = 0.046$. The length of both time series is $N = 200$. The variance $\sigma^2_\varepsilon = \sigma^2$, the window length L and the sample size N_s vary as in the example *H_0 hypothesis testing.*

The reconstruction of the time series is based on first two eigentriples, as they correspond to the signal $\sin(\pi n\omega)$. The analysis of the time series have shown that these are the components which contain the signal information.

The first eigentriple vary depending on the noise level, which is time series parameter, and the choice of the window length L. For example, for the series (3.1) with $\sigma^2 = 2$ for the SSA procedure first eigentriple weight varies from 12.5% to 17.5%. The smaller is the window length, the bigger weight first eigentriple has. The second eigentriple is of a slightly smaller weight, $10.38 - 14.42\%$ (same dependence on the window length). As for MSSA, we get following weights, the first eigentriple varies $18.7 - 21.1\%$ and the second one varies $14.7 - 17\%$. Values of the loss functions, see Tables 3.7-3.9, show a small improvement in the forecast obtained with MSSA for a higher noise variance σ^2. However for small noise level, the SSA and the MSSA forecasts seems to be equally good. The explanation for this lies in the simplicity of chosen time series, where the signal can be easily recognized without using this support series. Obviously, ratios of the loss functions are approximately one.

We also observe the same relation between loss functions, which was discussed earlier in *H_0 hypothesis testing.*

The next example is to show the goodness of loss functions for the time series, where the signal in main series is represented by combination of two simple sines and the support series correspond to the signal in main series just partially.

Combination of signals with different choice of eigentriples. We start by taking time series (3.2) to be the main series and (3.3) to be the support series with $\sigma^2_\zeta = 0$, both of length $N = 200$. There is no delay τ in the support series and the frequency ω is chosen to be the same as in previously described examples, i.e. $\omega = 0.046$. This

Tab. 3.7: The loss function $S^{(1)}$ for $N_s = 3000$ with the window length $L = 50$

σ^2	SSA	MSSA_+	MSSA
1	0.96	0.92	0.94
2	1.02	0.94	0.95
0.5	0.95	0.92	0.93
0.01	0.92	0.91	0.92
0	0.92	0.91	0.92

Tab. 3.8: The loss function $S^{(2)}$ for $N_s = 3000$ with the window length $L = 50$

σ^2	SSA	MSSA_+	MSSA
1	1.05	1.03	1.02
2	2.13	2.07	2.07
0.5	0.52	0.51	0.51
0.01	0.01	0.01	0.01
0	0.00	0.00	0.00

Tab. 3.9: The loss function $S^{(3)}$ for $N_s = 3000$ with the window length $L = 50$)

σ^2	SSA	MSSA_+	MSSA
1	0.05	0.03	0.03
2	0.11	0.05	0.05
0.5	0.03	0.01	0.01
0.01	0.00	0.00	0.00
0	0.00	0.00	0.00

example illustrates how the choice of eigentriples effects the measure based on the loss functions.

The signal of the main series f_n is a combination of two sines with different frequencies and could be described by the first four eigentriples.

We consider two cases, where forecast is based on either the first two or the first four eigentriples.

Naturally, two eigentriples do not give a proper reconstruction of the main series, as they describe just one part of the compound signal. But in this case we are able to study to what consequence the wrong choice of eigentriples leads to (for all three SSA-based procedures).

The first value we are looking at is empirical mean obtained for all scenarios for the 1^{st} forecast point ($N_s = 3000$). With generated data we have the advantage of knowing a real signal value at the 1^{st} forecast point. If the forecast is close to the real value, the empirical mean of forecast sample should be close to it as well. For this example with two eigentriples, empirical mean varies between -0.837 and 0.229 (see Table 3.14), while the signal value is -1.3525.

Thus, in this particular case the empirical mean for all SSA-based procedures, with any noise level and any window length is far off the signal value, i.e. even if the forecasts are stable, the mean was estimated incorrectly and obtained forecasts are not accurate. This shows the importance of choosing right number of eigentriples.

As the mean of the forecast is misleading, we are not looking at the values of the loss function $S^{(3)}$. Nevertheless, the relations between $S^{(1)}$ and $S^{(2)}$ described in previous examples hold. Tables 3.15 - 3.17 show that the results depend on the window length and the noise level. The smaller is the window length, the smaller is the loss function $S^{(1)}$ for SSA, and for both MSSA it varies more for $L = 100$, but gives more stable result for $L = 50$, and very stable result for $L = 25$. If we compare SSA and MSSA, it seems that for the larger window length MSSA is better than SSA, but as the window length decreases, $S^{(1)}$ gives a smaller error value for SSA, then for any of MSSA.

However, if we look at $S^{(3)}$ values in Table 3.18 it shows completely different tendency (for any noise level $S^{(3)}_{MSSA,MSSA_+} < S^{(3)}_{SSA}$), which leads us to a conclusion, that the ratio test for the loss function $S^{(3)}$ might be overoptimistic in case, when there are not enough eigentriples taken.

Wrong signal reconstructions leads to wrong mean estimation, which affects the relation between $S^{(1)}$ and $S^{(3)}$: $S^{(1)} \not\approx S^{(3)}$ (see Tables 3.16 and (3.18)).

With four eigentriples we get a reconstruction, which is close to the real signal of the main series. The empirical mean obtained for the scenarios, where we used four eigentriples, varies from -1.329 to -1.183, while the real value of the signal is -1.3525. Therefore, obtained distributions are closer to the real value of the signal, then distributions obtained by scenarios with two eigentriples (data is not shown).

The values in Tables 3.10, 3.11 indicate the similarity of MSSA and MSSA$_+$. For good signal extraction there is no significant improvement in forecast, when we use the future

Tab. 3.10: The loss function $S^{(1)}_{N+1}$ for sample size $N_s = 3000$

	SSA			MSSA		
σ^2 \ L	100	50	25	100	50	25
1	0.106	0.106	0.112	0.087	0.081	0.114
2	0.242	0.223	0.228	0.196	0.171	0.233
0.5	0.053	0.055	0.061	0.0439	0.039	0.061
0.01	0.003	0.010	0.019	0.002	0.002	0.017
0	0.002	0.009	0.018	0.001	0.001	0.017

Tab. 3.11: The loss function $S^{(1)}_{N+1}$ for sample size $N_s = 3000$

	MSSA_+		
σ^2 \ L	100	50	25
1	0.086	0.078	0.112
2	0.193	0.163	0.228
0.5	0.043	0.039	0.061
0.01	0.002	0.006	0.019
0	0.001	0.005	0.018

Tab. 3.12: $S^{(1)}_{N+1}$ ($N_s = 3000$)

	$S^{(1)}_{MSSA}/S^{(1)}_{SSA}$			$S^{(2)}_{MSSA}/S^{(2)}_{SSA}$		
σ^2 \ L	100	50	25	100	50	25
1	0.799	0.737	0.672	0.980	0.973	0.951
2	0.778	0.724	0.677	0.972	0.965	0.945
0.5	0.806	0.716	0.682	0.979	0.975	0.947
0.01	0.638	0.304	0.668	0.925	0.657	0.741
0	0.510	0.247	0.666	0.510	0.247	0.666

Tab. 3.13: The loss function $S^{(3)}_{N+1}$ for sample size $N_s = 3000$

	SSA			MSSA		
σ^2 \ L	100	50	25	100	50	25
1	0.104	0.098	0.133	0.087	0.080	0.098
2	0.241	0.217	0.301	0.196	0.170	0.219
0.5	0.052	0.047	0.065	0.043	0.038	0.046
0.01	0.001	0.0009	0.0012	0.0008	0.0007	0.0009
0	0.00	0.00	0.00	0.00	0.00	0.00

Tab. 3.14: The empirical mean of 201^{st} forecast point when two components are chosen with the signal value being -1.3525

	SSA			MSSA		
σ^2 \ L	100	50	25	100	50	25
1	-0.407	-0.691	-0.431	-0.362	-0.344	-0.416
2	-0.516	-0.676	-0.463	-0.342	-0.362	-0.448
0.5	-0.266	-0.686	-0.423	-0.376	-0.344	-0.408
0.01	0.228	-0.753	-0.410	-0.377	-0.341	-0.395
0	0.229	-0.747	-0.837	-0.377	-0.340	-0.394

information from the support series g, i.e. g_F (see 3.1). Thus, for the modeled time series, we are going to skip the MSSA$_+$ forecast algorithm.

Comparing values in Tables 3.10 and 3.11 with the ones in Tables 3.15 - 3.17, we see the obvious improvement for all implemented scenarios with four eigentriples.

Obtained values in Tables 3.10, 3.13 support the statement that $S^{(1)} \approx S^{(3)}$, which is equivalent to $\mu_{N+j} \approx f_{N+j,sig}$ (see notation in Section 3.2).

Looking at the ratios of the loss function $S^{(2)}$ (3.12), and recalling the fact that $S^{(2)} \approx S^{(2)} + \sigma^2$, then $S^{(2)}_{MSSA}/S^{(2)}_{SSA}$ can be rewritten as

$$\frac{S^{(2)}_{MSSA}}{S^{(2)}_{SSA}} \approx \frac{S^{(1)}_{MSSA} + \sigma^2_{MSSA}}{S^{(1)}_{SSA} + \sigma^2_{SSA}}. \tag{3.9}$$

If $S^{(1)}_{(M)SSA} << \sigma^2_{(M)SSA}$, then the noise is overtaking in the ratio (3.9). Therefore, even if causality is obvious, it can be not recognized for a high noise level. Values in Table (3.12) support this statement. For $\sigma^2 \geq 0.5$ the ratios $\frac{S^{(2)}_{MSSA}}{S^{(2)}_{SSA}} \approx 1$.

If $\sigma^2 = 0$, then $\frac{S^{(1)}_{MSSA}}{S^{(1)}_{SSA}} = \frac{S^{(2)}_{MSSA}}{S^{(2)}_{SSA}}$.

Tab. 3.15: The loss function $S^{(1)}$ for sample size $N_s = 3000$ with window length $L = 100$

σ^2	SSA	MSSA$_+$	MSSA
1	1.17	1.02	1.02
2	1.02	1.12	1.12
0.5	1.44	0.97	0.97
0.01	2.50	0.95	0.95
0	2.50	0.95	0.95

Tab. 3.16: The loss function $S^{(1)}$ for sample size $N_s = 3000$ with window length $L = 50$

σ^2	SSA	MSSA$_+$	MSSA
1	0.56	1.03	1.05
2	0.66	1.04	1.05
0.5	0.52	1.02	1.03
0.01	0.36	1.01	1.02
0	0.37	1.01	1.03

Tab. 3.17: The loss function $S^{(1)}$ for sample size $N_s = 3000$ with window length $L = 25$

σ^2	SSA	MSSA$_+$	MSSA
1	0.53	0.89	0.92
2	0.63	0.89	0.92
0.5	0.47	0.88	0.91
0.01	0.28	0.89	0.92
0	0.27	0.89	0.92

Tab. 3.18: The loss function $S^{(3)}$ for sample size $N_s = 3000$ with window length $L = 50$

σ^2	SSA	MSSA$_+$	MSSA
1	0.12	0.03	0.03
2	0.20	0.06	0.07
0.5	0.07	0.01	0.01
0.01	0.0006	0.0002	0.0002
0	0.00	0.00	0.00

Tab. 3.19: The loss function $S^{(1)}$ for sample size $N_s = 1000$ with noise level $\sigma^2 = 1$

σ^2	2			0.5		
τ \ L	100	50	25	100	50	25
SSA						
0	0.22	0.23	0.30	0.05	0.05	0.09
0.1π	0.23	0.23	0.30	0.05	0.06	0.11
0.2π	0.24	0.22	0.33	0.05	0.06	0.12
0.3π	0.21	0.23	0.32	0.05	0.06	0.11
0.4π	0.24	0.22	0.32	0.06	0.06	0.10
0.5π	0.22	0.22	0.31	0.05	0.05	0.08
MSSA						
0	0.18	0.18	0.22	0.04	0.04	0.06
0.1π	0.18	0.17	0.21	0.04	0.04	0.07
0.2π	0.19	0.17	0.23	0.04	0.04	0.08
0.3π	0.17	0.17	0.21	0.04	0.04	0.07
0.4π	0.19	0.16	0.22	0.04	0.04	0.06
0.5π	0.18	0.17	0.21	0.04	0.04	0.05

The last example was designed to study sensitivity of the MSSA forecast algorithm to the delay of the signal in the support series.

Shifted signal. Suppose we have two time series, main time series (3.2) and the support one (3.3) with the delay $\tau \in \{0.1\pi, 0.2\pi, 0.3\pi, 0.4\pi, 0.5\pi\}$. The rest of parameters are chosen similarly to previous examples, $N = 200$, $\sigma^2 = 1$, $L \in \{25, 50, 100\}$ and $N_s = 1000$.

We already defined the relation between three introduced loss functions in the previous examples, which holds for this one as well. To see how the delay reflects on the loss functions it is enough to check values just for one of them, for example, $S^{(1)}$ (see Table 3.19). The values in this Table suggest that the shift does not effect the causality between time series (3.2) and (3.3).

3.3 Normality tests and F-test

Theory

The F-test is used to test the hypothesis that two samples arise from the same normal distribution. If the data are normally distributed, they can be reasonably described by their mean and variance only. In terms of the question of causality, smaller dispersion means a more stable forecast.

The F-test checks whether the variances of two normally distributed samples are significantly different. As we are only interested in the variances here, the mean can be eliminated from both distributions. Suppose we have two samples of observations $\{x_i\}$, $\{y_i\}$, then we can introduce values $\{\alpha_i\}$ and $\{\beta_i\}$ $(i \in \{1, ..., n\})$:

$$\alpha_i = x_i - \mu(x) \tag{3.10}$$

$$\beta_i = y_i - \mu(y) \tag{3.11}$$

where $\mu(x)$, $\mu(y)$ are the empirical means of the corresponding samples. The F-test statistics F is

$$F = \frac{\sigma_{\alpha_i}}{\sigma_{\beta_i}},$$

which corresponds to the ratio of the loss functions $S^{(3)}$ when they are calculated from the samples coming from normal distributions.

Under the null hypothesis we assume that the variances of two modified samples $\{\alpha_i\}$, $\{\beta_i\}$ are the same, i.e. both come from the same normal distribution. If they are significantly different, we reject the null hypothesis, and say that one method is outperforming the another one. In other words, smaller variance (smaller standard error) gives more stable forecast. The F-test also can be used as a tool for checking the sensitivity of the forecast, obtained by the (M)SSA algorithms, to changes in the procedure and the time series parameters.

Several observations arise from F-test simulations while varying trials parameters. The F-test p-value depend strongly on the sample size N_s of forecast distribution; the F statistics, on the other hand, converge to some stable ratio of variances for large N_s. Moreover, if there is no normality, the ratio of variances is still a reasonable measure to check forecast stability.

As was mentioned previously, the F-test can be applied only to normally distributed data. To test that data comes from normal distribution, we use following tests for normality *Kolmogorov-Smirnov* (KS) or *Anderson-Darling* (AD) test.

Kolmogorov-Smirnov Test. Before describing KS test, it is worth mentioning the fact that this test can be used for two different purposes: checking if two samples have the same unspecified distribution or checking if a sample comes from a normal distribution. In this section we are interested in normality.

For KS test, assuming that values of a sample come from a normal distribution, one takes the maximum difference between empirical cumulative distribution functions and the cdf of the standard normal distribution $N(\mu, \sigma^2)$.

$$D_n = \sup_t \mid F_x(t) - F(t) \mid, \tag{3.12}$$

where F_x are empirical cdf of some set of observations $\{x_i\}$ of size n and F is a cumulative distribution function (cdf) of a theoretical (in this case, normal) distribution.
In fact, under the null hypothesis, it is true that $D_n = O(\frac{1}{\sqrt{n}}), n \to \infty$ [3].

Denote the significance level as α and let $D_{stat} = \sqrt{n} D_n$. Then the critical value D_α is chosen so that the type I error probability is $P(D_{stat} > D_\alpha) = \alpha$.

For $n \to \infty$ the asymptotic distribution is given as

$$P(D_{stat} > D_\alpha) = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2D_\alpha^2}, \tag{3.13}$$

see [3].

For this test we are satisfied with significance level being $\alpha = 0.05$ (type I error: if the null hypothesis is true and we still reject it).

For KS normality test we compare given empirical cdf F_x of some i.i.d. observations with a standard normal distribution. For a fair comparison (as we are interested if the values come from any normal distribution) we normalize data $x_1, ..., x_n$ by introducing variables ξ_i such that

$$\xi_i = \frac{x_i - \mu(x)}{\sigma(x)}, \tag{3.14}$$

where $i \in \{1, ..., N_s\}$, $\mu(x)$ is the empirical mean of $\{x_i\}_1^n$ and $\sigma(x)$ is its dispersion.

Under the null hypothesis we make the assumption that $\xi_i \sim N(0, 1)$, and we reject the null hypothesis if $p < 0.05$.

Anderson-Darling test. AD test was derived from the *Cramer von Mises test.* Both tests, like KS test, are based on the distance comparison between empirical cumulative distribution functions. But while KS test is looking at the maximum difference between empirical cumulative distribution functions, AD and von Mises tests are looking at averaged weighted difference, but using slightly different weight functions. In general, both tests are described with the same expression

$$n\int_{-\infty}^{\infty}(F(t) - F_n(t))^2\phi(t)d(F(t)) \tag{3.15}$$

Here F is the distribution used for comparison, and F_n is the empirical cdf of known observations. The weight function for the Cramer von Mises test is $\phi(t) \equiv 1$, so that its statistic W^2 is a measure of the mean square difference between empirical and normal (or another theoretical) distributions; and for the AD test the weight function is $\phi(x) = [F(t)(1 - F(t))]^{-1}$, which gives more weight to differences in tails of the distribution, than to the differences in its middle part [23].

The AD test uses the statistic A such that

$$A^2 = -n - S, \text{ where } S = \sum_{k=1}^{n}\frac{2k-1}{n}(\ln F(x_k) + \ln(1 - F(x_{n+1-k}))) \tag{3.16}$$

and A statistic is compared with the critical value of asymptotic distribution

$$P(A^2 < A_\alpha^2) = \frac{\sqrt{2\pi}}{A_\alpha^2}\sum_{j=0}^{\infty}\binom{-\frac{1}{2}}{j}(4j+1)e^{-((4j+1)^2\pi^2)/(8A_\alpha^2)}\int_0^{\infty}e^{x/(8(\omega^2+1))-(4j+1)^2\pi^2\omega^2/(8A_\alpha^2)}d\omega, \tag{3.17}$$

where α is the significance level and A^2_α is the corresponding critical value. This asymptotic distribution is suitable for the AD weighted function [2, 3].

If we have a sample of observations x_i and we want to find out if this sample came from a normal distribution by using AD test, we can compare it with the standard normal distribution. Similarly to KS test, for a fair comparison x_i are needed to be normalized first as in 3.14. The p value is calculates as follows $p = 1 - P(A^2 < A^2_\alpha)$. If $p < 0.05$, we reject the null hypothesis.

Remark KS and AD tests described in this section can be used not only for investigating if some data come from a particular theoretical distribution, but also for the direct comparison of two samples. Under the null hypothesis we assume that two empirical cumulative distribution functions are the same, i.e. $F_x = F_y$. Thus, the alternative hypothesis for the two-sided test is $H_1 : F_x \neq F_y$. The only difference to normality tests is that we are comparing two samples with each other, not a sample with a distribution.

Here we do not make any assumptions about how the data are distributed, which gives a big advantage, when one is dealing with real data.

In terms of SSA/MSSA approaches, if the H_0 hypothesis holds, then the distributions of the forecast obtained by SSA and MSSA are the same and there is no benefit in using the support series.

Practice

Using the setup of the H_0 *hypothesis testing* from the Section 3.2 we examine the goodness of the F-test performance under the null hypothesis.

H_0 *hypothesis testing.* Before applying F-test we need to check if the obtained forecast values are normally distributed. For this test we have chosen the noise level, or variance, to be $\sigma^2 = 1$ and see if the result stays consistent under two changes: the change of the window length L and the change of the sample size N_s. Tables 3.20 and 3.21 contain p-values of two normality tests for two samples obtained by 3 scenarios ran with different window length L.

After satisfying the condition of normality, we can apply the F-test. Some of the results are presented in Table 3.22. The p-values for the scenarios with window length $L = 100$, $N_s = 200$ trials and with $L = 25$, $N_s = 300$ are relatively small ($p \leq 0.05$), hence, we may come to the conclusion that samples in this case come from different distributions, but for this test we expect an error in approximately 5% scenarios. Other p-values in this table do not point to any significant difference between the two samples, from what we are concluding that samples come from the same normal distribution.

As it was mentioned in the remark in section 3.3 KS test can be also used for direct distribution comparison. In fact, it seems that KS test gives more consistent result in comparison with F-test ($\min p = 0.21$, see Table 3.23). The KS test supports the null hypothesis that values obtained by two run of SSA come from the same distribution. Note, that here we do not check for normality. The only information that we get out of performing KS test for direct distribution comparison is that data come from the same distribution.

Tab. 3.20: Kolmogorov-Smirnov Normality Test p-values ($\sigma^2 = 1$)

N_s \ L	100	50	25
100	0.96 (0.91)	0.96 (0.95)	0.95 (0.98)
200	0.74 (0.70)	0.58 (0.68)	0.80 (0.96)
300	0.92 (0.97)	0.90 (0.82)	0.97 (0.99)
400	0.97 (0.91)	0.22 (0.93)	0.998 (0.76)
500	0.76 (0.64)	**0.06** (0.77)	0.95 (0.98)
1000	0.96 (0.94)	0.99 (0.91)	0.90 (0.87)
1500	0.63 (0.70)	0.70 (0.29)	0.47 (0.91)
2000	0.90 (0.18)	0.92 (0.39)	0.98 (0.97)
2500	0.28 (0.55)	0.997 (0.75)	0.74 (0.80)
3000	0.25 (0.75)	0.57 (0.32)	0.57 (0.96)

Tab. 3.21: Anderson-Darling Normality Test p-values ($\sigma^2 = 1$)

N_s \ L	100	50	25
100	0.94 (0.41)	0.67 (0.89)	0.81 (0.61)
200	0.35 (0.46)	0.04 (0.13)	0.69 (0.81)
300	0.83 (0.57)	0.23 (0.28)	0.80 (0.88)
400	0.89 (0.29)	0.05 (0.85)	0.88 (0.10)
500	0.27 (0.44)	**0.001** (0.06)	0.81 (0.85)
1000	0.80 (0.59)	0.72 (0.54)	0.39 (0.50)
1500	0.12 (0.15)	0.17 (0.09)	**0.009** (0.81)
2000	0.90 (**0.01**)	0.56 (**0.04**)	0.85 (0.89)
2500	**0.005** (0.07)	0.88 (0.38)	0.46 (0.19)
3000	**0.0007** (0.12)	**0.03** (**0.003**)	**0.04** (0.61)

Tab. 3.22: F-Test p-values with noise level $\sigma^2 = 1$

N_s \ L	100	50	25
100	0.70	0.40	0.55
200	0.05	0.54	0.26
300	0.73	0.94	0.002
400	0.82	0.45	0.59
500	0.48	0.10	0.80
1000	0.93	0.94	0.76
1500	0.96	0.99	0.36
2000	0.67	0.12	0.70
2500	0.50	0.40	0.71
3000	0.67	0.22	0.67

Tab. 3.23: Kolmogorov-Smirnov Test p-values with noise level $\sigma^2 = 1$

σ^2 \ L	100	50	25
100	0.97	0.97	0.99
200	0.33	0.47	0.92
300	0.89	0.79	0.21
400	0.98	0.96	0.75
500	0.82	0.29	0.96
1000	0.85	0.76	0.83
1500	0.996	0.37	0.65
2000	0.61	0.51	0.92
2500	0.72	0.97	0.998
3000	0.93	0.39	0.84

The other way of assessing the performance of statistical tests, described in this section, F-test (AD and KS tests) is by trying them out in case when null hypothesis is likely to be rejected, therefore causality is not expected. For this purpose we take two time series with different structures.

Rejecting H_0 hypothesis example. Suppose we have main time series (3.1) and the support time series (3.4), both of length $N = 200$. The analysis of the series using Caterpillar software tools have shown that it is convenient to choose the following (M)SSA procedure parameters: the window length $L \in \{20, 100\}$ and first three eigentriples r, i.e. the set of eigentriples $I_r = \{1, 2, 3\}$. For this example we have chosen to run the procedure $N_s \in \{1000, 3000\}$ times.

For small window length $L \leq 10$ the separation of two main trends of two chosen time series seem to be more difficult, i.e. we cannot obtain approximately clean signal reconstruction.

First eigentriple $(53.4 - 75.5\%)$ corresponds to the exponent reconstruction, the second $(2.5 - 4.5\%)$ and the third $(2.3 - 3.9\%)$ to the sine wave reconstruction. The weight of the first eigentriple decrease as the window length increases.

Here statistical tests do not give any structural sets of results. Nevertheless, the normality is supported by KS test in most of the scenarios. Assuming that the data come from a normal distribution, according to the KS test, we can apply the F-test. The F-test rejects the similarity of the samples for most of the applied scenarios (all p-values, apart from one, are practically zero). But if we look at the F statistic, it is clear that $var_{SSA} < var_{MSSA} \Rightarrow$ SSA obtained forecast is more concentrated, therefore, more stable (see Table 3.24).

The next example is to see the performance of the tests in case of simple and similar structured series. For this example we are using the setup of *Simple signal* example, described in Section 3.2.

Simple signal. First we perform the KS and AD normality tests to see if obtained forecasts data is normally distributed. Varying the noise level in the mains series σ^2 and the window length L we look at the p-values of both normality tests for the large sample size $N_s = 3000$, see Tables 3.25 - 3.27. The first thing that catches the eye is the difference between KS and AD tests performances. KS test gives the consistent result for all three forecasts, obtained by SSA-based algorithms. However, AD normality test seems to reject null hypothesis in roughly half of the cases and seem to be not consistent for this example.

Based on the KS normality test, we assume that all three forecasts obtained come from normal distribution. Now, as the condition of normality is satisfied we may use F-test. Looking at the F-test statistic, Table 3.28, we see that the samples obtained are rather different. The F-test p-values obtained are all zeros and if we look at Table 3.28, we see that MSSA clearly gives more stable result, with narrower variance. In fact, the additional information which is used when we run the MSSA$_+$ algorithm does not make much difference for the F statistic.

Tab. 3.24: Normality and F tests p-values

L	100	20	100	20	100	20	100	20
σ^2	1		4		0.7		0.1	
Anderson Darling Normality test:								
SSA								
1000	0.12	**0.02**	**0.00**	0.07	0.06	0.52	0.82	0.63
3000	0.71	0.095	0.08	**0.00**	**0.03**	**0.02**	0.67	**0.00**
MSSA								
1000	**0.03**	0.45	**0.00**	**0.01**	0.002	**0.00**	0.18	0.10
3000	0.24	0.24	0.67	0.51	0.09	**0.03**	0.64	0.98
Kolmogorov-Smirnov Normality test:								
SSA								
1000	0.47	0.10	0.08	0.60	0.71	0.83	0.98	0.86
3000	0.98	0.71	0.72	0.43	0.66	0.49	0.91	0.35
MSSA								
1000	0.48	0.87	0.14	0.45	0.26	**0.001**	0.61	0.34
3000	0.61	0.68	0.92	0.80	0.61	0.22	0.66	0.94
F-test:								
p-value								
1000	0.00	0.00	0.46	0.00	0.00	0.00	0.00	0.00
3000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
var_{MSSA}/var_{SSA}								
1000	1.02	1.13	1.01	1.03	1.31	1.55	6.15	5.72
3000	3.55	3.78	3.06	3.47	3.44	3.77	3.76	3.87

Tab. 3.25: KS (AD) Normality Test p-values, SSA for sample size $N_s = 3000$

σ^2 \ L	100	50	25
1	0.14 (0.00)	0.29 (0.001)	0.71 (0.40)
2	0.47 (0.13)	0.63 (0.06)	0.13 (0.003)
0.5	0.35 (0.104)	0.75 (0.34)	0.88 (0.42)
0.01	0.73 (0.20)	0.92 (0.47)	0.83 (0.52)

Tab. 3.26: KS (AD) Normality Test p-values, MSSA_+ for sample size $N_s = 3000$

σ^2 \ L	100	50	25
1	0.11 (0.00)	0.17 (0.001)	0.43 (0.01)
2	0.59 (0.12)	0.99 (0.75)	0.25 (0.001)
0.5	0.63 (0.08)	0.17 (0.01)	0.27 (0.04)
0.01	0.74 (0.22)	0.68 (0.49)	0.86 (0.39)

Tab. 3.27: KS (AD) Normality Test p-values, MSSA for sample size $N_s = 3000$

σ^2 \ L	100	50	25
1	0.10 (0.00)	0.16 (0.001)	0.42 (0.009)
2	0.59 (0.12)	0.99 (0.72)	0.24 (0.001)
0.5	0.60 (0.08)	0.18 (0.01)	0.27 (0.04)
0.01	0.76 (0.22)	0.68 (0.49)	0.86 (0.39)

Tab. 3.28: F-Test statistics $\frac{var_{MSSA_+}}{var_{SSA}}$ ($\frac{var_{MSSA}}{var_{SSA}}$) for sample size$N_s = 3000$

σ^2 \ L	100	50	25
1	0.63 (0.63)	0.48 (0.48)	0.40 (0.40)
2	0.59 (0.59)	0.48 (0.48)	0.42 (0.42)
0.5	0.63 (0.64)	0.49 (0.49)	0.40 (0.40)
0.01	0.64 (0.64)	0.48 (0.48)	0.38 (0.38)

Tab. 3.29: p-values for F test for different sample sizes with noise level $\sigma^2 = 1$

	MSSA$_+$ vs SSA			MSSA vs SSA		
N_s \ L	100	50	25	100	50	25
100	0.35	0.44	0.11	0.40	0.65	0.15
200	0.11	0.06	0.01	0.14	0.20	0.02
300	0.05	0.01	0.00	0.08	0.05	0.00
400	0.06	0.01	0.00	0.10	0.07	0.00
500	0.02	0.00	0.00	0.05	0.01	0.00

Moving on two less simple signal structures, the next example will illustrate how the tests from this section perform in case when the main series is slightly more complicated and the support series contain just partial information about main series.

Using the setup of *Combination of signals with different choice of eigentriples* from Section 3.2. AD test supports the normality in nearly $80 - 87\%$ of scenarios. The smallest p-value for KS normality test is $p = 0.05$ (MSSA$_+$: $N_s = 3000$, $\sigma^2 = 1$, $L = 50$), therefore all the distribution obtained can be considered as normal, according to KS test. Therefore, according to KS test, the condition for using the F-test is satisfied.

Values in Table 3.29 show that the F test p-values rapidly converge to zero, as $N_s \longrightarrow \infty$. In fact, for $N_s \geq 1000$ p-values are practically zeros. On the other hand, the F-statistics values are very stable for both MSSA algorithms and do converge to a number greater than zero see Table 3.30.

Direct comparison of the given samples (SSA vs MSSA$_+$ and SSA vs MSSA) using KS test produces p-values, which converge to zero as $N_s \longrightarrow \infty$ (see Table 3.31).

There is no contradiction between F-test and direct comparison KS test; both show significant difference between distributions, obtained by SSA and MSSA (or MSSA$_+$).

Comparing obtained values in Tables 3.29-3.31, we observe similar results for MSSA$_+$ and MSSA, therefore, we can consider these two methods being approximately equivalent in this particular set of scenarios.

In the last example we are looking at the modeled time series to see if MSSA is sensitive to the delay of the signal in the support series. We use the setup of *Shifted signal* example described in Section 3.2.

Shifted signal. Both AD and KS tests support the null hypothesis for SSA and MSSA obtained forecast distributions similarly to the example of the same system of time series, but with $\tau = 0$ (for example, see p-values for scenarios with noise level $\sigma^2 = 1$, Table 3.32). AD test, similarly to the previous examples, is more sensitive test and it does not support the normality for all scenarios.

Tab. 3.30: F-test F statistic for different sample sizes with noise level $\sigma^2 = 1$

	var_{MSSA_+}/var_{SSA}			var_{MSSA}/var_{SSA}		
N_s \ L	100	50	25	100	50	25
100	0.83	0.85	0.72	0.84	0.91	0.75
200	0.79	0.77	0.69	0.81	0.83	0.71
300	0.80	0.74	0.68	0.82	0.79	0.70
400	0.83	0.77	0.69	0.85	0.84	0.71
500	0.82	0.74	0.69	0.84	0.79	0.71
1000	0.81	0.72	0.70	0.83	0.78	0.72
1500	0.81	0.74	0.69	0.83	0.80	0.72
2000	0.80	0.74	0.70	0.82	0.80	0.72
2500	0.81	0.76	0.70	0.83	0.82	0.72
3000	0.82	0.75	0.71	0.83	0.81	0.73

Tab. 3.31: p-values for KS test for different sample sizes. $\sigma^2 = 1$, $L = 25$

	MSSA_+ vs SSA			MSSA vs SSA		
N_s \ L	100	50	25	100	50	25
100	0.97	0.99	0.47	0.97	0.99	0.47
200	0.96	0.47	0.33	0.92	0.54	0.33
300	0.45	0.65	0.25	0.45	0.72	0.34
400	0.86	0.58	0.52	0.86	0.70	0.70
500	0.72	0.46	0.26	0.82	0.77	0.26
1000	0.47	0.03	0.05	0.61	0.12	0.12
1500	0.31	0.09	0.05	0.45	0.33	0.07
2000	0.18	0.04	0.01	0.24	0.09	0.01
2500	0.11	0.06	0.00	0.22	0.14	0.00
3000	0.09	0.01	0.00	0.13	0.06	0.01

Tab. 3.32: Normality tests p-values for the 201 point with $\sigma^2 = 1$

Method	SSA			MSSA		
τ \ L	100	50	25	100	50	25
KS						
0	0.87	0.56	0.82	0.50	0.73	0.95
0.1π	0.17	0.92	0.42	0.26	0.78	0.67
0.2π	0.49	0.16	1.00	0.52	0.21	0.78
0.3π	0.20	0.42	0.90	0.63	0.67	0.99
0.4π	0.79	0.73	0.26	0.77	0.83	0.25
0.5π	0.95	0.21	0.84	0.77	0.29	0.97
AD						
0	0.37	0.12	0.51	0.29	0.36	0.64
0.1π	0.00	0.66	0.02	0.00	0.35	0.09
0.2π	0.12	0.00	0.98	0.25	0.00	0.31
0.3π	0.03	0.002	0.63	0.14	0.02	0.75
0.4π	0.39	0.50	0.02	0.45	0.35	0.01
0.5π	0.75	0.01	0.73	0.60	0.02	0.65

The F-test p-values point at the significant difference between the obtained distribution, despite their normality (according to KS test). As was mentioned before, F-test p-value is an easily manipulated parameter, i.e. the bigger are the samples, the bigger is their difference. On the other hand, due to its stability, F-statistics may be considered as a possible measurement of causality. Table 3.33 suggests that the causality does not change and does not become weaker with a shift in the support series.

3.4 Sign tests

Theory

For each implementation of each scenario, described in 3.1, we obtain several values that are lately used for the comparison: SSA and MSSA forecast values, signal and actual noisy value of the forecast point.

Before describing sign test we need to mention another group of loss functions (3.18), which are different from loss functions (3.5). The sign test is based on the comparison of the magnitude of the two loss functions $e^{(i)}_{N+j,SSA}$, $e^{(i)}_{N+j,MSSA}$ for each trial separately. Note that here we are not comparing empirical cumulative distribution functions, we are comparing them pairwise for each individual trial

$$e^{i}_{N+j,(M)SSA} = \mid \hat{f}^{(i,(M)SSA)}_{N+j} - f_{N+j,comp} \mid, \tag{3.18}$$

Tab. 3.33: F-tests p-values and statistics for the $(N+1)$ point. $\sigma^2 = 1$

τ \ L	100	50	25
p			
0	0.00	0.00	0.00
0.1π	0.002	0.001	0.00
0.2π	0.001	0.00	0.00
0.3π	0.00	0.005	0.00
0.4π	0.001	0.00	0.00
0.5π	0.001	0.00	0.00
F			
0	0.83	0.78	0.72
0.1π	0.83	0.82	0.72
0.2π	0.82	0.79	0.73
0.3π	0.80	0.84	0.72
0.4π	0.82	0.79	0.74
0.5π	0.81	0.78	0.73

where $f_{N+j,comp} = f_{N+j,sig}$. By looking at the differences

$$d_i^j = e^{(i)}_{N+j,SSA} - e^{(i)}_{N+j,MSSA}, \tag{3.19}$$

one can estimate for which method the loss function (3.18) is smaller or bigger. Introducing the indicator function

$$I_+(d_j^i) = \begin{cases} 1 & \text{if } d_j^i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.20}$$

we can count how many times MSSA improves upon SSA. The frequency of MSSA being better than SSA is calculated as follows:

$$\omega_{N_s} = \frac{\sum_{i=1}^{N_s} I_+(d_j^i)}{N_s} \tag{3.21}$$

If the frequency is above 0.5 then we infer causality.

The sign test is a simple, nice, robust test, but does not take into account that some differences d_j^i are very small compared to others. If we want to distinguish d_j^i differences, it is reasonable to use *Wilcoxon's Signed-Rank Test.*

Wilcoxon's Signed-Rank test is similar to the sign test, but here we consider not only the sign of errors differences, but the magnitude of those differences as well. All the differences are ranked, i.e. ordered by increasing absolute value, and the indicator function is multiplied with a weight given by the rank number. The smallest difference receives the smallest weight. Both positive and negative d_j^i are ranked, but counted separately,

$$W_+ = \sum_i I_+(d_i^j) R_i^+. \tag{3.22}$$

Tab. 3.34: H_0 hypothesis test. Frequency of Sign test ω with noise level $\sigma^2 = 1$

N_s \ L	100	50	25
100	0.45	0.52	0.47
200	0.55	0.53	0.53
300	0.50	0.47	0.50
400	0.54	0.53	0.51
500	0.51	0.54	0.49
1000	0.49	0.49	0.52
1500	0.51	0.50	0.48
2000	0.51	0.51	0.50
2500	0.50	0.50	0.51
3000	0.50	0.51	0.50

Lets define another indicator function:

$$I_-(d_j^i) = \begin{cases} 1 & \text{if } d_j^i < 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.23}$$

and

$$W_- = \sum_i I_-(d_i^j) R_i^- \tag{3.24}$$

where R_i^+, R_i^- are ranks of d_i^j. The smallest R_i^+ is given to the smallest positive and R_i^- is given to the smallest negative d_i^j. The statistics for Wilcoxon's Signed-Rank two-sided test is $S = \min(W_+, W_-)$.

Comparing the statistic S to the critical value S_{cv} we accept the null hypothesis H_0 if $S > S_{cv}$ and reject it otherwise.

In case of causality study, we are interested in one-sided test, therefore we consider just W_+.

It is worth mentioning that $S \sim N(\mu, \sigma^2)$ as $n \to \infty$, where $\mu = n(2n+1)/2$, $\sigma^2 = n^2(2n+1)/12$. Calculating $Z = (S - \mu)/\sigma$ we can find a p-value using standard normal distribution.

Practice

Here we also start from the H_0 *hypothesis testing* from the Section 3.2 to see the performance of sign tests in case of null hypothesis is true.

H_0 hypothesis testing. The sign test produces the frequency of one SSA forecast in a first run outperforming SSA forecast in a second run. The values of frequencies in Table 3.34 point out at the stability and consistency of the sign test result. Moreover, the frequency values become nearly exact half as we increase the sample size N_s.

Tab. 3.35: MSSA frequencies of sign test

L	100	20	100	20	100	20	100	20
N_s \ σ^2	1		4		0.7		0.1	
1000	0.46	0.45	0.37	0.47	0.44	0.42	0.22	0.20
3000	0.60	0.45	0.57	0.55	0.61	0.43	0.76	0.33

Tab. 3.36: Frequency ω for sign test, where sample size $N_s = 3000$

σ^2 \ L	100	50	25
2	0.61	0.61	0.63
1	0.53	0.53	0.45
0.5	0.51	0.53	0.54
0.01	0.49	0.50	0.50

We see that the sign test perform well under the null hypothesis assumption. In the following example we are looking if its performance is as good when the null hypothesis is likely to be rejected. For the next example we use a *Rejecting H_0 hypothesis example* setup from Section 3.3.

Rejecting H_0 hypothesis example. For this example we observe that a bigger frequency ω appear for a larger sample size, i.e. $\omega_{1000} \leq \omega_{3000}$, see Table 3.35. As the oscillation of the frequency still stays around 0.5 the sign for this particular example is not conclusive.

For the next example we use the setup *Simple signal* from the Section 3.2.

Simple signal. The sign test resultant frequencies suggest that there is no significant improvement of the forecast for the relatively small noise level with variance $\sigma^2 < 2$ in the mains series. But if the noise is sufficiently large (see Table 3.36), the MSSA seems to help to catch the signal better and as the result improve the forecast of the main time series.

Moving on to the next example, where we look at the series with the signal consisting of two sines with different frequencies. The support series signal partially corresponds to the signal of the mains series.

Using the setup of *Combination of signals with different choice of eigentriples* in Section 3.2. Here we are comparing SSA with MSSA procedure and skip the comparison with MSSA_+, due to similarity of MSSA and MSSA_+. The MSSA has higher frequency for all obtained scenarios. Moreover, the frequency is quite stable for noise level with $\sigma^2 > 0.5$ and is $\omega \cong 1$ for $\sigma^2 \leq 0.5$ (see Table 3.37).

Tab. 3.37: The MSSA frequency ω of sign test for different sample sizes.

σ^2	1			2			0.5			0.01		
N_s \ L	100	50	25	100	50	25	100	50	25	100	50	25
100	0.76	0.62	0.64	0.61	0.68	0.60	0.63	0.68	0.70	0.89	1.00	1.00
200	0.67	0.64	0.71	0.71	0.62	0.67	0.71	0.68	0.72	0.90	0.99	1.00
300	0.71	0.65	0.74	0.65	0.62	0.70	0.70	0.67	0.69	0.90	1.00	1.00
400	0.65	0.63	0.68	0.68	0.61	0.70	0.66	0.65	0.75	0.88	0.99	1.00
500	0.65	0.64	0.71	0.66	0.63	0.70	0.67	0.64	0.70	0.87	0.99	1.00
1000	0.67	0.66	0.70	0.66	0.63	0.68	0.67	0.65	0.71	0.89	0.99	1.00
1500	0.67	0.64	0.70	0.64	0.62	0.69	0.67	0.64	0.70	0.87	0.99	1.00
2000	0.68	0.65	0.69	0.67	0.63	0.67	0.67	0.65	0.70	0.88	0.99	1.00
2500	0.65	0.64	0.69	0.67	0.63	0.68	0.68	0.65	0.71	0.89	0.99	1.00
3000	0.66	0.64	0.69	0.67	0.64	0.68	0.67	0.65	0.72	0.88	0.99	1.00

Tab. 3.38: The MSSA frequency ω of the 201 point for sign test with $\sigma^2 = 1$

τ \ L	100	50	25
0	0.672	0.656	0.699
0.1π	0.664	0.651	0.716
0.2π	0.688	0.663	0.746
0.3π	0.69	0.677	0.727
0.4π	0.673	0.679	0.748
0.5π	0.699	0.666	0.733

For the last example we use the setup of *Shifted signal* example in Section 3.2.

Shifted signal. The sign test frequency values varies between 0.65 and 0.75 (see Table 3.38). Roughly speaking, the frequency ω approximately stays the same. It is worth mentioning that $\omega > 0.5$ for all applied scenarios, therefore the MSSA forecast algorithm is considered to be better (improves the forecast) for the sign test.

3.5 Distribution comparison

Theory

The distribution comparison can be used, when the time series are constructed with the idea of noise permutation, as it is in our setup. As the normally distributed forecast is not always the case, we can look at the distributions of SSA/MSSA forecasts and compare them directly to each other. There are different ways of doing it, but we are are going to concentrate on the comparison of the empirical cumulative distribution functions.

The comparison of empirical cumulative distribution functions is well described in [34]. For this comparison data is represented by the loss functions (3.18) and (3.28). The error in

(3.28) is calculated with respect to the median. Here the median is taken as an alternative measure to the empirical mean, as we do not assume normality for this test.

Under the null hypothesis we assume that distributions of the $(N+j)^{th}$ forecast point are the same $H_0 : F_{SSA} = F_{MSSA}$, where F is the cumulative distribution function. If the null hypothesis gets rejected we distinguish four typical cases [34]:

1. stochastic dominance,
2. difference in variances,
3. difference in tails,
4. different structure.

For samples we are using X and Y notation, μ for the mean, σ for the standard deviation, γ for the skewness and β for the kurtosis. Adopting definitions given in [34] for our studies, we are looking at the following cases.

The *stochastic dominance* gives a clear picture of one method being dominant over another one. If one empirical cdf is above the other, then the one that is below dominates the other one. In terms of distribution parameters, when empirical means are different ($\mu_X \neq \mu_Y$), but variances are the same $\sigma_X = \sigma_Y$ the shapes of cfds are similar. Cdfs functions don't have any intersections when variances differ, $\sigma_X \neq \sigma_Y$ and $\mu_X < \mu_Y$. In fact, empirical cumulative distribution functions in both described cases, do not intersect.

Difference in variance case is suitable not only for normally distributed data comparison. If we are looking at the concentration of the distribution, the difference in variance can be treated as a stochastic dominance of one method over another one: for accuracy the small variance shows more precise forecast and for the stability the more concentrated result indicates more stable forecast. Thus, in both cases, one forecast distribution dominates another one. In terms of empirical cumulative distribution functions, difference in variance can be recognized on the graph by one intersection of empirical cumulative distribution functions (i.e. $\sigma_X \neq \sigma_Y$ and $\mu_X \approx \mu_Y$).

Difference in tails is determined by moments of higher order, such as skewness and kurtosis. Here mean and standard deviation are not significantly different. If $\gamma_X < \gamma_Y$, the longer tails of F_X and F_Y are in opposite directions. The difference in kurtosis measure $\beta_X < \beta_Y$ indicates the difference in tails too (in this case, if there is no skewness, one distribution has longer tails on both sides). Difference in tails can be recognized by two crossings of given cdfs.

Difference in structure indicates two completely different distributions, which have multiple crossings of their empirical cumulative distribution functions. In this case it is reasonable to use some tests to show any measurable difference. For example, one can still use the sign test to estimate in how many cases one method outperform another one.

Two following tests, dominance in accuracy and stability, are similar to the sign test, although they are based on the comparison of empirical cumulative distribution functions of two distributions. Note that in this test *we do not compare them pairwise from trial to trial.* We first sort these samples in increasing order and compare them afterwards.

Dominance in accuracy. This test points out which method is more accurate. By accuracy we understand the distance between the signal and the obtained forecast values. In other words, it verifies the most accurate forecast out of two, SSA and MSSA forecasts. To compare two empirical cumulative distribution functions of the loss functions (3.18), in which *comp* stands for signal, i.e. $e^i_{N+j,(M)SSA} = \mid \hat{f}^{(i,(M)SSA)}_{N+j} - f_{N+j,signal} \mid$, we look at their difference

$$d^j_{i,acc} = e^{<i>}_{N+j,SSA} - e^{<i>}_{N+j,MSSA}, \tag{3.25}$$

where $< i >$ is notation for order statistics, i.e. $e^{<i-1>}_{N+j,(M)SSA} \leq e^{<i>}_{N+j,(M)SSA} \forall i \in N_s$.

Here we detect dominance or relative dominance of one method over another by calculating the frequency of one loss function being smaller then the other one

$$\omega_{N_s,acc} = \frac{\sum_{i=1}^{Ns} I_+(d^i_{j,acc})}{N_s}. \tag{3.26}$$

If $\omega_{N_s,acc} = 1$, then we conclude that there is clear dominance of the MSSA forecast distribution (MSSA is always more accurate then SSA, MSSA empirical cdf is always below SSA); if $\omega_{N_s,acc} = 0$ then SSA dominates MSSA. If $\omega > 0.5$, then MSSA is dominating SSA in more cases and overall still dominates and vice versa.

Dominance in stability. This test is used to indicate which forecast, SSA or MSSA, is more stable. The forecast is considered to be stable if its value does not change much from trial to trial. In fact, for the real time series, when we do not know the actual (or signal) value of the forecast, it is more reasonable to use the median to estimate the stability of the obtained forecast. Here we look at the difference

$$d^j_{i,stab} = m^{<i>}_{N+j,SSA} - m^{<i>}_{N+j,MSSA}, \tag{3.27}$$

where $m^{<i>}_{N+j,(M)SSA}$ comes from

$$m^i_{N+j,(M)SSA} = \left| \hat{f}^{(i,(M)SSA)}_{N+j} - \widetilde{m}^{(M)SSA}_{N+j} \right|. \tag{3.28}$$

and $\widetilde{m}^{(M)SSA}_{N+j}$ is the median of the forecast of $(N+j)^{th}$ point. Using the same indicator function (3.20) for $d^j_{i,stab}$ one can calculate the frequency of trials where the MSSA forecast turns out to be more stable then the SSA forecast,

$$\omega_{N_s,stab} = \frac{\sum_{i=1}^{Ns} I_+(d^i_{j,stab})}{N_s}. \tag{3.29}$$

Similarly to the dominance in accuracy test if $\omega_{N_s,stab} = 1$, then we say that MSSA forecast dominates in terms of stability over SSA; if $\omega_{N_s,stab} = 0$ then the dominance is other way around.

The cumulative distribution comparison is more of a visual technique. Thus, the logical question rise if it is possible to have quantitative measure, or a characteristic number, which will indicate cases we are dealing with.

Remark We suggest the following measures which could help to indicate the type of relationship between empirical cumulative distribution functions. Define

$$\mathsf{K} = \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{s_i^j}{d_{i+}^j} \tag{3.30}$$

and

$$\mathsf{K}_{\shortparallel} = \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{\left| d_i^j \right|}{s_{i+}^j} \tag{3.31}$$

where s_{i+}^j is the sum of corresponding differences of d_i^j for accuracy

$$s_{i+}^j = e_{N+j,SSA}^{<i>} + e_{N+j,MSSA}^{<i>}, \tag{3.32}$$

and is the sum of corresponding differences of d_i^j for stability

$$s_{i+}^j = m_{N+j,SSA}^{<i>} + m_{N+j,MSSA}^{<i>}, \tag{3.33}$$

Measure (3.30) and (3.31) estimate the average distance between cdfs. Using both numerical measures together we can define between stochastic dominance and difference in variance or tails.

Each of the measures can be calculated both for stability and accuracy dominance. Both measures are ratios of *ordered* relations between the differences and the sums of loss functions (3.18) and (3.28) for accuracy and stability respectively. The division by the sum provides the invariant scaling of the time series.

The remaining difficulty is to define what can be considered as a small or a big distance between chosen empirical cumulative distribution functions.

Thus, we consider two cases, $\mid \mathsf{K} \mid << \mathsf{K}_{\shortparallel}$ and $\mid \mathsf{K} \mid \approx \mathsf{K}_{\shortparallel}$.

1. $\mid \mathsf{K} \mid << \mathsf{K}_{\shortparallel}$
   - $\mathsf{K}, \mathsf{K}_{\shortparallel}$ small: distributions are very alike or the same;
   - $\mathsf{K} << \mathsf{K}_{\shortparallel}$ and $\mathsf{K}_{\shortparallel}$ is large: difference in structure.
2. $\mid \mathsf{K} \mid \approx \mathsf{K}_{\shortparallel}$
   - $\mathsf{K}, \mathsf{K}_{\shortparallel}$ small: dominance of one distribution over another, but still very close;
   - $\mathsf{K}, \mathsf{K}_{\shortparallel}$ are large: dominance of one distribution over another, difference is more noticeable.

The measures K and $\mathsf{K}_{\shortparallel}$ should be considered more as indicators and used altogether with the visual tools, such as empirical cumulative distribution function illustrations.

Practice

Again starting from the H_0 *hypothesis testing* from the section 3.2 we look how both dominance tests for stability and accuracy perform.

Tab. 3.39: p-values for the dominance accuracy test frequency ω_{acc} ($\sigma^2 = 1$)

σ^2 \ L	100	50	25
100	0.04	0.82	0.13
200	1.00	0.86	0.82
300	0.53	0.87	0.57
400	0.91	0.61	0.52
500	0.31	0.91	0.13
1000	0.14	0.35	0.80
1500	0.58	0.45	0.31
2000	0.85	0.83	0.62
2500	0.49	0.25	0.83
3000	0.004	0.80	0.44

H_0 hypothesis testing. The resultant frequency values in Tables 3.39 - 3.40 do not have any particular pattern and seem to be random. To interpret the values for stability and accuracy dominance it is helpful to look at actual distributions of obtained values (3.18) and (3.28). Both probability distribution functions, pdf for accuracy error (3.18) and for stability error (3.28), were plotted for two sample sizes, the smallest one $N_s = 100$ and the biggest one $N_s = 3000$.

We can see from Figures 3.1 - 3.3 that the distributions obtained, both for stability and accuracy check, are approximately the same and, in fact, are both normal. For the accuracy test the mean of the accuracy error is approximately $\mu_{acc} \approx 1.5$ and the standard deviation varies slightly: for sample size $N_s = 100$, for the first run $\sigma_{SSA1} = \pm 0.1573$ and for the second run $\sigma = \pm 0.1751$. For the largest sample size $N_s = 3000$ the variance is more stable. In particular, standard deviation for both runs is $\sigma_{SSA} \approx 0.16$.

The differences between these distributions are very small overall, consequently the further analysis seem to be irrelevant and we conclude that both samples come from the same normal distribution. Also, if differences (3.25) and (3.27) are practically zeros then dominance analysis could be put aside.

Tab. 3.40: p-values for the dominance stability test frequency ω_{stab} ($\sigma^2 = 1$)

σ^2 \ L	100	50	25
100	0.49	0.92	0.20
200	0.97	0.71	0.94
300	0.53	1.00	0.51
400	0.94	0.82	0.04
500	0.49	0.98	0.40
1000	0.14	0.35	0.80
1500	0.58	0.45	0.31
2000	0.78	0.81	0.63
2500	0.57	0.28	0.82
3000	0.08	0.76	0.21

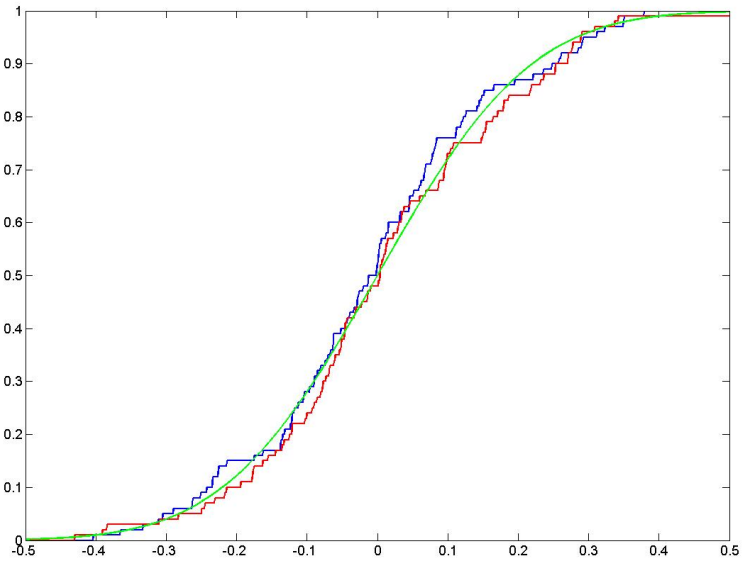


Fig. 3.1: Accuracy cdf of $d^j_{i,acc}$ in red and blue (two runs of SSA) vs normal distribution $\varepsilon \sim N(0, \sigma^2_{d^j_{i,acc}})$. $N = 200, L = 100, \sigma = 1, N_s = 100$

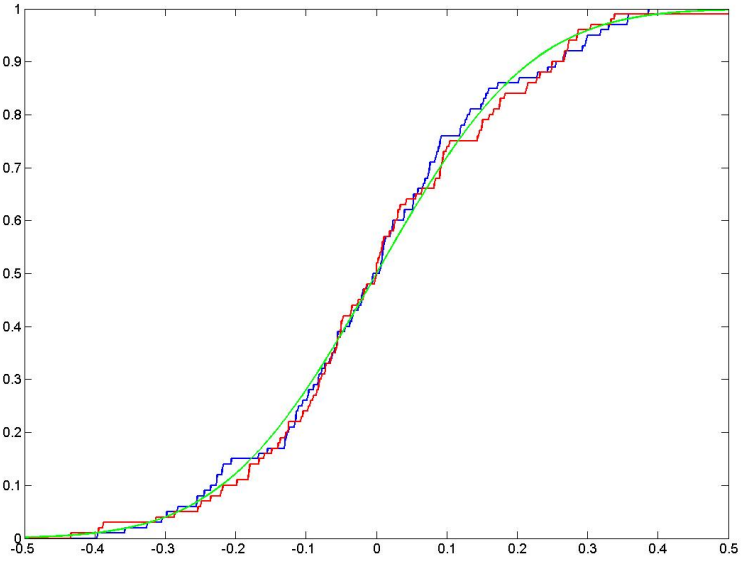


Fig. 3.2: Stability cdf of $d^j_{i,stab}$ in red and blue vs normal distribution $\varepsilon \sim N(0, \sigma^2_{d^j_{i,stab}})$. $N = 200, L = 100, \sigma = 1, N_s = 100$

Now, using suggested numerical measures K and $\mathsf{K}_{\shortparallel}$ we can check if the measures agree with both dominance tests. The values for the measure (3.30) and (3.31) do not tell much if one looks at them separately. However, looking at the values obtained for this measure in Tables 3.41-3.45, we see that in most cases $\mathsf{K} \approx \mathsf{K}_{\shortparallel}$, but there are some cases, where $\mathsf{K} \geq \mathsf{K}_{\shortparallel}$ or $\mathsf{K} \leq \mathsf{K}_{\shortparallel}$. To see a clearer picture of the measures relation, we put two additional Tables 3.43 and 3.46. The highlighted values in this tables correspond to the cdfs shown in Figures 3.1 - 3.3. For a small sample size $N_s = 100$ we see a small shift between two SSA cdfs, but for a large sample size we see that cdfs are nearly identical and get closer to standard normal distribution. Now looking at the ratio for the accuracy dominance test in (Table 3.43), we see that tthe values for these two sample are 0.99 and 1 respectively. If we look at Tables 3.41 and 3.42 we see that the corresponding values of the measures K and $\mathsf{K}_{\shortparallel}$ are small. So we conclude that two distributions are very alike, although there is a possibility of weak dominance. Overall the obtained values and conclusions drawn from them do not contradict the picture we see on Figures 3.1, 3.3. Similarly with stability measure, we see that the that there ratios are less then 1 (see highlighted values in Table 3.46) and then looking at Tables 3.44 and 3.45 we see that values obtained for the corresponding setups are small and therefore both distributions obtained are very alike or the same. This conclusion also support the picture we see on Figures 3.2, 3.4.

The next example is to show the performance of dominance tests when the null hypoth-

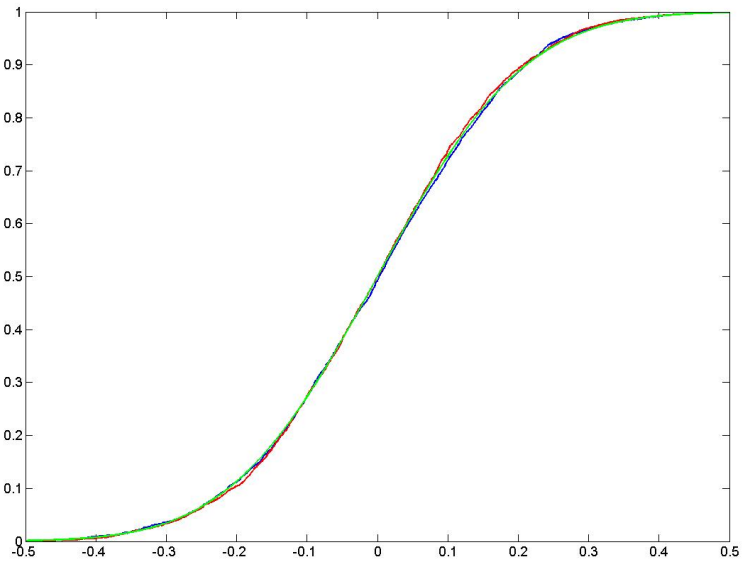


Fig. 3.3: Accuracy cdf of $d^j_{i,acc}$ in red and blue vs normal distribution $\varepsilon \sim N(0, \sigma^2_{d^j_{i,acc}})$. $N = 200, L = 100, \sigma = 1, N_s = 3000$

Tab. 3.41: Accuracy dominance measure K^{acc} with noise level $\sigma^2 = 1$

N_s \ L	100	50	25
100	**-0.108**	0.010	-0.046
200	0.168	0.075	0.034
300	0.015	0.062	0.042
400	0.049	0.046	-0.071
500	-0.002	0.051	-0.017
1000	-0.016	-0.020	0.027
1500	0.024	-0.001	-0.006
2000	0.022	0.014	0.0004
2500	0.012	-0.011	0.038
3000	**-0.030**	0.021	0.004

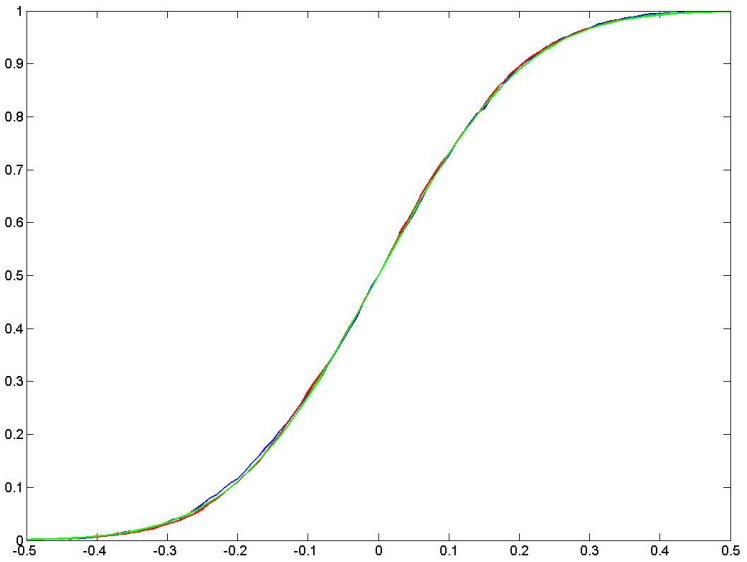


Fig. 3.4: Stability cdf of $d^j_{i,stab}$ in red and blue vs normal distribution $\varepsilon \sim N(0, \sigma^2_{d^j_{i,stab}})$. $N = 200, L = 100, \sigma = 1, N_s = 3000$

Tab. 3.42: Accuracy dominance measure $\mathsf{K}^{acc}_{||}$ with noise level $\sigma^2 = 1$

N_s \ L	100	50	25
100	**0.109**	0.053	0.063
200	0.168	0.083	0.077
300	0.039	0.097	0.052
400	0.052	0.062	0.073
500	0.038	0.088	0.036
1000	0.022	0.026	0.034
1500	0.034	0.013	0.018
2000	0.024	0.024	0.016
2500	0.021	0.025	0.04
3000	**0.030**	0.023	0.013

Tab. 3.43: Accuracy dominance measures ratio $\frac{\mathsf{K}_{\mathrm{II}}^{acc}}{\mathsf{K}^{acc}}$ with noise level $\sigma^2 = 1$

N_s \ L	100	50	25
100	**0.991**	0.189	0.73
200	1	0.9	0.442
300	0.385	0.639	0.808
400	0.942	0.742	0.973
500	0.053	0.58	0.472
1000	0.727	0.769	0.794
1500	0.706	0.0769	0.333
2000	0.917	0.583	0.025
2500	0.571	0.44	0.95
3000	**1**	0.913	0.308

Tab. 3.44: Stability dominance measure K^{stab} with noise level $\sigma^2 = 1$

N_s \ L	100	50	25
100	**-0.044**	0.061	-0.082
200	0.120	0.026	0.058
300	0.030	0.117	0.005
400	0.038	0.019	0.012
500	0.026	0.117	-0.005
1000	-0.027	0.011	0.01
1500	0.02	-0.003	0.001
2000	0.023	0.015	0.005
2500	0.014	-0.019	0.018
3000	**-0.014**	0.007	-0.009

Tab. 3.45: Stability dominance measure $\mathsf{K}_{\shortparallel}^{stab}$ with noise level $\sigma^2 = 1$

N_s \ L	100	50	25
100	**0.080**	0.064	0.089
200	0.131	0.052	0.076
300	0.051	0.117	0.038
400	0.048	0.031	0.032
500	0.0454	0.117	0.025
1000	0.031	0.031	0.02
1500	0.024	0.01	0.030
2000	0.025	0.025	0.014
2500	0.022	0.025	0.021
3000	**0.017**	0.021	0.014

Tab. 3.46: Stability dominance measures ratio $\frac{|\mathsf{K}^{stab}|}{\mathsf{K}_{\shortparallel}^{stab}}$ with noise level $\sigma^2 = 1$

N_s \ L	100	50	25
100	**0.55**	0.953	0.921
200	0.916	0.5	0.763
300	0.588	1	0.132
400	0.792	0.613	0.375
500	0.573	1	0.2
1000	0.871	0.355	0.5
1500	0.833	0.3	0.033
2000	0.92	0.6	0.357
2500	0.636	0.76	0.857
3000	**0.824**	0.333	0.643

Tab. 3.47: MSSA frequencies for Dominance tests

L	100	20	100	20	100	20	100	20
σ^2	1		4		0.7		0.1	
Dominance in accuracy								
N_s								
1000	0.00	0.07	0.22	0.03	0.09	0.02	0.00	0.00
3000	0.26	0.26	0.28	0.27	0.24	0.26	0.23	0.29
Dominance in stability								
N_s								
1000	0.01	0.00	0.19	0.15	0.00	0.06	0.00	0.00
3000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

esis is likely to be rejected. The setup is the same as for *Rejecting H_0 hypothesis example* in Section 3.3.

Both tests, accuracy and stability dominance tests, are rejecting the null hypothesis of two samples being from one distribution. In fact, we see clear dominance of SSA obtained forecast,see Table 3.47. In particular, for $N_s = 3000$ the SSA forecast outperformed the MSSA forecast in each trial.

Using the setup *Simple signal* from the Section 3.2 we are now looking at the performance of dominance tests when SSA and MSSA are applied to the time series with same signal.

Simple signal. From the dominance stability and accuracy frequencies given in Tables 3.48 and 3.49 we see that in terms of accuracy SSA and MSSA are equally good, although in terms of stability MSSA shows clear dominance.

Table 3.49 with accuracy dominance frequencies shows that for the sufficiently large noise the MSSA improves the accuracy of the forecast. Overall, the pattern for this example is following, the MSSA seem to be more stable and overtakes for a higher noise level in accuracy, although for small noise level SSA seem to be more accurate.

The results of accuracy and stability dominance tests could be easily visualized with cdfs, see Figure 3.5 for stability and Figure 3.6 for accuracy. We see that accuracy and dominance tests give very similar variance, although mean of each distribution is slightly shifted from zeros for accuracy to the right $\mu_{acc,SSA} = 0.004$,$\mu_{acc,MSSA} = 0.002$ and for stability to the left $\mu_{stab,SSA} = -0.006$,$\mu_{stab,MSSA} = -0.008$.

Moving on to the accuracy and stability measure suggested in this section in theoretical part (3.30, 3.31) and their ratios. The highlighted values in Tables 3.50-3.54 correspond to the empirical cumulative distribution functions illustrated on Figures 3.5 - 3.6. We see that all values obtained are small, and judging by ratios in 3.52 and 3.55 we see that we are in the situation when $|\mathsf{K}| < \mathsf{K}_{\shortparallel}$ for both, dominance and accuracy. Therefore the distributions are very alike, which is what we see on Figures 3.5-3.6.

Tab. 3.48: Dominance stability test frequency ω ($N_s = 3000$)

σ^2 \ L	100	50	25
1	0.97	1	0.99
2	1.00	1.00	0.99
0.5	0.99	1.00	1.00
0.01	1.00	1.00	1.00

Tab. 3.49: Dominance accuracy test frequency ω ($N_s = 3000$)

σ^2 \ L	100	50	25
1	0.49	0.53	0.53
2	0.55	0.53	0.53
0.5	0.50	0.51	0.53
0.01	0.49	0.47	0.50

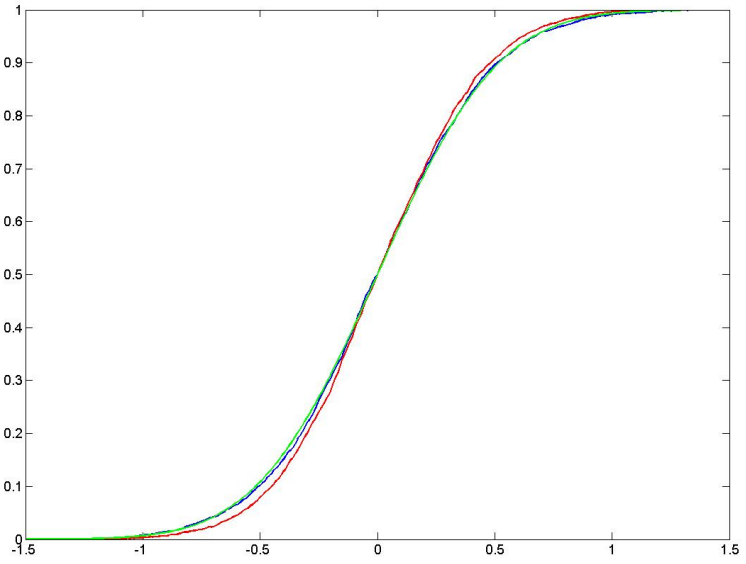


Fig. 3.5: Stability cdf $m^i_{201,SSA}$ in blue ($\sigma^2_{m_{201,SSA}} = 0.12$), $m^i_{201,MSSA}$ in red ($\sigma^2_{m_{201,MSSA}} = 0.07$), where $i \in \{1, ..., 200\}$, $L = 100$, $\sigma^2 = 2$, $N_s = 3000$ and normal cdf with SSA forecast mean and variance

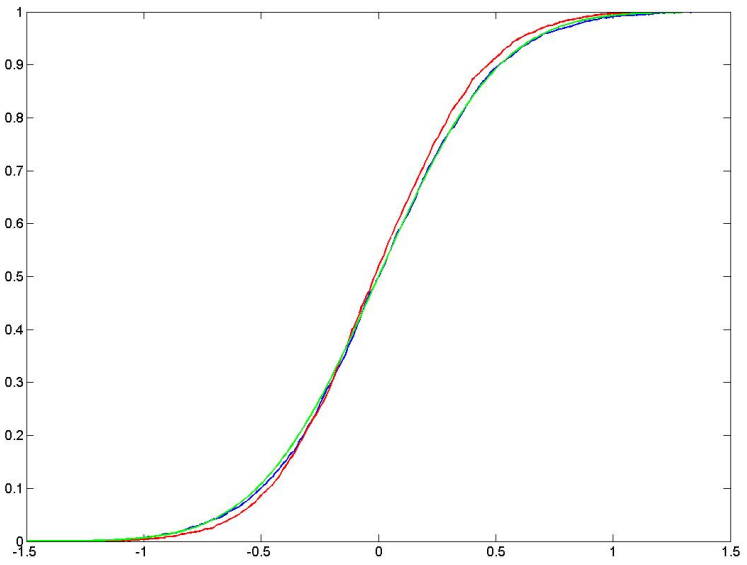


Fig. 3.6: Accuracy $e^i_{201,SSA}$ in blue ($\sigma^2_{e_{201,SSA}} = 0.12$), $e^i_{201,MSSA}$ in red ($\sigma^2_{e_{201,MSSA}} = 0.072$), where $i \in \{1, ..., 200\}$, $L = 100$, $\sigma^2 = 2$, $N_s = 3000$ and normal cdf with SSA forecast mean and variance

Tab. 3.50: Accuracy dominance measure K^{acc} for sample size $N_s = 3000$

σ^2 \ L	100	50	25
2	**-0.0116**	0.0027	-0.0073
1	-0.0073	0.0050	-0.0039
0.5	-0.0014	-0.0051	0.0051
0.01	-0.0003	0.0001	0.0002

Tab. 3.51: Accuracy dominance measure $\mathsf{K}^{acc}_{||}$ for sample size $N_s = 3000$

σ^2 \ L	100	50	25
2	**0.0146**	0.0075	0.0159
1	0.0081	0.0062	0.0056
0.5	0.0049	0.0054	0.0064
0.01	0.0004	0.0005	0.0004

Tab. 3.52: Accuracy dominance measures ratio $\frac{\mathsf{K}_{\shortparallel}^{acc}}{\mathsf{K}^{acc}}$ for sample size $N_s = 3000$

σ^2 \ L	100	50	25
2	**1.26**	2.78	2.18
1	1.11	1.24	1.44
0.5	3.5	1.06	1.26
0.01	1.33	5	2

Tab. 3.53: Stability dominance measure K^{stab} for sample size $N_s = 3000$

σ^2 \ L	100	50	25
2	**-0.005**	0.007	-0.0255
1	0.0217	-0.0175	-0.0327
0.5	-0.0474	-0.0092	0.027
0.01	0.0007	0.0281	0.0007

Tab. 3.54: Stability dominance measure $\mathsf{K}_{\shortparallel}^{stab}$ for sample size $N_s = 3000$

σ^2 \ L	100	50	25
2	**0.0179**	0.0199	0.0379
1	0.0352	0.0234	0.0387
0.5	0.0477	0.0054	0.0064
0.01	0.0004	0.0005	0.0004

Tab. 3.55: Stability dominance measures ratio $\frac{|\mathsf{K}^{stab}|}{\mathsf{K}_{\shortparallel}^{stab}}$ for sample size $N_s = 3000$

σ^2 \ L	100	50	25
2	**0.28**	0.35	0.67
1	0.62	0.75	0.85
0.5	0.99	1.70	4.22
0.01	1.75	56.2	1.75

Tab. 3.56: Dominance frequencies for the $(N+1)$ point. $\sigma^2 = 1$

τ \ L	100	50	25
ω_{acc}			
0	0.956	0.977	0.941
0.1π	0.978	1	0.973
0.1π	0.997	0.982	0.959
0.1π	1	0.946	0.996
0.1π	0.909	1	0.994
0.1π	0.994	0.966	0.995
ω_{stab}			
0	0.987	0.992	0.949
0.1π	0.998	0.984	0.952
0.1π	0.927	0.958	0.94
0.1π	0.961	1	0.989
0.1π	0.994	0.941	0.998
0.1π	0.999	0.997	0.959

Remark For *Combination of signals with different choice of eigentriples* defined in Section 3.2 both dominance tests produce very promising results as well. Nearly for any choice of time series and procedure parameters frequency for accuracy and stability dominance tests is $\omega_{acc,stab} \cong 1$, respectively.

The last example is to see how the delay effects the dominance tests. We use the setup *Shifted signal* example from Section 3.2.

Shifted signal. As for the dominance in accuracy and stability, this test show nearly clear stochastic dominance of the MSSA algorithm (see Table 3.56).

All previously described tests point at the fact of causality, if it exists, being independent from the delay in the support series. Moreover this example shows that dominance tests for accuracy and stability seem to be a good measure of causality. All the examples, and this one is not exception, were built so that the presence or absence of causality is obvious. Here the support series was presented by the part of the signal of main series with noise, i.e. we expect causality.

Both dominance tests gave convincing results for the case, where we did not expect causality at all (see 3.47).

3.6 A relationship between loss functions

Here are several main observations made on loss functions in Section 3.2. In Section 3.2 we observed that loss functions $S^{(1)}$ and $S^{(3)}$ are approximately the same if the choice of eigentriples for the analysis was made accordingly to the signal. In other words, the mean

of the forecast point distribution gives a value close to the actual signal value of the time series.

We also observe following relationship between loss functions $S^{(1)}$ and $S^{(2)}$

$$S^{(2)} \approx S^{(1)} + \sigma^2.$$

In fact, values of the loss function $S^{(2)}$ seem to reflect the standard deviation of the noise directly. Here σ^2 correspond to the variance of the white noise in the main time series (for example see Table 3.1,3.3, 3.7, 3.8 and other tables from the Section 3.2).

The last statement can be explained as follows.

Lemma 3.6.1. *Suppose we have time series of the following form*

$$\begin{array}{rr} \textit{main series} & f_n = s_n + \sigma\varepsilon_n \\ \textit{support series (only for MSSA)} & g_n \end{array} \tag{3.34}$$

where s_n and g_n correspond to signal in the main and the support series respectively and $\varepsilon_n \sim N(0,1)$ is i.i.d. white noise. Then, assuming that the forecast errors are approximately independent of the ε_n, then

$$S^{(2)} \approx S^{(1)} + \sigma^2 \tag{3.35}$$

where $S^{(1)}$ (3.5) and $S^{(2)}$ (3.6) are loss functions, and σ^2 is the variance of the noise in the main series

Proof. Recalling the expressions for $S^{(1)}$ (3.5) and $S^{(2)}$ (3.6) we can look at the difference $S^{(1)} - S^{(2)}$. We are using the following facts: the signal value $f^{(i)}_{N+j,s}$ is independent of the trial index i; the forecast value is $\hat{f}^{(i)}_{N+j} = f^{(i)}_{N+j,s} + \tilde{\varepsilon}^{(i)}_{N+j}$, where $\tilde{\varepsilon}^{(i)}_{N+j}$ is the forecast error at $(N+j)^{th}$ point for the i^{th} trial; the actual value of the time series $f^{(i)}_{N+j} = f^{(i)}_{N+j,s} + \sigma\varepsilon^{(i)}_{N+j}$, where $\varepsilon \sim N(0,1)$. If we rewrite now $S^{(1)} - S^{(2)}$, substituting all the described values, we will get

$$\begin{aligned} S^{(1)} - S^{(2)} &= \frac{\sum_{i=1}^{N_s} (\hat{f}^{(i)}_{N+j} - f^{(i)}_{N+j,sig})^2}{N_s} - \frac{\sum_{i=1}^{N_s} (\hat{f}^{(i)}_{N+j} - f^{(i)}_{N+j})^2}{N_s} \\ &= \frac{\sum_{i=1}^{N_s} -2f^{(i)}_{N+j,sig}\hat{f}^{(i)}_{N+j} + {f^{(i)}_{N+j,sig}}^2 + 2\hat{f}^{(i)}_{N+j}f^{(i)}_{N+j} - {f^{(i)}_{N+j}}^2}{N_s} \\ &= \frac{2\sum_{i=1}^{N_s}(\tilde{\varepsilon}^{(i)}_{N+j}\varepsilon^{(i)}_{N+j}) - \sum_{i=1}^{N_s} {\varepsilon^{(i)}_{N+j}}^2}{N_s}. \end{aligned} \tag{3.36}$$

Due to independence of two random variables $\tilde{\varepsilon}^{(i)}_{N+j}$ and $\varepsilon^{(i)}_{N+j}$, the expectation of their product is the product of their expectations, and expectation of ε_n is zero, and therefore

$$S^{(1)} - S^{(2)} \approx -\frac{\sum_{i=1}^{N_s} {\varepsilon^{(i)}_{N+j}}^2}{N_s} \Leftrightarrow S^{(2)} \approx S^{(1)} + \sigma_2^2, \tag{3.37}$$

where σ_2^2 is the variance of ε_{N+j} white noise. □

Proposition 3.6.2. *Suppose we have time series of the from (3.34). If* $\sigma^2_{SSA} >> S^{(1)}_{SSA}$ *and* $\sigma^2_{MSSA} >> S^{(1)}_{MSSA}$ *then the ratio of loss functions* $\frac{S^{(2)}_{MSSA}}{S^{(2)}_{SSA}}$ *is approximately equal to the ratio of the variances of the noise in main series.*

Proof. Recalling the statement of Lemma 3.34, we can rewrite the ratio $\frac{S^{(2)}_{MSSA}}{S^{(2)}_{SSA}}$ as follows

$$\frac{S^{(2)}_{MSSA}}{S^{(2)}_{SSA}} \approx \frac{S^{(1)}_{MSSA} + \sigma^2_{MSSA}}{S^{(1)}_{SSA} + \sigma^2_{SSA}}. \tag{3.38}$$

If $\sigma^2_{(M)SSA} >> S^{(1)}_{(M)SSA}$ then

$$\frac{S^{(1)}_{MSSA} + \sigma^2_{MSSA}}{S^{(1)}_{SSA} + \sigma^2_{SSA}} \approx \frac{\sigma^2_{MSSA}\left(1 + \frac{S^{(1)}_{MSSA}}{\sigma^2_{MSSA}}\right)}{\sigma^2_{SSA}\left(1 + \frac{S^{(1)}_{SSA}}{\sigma^2_{SSA}}\right)} \approx \frac{\sigma^2_{MSSA}}{\sigma^2_{SSA}} \tag{3.39}$$

Hence, the loss functions ratio becomes a ratio of two noise variances. □

3.7 Summary

We looked at several statistical tests and tools to estimate the forecast, obtained by the SSA and MSSA forecast algorithms. Following the definition of causality, the forecast estimation consequently helps to detect it.

Firstly we have generated time series of four types, so that it is possible to see the performance of statistical tests in different cases.

In Section 3.2 we have constructed three loss functions, so that it is possible to estimate the error in forecast using the signal values at the forecast point, the actual values of the time series at this point or using empirical mean calculated from obtained forecast distribution (see 3.5).

The most common causality test is F-test. The F-test is based on the idea of comparing two obtained variances by looking at their ratio. In fact, the ratios of the loss functions $S^{(3)}$ is the F-test statistic. To apply F-test we first need to make sure that the data satisfies the properties of normal distribution. For this purposes we have used two normality tests - KS and AD tests.

The observation on normality tests have shown that AD test is more sensitive than KS test (for example, see Table 3.25). We assume that KS test is good enough for the purposes of the study of this section. Therefore, we are able to use F-test, if data passes KS test and in most cases it did.

The F-test, in turn, performed according to our expectations. However, it seems that if we increase the sample size, the p-value of F-test tends to zero, as F statistic seems to be independent on sample size. For example, see Tables 3.29-3.30. Therefore, the F-test statistic (ratio of variances) appear to be a more formative indicator for causality then the actual test p-value, which depends on the (arbitrary) sample size.

The sign test is a simple calculation of how many times one method outperforms the other one. In our case it is how many times MSSA outperform SSA forecast algorithm. In most cases sign test supports our assumptions made about existence of causality between time series. However, in case when we do not expect any causal relationship, the result of the test is inconclusive (see Table 3.35).

The test to see which of two methods dominates, as described in Section 3.5, is based on the idea of direct comparison of empirical distributions. The main advantage of this method is that although we are not making assumptions about the distributions our data come from,we still are able to compare them and see if one gives a better result than the other one. We defined four cases of distinctions between two distributions in Section 3.5: stochastic dominance, difference in variance, difference in tails and difference in structure. We studied the dominance tests from two angles, from accuracy point of view and from the point of view of stability. Also, we have introduced possible indication functions, which help to define the case we are dealing with.

4. LINEARIZED MSSA AND CAUSALITY

4.1 MSSA analogue of Granger causality

It is our aim to develop a new approach to causality based on the SSA/MSSA decomposition of time series. The fundamental difference is that the underlying model is not autoregressive, but a deterministic linear recurrence with added noise (although the coefficients in this model still contain some randomness). In this model the noise does not carry any information on the model parameters, in contrast to the VAR model (2.32), which could be reduced to considering the propagation of the noise only.

As for Granger's concept, the whole idea is based on the fact, that the underlying model is some white noise based stationary process (see section 2.2).

More explicitly, the bivariate MSSA approach provides the following model. Let X_t and Y_t be some real time series; then after MSSA decomposition they are modeled as

$$\begin{aligned} X_t &= \widetilde{X}_t + \varepsilon_t \\ Y_t &= \widetilde{Y}_t + \eta_t, \end{aligned} \tag{4.1}$$

where $\widetilde{X}_t$, $\widetilde{Y}_t$ are their reconstructions, obtained with the MSSA linear recurrence formulas

$$\begin{aligned} \widetilde{\widetilde{X}_t} &= \sum_{j=1}^{m} a_j \widetilde{X}_{t-j} + \sum_{j=1}^{m} b_j \widetilde{Y}_{t-j} \\ \widetilde{\widetilde{Y}_t} &= \sum_{j=1}^{m} c_j \widetilde{X}_{t-j} + \sum_{j=1}^{m} d_j \widetilde{Y}_{t-j} \end{aligned} \tag{4.2}$$

and ε_t, η_t are residuals

$$\begin{aligned} \varepsilon_t &= X_t - \widetilde{\widetilde{X}_t} \\ \eta_t &= Y_t - \widetilde{\widetilde{Y}_t}. \end{aligned}$$

The main difference between this model obtained from MSSA and the AR model is that in the AR model the noise is driving the model, and in the model constructed with the MSSA approach the noise is excluded from the recurrence and is added as residuals after the model is built. In other words, the main difference is in the way of noise entering the model $\varepsilon_{t-1}, \eta_{t-1}$ do not affect X_t or Y_t for any t.

Using the MSSA approach, one is trying to separate noise from the signal and noise is treated as an irrelevant part of the model, which does not contain any useful information. That is, in this approach noise does not play crucial role in building up the model.

Considering the MSSA model in a vector form

$$X_t = R_{11}^{\intercal}\widetilde{\mathbb{X}}_{t-1} + R_{12}^{\intercal}\widetilde{\mathbb{Y}}_{t-1} \tag{4.3}$$

$$Y_t = R_{21}^{\intercal}\widetilde{\mathbb{X}}_{t-1} + R_{22}^{\intercal}\widetilde{\mathbb{Y}}_{t-1} \tag{4.4}$$

where $\widetilde{\mathbb{X}}_{s-1} = (\widetilde{X}_{s-L}, \widetilde{X}_{s-L+1}, ..., \widetilde{X}_{s-1})$, $\widetilde{\mathbb{Y}}_{s-1} = (\widetilde{Y}_{s-L}, \widetilde{Y}_{s-L+1}, ..., \widetilde{Y}_{s-1})$ and $R_{11}, R_{12}, R_{21}, R_{22}$ are the recurrence vectors obtained from (1.56), we have

$$\begin{aligned}
R_{11} &= \frac{1}{\Delta}\left((1-\sum_k \eta_{k,2L}^2)\sum_k \eta_{k,L}\eta_k^{(1)} + (\sum_k \eta_{k,2L})\sum_k \eta_k^{(1)}\eta_{k,2L})\right) \\
R_{12} &= \frac{1}{\Delta}\left((1-\sum_k \eta_{k,2L}^2)\sum_k \eta_{k,L}\eta_k^{(2)} + (\sum_k \eta_{k,2L})\sum_k \eta_k^{(2)}\eta_{k,2L})\right) \\
R_{21} &= \frac{1}{\Delta}\left((1-\sum_k \eta_{k,L}^2)\sum_k \eta_{k,L}\eta_k^{(1)} + (\sum_k \eta_{k,2L})\sum_k \eta_k^{(1)}\eta_{k,2L})\right) \\
R_{22} &= \frac{1}{\Delta}\left((1-\sum_k \eta_{k,L}^2)\sum_k \eta_{k,L}\eta_k^{(2)} + (\sum_k \eta_{k,2L})\sum_k \eta_k^{(2)}\eta_{k,2L})\right)
\end{aligned} \tag{4.5}$$

where Δ, see (4.8) below, is the determinant of the matrix A defined in (1.54).

4.2 Linearized MSSA

Defining the MSSA forecast of the time series x with the support of time series y as $f(x \mid y)$ we can now more formally describe one difficulty of the MSSA approach with regard to measuring the causality.

A general problem with MSSA is the lack of homogeneity. In other words, if we change the support series by multiplying it by some constant, the resulting forecast of the main series may vary. This statement can be illustrated by a simple example.

Consider the extreme case of some finite main time series x_t and a proportional support series λx_t, so that their Hankel matrices are $\mathbb{X}$ and $\lambda\mathbb{X}$ respectively, both of $L \times K$ size. Then the eigenvectors and eigenvalues of the matrix

$$S = \begin{pmatrix}\mathbb{X} & \lambda\mathbb{X}\end{pmatrix}\begin{pmatrix}\mathbb{X}^{\intercal} \\ \lambda\mathbb{X}^{\intercal}\end{pmatrix} = \begin{pmatrix}\mathbb{X}\mathbb{X}^{\intercal} & \lambda\mathbb{X}\mathbb{X}^{\intercal} \\ \lambda\mathbb{X}\mathbb{X}^{\intercal} & \lambda^2\mathbb{X}\mathbb{X}^{\intercal}\end{pmatrix} = \begin{pmatrix}1 & \lambda \\ \lambda & \lambda^2\end{pmatrix} \otimes \mathbb{X}\mathbb{X}^{\intercal}$$

are $(\lambda^2+1)\mu_1, ..., (\lambda^2+1)\mu_L, 0\mu_1, ..., 0\mu_L$, where $\mu_1, ..., \mu_L$ are the eigenvalues of $\mathbb{X}\mathbb{X}^{\intercal}$. Eigenvectors corresponding to non-zero eigenvalues are $\begin{pmatrix}V \\ \lambda V\end{pmatrix}$, where $V = \begin{pmatrix}V^{\triangledown} \\ V_L\end{pmatrix}$ is of $L \times 1$ size and $V^{\triangledown}$ contains first $(L-1)$ elements of the vector and V_L is the L^{th} element

of vector V. Substituting these eigenvectors into MSSA recurrence vectors (4.5), we get

$$R_{11} = \frac{1}{\Delta}\left((1-\lambda^2\sum_k V_{k,L}^2)\sum_k V_{k,L}V_k^{\triangledown} + \lambda^2(\sum_k V_{k,L})\sum_k V_k^{\triangledown}V_{k,L})\right), \tag{4.6}$$

$$R_{12} = \frac{1}{\Delta}\left((1-\lambda^2\sum_k V_{k,L}^2)\lambda\sum_k V_{k,L}V_k^{\triangledown} + \lambda^3(\sum_k V_{k,L})\sum_k V_k^{\triangledown}V_{k,L})\right), \tag{4.7}$$

where

$$\Delta = (1-\sum_k V_{k,L}^2)(1-\lambda\sum_k V_{k,L}^2) - \lambda^2(\sum_k V_{k,L}^2)(\sum_l V_{l,L}^2). \tag{4.8}$$

Causality appears between time series structures (signals), which include trends, periodicities and exclude noise. Therefore, one would expect that the forecast model coefficients should be independent of $\lambda \in \mathbb{R}$. In other words, $\hat{f}(x_t \mid y_t) = \lambda\hat{f}(x_t \mid \lambda x_t)$. But the above substitutions (4.6)-(4.8) shows that $\hat{f}(x_t \mid y_t) \neq \lambda\hat{f}(x_t \mid \lambda y_t)$. This can be illustrated by the following example.

Suppose we have a time series of length N:

$$\begin{cases} x_t = sin(\pi n\omega) + sin(3\pi n\omega) + \varepsilon_t \\ y_t = \lambda sin(\pi n\omega) \end{cases}$$

where ε_t is Gaussian white noise with variance σ_ε.

Consider $N = 200$, $\sigma_\varepsilon^2 = 1$, $\lambda_1 = 0.1$, $\lambda_2 = 10$. One would assume that the MSSA based forecast of x_t does not vary on λ. Hence, we expect $\hat{f}(x_t|\lambda_1 y_t) \approx \hat{f}(x_t|\lambda_2 y_t)$. Applying the MSSA procedure for the same set of generated data (changing just multiplier λ) we obtain the forecast for the 201^{st} value x_{201}: $\hat{f}(x_t|0.1y_t) = -0.3612$, and $\hat{f}(x_t|10y_t) = -0.2864$. Obviously, $\hat{f}(x_t|0.1y_t) \not\approx \hat{f}(x_t|10y_t)$.

This lack of homogeneity in the support series is a fundamental difficulty of MSSA. Consider two time series of quantities which have different units, e.g. x_t is a price measured in £ and y_t is a weight measured in kilograms. Clearly then a_j in (4.2) is purely numerical and b_j are measured in pounds per kilogram units (£/kg). One should think that the analysis should be no different if Y_t is measured not in kilograms, but in tonnes instead, i.e. if the values of Y_t are just multiplied by 10^{-3}, then the values of b_j should be multiplied by 10^3 and should be measured in pounds per tonne, while a_j should stay unchanged. However, this is not the case for standard MSSA due to the non-linearity of the spectral decomposition of the lag-covariance matrix.

We therefore introduce *linearized MSSA*, which is linear in the support series. For this purpose we essentially take the (functional) derivative of MSSA with respect to the support series.

Proposition 4.2.1. *Suppose we have main series x_t and support series y_t, $t \in \mathbb{N}$, both finite. Let ϵ be a small parameter. Note, that we do not make any assumptions on the structure of time series.*

Let $\mathbb{X}_1$ be the Hankel matrix of x_t, and $\mathbb{X}_2$ be the Hankel matrix of y_t (note that then the Hankel matrix of ϵy_t is $\epsilon\mathbb{X}_2$). Using the first order perturbation theory, neglecting terms of ϵ^2 and higher, we get the MSSA lag-covariance matrix

$$\begin{bmatrix}\mathbb{X}_1\\ \epsilon\mathbb{X}_2\end{bmatrix}\begin{bmatrix}\mathbb{X}_1 & \epsilon\mathbb{X}_2\end{bmatrix} = \begin{bmatrix}\mathbb{X}_1\mathbb{X}_1^\intercal & \epsilon\mathbb{X}_1\mathbb{X}_2^\intercal\\ \epsilon\mathbb{X}_2\mathbb{X}_1^\intercal & \epsilon^2\mathbb{X}_2\mathbb{X}_2^\intercal\end{bmatrix} \approx \begin{bmatrix}\mathbb{X}_1\mathbb{X}_1^\intercal & \epsilon\mathbb{X}_1\mathbb{X}_2^\intercal\\ \epsilon\mathbb{X}_2\mathbb{X}_1^\intercal & 0\end{bmatrix} = A + \epsilon B \tag{4.9}$$

where

$$A = \begin{bmatrix}\mathbb{X}_1\mathbb{X}_1^\intercal & 0\\ 0 & 0\end{bmatrix}$$

and

$$\epsilon B = \epsilon\begin{bmatrix}0 & \mathbb{X}_1\mathbb{X}_2^\intercal\\ \mathbb{X}_2\mathbb{X}_1^\intercal & 0\end{bmatrix}.$$

The matrix A represents the unperturbed system, where the support series is equal to zero with eigenvectors $u = \begin{pmatrix}\eta\\ 0\end{pmatrix}$, where η is the eigenvector of matrix $\mathbb{X}_1\mathbb{X}_1^\intercal$

Then the approximation of the recurrence vectors (1.60) - (1.63) to the first order with respect to ϵ is as following

$$\begin{aligned}
R_{11}(\epsilon) &= \frac{1}{\Delta_\epsilon}\sum_k \eta_{k,L}\eta_k^{(L-1)},\\
R_{12}(\epsilon) &= \frac{1}{\Delta_\epsilon}\epsilon\sum_k \eta_{k,L}\gamma_k^{(L-1)},\\
R_{21}(\epsilon) &= \frac{1}{\Delta_\epsilon}\epsilon\sum_k \left(\gamma_{k,L}(1-\sum_l \eta_{l,L}^2) + \eta_{k,L}(\sum_l \eta_{l,L}\gamma_{l,L})\right)\eta_k^{(L-1)},\\
R_{22}(\epsilon) &= \frac{1}{\Delta_\epsilon}\epsilon^2\sum_k \left(\gamma_{k,L}(1-\sum_l \eta_{l,L}^2) + \eta_{k,L}(\sum_l \eta_{l,L}\gamma_{l,L})\right)\gamma_k^{(L-1)},
\end{aligned}$$

where $\gamma = \frac{1}{\lambda}\mathbb{X}_2\mathbb{X}_1^\intercal\eta$

Proof. Matrix A is an equivalent to the matrix $\mathbb{X}\mathbb{X}^\intercal$ in the SSA case, in which we have the matrix $\mathbb{X}\mathbb{X}^\intercal$ for eigendecomposition.

Let λ be the eigenvalue of A and u its corresponding eigenvector. The relationship between the eigenvalues and eigenvectors of A and $\mathbb{X}\mathbb{X}^\intercal$ is straightforward. Specifically, matrix A have eigenvectors of the form $u = \begin{pmatrix}\eta\\ 0\end{pmatrix}$, where η is an eigenvector of $\mathbb{X}\mathbb{X}^\intercal$ and 0 in this case is representing L extra zeros. Each eigenvalue of $\mathbb{X}\mathbb{X}^\intercal$ is also an eigenvalue of A, but matrix A also has additional eigenvalues which are equal to zero.

Now we consider the perturbation $A + \epsilon B$ of A. Performing spectral decomposition of the unperturbed matrix A, one finds the eigenvalues λ and eigenvectors u of the matrix, so that

$$Au = \lambda u.$$

For the perturbed matrix $A + \epsilon B$, there are eigenvectors u_ϵ, and eigenvalues λ_ϵ analytical in ϵ

$$u_\epsilon = u + \epsilon\nu + \epsilon^2\nu_2 + ... \tag{4.10}$$

and

$$\lambda_\epsilon = \lambda + \epsilon\mu_1 + \epsilon^2\mu_2 + ... \tag{4.11}$$

such that

$$(A + \epsilon B)u_\epsilon = \lambda_\epsilon u_\epsilon, \tag{4.12}$$

see [22, Ch.II, Theorem 2.3],[31, p.245-246].

Substituting the power series (4.10) and (4.11) for λ_ϵ and u_ϵ into (4.12), we find

$$Au + \epsilon(A\nu + Bu) + O(\epsilon^2) = \lambda u + \epsilon(\lambda\nu + \mu_1 u) + O(\epsilon^2). \tag{4.13}$$

Knowing that

$$Au - \lambda u = 0 \tag{4.14}$$

and comparing coefficients of ϵ in (4.13), we get

$$A\nu + Bu = \lambda\nu + \mu_1 u. \tag{4.15}$$

Multiplying both sides by $u^\intercal$ from the left, we have

$$u^\intercal A\nu + u^\intercal Bu = \lambda u^\intercal\nu + \mu_1 u^\intercal u,$$

i.e.

$$u^\intercal(A - \lambda)\nu = -u^\intercal Bu + \mu_1 u^\intercal u,$$

and using the fact that $(A - \lambda)$ is a symmetric matrix and (4.14)

$$u^\intercal(A - \lambda)\nu = ((A - \lambda)u^\intercal)^T\nu = ((A - \lambda)u)^T\nu = 0,$$

we get

$$-u^\intercal Bu + \mu_1 u^\intercal u = 0.$$

Thus, we determine the first order perturbation term for the eigenvalue λ_ϵ

$$\mu_1 = \frac{u^\intercal B u}{u^\intercal u}.$$

Substituting this expression for μ_1 back into (4.15), we find

$$A\nu + Bu = \lambda\nu + \frac{u^\intercal B u}{u^\intercal u} u.$$

Writing out the matrices A, B and using normalized eigenvectors η of $\mathbb{X}_1\mathbb{X}_1^\intercal$, $\eta^\intercal\eta = 1$, and hence $u^\intercal u = 1$, we have

$$\begin{pmatrix} \mathbb{X}_1\mathbb{X}_1^\intercal - \lambda & 0 \\ 0 & -\lambda \end{pmatrix} \nu = \begin{pmatrix} 0 & \mathbb{X}_1\mathbb{X}_2^\intercal \\ \mathbb{X}_2\mathbb{X}_1^\intercal & 0 \end{pmatrix} \begin{pmatrix} \eta \\ 0 \end{pmatrix} + \underbrace{\begin{pmatrix} \eta \\ 0 \end{pmatrix}^\intercal \begin{pmatrix} 0 & \mathbb{X}_1\mathbb{X}_2^\intercal \\ \mathbb{X}_2\mathbb{X}_1^\intercal & 0 \end{pmatrix} \begin{pmatrix} \eta \\ 0 \end{pmatrix}}_{=0} \begin{pmatrix} \eta \\ 0 \end{pmatrix}. \quad (4.16)$$

Thus,

$$\begin{pmatrix} \mathbb{X}_1\mathbb{X}_1^\intercal - \lambda & 0 \\ 0 & -\lambda \end{pmatrix} \nu = \begin{pmatrix} \mathbb{X}_2\mathbb{X}_1^\intercal \eta \\ 0 \end{pmatrix}. \quad (4.17)$$

Writing $\nu = \begin{pmatrix} \nu^\top \\ \nu^\perp \end{pmatrix}$, we can derive $\nu^\top$, $\nu^\perp$ from (4.17)

$$\begin{cases} \mathbb{X}_1\mathbb{X}_1^\intercal \nu^\top = \lambda\nu^\top \\ -\lambda\nu^\perp = \mathbb{X}_2\mathbb{X}_1^\intercal \nu \end{cases}$$

where

$$\nu^\top = c\eta$$
$$-\nu^\perp = \frac{1}{\lambda}\mathbb{X}_2\mathbb{X}_1^\intercal \nu.$$

Hence, the first order perturbation term for the eigenvector is

$$\nu = \begin{pmatrix} c\eta \\ \frac{1}{\lambda}\mathbb{X}_2\mathbb{X}_1^\intercal \eta \end{pmatrix}, \quad (4.18)$$

where c is some constant.

Substituting u and ν to (4.10), we get the expression for the eigenvector of the perturbed matrix $(A + \epsilon B)$ with one unknown c.

We can calculate the constant c from the normalization of the eigenvector u_ϵ. Indeed,

$$u_\epsilon^\intercal u_\epsilon = 1 \Leftrightarrow \underbrace{u^\intercal u}_{=1} + \epsilon\nu^\intercal u + \epsilon u^\intercal \nu + \epsilon^2 \nu^\intercal \nu = 1, \quad (4.19)$$

therefore

$$\nu^\intercal u + u^\intercal \nu = -\epsilon\nu^\intercal \nu. \quad (4.20)$$

Now for (4.18), we have

$$u^\intercal \nu = c\eta^\intercal \eta = \nu^\intercal u$$

and

$$\nu^\intercal \nu = c^2 \eta^\intercal \eta + \frac{1}{\lambda^2} \eta^\intercal \mathbb{X}_1 \mathbb{X}_2^\intercal \mathbb{X}_2 \mathbb{X}_1^\intercal \eta.$$

Thus, using

$$u^\intercal \nu = \frac{\epsilon}{2} \nu^\intercal \nu \Leftrightarrow \nu^\intercal \nu = \frac{2}{\epsilon} u^\intercal \nu = \frac{2}{\epsilon} c\eta^\intercal \eta,$$

the formula (4.20) can be rewritten as

$$c \underbrace{\eta^\intercal \eta}_{=1} = \frac{\epsilon}{2}(c^2 + \frac{1}{\lambda^2} \eta^\intercal \mathbb{X}_1 \mathbb{X}_2^\intercal \mathbb{X}_2 \mathbb{X}_1^\intercal \eta).$$

By multiplying both sides by $\frac{2}{\epsilon}$ and adding and subtracting $\frac{1}{\epsilon^2}$, we get

$$\underbrace{c^2 - \frac{2}{\epsilon} c + \frac{1}{\epsilon^2}}_{(c-\frac{1}{\epsilon})^2} - \frac{1}{\epsilon^2} + \frac{1}{\lambda^2} \eta^\intercal \mathbb{X}_1 \mathbb{X}_2^\intercal \mathbb{X}_2 \mathbb{X}_1^\intercal \eta = 0,$$

$$(c - \frac{1}{\epsilon}) = \pm \sqrt{\frac{1}{\epsilon^2} - \frac{1}{\lambda^2} \eta^\intercal \mathbb{X}_1 \mathbb{X}_2^\intercal \mathbb{X}_2 \mathbb{X}_1^\intercal \eta}$$

and we get two roots

$$c = \frac{1}{\epsilon} \pm \frac{1}{\epsilon} \sqrt{1 - \frac{\epsilon^2}{\lambda^2} \eta^\intercal \mathbb{X}_1 \mathbb{X}_2^\intercal \mathbb{X}_2 \mathbb{X}_1^\intercal \eta}$$

where

$$\sqrt{1 - \tfrac{\epsilon^2}{\lambda^2} \eta^\intercal \mathbb{X}_1 \mathbb{X}_2^\intercal \mathbb{X}_2 \mathbb{X}_1^\intercal \eta} = 1 + \tfrac{1}{2} \tfrac{\epsilon^2}{\lambda^2} \eta^\intercal \mathbb{X}_1 \mathbb{X}_2^\intercal \mathbb{X}_2 \mathbb{X}_1^\intercal \eta + O\left((\tfrac{\epsilon^2}{\lambda^2} \eta^\intercal \mathbb{X}_1 \mathbb{X}_2^\intercal \mathbb{X}_2 \mathbb{X}_1^\intercal \eta)^2 \right)$$

Hence the constant c can be expressed as

$$c \approx \frac{1}{\epsilon} \pm \frac{1}{\epsilon} (1 + \frac{1}{2} \frac{\epsilon^2}{\lambda^2} \eta^\intercal \mathbb{X}_1 \mathbb{X}_2^\intercal \mathbb{X}_2 \mathbb{X}_1^\intercal \eta).$$

But as $\frac{1}{\epsilon} \to \infty$ ($\epsilon \to 0$),the following root is not suitable since

$$\frac{1}{\epsilon} + \frac{1}{\epsilon} (1 + \frac{1}{2} \frac{\epsilon^2}{\lambda^2} \eta^\intercal \mathbb{X}_1 \mathbb{X}_2^\intercal \mathbb{X}_2 \mathbb{X}_1^\intercal \eta) \to \infty.$$

Therefore, the final expression for c is

$$c \approx -\frac{1}{2} \frac{\epsilon}{\lambda^2} \eta^\intercal \mathbb{X}_1 \mathbb{X}_2^\intercal \mathbb{X}_2 \mathbb{X}_1^\intercal \eta.$$

Recalling the equation for the eigenvector of the perturbed matrix $A + \epsilon B$ (4.10) and the equation of the first order perturbation term (4.18), the final expression for the eigenvector to first order in ϵ is

$$u_\epsilon = \begin{pmatrix} \eta(1 - \underbrace{\frac{\epsilon^2}{2\lambda^2}\eta^\intercal \mathbb{X}_1 \mathbb{X}_2^\intercal \mathbb{X}_2 \mathbb{X}_1^\intercal \eta)}_{\approx 0} \\ \\ \frac{\epsilon}{\lambda}\mathbb{X}_2\mathbb{X}_1^\intercal\eta \end{pmatrix} = \begin{pmatrix} \eta \\ \frac{\epsilon}{\lambda}\mathbb{X}_2\mathbb{X}_1^\intercal\eta \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{\epsilon}{\lambda}\mathbb{X}_2\mathbb{X}_1^\intercal \end{pmatrix} \otimes \eta.$$

Defining $\gamma = \frac{1}{\lambda}\mathbb{X}_2\mathbb{X}_1^\intercal\eta$, the eigenvector can be rewritten as

$$u_\epsilon = \begin{pmatrix} \eta^{(L-1)} \\ \eta_L \\ \epsilon\gamma^{(L-1)} \\ \epsilon\gamma_L \end{pmatrix},$$

where $\eta^{(L-1)}$ is the first $(L-1)$ elements, η_L is L^{th} element, $\epsilon\gamma^{(L-1)}$ contain elements from $(L+1)$ to $(2L-1)$, and $\epsilon\gamma_L$ is the $2L^{th}$ element.

Thus, the approximation of the vectors R from (4.5), omitting ϵ^2 and higher order, is as following

$$\begin{aligned} R_{11}(\epsilon) &= \frac{1}{\Delta_\epsilon}\sum_k \eta_{k,L}\eta_k^{(L-1)} = R_{SSA}, \\ R_{12}(\epsilon) &= \frac{1}{\Delta_\epsilon}\epsilon\sum_k \eta_{k,L}\gamma_k^{(L-1)}, \\ R_{21}(\epsilon) &= \frac{1}{\Delta_\epsilon}\epsilon\sum_k \left(\gamma_{k,L}(1 - \sum_l \eta_{l,L}^2) + \eta_{k,L}(\sum_l \eta_{l,L}\gamma_{l,L})\right)\eta_k^{(L-1)}, \\ R_{22}(\epsilon) &= \frac{1}{\Delta_\epsilon}\epsilon^2\sum_k \left(\gamma_{k,L}(1 - \sum_l \eta_{l,L}^2) + \eta_{k,L}(\sum_l \eta_{l,L}\gamma_{l,L})\right)\gamma_k^{(L-1)}, \end{aligned} \tag{4.21}$$

where Δ_ϵ is

$$\Delta_\epsilon = \det A(\epsilon) = (1 - \sum_k \eta_{k,L}^2)(1 - \sum_k \epsilon^2\gamma_{k,L}^2) - \epsilon^2(\sum_k \eta_{k,L}\gamma_{k,L}).$$

□

If $\epsilon \to 0$ then to first order we can omit ϵ^2 and higher terms and obtain

$$\det A(\epsilon) = 1 - \sum_k \eta_{k,L}^2 = \det A_{(SSA)}.$$

The fact that $R_{11}(\epsilon) = R_{SSA}$ can be explained by the small impact of the support series in obtaining the first linear recurrence vector $R_{11}(\epsilon)$.

By analyzing the recurrence vectors (4.21), one can notice that for the linearized MSSA the support series Y enters the recurrence vectors *linearly*, i.e. only via the eigenvector $\gamma = \frac{1}{\lambda}\mathbb{X}_2\mathbb{X}_1\eta$. Thus, the linearized MSSA requires SSA procedure being applied to the main series X only (in other words, SVD of the matrix $\mathbb{X}_1$) and the constructing of the Hankel matrix of the support series $\mathbb{X}_2$.

Remark *Series reconstruction.* It is worth mentioning that the matrix $\mathbb{X}_2\mathbb{X}_2^T$ is omitted from (4.9). Hence, the reconstruction of the support series cannot be obtained using the linearized MSSA procedure. The linearized MSSA can be treated as one-sided analysis of the main time series and the effect of the support series on it. But we can not fully analyze the support series using this approach. However, we are interested only in measuring the causality effect of the support series on the main one, not the other way around.

The example below illustrates the approximation of the forecast, obtained by the linearized MSSA LRF in comparison with the MSSA forecast.

Suppose we have two time series:

$$\begin{cases} x_n = \sin(\pi n\omega) + \sin(3\pi n\omega) + \sigma\varepsilon_n \\ y_n = \epsilon\sin(\pi n\omega) \end{cases} \tag{4.22}$$

where $n \in \{1, ..., N\}$, $N = 200$ and $\varepsilon_n \sim N(0, 1)$.

Figure 4.1 illustrates linearized MSSA and full MSSA against real signal; the green line represents x_{N+m}, where $m \in \{1, ..., M\}$ time series. One can see that visually it looks like the linearized version of MSSA and is very close to the original MSSA procedure.

For illustration purposes we take $M = 200$. Figure 4.1 shows the performance of linearized MSSA along with standard MSSA for different noise level, $\sigma = 0, 0.5, 1$ in main series, and different perturbation term ϵ. For these particular example linearized MSSA gives a good approximation of standard MSSA.

4.3 Constructing a causality measure

Recalling the LRF for the main series F (1.56), in its vector form (4.3) is

$$\hat{f}_{N+m} = R_{11}^{\intercal}(\epsilon)\widetilde{f}_{N-L+1,N} + R_{12}^{\intercal}(\epsilon)\widetilde{g}_{N-L+1,N}, \tag{4.23}$$

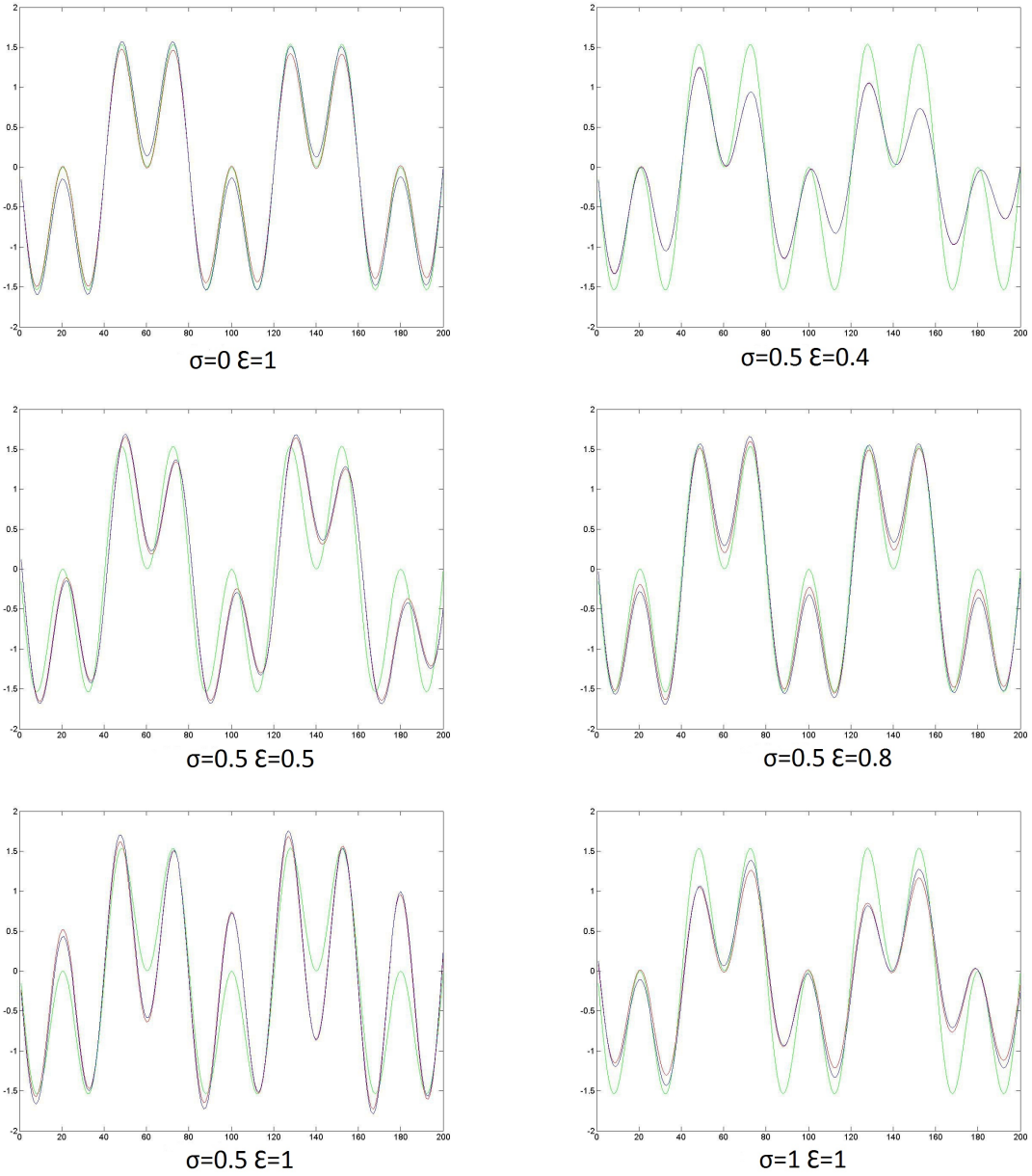
where $\widetilde{f}_{N-L+1,N}$ and $\widetilde{g}_{N-L+1,N}$ are the last $(L-1)$ terms of the known reconstructed main series F and support series G, respectively.

Intuitively, the support series "cause" the main series if it has a substantial "weight" in the LRF, in other words, if the recurrence vector $R_{12}(\epsilon)$ is appreciably larger than zero.

There are several ways to construct such measure. One can compare recurrence vectors directly by taking the normalized ratio of their norms. The normalization is needed in order to avoid the scaling problems (as exemplified when the two time series are measured in different units). The first ratio $M_{test,1}$ is

$$M_{test,1} = \sqrt{\frac{\| R_{12}(\epsilon) \|^2}{\| R_{11}(\epsilon) \|^2}}\sqrt{\frac{\sum_{k=1}^{r}\lambda_{1k}}{\sum_{k=1}^{r}\lambda_{2k}}}, \tag{4.24}$$

Fig. 4.1: Part of the original series (blue), MSSA (red), linearized MSSA (black)



where λ_{1k} are eigenvalues of $\mathbb{X}_1\mathbb{X}_1^T$, λ_{2k} are eigenvalues of $\epsilon\mathbb{X}_2\mathbb{X}_2^T$, and r corresponds to the choice of principal components for reconstruction, and $\mathbb{X}_1$, $\epsilon\mathbb{X}_2$ are Hankel matrices of the main series F and the support series G, respectively.

However, the above measure does not fully meet the Granger causality definition, which implies that two identical time series are not causal. To see how much in the $M_{test,1}$ measure is due to identity, we need to construct another measure.
Here we are calculating the measure by comparing the norm of the orthogonal complement of $R_{12}(\epsilon)$ - $\widetilde{R}_{12}(\epsilon)$

$$\| \widetilde{R}_{12}(\epsilon) \| = \left\| R_{12}(\epsilon) - \frac{(R_{11}^{\intercal}(\epsilon)R_{12}(\epsilon))R_{11}(\epsilon)}{\| R_{11}(\epsilon) \|^2} \right\| \tag{4.25}$$

with the norm of the recurrence vector R_{11} by looking at their ratio

$$M_{test,2} = \frac{\| \widetilde{R}_{12}(\epsilon) \|}{\| R_{11}(\epsilon) \|} = \sqrt{\frac{\| R_{12}(\epsilon) \|^2}{\| R_{11}(\epsilon) \|^2} - \frac{(R_{12}^T(\epsilon)R_{11})^2(\epsilon)}{\| R_{11}(\epsilon) \|^4}}, \tag{4.26}$$

$$\sqrt{\frac{\|R_{12}(\epsilon)\|^2}{\| R_{11}(\epsilon) \|^2} - \frac{(R_{12}^T(\epsilon)R_{11})^2(\epsilon)}{\| R_{11}(\epsilon) \|^4}} \sqrt{\frac{\sum_{k=1}^{r} \lambda_{1k}}{\sum_{k=1}^{r} \lambda_{2k}}}. \tag{4.27}$$

Taking the ratio of the orthogonal compliment of the recurrence vector $R_{12}(\epsilon)$ ($\widetilde{R}_{12}(\epsilon)$) and the recurrence vector $R_{11}(\epsilon)$ allows not only to compare the sizes of the two vectors, but their directions as well. Vectors themselves can be different, but they could have same or very similar direction. If we take $M_{test,1}$ as a measure of causality, we are not able to distinguish if the causality is due mostly to the identity of the time series or actual causality, which is not the same thing according to Granger (see [16]).

Following Granger's definition of causality, two identical time series are not causal. Therefore, one of the measures, $M_{test,2}$, is constructed to exclude the case of identity.

Assume a trivial example, when we have two identical time series F and its duplicate with the Hankel matrix of F with window length L being $\mathbb{X}$. Then the eigenvalues and the eigenvectors of the matrix

$$S = \begin{pmatrix} \mathbb{X} & \mathbb{X} \end{pmatrix} \begin{pmatrix} \mathbb{X} \\ \mathbb{X} \end{pmatrix} = \begin{pmatrix} \mathbb{X}\mathbb{X}^{\intercal} & \mathbb{X}\mathbb{X}^{\intercal} \\ \mathbb{X}\mathbb{X}^{\intercal} & \mathbb{X}\mathbb{X}^{\intercal} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \otimes \mathbb{X}\mathbb{X}^{\intercal} \tag{4.28}$$

are $0\mu_1, ..., 0\mu_L, 2\mu_1, ..., 2\mu_L$, where $\mu_1, ..., \mu_L$ are the eigenvalues of $XX^{\intercal}$. Eigenvectors, corresponding to non-zero eigenvalues are $\begin{pmatrix} V \\ V \end{pmatrix}$, where $V = \begin{pmatrix} V^{\triangledown} \\ V_L \end{pmatrix}$ is $L \times 1$ size and $V^{\triangledown}$ is the vector of the first $(L-1)$ elements of the vector and V_L is the L^{th} element of vector V. Substituting these eigenvectors into the MSSA recurrence vectors (4.5), we get

$$R_{11} = \frac{1}{\Delta} \left((1 - \sum_k V_{k,L}^2) \sum_k V_{k,L} V_k^{\triangledown} + (\sum_k V_{k,L}) \sum_k V_k^{\triangledown} V_{k,L}) \right) = R_{12}. \tag{4.29}$$

By substituting the recurrence vectors into (4.24) and (4.26) we get $M_{test,1} = 1$ and $M_{test,2} = 0$. $M_{test,1}$ indicates that two time series have similarities, and $M_{test,2}$ shows that the similarity in this case is all due to the identity of two time series.

In this chapter we discussed difference between the autoregressive model and a suggested analogous SSA-based model, which leads to a fundamental difference between Granger causality concept and the one described here. Granger's idea of measuring causality is fully based on the fact that the underlying model is stationary noise model and the SSA based model is a deterministic linear recurrence with added arbitrary noise. Specifically, the residuals in the SSA model do not carry any information on the model parameters, as it does in AR models. Noise in the SSA approach is treated as irrelevant part and does not play a crucial role in the analysis, as it is in autoregressive models.

However, to construct a measure for causality for the SSA/MSSA model, we first need to deal with non-linearity of the MSSA approach. We introduce the so-called linearized MSSA procedure, which corresponds to the linear approximation of the standard MSSA to the first order. The example given in this chapter illustrates the goodness of linearized MSSA approximation.

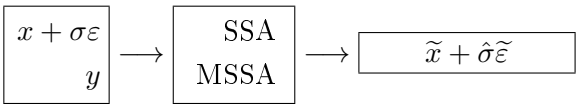
We also introduced tentative causality measures, which are based on the comparison of recurrence vectors $R_{11}(\epsilon)$ and $R_{12}(\epsilon)$ to be used along with each other to satisfy the definition of causality introduced by Granger. The first causality measure $M_{test,1}$ is a normalized ratio of the recurrence vectors norms, which allows to see if the recurrence vector $R_{12}(\epsilon)$ is substantial in the forecast LRF (4.23). The second measure $M_{test,2}$ is a ratio of the orthogonal complement of the recurrence vector $R_{12}(\epsilon)$ and the recurrence vector $R_{11}(\epsilon)$. This measure allows us to compare recurrence vectors not only by size, but also by the directions as well.

5. STABILITY ANALYSIS

In this chapter we focus on the relationship between initial time series that we analyze through the SSA mechanism and the resultant output of this analysis, i.e. the time series reconstruction and its forecast. This relationship is closely connected to the F-test. Most common causality tests, used for stationary autoregressive model, are also related to the F-test, as it was discussed in section 2.3. We used F-test as a tool for detecting causality in SSA/MSSA model, see section 3.3.

F-test compares the resultant variance in the forecast of the time series, obtained with SSA/MSSA LRFs. These variances, in turn, arise from randomness in the initial time series. Hence, we can try to find a connection between the input and output noise by considering how the forecast of SSA/MSSA changes as we add a random perturbation to the initial time series, which is the key idea of the stability analysis in this chapter.

If we represent the SSA procedure more visually, we have the following picture



Noise in the support series is omitted for the purposes of studying the effect of the noise in the main series

$$\sigma_\varepsilon \quad \overset{?}{\longleftrightarrow} \quad \hat{\sigma}_{\widetilde{\varepsilon}}.$$

In this chapter we study how exactly SSA mechanism works and how the perturbation of the initial time series reflects on the error of the obtained reconstruction and further of the forecast as well.

We begin the chapter from constructing the linearization of the univariate SSA procedure. This construction is similar to the construction of linearized MSSA, although with SSA we consider perturbation in the main series itself. Here we are also considering perturbation of the main series to the first order to see if it is possible to derive an expression, which connects the perturbation of the series with the error in the reconstruction and forecast.

We come to a better understanding of noise propagation by studying the effect of the perturbation at three stages. These stages are: projector construction, time series reconstruction and forecast. Firstly, we deal with the noise propagation at the stage of constructing the projector, obtained from perturbed eigenvectors components. Second stage is reconstruction, where the noise comes through the Hankel matrix and perturbed eigenvectors. And finally, the forecast, where the noise comes in through the recurrence

vectors and through reconstruction. In general, at each stage noise comes through the Hankel matrix and/or through the obtained eigenvectors. On each stage we assess the size of the noise effect and see if any of these effects are dominant, so that less dominant ones could be neglected in comparison. Thus, we find it useful to look at noise propagation stages separately. By applying these effects individually, we can study their reflection on reconstructions, recurrence vectors and forecasts.

We use both simulated data and real data examples to study the effects described above, which give rise to assumptions on stability of the time series reconstruction and obtained recurrence vectors under perturbation effects.

In this chapter we apply convolution representation of univariate and bivariate SSA to derive an expression, which reflects the connection between a random perturbation of the time series and the error in the reconstruction and the forecast. Convolution makes the study of variance possible, as it gives rise to an expression, where the statistical independence of the input noise is preserved.

5.1 Linearization of SSA

Proposition 5.1.1. *Consider an unperturbed time series x_n and its perturbation $\sigma\varepsilon_n$, where $\varepsilon_n \sim N(0,1)$ is i.i.d.. Let $\mathbb{X}$ be the Hankel matrix of unperturbed time series x_n with window length L, and $\mathbb{N}$ the Hankel matrix of the perturbation term ε_n. Let*

$$\mathbb{Z} = \mathbb{X}\mathbb{N}^T + \mathbb{N}\mathbb{X}^T. \tag{5.1}$$

Performing the spectral decomposition of the unperturbed matrix $\mathbb{X}\mathbb{X}^\intercal$, we find its eigenvalues and eigenvectors

$$\mathbb{X}\mathbb{X}^\intercal\eta_k = \lambda\eta_k,$$

where eigenvectors η_k form the orthonormal basis $\eta_1, .., \eta_L$ in $\mathbb{R}^L$ with corresponding eigenvalues $\lambda_1, ..., \lambda_L$.

Let $R = (a_{L-1}, ..., a_1)$ be the recurrence vector, obtained from the unperturbed time series x_n. Then the approximation of the SSA recurrence vector, obtained from the perturbed time series $x_n + \sigma\varepsilon_n$, to the first order, is

$$R(\sigma) = R + \sigma\left(\mathbf{c}R + \widetilde{R}\right) + O(\sigma^2),$$

where

$$\mathbf{c} = \frac{2\sum_{k=1}^{r}\sum_{i=r+1}^{L}\alpha_{i,k}\eta_{i,L}\eta_{k,L}}{1-\sum_{k=1}^{r}\eta_{k,L}^2},$$
$$\widetilde{R} = \frac{\sum_{k=1}^{r}\sum_{i=r+1}^{L}\alpha_{i,k}(\eta_{i,L}\eta_k^{(L-1)} + \eta_{k,L}\eta_i^{(L-1)})}{1-\sum_{k=1}^{r}\eta_{k,L}^2}$$

Proof. For the perturbed matrix

$$(\mathbb{X}+\sigma\mathbb{N})(\mathbb{X}+\sigma\mathbb{N})^T = (\mathbb{X}\mathbb{X}^T+\sigma\mathbb{Z})+O(\sigma^2)$$

there are eigenvectors γ_σ and eigenvalues λ_σ analytical in σ

$$\gamma_{\sigma,k} = \eta_k + \sigma\nu_{1,k} + \dots \tag{5.2}$$

and

$$\lambda_{\sigma,k} = \lambda_k + \sigma\mu_{1,k} + \dots \tag{5.3}$$

such that

$$(\mathbb{X}\mathbb{X}^T+\sigma\mathbb{Z})\gamma_{\sigma,k} = \lambda_{\sigma,k}\gamma_{\sigma,k}. \tag{5.4}$$

Substituting power series (5.2) and (5.3) into (5.4), we get

$$\mathbb{X}\mathbb{X}^T\eta_k + \sigma\mathbb{X}\mathbb{X}^T\nu_{1,k} + \sigma\mathbb{Z}\eta_k + O(\sigma^2) = \lambda_k\eta_k + \sigma(\lambda_k\nu_{1,k}+\mu_{1,k}\eta_k) + O(\sigma^2), \tag{5.5}$$

$$XX^T\eta_k - \lambda_k\eta_k + \sigma\mathbb{X}\mathbb{X}^T\nu_{1,k} + \sigma\mathbb{Z}\eta_k + O(\sigma^2) = \sigma(\lambda_k\nu_{1,k}+\mu_{1,k}\eta_k) + O(\sigma^2). \tag{5.6}$$

Using the equality

$$\mathbb{X}\mathbb{X}^T\eta_k - \lambda_k\eta_k = 0 \tag{5.7}$$

dividing both sides of (5.6) by σ, and considering the limit as $\sigma \to 0$, we find that

$$\mathbb{X}\mathbb{X}^T\nu_{1,k} + \mathbb{Z}\eta_k = \lambda_k\nu_{1,k} + \mu_{1,k}\eta_k. \tag{5.8}$$

Multiplying (5.8) by η_k^T from the left gives

$$\eta_k^T(\mathbb{X}\mathbb{X}^T-\lambda_k)\nu_{1,k} = \mu_{1,k}\underbrace{\eta_k^T\eta_k}_{=1} - \eta_k^T\mathbb{Z}\eta_k, \tag{5.9}$$

where

$$\eta_k^T(\mathbb{X}\mathbb{X}^T-\lambda_k) = ((\mathbb{X}\mathbb{X}^T-\lambda_k)^T\eta_k)^\intercal = 0^\intercal = 0.$$

Finally, we get the expression for the 1^{st} order eigenvalue perturbation term $\mu_{1,k}$,

$$\mu_{1,k} = \eta_k^T\mathbb{Z}\eta_k. \tag{5.10}$$

The term $\nu_{1,k}$ can be expressed in terms of basis eigenvectors η_i

$$\nu_{1,k} = \sum_{i=1}^{L}\alpha_{i,k}\eta_i.$$

Substituting $\mu_{1,k}$, $\nu_{1,k}$ into (5.8), we get

$$(\mathbb{X}\mathbb{X}^T - \lambda_k)\sum_{i=1}^{L}\alpha_{i,k}\eta_i = (\mu_{1,k} - \mathbb{Z})\eta_k, \tag{5.11}$$

knowing that

$$(\mathbb{X}\mathbb{X}^T - \lambda_k)\alpha_{i,k}\eta_i = \alpha_{i,k}(\mathbb{X}\mathbb{X}^T\eta_i - \lambda_k\eta_i) \tag{5.12}$$

since (5.7)

$$\alpha_{i,k}(\mathbb{X}\mathbb{X}^T\eta_i - \lambda_k\eta_i) = 0, \text{ if } i = k.$$

Therefore, $\alpha_{k,k}$ is arbitrary, but for simplicity we choose $\alpha_{k,k} = 0$. Substituting (5.12) into (5.11) and multiplying both sides of (5.11) by η_j^T from the left, we get

$$\sum_{i=1}^{L}\alpha_{i,k}\eta_j^T(\lambda_i - \lambda_k)\eta_i = \eta_j^T(\mu_{1,k} - \mathbb{Z})\eta_k. \tag{5.13}$$

Using the property of orthonormal basis vectors, we know that

$$< \eta_j^T, \eta_i >= \begin{cases} 0, \text{ if } i \neq j \\ 1, \text{ if } i = j. \end{cases}$$

Therefore, we can find $\alpha_{j,k}$ coefficients from (5.13):

$$\alpha_{j,k}(\lambda_j - \lambda_k) = \eta_j^T(\mu_{1,k} - \mathbb{Z})\eta_k,$$
$$\alpha_{j,k} = \frac{\eta_j^T(\mu_{1,k} - \mathbb{Z})\eta_k}{\lambda_j - \lambda_k} = \frac{\mu_{1,k}\eta_j^T\eta_k - \eta_j^T\mathbb{Z}\eta_k}{\lambda_j - \lambda_k}.$$

Hence, we get the final expression for $\alpha_{j,k}$

$$\alpha_{j,k} = -\frac{\eta_j^T\mathbb{Z}\eta_k}{\lambda_j - \lambda_k} = -\alpha_{k,j} \tag{5.14}$$

since $\mathbb{Z}$ is symmetric, we have the antisymmetry in $\alpha_{j,k}$ coefficients. The expression for the eigenvector of the perturbed system to first order is

$$\gamma_{k,\sigma} = \eta_k + \sigma\sum_{i=1}^{L}\alpha_{i,k}\eta_i = \eta_k + \sigma(\underbrace{\alpha_{k,k}\eta_k}_{=0} + \sum_{i=1,i\neq k}^{L}\alpha_{i,k}\eta_i). \tag{5.15}$$

The recurrence vector for SSA (1.42) is

$$R(\sigma) = \frac{\sum_{k=1}^{r}\gamma_{k,\sigma,L}\gamma_{k,\sigma}^{(L-1)}}{1 - \sum_{k=1}^{r}\gamma_{k,\sigma,L}^2}. \tag{5.16}$$

Substituting $\gamma_{k,\sigma}$, we get

$$R(\sigma) = \frac{\sum_{k=1}^{r}\left(\eta_{k,L}\eta_k^{L-1} + \sigma\sum_{i=1,i\neq k}^{L}\alpha_{i,k}(\eta_{i,L}\eta_k^{(L-1)} + \eta_{k,L}\eta_i^{(L-1)})\right) + O(\sigma^2)}{1 - \sum_{k=1}^{r}(\eta_{k,L}^2 + 2\sigma\sum_{i=1,i\neq k}^{r}\alpha_{i,k}\eta_{i,L}\eta_{k,L}) + O(\sigma^2)}. \tag{5.17}$$

Using the antisymmetry of $\mathbb{Z}$ (5.1) and coefficients $\alpha_{j,k}$ (5.14) we can simplify the expression for the vector R_σ (5.17).
Due to antisymmetry, we have

$$\alpha_{i,k}(\eta_{i,L}\eta_k^{(L-1)} + \eta_{k,L}\eta_i^{(L-1)}) + \underbrace{\alpha_{k,i}}_{=-\alpha_{i,k}}(\eta_{k,L}\eta_i^{(L-1)} + \eta_{i,L}\eta_k^{(L-1)}) = 0 \tag{5.18}$$

and the same is for

$$\alpha_{i,k}\eta_{i,L}\eta_k + \alpha_{k,i}\eta_{k,L}\eta_i = \alpha_{i,k}\eta_{i,L}\eta_k - \alpha_{i,k}\eta_{k,L}\eta_i = 0. \tag{5.19}$$

Therefore, substituting (5.18) and (5.19) into (5.17), we get slightly simpler expression for $R(\sigma)$

$$R(\sigma) = \frac{\sum_{k=1}^{r}\eta_{k,L}\eta_k^{(L-1)} + \sigma(\sum_{k=1}^{r}\sum_{i=r+1}^{L}\alpha_{i,k}(\eta_{i,L}\eta_k^{(L-1)} + \eta_{k,L}\eta_i^{(L-1)})) + O(\sigma^2)}{1 - \sum_{k=1}^{r}\eta_{k,L}^2 - 2\sigma\sum_{k=1}^{r}\sum_{i=r+1}^{L}\alpha_{i,k}\eta_{i,L}\eta_{k,L} + O(\sigma^2)}. \tag{5.20}$$

Looking at the denominator of the recurrence vector $R(\sigma)$

$$\frac{1}{1 - \sum_{k=1}^{r}\eta_{k,L}^2 - 2\sigma\sum_{k=1}^{r}\sum_{i=r+1}^{L}\alpha_{i,k}\eta_{i,L}\eta_{k,L} + O(\sigma^2)}$$

$$= \frac{1}{1 - \sum_{k=1}^{r}\eta_{k,L}^2} \quad \frac{1}{1 - \frac{2\sigma\sum_{k=1}^{r}\sum_{i=r+1}^{L}\alpha_{i,k}\eta_{i,L}\eta_{k,L}}{1-\sum_{k=1}^{r}\eta_{k,L}^2} + O(\sigma^2)}$$

$$= \frac{1}{1 - \sum_{k=1}^{r}\eta_{k,L}^2}\left(1 + \frac{2\sigma\sum_{k=1}^{r}\sum_{i=r+1}^{L}\alpha_{i,k}\eta_{i,L}\eta_{k,L}}{1 - \sum_{k=1}^{r}\eta_{k,L}^2} + O(\sigma^2)\right)$$

and substituting this expression instead of denominator, we get

$$\boxed{R(\sigma) = R + \sigma\left(\mathbf{c}R + \widetilde{R}\right) + O(\sigma^2)}, \tag{5.21}$$

where

$$\mathbf{c} = \frac{2\sum_{k=1}^{r}\sum_{i=r+1}^{L}\alpha_{i,k}\eta_{i,L}\eta_{k,L}}{1 - \sum_{k=1}^{r}\eta_{k,L}^2} \tag{5.22}$$

and

$$\widetilde{R} = \frac{\sum_{k=1}^{r}\sum_{i=r+1}^{L}\alpha_{i,k}(\eta_{i,L}\eta_k^{(L-1)} + \eta_{k,L}\eta_i^{(L-1)})}{1-\sum_{k=1}^{r}\eta_{k,L}^2}. \tag{5.23}$$

Thus, $R = (a_{L-1}, ..., a_1)$ and $\widetilde{R} = (b_{L-1}, ..., b_1)$, where a_i and b_i are the coefficients of the forecast LRF.
The approximation of the linear recurrence vector $R(\sigma)$ is now fully linear with respect to the parameter σ. □

If the approximation is good, one would expect that

$$\frac{R_\sigma - R}{\sigma} - (\mathbf{c}R + \widetilde{R}) \sim O(\sigma), \tag{5.24}$$

where R_σ is the standard SSA linear recurrence vector obtained for the perturbed time series $x_n + \sigma\varepsilon_n$ and $(\mathbf{c} + \widetilde{R})$ is the correction term of the linearization of SSA recurrence vector.

Suppose we have an unperturbed time series

$$x_n = sin(\pi n\omega), \tag{5.25}$$

where $n \in \{1, ..., 200\}$, $\omega = 1/22$, and perturbation $\sigma\varepsilon$, where $\sigma \in \{0.1 \cdot 10^{-8}$, $0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$. The curves plotted in Figure 5.1 correspond to the difference (5.38) for given σ. The smaller is the variance, the narrower the curve.

For a very small variance ($\sigma \leq 0.01$) the difference (5.38) does not change much and seems to converge. However, it does not converge to zero as expected. Nevertheless, the convergence limit is of order 10^{-4}, see Figure 5.1, and is small enough to consider the approximation to be good.

Knowing the eigenvalues and eigenvectors, it is possible to find the reconstruction of the initial perturbed time series $x_n + \sigma\varepsilon_n$.

The reconstruction of unperturbed time series comes from the sum of the first r elementary matrices of the decomposition of the Hankel matrix $\mathbb{X}$

$$\mathbf{X} = \sum_{k=1}^{r}\mathbf{X}_k(0), \tag{5.26}$$

where the elementary matrix is

$$\mathbf{X}_k(0) = \eta_k\eta_k^T\mathbb{X} \tag{5.27}$$

To get the reconstruction for the perturbed time series $x_n(\sigma)$ using the linearized SSA approach, the elementary matrices are calculated now using the perturbed eigenvectors

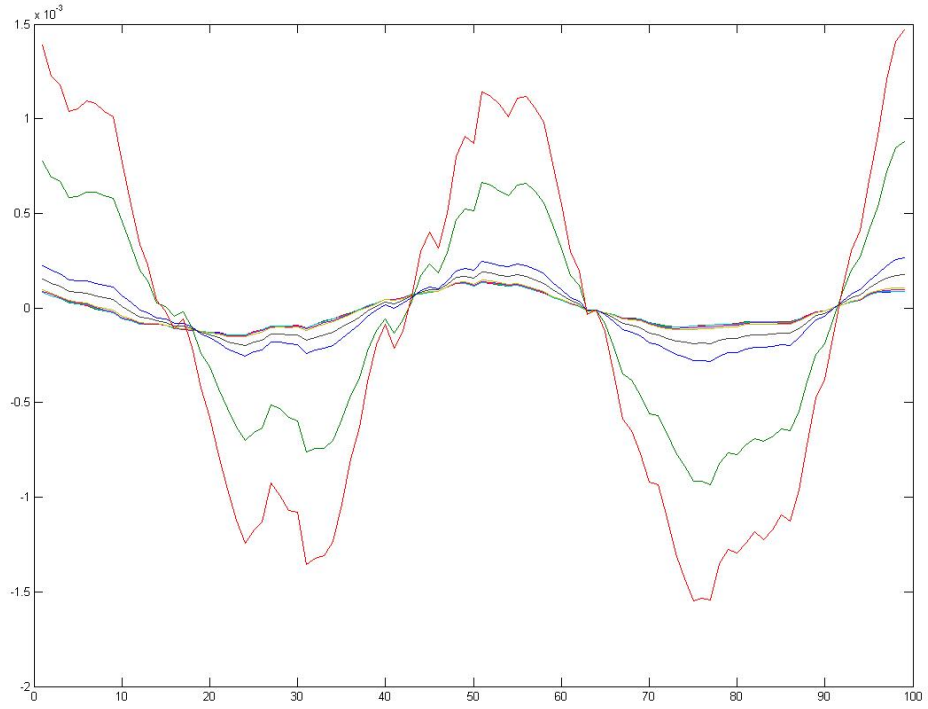


Fig. 5.1: Plot shows the remainder of linearized vector (5.24) for (5.25), $L = 200$, $r = 2$; $\sigma = 0.1$ (red), $\sigma = 0.05$ (green), $\sigma = 0.01$ (blue), $\sigma = 0.005$ (black), $\sigma \leq 0.001$ the rest

$\gamma_{\sigma,i}$

$$\begin{aligned}\mathbf{X}_k(\sigma) &= \gamma_{\sigma,k}\gamma_{\sigma,k}^T\mathbb{X} \\ &= (\eta_k + \sigma\sum_{i=1,i\neq k}^{L}\alpha_{i,k}\eta_i + O(\sigma^2))(\eta_k + \sigma\sum_{i=1,i\neq k}^{L}\alpha_{i,k}\eta_i + O(\sigma^2))^T\mathbb{X} \\ &= \eta_k\eta_k^T X + \sigma\sum_{i=1,i\neq k}^{L}\alpha_{i,k}(\eta_i\eta_k^T + \eta_k\eta_i^T)\mathbb{X} + O(\sigma^2) \\ &= \mathbf{X}_k(0) + \sigma\sum_{i=1,i\neq k}^{L}\alpha_{i,k}(\eta_i\eta_k^T + \eta_k\eta_i^T)\mathbb{X} + O(\sigma^2)\end{aligned}$$

The second term corresponds to the first order change in the reconstruction of the series due to its perturbation.

Grouping elementary matrices $\mathbf{X}_k(\sigma)$, corresponding to first r principal components, we get the resultant matrix

$$\mathbf{X}(\sigma) = \sum_{i=1}^{r}\mathbf{X}_i(\sigma).$$

Finally, after applying the diagonal averaging to the matrix $\mathbf{X}(\sigma)$, we get the reconstructed time series $\widetilde{x}_n$. To calculate the forecast we substitute the vector (5.21) into the linear recurrence formula (1.41) and use the reconstruction series $\widetilde{x}_n$

$$\widetilde{x}_i(\sigma) = \widetilde{x}_i(0) + \sigma\widetilde{\varepsilon}_i + O(\sigma^2)$$

The forecast for the $N+1$ point is calculated using the LRF

$$\widetilde{x}_{N+1}(\sigma) = \sum_{k=1}^{L-1}\big(a_k + \sigma(\mathbf{c}a_k + b_k) + O(\sigma^2))(\widetilde{x}_{N-k+1}(0) + \sigma\widetilde{\varepsilon}_{N-k+1} + O(\sigma^2)\big).$$

Opening the brackets, we get

$$\begin{aligned}\widetilde{x}_{N+1}(\sigma) &= \sum_{k=1}^{L-1}\Big(a_k\widetilde{x}_{N-k+1}(0) + \sigma\big((\mathbf{c}a_k + b_k)\widetilde{x}_{N-k+1}(0) + a_k\widetilde{\varepsilon}_{N-k+1}\big)\Big) + O(\sigma^2) \\ &= \sum_{k=1}^{L-1}a_k\widetilde{x}_{N-k+1}(0) + \sigma\sum_{k=1}^{L-1}\big((\mathbf{c}a_k + b_k)\widetilde{x}_{N-k+1}(0) + a_k\widetilde{\varepsilon}_{N-k+1}\big) + O(\sigma^2).\end{aligned} \tag{5.28}$$

Considering the forecast of the $N+1$ point of an unperturbed time series to be $\widetilde{x}_{N+1}(0)$, (5.28) can be expressed as a sum of the unperturbed forecast and the correction term

$$\widetilde{x}_{N+1}(\sigma) = \widetilde{x}_{N+1}(0) + \sigma\sum_{k=1}^{L-1}\big((\mathbf{c}a_k + b_k)\widetilde{x}_{N-k+1}(0) + a_k\widetilde{\varepsilon}_{N-k+1}\big) + O(\sigma^2). \tag{5.29}$$

Analyzing the recurrence vector (5.21), one can see that the noise appears only in $\alpha_{j,k}$ coefficients. Recalling (5.14), one can investigate how the noise (considered to be white noise) and its variance affect the recurrence vector. Firstly we look at the numerator of the coefficient $\alpha_{j,k}$

$$\eta_I^T \mathbb{X}\mathbb{N}^T \eta_{II} = \sum_{k=1}^{L} \eta_{II,k} \sum_{i=1}^{L} \eta_{I,i} (\mathbb{X}\mathbb{N}^T)_{ik}. \tag{5.30}$$

Recalling that $\mathbb{X}$ and $\mathbb{N}$ are Hankel matrices, we can write each element of these matrices as $\mathbb{X}_{ij} = x_{i+j-1}$ and $\mathbb{N}_{jk} = n_{j+k-1}$, respectively. Thus, we can rewrite (5.30) as

$$\eta_I^T \mathbb{X}\mathbb{N}^T \eta_{II} = \sum_{k=1}^{L} \eta_{II,k} \sum_{i=1}^{L} \eta_{I,i} \sum_{j=1}^{K} x_{i+j-1} n_{j+k-1}. \tag{5.31}$$

Let $m = j + k - 1$, then (5.31) becomes

$$\sum_{k=1}^{L} \eta_{II,k} \sum_{i=1}^{L} \eta_{I,i} \sum_{m=k}^{K+k-1} x_{i+m-k} n_m \tag{5.32}$$

The aim of these manipulations is to see how the noise affects the future forecast of the series. However, noise comes into the forecast not only through reconstruction of the time series, but through the recurrence vector as well, specifically through $\alpha_{j,k}$ coefficients. We know that the noise satisfies properties of white noise. If it is possible to rewrite the numerator as a sum of n_m with some coefficients, we can treat it as i.i.d. random variable and therefore calculate its variance.

Now we can rewrite (5.32) as

$$\begin{aligned} \sum_{m=1}^{L} \sum_{k=1}^{m} n_m \sum_{i=1}^{L} \eta_{II,k} \eta_{I,i} x_{i+m-k} + \sum_{m=L+1}^{K-1} \sum_{k=1}^{L} n_m \sum_{i=1}^{L} \eta_{II,k} \eta_{I,i} x_{i+m-k} \\ + \sum_{m=K}^{K+L-1} \sum_{k=m-K+1}^{L} n_m \sum_{i=1}^{L} \eta_{II,k} \eta_{I,i} x_{i+m-k}. \end{aligned} \tag{5.33}$$

Since $n_m \sim N(0, \sigma^2)$ and i.i.d. we can find the variance of $\eta_I^T \mathbb{X}\mathbb{N}^T \eta_{II}$.

In principle, it is possible to derive the expression for the variance of noise $\widetilde{\varepsilon}_n$, but in the obtained sum (5.33) is derived so that the statistical independence of the input noise n_m is obscured, which makes it difficult to derive the expression for variance of the forecast. If we try to plug in the coefficients $\alpha_{j,k}$ with their rewritten numerator expressions into (5.28), recalling that they enter the expression through $\mathbf{c}$ (5.22), we get even more complicated formula.

This expression illustrates the difficulties one meets while working out the relationship between input and output noise. Thus, next step is to consider each part of the SSA procedure that perturbation affects and see if it is possible to separate these effects.

5.2 Reconstruction stability.

From the forecast LRF (1.41), (1.45), (1.58), (1.59) and the way they were derived one can see that the input noise is coming into the forecast not only through the time series reconstruction, but also through the recurrence vectors as well. The input-output noise relation is non linear due to the time series reconstruction process and complicates the understanding of this relation.

We begin from looking at the univariate case. Suppose we have an unperturbed time series x_n with Hankel matrix $\mathbb{X}$ and its perturbation $\sigma\varepsilon_n$ with Hankel matrix $\sigma\mathbb{N}$. The matrix $\mathbb{Z}$ is denoted as in (5.1). Recall that eigenvalues and eigenvectors for the perturbed time series are calculated from (5.4).

The perturbation of the time series affects the reconstruction of the time series x_n directly and linearly through the Hankel matrix $\sigma\mathbb{N}$, and indirectly and non-linearly by the noise coming through perturbed eigenvectors γ, i.e. the projection of Hankel matrix $\mathbb{X}+\sigma\mathbb{N}$. To study these effects in the projection step (5.26), (5.27), (1.19) individually, we need to look at them separately.

The reconstruction resulting from the grouping of the elementary matrices

$$\eta_i\eta_i^T\mathbb{X} \tag{5.34}$$

is the reconstruction of the signal $\widetilde{x}^{(1)}$ with no perturbation in the time series or in the eigenvectors η_i, used to build this reconstruction, i.e. this is SSA of the unperturbed time series.

The reconstruction $\widetilde{x}^{(2)}$ resulting from the grouping of the elementary matrices

$$\gamma_i\gamma_i^T(\mathbb{X}+\sigma\mathbb{N}) \tag{5.35}$$

is considered to be the reconstruction of the double perturbation effect, i.e. this is SSA of perturbed time series. Both, $\widetilde{x}^{(1)}$ and $\widetilde{x}^{(2)}$ are standard SSA reconstructions of time series x_n and $x_n+\sigma\varepsilon_n$, respectively.

Two following constructions are not exactly reconstructions. However, they are needed to study the influence of the perturbation of time series.

Thus, next reconstruction $\widetilde{x}^{(3)}$ is based on the effect of the Hankel matrix perturbation only, but using the unperturbed eigenvectors

$$\eta_i\eta_i^T(\mathbb{X}+\sigma\mathbb{N}), \tag{5.36}$$

and the last one, $\widetilde{x}^{(4)}$, uses elementary matrices resulting from perturbed eigenvectors, but applied to the unperturbed time series

$$\gamma_i\gamma_i^T\mathbb{X}. \tag{5.37}$$

Thus we have:

- unperturbed SSA reconstruction,
- perturbed SSA reconstruction,

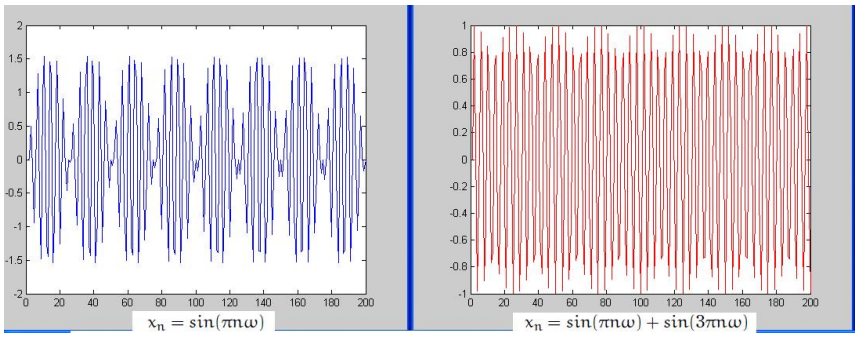


Fig. 5.2: Unperturbed time series (5.39) and (5.40)

- two mixed cases, which help to study the effect of perturbation through eigenvectors and the Hankel matrix.

For comparison purposes we take the difference between the reconstruction of time series with any mentioned above perturbation effect and the one, which does not involve any perturbations

$$\widetilde{x}^{(i)} - \widetilde{x}^{(j)}, \text{ where } i, j \in \{1, 2, 3, 4\}. \tag{5.38}$$

We begin with analyzing simple time series. We take two simple generated time series

$$x_n = sin(\pi n \omega) \tag{5.39}$$

$$x_n = sin(\pi n \omega) + sin(3\pi n \omega) \tag{5.40}$$

Note that both time series were generated with MATLAB using following commands

$$x = sin(linspace(0, 300, 200))' + sin(3 * linspace(0, 300, 200))'$$

$$y = sin(linspace(0, 300, 200))',$$

Figure 5.2 illustrates both unperturbed time series, which were generated using with MATLAB.

Firstly, we take (5.39) and perturb it with i.i.d. noise $\sigma\varepsilon_n \sim N(0, \sigma^2)$.

As in previous tests we choose $n \in (1, ..., 200)$, $\omega = 1.5$. The standard deviation σ is chosen to be 0.5. For this generated example, we have chosen window lengths $L \in \{10, 50\}$ for illustrative purposes and to see if the change in the window length effects the overall picture. The time series itself has a straightforward structure, hence the reconstruction requires only first two eigentriples, $r = 2$.

Figures 5.3-5.6 illustrate the reconstructions differences (5.38) with the parameters, described above, i.e. σ for the times series perturbation, and L and r for analysis.

Secondly, we analyze example (5.40) to see if the results obtained for the (5.39) example hold. The setup for analyzing is the same as above, the only change is that for this example we need first 4 eigentriples for the appropriate reconstruction. The reconstruction differences are illustrated in Figures 5.7-5.10.

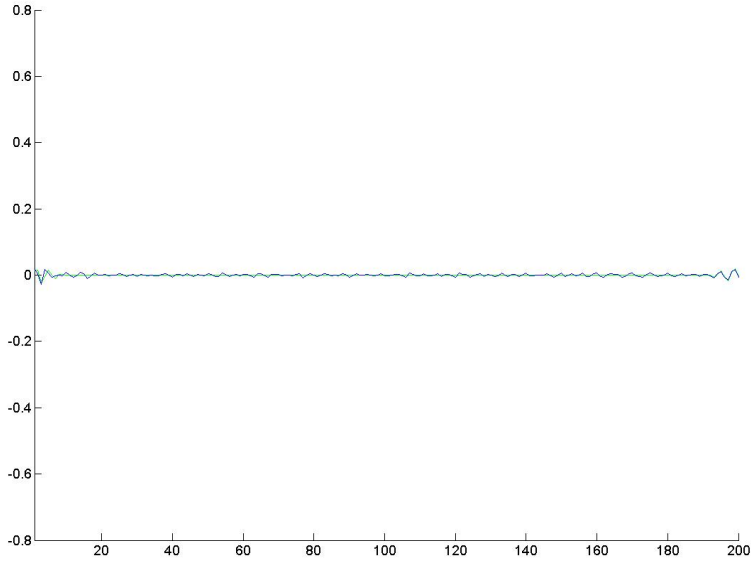


Fig. 5.3: Reconstructions differences for the model (5.39) with perturbation $\sigma\varepsilon$, $\widetilde{x}_n^{(4)} - \widetilde{x}_n^{(1)}$ (green), $\widetilde{x}_n^{(2)} - \widetilde{x}_n^{(3)}$ (blue), L=10

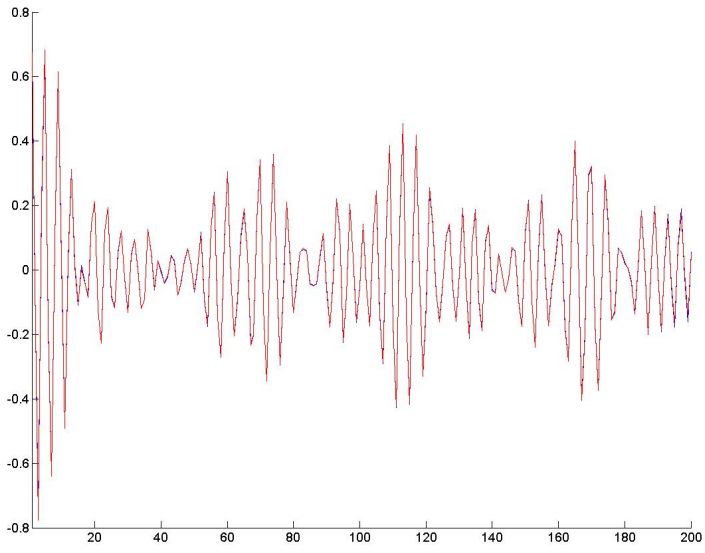


Fig. 5.4: Reconstructions differences for the model (5.39) with perturbation $\sigma\varepsilon$, $\widetilde{x}_n^{(3)} - \widetilde{x}_n^{(1)}$ (blue), $\widetilde{x}_n^{(2)} - \widetilde{x}_n^{(1)}$ (red), L=10

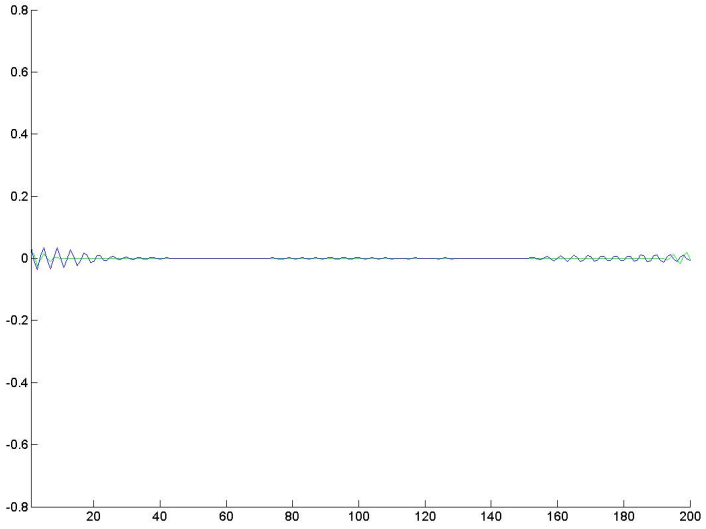


Fig. 5.5: Reconstructions differences for the model (5.39) with perturbation $\sigma\varepsilon$, $\widetilde{x}_n^{(4)} - \widetilde{x}_n^{(1)}$ (green), $\widetilde{x}_n^{(2)} - \widetilde{x}_n^{(3)}$ (blue), L=50

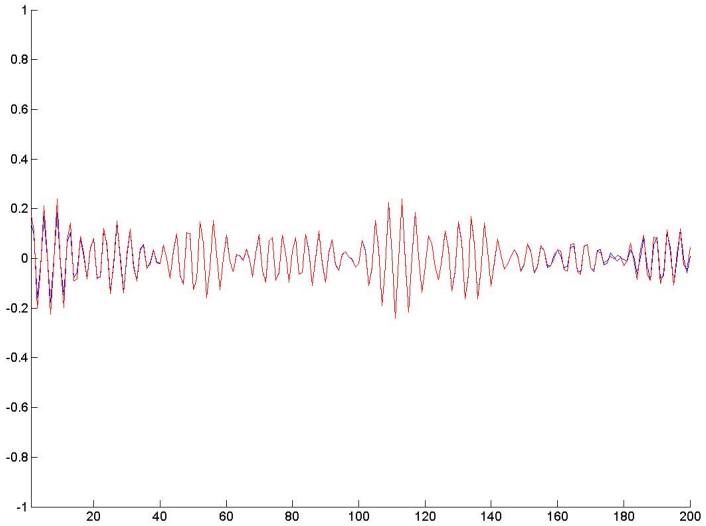


Fig. 5.6: Reconstructions differences for the model (5.39) with perturbation $\sigma\varepsilon$, $\widetilde{x}_n^{(3)} - \widetilde{x}_n^{(1)}$ (blue), $\widetilde{x}_n^{(2)} - \widetilde{x}_n^{(1)}$ (red), L=50

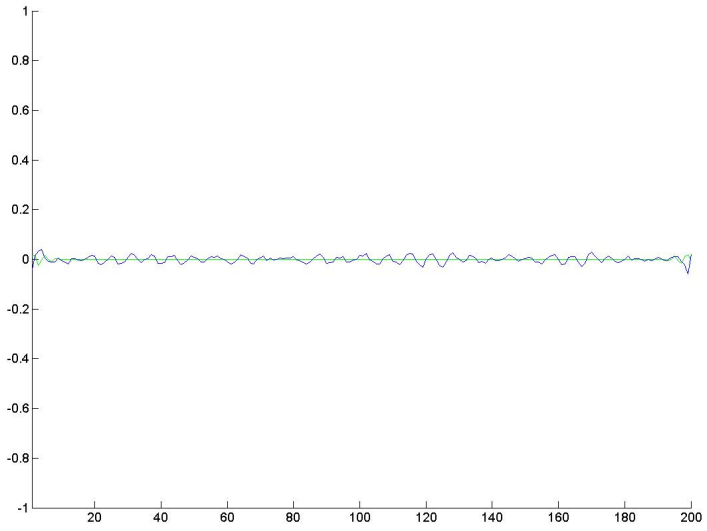


Fig. 5.7: Reconstructions differences for the model (5.40) with perturbation $\sigma\varepsilon$, $\widetilde{x}_n^{(4)} - \widetilde{x}_n^{(1)}$ (green), $\widetilde{x}_n^{(2)} - \widetilde{x}_n^{(3)}$ (blue), L=10

Several observations can be made from Figures 5.3 - 5.10.
The difference

$$\widetilde{x}_n^{(2)} - \widetilde{x}_n^{(3)} \approx 0 \text{ for } n \in \{L+1, ..., N-L\} \tag{5.41}$$

holds through the whole reconstruction, apart from the first and last L interval. Both reconstructions are resulting from the grouping of elementary matrices (5.36) and (5.35), respectively. Therefore, we can write the expression, which is equivalent to (5.41) in terms of the sum of elementary matrices

$$\sum_{i=1}^{r} \gamma_i \gamma_i^T (\mathbb{X} + \sigma\mathbb{N}) - \sum_{i=1}^{r} \eta_i \eta_i^T (\mathbb{X} + \sigma\mathbb{N}) = \sum_{i=1}^{r} (\gamma_i \gamma_i^T - \eta_i \eta_i^T)(\mathbb{X} + \sigma\mathbb{N}) \approx 0, \tag{5.42}$$

which fits in with the observation that the difference in the eigenvectors is small, that it can be neglected in comparison with the perturbation coming through the Hankel matrix $\mathbb{X} + \sigma\mathbb{N}$.

Similar behaviour is observed for the difference

$$\widetilde{x}_n^{(4)} - \widetilde{x}_n^{(1)} \approx 0 \text{ for } n \in \{L+1, ..., N-L\}, \tag{5.43}$$

which is approximately zero through all series, apart from the first and last L interval.

Since (5.41) and (5.43) are true, one would assume that the change in perturbed eigenvectors γ, as was already mentioned, is not crucial and, therefore, the reconstruction can

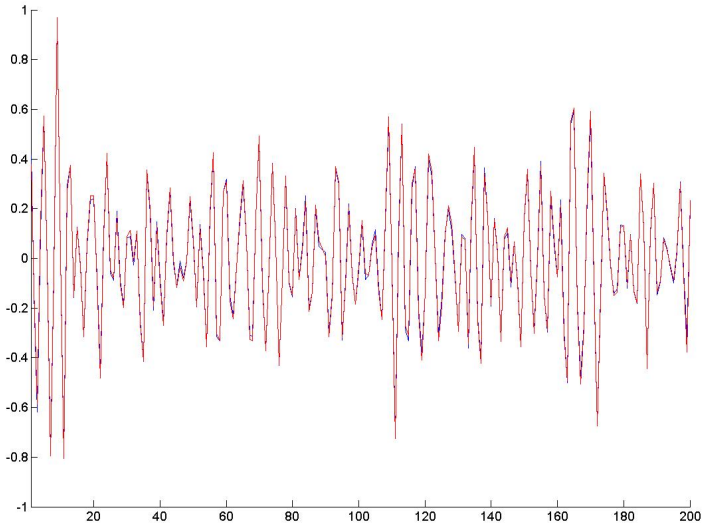


Fig. 5.8: Reconstructions differences for the model (5.40) with perturbation $\sigma\varepsilon$, $\widetilde{x}_n^{(3)} - \widetilde{x}_n^{(1)}$ (blue), $\widetilde{x}_n^{(2)} - \widetilde{x}_n^{(1)}$ (red), L=10

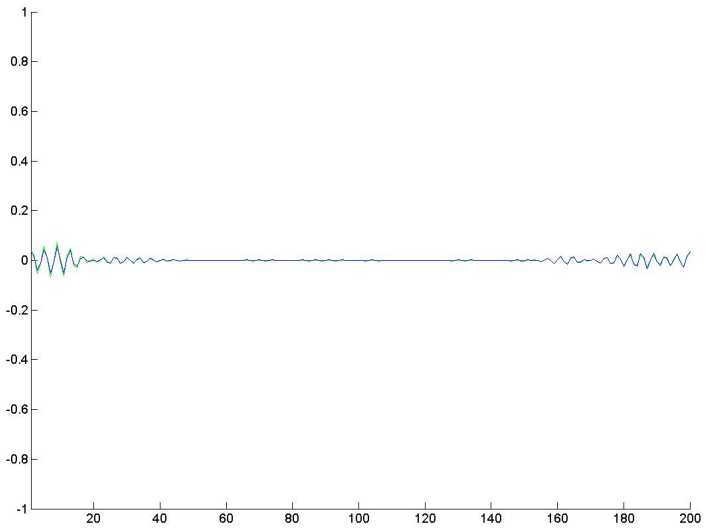


Fig. 5.9: Reconstructions differences for the model (5.40) with perturbation $\sigma\varepsilon$, $\widetilde{x}_n^{(4)} - \widetilde{x}_n^{(1)}$ (green), $\widetilde{x}_n^{(2)} - \widetilde{x}_n^{(3)}$ (blue), L=50

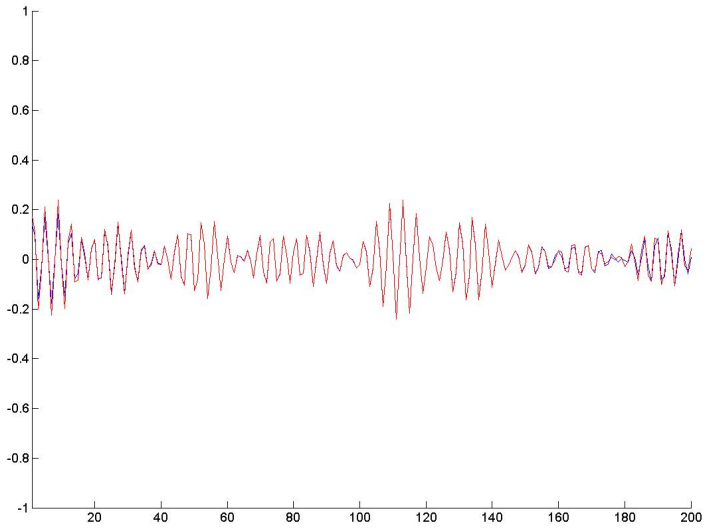


Fig. 5.10: Reconstructions differences for the model (5.40) with perturbation $\sigma\varepsilon$, $\widetilde{x}_n^{(3)} - \widetilde{x}_n^{(1)}$ (blue), $\widetilde{x}_n^{(2)} - \widetilde{x}_n^{(1)}$ (red), L=50

be obtained with unperturbed eigenvectors η with a minor loss in the middle part of the series. However, the minor loss in this case could be explained by the simplicity of the time series structure.

Now we look at two other differences

$$x_n^{(3)} - x_n^{(1)}$$
$$x_n^{(2)} - x_n^{(1)},$$

which correspond to the differences of sums of elementary matrices, respectively

$$\begin{aligned} \sum_{i=1}^{r} \gamma_i\gamma_i^T(\mathbb{X}+\sigma\mathbb{N}) - \sum_{i=1}^{r} \eta_i\eta_i^T\mathbb{X} &= \sum_{i=1}^{r}(\gamma_i\gamma_i^T - \eta_i\eta_i^T)\mathbb{X} + \sigma\gamma_i\gamma_i^T\mathbb{N} \\ \sum_{i=1}^{r} \eta_i\eta_i^T(\mathbb{X}+\sigma\mathbb{N}) - \sum_{i=1}^{r} \eta_i\eta_i^T\mathbb{X} &= \sigma\sum_{i=1}^{r} \eta_i\eta_i^T\mathbb{N}. \end{aligned} \tag{5.44}$$

Figures 5.4, 5.6, 5.8, 5.10 illustrate that these differences are not zero and in fact are sufficiently larger than the difference in (5.43). The differences in terms of elementary matrices (5.44) point out that the noise in the Hankel matrix $\mathbb{N}$ has a large impact on the reconstruction of the time series and hence the effect of this perturbation has more weight in comparison with the perturbation coming through the eigenvectors.

Both time series above were generated and have simple structures. To justify the omission of the perturbation in the eigenvectors, we look at the real data examples in section 5.5.

So far, we assume that it is possible to neglect the change in eigenvectors due to perturbation comparing to the change due to Hankel matrix $\sigma\mathbb{N}$. The purpose of this action is to simplify the expression (5.28) by neglecting unessential changes due to perturbation, so that one could focus on more substantial effect, caused by noise.

Remark *Bivariate SSA reconstruction stability.* For simplicity, suppose we have main series x_n with the perturbation $\sigma\varepsilon_n$, such that $\mathbb{X}_1$ is unperturbed Hankel matrix (i.e. $\sigma = 0$) and $\mathbb{X}_1(\sigma)$ is the perturbed one. The support series y_n has Hankel matrix $\mathbb{X}_2$. Here we study 4 cases as for univariate SSA, which are unperturbed MSSA reconstruction of series x_n, $\widetilde{x}_n^{(1),MSSA}$, perturbed MSSA reconstruction of the main series $\widetilde{x}_n^{(2),MSSA}$ and two mixed cases. First mixed case, where the reconstruction $\widetilde{x}_n^{(3),MSSA}$ comes from diagonal averaging of grouped elementary matrices

$$\eta_i^{MSSA}\left(\eta_i^{MSSA}\right)^{\mathsf{T}}\begin{pmatrix}\mathbb{X}_1(\sigma)\\ \mathbb{X}_2\end{pmatrix}$$

and reconstruction $\widetilde{x}_n^{(4),MSSA}$, which comes from diagonal averaging of grouped elementary matrices

$$\gamma_i^{MSSA}\left(\gamma_i^{MSSA}\right)^{\mathsf{T}}\begin{pmatrix}\mathbb{X}_1\\ \mathbb{X}_2\end{pmatrix},$$

where $\eta^{MSSA} = \begin{pmatrix}\eta^{(x)}\\ \eta^{(y)}\end{pmatrix}$ and $\eta^{MSSA} = \begin{pmatrix}\gamma^{(x(\sigma))}\\ \gamma^{(y(\sigma))}\end{pmatrix}$ both of size $2L$ with $\eta^{(x)}$ and $\gamma^{(x(\sigma))}$ corresponding to the first L terms of unperturbed vector η^{MSSA} and perturbed eigenvector γ^{MSSA}, respectively and $\eta^{(y)}$ and $\gamma^{(y(\sigma))}$ are corresponding to last L terms of unperturbed vector η^{MSSA} and perturbed eigenvector γ^{MSSA}, respectively.

However, if we look at the convolution expression for the main series reconstruction (see 1.33)

$$\mathbb{H}\left(\sum_{k=0}^{L-1}\sum_{j=0}^{L-1}p_{k,j}^1 S^{j-k}x + \sum_{k=0}^{L-1}\sum_{j=0}^{L-1}p_{k,j}^2 S^{j-k}y\right)$$

we see that perturbation appear in the initial time series $x_n + \sigma\varepsilon_n$ and the projectors p^1 and p^2 (1.31).

Taking a simple example with time series

$$\begin{aligned} x_n &= \sin(\pi n\omega) + 0.5\varepsilon_n\\ y_n &= \sin(\pi n\omega)\end{aligned} \tag{5.45}$$

of the length $N = 200$ with $\omega = 1.5$ and i.i.d. perturbation $0.5\varepsilon_n \sim N(0, 0.5)$ and calculating $\widetilde{x}_n^{(1),MSSA}, \widetilde{x}_n^{(2),MSSA}, \widetilde{x}_n^{(3),MSSA}\widetilde{x}_n^{(4),MSSA}$ we see in Figures 5.11 and 5.12 that the difference

$$\widetilde{x}_n^{(4),MSSA} - \widetilde{x}_n^{(1),MSSA}$$

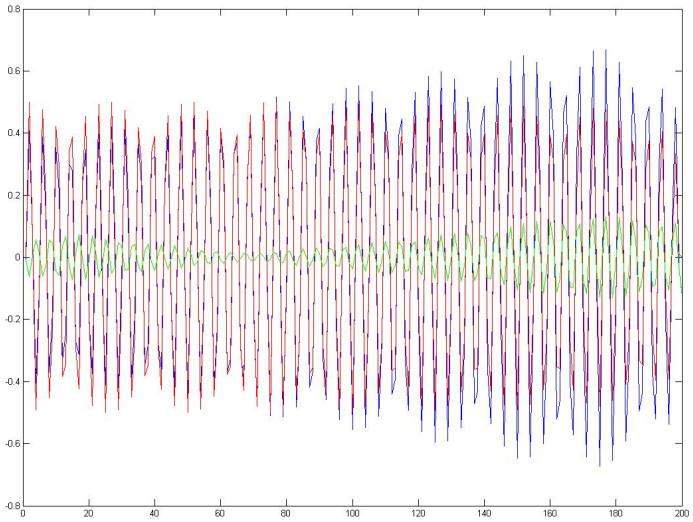


Fig. 5.11: Reconstructions differences for the model (5.45) $\widetilde{x}_n^{(4),MSSA} - \widetilde{x}_n^{(1),MSSA}$ (green), $\widetilde{x}_n^{(3),MSSA} - \widetilde{x}_n^{(1),MSSA}$ (blue), $\widetilde{x}_n^{(2),MSSA} - \widetilde{x}_n^{(1),MSSA}$ (red), L=10, r=2

does not go to zero, as expected and differences

$$\widetilde{x}_n^{(3),MSSA} - \widetilde{x}_n^{(1),MSSA} \not\approx \widetilde{x}_n^{(2),MSSA} - \widetilde{x}_n^{(1),MSSA}$$

have similar oscillations but the difference between them is also not zero.

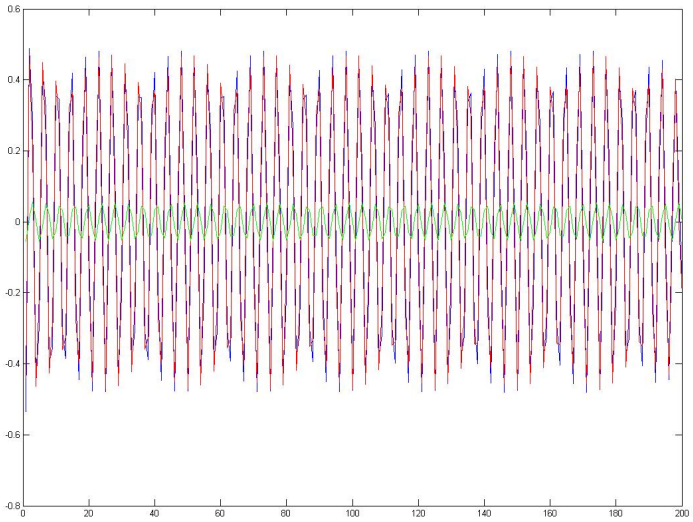


Fig. 5.12: Reconstructions differences for the model (5.45) $\widetilde{x}_n^{(4),MSSA} - \widetilde{x}_n^{(1),MSSA}$ (green), $\widetilde{x}_n^{(3),MSSA} - \widetilde{x}_n^{(1),MSSA}$ (blue), $\widetilde{x}_n^{(2),MSSA} - \widetilde{x}_n^{(1),MSSA}$ (red), L=100, r=2

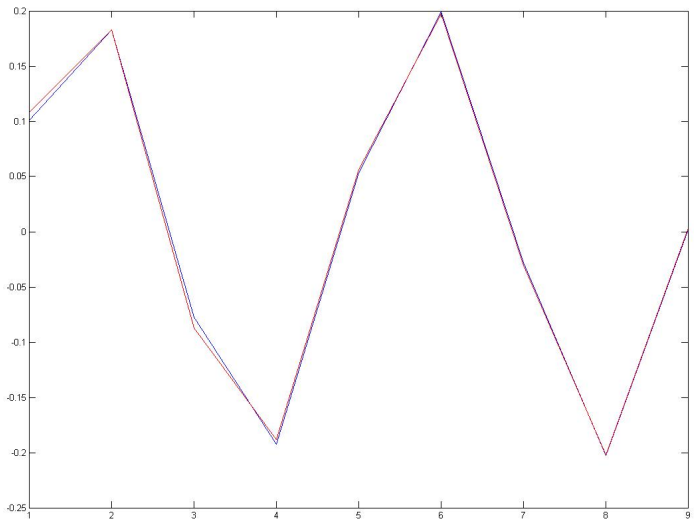


Fig. 5.13: The Recurrence Vectors obtained from SSA of (5.39) without and with perturbation $\sigma\varepsilon$, R_η (blue) and R_γ (red), respectively, L=10

5.3 Recurrence vectors stability: univariate case.

As was mentioned in the previous section the variance of the perturbation depends not only on the perturbation in the reconstruction of the series, but also is affected by the perturbation in the recurrence vectors.

The recurrence vectors approximately stay the same under small perturbation (see Figures 5.13-5.16). We may assume that the change in the recurrence vectors under perturbation can be also neglected without dramatic loss for further analysis and forecast. However, in the forecast LRF the whole recurrence vector is multiplied by the last $(L-1)$ reconstruction terms, so that the small differences at each point add up in LRF and one can think that it may lead to a bigger difference, especially if the window length L is large enough. Nevertheless, in section 5.4 we observe the opposite situation, when the larger window length actually gives a better approximation of the forecast.

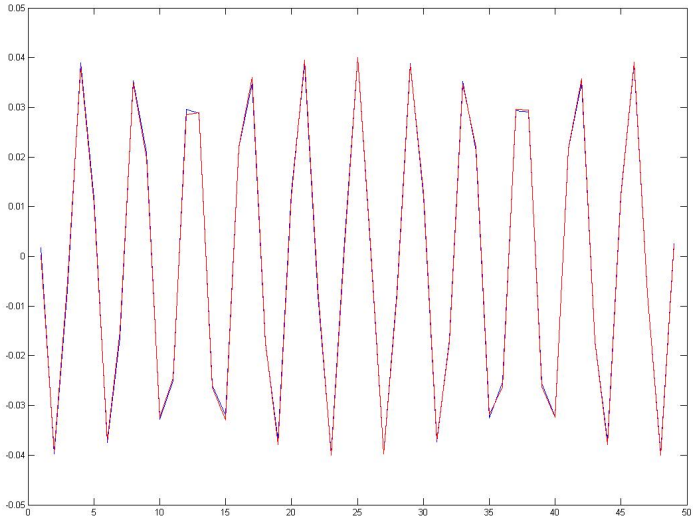


Fig. 5.14: The Recurrence Vectors from SSA of (5.39) without and with perturbation, $\sigma\varepsilon$, R_η (blue), R_γ (red), respectively, L=50

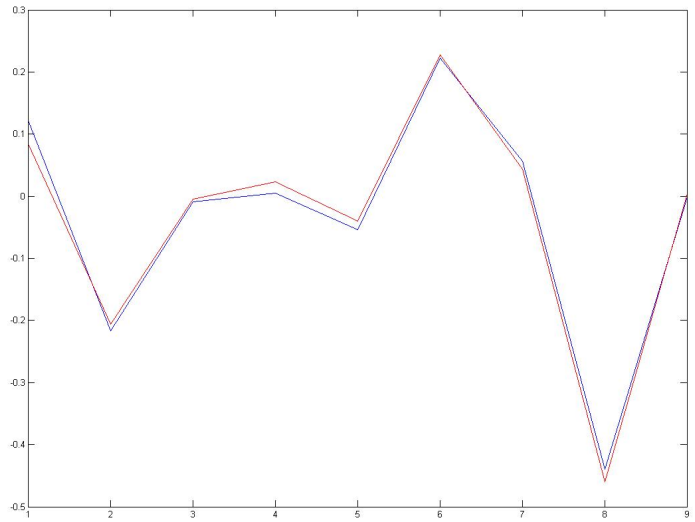


Fig. 5.15: The Recurrence Vectors from SSA of (5.40) without and with perturbation, $\sigma\varepsilon$, R_η (blue), R_γ (red), respectively, L=10

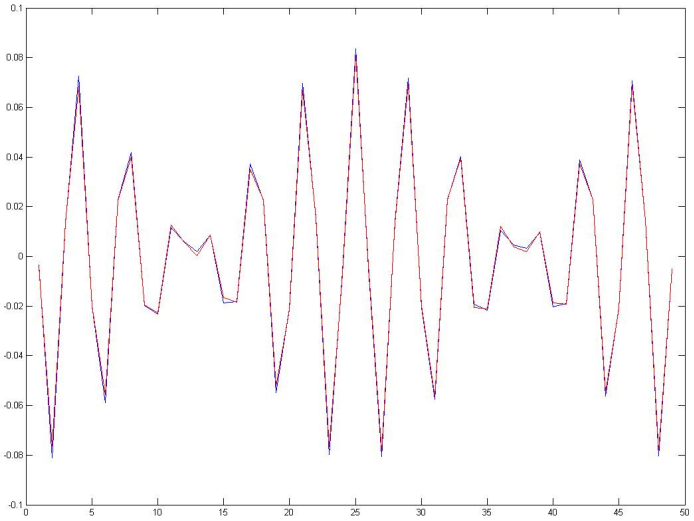


Fig. 5.16: The Recurrence Vectors from SSA of (5.40) without and with perturbation, $\sigma\varepsilon$, R_η (blue), R_γ (red), respectively, L=50

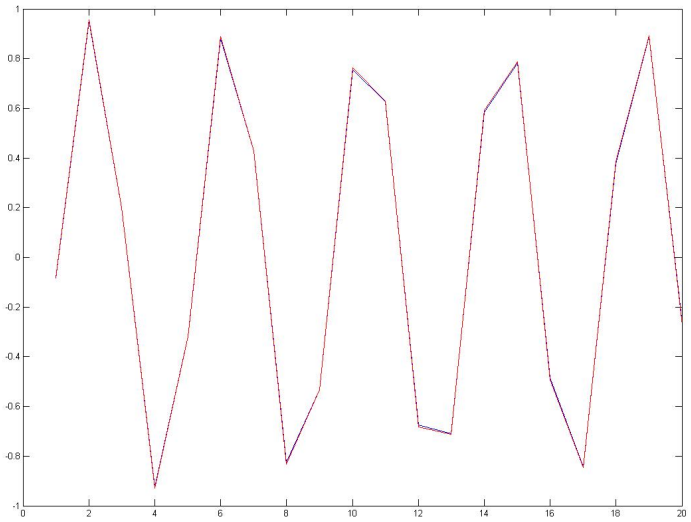


Fig. 5.17: Forecast $(\hat{x}_{181}, ..., \hat{x}_{201})$ of the main series (5.39) with perturbation $\sigma\varepsilon_n$ using unperturbed R_η (blue) and perturbed R_γ (red), L=50

5.4 Forecast stability: univariate case.

In this section we want to verify the assumption that the omission of the perturbation in reconstruction and in the recurrence vectors do not have a tangible impact on the forecast. In section 5.2 we establish that the reconstructions of the perturbed time series using perturbed and unperturbed eigenvectors look approximately the same on the interval $(L+1, ..., N-L)$. Hence, for stability analysis of the forecast, a sensible thing to do is to start the forecast from the point $(N-L)$. Here we do not wish to forecast the future, but rather we are interested in F-test statistic and its connection to the noise in the initial time series.

For a bigger window length the forecast for both time series generated stays approximately the same (see Figures 5.17,5.18). For this examples the choice of smaller window length $L = 10$ leads to a larger forecast error (see Figures 5.19,5.20). According to Figure 5.2 the window length $L = 10$ describes only half of the period, therefore, it does not capture the regularity of the behaviour, while $L = 50$ is enough for the whole period description.

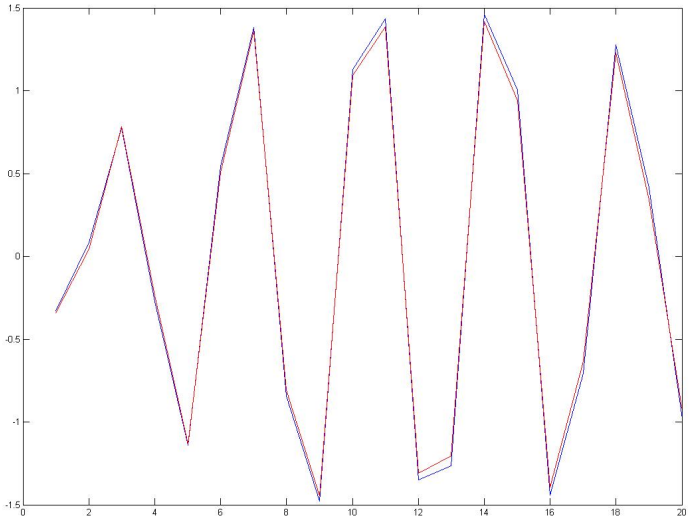


Fig. 5.18: Forecast $(\hat{x}_{181}, ..., \hat{x}_{201})$ of the main series (5.40) with perturbation $\sigma\varepsilon_n$ using unperturbed R_η (blue) and perturbed R_γ (red), L=50

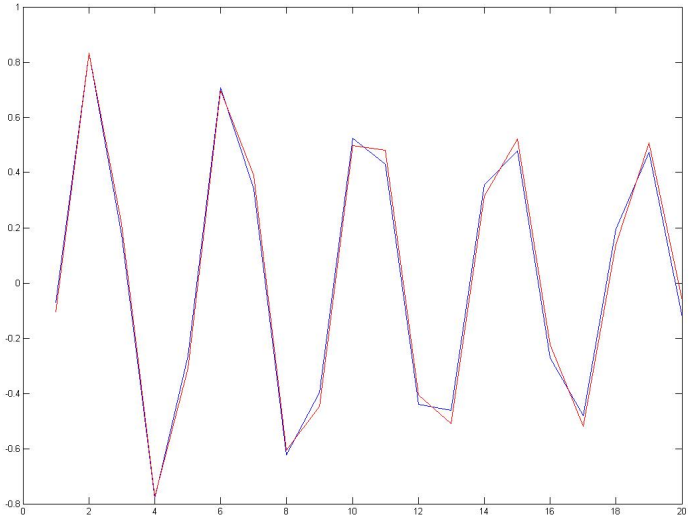


Fig. 5.19: Forecast $(\hat{x}_{181}, ..., \hat{x}_{201})$ of the main series (5.39) with perturbation $\sigma\varepsilon_n$ using unperturbed R_η (blue) and perturbed R_γ (red), L=10

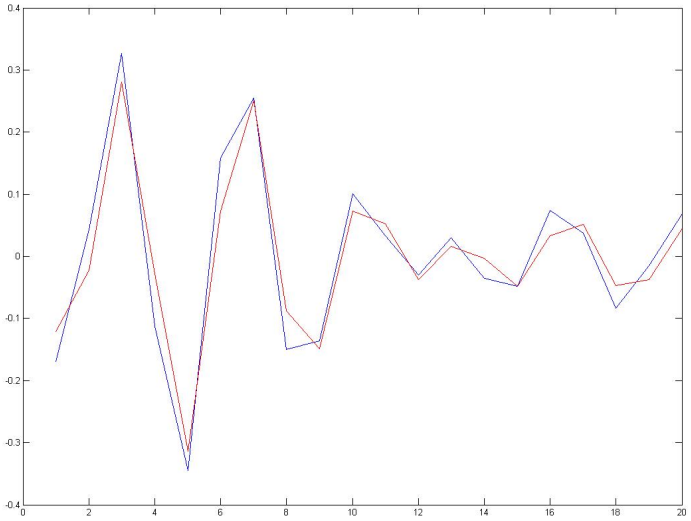


Fig. 5.20: Forecast $(\hat{x}_{181}, ..., \hat{x}_{201})$ of the main series (5.40) with perturbation $\sigma\varepsilon_n$ using unperturbed R_η (blue) and perturbed R_γ (red), L=10

5.5 Stability check performed for real data: univariate case

To validate neglecting changes in reconstructions, recurrence vectors and forecast under perturbation discussed in this chapter, we need to check it for some real time series.

The real data presented here is monthly sales of dry Australian wine for the period 1980 - 1994 [21]. Here we take dry wine time series. The analysis of the dry wine time series is done as in [11, Chapter II], where the same time series was used. The length of the series is $N = 187$ and the natural period is equal to one year, i.e. 12 months (12 points). It is natural to choose a window length to be a multiple of 12. According to the book [11, p. 138], the optimal window length to obtain the structure of the time series is $L = 24$ and the number of eigentriples, which correspond to the trend and main periodics, is the first five.

Here we consider the *signal* to be based on the first 5 principal components, and the leftover residuals correspond to a noise term.

There are several approaches to look over the stability of given data reconstruction.

We begin with defining the series signal and residuals and then work with obtained model. We define the time series structure as follows. The main series $\widetilde{x}_{1-5}$ is smoothened dry wine series, based of first the 5 eigentriples, residuals $\varepsilon = x - \widetilde{x}_{1-5}$, where x is initial dry wine time series, correspond to the natural noise. Based on this structure, we now set up perturbed time series in two different ways.

1. We take $\widetilde{x}_{1-5}$ to be the unperturbed time series and perturb it with permuted residuals ε. The permutation of the residuals destroys the initial correlations in it, but keeps its initial mean, variance and distribution.

2. We take the initial dry wine series, $\widetilde{x}_{1-5} + \varepsilon$, and permute it with the artificial generated white noise $\xi \sim WN(0, \sigma^2)$ with $\sigma = 200$.

The wine time series, the reconstruction based on the first 5 principal components ($L = 24$) and residuals are illustrated in Figure 5.21.

The differences between reconstructions shown in Figure 5.22 (original residuals) and in Figure 5.24 (permuted residuals) supports the statement that reconstructions $x^{(1)}$ and $x^{(4)}$ are mostly identical apart from the first and last window length interval. Also, if we permute given time series by some generated noise we will get the same result (see Figure 5.5). The perturbation for the example shown in Figure 5.5 is i.i.d. white noise with the variance $\sigma^2 = 40000$.

For all three suggested approaches we checked how the recurrence vector change under each perturbation. As one can see from Figures 5.23, 5.25, 5.5 the change is minor and could be omitted.

Although the recurrence vectors do not change dramatically under perturbation, the forecast seems to change more significantly if we permute the original residuals (see 5.5) and much less if one adds artificial Gaussian white noise to the time series (5.5).

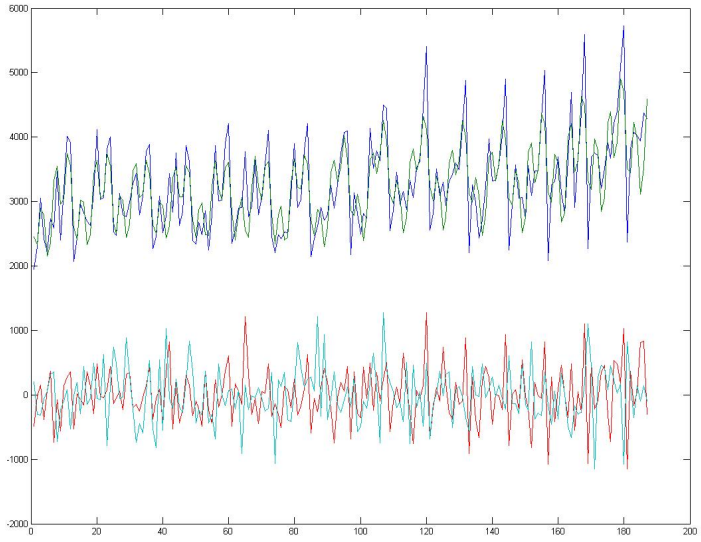


Fig. 5.21: The initial Australian dry wine time series (blue), signal (green), noise (red), permuted noise (light blue)

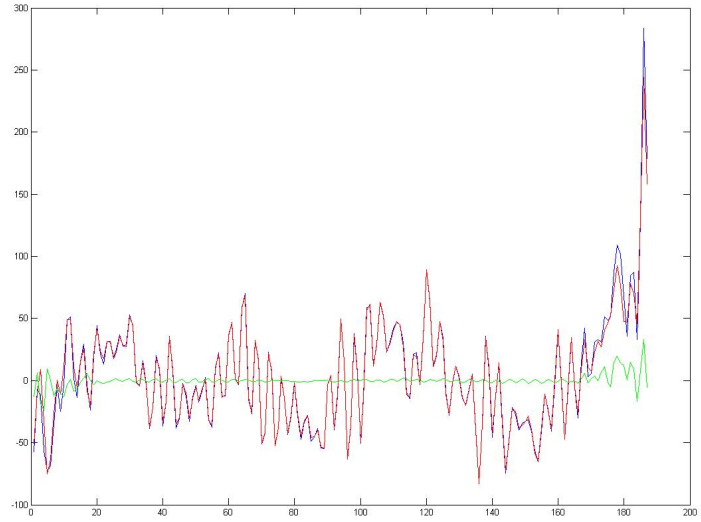


Fig. 5.22: Reconstructions differences: $\widetilde{x}_n^{(2)} - \widetilde{x}_n^{(1)}$ (blue), $\widetilde{x}_n^{(3)} - \widetilde{x}_n^{(1)}$ (red), $\widetilde{x}_n^{(4)} - \widetilde{x}_n^{(1)}$ (green)

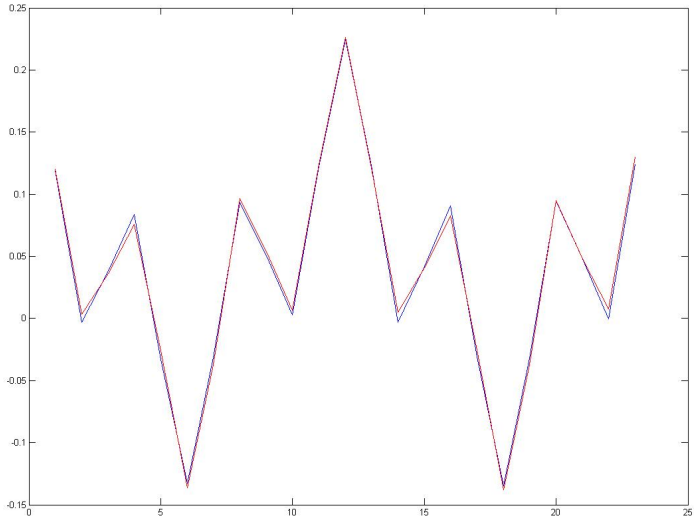


Fig. 5.23: Recurrence Vectors: R_η (blue), R_γ (red)

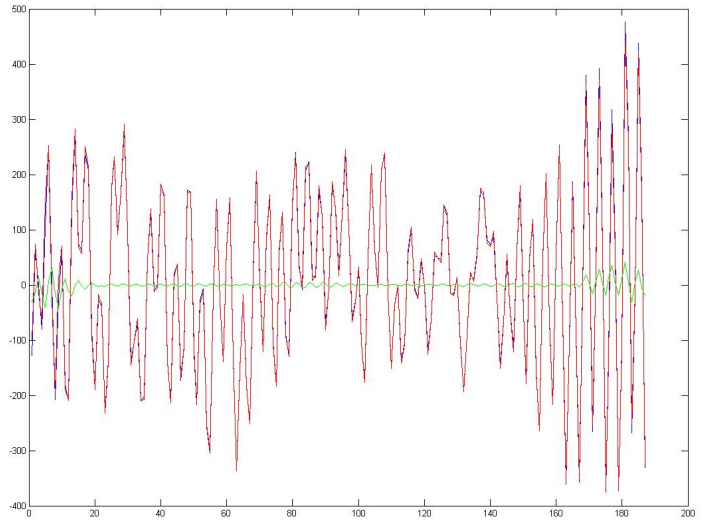


Fig. 5.24: Reconstructions differences: $\widetilde{x}_n^{(2)} - \widetilde{x}_n^{(1)}$ (blue), $\widetilde{x}_n^{(3)} - \widetilde{x}_n^{(1)}$ (red), $\widetilde{x}_n^{(4)} - \widetilde{x}_n^{(1)}$ (green)

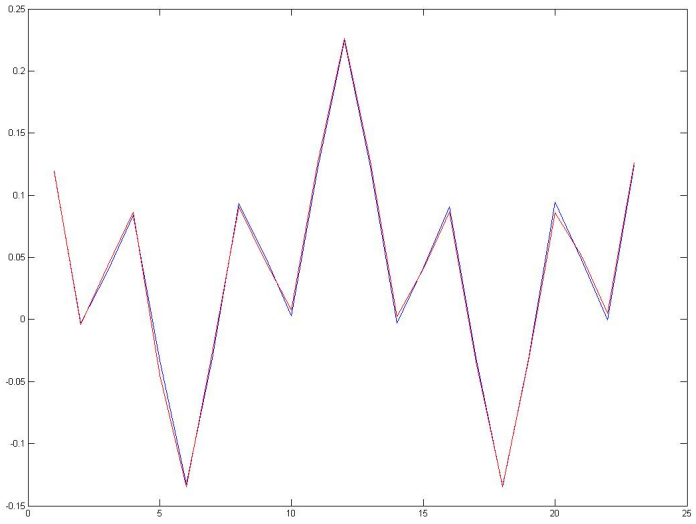


Fig. 5.25: Recurrence Vectors R_η (blue) and R_γ (red)

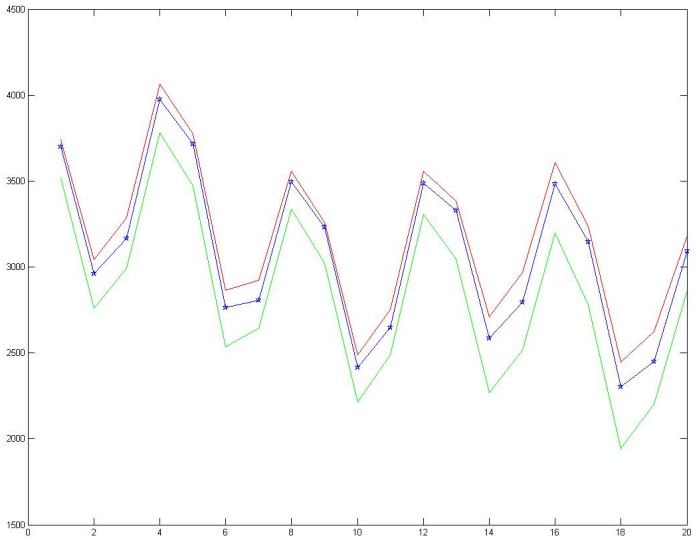


Fig. 5.26: Forecast $(\hat{x}_{128}, ..., \hat{x}_{148})$ with unperturbed R (blue), $(\hat{x}_{128}, ..., \hat{x}_{148})$ with perturbed R and original residuals (red), $(\hat{x}_{128}, ..., \hat{x}_{148})$ with perturbed R and permuted residuals (green)

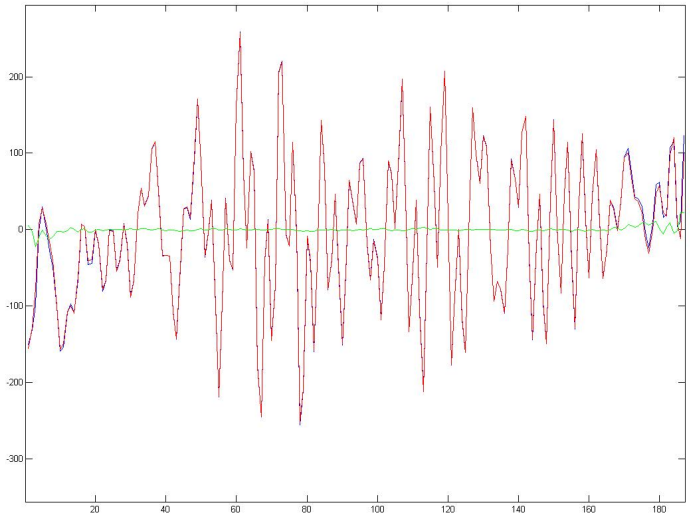


Fig. 5.27: Reconstructions differences: $\widetilde{x}_n^{(2)} - \widetilde{x}_n^{(1)}$ (blue), $\widetilde{x}_n^{(3)} - \widetilde{x}_n^{(1)}$ (red), $\widetilde{x}_n^{(4)} - \widetilde{x}_n^{(1)}$ (green)

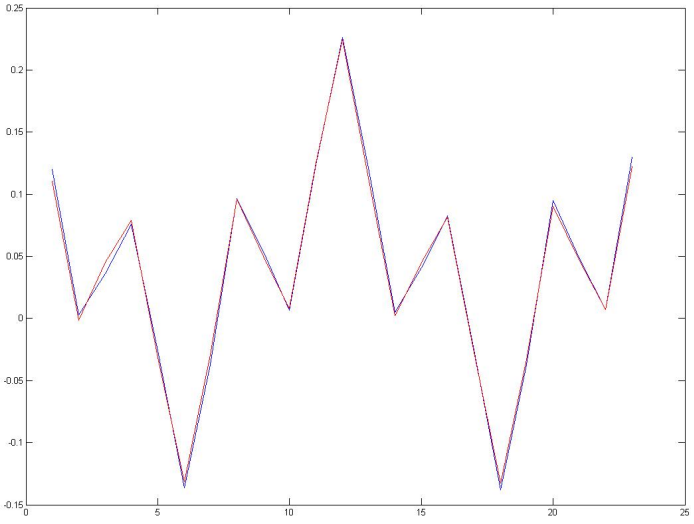


Fig. 5.28: Recurrence Vectors: R_η (blue), R_γ (red)

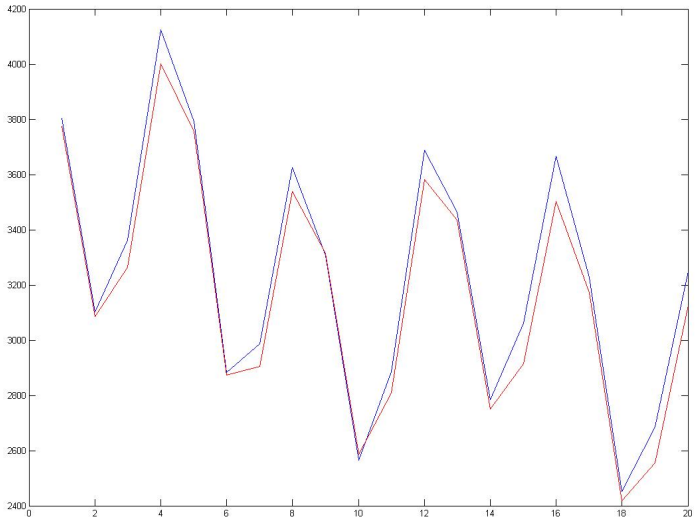


Fig. 5.29: Forecast $(\hat{x}_{128}, ...\hat{x}_{148})$ with R_η (blue), $(\hat{x}_{128}, ..., \hat{x}_{148})$ with R_γ (red)

5.6 *Analysis of noise variance: univariate case*

The convolution representation of SSA makes it is possible to study noise propagation in more transparent way.

Suppose we have some perturbed time series

$$x_n + \sigma\varepsilon_n$$

where x_n correspond to signal and ε_n is i.i.d. Gaussian white noise. Using the convolution expression for the SSA reconstruction (1.23) with some coefficients $q_i(\sigma)$ and substituting our perturbed time series, we get

$$\widetilde{x}_n(\sigma) = \sum_{m\in\mathbb{Z}} q_{n-m}(\sigma)(x_m + \sigma\varepsilon_m) \tag{5.46}$$

$$= \sum_{m\in\mathbb{Z}} q_{n-m}(\sigma)x_m + \sigma \sum_{m\in\mathbb{Z}} q_{n-m}(\sigma)\varepsilon_m \tag{5.47}$$

Applying the LRF (1.41) to the reconstruction (5.46), the expression for the forecast $\hat{x}_{N+1}(\sigma)$ may be written as

$$\hat{x}_{N+1}(\sigma) = \sum_{i=1}^{L-1} a_i(\sigma) \left(\sum_{m\in\mathbb{Z}} q_{N+1-i-m}(\sigma)x_m + \sigma \sum_{m\in\mathbb{Z}} q_{N+1-i-m}(\sigma)\varepsilon_m \right) \tag{5.48}$$

$$= \sum_{i=1}^{L-1} a_i(\sigma) \sum_{m\in\mathbb{Z}} q_{N+1-i-m}(\sigma)x_m + \sigma \sum_{i=1}^{L-1} a_i(\sigma) \sum_{m\in\mathbb{Z}} q_{N+1-i-m}(\sigma)\varepsilon_m \tag{5.49}$$

From (5.49) we see that both kernel coefficients $q_i(\sigma)$ and recurrence vector coefficients $a_i(\sigma)$ depend on noise of the initial time series, which complicates the process of estimating the variance of the noise in the reconstruction and in the forecast.

The stability analysis has shown that under small perturbation it is possible to ignore the changes in time series reconstructions and recurrence vectors. In terms of convolution SSA it means that under small perturbation we can ignore the changes in coefficients q_i.

Based on the stability analysis in Chapter 5 we can make an assumption that recurrence vector coefficients and the convolution kernel coefficients do not change under perturbation $\sigma\varepsilon$. This assumption gives rise to the following proposition.

Proposition 5.6.1. *Suppose we have a time series*

$$x_n(\sigma) = x_n + \sigma\varepsilon_n \quad (n \in \mathbb{Z}),\ s.t.\ x_n = \varepsilon_n = 0,\ if\ \notin \{1, ..., N\},$$

where x_n *correspond to unperturbed time series and* $\varepsilon_n \sim N(0,1)$ *i.i.d. if* $n \in \{1, ..., N\}$.

Let (5.49) be the LRF for the forecast point $x_{N-L+1}(\sigma)$. *Then assuming that the convolution kernel coefficients do not change under perturbation the variance of forecast error can be written as*

$$var(\hat{\varepsilon}_{N-L+1}) = \sigma^2 \|R \star q\|_2^2$$

where R *and* q *are obtained from the unperturbed time series* x_n.

Proof. Assuming that convolution kernel coefficients do not change under perturbation we can substitute coefficients q_i obtained from the unperturbed time series instead of $q_i(\sigma)$ into (5.47), and get

$$\widetilde{x}_n(\sigma) = \sum_{m\in\mathbb{Z}} q_{n-m}x_m + \sigma \sum_{m\in\mathbb{Z}} q_{n-m} \tag{5.50}$$

$$= \widetilde{x}_n(0) + \sigma\widetilde{\varepsilon}_m \tag{5.51}$$

where

$$\widetilde{\varepsilon}_n = \sum_{m\in\mathbb{Z}} q_{n-m}\varepsilon_m \tag{5.52}$$

$$\widetilde{\varepsilon}_n = \sum_{m\in\mathbb{Z}} q_{n-m}\varepsilon_m \tag{5.53}$$

is perturbation in the reconstruction $\widetilde{x}_n(\sigma)$. The variance of the perturbation in reconstruction of $\widetilde{x}_n(\sigma)$ could be easily calculated, using the property of independent random variables. And the variance of the perturbation in reconstruction at each point is

$$var(\widetilde{\varepsilon}_n) = \sum_{m\in\mathbb{Z}} \mid q_{n-m} \mid_2^2 \tag{5.54}$$

The forecast is obtained using LRF (1.41)

$$\hat{x}_{N-L+1}(\sigma) = \sum_{i=1}^{L-1} a_i \widetilde{x}_{N-L+1-i}(\sigma) \tag{5.55}$$

where a_i are the terms of the recurrence vector R (1.42). Substituting (5.47) into (5.55) we get

$$\hat{x}_{N-L+1}(\sigma) = \sum_{i=1}^{L-1} a_i \Big(\sum_{m\in\mathbb{Z}} q_{N-L+1-i-m} x_m + \sigma \sum_{m\in\mathbb{Z}} q_{N-L+1-i-m}\varepsilon_m\Big) \tag{5.56}$$

$$= \sum_{i=1}^{L-1} a_i \sum_{m\in\mathbb{Z}} q_{N-L+1-i-m} x_m + \sigma \sum_{i=1}^{L-1} a_i \sum_{m\in\mathbb{Z}} q_{N-L+1-i-m}\varepsilon_m \tag{5.57}$$

$$= \hat{x}_{N-L+1}(0) + \sigma \sum_{i=1}^{L-1} a_i \sum_{m\in\mathbb{Z}} q_{N-L+1-i-m}\varepsilon_m \tag{5.58}$$

Thus

$$\hat{\varepsilon}_{N-L+1} = \sigma \sum_{i=1}^{L-1} a_i \sum_{m\in\mathbb{Z}} q_{N-L+1-i-m}\varepsilon_m \tag{5.59}$$

$$= \sigma \sum_{m\in\mathbb{Z}} \Big(\sum_{i=1}^{L-1} a_i q_{N-L+1-i-m}\Big)\varepsilon_m \tag{5.60}$$

is a perturbation term of the forecast, *under simplifying assumptions.*

Before calculating the variance of the output noise, let us recall the assumptions that have been made.

First of all, the input noise is the i.i.d. and comes from normal distribution with zero mean and variance σ^2. Secondly, we assume that the change in eigenvectors is not crucial due to perturbation and therefore we omit these changes, so that eigenvectors are fixed. And the last thing, that is worth mentioning is that due to the second assumption the convolution coefficients q_m are also fixed and do not depend on noise.

So since it is possible to separate the perturbation term $\sigma\hat{\varepsilon}_{N-L+1}$ from the signal term $\hat{x}_{N-L+1}(0)$, we can find the variance of $\hat{\varepsilon}_{N-L+1}$

$$var(\hat{\varepsilon}_{N-L+1}) = \sigma^2 \sum_{m\in\mathbb{Z}} \mid \sum_{i=1}^{L-1} a_i q_{N-L+1-i-m} \mid^2 \tag{5.61}$$

$$= \sigma^2 \sum_{k\in\mathbb{Z}} \mid \sum_{i=1}^{L-1} a_i q_{k-i} \mid^2 \quad \text{where } k = N-L+1-m \tag{5.62}$$

which is the ℓ^2 norm of the convolution of the recurrence vector with the reconstruction kernel,

$$\boxed{var(\hat{\varepsilon}_{N-L+1}) = \sigma^2 \|R \star q\|_2^2} \tag{5.63}$$

□

According to the stability analysis we can substitute the perturbed coefficients with unperturbed.

Although, the recurrence vector R_{11} behaviour is stable under small perturbation, there are certain instability problems with R_{12}, variance analysis for the real time series further in this section. It is also interesting to look at the behaviour of the constructed variance of the MSSA forecast under assumptions that minor changes in recurrence vector and convolution coefficients can be neglected and what actually is causing the instability in the R_{12} recurrence vector.

Suppose now we add a noise perturbation to the

$$x_n(\sigma) = x_n + \sigma\varepsilon_n \tag{5.64}$$

where x_n corresponds to signal and ε_n corresponds to Gaussian white noise, $\varepsilon_n \sim N(0,1)$ i.i.d..

Recalling MSSA LRF for the forecast (1.58) and its recurrence vectors R_{11} (1.60) and R_{12} (1.61) one can find the forecast $\hat{x}_{N-L+1}(\sigma)$:

$$\hat{x}_{N-L+1}(\sigma) = \sum_{i=1}^{L-1} a_{1,i}\widetilde{x}_{N-L+1-i}(\sigma) + \sum_{i=1}^{L-1} b_{1,i}\widetilde{y}_{N-L+1-i}(\sigma) \tag{5.65}$$

where $a_{1,i}, b_{1,i}$ are terms of two recurrence vectors R_{11} (1.60) and R_{12} (1.61) respectively.

Note that we are operating under following assumption that recurrence vectors R_{11} and R_{12} are obtained from the unperturbed time series x_n, i.e. fixed.

Substituting reconstructions (1.35) into (5.65) we get

$$\begin{aligned}\hat{x}_{N-L+1}(\sigma) = &\sum_{i=1}^{L-1} a_{1,i}\Big(\sum_{m\in\mathbb{Z}} q^1_{N-L+1-i-m}x_m(\sigma) + \sum_{m\in\mathbb{Z}} q^2_{N-L+1-i-m}y_m\Big)\\ &+ \sum_{i=1}^{L-1} b_{1,i}\Big(\sum_{m\in\mathbb{Z}} q^3_{N-L+1-i-m}x_m(\sigma) + \sum_{m\in\mathbb{Z}} q^4_{N-L+1-i-m}y_m\Big)\end{aligned} \tag{5.66}$$

Then substituting (5.64) into (5.66), we get

$$\begin{aligned}\hat{x}_{N-L+1}(\sigma) = &\sum_{i=1}^{L-1} a_{1,i}\Big(\sum_{m\in\mathbb{Z}} q^1_{N-L+1-i-m}x_m + \sigma\sum_{m\in\mathbb{Z}} q^1_{N-L+1-i-m}\varepsilon_m + \sum_{m\in\mathbb{Z}} q^2_{N-L+1-i-m}y_m\Big)\\ &+ \sum_{i=1}^{L-1} b_{1,i}\Big(\sum_{m\in\mathbb{Z}} q^3_{N-L+1-i-m}x_m + \sigma\sum_{m\in\mathbb{Z}} q^3_{N-L+1-i-m}\varepsilon_m + \sum_{m\in\mathbb{Z}} q^4_{N-L+1-i-m}y_m\Big).\end{aligned} \tag{5.67}$$

$$= \hat{x}_{N-L+1}(0) + \hat{\varepsilon}_{N-L+1} \tag{5.68}$$

Define

$$\hat{\varepsilon}_{N-L+1} = \sigma \sum_{i=1}^{L-1} a_{1,i} \sum_{m\in\mathbb{Z}} q^1_{N-L+1-i-m}\varepsilon_m + \sigma \sum_{i=1}^{L-1} b_{1,i} \sum_{m\in\mathbb{Z}} q^3_{N-L+1-i-m}\varepsilon_m \tag{5.69}$$

$$= \sigma \sum_{m\in\mathbb{Z}} \sum_{i=1}^{L-1} (a_{1,i} q^1_{N-L+1-i-m} + b_{1,i} q^3_{N-L+1-i-m})\varepsilon_m \tag{5.70}$$

as a perturbation term.

Similarly to the univariate case, we can now calculate the variance of the perturbation in the forecast $\hat{x}_{N-L+1}(\sigma)$

$$var(\hat{\varepsilon}_{N-L+1}) = \sigma^2 \sum_{m\in\mathbb{Z}} \mid \sum_{i=1}^{L-1} (a_{1,i} q^1_{N-L+1-i-m} + b_{1,i} q^3_{N-L+1-i-m}) \mid^2 \tag{5.71}$$

$$= \sigma^2 \sum_{k\in\mathbb{Z}} \mid \sum_{i=1}^{L-1} (a_{1,i} q^1_{k-i} + b_{1,i} q^3_{k-i}) \mid^2 \text{ , where } k = N-L+1-m \tag{5.72}$$

which is the ℓ^2 norm of the sum of convolutions of recurrence and reconstruction vectors

$$\boxed{var(\hat{\varepsilon}_{N-L+1}) = \sigma^2 \| R_{11} \star q^1 + R_{12} \star q^3 \|_2^2} \tag{5.73}$$

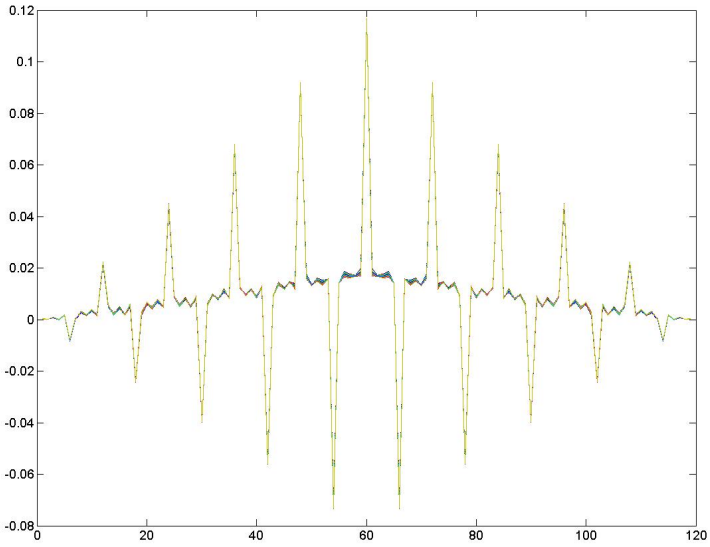


Fig. 5.30: Setup 1. Generated noise added to initial time series. SSA convolution coefficients q.

5.8 Stability analysis and convolution based causality measure

Stability of convolution kernel coefficients and recurrence vectors on the real data example Before estimating variances of the output noise (5.63,5.73) we need to make sure that the omitting changes under perturbation assumption holds for the real data.

Let red wine series be main series x and sparkling wine to be support series y. For both time series $L = 60$ and $r = 7$ seem to be optimal choice of SSA parameters. Let us define 3 following setups for this test

1. *signal* - initial x, *perturbation* - generated white noise $\zeta \sim WN(0, 100)$ (generated $N_s = 1000$ times), *support* - initial y;

2. *signal* - reconstruction based on first 7 principal components ($L = 60$) $\widetilde{x}_{1-7}$, *perturbation* - leftover residuals, permuted $N_s = 256$ times, *support* - reconstruction based on first 7 principal components ($L = 60$) $\widetilde{y}_{1-7}$;

3. *signal* - initial $\widetilde{x}_{1-7}$, *perturbation* - generated white noise $\zeta \sim WN(0, 100)$ (generated $N_s = 1000$ times), *support* - initial $\widetilde{y}_{1-7}$;

For each setup we want to investigate the behavior of the convolution coefficients q, q^1, q^3 and the recurrence vectors R, R_{11}, R_{12}.

For the Setup 1 (see Figures 5.30-5.35) and 3 (see 5.42-5.47) we observe good stable behaviour, especially for the setup 3, where we have smoothened time series (the signal is based on the first 5 eigentriples) and added Gaussian white noise. However, in case

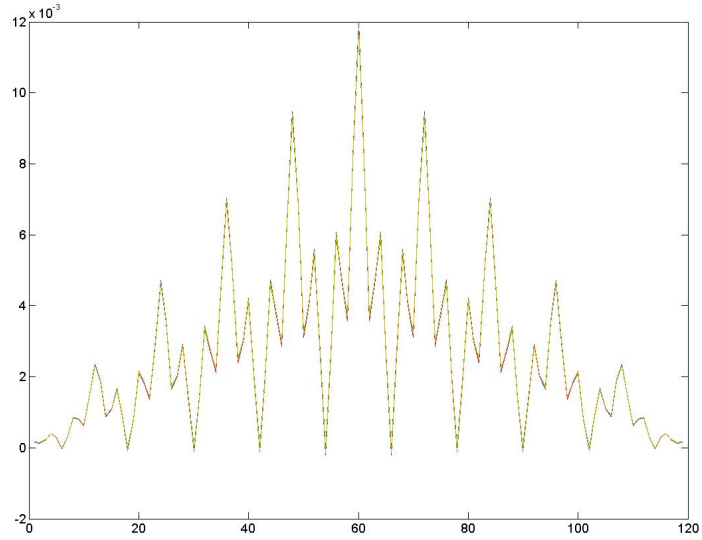


Fig. 5.31: Setup 1. Generated noise added to initial time series. MSSA convolution coefficients q^1.

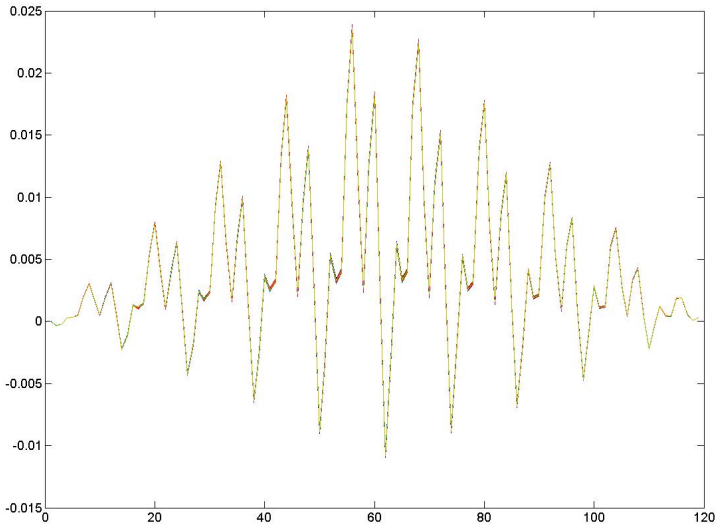


Fig. 5.32: Setup 1. Generated noise added to initial time series. MSSA convolution coefficients q^3.

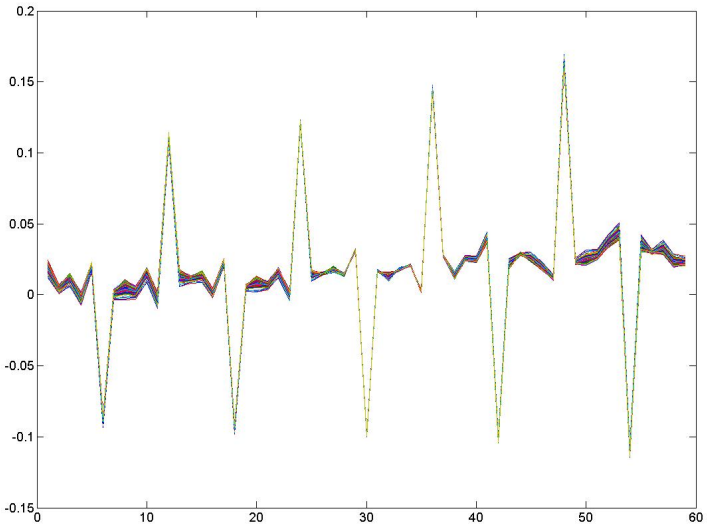


Fig. 5.33: Setup 1. Generated noise added to initial time series. SSA recurrence vector R.

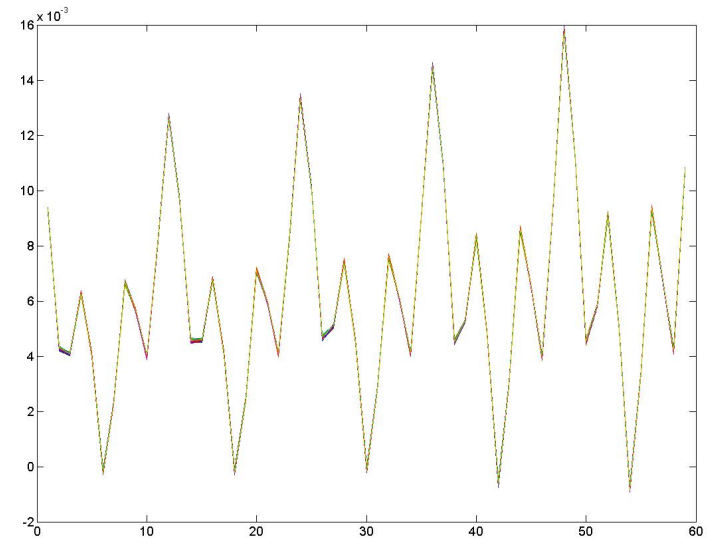


Fig. 5.34: Setup 1. Generated noise added to initial time series. MSSA recurrence vector R_{11}.

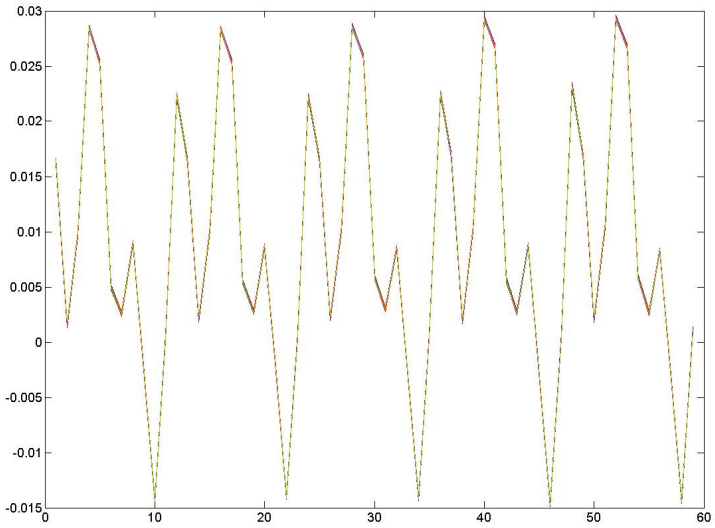


Fig. 5.35: Setup 1. Generated noise added to initial time series. MSSA recurrence vector R_{12}.

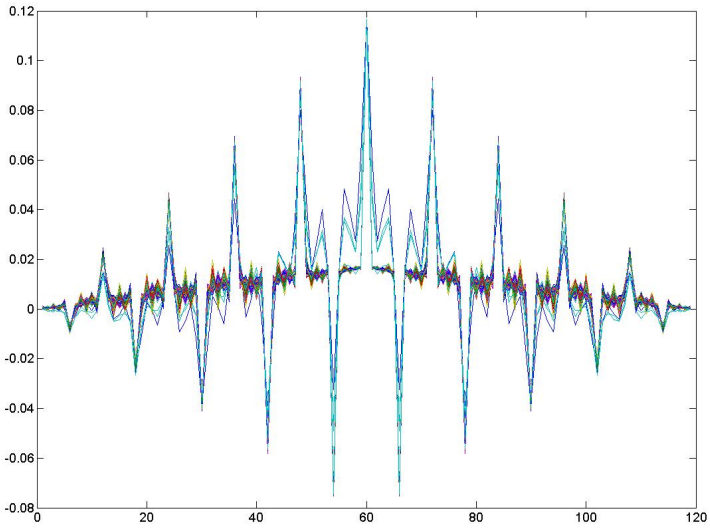


Fig. 5.36: Setup 2. SSA convolution coefficients q.

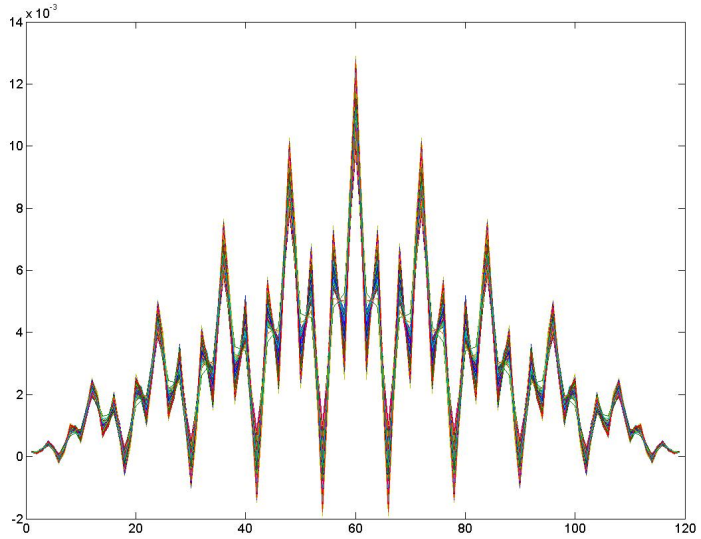


Fig. 5.37: Setup 2. MSSA convolution coefficients q^1.

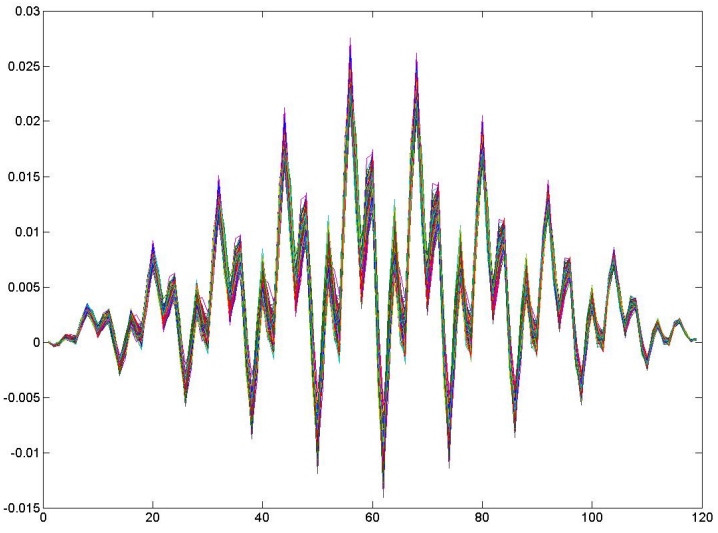


Fig. 5.38: Setup 2. MSSA convolution coefficients q^3.

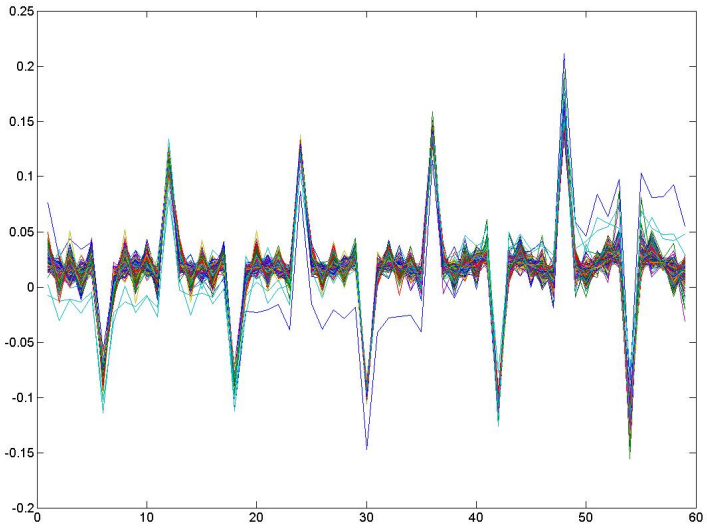


Fig. 5.39: Setup 2. Time series with permuted natural noise. SSA recurrence vector R.

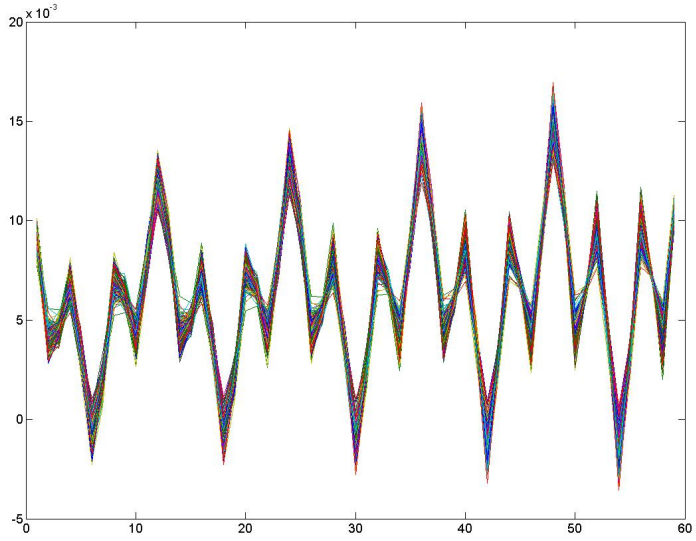


Fig. 5.40: Setup 2. Time series with permuted natural noise. MSSA recurrence vector R_{11}.

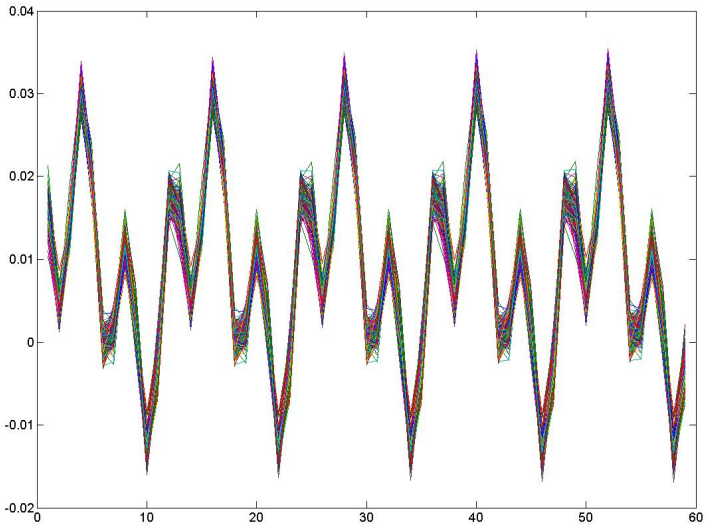


Fig. 5.41: Setup 2. Time series with permuted natural noise. MSSA recurrence vector R_{12}.

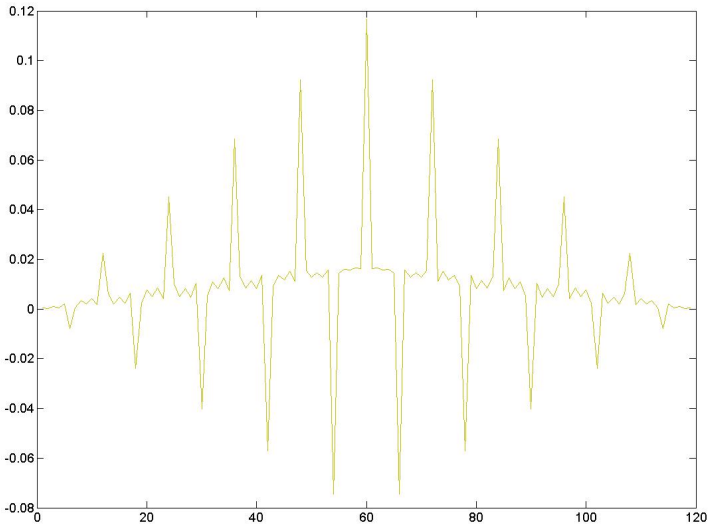


Fig. 5.42: Setup 3. Time series with permuted natural noise. SSA convolution coefficients q.

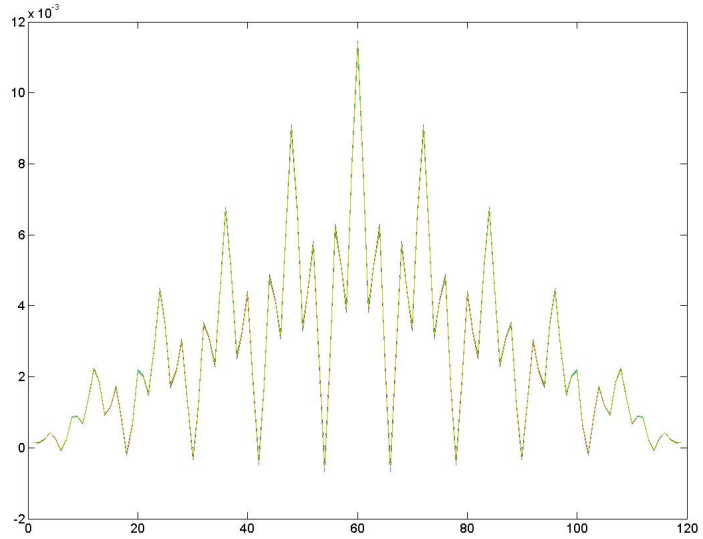


Fig. 5.43: Setup 3. Time series with permuted natural noise. MSSA convolution coefficients q^1.

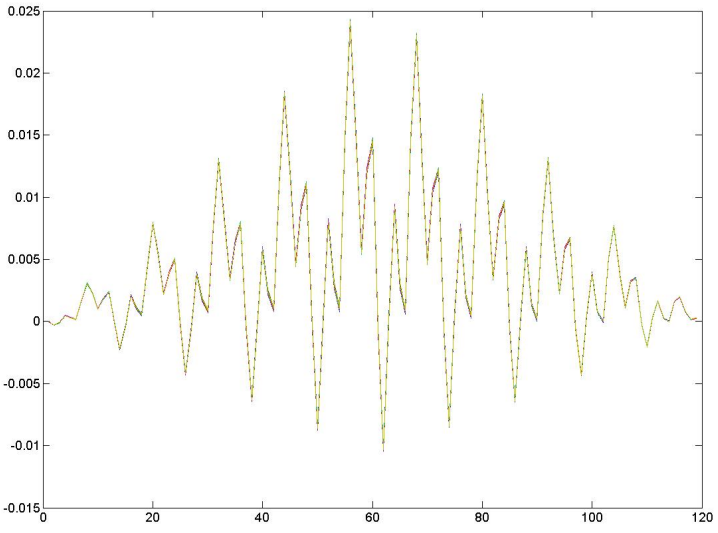


Fig. 5.44: Setup 3. Time series with permuted natural noise. MSSA convolution coefficients q^3.

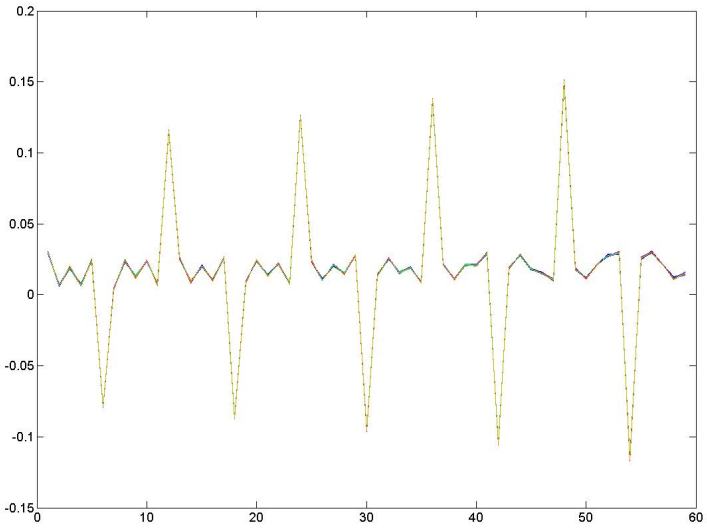


Fig. 5.45: Setup 3. Generated noise added to the reconstruction. SSA recurrence vector R.

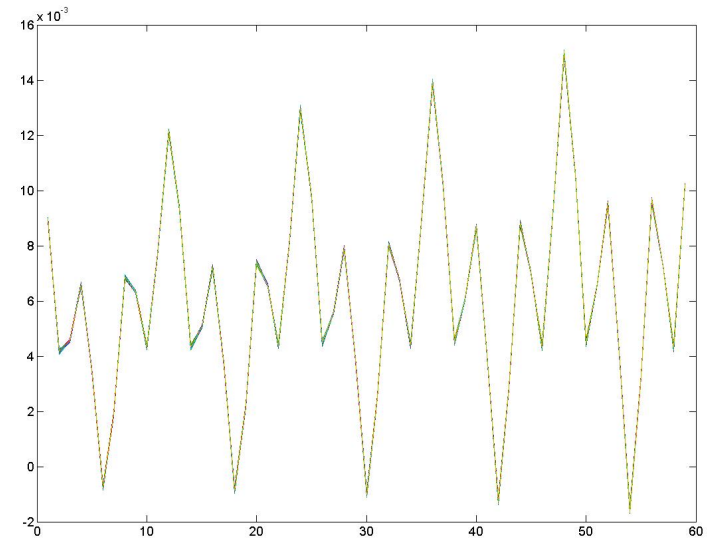


Fig. 5.46: Setup 3. Generated noise added to the reconstruction. MSSA recurrence vector R_{11}.

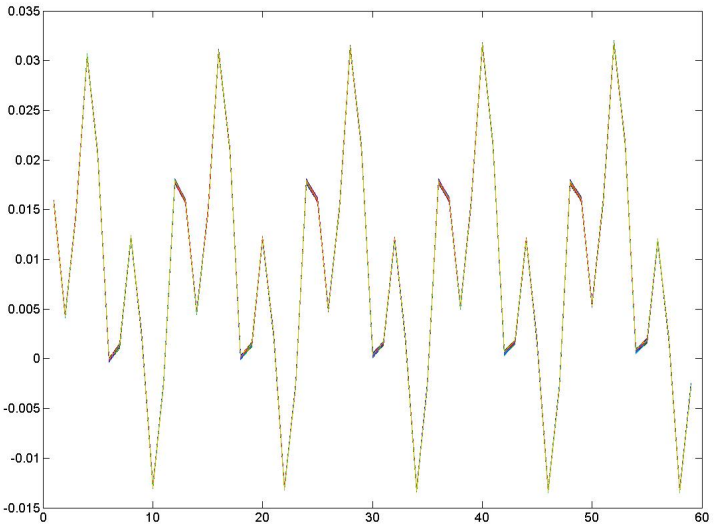


Fig. 5.47: Setup 3. Generated noise added to the reconstruction. MSSA recurrence vector R_{12}.

when we try to destroy the correlation in natural noise of the series we observe instability, see Figures 5.36-5.41. Thus, in the following section we first smooth the data and add Gaussian noise to it to do further stability analysis. Note that each graph is plotted for $N_s = 1000$ trials. Thus, the fatness of the graphs gives the indication of the stability. Note that the condition of noise being Gaussian seems to improve the stability.

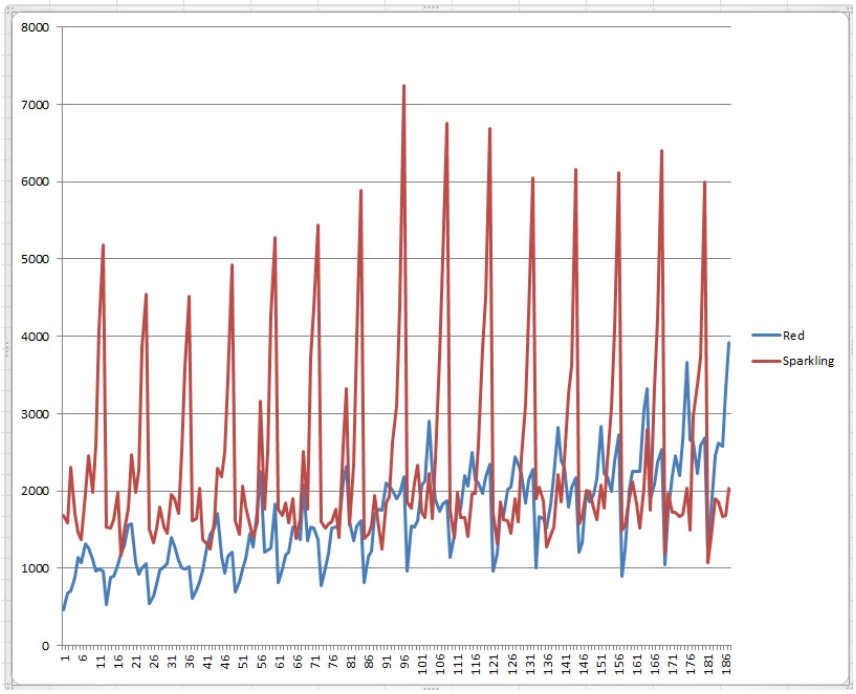


Fig. 5.48: Red wine (red) and sparkling wine (blue), 1980-1994.

Stability of forecast error variances on the real data example

The time series is taken from the same data [21]. For this example we have chosen red and sparkling wine time series of the length $N = 187$, as visually the periodics of both time series look the same, see Figure 5.48.

Before running the trials to examine the goodness of output noise variances (5.73,5.63), the data need to be analyzed. According to the book [11] [p.138-139], the optimal parameters for both, the red and the sparkling wine, time series are as following: window length is $L = 60$ and number of principal components $r = 7$.

Each trial of the test has the following setup:

- main series x - smoothened red wine series, based on the first 7 eigentriples (obtained with SSA),
- support series y - smoothened sparkling wine series, based of the first 7 eigentriples (obtained with SSA),
- $\varepsilon \sim WN(0, \sigma^2)$ - pseudo-random generated white noise using MATLAB with $\sigma = 10$ (generated $N_s = 1000$ times);
- $\rho \in \{0.1, 0.2, ..., 1.1\}$ is the support series multiplier.

Note that the added noise does not change when we switch to different ϱ

Tab. 5.1: SSA Red Wine Forecast Measures

σ_{SSA}	$\sigma^2\lvert R^s * q\rvert$	$var(\mathcal{C})$ (5.77)	$var(\hat{x}_{128}^{SSA})$
7.9911	6.7355	7.2037	7.9991

So the time series for each trial are

$$\begin{cases} x_n + \varepsilon_n \\ \rho y_n \end{cases} \qquad n \in \{1, ..., 200\},\ n \in \{1, ..., 187\} \tag{5.74}$$

We define several expressions, which help us to study the forecast stability

$$\mathcal{A} = \sum_{i=1}^{L-1} (a_{1,i}^s \widetilde{x}_{128-i} + b_{1,i} \widetilde{y}_{128-i}) \tag{5.75}$$

$$\mathcal{B} = \sum_{i=1}^{L-1} (a_{1,i}^s \widetilde{x}_{128-i} + b_{1,i}^s \widetilde{y}_{128-i}) \tag{5.76}$$

$$\mathcal{C} = \sum_{i=1}^{L-1} (c_{1,i}^s \widetilde{x}_{128-i}^{SSA}) \tag{5.77}$$

where $\widetilde{x}$, $\widetilde{y}$ are reconstructions obtained from MSSA, $\widetilde{x}^{SSA}$ is a reconstruction obtained from SSA, a_i and b_i are the coefficients of the recurrence vectors R_{11} and R_{12} respectively, and a_i^s,b_i^s,c_i^s are the coefficients of *fixed (unperturbed)* recurrence vectors R_{11}^s, R_{12}^s, R^s obtained from the MSSA(SSA) applied to the unperturbed time series. The unperturbed time series in this case correspond to x_n for SSA; for MSSA the unperturbed time series correspond to (5.74), where $\varepsilon_n = 0$.

First, looking at SSA resulting Table 5.1, one could see that the theoretical measure (5.63) is not far off $\mathcal{C}$ and could be its reasonable estimate.

Leaving the values for $var(\mathcal{A})$ for MSSA aside for a moment, we observe that the variance (5.73) in Table 5.2 decrease dramatically as ρ grows, while the empirical variance $var(\hat{x}_{128})$ seem to be more or less stable, apart from the outstanding value for $\rho = 0.18$. Although, ignoring $\rho = 0.18$ values for a moment, we see it decreases till certain ρ (0.8), and then we observe slight increase in empirical variance. We also observe that the variance $var(\mathcal{B})$ has a similar tendency as (5.73). In fact, recalling (5.73), one can see that

$$var(\mathcal{B}) \approx var(\hat{\varepsilon}_{N+1}^{MSSA}). \tag{5.78}$$

Nevertheless, we do not observe convergence to zero for $var(\hat{x}_{128})$ and $var(\mathcal{A})$. The only difference between $var(\mathcal{A})$ and $var(\mathcal{B})$ is the choice of recurrence vector in the second term. It makes a dramatic difference if one uses R_{12}^s instead R_{12}. Hence, we are able to pinpoint the problem with the theoretically constructed variance for MSSA output noise, which is the instability of the R_{12} recurrence vector under the perturbation ε.

Also, for this analysis we came across very unstable behaviour of all the variances, mentioned in Table 5.2, apart from the estimate of variance (5.63), based on the convolution approach. This instability is caused by the eigenvectors swap due to the fact, that

Tab. 5.2: MSSA Red Wine Forecast Measures

ρ	$\sigma^2\lvert R^s_{11} * q^1 + R^s_{12} * q^3\rvert$	$var(\mathcal{A})$ (5.75)	$var(\hat{x}_{128})$	$var(\mathcal{B})$ (5.76)
0.1	5.6975	6.6278	7.4768	6.5749
0.17	4.5769	**114.5941**	8.1574	7.1851
0.18	2.6169	**3146.2**	**3898.9**	**997.9766**
0.19	2.5575	**194.1092**	7.3854	3.6954
0.2	2.4359	**30.4935**	6.7346	3.2459
0.22	2.1989	9.5182	5.9248	3.1407
0.3	1.5222	5.1692	4.7625	1.9477
0.4	1.0386	4.5712	4.6748	1.2182
0.5	0.7341	4.371	4.7254	0.7938
0.6	0.5253	4.3115	4.8085	0.537
0.7	0.3792	4.3248	4.8949	0.3738
0.8	0.2764	4.3794	4.978	0.2663
0.9	0.2039	4.4572	5.0574	0.1934
1	0.1523	4.5463	5.1319	0.1429
1.1	0.1152	4.6383	5.1994	0.1073

corresponding eigenvalues are approximately the same. In this particular case, our cut-off point is 7 principal components, i.e. we use 7 principal components to obtain a forecast base. The eigenvalues for principal components 7 and 8 are very close to each other for $\rho = 0.18$, which is not the case for example for $\rho = 0.3$. We can see from Figure 5.49 that distributions of eigenvalues 7 and 8 are intersecting and in fact seem to be very alike. However for $\rho = 0.3$ eigenvalues 7 and 8 have no intersections in distribution whatsoever (see Figure 5.50).

This particular example with different multiplier in support series illustrates that the comparison of the variances is not always a fair test for causality. In order to see if the variances do not mislead in terms of causality it is useful to compare sets of variances obtained for different scaling.

Now by generating a support time series, such that it has no common trend or periodics, we look at the variances behaviour of red wime forecast when causality is not expected. For this example we have generated the following time series of the length $N = 187$ using MATLAB

$$y_n = 1000 \left(\sin \frac{1.61 \cdot 12n}{2\pi} \right) + 500 \tag{5.79}$$

The analyzed time series are illustrated on Figure 5.51.
In Table 5.3 we observe nearly opposite behaviour of the variances to the one shown in Table 5.2. The suggested forecast error variance (5.73) does not decrease as rapidly as in Table 5.2, moreover, its seems to stabilize at value 6.6. The variance $var(\mathcal{A})$ in Table 5.2 has a sharp increase from $\rho = 0.1$ to $\rho = 0.17$, then decreases with ρ increasing to 0.6 and slightly increase for ρ in range 0.6 to 1.1. However, we observe the increasing trend in Table 5.3. The forecast variance $var(\hat{x}_{128})$ for both examples, when support series is sparkling

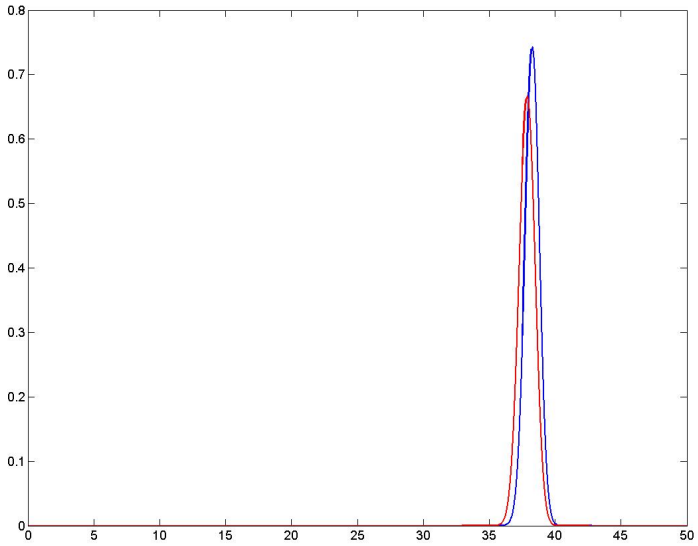


Fig. 5.49: Distributions of eigenvalues 7 (blue) and 8 (red). $\rho = 0.18$

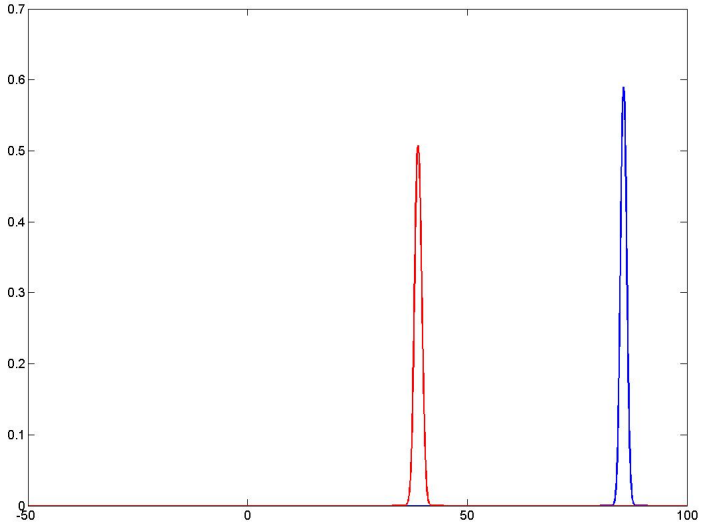


Fig. 5.50: Distributions of eigenvalues 7 (blue) and 8 (red). $\rho = 0.3$

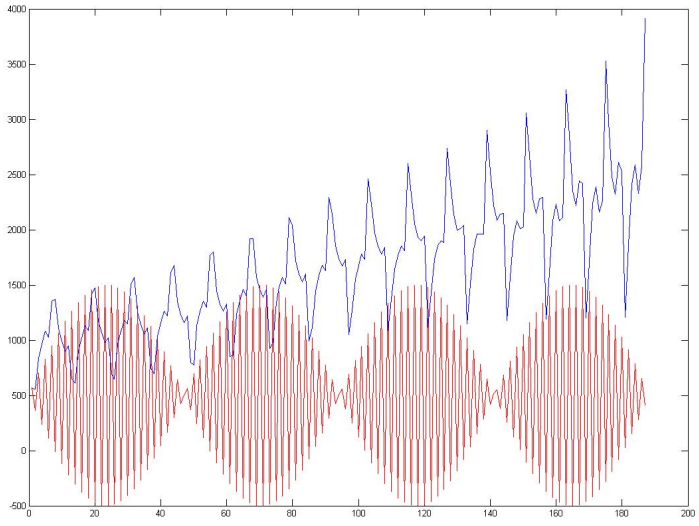


Fig. 5.51: Red Wine Time Series and (5.79) (red)

Tab. 5.3: MSSA Red Wine Forecast Measures with support series (5.79)

ρ	$\sigma^2\lvert R^s_{11} * q^1 + R^s_{12} * q^3\rvert$	$var(\mathcal{A})$ (5.75)	$var(\hat{x}_{128})$	$var(\mathcal{B})$ (5.76)
0.1	7.8661	8.4545	8.6397	7.8798
0.2	7.1994	7.952	8.1232	7.5153
0.3	6.7942	8.8551	7.7428	7.1967
0.4	6.6424	11.6653	7.5411	6.9916
0.5	6.5881	15.1699	7.4265	6.8687
0.6	6.5639	18.2412	7.3044	6.793
0.7	6.5341	22.2217	9.9216	6.7272
0.8	6.5747	22.6167	8.2548	6.7453
0.9	6.5809	24.0218	8.3286	6.7349
1	6.5853	25.0824	8.3883	6.7274
1.1	6.5888	25.8982	8.4366	6.7218
10	6.6064	30.0821	8.7161	6.6967
100	6.6067	30.1362	8.7201	6.6964

wine time series and simulated time series (5.79), oscillates but not dramatically. In Table 5.3 we see that $var(\mathcal{B})$ decreases rapidly while ρ increases, but for the counterexample (see Figure 5.51, Table 5.3) the variance $var(\mathcal{B})$ is stable.

The focus of this chapter was the relationship between the initial input noise of the time series and the resultant output noise after SSA/MSSA analysis. We studied the exact effects SSA mechanism has on the formation of the error in the obtained forecast.

We introduced the linearization of the SSA procedure, naturally to the first order, when the perturbation comes in the main series. The purpose of the linearization was to come to a better understanding of noise propagation through the SSA approach and see if it was possible to derive an expression, that connected noise in the initial time series with the error in the forecast. In its complete form, this expression tends to obscure statistical independence of input noise, which made further derivation difficult.

We studied the noise propagation at three SSA/MSSA stages: projector construction, time series reconstruction and forecast. At each stage we assessed the size of perturbation effect and checked if any of them are dominant. The analysis have shown that the perturbation coming through the perturbed eigenvectors can be neglected due its unsubstantial influence on the reconstruction, while the perturbation coming through the Hankel matrix plays a crucial role in reconstructions. To validate this point we did three stages stability analysis for simulated and real data.

Thus, for further stability analysis we assumed that the recurrence vectors are fixed, although some instability occurred for the recurrence vector R_{12}, which are revealed in Sections 5.8. Nevertheless, we assume that all recurrence vectors stay constant.

We analyzed the variance of the forecast of red wine time series with two different support series, sparkling wine support series and simulated support series (5.79). We expect causality in case when support series is sparkling wine, as visually their structures have similarities, however we do not expect any causality between red wine time series and time series (5.79). We studied the forecast variance from different angles by fixing the components in the forecast LRF, see (5.75, 5.76, 5.77). Curiously, the theoretical forecast variance, suggested in this chapter (5.73) has a homogeneous behaviour comparing to other variances and seem to be the only measure, which is not affected by the overlapping distributions of cut-off eigenvalue with the eigenvalue next to the cut-off one. However, in the counterexample, which is set so that there is no causality between time series, the variance (5.73) is stable for all chosen ρ and does not change even if the multiplier ρ is much bigger then 1 (see Table 5.3, values for $\rho = 10, 100$).

For these particular examples we observed that the introduced variance decreases rapidly with increasing multiplier if causality takes place and is stable if causality is absent. The comparison of two introduced theoretical variances, (5.63) and (5.73) may be a fair way to detect causality, as it is not affected by the possible eigenvectors swap, as was described above. Thus, the ratio of variances (5.63) and (5.73) is competitive to the ratio of variances used as F-test statistic.

To summarize last two paragraphs, on the basis of simplifying assumptions we derived a theoretical model for the output variance, which is based on the recurrence vectors and convolution. The implementation of this model showed the most stable and consistent

result, comparing to other suggested variants. Moreover, the comparison of the output variance (5.63) and (5.73) is competitive to the ratio of variances used as F-test statistic.

CONCLUSIONS

Achieved goals

This thesis is dedicated to the study of two major concepts in time series analysis and in econometrics, i.e. of Singular Spectrum Analysis, and of Granger's causality concept, which originally is applicable only to the stationary autoregressive model.

The goal of the work was to to combine Singular Spectrum Analysis technique with Granger's fundamental idea in such a way that it is possible to find a suitable measure for causality between two time series, without stationary autoregressive model constraints. Granger's concept of causality is founded upon the ability to forecast, so SSA seems to be a good starting point to build a new causality test, as it provides flexible tools for forecasting.

This study revealed that the statistical Granger's causality tests, which should be distinguished from Granger's theoretical concept of causality coherence, are sensitive to chosen parameters, such as the time lag used in the underlying linear regression model and time shifts in the support series. Moreover, the majority of Granger causality tests are closely related to the F-test, described in Chapter 3.

We applied the F-test, sign test and dominance tests to the quality of the MSSA forecast, compared with the SSA forecast, in a simulation study see if causality can be identified. F-test and sign test performance supported our expectations.

For the dominance tests we introduced indicator measures, which are useful to help decide which case of dominance occurs in a given data set.

In general, F-test, sign test and dominance tests along with direct distribution comparison did not contradict each other, and results in most cases were consistent.

These calculations also revealed a certain regular pattern, appearing between the loss function comparing the forecast with the real signal of the time series $S^{(1)}$, and the loss function comparing the forecast with the actual value of the time series $S^{(2)}$, i.e.

$$S^{(2)} \approx S^{(1)} + \sigma^2$$

with σ^2 being the variance of the noise in the main series.

We derived a novel approach to causality using Granger's idea and built a tentative causality measure modeled more closely on Granger's concept, but using SSA/MSSA which is suitable for time series outside the class of stationary autoregressive time series. This measure arises from the idea of so-called linearized MSSA, which was derived in order to avoid a fundamental scaling problem one comes across when dealing with standard MSSA.

The first order stability analysis of the SSA reconstruction and forecast algorithms under perturbations in the main time series showed that the noise component coming through

the perturbed eigenvectors can be neglected due to its small effect on the reconstruction, while the perturbation coming through the Hankel matrix has a dramatic impact on the reconstruction.

The real data study revealed that natural noise leads to a greater instability than Gaussian noise.

Thus, for further stability analysis we assumed that the recurrence vectors are fixed. This assumption was shown to be valid empirically in some simulated and real data tests. However, some remaining instability occurred for the MSSA partial recurrence vector R_{12}, as discussed and shown on examples in Sections 5.8. Nevertheless, we assume that all recurrence vectors stay constant in our theoretical analysis of stability under perturbation, in order to obtain a manageable formula.

This theoretical (5.73) forecast variance for MSSA has a smooth behaviour under scaling of the support series, comparing to other variances considered, and seems to be the only measure which is not affected by the overlapping distributions of cut-off eigenvalue corresponding to the eigentriple with the smallest weight included in reconstruction, with the eigenvalue next to the cut-off one. In the case of causality, the theoretical variance decreased markedly with increasing support series multiplier. However, in the non-causality case the variance (5.73) is approximately constant for all chosen support series multipliers, even if it is large.

We suggested the comparison of two introduced theoretical variances, (5.63) and (5.73) to be a suitable way to detect causality, as it is not affected by the possible eigenvectors swap, as was described in Chapter 5. The ratio of variances (5.63) and (5.73) could be a foundation for a competitive measure for causality detection.

Future work

In Chapter 4 we introduced so-called linearized MSSA. The background of this idea is following. If we consider causality to be phenomenon occurring between time series structures, we would expect that these structures do not change with regard to scaling. Nevertheless, if we run full MSSA procedure, we get different forecasts for the main series, depending non-linearly on the multiplier used in the support series; using the linearized MSSA we avoid this problem.

However, in the linearized MSSA we fully omit the autocovariance matrix $\mathbb{X}_2\mathbb{X}_2^{\intercal}$ obtained for the support series and to the analysis of the matrix

$$\begin{pmatrix} \mathbb{X}_1\mathbb{X}_1^{\intercal} & \mathbb{X}_1\mathbb{X}_2^{\intercal} \\ \mathbb{X}_2\mathbb{X}_1^{\intercal} & 0 \end{pmatrix},$$

from which arise following issues. First of all, by omitting the matrix $\mathbb{X}_2\mathbb{X}_2^{\intercal}$ we make it impossible to get a full reconstruction of the support series, which is substantial to obtain the MSSA forecast. Therefore, we we cannot do the stability analysis performed in Chapter 5. So far, we have got a good linear approximation of the recurrence vectors of the MSSA procedure. It is interesting to construct the full linearized MSSA mechanism and use it for further causality studies.

It is a challenging task to combine the ideas of the linearized MSSA and the convolution approach together with the approximate SSA suggested in paper [26].

One of the ideas in the paper [26] was to build an approximate projector without doing spectral analysis. The eigentriples in this case are given weight functions, so that if the explicit eigenvalues change places the projector still changes continuously.

The combination of these approaches may lead to a better understanding of noise propagation as well as to improve established theoretical models for resultant output variances.

The theoretical variance was built on the assumption that recurrence vectors and convolution kernels are stable and empirically we showed that the stability is valid for SSA recurrence vector R and R_{11}. In the future it is interesting to see to which extent the stability assumptions made in Chapter 5 do hold and whether there are any boundaries for their validity. It would be of a great importance for this question to know if these stability assumptions have a theoretical proof behind them.

One of possible future directions is to verify the range of validity of suggested simplifying assumptions of stability for recurrence vectors and convolution kernels for a wider range of time series.

Another possible future research can be dedicated to study of the instability of the recurrence vector R_{12}. It may be a sufficient contribution to stability analysis of the theoretical variance model to investigate the causes of R_{12} instability and may lead to a better, improved version of causality measure.

BIBLIOGRAPHY

[1] T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit" Criteria based on Stochastic Processes. *Ann. Math. Statistics*, 23:193–212, 1952.

[2] T. W. Anderson and D. A. Darling. A Test of Goodness of Fit. *Ann. Math. Statistics*, 32:765–769, 1954.

[3] L. N. Bol′shev and N. V. Smirnov. *Tablitsy Matematicheskoi Statistiki*. "Nauka", Moscow, third edition, 1983. With an afterword by Yu. V. Prokhorov and D. M. Chibisov, With a commentary and bibliography by D. S. Shmerling.

[4] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, fourth edition, 2008. Forecasting and control.

[5] P. T. Brandt. Markov-Switching, Bayesian, Vector Autoregression Models. *CRAN*, 2013.

[6] D. S. Broomhead and G. P. King. Extracting Qualitative Dynamics from Experimental Data. *Phys. D*, 20(2-3):217–236, 1986.

[7] D. S. Broomhead and G. P. King. On the Qualitative Analysis of Experimental Dynamical Systems. In *Nonlinear phenomena and chaos (Malvern, 1985)*, Malvern Phys. Ser., pages 113–144. Hilger, Bristol, 1986.

[8] V. P. Chistyakov. *Kurs Teorii Veroyatnostei*. "Nauka", Moscow, second edition, 1982. Cahiers Mathématiques Montpellier [Montpellier Mathematical Reports], 27.

[9] F. X. Diebold and R. Mariano. Comparing Predictive Accuracy. *Journal for Business and Economic Statistics*, 13(2-3):253–265, 1995.

[10] C. Diks and J. DeGoede. A General Nonparametric Bootstrap Test for Granger Causality. In *Global analysis of dynamical systems*, pages 391–403. Inst. Phys., Bristol, 2001.

[11] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky. *Analysis of time series structure*, volume 90 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, 2001. SSA and Related Techniques.

[12] N. Golyandina, A. Pepelyshev, and A. Steland. New Approaches to Nonparametric Density Estimation and Selection of Smoothing Parameters. *Comput. Statist. Data Anal.*, 56(7):2206–2218, 2012.

[13] N. Golyandina and D. Stepanov. SSA-based Approaches to Analysis and Forecast of Multidimensional Time Series. *Proceedings of the 5th St.Petersburg Workshop on Simulation*, pages 293–298, 2005.

[14] N. Golyandina and A. Zhigljavsky. *Singular Spectrum Analysis for Time Series*. Springer Briefs in Statistics. Springer, Heidelberg, 2013.

[15] C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–38, July 1969.

[16] C. W. J. Granger. Testing for Causality: a Personal Viewpoint. *J. Econom. Dynamics Control*, 2(4):329–352, 1980.

[17] C. W. J. Granger. Some Recent Developments in a Concept of Causality. *J. Econometrics*, 39(1-2):199–211, 1988.

[18] D. K. Guilkey and M. K. Salemi. Small Sample Properties of Three Tests for Granger-Causal Ordering in a Bivariate Stochastic System. *The Review of Economics and Statistics*, 64(4):668–80, 1982.

[19] J. D. Hamilton. *Time series analysis*. Princeton University Press, Princeton, NJ, 1994.

[20] H. Hassani, A. Zhigljavsky, K. Patterson, and A. S. Soofi. A Comprehesnive Causality Test based on the Singular Spectrum Analysis. *in Causality in the sciences*, pages 379–403, 2011. Causality.

[21] R. J. Hyndman. Time Series Online Data library. Monthly Australian Wine Sales.

[22] T. Kato. *Perturbation Theory for Linear Operators*. Die Grundlehren der mathematischen Wissenschaften, Band 132. Springer-Verlag New York, Inc., New York, 1966.

[23] F. Laio. Cramer-von Mises and Anderson-Darling Goodness of Fit Tests for Extreme Value Distributions with Unknown Parameters. *Water Resour. Res.*, 40, 2004.

[24] I. Markovsky and K. Usevich. Structured Low-Rank Approximation with Missing Data. *SIAM J. Matrix Anal. Appl.*, 34(2):814–830, 2013.

[25] R. Mendezes, A. Dionisio, and H. Hassani. On the Globalization of Stock Markets: an Application of VECM, SSA technique and Mutual Information to the G7. *Preprinted*, 2010.

[26] V. Moskvina and K. M. Schmidt. Approximate Projectors in singular Spectrum Analysis. *SIAM J. Matrix Anal. Appl.*, 24(4):932–942 (electronic), 2003.

[27] R. F. Nau. *Online Course*.

[28] V. Nekrutkin. Perturbation Expansions of Signal Subspaces for Long Signals. *Stat. Interface*, 3(3):297–319, 2010.

[29] K. Patterson, H. Hassani, S. Heravi, and A. Zhigljavsky. Multivariate Singular Spectrum Analysis for Forecasting Revisions to Real-time Data. *J. Appl. Stat.*, 38(10):2183–2211, 2011.

[30] D. S. G. Pollock. *A Handbook of Time Series Analysis, Signal Processing and Dynamics*. Signal Processing and its Applications. Academic Press Inc., San Diego, CA, 1999. With 1 CD-ROM (Windows, Macintosh and UNIX).

[31] I.L. Schiff. *Quantum Mechanics*. McGraw-Hill Book Company, third edition, 1986.

[32] C. Sims. Money, Income, and Causality. *American Economic Review*, 62(4):540–52, Sept 1972.

[33] T. J. Harris Y. Hui. Filtering and Frequency Interpretation of Singular Spectrum Analysis. *Physica D*, 239:1958–1967, 2010.

[34] A. Zhigljavsky and V. Solntsev. *Oral Care Statistical Studies*. University of St. Petersburg, 1995.