SPRINGER BRIEFS IN STATISTICS

Nina Golyandina
Anatoly Zhigljavsky

Singular Spectrum Analysis for Time Series

Springer

SpringerBriefs in Statistics

For further volumes:
http://www.springer.com/series/8921

Nina Golyandina · Anatoly Zhigljavsky

Singular Spectrum Analysis for Time Series

Springer

Nina Golyandina
Department of Mathematics
St. Petersburg University
St. Petersburg
Russia

Anatoly Zhigljavsky
School of Mathematics
Cardiff University
Cardiff
UK

ISSN 2191-544X
ISBN 978-3-642-34912-6
DOI 10.1007/978-3-642-34913-3
Springer Heidelberg New York Dordrecht London

ISSN 2191-5458 (electronic)
ISBN 978-3-642-34913-3 (eBook)

Library of Congress Control Number: 2012953018

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Chapter 1
Introduction

1.1 Preliminaries

Singular spectrum analysis (SSA) is a technique of time series analysis and forecasting. It combines elements of classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing. SSA aims at decomposing the original series into a sum of a small number of interpretable components such as a slowly varying trend, oscillatory components and a 'structureless' noise. It is based on the singular value decomposition (SVD) of a specific matrix constructed upon the time series. Neither a parametric model nor stationarity-type conditions have to be assumed for the time series. This makes SSA a model-free method and hence enables SSA to have a very wide range of applicability.

The present book is fully devoted to the methodology of SSA. It exhibits the huge potential of SSA and shows how to use SSA both safely and with maximum effect.

Potential readers of the book. (a) Professional statisticians and econometricians; (b) specialists in any discipline where problems of time series analysis and forecasting occur; (c) specialists in signal processing and those needed to extract signals from noisy data; (d) PhD students working on topics related to time series analysis; (e) students taking appropriate MSc courses on applied time series analysis; (f) anyone interested in the interdisciplinarity of statistics and mathematics.

Historical remarks. The first publication, which can be considered as one of the origins of SSA (and more generally of the subspace-based methods of signal processing), can be traced back to the eighteenth century [28].

The commencement of SSA is usually associated with publication in 1986 of the papers [4, 5] by Broomhead and King. Since then SSA has received a fair amount of attention in literature. Additionally to [4, 5] the list of most cited papers on SSA published in the 1980s and 1990s includes [2, 10, 32, 33].

There are three books fully devoted to SSA, [8, 9, 14]. The book [9] is well written but it only provides a very elementary introduction to SSA. The volume [8] is a collection of papers written entirely by statisticians based at that time at St.Petersburg university. All these papers are devoted to the so-called 'Caterpillar'

N. Golyandina and A. Zhigljavsky, *Singular Spectrum Analysis for Time Series*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-34913-3_1,

methodology (the words 'Caterpillar' or 'Gusenitsa' is due to the association with the moving window). This methodology is a version of SSA that was developed in the former Soviet Union independently (the 'iron curtain effect') of the mainstream SSA. The work on the 'Caterpillar' methodology has started long after publication of [28] but well before 1986, the year of publication of [4] and [5].

The main difference between the main-stream SSA of [2, 4, 5, 10, 32, 33] and the 'Caterpillar' SSA is not in the algorithmic details but rather in the assumptions and in the emphasis in the study of SSA properties. To apply the mainstream SSA, one often needs to assume some kind of stationarity of the time series and think in terms of the 'signal plus noise' model (where the noise is often assumed to be 'red'). In the 'Caterpillar' SSA, the main methodological stress is on separability (of one component of the series from another one) and neither the assumption of stationarity nor the model in the form 'signal plus noise' are required.

The main methodological principles described in [8] have been further developed in the monograph [14]. The publication of [14] has helped to attract much wider attention to SSA from the statistical circles as well as many other scientific communities. During the last 10 years much new SSA-related research has been done and many new successful applications of SSA have been reported. A recent special issue of 'Statistics and Its Interface' [35] gives an indication of how much progress in theoretical and methodological developments of SSA, as well as its applications, has been achieved in recent years. The SSA community regularly organizes international workshops on SSA. The latest SSA workshop was held in Beijing in May 2012, see http://www.cefs.ac.cn/express/SSA.html.

The research on the theory and methodology of SSA performed in the last two decades has resulted in a rather pleasing state of affairs: (i) the existence of an active SSA community and (ii) the existence of a general methodology of SSA rather than simply a collection of many different SSA algorithms. This methodology unifies different versions of SSA into a very powerful tool of time series analysis and forecasting. Description of SSA methodology is the sole purpose of the present book.

Correspondence between the present book and [14]. Some entirely new topics are included (for example, Sect. 3.7–3.9) but a few topics thoroughly described in [14] are not considered at all (see, for example, [14, Chap. 3]). This volume is fully devoted to the methodology of SSA unlike [14], where many theoretical issues were also considered. The material is correspondingly revised in view of the new objectives. The main aim of [14] is to establish SSA as a serious subject. There is no need to do it now and the aspiration of this book is to show the power and beauty of SSA to as wide audience as possible.

Several reasons why SSA is still not very popular among statisticians. First reason is tradition: SSA is not a classical statistical method, and therefore many people are simply not aware of it. Second, SSA demands more computing power than the traditional methods.Third, many people prefer model-based statistical techniques where calculations are automatic and do not require the computer-analyst interaction. Finally, SSA is sometimes too flexible (especially when analyzing multivariate series) and therefore has too many options which are difficult to formalize.

Links between SSA and other methods of time series analysis. SSA has no links with ARIMA, GARCH and other methods of this type and also with wavelets. However, SSA has very close links with some methods of signal processing and with methods of multivariate statistics like principal component analysis and projection pursuit; see Sect. 2.5.4 and 3.8.The so-called Empirical Mode Decomposition (EMD), see [20], is intended to solve similar problems to SSA but there are significant conceptual and methodological differences between SSA and EMD.

Structure of the next three sections. In the next section, we give a short introduction into SSA methodology and simultaneously into the material of the present book. Then we mention important issues related to SSA, which did not find their way into this book. Finally, we provide a list of most common symbols and acronyms.

Acknowledgements. The authors are very much indebted to Vladimir Nekrutkin, their coauthor of the monograph [14]. His contribution to the methodology and especially theory of SSA cannot be underestimated. The authors very much acknowledge many useful comments made by Jon Gillard. The authors are also grateful to former and current Ph.D. students and collaborators of Nina Golyandina: Konstantin Usevich (a specialist in algebraic approach to linear recurrence relations), Theodore Alexandrov (automatic SSA), Andrey Pepelyshev (SSA for density estimation), Anton Korobeynikov (fast computer implementation of SSA), Eugene Osipov and Marina Zhukova (missing data imputation), and Alex Shlemov (SSA filtering). Help of Alex Shlemov in preparation of figures is very much appreciated.

1.2 SSA Methodology and the Structure of the Book

The present volume has two chapters. In Chap. 2, SSA is typically considered as a model-free method of time series analysis. The applications of SSA dealt with in Chap. 3 (including forecasting) are model based and use the assumption that the components of the original time series extracted by SSA satisfy linear recurrence relations.

The algorithm of Basic SSA (Sect. 2.1). A condensed version of Basic SSA (which is the main version of SSA) can be described as follows.

Let $\mathbb{X}_N = (x_1, \ldots, x_N)$ be a time series of length N. Given a window length L $(1 < L < N)$, we construct the L-lagged vectors $X_i = (x_i, \ldots, x_{i+L-1})^{\mathrm{T}}$, $i = 1, 2, \ldots, K$, where $K = N - L + 1$, and compose these vectors into the trajectory matrix $\mathbf{X}$.

The columns X_j of $\mathbf{X}$ can be considered as vectors in the L-dimensional space R^L. The eigendecomposition of the matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ (equivalently, the SVD of the matrix $\mathbf{X}$) yields a collection of L eigenvalues and eigenvectors. A particular combination of a certain number r of these eigenvectors determines an r-dimensional subspace $\mathcal{L}_r$ in R^L, $r < L$. The L-dimensional data $\{X_1, \ldots, X_K\}$ is then projected onto the subspace $\mathcal{L}_r$ and the subsequent averaging over the diagonals yields some Hankel matrix $\widetilde{\mathbf{X}}$. The time series $(\tilde{x}_1, \ldots, \tilde{x}_N)$, which is in the one-to-one correspondence

with matrix $\widetilde{\mathbf{X}}$, provides an approximation either the whole series $\mathbb{X}_N$ or a particular component of $\mathbb{X}_N$.

Basic SSA and models of time series (Sect. 2.3). As a non-parametric and model-free method, Basic SSA can be applied to any series. However, for interpreting results of analysis and making decisions about the choice of parameters some models may be useful. The main assumption behind Basic SSA is the assumption that the time series can be represented as a sum of different components such as trend (which we define as any slowly varying series), modulated periodicities, and noise. All interpretable components can be often approximated by time series of small rank, and hence can be described via certain LRRs (linear recurrence relations). Separating the whole series into these components and analysis of the LRRs for interpretable components helps in getting reliable and useful SSA results.

Potential of Basic SSA (Sect. 2.2 and also Sect. 3.7, 3.8 and 3.9). The list of major tasks, which Basic SSA can be used for, includes smoothing, noise reduction, extraction of trends of different resolution, extraction of periodicities in the form of modulated harmonics, estimation of volatility, etc. These tasks are considered in Sect. 2.2. The following more advanced abilities (but model-based) of SSA are considered in the final three sections of Chap. 3: the use of SSA for filling in missing values is considered in Sect. 3.7; add-ons to Basic SSA permitting estimation of signal parameters are considered in Sect. 3.8; finally, Basic SSA as a filtration tool is studied in Sect. 3.9. Methodologically, these last three topics are closely linked with the problem of SSA forecasting. Note also that all major capabilities of Basic SSA are illustrated on real-life time series.

Choice of parameters in Basic SSA (Sect. 2.4). There are two parameters to choose in Basic SSA: the window length L and the group of r indices which determine the subspace $\mathcal{L}_r$. A rational or even optimal choice of these parameters should depend on the task we are using SSA for. The majority of procedures require interactive (including visual) identification of components. An automatic choice of parameters of Basic SSA could be made, see Sect. 2.4.5. However, the statistical procedures for making this choice are modelbased. Success in using the corresponding versions of SSA depends on the adequacy of the assumed models and especially on achieving good separability of the time series components.

Toeplitz SSA (Sect. 2.5.3). Basic SSA can be modified and extended in many different ways, see Sect. 2.5. As a frequently used modification of Basic SSA, consider a common application of SSA for the analysis of stationary series, see Sect. 2.5.3. Under the assumption that the series $\mathbb{X}_N$ is stationary, the matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ of Basic SSA can be replaced with the so-called lag-covariance matrix $\mathbf{C}$ whose elements are $c_{ij} = \frac{1}{N-k}\sum_{t=1}^{N-k} x_t x_{t+k}$ with $i, j = 1, \ldots, L$ and $k = |i - j|$. In the book, this version of SSA is called 'Toeplitz SSA'.[1] Unsurprisingly, if the original series is stationary then Toeplitz SSA slightly outperforms Basic SSA. However, if the series

[1] In the literature on SSA, Basic SSA is sometimes called BK SSA and what we call 'Toeplitz SSA' is called VG SSA; here BK and VG stand for Broomhead & King [4, 5] and Vautard & Ghil [32], respectively.

is not stationary then the use of Toeplitz SSA may yield results which are simply wrong.

SSA–ICA (Sect. 2.5.4). Another important modification of Basic SSA can be viewed as a combination of SSA and Independent Component Analysis (ICA), see Sect. 2.5.4. This algorithm, called SSA–ICA, helps to separate time series components that cannot be separated with the help of the SVD alone, due to the lack of strong separability. Despite we deal with deterministic time series components but ICA is developed for dealing with random series and processes, the algorithm of ICA can be formally applied for achieving a kind of independence of components. Note that it is not a good idea to use the ICA as a full replacement of the SVD since the ICA is a much less stable procedure than the SVD. Therefore, a two-stage procedure is proposed, where the SVD is performed for the basic decomposition and then some version of the ICA (or the projection pursuit) is applied to those components which are produced by the SVD but remain mixed up.

Computational aspects of SSA (Sects. 2.5.6 and 2.5.7). If L is very large then the conventional software performing SVD decomposition may be computationally costly. In this case, the Partial SVD and other techniques can be used for performing fast computations. This can be achieved either by using clever implementations (Sect. 2.5.6) or by replacing the SVD with simpler procedures (Sect. 2.5.7).

SSA forecasting (Sects. 3.1 – 3.6). Time series forecasting is an area of huge practical importance and Basic SSA can be very effective for forecasting. The main idea of SSA forecasting is as follows.

Assume that $\mathbb{X}_N = \mathbb{X}_N^{(1)} + \mathbb{X}_N^{(2)}$ and we are interested in forecasting of $\mathbb{X}_N^{(1)}$. If $\mathbb{X}_N^{(1)}$ is a time series of finite rank, then it generates some subspace $\mathcal{L}_r \subset \mathsf{R}^L$. This subspace reflects the structure of $\mathbb{X}_N^{(1)}$ and can be taken as a base for forecasting. Under the conditions of separability between $\mathbb{X}_N^{(1)}$ and $\mathbb{X}_N^{(2)}$ (these conditions are discussed throughout the volume; see, for example, Sects. 2.3.3, 2.4, 3.3.1 and 3.5), Basic SSA is able to accurately approximate $\mathcal{L}_r$ and hence it yields an LRR which approximates the true LRR and can be directly used as a forecasting formula. This method of forecasting is called recurrent forecasting and considered in Sect. 3.3, as well as few other sections. Alternatively, we may use the so-called 'vector forecasting' which main idea is in the consecutive construction of the vectors $X_i = (x_i, \dots, x_{i+L-1})^{\mathrm{T}}$, for $i = K+1, K+2, \dots$ so that they lie as close as possible to the subspace $\mathcal{L}_r$.

Short-term forecasting makes very little use of the model while responsible forecasting for long horizons is only possible when an LRR is built by SSA and the adequacy of this LRR is testified. As demonstrated in Sect. 3.3, in addition to the LRRs, SSA forecasting methods use the characteristic polynomials associated with these LRRs. The precision of SSA forecasting formulas depends on the location of the roots of these polynomials. In Sect. 3.2 we provide an overview of the relations between LRRs, the characteristic polynomials and their roots and discuss properties of the so-called min-norm LRRs which are used for estimating parameters of the signal (see Sect. 3.8), in addition to forecasting.

In forecasting methodology, the construction of confidence intervals for the forecasts is often an essential part of the procedure. Construction of these intervals for

SSA forecasts is discussed in Sect. 3.4. Despite SSA itself being a model-free technique, for building confidence intervals we need to make certain assumptions such as that the residual series is a realization of a stochastic white noise process.

In Sect. 3.5 we give recommendations on the choice of forecasting parameters and in Sect. 3.6 we discuss results of a case study. We argue that stability of forecasts is the major aim we have to try to achieve in the process of building forecasts. Forecast stability is highly related to the forecast precision and forecast reliability.

SSA for missing value imputation (Sect. 3.7). Forecasting can be considered as a special case of missing value imputation if we assume that the missing values are located at the end of the series. We show how to extend some SSA forecasting procedures (as well as methods of their analysis) to this more general case.

Parameter estimation in SSA and signal processing (Sect. 3.8). Although there are many similarities between SSA and the subspace-based methods of signal processing, there is also a fundamental difference between these techniques. This difference lies in the fact that the model of the form 'signal plus noise' is obligatory in signal processing; consequently, the main aim of the signal processing methods is the estimation of the parameters of the model (which is usually the sum of damped sinusoids). The aims of SSA analysis are different (for instance, splitting the series into components or simply forecasting) and the parameters of the approximating time series are of secondary importance. This fundamental difference between the two approaches leads, for example, to different recommendations for the choice of the window length L: a typical recommendation in Basic SSA is to choose L reasonably large while in the signal processing methods L is typically relatively small.

Causal SSA (Sect. 3.9.5). Causal SSA (alternatively, Last Point SSA) can be considered as an alternative to forecasting. In Causal SSA, we assume that the points in the time series $\mathbb{X}_\infty = (x_1, x_2, \ldots)$ arrive sequentially, one at a time. Starting at some $M_0 > 0$, we apply Basic SSA with fixed window length and the grouping rule to the series $\mathbb{X}_M = (x_1, \ldots, x_M)$ for all $M \geq M_0$. We then monitor how SSA reconstruction of previously obtained points of the series change as we increase M (this is called redrawing). The series consisting of the last points of the reconstructions is the result of Causal SSA. The delay of the Causal SSA series reflects the quality of forecasts based on the last points of the reconstructions. Additionally to the redrawings of the recent points of the reconstructions, this delay can serve as an important indicator of the proper choice of the window length, proper grouping and in general, predictability of the time series. This could be of paramount importance for the stock market traders when they try to decide whether a particular stock is consistently decreasing/increasing its value or they only observe market fluctuations.

1.3 SSA Topics Outside the Scope of This Book

Theory of SSA. For the basic theory of SSA we refer to the monograph [14]. Since the publication of that book, several influential papers on theoretical aspects of SSA have been published. The main theoretical paper on perturbations in SSA and

subspace-based methods of signal processing is [26]. Another important theoretical paper is [31], where the concept of SSA separability is further developed (relative to [14]) and studied through the apparatus of the roots of characteristic polynomials of the linear recurrence relations of SSA approximation of the signal (in the 'signal plus noise' model). Elements of the theory of SSA are also discussed in [12].

SSA for change-point detection and subspace tracking. Assume that the observations $x_1, x_2, \ldots$ of the series arrive sequentially in time and we apply Basic SSA to the observations at hand. Then we can monitor the distances from the sequence of the trajectory matrices to the r-dimensional subspaces we construct and also the distances between these r-dimensional subspaces. Significant changes in any of these distances may indicate a change in the mechanism generating the time series. Note that this change in the mechanism does not have to affect the whole structure of the series but rather only a few of its components. For some references we refer to [14, Chap. 3] and [25] and the website http://www.cf.ac.uk/maths/subsites/stats/changepoint/. Recently, the method developed in [25] has found applications in robotics, see [24].

Monte-Carlo SSA. A typical SSA assumption about the noise in the signal plus noise model is the association of noise with a structure-less series (that is, a series which cannot be well approximated by a time series of finite rank). If we assume that the noise is stochastic and red (that is, AR(1) model) then the so-called Monte Carlo SSA is a common technique. In this version of SSA, special tests based on the Monte Carlo simulations are devised for testing the hypothesis of the presence of a weak signal on the background of a large noise, see [2].

SSA for density estimation. As shown in [15], SSA could be used for non-parametric density estimation and can produce estimates that are more accurate than the celebrated Kernel density estimates.

SSA for multivariate time series. Multivariate (or multichannel) SSA (shortly, MSSA) [8, 9] is a direct extension of the standard SSA for simultaneous analysis of several time series. Assume that we have two series, $\mathbb{X}_N = (x_1, \ldots, x_N)$ and $\mathbb{Y}_N = (y_1, \ldots, y_N)$. Let $\mathbf{X}$ be the trajectory matrix of the series $\mathbb{X}_N$ and $\mathbf{Y}$ be the trajectory matrix of $\mathbb{Y}_N$ (both matrices have size $L \times K$). Then the (joint) trajectory matrix of the two-variate series $(\mathbb{X}_N, \mathbb{Y}_N)$ can be defined as either $\mathbf{Z} = [\mathbf{X} : \mathbf{Y}]$ (matrix of size $L \times 2K$) or $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$ (matrix of size $2L \times K$). Matrix $\mathbf{Z}$ is block-Hankel rather than simply Hankel. Other stages of MSSA are identical to the ones of the univariate SSA except that we build a block-Hankel (rather than ordinary Hankel) approximation $\widetilde{\mathbf{Z}}$ to the trajectory matrix $\mathbf{Z}$.

MSSA may be very useful for analyzing several series with common structure, see http://www.gistatgroup.com/gus/mssa2.pdf. MSSA could also be used for establishing a causality between two series. Following the lines of Granger [16] we say that the absence of causality of $\mathbb{Y}_N$ on $\mathbb{X}_N$ means that the knowledge of $\mathbb{Y}_N$ does not improve the quality of forecasts of $\mathbb{X}_N$. The MSSA causality is discussed in [18, 19].

2D-SSA for image processing. 2D-SSA is a straightforward extension of Basic SSA and MSSA for analyzing images. The only difference between these three

versions of SSA is in the construction of the trajectory matrix, see [8, 13, 29]. Note however that the moving window in 2D-SSA is a rectangle and the window length becomes a product of two numbers. This implies that the size of the trajectory matrix could be very large and a clever implementation of the SVD becomes essential.

Comparison of SSA with other techniques. SSA is compared with some subspace-based techniques of signal processing in [12, 26] and [11]. Numerical comparison of SSA with ARIMA and other classical methods of time series analysis can be found in several papers of the volume [35] and in many papers devoted to applications of SSA, see for example [3, 17, 21, 27, 30].

Application areas. SSA has proved to be very useful and has become a standard tool in the analysis of climatic, meteorological and geophysical time series; see, for example, [10, 32] (climatology), [34] (meteorology), [7] (marine science), [22] (geophysics); for more references, see [1, 2, 8–10, 14, 22, 32, 33] and the papers in [35]. More recent areas of application of SSA include engineering, image processing, medicine, actuarial science and many other fields; for references see, for example, [3, 23, 24] and various papers in [35]. A special case is econometrics where SSA was basically unknown only a few years ago but recently it has made a tremendous advancement and is becoming more and more popular; see, for example, [6, 17–19, 27].

1.4 Common Symbols and Acronyms

SVD	singular value decomposition
LRR	linear recurrence relation
SSA	Singular Spectrum Analysis
$\mathbb{X}$ or $\mathbb{X}_N$	time series
$\mathbb{X}_N = (x_1, \dots, x_N)$	time series of length N
$\mathbb{X}_\infty = (x_1, x_2, \dots)$	infinite time series
N	length of time series
L	window length
$K = N - L + 1$	the number of L-lagged vectors obtained from $\mathbb{X}_N$
$X_i = (x_i, \dots, x_{i+L-1})^{\mathrm{T}}$	ith L-lagged vector obtained from $\mathbb{X}_N$
$\mathbf{X} = [X_1 : \dots : X_K]$	trajectory matrix with columns X_i
$\|\|\mathbf{X}\|\|_{\mathrm{F}}$	Frobenius matrix norm
$\operatorname{rank} \mathbf{X}$	rank of the matrix $\mathbf{X}$
$\mathcal{X} = \mathcal{X}^{(L)}(\mathbb{X}_N)$	L-trajectory space of a time series $\mathbb{X}_N$
$\operatorname{rank}_L(\mathbb{X}_N)$	L-rank of a time series $\mathbb{X}_N$
$\mathscr{H}$	hankelization operator
λ_i	ith eigenvalue of the matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$
U_i	ith eigenvector of the matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$

$V_i = \mathbf{X}^{\mathrm{T}} U_i / \sqrt{\lambda_i}$	ith factor vector of the matrix $\mathbf{X}$
$(\sqrt{\lambda_i}, U_i, V_i)$	ith eigentriple of the SVD of the matrix $\mathbf{X}$
$\mathbf{I}_{\mathrm{M}}$	identity $M \times M$ matrix
R^L	Euclidean space of dimension L
$\mathfrak{L}_r$	r-dimensional linear subspace of R^L
$\mathrm{span}(P_1, \ldots, P_n)$	linear subspace spanned by vectors $P_1, \ldots, P_n$
$\rho^{(w)}$	weighted correlation between two series
Π_x^N	periodogram of a time series $\mathbb{X}_N$

References

1. Alexandrov T, Golyandina N, Spirov A (2008) Singular spectrum analysis of gene expression profiles of early drosophila embryo: Exponential-in-distance patterns. Res Lett in Signal Process 2008:1–5
2. Allen M, Smith L (1996) Monte Carlo SSA: Detecting irregular oscillations in the presence of colored noise. J Clim 9(12):3373–3404
3. Azulay DO, Brain P, Sultana SR (2011) Characterisation of very low frequency oscillations in laser Doppler perfusion signals with a singular spectrum analysis. Microvasc Res 81(3):239–244
4. Broomhead D, King G (1986) Extracting qualitative dynamics from experimental data. Physica D 20:217–236
5. Broomhead D, King G (1986b) On the qualitative analysis of experimental dynamical systems. In: Sarkar S (ed) Nonlinear Phenomena and Chaos. Adam Hilger, Bristol, pp 113–144
6. de Carvalho M, Rodrigues PC, Rua A (2012) Tracking the US business cycle with a singular spectrum analysis. Econ Lett 114(1):32–35
7. Colebrook JM (1978) Continuous plankton records - zooplankton and evironment, northeast Atlantic and North Sea, 1948–1975. Oceanol Acta 1:9–23
8. Danilov D, Zhigljavsky A (eds) (1997) Principal components of time series: the "Caterpillar" method. St.Petersburg Press, St. Petersburg (in Russian)
9. Elsner JB, Tsonis AA (1996) Singular spectrum analysis: a new tool in time series analysis. Plenum, New York
10. Fraedrich K (1986) Estimating dimensions of weather and climate attractors. J Atmos Sci 43:419–432
11. Gillard J (2010) Cadzow's basic algorithm, alternating projections and singular spectrum analysis. Stat Interface 3(3):335–343
12. Golyandina N (2010) On the choice of parameters in singular spectrum analysis and related subspace-based methods. Stat Interface 3(3):259–279
13. Golyandina N, Usevich K (2010) 2D-extension of singular spectrum analysis: algorithm and elements of theory. In: Olshevsky V, Tyrtyshnikov E (eds) Matrix methods: theory, algorithms and applications. World Scientific Publishing, Algorithms and Applications, pp 449–473
14. Golyandina N, Nekrutkin V, Zhigljavsky A (2001) Analysis of time series structure: SSA and related techniques. Chapman&Hall/CRC, Boca Raton
15. Golyandina N, Pepelyshev A, Steland A (2012) New approaches to nonparametric density estimation and selection of smoothing parameters. Comput Stat Data Anal 56(7):2206–2218
16. Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37(3):424–438
17. Hassani H, Heravi S, Zhigljavsky A (2009) Forecasting European industrial production with singular spectrum analysis. Int J Forecast 25(1):103–118
18. Hassani H, Zhigljavsky A, Patterson K, Soofi AS (2011) A comprehensive causality test based on the singular spectrum analysis. In: Russo F, Williamson J (eds) Illari PM. Oxford University press, Causality in the Sciences, pp 379–404

19. Hassani H, Heravi S, Zhigljavsky A (2012) Forecasting UK industrial production with multivariate singular spectrum analysis. J Forecast. doi:10.1002/for.2244
20. Huang N et al (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. R Soc Lond Proc Ser A 454:903–995
21. Kapl M, Mueller W (2010) Prediction of steel prices: a comparison between a convential regression model and MSSA. Stat Interface 3(3):369–375
22. Kondrashov D, Ghil M (2006) Spatio-temporal filling of missing points in geophysical data sets. Nonlinear Process, Geophys 13(2):151–159
23. Mahecha MD, Fürst LM, Gobron N, Lange H (2010) Identifying multiple spatiotemporal patterns: A refined view on terrestrial photosynthetic activity. Pattern Recogn Lett 31(14):2309–2317
24. Mohammad Y, Nishida T (2011) On comparing SSA-based change point discovery algorithms. IEEE SII pp 938–945
25. Moskvina V, Zhigljavsky A (2003) An algorithm based on singular spectrum analysis for change-point detection. Commun Stat Simul Comput 32:319–352
26. Nekrutkin V (2010) Perturbation expansions of signal subspaces for long signals. Stat Interface 3:297–319
27. Patterson K, Hassani H, Heravi S, Zhigljavsky A (2011) Multivariate singular spectrum analysis for forecasting revisions to real-time data. J of Appl Stat 38(10):2183–2211
28. de Prony G (1795) Essai expérimental et analytique sur les lois de la dilatabilité des fluides élastiques et sur celles de la force expansive de la vapeur de l'eau et la vapeur de l'alkool à différentes températures. J de l'Ecole Polytechnique 1(2):24–76
29. Rodríguez-Aragón L, Zhigljavsky A (2010) Singular spectrum analysis for image processing. Stat Interface 3(3):419–426
30. Tang Tsz Yan V, Wee-Chung L, Hong Y (2010) Periodicity analysis of DNA microarray gene expression time series profiles in mouse segmentation clock data. Stat Interface 3(3):413–418
31. Usevich K (2010) On signal and extraneous roots in singular spectrum analysis. Stat Interface 3(3):281–295
32. Vautard M, Ghil M (1989) Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. Physica D 35:395–424
33. Vautard R, Yiou P, Ghil M (1992) Singular-spectrum analysis: a toolkit for short, noisy chaotic signals. Physica D 58:95–126
34. Weare BC, Nasstrom JS (1982) Examples of extended empirical orthogonal function analyses. Mon Weather Rev 110(6):481485
35. Zhigljavsky A (Guest Editor) (2010) Special issue on theory and practice in singular spectrum analysis of time series. Stat Interface 3(3)

Chapter 2
Basic SSA

2.1 The Main Algorithm

2.1.1 Description of the Algorithm

Consider a real-valued time series $\mathbb{X} = \mathbb{X}_N = (x_1, \ldots, x_N)$ of length N. Assume that $N > 2$ and $\mathbb{X}$ is a nonzero series; that is, there exists at least one i such that $x_i \neq 0$. Let L $(1 < L < N)$ be some integer called *the window length* and $K = N - L + 1$.

Basic SSA is an algorithm of time series analysis which is described below. This algorithm consists of two complementary stages: decomposition and reconstruction.

2.1.1.1 First Stage: Decomposition

1st step: Embedding
To perform the *embedding* we map the original time series into a sequence of lagged vectors of size L by forming $K = N - L + 1$ *lagged vectors*

$$X_i = (x_i, \ldots, x_{i+L-1})^{\mathrm{T}} \quad (1 \le i \le K)$$

of size L. If we need to emphasize the size (dimension) of the vectors X_i, then we shall call them *L-lagged vectors*.

The *L-trajectory matrix* (or simply *the trajectory matrix*) of the series $\mathbb{X}$ is

$$\mathbf{X} = [X_1 : \ldots : X_K] = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} x_1 & x_2 & x_3 & \ldots & x_K \\ x_2 & x_3 & x_4 & \ldots & x_{K+1} \\ x_3 & x_4 & x_5 & \ldots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \ldots & x_N \end{pmatrix}. \qquad (2.1)$$

N. Golyandina and A. Zhigljavsky, *Singular Spectrum Analysis for Time Series*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-34913-3_2,

The lagged vectors X_i are the columns of the trajectory matrix $\mathbf{X}$. Both, the rows and columns of $\mathbf{X}$ are subseries of the original series.

The (i, j)th element of the matrix $\mathbf{X}$ is $x_{ij} = x_{i+j-1}$ which yields that $\mathbf{X}$ has equal elements on the 'antidiagonals' $i + j = \text{const}$. (Hence the trajectory matrix is a *Hankel matrix*.) Formula (2.1) defines a one-to-one correspondence between the trajectory matrix of size $L \times K$ and the time series.

2nd step: Singular value decomposition (SVD)

At this step, we perform the singular value decomposition (SVD) of the trajectory matrix $\mathbf{X}$. Set $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ and denote by $\lambda_1, \ldots, \lambda_L$ the *eigenvalues* of $\mathbf{S}$ taken in the decreasing order of magnitude ($\lambda_1 \geq \ldots \geq \lambda_L \geq 0$) and by $U_1, \ldots, U_L$ the orthonormal system of the *eigenvectors* of the matrix $\mathbf{S}$ corresponding to these eigenvalues.

Set $d = \operatorname{rank} \mathbf{X} = \max\{i, \text{such that } \lambda_i > 0\}$ (note that in real-life series we usually have $d = L^*$ with $L^* = \min\{L, K\}$) and $V_i = \mathbf{X}^{\mathrm{T}} U_i / \sqrt{\lambda_i}$ $(i = 1, \ldots, d)$. In this notation, the SVD of the trajectory matrix $\mathbf{X}$ can be written as

$$\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d, \tag{2.2}$$

where $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$. The matrices $\mathbf{X}_i$ have rank 1; such matrices are sometimes called *elementary matrices*. The collection $(\sqrt{\lambda_i}, U_i, V_i)$ will be called ith *eigentriple* (abbreviated as ET) of the SVD (2.2).

2.1.1.2 Second Stage: Reconstruction

3rd step: Eigentriple grouping

Once the expansion (2.2) is obtained, the grouping procedure partitions the set of indices $\{1, \ldots, d\}$ into m disjoint subsets $I_1, \ldots, I_m$.

Let $I = \{i_1, \ldots, i_p\}$. Then the resultant matrix $\mathbf{X}_I$ corresponding to the group I is defined as $\mathbf{X}_I = \mathbf{X}_{i_1} + \ldots + \mathbf{X}_{i_p}$. The resultant matrices are computed for the groups $I = I_1, \ldots, I_m$ and the expansion (2.2) leads to the decomposition

$$\mathbf{X} = \mathbf{X}_{I_1} + \ldots + \mathbf{X}_{I_m}. \tag{2.3}$$

The procedure of choosing the sets $I_1, \ldots, I_m$ is called *eigentriple grouping*. If $m = d$ and $I_j = \{j\}$, $j = 1, \ldots, d$, then the corresponding grouping is called *elementary*.

4th step: Diagonal averaging

At this step, we transform each matrix $\mathbf{X}_{I_j}$ of the grouped decomposition (2.3) into a new series of length N. Let $\mathbf{Y}$ be an $L \times K$ matrix with elements y_{ij}, $1 \leq i \leq L$, $1 \leq j \leq K$. Set $L^* = \min(L, K)$, $K^* = \max(L, K)$ and $N = L + K - 1$. Let $y^*_{ij} = y_{ij}$ if $L < K$ and $y^*_{ij} = y_{ji}$ otherwise. By making the *diagonal averaging* we transfer the matrix $\mathbf{Y}$ into the series $y_1, \ldots, y_N$ using the formula

$$y_k = \begin{cases} \dfrac{1}{k}\displaystyle\sum_{m=1}^{k} y^*_{m,k-m+1} & \text{for } 1 \le k < L^*, \\ \dfrac{1}{L^*}\displaystyle\sum_{m=1}^{L^*} y^*_{m,k-m+1} & \text{for } L^* \le k \le K^*, \\ \dfrac{1}{N-k+1}\displaystyle\sum_{m=k-K^*+1}^{N-K^*+1} y^*_{m,k-m+1} & \text{for } K^* < k \le N. \end{cases} \tag{2.4}$$

This corresponds to averaging the matrix elements over the 'antidiagonals' $i + j = k + 1$: the choice $k = 1$ gives $y_1 = y_{1,1}$, for $k = 2$ we have $y_2 = (y_{1,2} + y_{2,1})/2$, and so on. Note that if the matrix $\mathbf{Y}$ is the trajectory matrix of some series $(z_1, \ldots, z_N)$, then $y_i = z_i$ for all i.

Diagonal averaging (2.4) applied to a resultant matrix $\mathbf{X}_{I_k}$ produces a *reconstructed series* $\widetilde{\mathbb{X}}^{(k)} = (\widetilde{x}_1^{(k)}, \ldots, \widetilde{x}_N^{(k)})$. Therefore, the initial series $x_1, \ldots, x_N$ is decomposed into a sum of m reconstructed series:

$$x_n = \sum_{k=1}^{m} \widetilde{x}_n^{(k)} \qquad (n = 1, 2, \ldots, N)\,. \tag{2.5}$$

The reconstructed series produced by the elementary grouping will be called *elementary reconstructed series*.

Remark 2.1 The Basic SSA algorithm has a natural extension to the complex-valued time series: the only difference in the description of the algorithm is the replacement of the transpose sign with the sign of complex conjugate.

2.1.2 Analysis of the Four Steps in Basic SSA

The formal description of the steps in Basic SSA requires some elucidation. In this section we briefly discuss the meaning of the procedures involved.

2.1.2.1 Embedding

Embedding is a mapping that transfers a one-dimensional time series $\mathbb{X} = (x_1, \ldots, x_N)$ into the multidimensional series $X_1, \ldots, X_K$ with vectors $X_i = (x_i, \ldots, x_{i+L-1})^{\mathrm{T}} \in \mathsf{R}^L$, where $K = N - L + 1$. The parameter defining the embedding is the *window length* L, an integer such that $2 \le L \le N - 1$. Note that the trajectory matrix (2.1) possesses an obvious symmetry property: the transposed matrix $\mathbf{X}^{\mathrm{T}}$ is the trajectory matrix of the same series $(x_1, \ldots, x_N)$ with window length equal to K rather than L.

Embedding is a standard procedure in time series analysis, signal processing and the analysis of non-linear dynamical systems. For specialists in dynamical systems, a common technique is to obtain the empirical distribution of all the pairwise distances between the lagged vectors X_i and X_j and then calculate the so-called correlation dimension of the series. This dimension is related to the fractal dimension of the attractor of the dynamical system that generates the time series; see, for example, [32] and [33]. Note that in this approach, L must be relatively small and K must be very large (formally, $K \to \infty$). Similarly, in the so-called Structural Total Least Squares (STLS) with Hankel matrix structure, the usual practice is to choose $L = r + 1$, where r is the guessed rank of the approximation matrix, see [24, 26, 27].

In SSA, the window length L should be sufficiently large. In particular, the value of L has to be large enough so that each L-lagged vector incorporates an essential part of the behaviour of the initial series $\mathbb{X} = (x_1, \ldots, x_N)$. The use of large values of L gives us a possibility of considering each L-lagged vector X_i as a separate series and investigating the dynamics of certain characteristics for this collection of series. We refer to Sect. 2.4.3 for a discussion on the choice of L.

2.1.2.2 Singular Value Decomposition (SVD)

The SVD can be described in different terms and be used for different purposes. Let us start with general properties of the SVD which are important for SSA.

As was already mentioned, the SVD of an arbitrary nonzero $L \times K$ matrix $\mathbf{X} = [X_1 : \ldots : X_K]$ is a decomposition of $\mathbf{X}$ in the form

$$\mathbf{X} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}, \tag{2.6}$$

where λ_i $(i = 1, \ldots, L)$ are eigenvalues of the matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ arranged in order of decrease, $d = \max\{i, \text{ such that } \lambda_i > 0\} = \operatorname{rank} \mathbf{X}$, $\{U_1, \ldots, U_d\}$ is the corresponding orthonormal system of the eigenvectors of the matrix $\mathbf{S}$, and $V_i = \mathbf{X}^{\mathrm{T}} U_i / \sqrt{\lambda_i}$.

Standard SVD terminology calls $\sqrt{\lambda_i}$ the *singular values*; the U_i and V_i are the *left* and *right singular vectors* of the matrix $\mathbf{X}$, respectively. If we define $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$, then the representation (2.6) can be rewritten in the form (2.2), i.e. as the representation of $\mathbf{X}$ as a sum of the elementary matrices $\mathbf{X}_i$.

If all eigenvalues have multiplicity one, then the expansion (2.2) is uniquely defined. Otherwise, if there is at least one eigenvalue with multiplicity larger than 1, then there is a freedom in the choice of the corresponding eigenvectors. We shall assume that the eigenvectors are somehow chosen and the choice is fixed.

The equality (2.6) shows that the SVD possesses the following property of symmetry: $V_1, \ldots, V_d$ form an orthonormal system of eigenvectors for the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ corresponding to the same eigenvalues λ_i. Note that the rows and columns of the trajectory matrix are subseries of the original time series. Therefore, the left and right

singular vectors also have a temporal structure and hence can also be regarded as time series.

The SVD (2.2) possesses a number of optimal features. One of these properties is as follows: among all matrices $\mathbf{X}^{(r)}$ of rank $r < d$, the matrix $\sum_{i=1}^{r} \mathbf{X}_i$ provides the best approximation to the trajectory matrix $\mathbf{X}$, so that $||\mathbf{X} - \mathbf{X}^{(r)}||_{\mathrm{F}}$ is minimum.

Here and below the (*Frobenius*) *norm* of a matrix $\mathbf{Y}$ is $||\mathbf{Y}||_{\mathrm{F}} = \sqrt{\langle \mathbf{Y}, \mathbf{Y} \rangle_{\mathrm{F}}}$, where the inner product of two matrices $\mathbf{Y} = \{y_{ij}\}_{i,j=1}^{q,s}$ and $\mathbf{Z} = \{z_{ij}\}_{i,j=1}^{q,s}$ is defined as

$$\langle \mathbf{Y}, \mathbf{Z} \rangle_{\mathrm{F}} = \sum_{i,j=1}^{q,s} y_{ij} z_{ij}.$$

For vectors the Frobenius norm is the same as the conventional Euclidean norm.

Note that $||\mathbf{X}||_{\mathrm{F}}^2 = \sum_{i=1}^{d} \lambda_i$ and $\lambda_i = ||\mathbf{X}_i||_{\mathrm{F}}^2$ for $i = 1, \ldots, d$. Thus, we shall consider the ratio $\lambda_i / ||\mathbf{X}||_{\mathrm{F}}^2$ as the characteristic of the contribution of the matrix $\mathbf{X}_i$ in the expansion (2.2) to the whole trajectory matrix $\mathbf{X}$. Consequently, $\sum_{i=1}^{r} \lambda_i / ||\mathbf{X}||_{\mathrm{F}}^2$, the sum of the first r ratios, is the characteristic of the optimal approximation of the trajectory matrix by the matrices of rank r or less. Moreover, if $\lambda_r \neq \lambda_{r+1}$ then $\sum_{i=r+1}^{d} \lambda_i$ is the distance between the trajectory matrix $\mathbf{X}$ and the set of $L \times K$ matrices of rank $\leq r$.

Let us now consider the trajectory matrix $\mathbf{X}$ as a sequence of L-lagged vectors. Denote by $\mathcal{X}^{(L)} \subset \mathsf{R}^L$ the linear space spanned by the vectors $X_1, \ldots, X_K$. We shall call this space the *L-trajectory space* (or, simply, *trajectory space*) of the series $\mathbb{X}$. To emphasize the role of the series $\mathbb{X}$, we use notation $\mathcal{X}^{(L)}(\mathbb{X})$ rather than $\mathcal{X}^{(L)}$. The equality (2.6) shows that $\mathscr{U} = (U_1, \ldots, U_d)$ is an orthonormal basis in the d-dimensional trajectory space $\mathcal{X}^{(L)}$.

Setting $Z_i = \sqrt{\lambda_i} V_i, i = 1, \ldots, d$, we can rewrite the expansion (2.6) in the form $\mathbf{X} = \sum_{i=1}^{d} U_i Z_i^{\mathrm{T}}$, and for the lagged vectors X_j we have $X_j = \sum_{i=1}^{d} z_{ji} U_i$, where the z_{ji} are the components of the vector Z_i. This means that the vector Z_i is composed of the ith components of lagged vectors X_j represented in the basis $\mathscr{U}$.

Let us now consider the transposed trajectory matrix $\mathbf{X}^{\mathrm{T}}$. Introducing $Y_i = \sqrt{\lambda_i} U_i$ we obtain the expansion $\mathbf{X}^{\mathrm{T}} = \sum_{i=1}^{d} V_i Y_i^{\mathrm{T}}$, which corresponds to the representation of the sequence of K-lagged vectors in the orthonormal basis $V_1, \ldots, V_d$. Thus, the SVD gives rise to two dual geometrical descriptions of the trajectory matrix $\mathbf{X}$.

The optimal feature of the SVD considered above may be reformulated in the language of multivariate geometry for the L-lagged vectors as follows. Let $r < d$. Then among all r-dimensional subspaces $\mathcal{L}_r$ of R^L, the subspace spanned by $U_1, \ldots, U_r$ approximates these vectors in the best way; that is, the minimum of $\sum_{i=1}^{K} \mathrm{dist}^2(X_i, \mathcal{L}_r)$ is attained at $\mathcal{L}_r = \mathrm{span}(U_1, \ldots, U_r)$. The ratio $\sum_{i=1}^{r} \lambda_i / \sum_{i=1}^{d} \lambda_i$ is the characteristic of the best r-dimensional approximation of the lagged vectors.

Another optimal feature relates to the properties of directions determined by the eigenvectors $U_1, \ldots, U_d$. Specifically, the first eigenvector U_1 determines the direction such that the variation of the projections of the lagged vectors onto this direction is maximum. Every subsequent eigenvector determines a direction that is orthogonal to all previous directions, and the variation of the projections of the lagged vectors onto this direction is also maximum. It is, therefore, natural to call the direction of ith eigenvector U_i the *ith principal direction*. Note that the elementary matrices $\mathbf{X}_i = U_i Z_i^{\mathrm{T}}$ are built up from the projections of the lagged vectors onto ith directions.

This view on the SVD of the trajectory matrix composed of L-lagged vectors and an appeal to associations with *principal component analysis* lead us to the following terminology. We shall call the vector U_i the ith (principal) *eigenvector*, the vectors V_i and $Z_i = \sqrt{\lambda_i} V_i$ will be called the *ith factor vector* and the *ith principal component*, respectively.

Remark 2.2 The SVD of the trajectory matrices used in Basic SSA is closely related to Principal Component Analysis (PCA) in multivariate analysis and the Karhunen-Loeve (KL) decomposition in the analysis of stationary time series. However, the SVD approach in SSA uses the specificity of the Hankel structure of the trajectory matrix: indeed, the columns and rows of this matrix have the same temporal sense as all they are subseries of the original series. This is not so in PCA and KL.

Remark 2.3 In general, any orthonormal basis $P_1, \ldots, P_d$ of the trajectory space can be considered in place of the SVD-generated basis consisting of the eigenvectors $U_1, \ldots, U_d$. In this case, the expansion (2.2) takes place with $\mathbf{X}_i = P_i Q_i^{\mathrm{T}}$, where $Q_i = \mathbf{X}^{\mathrm{T}} P_i$. One of the examples of alternative bases is the basis of eigenvectors of the autocovariance matrix in Toeplitz SSA, see Sect. 2.5.3. Other examples can be found among the methods of multivariate statistics such as Independent Component Analysis and Factor Analysis with rotation, see Sect. 2.5.4.

For further discussion concerning the use of other procedures in place of SVD, see Sect. 2.5.7.

2.1.2.3 Grouping

Let us now comment on the grouping step, which is the procedure of arranging the matrix terms $\mathbf{X}_i$ in (2.2). Assume that $m = 2$, $I_1 = I = \{i_1 \ldots, i_r\}$ and $I_2 = \{1, \ldots, d\} \backslash I$, where $1 \le i_1 < \ldots < i_r \le d$.

The purpose of the grouping step is the separation of additive components of time series. Let us discuss the very important concept of separability in detail. Suppose that the time series $\mathbb{X}$ is a sum of two time series $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$; that is, $x_i = x_i^{(1)} + x_i^{(2)}$ for $i = 1, \ldots, N$. Let us fix the window length L and denote by $\mathbf{X}$, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ the L-trajectory matrices of the series $\mathbb{X}$, $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$, respectively.

Consider an SVD (2.2) of the trajectory matrix $\mathbf{X}$. (Recall that if all eigenvalues have multiplicity one, then this expansion is unique.) We shall say that the series

$\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$ are (weakly) *separable by the decomposition* (2.2), if there exists a collection of indices $I \subset \{1, \ldots, d\}$ such that $\mathbf{X}^{(1)} = \sum_{i \in I} \mathbf{X}_i$ and consequently $\mathbf{X}^{(2)} = \sum_{i \notin I} \mathbf{X}_i$.

In the case of separability, the contribution of $\mathbf{X}^{(1)}$, the first component in the expansion $\mathbf{X} = \mathbf{X}^{(1)} + \mathbf{X}^{(2)}$, is naturally measured by the share of the corresponding eigenvalues: $\sum_{i \in I} \lambda_i / \sum_{i=1}^{d} \lambda_i$.

The separation of the series by the decomposition (2.2) can be looked at from different perspectives. Let us fix the set of indices $I = I_1$ and consider the corresponding resultant matrix $\mathbf{X}_{I_1}$. If this matrix, and therefore $\mathbf{X}_{I_2} = \mathbf{X} - \mathbf{X}_{I_1}$, are Hankel matrices, then they are necessarily the trajectory matrices of certain time series that are separable by the expansion (2.2).

Moreover, if the matrices $\mathbf{X}_{I_1}$ and $\mathbf{X}_{I_2}$ are close to some Hankel matrices, then there exist series $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$ such that $\mathbb{X} = \mathbb{X}^{(1)} + \mathbb{X}^{(2)}$ and the trajectory matrices of these series are close to $\mathbf{X}_{I_1}$ and $\mathbf{X}_{I_2}$, respectively (the problem of finding these series is discussed below). In this case we shall say that the series are *approximately separable*.

Therefore, the purpose of the grouping step (that is, the procedure of arranging the indices $1, \ldots, d$ into groups) is to find the groups $I_1, \ldots, I_m$ such that the matrices $\mathbf{X}_{I_1}, \ldots, \mathbf{X}_{I_m}$ satisfy (2.3) and are close to certain Hankel matrices.

Let us now look at the grouping step from the viewpoint of multivariate geometry. Let $\mathbf{X} = [X_1 : \ldots : X_K]$ be the trajectory matrix of a time series $\mathbb{X}$, $\mathbb{X} = \mathbb{X}^{(1)} + \mathbb{X}^{(2)}$, and the series $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$ are separable by the decomposition (2.2), which corresponds to splitting the index set $\{1, \ldots, d\}$ into I and $\{1, \ldots, d\} \backslash I$.

The expansion (2.3) with $m = 2$ means that $U_1, \ldots, U_d$, the basis in the trajectory space $\mathcal{X}^{(L)}$, is being split into two groups of basis vectors. This corresponds to the representation of $\mathcal{X}^{(L)}$ as a product of two orthogonal subspaces (*eigenspaces*) $\mathcal{X}^{(L,1)} = \operatorname{span}(U_i, i \in I)$ and $\mathcal{X}^{(L,2)} = \operatorname{span}(U_i, i \notin I)$ spanned by $U_i, i \in I$, and $U_i, i \notin I$, respectively.

Separability of two series $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$ means that the matrix $\mathbf{X}_I$, whose columns are the projections of the lagged vectors $X_1, \ldots, X_K$ onto the eigenspace $\mathcal{X}^{(L,1)}$, is exactly the trajectory matrix of the series $\mathbb{X}^{(1)}$.

Despite the fact that several formal criteria for separability can be introduced, the whole procedure of splitting the terms into groups (i.e., the grouping step) is difficult to formalize completely. This procedure is based on the analysis of the singular vectors U_i, V_i and the eigenvalues λ_i in the SVD expansions (2.2) and (2.6). The principles and methods of identifying the SVD components for their inclusion into different groups are described in Sect. 2.4.

Since each matrix component of the SVD is completely determined by the corresponding eigentriple, we shall talk about grouping of the eigentriples rather than grouping of the elementary matrices $\mathbf{X}_i$.

Note also that the case of two series components ($m = 2$) considered above is often more sensibly regarded as the problem of separating out a single component rather than the problem of separation of two terms. In this case, we are interested in only one group of indices, namely I.

In the problems of signal processing, the series $\mathbb{X}^{(1)}$ is interpreted as a signal. In these problems, we often choose $I_1 = \{1, \ldots, r\}$ for some r and call $\mathfrak{X}^{(1)}$ the *signal subspace*.

2.1.2.4 Diagonal Averaging

If the components of the series are separable and the indices are being split accordingly, then all the matrices in the expansion (2.3) are the Hankel matrices. We thus immediately obtain the decomposition (2.5) of the original series: for all k and n, $\widetilde{x}_n^{(k)}$ is equal to all entries $x_{ij}^{(k)}$ along the antidiagonal $\{(i, j)$, such that $i + j = n + 1\}$ of the matrix $\mathbf{X}_{I_k}$.

In practice, however, this situation is not realistic. In the general case, no antidiagonal consists of equal elements. We thus need a formal procedure of transforming an arbitrary matrix into a Hankel matrix and therefore into a series. As such, we shall consider the procedure of *diagonal averaging,* which defines the values of the time series $\widetilde{\mathbb{X}}^{(k)}$ as averages for the corresponding antidiagonals of the matrices $\mathbf{X}_{I_k}$.

It is convenient to represent the diagonal averaging step with the help of the *hankelization* operator $\mathscr{H}$. This operator acts on an arbitrary $L \times K$-matrix $\mathbf{Y} = (y_{ij})$ in the following way: for $A_s = \{(l, k) : l + k = s, 1 \le l \le L, 1 \le k \le K\}$ and $i + j = s$ the element $\widetilde{y}_{ij}$ of the matrix $\mathscr{H}\mathbf{Y}$ is

$$\widetilde{y}_{ij} = \sum_{(l,k) \in A_s} y_{lk} \Big/ |A_s|,$$

where $|A_s|$ denotes the number of elements in the set A_s.

The hankelization is an optimal procedure in the sense that the matrix $\mathscr{H}\mathbf{Y}$ is closest to $\mathbf{Y}$ (with respect to the Frobenius matrix norm) among all Hankel matrices of the corresponding size [14, Proposition 6.3]. In its turn, the Hankel matrix $\mathscr{H}\mathbf{Y}$ defines the series uniquely by relating the values in the antidiagonals to the values in the series.

By applying the hankelization procedure to all matrix components of (2.3), we obtain another expansion:

$$\mathbf{X} = \widetilde{\mathbf{X}}_{I_1} + \ldots + \widetilde{\mathbf{X}}_{I_m}, \tag{2.7}$$

where $\widetilde{\mathbf{X}}_{I_l} = \mathscr{H}\mathbf{X}_{I_l}$.

A sensible grouping leads to the decomposition (2.3) where the resultant matrices $\mathbf{X}_{I_k}$ are almost Hankel ones. This corresponds to approximate separability and implies that the pairwise inner products of different matrices $\widetilde{\mathbf{X}}_{I_k}$ in (2.7) are small.

Since all matrices on the right-hand side of the expansion (2.7) are Hankel matrices, each matrix uniquely determines the time series $\widetilde{\mathbb{X}}^{(k)}$ and we thus obtain (2.5), the decomposition of the original time series.

Note that by linearity $\mathscr{H}\mathbf{X}_I = \sum_{i \in I} \mathscr{H}\mathbf{X}_i$ and hence the order in which the Grouping and the Diagonal Averaging steps appear in Basic SSA can be reversed.

The procedure of computing the time series $\widetilde{\mathbb{X}}^{(k)}$ (that is, building up the group I_k plus diagonal averaging of the matrix $\mathbf{X}_{I_k}$) will be called *reconstruction of a series component* $\widetilde{\mathbb{X}}^{(k)}$ *by the eigentriples* with indices in I_k. In signal processing problems with $I_1 = \{1, \ldots, r\}$, we can say that the signal is reconstructed by the r leading eigentriples.

2.2 Potential of Basic SSA

In this section we start discussing examples that illustrate main capabilities of Basic SSA . Note that terms such as 'trend', 'smoothing', 'signal', and 'noise' are used here in their informal, common-sense meaning and will be commented on later.

2.2.1 Extraction of Trends and Smoothing

2.2.1.1 Trends of Different Resolution

The example 'Production' (crude oil, lease condensate, and natural gas plant liquids production, monthly data from January 1973 to September 1997, $N = 297$) shows the capabilities of SSA in extraction of trends that have different resolutions. Though the series has a seasonal component (and the corresponding component can be extracted together with the trend component), for the moment we do not pay attention to periodicities.

Taking the window length $L = 120$ we see that the eigentriples 1–3 correspond to the trend. By choosing these eigentriples in different combinations we can find different trend components.

Figure 2.1 demonstrates two alternatives in the trend resolution. The leading eigentriple gives a general tendency of the series (Fig. 2.1a). The three leading eigentriples describe the behaviour of the data more accurately (Fig. 2.1b) and show not only the general decrease of production, but also its growth from the middle 70s to the middle 80s.

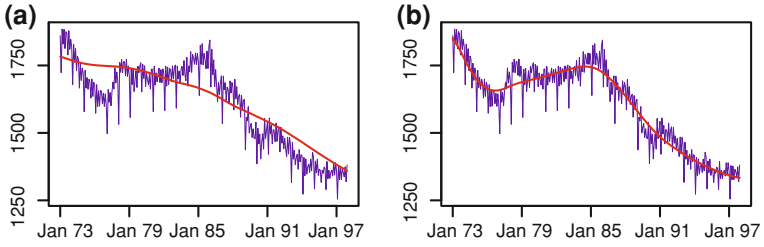

Fig. 2.1 Production: trends of different resolution. **a** General tendency (rough trend). **b** Accurate trend

2.2.1.2 Smoothing

The series 'Tree rings' (tree ring width, annual, from 42 B.C. to 1970) were collected by R. Tosh and has the ID code ITRDB CA051 in International Tree Ring Data Bank (http://www.ncdc.noaa.gov/paleo/treering.html). The time series looks like an autoregressive process. If the ARMA-type model is accepted, then it is often meaningless to look for any trend or periodicities. However, we can smooth the series with the help of Basic SSA. Figure 2.2a shows the initial series and the result of its SSA smoothing, which is obtained by choosing the leading 3 eigentriples with window length $L = 100$. Figure 2.2b depicts the residuals.

Another example demonstrating SSA as a smoothing technique uses the 'White dwarf' data, which contains 618 point measurements of the time variation of the intensity of the white dwarf star PG1159-035 during March 1989. The data is discussed in [9]. The whole series can be described as a smooth quasi-periodic curve with a noise component.

Using Basic SSA with window length $L = 100$ and choosing the leading 11 eigentriples for the reconstruction, we obtain the smooth curve of Fig. 2.3a (thick line). The residuals (Fig. 2.3b) seem to have no evident structure (to simplify the visualization of the results these figures present only a part of the series). Further analysis shows that the residual series can be regarded as a Gaussian white noise, though it does not contain very low frequencies. Thus, we can assume that in this case the smoothing procedure leads to noise reduction and the smooth curve in Fig. 2.3a describes the signal.

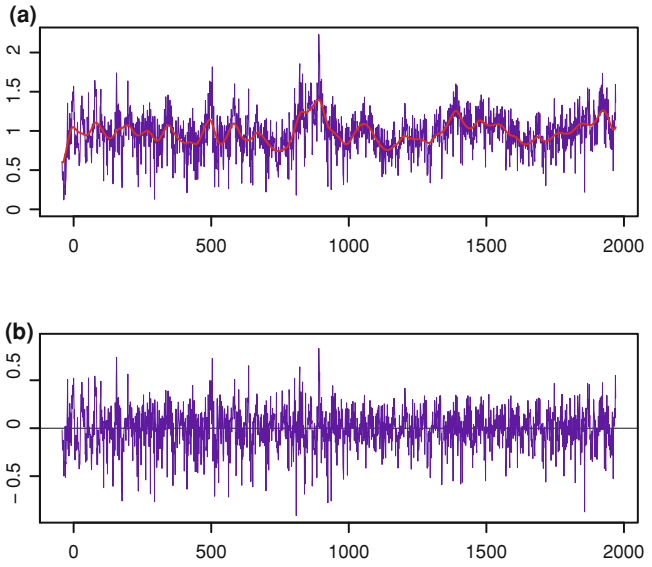

Fig. 2.2 Tree rings. **a** Smoothed series. **b** Residuals

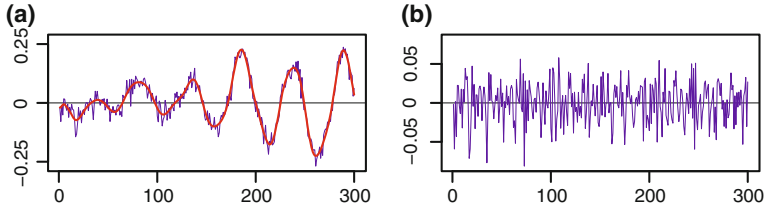

Fig. 2.3 White dwarf. **a** Smoothed series. **b** Residuals

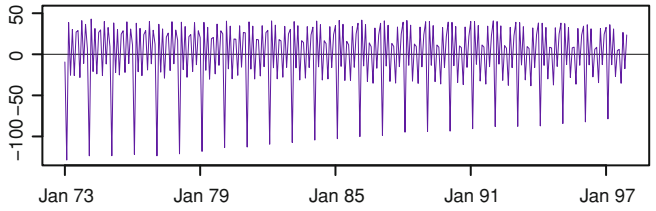

Fig. 2.4 Production: the seasonal component

2.2.2 Extraction of Periodic Components

2.2.2.1 Extraction of Seasonality Components

Let us consider the extraction of seasonality components from the 'Production' data that was discussed in Sect. 2.2.1.1.

Again, choose $L = 120$. Simultaneously with trend we are able to extract seasonal components, gathering the harmonics produced by the fundamental period 12 (12 (ET19–20), 6 (ET15–16), 4 (ET9–10), 3 (ET13–15), 2.4 (ET4–5), and 2-months (ET7) harmonics). The resulting seasonal component is depicted in Fig. 2.4. This example demonstrates that SSA can perform seasonal adjustment even for time series with complex and changing seasonal behaviour.

2.2.2.2 Extraction of Cycles with Small and Large Periods

The series 'Births' (number of daily births, Quebec, Canada, from January 1, 1977 to December 31, 1990) is discussed in [17]. It shows, in addition to a smooth trend, two cycles of different ranges: a one-year periodicity and a one-week periodicity.

Both periodicities (as well as the trend) can be simultaneously extracted by Basic SSA with window length $L = 365$. Figure 2.5 shows the one-year cycle of the series added to the trend (white line) on the background of the 'Births' series from 1981 to 1990. Note that the form of this cycle varies in time, though the main two peaks (spring and autumn) remain stable. The trend corresponds to the leading eigentriple

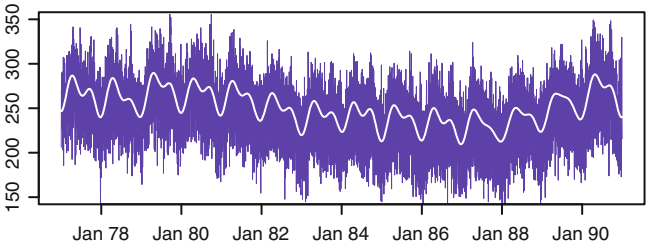

Fig. 2.5 Births: initial time series and its annual periodicity

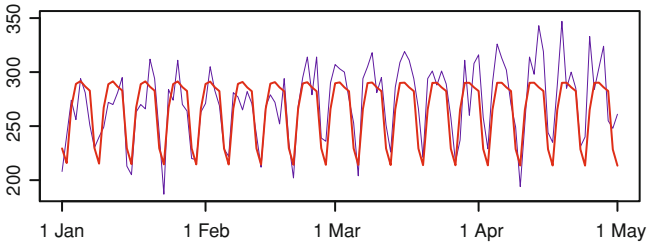

Fig. 2.6 Births: one-week periodicity

(ET1), while the one-year periodic component is reconstructed from ET 6–9. The eigentriples 12–19 also correspond to the fundamental period 365. However, they are unstable due to the small (with respect to the period value) window length.

Figure 2.6 demonstrates the one-week cycle on the background of the initial series for the first four months of 1977. This cycle corresponds to ET 2–5 and ET 10–11. The stability of the one-week periodicity does not seem to be related to the biological aspects of the birth-rate.

2.2.3 Complex Trends and Periodicities with Varying Amplitudes

The 'US unemployment' series (unemployment of females (16–19 years) in thousands, US, monthly, from 1948 to 1981, [5]) serves as an example of SSA capability of extracting complex trends simultaneously with amplitude-modulated periodicities. The result of extraction is presented in Fig. 2.7a (the initial series and the reconstructed trend) and in Fig. 2.7b (seasonality).

The window length was taken as $L = 60$. Such a moderate window length was chosen in order to simplify the capture of the complex form of the trend and complex modulation of the seasonality. The trend is reconstructed from the ET 1, 8, 13, 14, while the ET with numbers 2–7, 9–12 and 16 describe the seasonality.

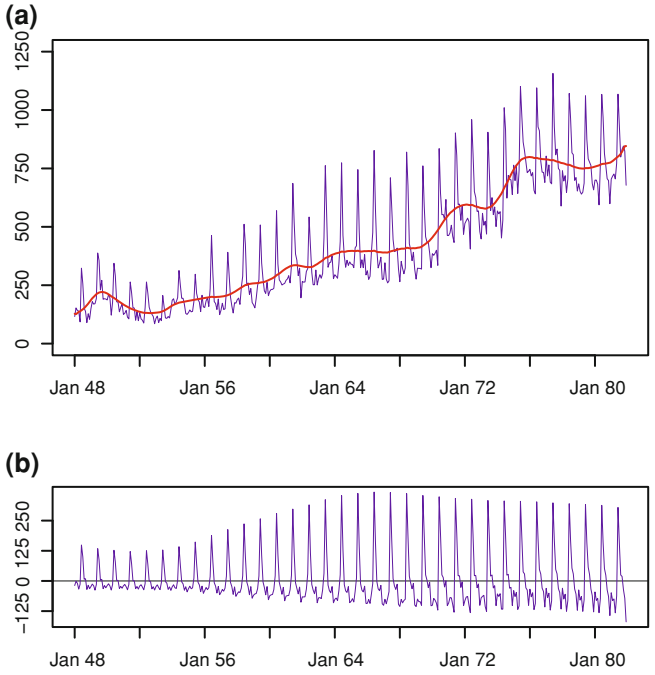

Fig. 2.7 US unemployment: decomposition. **a** Trend. **b** Seasonality

2.2.4 Finding Structure in Short Time Series

The series 'War' (U.S. combat deaths in the Indochina war, monthly, from 1966 to 1971, [20, Table 10]) is chosen to demonstrate the capabilities of SSA in finding a structure in short time series.

We have chosen $L = 18$. It is easy to see (Fig. 2.8a) that the two leading eigentriples describe perfectly the trend of the series (thick line on the background of the initial data). This trend relates to the overall involvement of U.S. troops in the war.

Figure 2.8c shows the component of the initial series reconstructed from the ET 3–4. There is little doubt that this is an annual oscillation modulated by the war intensity. This oscillation has its origin in the climatic conditions of South-East Asia: the summer season is much more difficult for any activity than the winter one.

Two other series components, namely that of the quarterly cycle corresponding to the ET 5–6 (Fig. 2.8c) and the omitted 4-months cycle, which can be reconstructed from the ET 7–8, are both modulated by the war intensity and both are less clear for interpretation. Nevertheless, if we add all these effects together (that is, reconstruct the series component corresponding to the eight leading eigentriples), a perfect agreement between the result and the initial series becomes apparent: see Fig. 2.8b with the thick line corresponding to the reconstruction.

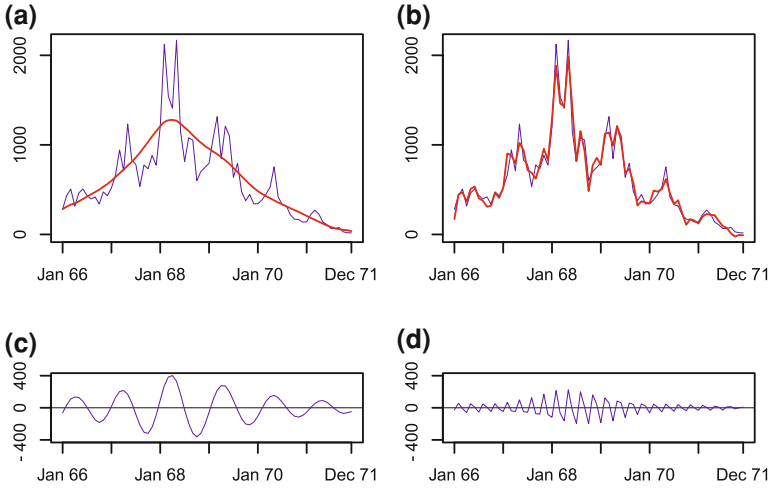

Fig. 2.8 War: structure of approximation. **a** Trend. **b** Approximation. **c** Annual cycle. **d** Quarterly cycle

2.2.5 Envelopes of Oscillating Signals and Estimation of Volatility

The capabilities of SSA in separating signals with high and low frequencies can be used in a specific problem of enveloping highly oscillating sequences with slowly varying amplitudes.

Let $x_n = A(n)\cos(2\pi\omega n)$, where the period $T = 1/\omega$ is not large in comparison with slowly varying $A(n)$. Define

$$y_n \stackrel{\text{def}}{=} 2x_n^2 = A^2(n) + A^2(n)\cos(4\pi\omega n). \tag{2.8}$$

Since $A^2(n)$ is slowly varying and the second term on the right-hand side of (2.8) oscillates rapidly, we can gather slowly varying terms of SSA decomposition for y_n and therefore approximately extract the term $A^2(n)$ from the series (2.8). All we need to do then is to take the square root of the extracted term.

Let us illustrate this technique. Consider the square of the annual periodicity of the 'Germany unemployment' series (Fig. 2.33b in Sect. 2.5.5) multiplied by 2 and denote it by $\mathbb{Y}$. Taking window length $L = 36$ and reconstructing the low-frequency part of the time series $\mathbb{Y}$ from the eigentriples 1, 4, 7 and 10, we obtain an estimate of $A^2(n)$ (the reconstructed series are depicted in Fig. 2.9a by the thick line; the thin line corresponds to the series $\mathbb{Y}$). By taking the square root of the estimate we obtain the result (see Fig. 2.9b).

Very similarly we can analyze the dynamics of the variance of a heteroscedastic noise. Let $x_n = A(n)\varepsilon_n$, where ε_n is the white normal noise with zero mean and unit variance and $A(n)$ is a slowly changing function. Since $A^2(n) = \mathbf{D}x_n = \mathbf{E}x_n^2$, the trend extracted from the series $\mathbb{Y}$ with $y_n = x_n^2$ provides the estimate of the variance.

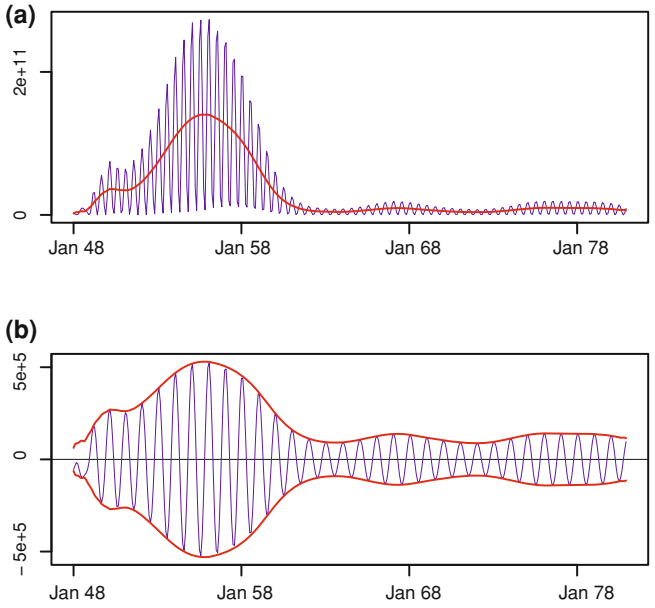

Fig. 2.9 Germany unemployment: envelope construction. **a** Squared annual periodicity and its trend. **b** Annual periodicity and its envelope

2.3 Models of Time Series and SSA Objectives

In the previous section the terms 'trend', 'smoothing', 'amplitude modulation' and 'noise' were used without any explanation of their meaning. In this section we shall provide related definitions and corresponding discussions. We shall also describe the major tasks that can be attempted by Basic SSA . Examples of application of Basic SSA for performing these tasks have been considered above in Sect. 2.2.

2.3.1 SSA and Models of Time Series

2.3.1.1 Models of Time Series and Periodograms

Formally, SSA can be applied to an arbitrary time series. However, a theoretical study of its properties requires specific considerations for different classes of series. Moreover, different classes assume different choices of parameters and expected results. We thus start this section with description of several classes of time series, which are natural for SSA treatment, and use these classes to discuss the important concept of (approximate) separability defined earlier in Sect. 2.1.2.3. (For the theoretical aspects of separability see [14].)

Since the main purpose of SSA is to make a decomposition of the series into additive components, we always implicitly assume that this series is a sum of several simpler series. These 'simple' series are the objects of the discussion below. Note also that here we only consider deterministic time series, including those that can be regarded as 'noise'.

(a) Periodogram

For a description of the time series $\mathbb{X}_N = (x_1, \ldots, x_N)$ in terms of frequencies it is convenient to use the language of the *Fourier expansion* of the initial series. This is the expansion

$$x_n = c_0 + \sum_{k=1}^{\lfloor N/2 \rfloor} \Big(c_k \cos(2\pi n\, k/N) + s_k \sin(2\pi n\, k/N) \Big), \tag{2.9}$$

where N is the length of the series, $1 \le n \le N$, and $s_{N/2} = 0$ for even N. The zero term c_0 is equal to the average of the series, so that if the series is centred, then $c_0 = 0$. Let $A_k^2 = c_k^2 + s_k^2$. Another form of (2.9) is

$$x_n = c_0 + \sum_{k=1}^{\lfloor N/2 \rfloor} A_k \cos(2\pi n\, k/N + \varphi_k).$$

We define the *periodogram* as

$$\Pi_x^N(k/N) = \begin{cases} c_0^2 & \text{for } k = 0, \\ (c_k^2 + s_k^2)/2 & \text{for } 0 < k < N/2, \\ c_{N/2}^2 & \text{for } k = N/2. \end{cases} \tag{2.10}$$

The last case is only possible if N is even. The normalization in the definition (2.10) is chosen to obtain

$$||\mathbb{X}_N||_{\mathrm{F}}^2/N = \sum_{k=0}^{\lfloor N/2 \rfloor} \Pi_x^N(k/N). \tag{2.11}$$

Some other normalizations of the periodograms are known in literature and could be useful as well. The equality (2.11) implies that the value (2.10) of the periodogram at the point k/N describes the influence of the harmonic components with frequency $\omega = k/N$ into the sum (2.9).

The collection of frequencies $\omega_k = k/N$ with positive powers is called *support of the periodogram*. If the support of a certain periodogram belongs to some interval $[a, b]$, then this interval is called the *frequency range of the series*.

Formally, the periodogram of the series is an analogue of the spectral measure for stationary series. Asymptotically, if the series is stationary, then the periodograms

approximate the spectral measures (see [14, Theorem 6.4]). The periodogram can also be helpful for a general description of an arbitrary time series. For example, trends can be described as finite subseries of a stationary low-frequency time series.

The drawback of the periodogram analysis is its low resolution. In particular, the perodogram can not distinguish frequencies that differ on any amount that is smaller than $1/N$. For short series the grid $\{j/N,\ \ j = 0, \dots, \lfloor N/2 \rfloor\}$ is a poor approximation to the whole range of frequencies $[0, 1/2]$, and the periodogram may not reveal a periodic structure of the series components.

(b) Stationary series

An infinite series (not necessarily stochastic) $\mathbb{X}_\infty = (x_1, x_2, \dots, x_N, \dots)$ is called *stationary* if for all nonnegative integers k, m we have

$$\frac{1}{N}\sum_{j=1}^{N} x_{j+k}x_{j+m} \underset{N\to\infty}{\longrightarrow} R(k-m), \tag{2.12}$$

where the (even) function $R(\cdot)$ is called the *covariance function* of the series $\mathbb{X}$ (the convergence in (2.12) is either deterministic or weak probabilistic depending on whether the series is deterministic or stochastic). Below, when discussing stationarity, we shall always assume that $\frac{1}{N}\sum_{j=1}^{N} x_{j+k} \to 0$ (as $N\to\infty$) holds for any k, which is the zero-mean assumption for the original series.

The covariance function can be represented through the spectral measure, which determines properties of the corresponding stationary series in many respects. The periodogram of a finite series $\mathbb{X}_N$ provides the estimate of the spectral density of $\mathbb{X}_\infty$.

A stationary series $\mathbb{X}_\infty$ with discrete spectral measure m_x can normally be written as

$$x_n \sim \sum_k a_k \cos(2\pi\omega_k n) + \sum_k b_k \sin(2\pi\omega_k n), \quad \omega_k \in (0, 1/2], \tag{2.13}$$

where $a_k = a(\omega_k)$, $b_k = b(\omega_k)$, $b(1/2) = 0$ and the sum $\sum_k (a_k^2 + b_k^2)$ converges. (Note that $a(1/2) \neq 0$ if one of the ω_k is exactly $1/2$.) The form (2.13) for the series $\mathbb{X}_\infty$ means the measure m_x is concentrated at the points $\pm\omega_k$, $\omega_k \in (0, 1/2)$, with the weights $(a_k^2 + b_k^2)/4$. The weight of the point $1/2$ equals $a^2(1/2)$.

A series of the form (2.13) will be called *almost periodic*. *Periodic* series correspond to a spectral measure m_x concentrated at the points $\pm j/T$ ($j = 1, \dots, \lfloor T/2 \rfloor$) for some integer T. In terms of the representation (2.13), this means that the number of terms in this representation is finite and all frequencies ω_k are rational.

Almost periodic series that are not periodic are called *quasi-periodic*. For these series the spectral measure is discrete, but it is not concentrated on the nodes of any grid of the form $\pm j/T$. The *harmonic* $x_n = \cos(2\pi\omega n)$ with irrational ω provides an example of a quasi-periodic series.

Aperiodic (in other terminology, *chaotic*) series are characterized by a spectral measure that does not have atoms. In this case, one usually assumes the existence of the *spectral density*: $m_x(d\omega) = p_x(\omega)d\omega$. Aperiodic series are often used as models for *noise*. If the spectral density of an aperiodic stationary series is constant, then this series is called *white noise*. Note that the white noise series does not have to be stochastic. In many cases, real-life stationary series have both components, periodic (or quasi-periodic) and noise (aperiodic) components.

It is difficult, if not impossible, while dealing with finite series, to make a distinction between a periodic series with large period and a quasi-periodic series. Moreover, on finite time intervals aperiodic series are almost indistinguishable from a sum of harmonics with wide spectrum and small amplitudes.

(c) Amplitude-modulated periodicities

The definition of stationarity is asymptotic. This asymptotic nature has both advantages (for example, a rigorous mathematical definition allows an illustration of the main concepts by model examples) and disadvantages (for example, it is impossible to check the assumption of stationarity using only finite data).

There are numerous deviations from stationarity. We consider only two classes of nonstationary time series which we describe at a qualitative level. Specifically, we consider amplitude-modulated periodic series and series with trends. The choice of these two classes is related to their practical significance and importance for SSA.

The trends are dealt with in the next subsection. Here we discuss the *amplitude-modulated* periodic signals; that is, series of the form $x_n = A(n)y_n$, where y_n is a periodic sequence and $A(n) \geq 0$. Usually it is assumed that on the given time interval ($1 \leq n \leq N$) the function $A(n)$ varies much slower than the low-frequency harmonic component of the series y_n.

Series of this kind are typical in economics where the period of the harmonics y_n is related to seasonality, but the amplitude modulation is determined by the long-term tendencies. Similar interpretation seems to be true for the example 'War', where the seasonal component of the combat deaths (Fig. 2.8c, d) is likely to be modulated by the intensity of the military activities.

Let us discuss the periodogram analysis of the amplitude-modulated periodic signals, temporarily restricting ourselves to the amplitude-modulated harmonic

$$x_n = A(n)\cos(2\pi\omega n + \theta), \quad n = 1, \ldots, N. \tag{2.14}$$

Unless the series (2.14) is too short, its periodogram is supported on a short frequency interval containing ω. Indeed, for large $\omega_1 \approx \omega_2$ the sum

$$\cos(2\pi\omega_1 n) + \cos(2\pi\omega_2 n) = 2\cos\Big(\pi(\omega_1 - \omega_2)n\Big)\cos\Big(\pi(\omega_1 + \omega_2)n\Big)$$

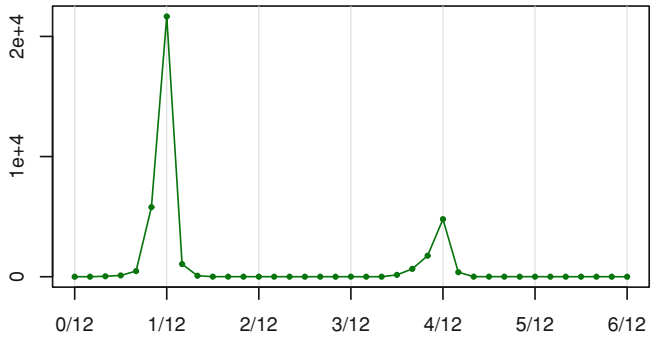

Fig. 2.10 War: periodogram of the main seasonality component

is a product of a slowly varying sequence $A(n) = 2\cos\Big(\pi(\omega_1 - \omega_2)n\Big)$ and a harmonic with large frequency $(\omega_1 + \omega_2)/2$. The oscillatory nature of the sequence $A(n)$ cannot be seen for small N.

Figure 2.10 depicts the periodogram of the main seasonal (annual plus quarterly) component of the series 'War' (Sect. 2.2.4). We can see that the periodogram is supported at around two main seasonal frequencies. However, it is not totally concentrated at these two points; this is caused by the amplitude modulation.

The discussion above implies that the appearance of exactly the same modulation can be caused by two different reasons: either it is the 'true' modulation or the modulation is spurious and originates from the closeness of the frequencies of the harmonic components of the original series.

Another reason for the frequencies spreading around the main frequency is the discreteness of the periodogram grid $\{k/N\}$: if the frequency ω of a harmonic does not belong to the grid, then it spreads around the grid giving large positive values to two or more frequencies on the grid points next to ω.

Note that since the length of the 'War' series is proportional to 12, the frequencies $\omega = 1/12$ and $\omega = 1/3$, which correspond to annual and quarterly periodicities, fall exactly on the periodogram grid $\{k/36,\ k = 1, \ldots, 18\}$.

It is evident that not only periodic series can be modulated by the amplitude; the same can happen to the quasi-periodic and chaotic sequences. However, identification of these cases by means of the periodogram analysis is much more difficult.

(d) Trends

There is no commonly accepted definition of the concept 'trend'. Common approaches for defining trend either need postulating a parametric model (this would allow the consideration of linear, exponential and logistic trends, among others) or consider the trend as a solution of an approximation problem, without any concerns about the tendencies; the most popular kind of trend approximation is the polynomial approximation.

In SSA framework, such meanings of the notion 'trend' are not suitable just because Basic SSA is a model-free, and hence nonparametric, method. In general, an appropriate definition of trend for SSA defines the trend as an additive component of the series which is (i) not stationary, and (ii) 'slowly varies' during the whole period of time that the series is being observed (compare [8, Chap. 2.12]).

Note that we have already collected oscillatory components of the series into a separate class of (centred) stationary series and therefore the term 'cyclical trend' does not make much sense for us.

Let us now discuss some consequences of this understanding of the notion 'trend'. The most important is the nonuniqueness of the solution to the problem 'trend identification' or 'trend extraction' in its nonparametric setup. This nonuniqueness has already been illustrated by the example 'Production', where Fig. 2.1 depicts two forms of the trend: a trend that explains a general tendency of the series (Fig. 2.1a) and a detailed trend (Fig. 2.1b).

Furthermore, for finite time series, a harmonic component with very low frequency is practically indistinguishable from a trend (it can even be monotone on a finite time interval). In this case, supplementary subject-related information about the series can be decisive for the problem of distinguishing trend from the periodicity. For instance, even though the reconstructed trend in the example 'War' (see Fig. 2.8a) looks like a periodicity observed over a time interval that is less than half of the period, there is no question of periodicity in this case.

In the language of frequencies, any trend generates large powers in the region of low-frequencies in the periodogram. Moreover, we have assumed that any stationary series is centred. Therefore, the average of all terms x_n of any series $\mathbb{X}$ is always added to its trend. On the periodogram, a nonzero constant component of the series corresponds to an atom at zero.

(e) Additive components of time series: case study

Summarizing, a general descriptive model of the series that we use in SSA methodology is an additive model where the components of the series are trends, oscillations and noise components. In addition, the oscillatory components are subdivided into periodic and quasi-periodic, while the noise components are aperiodic series. Amplitude modulation of the oscillatory and noise components is permitted. The sum of all additive components, except for the noise, will be referred to as *signal*.

Example 2.1 Let us consider the 'Rosé wine' series (monthly rosé wine sales, Australia, from July 1980 to June 1994, thousands of litres). Figure 2.11 depicts the series itself (the thin line) and Fig. 2.12 presents its periodogram.

Figure 2.11 shows that the series 'Rosé wine' has a decreasing trend and an annual periodicity of a complex form. Figure 2.12 shows the periodogram of the series; it seems reasonable to assume that the trend is related to the large values at the low-frequency range, and the annual periodicity is related to the peaks at the frequencies

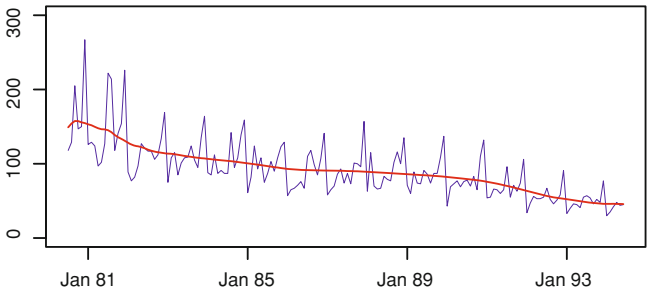

Fig. 2.11 Rosé wine: initial time series and the trend

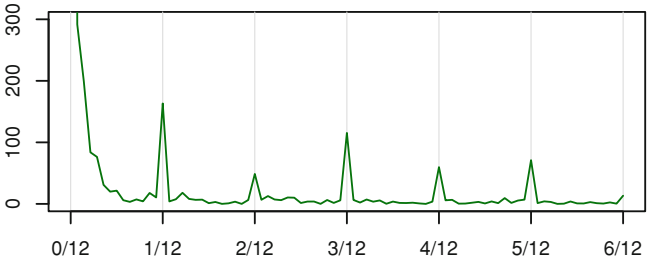

Fig. 2.12 Rosé wine: periodogram for the series

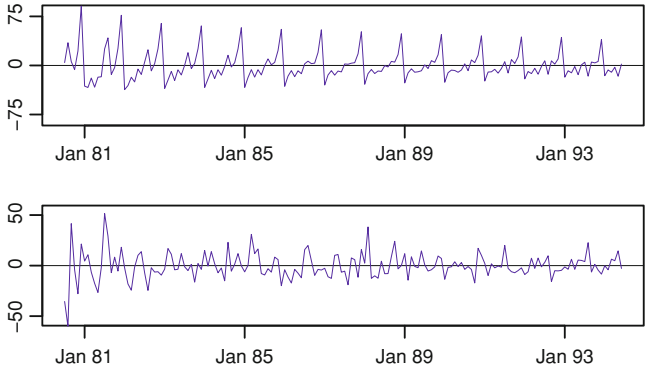

Fig. 2.13 Rosé wine: two components of the series

1/12, 1/6, 1/4, 1/3, 1/2.4, and 1/2. The non-regularity of powers of these frequencies indicates a complex form of the annual periodicity.

We have applied Basic SSA with window length $L = 84$. Figure 2.13 depicts two additive components of the 'Rosé wine' series: the seasonal component (top graph), which is described by the ET 2–11, 13 and the residual series. The trend component (thick line in Fig. 2.11) is reconstructed from the ET 1, 12, 14.

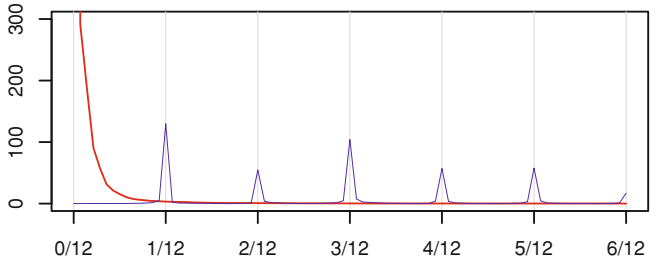

Fig. 2.14 Rosé wine: periodograms of the trend and the seasonal component

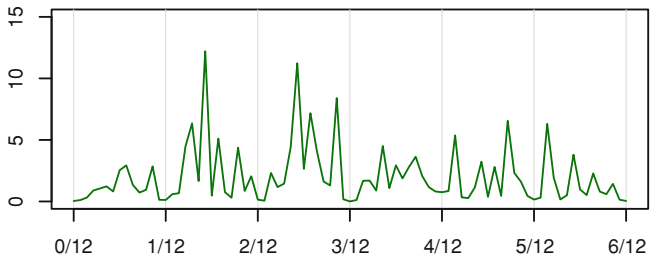

Fig. 2.15 Rosé wine: periodogram of the residuals

Periodogram analysis demonstrates that the expansion of the series into three parts is indeed related to the separation of the spectral range into three regions: low frequencies correspond to the trend (the thick line in Fig. 2.14), the frequencies describing the seasonalities correspond to the periodic component (Fig. 2.14, the thin line), and the residual series (which can be regarded as noise) has all the other frequencies (Fig. 2.15). Note that the periodograms of the whole series (see Fig. 2.12), its trend and the seasonal component (see Fig. 2.14) are presented on the same scale.

2.3.1.2 Models of Time Series and Rank

In the framework of SSA, the structure of the time series is closely related to $d(L) = \operatorname{rank} \mathbf{X}$, the number of non-zero eigenvalues in the SVD of the trajectory matrix $\mathbf{X}$ (we shall call this number L-rank of the time series). If for some fixed d we have $d(L) = d$ for large enough L, then the time series is called a finite-rank time series of rank d (see [14, Chap. 5] for details). For such a series, we have $d(L) = \min(d, L)$ if $L \le K$.

For any time series of finite length, $d \le \min(L, K)$. If $d < \min(L, K)$, then the time series has a structure. A small value of d corresponds to a series with simple structure. In particular, if the time series component is of a small rank, then the grouping for its reconstruction is easier.

Let us consider several examples of time series models in terms of their rank. Note that the class of finite-rank time series includes sums of products of polynomials, exponentials and sinusoids.

Pure periodicities. Any sine-wave time series (so-called sinusoid) with frequency from the range $(0, 0.5)$ has rank 2, the saw-tooth sinusoid with frequency 0.5 has rank 1. Therefore, any almost periodic time series in the form (2.13) with finite number of addends has finite rank. Certainly, any periodic time series has finite rank. Aperiodic time series cannot have a finite rank.

Note that the simplicity of the sinusoid in the framework of SSA analysis depends on the number of the observed periods, while the fact that the rank of the sinusoid is equal to 2 is valid for the sinusoid of any frequency from $(0, 0.5)$.

Modulated periodicities. Modulation of periodicities can complicate or even destroy SSA structure of the series. As a rule, for an arbitrary modulation, the modulated sinusoid is not of finite rank. The cosine modulation $A(n)$ defined in (2.14) is an example where the rank increases from 2 to 4 but stays finite.

The only possible example of modulation that does not change the rank of the signal is the exponential modulation $A(n) = \exp(\alpha n) = \rho^n$ with $\rho = e^{\alpha}$. For example, the rank of an exponentially damped sinusoid is the same as that of the undamped sinusoid. This is the essential advantage of SSA relative to the standard methods like the Fourier analysis and allows processing of the time series without log-transformation. Also, this allows SSA to deal with periodicities whose shape is changing.

Let us consider the 'Fortified wine' series (monthly fortified wine sales, Australia, from July 1980 to June 1994, thousands of litres). Figure 2.16 depicts the series itself (the thin line) and the reconstructed seasonality (the thick line); here the window length is $L = 84$ and the reconstruction is performed by ET 2–11. One can see that the form of seasonality is changing. This means that the standard methods of analysis like Fourier analysis can not be applied, even after the log-transformation. Figure 2.17 shows different kinds of modulation of the extracted (by Basic SSA) sine waves that altogether define the seasonal behaviour of the 'Fortified wine' series.

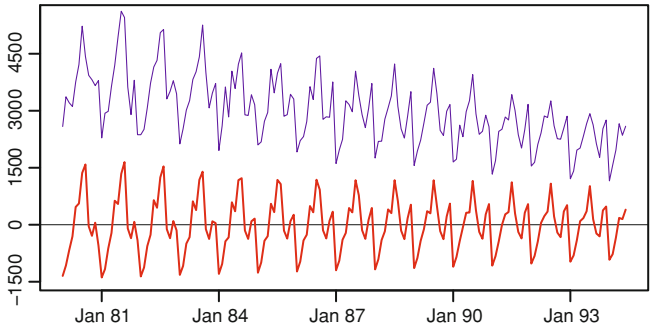

Fig. 2.16 Fortified wine: the initial time series and the reconstructed dynamic of the seasonality

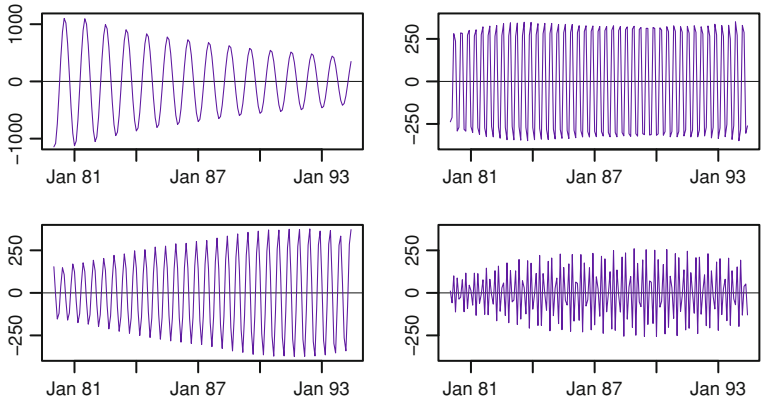

Fig. 2.17 Fortified wine: different behaviour of seasonal components

Trends. Trends have very different and, as a rule, non-structured behaviour; also, the trends make the main contribution towards the non-stationarity of the series. A typical trend (which is a slowly varying component of the series) can be accurately approximated by a series of finite rank. The list of slowly-varying series with simple SSA structure and small rank includes an exponential series (rank 1), a sinusoid with large period (rank 2), a linear series (rank 2) and polynomials of higher order (rank > 2).

2.3.1.3 Additive and Multiplicative Models

By the definition, an additive model of a series is a sum of components, while the multiplicative model is a product of positive components. Any multiplicative model can be easily transformed to the additive model by applying the log-transformation to the series.

SSA deals with time series that can be represented as sums of components. One may think that SSA can not be used for series represented via a multiplicative model. However, some series in a multiplicative model can be represented as sums with no extra transformation required. For example, let $x_n = t_n(1 + s_n)$, where $(t_1, \dots, t_N)$ is a trend and $(s_1, \dots, s_N)$ is a sinusoid with amplitude smaller than 1 (this is needed for positivity of $1 + s_n$). It is easily seen that $x_n = t_n + t_n s_n$; that is, the initial time series can be considered as a sum of a trend and a modulated sinusoid. Therefore, the multiplicative model can be considered as an additive one with modulated oscillations and noise.

Thus, SSA can be applied to both additive and multiplicative models. Log-transformation can increase the accuracy only if the structure of the signal after the log-transformation is simpler (has smaller rank) or the separability is improved. Otherwise the log-transformation leads to a decrease of the accuracy of SSA analysis. As an example, the log-transformation always worsens the structure of the series with exponential trend.

2.3.1.4 Non-parametric Versus Parametric Models

To use Basic SSA we do not need to assume any model about the time series. Therefore, Basic SSA belongs to the class of nonparametric and model-free techniques. However, under the assumption of separability, a parametric model can be constructed based on SSA results. Let us demonstrate the idea.

Let the component $\mathbb{X}^{(1)}$ of the series $\mathbb{X} = \mathbb{X}^{(1)} + \mathbb{X}^{(2)}$ be L-separable and therefore have finite L-rank $r < \min(L, K)$. Let $\mathbb{X}^{(1)}$ be reconstructed by the r leading eigentriples, that is, $I_1 = \{1, \dots, r\}$. Denote $\mathcal{X}^{(1)} = \operatorname{span}(U_1, \dots, U_r)$ its trajectory space. If the Lth coordinate vector $\mathbf{e}_L = (0, \dots, 0, 1)^{\mathrm{T}} \notin \mathcal{X}^{(1)}$, then $\mathbb{X}^{(1)}$ is governed by a *linear recurrence relation* (LRR)

$$x_n^{(1)} = \sum_{j=1}^{r} a_j x_{n-j}^{(1)}, \quad n = r+1, \dots, N,$$

where the coefficients a_j are uniquely defined by the r-dimensional subspace $\mathcal{X}^{(1)}$, see [14, Chap. 5].

The coefficients a_j determine the complex numbers $\mu_1, \dots, \mu_r$ which are the roots of the characteristic polynomial of the LRR, see Sect. 3.2 (we assume, for simplicity, that all roots μ_j are different; the case where some of μ_j are equal is more complicated and corresponds to the polynomial modulation of the time series components). The time series $x_n^{(1)}$ can be written in terms of $\mu_1, \dots, \mu_r$ as

$$x_n^{(1)} = \sum_{j=1}^{r} C_j \mu_j^n \tag{2.15}$$

with some coefficients C_j (see Theorem 3.1 in Sect. 3.2). Note that since $\mathbb{X}$ is a real-valued time series, if $\mu_j \in \{\mu_1, \dots, \mu_r\}$ and μ_j is complex then there is complex-conjugate $\mu_k = \mu_j^*$ of μ_j among $\{\mu_1, \dots, \mu_r\}$. As we can write $\mu_j = \rho_j \exp(\mathrm{i}2\pi\omega_j)$, the set $\{\mu_j\}$ provides full information about the frequencies $\{\omega_j\}$. For known $\{\mu_j\}$, the coefficients C_j are determined by, for example, values $x_1^{(1)}, \dots, x_r^{(1)}$.

Since in practice there is no exact separability between time series components, many methods are developed to estimate coefficients of the parametric form of the time series component, see Sect. 3.8 for more information.

2.3.2 Classification of the Main SSA Tasks

Basic SSA can be very useful for solving the following problems of time series analysis: smoothing, extraction of trend and extraction of oscillatory components. The most general problem which Basic SSA may attempt to solve is that of finding

the whole structure of the series; that is splitting the series into several 'simple' and 'interpretable' components, and the noise component. Let us discuss all these problems separately.

1. *Trend extraction and smoothing*
   There is no clear distinction between the trend extraction and smoothing; for instance, the example 'US unemployment' (Fig. 2.7a) can at the same time be considered as an example of a refined trend extraction as well as smoothing.
   Neither of these two problems has exact meaning unless a parametric model is assumed. As a result, a large number of model-free methods can be applied to solve both of them. It is however convenient to distinguish between trend extraction and smoothing, at least on a qualitative level.
   Results of the trend extraction by Basic SSA are illustrated on the examples 'Production' (Fig. 2.1a, b), 'US unemployment' (Fig. 2.7a) and 'War' (Fig. 2.8a). The example 'Tree rings' (Fig. 2.2a) shows smoothing capabilities of Basic SSA (see also [4, 15]).
   Note that the problem of noise reduction is very similar to the problem of smoothing. The difference between these two problems is related to the conditions which the residual is expected to satisfy: for the noise reduction, the residual must not include any part of the signal whereas in the problem of smoothing the residual may include high-frequency periodic components.

2. *Extraction of oscillatory components*
   The general problem here is the identification and separation of oscillatory components of the series that do not constitute parts of the trend. In the parametric form (under the assumptions of zero trend, finite number of harmonics, and additive stochastic white noise), this problem is extensively studied in the classical spectral analysis theory.
   Basic SSA is a model-free method. Therefore, the result of Basic SSA extraction of a single harmonic component of a series is typically not a purely harmonic sequence. This is related to the fact that in practice we deal with an approximate separability rather than with the exact one (see Sect. 2.3.3).
   Basic SSA does not require assumptions about the number of harmonics and their frequencies. However, an auxiliary information about the initial series always makes the situation clearer and helps in choosing parameters of the method, see Sect. 2.4.2.1.
   Finally, SSA allows the possibility of amplitude modulation for the oscillatory components of the series. Examples 'War' (Sect. 2.2.4) and 'US unemployment' (Sect. 2.2.3) illustrate the capabilities of Basic SSA for the extraction of modulated oscillatory components.

3. *Splitting the series into 'simple' and 'interpretable' components and noise*
   This task can be thought of as a combination of two tasks considered above; specifically, the tasks of extraction of trend and extraction of periodic components. A specific feature of this task is that in the full decomposition the residual should consist of the noise only. Since model-free techniques often tend to find false

interpretable components in noise, it is highly recommended to have a clear explanation (obtained using an information additional to the time series data itself) for each signal component found.

2.3.3 Separability of Components of Time Series

As discussed above, the main purpose of SSA is the decomposition of the original series into a sum of series, so that each component in this sum can be identified as either a trend, periodic or quasi-periodic component (perhaps, amplitude-modulated), or noise.

The notion of separability of series plays a fundamental role in the formalization of this problem (see Sects. 2.1.2.3 and 2.1.2.4). Roughly speaking, an SSA decomposition of the series $\mathbb{X}$ can be useful and informative only if the resulting additive components of the series are (approximately) separable from each other.

Weak and strong separability

Let us fix the window length L, consider a certain SVD of the L-trajectory matrix $\mathbf{X}$ of the initial series $\mathbb{X}$ of length N, and assume that the series $\mathbb{X}$ is a sum of two series $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$, that is, $\mathbb{X} = \mathbb{X}^{(1)} + \mathbb{X}^{(2)}$.

In this case, separability of the series $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$ means (see Sect. 2.1.2.3) that we can split the matrix terms of the SVD of the trajectory matrix $\mathbf{X}$ into two different groups, so that the sums of terms within the groups give the trajectory matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ of the series $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$, respectively.

Since the SVD is not uniquely defined if there are multiple singular values, two types of separability can be considered. The separability is called *weak* if *there exists an SVD* of the trajectory matrix $\mathbf{X}$ such that we can split the SVD matrix terms into two different groups, so that the sums of terms within the groups give $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. The separability is called *strong*, if this is true *for any SVD* of the trajectory matrix.

For strong separability, it is necessary that the sets of eigenvalues produced by the SVDs of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ have no intersection. Strong separability implies the weak one and it is more desirable in practice. The absence of strong separability can be a serious problem for SSA. In Sect. 2.5.4 we develop a new method called SSA–ICA; this method can provide separability if there is no strong separability. Weak separability is easier to study and validate in practice. Although conditions for exact (weak) separability are rather restrictive, they can be extended to approximate separability and therefore be used in the practical analysis.

The following conditions are equivalent to the definition of weak separability of two series $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$:

1. any subseries of length L (and $K = N - L + 1$) of the series $\mathbb{X}^{(1)}$ is orthogonal to any subseries of the same length of the series $\mathbb{X}^{(2)}$ (the subseries of the time series are considered here as vectors); in term of trajectory matrices, $\mathbf{X}^{(1)}(\mathbf{X}^{(2)})^{\mathrm{T}} = \mathbf{0}_{LL}$ and $(\mathbf{X}^{(1)})^{\mathrm{T}}\mathbf{X}^{(2)} = \mathbf{0}_{KK}$;

2. the subspace $\mathcal{X}^{(L,1)}$ spanned by the columns of the trajectory matrix $\mathbf{X}^{(1)}$, is orthogonal to the subspace $\mathcal{X}^{(L,2)}$ spanned by the columns of the trajectory matrix $\mathbf{X}^{(2)}$, and similar orthogonality must hold for the subspaces $\mathcal{X}^{(K,1)}$ and $\mathcal{X}^{(K,2)}$ spanned by the rows of the trajectory matrices.

Characteristics of weak separability

Let $L^* = \min(L, K)$ and $K^* = \max(L, K)$. Introduce the weights

$$w_i = \begin{cases} i & \text{for } 0 \le i < L^*, \\ L^* & \text{for } L^* \le i \le K^*, \\ N - i + 1 & \text{for } K^* < i \le N. \end{cases} \tag{2.16}$$

The weight w_i in (2.16) is equal to the number of times the element x_i appears in the trajectory matrix $\mathbf{X}$ of the series $\mathbb{X} = (x_1, \ldots, x_N)$. Define the inner product of two series $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$ of length N as

$$\left(\mathbb{X}^{(1)}, \mathbb{X}^{(2)}\right)_w \stackrel{\text{def}}{=} \sum_{i=1}^{N} w_i x_i^{(1)} x_i^{(2)} \tag{2.17}$$

and call the series $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$ **w**-*orthogonal* if $\left(\mathbb{X}^{(1)}, \mathbb{X}^{(2)}\right)_w = 0$.

It follows from the separability conditions that separability implies **w**-orthogonality. To measure the degree of approximate separability between two series $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$ we introduce the so-called **w**-correlation

$$\rho^{(w)}(\mathbb{X}^{(1)}, \mathbb{X}^{(2)}) \stackrel{\text{def}}{=} \frac{\left(\mathbb{X}^{(1)}, \mathbb{X}^{(2)}\right)_w}{\|\mathbb{X}^{(1)}\|_w \|\mathbb{X}^{(2)}\|_w}. \tag{2.18}$$

We shall loosely say that two series $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$ are *approximately separable* if $\rho^{(w)}(\mathbb{X}^{(1)}, \mathbb{X}^{(2)}) \simeq 0$ for reasonable values of L (see [14, Sects. 1.5 and 6.1] for precise definitions). Note that the window length L enters the definitions of **w**-orthogonality and **w**-correlation, see (2.16).

Another qualitative measure of separability is related to the frequency structure of the time series components [14, Sect. 1.5.3]. It is sufficient (but not necessary) for weak separability that the supports of the periodograms of $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$ do not intersect. If the intersection of the supports is, in a sense, small then the separability becomes approximate. Note that the separability of frequencies is equivalent to weak separability for the stationary time series.

Separable time series

Although there are many results available (see [14, Sects. 1.5 and 6.1]) on exact separability for the time series of finite rank, exact separability mostly presents purely

theoretical interest. In practice, exact separability of components hardly ever occurs but an approximate separability can be achieved very often. It is very important in practice that the trend, oscillations and noise components are approximately separable for large enough time series and window lengths.

To illustrate the concept of separability consider an example of two sinusoids

$$x_n^{(1)} = A_1 \cos(2\pi n\omega_1 + \varphi_1), \qquad x_n^{(2)} = A_2 \cos(2\pi n\omega_2 + \varphi_2), \tag{2.19}$$

where $n = 1, \ldots, N$, $0 < \omega_i < 0.5$ and $\omega_1 \neq \omega_2$. Let $L \leq N/2$ be the window length and $K = N - L + 1$. These time series are weakly separable if $L\omega_i$ and $K\omega_i$ are integers (in other words, if L and K are divisible by the periods $T_i = 1/\omega_i$). The additional condition $A_1 \neq A_2$ implies strong separability, since the eigenvalues produced by the sinusoids are proportional to their squared amplitudes.

For large N and L two sinusoids are approximately weakly separable if $\omega_1 \neq \omega_2$; the divisibility of L and K by the periods is not necessary, although it can improve the separability. The quality of separability (that influences the accuracy of the reconstruction) depends on the magnitude of $|\omega_1 - \omega_2|$. Close frequencies need much larger time series lengths to obtain a sufficient level of separability.

Under the condition of approximate weak separability, closeness of amplitudes A_1 and A_2 can cause the lack of strong separability. Note also that the frequency interpretation of separability for sinusoids is adequate, since for large L the leakage at the periodograms of sinusoids is small.

2.4 Choice of Parameters in Basic SSA

In this section we discuss the role of parameters in Basic SSA and the principles for their selection. There are two parameters in Basic SSA: the first parameter is the window length L, and the second parameter is, loosely speaking, the way of grouping. In accordance with Sects. 2.3.1.1 and 2.3.2, we assume that the time series under consideration is a sum of a slowly varying trend, different oscillatory components, and a noise.

2.4.1 General Issues

2.4.1.1 Forms of Singular Vectors

We start with mentioning several theoretical results about the eigentriples of several 'simple' time series.

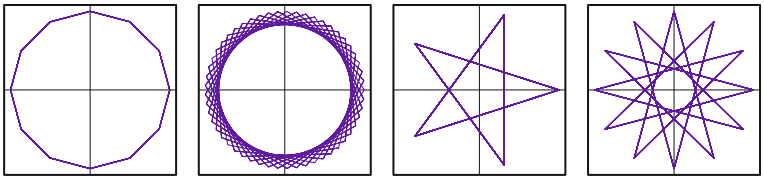

Fig. 2.18 Scatterplots of sines/cosines

Oscillations: exponential-cosine sequences

Consider the series

$$x_n = Ae^{\alpha n}\cos(2\pi\omega n + \varphi), \tag{2.20}$$

where $\omega \in (0, 1/2]$ and $\varphi \in [0, 2\pi)$. Depending on the values of parameters, the exponential-cosine sequence produces the following non-zero eigentriples:

1. *Exponentially modulated harmonic time series with frequency* $\omega \in (0, 1/2)$
   If $\omega \in (0, 1/2)$, then for any L and N the SVD of the trajectory matrix has two non-zero terms. Both eigenvectors (and factor vectors) have the form (2.20) with the same ω and α. In particular, for harmonic series ($\alpha = 0$), the eigenvectors and factor vectors are harmonic series with frequency ω.
2. *Exponentially modulated saw-tooth curve* ($\omega = 1/2$)
   If $\sin(\varphi) \neq 0$, then x_n is proportional to $(-e^{\alpha})^n$. If $\alpha = 0$, then $x_n = A(-1)^n = A\cos(\pi n)$. In this case, for any L the corresponding SVD has just one term. Both singular vectors have the same form as the initial series.

Let $\omega \neq 1/2$ and $\alpha = 0$. Then we have the pure harmonic series defined by (2.20) with $\alpha = 0$. It generates an SVD of order two with singular vectors having the same harmonic form.

Let us consider, for definiteness, the left singular vectors (that is, the eigenvectors) and assume an ideal situation, where $L\omega$ is integer. In this situation, the eigenvectors have the form of sine and cosine sequences with the same frequency ω and the same phases.

Figure 2.18 depicts pairwise scatterplots of four pairs of sin/cosine sequences with zero phases, the same amplitudes and frequencies $1/12$, $10/53$, $2/5$, and $5/12$. Clearly all the points lie on the unit circle. If $T = 1/\omega$ is an integer, then these points are the vertices of the regular T-vertex polygon. For the rational frequency $\omega = q/p < 1/2$ with relatively prime integers p and q, the points are the vertices of the regular p-vertex polygon.

Trends: exponential and polynomial series

1. *The exponential sequence* $x_n = e^{\alpha n}$. For any N and window length L, the trajectory matrix of the exponential series has only one eigentriple. Both singular vectors of this eigentriple are exponential with the same parameter α.

2. *A general polynomial series.* Consider a polynomial series of degree m:

$$x_n = \sum_{k=0}^{m} a_k n^k, \quad a_m \neq 0.$$

   For this series, the order of the corresponding SVD is $m + 1$ and all singular vectors are polynomials of degree not exceeding m.
3. *Linear series.* For a linear series $x_n = an+b, \quad a \neq 0$, with arbitrary N and L, the SVD of the L-trajectory matrix consists of two non-zero terms. The corresponding singular vectors are also linear series.

Note that the exponential-cosine and linear series (in addition to the sum of two exponential series with different rates) are the only series that have at most two non-zero terms in the SVD of their trajectory matrices for any series of length N and window length $L \geq 2$. This fact helps in their SSA identification as components of more complex series.

2.4.1.2 Predicting the Shape of Reconstructed Components

The shape of the eigentriples selected at the grouping stage can help us to predict the shape of the component which is going to be reconstructed from these eigentriples.

1. *If we reconstruct a component of a time series with the help of just one eigentriple and both singular vectors of this eigentriple have similar form, then the reconstructed component will have approximately the same form.* This means that when dealing with a single eigentriple we can often predict the behaviour of the corresponding component of the series. For example, if both singular vectors of an eigentriple resemble linear series, then the corresponding component is also almost linear. If the singular vectors have the form of an exponential series, then the trend has similar shape. Harmonic-like singular vectors produce harmonic-like components (compare this with the results for exponential-cosine series presented at the beginning of this section). This general rule also applies to some other properties of time series including monotonicity (monotone singular vectors generate monotone components of the series).
2. *If $L \ll K$ then the factor vector in the chosen eigentriple has a greater similarity with the component, reconstructed from this eigentriple, than the eigenvector.* Consequently we can approximately predict the result of reconstruction from a single eigentriple by taking into account only the factor vector.
3. *If we reconstruct a series with the help of several eigentriples and the periodograms of their singular vectors are (approximately) supported on the same frequency interval $[a, b]$, then the frequency power of the reconstructed series will be mainly supported on $[a, b]$.* This feature is similar to that of item 1 but concerns several eigentriples and is formulated in terms of the Fourier expansions.

2.4.1.3 Eigenvalues

Let us enumerate several features of singular values of trajectory matrices.

1. The larger the singular value of the eigentriple is, the bigger the weight of the corresponding component of the series. Roughly speaking, this weight may be considered as being proportional to the singular value.
2. By analogy with Principal Component Analysis (PCA), the share of the leading eigenvalues reflects the quality of approximation by the corresponding eigentriples. However, there is a significant difference between Basic SSA and PCA, since PCA performs centering of variables. Since Basic SSA does not perform centering, the share of eigenvalues as a measure of approximation may have little sense. As an example, consider the series $\mathbb{X} = (x_1, \ldots, x_{100})$ with

$$x_n = c + \cos(2\pi n/10) + 0.9\cos(2\pi n/5).$$

   For $L = 50$ and $c > 0.45$ the three leading components provide exact reconstruction of $\mathbb{Y}$ with $y_n = c + \cos(2\pi n/10)$. It may be natural to suggest that the quality of approximation of $\mathbb{X}$ by $\mathbb{Y}$ should not depend on the value of c. However, if we denote $p(c) = (\lambda_1 + \lambda_2 + \lambda_3)/(\lambda_1 + \ldots + \lambda_{50})$, then $p(0.5) \simeq 0.649$, $p(1) \simeq 0.787$ and $p(10) \simeq 0.996$.
3. For series $x_n = A\exp(\alpha n)\cos(2\pi\omega n)$, $\omega \in (0, 0.5)$, if $L\omega$ is integer, then both singular values coincide. If $\alpha \leq 0$ then for large N, L and $K = N - L + 1$, both singular values are close (formally, these values coincide asymptotically, as $L, K \to \infty$). Practically, they are close enough when L and K are several times larger than $T = 1/\omega$.

2.4.1.4 Elementary Reconstructed Components and w-Correlation Matrix

The elementary reconstructed series (recall that they correspond to elementary grouping $I_j = \{j\}$) reflect the final result of reconstruction. If we group two eigentriples, the ith and jth, then the reconstructed time series is equal to the sum of ith and jth elementary reconstructed components.

Let us use **w**-correlations as defined in Sect. 2.3.3 between elementary reconstructed components as separability measures.

While two singular vectors produced by a harmonic are orthogonal and have phase shift approximately equal to $\pi/2$, two associated elementary reconstructed series have approximately zero phase shift and therefore strongly **w**-correlated. If two time series components are strongly separable, then the elementary reconstructed components produced by them are **w**-orthogonal. Therefore, the **w**-correlation matrix $\{\rho_{ij}^{(w)}\}$ between elementary reconstructed components reflects the structure of the series detected by SSA.

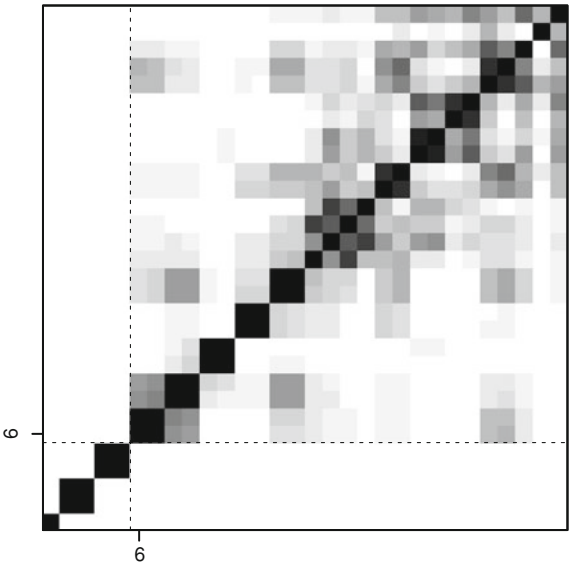

Fig. 2.19 Series (2.21): matrix of **w**-correlations

The **w**-correlation matrix for an artificial series $\mathbb{X}$ with

$$x_n = e^{n/400} + \sin(2\pi n/17) + 0.5\sin(2\pi n/10) + \varepsilon_n, \quad n = 1, \ldots, 340, \quad (2.21)$$

standard Gaussian white noise ε_n, and $L = 85$, is depicted in Fig. 2.19, where **w**-correlations for the first 30 reconstructed components are shown in 20-colour scale from white to black corresponding to the absolute values of correlations from 0 to 1.

The leading eigentriple describes the exponential trend, the two pairs of the subsequent eigentriples correspond to the harmonics, and the large sparkling square indicates the white noise components. Note that this is in full agreement with the theory of (asymptotic) separability.

2.4.2 Grouping for Given Window Length

Assume that the window length L is fixed and we have already made the SVD of the trajectory matrix of the original time series. The next step is to group the SVD terms in order to solve one of the problems discussed in Sect. 2.3.2. We suppose that this problem has a solution; that is, the corresponding terms can be found in the SVD, and the result of the proper grouping would lead to the (approximate) separation of the time series components (see Sect. 2.3.3).

Therefore, we have to decide what the proper grouping is and how to construct it. In other words, we need to identify the eigentriples corresponding to the time series

component we are interested in. Since each eigentriple consists of an eigenvector (left singular vector), a factor vector (right singular vector) and a singular value, this needs to be achieved using only the information contained in these vectors (considered as time series) and in the singular values.

2.4.2.1 Preliminary Analysis

The preliminary analysis of the time series is not necessary but it can be helpful for easier interpretation of the results of SSA processing.

The following steps can be performed.

1. Observe the time series as a whole.
   - One can inspect the general shape of trend, its complexity and hence how many trend components one can expect in the SVD expansion.
   - Based upon the form of the time series and its nature, one can expect some oscillations and their periods. For example, for seasonal monthly data, the period 12 is natural. If some period T is expected, then its divisors by integers (the result should be ≥ 2) are likely to be found in SSA decomposition. For monthly seasonal data they are $12, 6 = 12/2, 4 = 12/3, 3 = 12/4, 2.4 = 12/5$ and $2 = 12/6$.
2. Explore the time series periodogram.
   - Periodogram peaks reflect the expected periods that can be found in SSA decomposition.
   - Equal or close values at the peaks indicate a potential problem of the lack of strong separability.

For an example of a preliminary analysis of this kind, see the case study in Example 2.1 (Sect. 2.3.1.1), where Basic SSA was used to analyze the 'Rosé wine' series.

2.4.2.2 How to Group

For illustration, we provide references to the figures below in the description of the general recommendations. As an example, we consider the 'Fortified wine' series (Fig. 2.16), which has already been analysed in Sect. 2.3.1.2.

General recommendations

1. Inspect the one-dimensional graphs of eigenvectors, factor vectors or elementary reconstructed components. Find slowly varying components. Note that any slowly varying component can be corrupted by oscillations if the trend and oscillating components are not separated. Elementary reconstructed components show whether the oscillating component is suppressed by the diagonal averaging. Most likely, the presence of the mix-up between the components is caused by the lack

of strong separability. Changes in the window length and application of different preprocessing procedures can improve strong separability. All slowly varying components should be grouped into the trend group. Figure 2.20 shows both the trend eigenvector and the trend reconstruction.

2. Consider two-dimensional plots of successive eigenvectors. Find regular p-vertex polygons, may be, in the form of a spiral. Group the found pairs of eigentriples. The harmonic with period 2 produces 1 eigentriple and therefore can be found at the one-dimensional graphs of, say, eigenvectors as a saw-tooth graph. See Fig. 2.21 with scatterplots and the reconstructed series in Fig. 2.17.
3. If there is a fundamental period T (e.g. seasonality with period 12), then special efforts should be made at finding the harmonics with periods that are divisors of T. Also, to reconstruct the whole periodic component with given period T, the pairs with periods T/k, $k = 1, \ldots, \lfloor T/2 \rfloor$ should be grouped, see Fig. 2.16, where the reconstruction of the whole seasonality is depicted.
4. The **w**-correlation matrix of elementary components can help in finding the components if they are not well separated and the techniques described above were not successful. Blocks of two correlated components reflect a harmonic. A block of 4 correlated consequent components probably corresponds to two mixed pairs of harmonics. This can be checked by, for example, their periodogram analysis. Since noise is, in a sense, a mixture of many not-separable components, the **w**-correlation matrix can help to determine the number of components to identify.

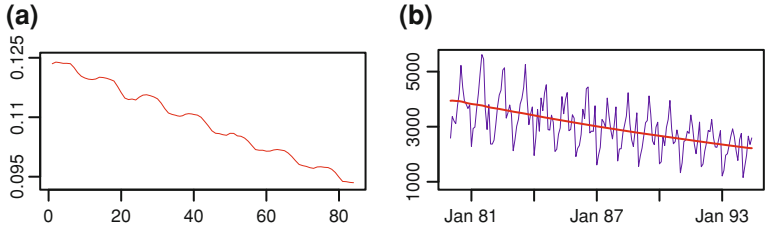

Fig. 2.20 Fortified wine: trend component. **a** Trend eigenvector. **b** Trend reconstruction and the initial series

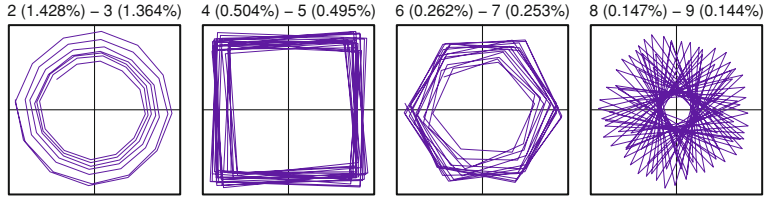

Fig. 2.21 Fortified wine: Scatterplots of eigenvector pairs corresponding to periods 12, 4, 6, 2.4

2.4.2.3 How to Check the Results of Reconstruction

1. Any statistical testing is only possible when some assumptions are made. It could be a parametric model of the signal and noise. Nonparametric models usually require availability of a sample taken from the same distribution. The SSA procedure positions itself as a model-free technique and therefore the justification of the results is complicated. Hence, the interpretability of the resultant series components is very important. For example, the extraction of the component with period 7 for monthly data is often more doubtful than, for example, half-year periodicity.
2. While signals could have very different forms and structures, noise frequently looks like white or rarer red noise. If there are reasons to assume a model of noise, then one can routinely test the corresponding hypothesis to confirm the results. In any case, the periodogram of the residual or their autocorrelation function can show if there is a part of the signal in the residual.
3. To test the specific hypothesis that the series is the red noise (AR(1) model with positive correlations), Monte Carlo SSA [2] may be used. The declared advantage of this test is its power with respect to the alternative of red noise corrupted by a signal for short time series.

2.4.2.4 Methods of Period Estimation

Since period estimation can be very useful in the process of identification of periodic components, let us enumerate several methods of estimation that can be applied within the framework of SSA.

1. A conventional method for frequency estimation is periodogram analysis. We can apply it for estimation of frequencies of eigenvectors, factor vectors as well as reconstructed components. This can be effective for long series (and for large window lengths if we want to consider eigenvectors). If the time series is short, then the resolution of the periodogram analysis is low.
2. We can estimate the period using both eigenvectors (or factor vectors) produced by a harmonic. If the eigenvectors have already been calculated, this method is very fast. Consider two eigentriples, which approximately describe a harmonic component with frequency $0 < \omega_0 < 0.5$. Then the scatterplot of their eigenvectors can be expressed as a two-dimensional curve with Euclidean components of the form

$$x(n) = r(n)\cos(2\pi\omega(n)n + \varphi(n)), \quad y(n) = r(n)\sin(2\pi\omega(n)n + \varphi(n)),$$

where the functions r, ω and φ are close to constants and $\omega(n) \approx \omega_0$. The polar coordinates of the curve vertices are $(r(n), \delta(n))$ with $\delta(n) = 2\pi\omega(n)n + \varphi(n)$. Since $\Delta_n \stackrel{\text{def}}{=} \delta(n+1) - \delta(n) \approx 2\pi\omega_0$, one can estimate ω_0 by averaging polar

angle increments Δ_n ($n = 1, \ldots, L$). The same procedure can be applied to a pair of factor vectors.

3. We can also use the subspace-based methods of signal processing including ESPRIT, MUSIC, and others, see Sect. 3.8. These methods have high resolution and can be applied to short time series if we were able to separate signal from noise accurately enough. An important common feature of these methods is that they do not require the sinusoids to be separated from each other.

2.4.3 Window Length

The window length L is the main parameter of Basic SSA: its inadequate choice would imply that no grouping activity will lead to a good SSA decomposition.

There is no universal rule for the selection of the window length. The main difficulty here is caused by the fact that variations in L may influence both weak and strong separability features of SSA, i.e., both the orthogonality of the appropriate subseries of the original series and the closeness of the singular values. However, there are several general principles for the selection of the window length L that have certain theoretical and practical grounds. Let us discuss these principles.

2.4.3.1 General Principles

1. The SVDs of the trajectory matrices, corresponding to the window lengths L and $K = N - L + 1$, are equivalent (up to the symmetry: left singular vectors $\leftrightarrow$ right singular vectors). Therefore, we can always assume $L \le N/2$.
2. Assuming $L \le N/2$, the larger L is, the more detailed is the decomposition of the time series. The most detailed decomposition is achieved when $L \simeq N/2$ unless the series has finite rank d, see Sect. 2.3.1.2. In this case, SSA decompositions with any L such that $d \le L \le N + 1 - d$ are equivalent.
3. Small window lengths act like smoothing linear filters of width $2L - 1$. For small L, the filter produced by the leading eigentriple is similar to the Bartlett filter with triangular coefficients (see Sect. 3.9.3).
4. The following are the effects related to weak separability.

   - As the results concerning weak separability of time series components are mostly asymptotic (when $L, K \to \infty$), in the majority of examples to achieve better (weak) separation one has to choose large window lengths. In other words, the use of small L could lead to a mix-up between components which otherwise would be interpretable. Unless two time series are deterministic and exactly separable, there is no convergence of the reconstruction error to zero if L is fixed and $K \to \infty$ (see for details [13]).
   - If the window length L is relatively large, then the (weak) separation is stable with respect to small perturbations in L.

- On the other hand, for specific series and tasks, some concrete recommendations can be given for the window length selection; these recommendations can be very useful for relatively small N (see Sect. 2.4.3.3 below).

5. It is hard to successfully overcome (only by varying L) the difficulty related to the closeness of singular values; that is, to the absence of strong separability when there is an approximate weak separability. Let us mention two general points related to the closeness of the singular values.
   - For the series with complex structure, too large values of L can lead to an undesirable decomposition of the series components of interest, which in turn may yield their mixing with other series components. This is an unpleasant possibility, especially since a significant reduction of L can lead to a poor quality of the (weak) separation.
   - Alternatively, sometimes in these situations even a small variation in the value of L can reduce mixing and lead to a better separation of the components and hence provide a transition from weak to strong separability.
6. Whatever the circumstances, it is always a good idea to repeat SSA analysis several times using different values of L.

2.4.3.2 Window Length for Extraction of Trends and Smoothing

1. Trends

In the problem of trend extraction, a possible contradiction between the requirements for weak and strong separability emerges most frequently.

Since trend is a relatively smooth curve, its separability from noise and oscillations requires large values of L. On the other hand, if the trend has a complex structure, then for very large L it can only be described using a substantial number of eigentriples with relatively small singular values. Some of these singular values could be close to those generated by oscillations and/or noise time series components.

This happens in the example 'Births', see Sect. 2.2.2.2, where the window length of order 1000 and more (the series length is 5113) leads to the situation where the components of the trend are mixed up with the components of the annual and half-year periodicities (other aspects relating to the choice of the window length in this example are discussed below).

If the trend is simple and dominates the rest of the series, then the choice of L does not present any difficulty (that is, L can be taken from a wide range). Let $\mathbb{X} = \mathbb{X}^{(1)} + \mathbb{X}^{(2)}$, where $\mathbb{X}^{(1)}$ is a trend and $\mathbb{X}^{(2)}$ is the residual. The notion of 'simplicity' can be understood as follows:

- From the theoretical viewpoint, the series $\mathbb{X}^{(1)}$ is well approximated by a series with finite and small rank d, see Sect. 2.3.1.2 for a description of the series of finite rank.

- We are interested in the extraction of the general tendency of the series rather than of the refined trend.
- In terms of frequencies, the periodogram of the series $\mathbb{X}^{(1)}$ is concentrated in the domain of small frequencies.
- In terms of SSA decomposition, the few first eigentriples of the decomposition of the trajectory matrix of the series $\mathbb{X}^{(1)}$ are enough for a reasonably good approximation of it, even for large L.

Assume also that the series $\mathbb{X}^{(1)}$ is much 'larger' than the series $\mathbb{X}^{(2)}$ (for instance, the inequality $\|\mathbb{X}^{(1)}\|_{\mathrm{F}} \gg \|\mathbb{X}^{(2)}\|_{\mathrm{F}}$ is valid).

Suppose that these assumptions hold and the window length L provides a certain (weak, approximate) separation between the time series $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$. Then we can expect that in the SVD of the trajectory matrix of the series $\mathbb{X}$ the leading eigentriples will correspond to the trend $\mathbb{X}^{(1)}$; i.e., they will have larger singular values than the eigentriples corresponding to $\mathbb{X}^{(2)}$. In other words, we expect strong separability to occur. Moreover, the window length L, sufficient for the separation, should not be very large in this case in view of the 'simplicity' of the trend.

This situation is illustrated by the example 'Production' (Fig. 2.1a, b), where both trend versions are described by the leading eigentriples. However, more refined versions of the trend can be difficult to extract.

Much more complicated situations arise if we want to extract a refined trend $\mathbb{X}^{(1)}$, when the residual $\mathbb{X}^{(2)}$ has a complex structure (for example, it includes a large noise component) with $\|\mathbb{X}^{(2)}\|_{\mathrm{F}}$ being large. Then large L can cause not only mixing of the ordinal numbers of the eigentriples corresponding to $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$ (this is the case in the 'US unemployment' example), but also closeness of the corresponding singular values, and therefore a lack of strong separability.

2. Smoothing

The recommendations concerning the selection of the window length for the problem of smoothing are similar to those for trend extraction. This is related to the fact that these two problems are closely related. Let us describe the effects of the window length in the language of frequencies.

If we treat smoothing as a removal of the high-frequency part of the series, then we have to choose L large enough to provide separation of this low-frequency part from the high-frequency one. If the powers of all low frequencies of interest are significantly larger than those of the high ones, then the smoothing problem is not difficult, and our only job is to gather several leading eigentriples. This is the case for the 'Tree rings' and 'White dwarf' examples of Sect. 2.2.1.2. Here, the larger L we take, the narrower the interval of low frequencies we can extract.

For instance, in Sect. 2.2.1.2, the smoothing of the series 'White dwarf' has been performed with $L = 100$, with the result of the smoothing being described by the leading 11 eigentriples. In the periodogram of the residuals (see Fig. 2.22a) we can see that for this window length the powers of the frequencies in the interval

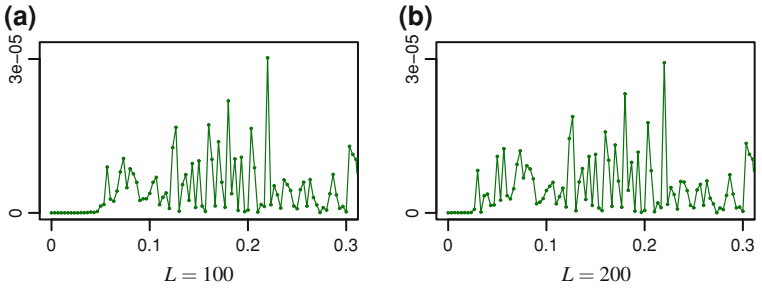

Fig. 2.22 White dwarf: periodograms of residuals. **a** $L = 100$. **b** $L = 200$

[0, 0.05] are practically zero. If we take $L = 200$ and 16 leading eigentriples for the smoothing, then this frequency interval is reduced to [0, 0.03] (see Fig. 2.22b). At the same time, for $L = 10$ and two leading eigentriples, the result of smoothing contains the frequencies from the interval [0, 0.09].

Visual inspection shows that all smoothing results look similar. Also, their eigenvalue shares are equal to $95.9\,\% \pm 0.1\,\%$. Certainly, this effect can be explained by the following specific feature of the series: its frequency power is highly concentrated in the narrow low-frequency region.

2.4.3.3 Window Length for Periodicities

The problem of choosing the window length L for extracting a periodic component $\mathbb{X}^{(1)}$ out of the sum $\mathbb{X} = \mathbb{X}^{(1)} + \mathbb{X}^{(2)}$ has certain specificities related to the correspondence between the window length and the period. These specificities are very similar for the pure harmonics, for complex periodicities and even for modulated periodicities. Thus, we do not consider these cases separately.

1. For the problem of extraction of a periodic component with period T, it is natural to measure the length of the series in terms of the number of periods: if $\mathbb{X}^{(1)}$ is asymptotically separable from $\mathbb{X}^{(2)}$, then to achieve the separation we must have, as a rule, the length of the series N such that the ratio N/T is at least several units.
2. For relatively short series, it is preferable to take into account the conditions for pure (nonasymptotic) separability (see Sect. 2.3.3); if one knows that the time series has a periodic component with integer period T (for example, $T = 12$), then it is advisable to take the window length L proportional to T. Note that from the theoretical viewpoint, $N-1$ must also be proportional to T.
3. In the case of long series, the requirement for L/T and $(N-1)/T$ to be integers is not that important. In this case, it is recommended to choose L as large as possible (for instance, close to $N/2$, if the computer facilities allow us to do this). Nevertheless, even in the case of long series it is recommended to choose L so that L/T is an integer.

4. If the series $\mathbb{X}^{(2)}$ contains a periodic component with period $T_1 \approx T$, then to extract $\mathbb{X}^{(1)}$ we generally need a larger window length than for the case when such a component is absent (see Sect. 2.3.3).

To demonstrate the effect of divisibility of L by T, let us consider the 'Eggs' example (eggs for a laying hen, monthly, U.S., from January 1938 to December 1940, [21, Chap. 45]). This series has a rather simple structure: it is the sum of an explicit annual oscillation (though not a harmonic one) and a trend, which is almost constant. However, this series is short and therefore the choice of L is very important.

The choice $L = 12$ allows us to extract simultaneously all seasonal components (12, 6, 4, 3, 2.4, and 2-months harmonics) as well as the trend. The graph in Fig. 2.23 depicts the initial series and its trend (thick line), which is reconstructed from the leading eigentriple.

Figures 2.24a, b depict the matrices of **w**-correlations for the full decomposition of the series with $L = 12$ and $L = 18$. It is clearly seen that for $L = 12$ the matrix is essentially diagonal, which means that the eigentriples related to the trend and different seasonal harmonics are almost **w**-uncorrelated. This means that the choice $L = 12$ allows us to extract all harmonic components of the series.

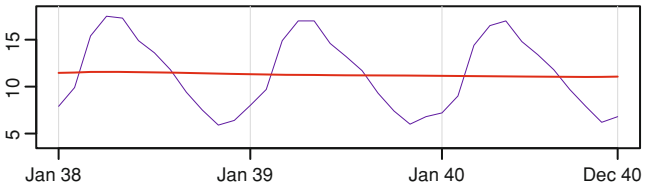

Fig. 2.23 Eggs: initial series and its trend

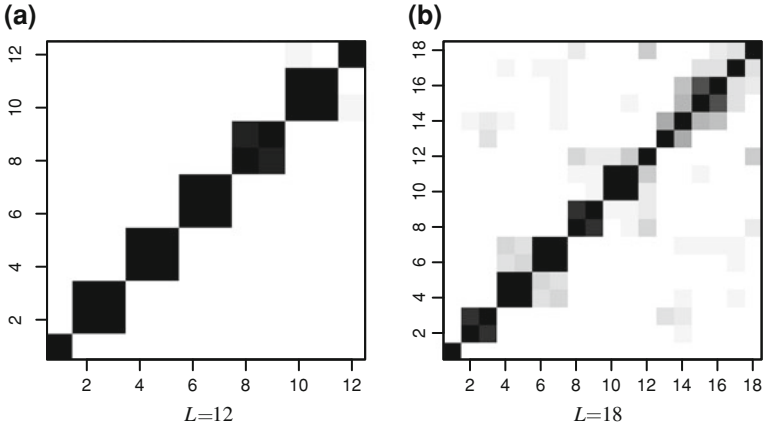

Fig. 2.24 Eggs: **w**-correlations. **a** $L = 12$. **b** $L = 18$

For $L = 18$ (that is, when the period 12 does not divide L), only the leading seasonality harmonics can be extracted properly.

The choice $L = 13$ would give results that are slightly worse than for $L = 12$, but much better than for $L = 18$. This confirms the robustness of the method with respect to small variations in L.

2.4.3.4 Refined Structure

In doing simultaneous extraction of different components from the whole series, all the aspects discussed above should be taken into account. For instance, in basically all examples of Sect. 2.2, where the periodicities were the main interest, the window length was a multiple of the periods. At the same time, if in addition trends were to be extracted, L was reasonably large (but smaller than $N/2$) to avoid the mix-up between the components.

To demonstrate the influence of the window length on the result of the decomposition, let us consider the example 'Births' (Sect. 2.2.2.2). In this series (daily data for about 14 years, $N = 5113$) there is a one-week periodicity ($T_1 = 7$) and an annual periodicity ($T_2 = 365$). Since $T_2 \gg T_1$, it is natural to take the window length as a multiple of T_2.

The choice $L = T_2$, as was shown in Sect. 2.2.2.2, guarantees a simultaneous extraction of both weekly and annual periodicities. Moreover, this window length also allows us to extract the trend of the series (see Fig. 2.25) using just one leading eigentriple. Note that these results are essentially the same as for $L = 364$ and $L = 366$.

At the same time, if we would choose $L = 3T_2 = 1095$ or $L = 7T_2 = 2555$, then the components of the trend will be mixed up with the components of the annual and half-year periodicities; this is a consequence of the complex shape of the trend and the closeness of the corresponding eigenvalues. Thus, choosing the values of L which are too large leads to the loss of strong separability.

If the problem of separation of the trend from the annual periodicity is not important, then values of L larger than 365 work well. If the window length is large, we can separate the global tendency of the series (trend + annual periodicity) from the

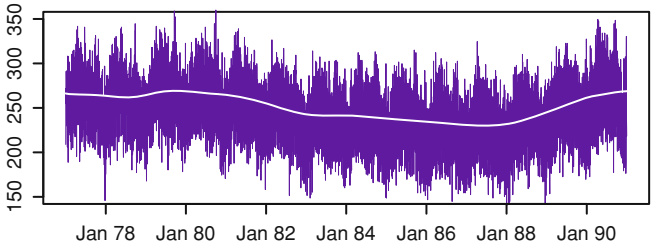

Fig. 2.25 Births: trend

weekly periodicity + noise even better than for $L = 365$ (for $L = 1095$ this component is described by several dozen eigentriples rather than by 5 eigentriples for $L = 365$). In this case, the weekly periodicity itself is perfectly separable from the noise as well.

In even more complex cases, better results are often achieved by the application of the so-called Sequential SSA, see Sect. 2.5.5. In Sequential SSA, after extraction of a component with certain L, Basic SSA with different value of L is applied again, to the residual series obtained in the first run of SSA.

2.4.4 Signal Extraction

2.4.4.1 Specifics of Extraction of the Signal

Sometimes, the structure of the deterministic component of the time series which can be called a signal is not important. In this case, the following three simple observations may help achieve better separation of the signal from noise.

1. Since we are interested in the signal as a whole, separability of signal components is not essential. As a consequence, for the signal containing a periodic component, divisibility of the window length by the period is not important for separation of the signal from noise. However, if the window length is divisible by the period, it is easier to identify the signal components.
2. Since the signal components are often dominating, the only parameter of grouping is the number r of the leading components related to the signal. This number can be estimated using the matrix of **w**-correlations between elementary reconstructed components. In the example 'White dwarf' (Sect. 2.2.1.2) with $L = 100$, the matrix of the absolute values of **w**-correlations of the reconstructed components produced from the leading 30 eigentriples is depicted in Fig. 2.26 in the manner of Fig. 2.19. Splitting all eigentriples into two groups, from the first to the 11th and the rest, gives a decomposition of the trajectory matrix into two almost orthogonal blocks, with the first block corresponding to the smoothed version of the original series and the second block corresponding to the residual, see Fig. 2.3a, b in Sect. 2.2.1.2.
3. The problem of extraction of signal of finite rank from noisy time series is very well elaborated. In particular, there are different methods of rank estimation (see below). These methods can be used while identifying the components in SSA.

2.4.4.2 Methods of Estimation of the Rank of the Signal

Two types of methods of rank estimation are used in signal processing. The first type is related to the so-called AIC-methods. They use some information criteria [36], are based on the maximum likelihood function and therefore could only be applied to the series with given parametric model of the residuals (usually, Gaussian noise).

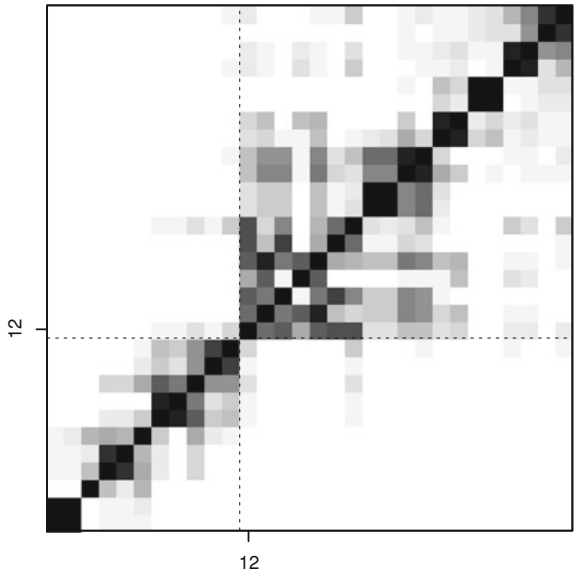

Fig. 2.26 White dwarf: matrix of **w**-correlations

The second type of methods can be applied for general series. Let the method estimate some time series characteristic. Then the accuracy of this estimation for different values of the assumed series rank r can point towards the proper value of r.

For example, the proper rank can be estimated on the base of the accuracy of forecasts of historical data. Or, more generally, one can consider several time series points as artificial missing values and their imputation accuracy serves as a characteristic for the choice of the best rank.

For signals of finite rank, specific methods can also be suggested. For example, the ESTER method [6] is based on features of the ESPRIT method as a method of parameter estimation (see for details Sects. 3.8.2 and 3.8.2.3).

2.4.5 Automatic Identification of SSA Components

While the choice of the window length is well supported by SSA theory, the procedure for choosing the eigentriples for grouping is much less formal.

Let us describe several tests for the identification of SSA components constituting parts of the trend or related to periodicities. We assume that the components to be identified are (approximately) separated from the rest of the series.

The tests described below can be used differently. First, these tests can provide some hints for making the grouping. This is a safe way of using the tests and we shall consider the tests from this viewpoint only. Second, the tests can be used as the base for the so-called *batch processing*. If there is a large set of similar time series,

then a part of them can be used for the threshold adjustment. Similar to many other automatic procedure, the results of SSA batch processing may be misleading as many deviations from the model assumptions are possible. Note also that any choice of the threshold should take into consideration the following two conflicting types of decision error: (i) not to choose the proper SVD components (it is more important), and (ii) to choose wrong SVD components. Certainly, to estimate probabilities of these errors, a stochastic model of the time series should be specified.

2.4.5.1 Grouping Based on w-Correlations

The first approach is based on the properties of the **w**-correlation matrix $\{\rho_{ij}^{(w)}\}$ for the separability identification, see Sect. 2.4.1.4. This idea was used in different SSA-processing procedures. For example, Bilancia and Campobasso [7] consider hierarchical clustering with the dissimilarity measure $1-|\rho_{ij}^{(w)}|$ and complete linkage, while Alonso and Salgado [3] use the k-means clustering procedure.

Let us consider two **w**-correlation matrices with full decompositions depicted in Figs. 2.19 and 2.26. The dissimilarity matrix consisting of $1 - |\rho_{ij}^{(w)}|$ along with the average linkage provides the proper split into two clusters for the White dwarf data. The first cluster consists of ET1–11 and the second cluster corresponds to noise. The same procedure for the example of Fig. 2.19 gives the first cluster consisting of ET1 only, while the complete linkage provides the cluster of ET1–5. Note that the division onto four groups (ET1; ET2,3; ET4,5; the rest) is the most appropriate for the average linkage. It seems that the average linkage is a good choice if the number of clusters is known. The choice of the number of clusters can be performed by the conventional tools of Cluster Analysis. Also, large **w**-correlations between grouped components from the clusters can help in distinguishing false clusters.

2.4.5.2 Identification of Trend

Since we define trend as any slowly-varying component of the time series, analysis of frequencies is a suitable tool for trend identification. The authors of [35] suggest to use the number of zero crossings or the Kendall's test to find slowly-varying eigenvectors. A rather general approach is to use the periodogram and consider the contribution of low frequencies as a test; see e.g. [1], where the emphasis is made on the procedure of an automatic choice of the identification thresholds.

Consider the periodogram (2.10) of a series $\mathbb{Y}$ of length M and define

$$T(\mathbb{Y};\omega_1,\omega_2) = \sum_{k:\omega_1\le k/M\le\omega_2} I_y^M(k/M), \tag{2.22}$$

where $I_y^M(k/M) = M\,\Pi_y^M(k/M)/||\mathbb{Y}||^2$, Π_y^M is defined in (2.10). In view of (2.11), $0 \le T(\mathbb{Y}; \omega_1, \omega_2) \le 1$ for any $0 \le \omega_1 \le \omega_2 \le 0.5$. Let us choose the bounding frequency ω_0, $0 \le \omega_0 \le 0.5$, and set up a threshold T_0, $0 \le T_0 \le 1$.

Below we formulate a generic test for deciding whether a given SSA component is slowly varying. This test can be applied to eigenvectors, factor vectors and elementary reconstructed components considered as time series. Let $\mathbb{Y}$ be the series we are going to test.

Trend test *T. A given component* $\mathbb{Y}$ *is related to the trend if* $T(\mathbb{Y}; 0, \omega_0) \ge T_0$.

The choice of the bounding frequency ω_0 depends on how we want the trend to look like. For example, for monthly data with possible seasonality it is recommended to choose $\omega_0 < 1/12$.

If we consider the results of trend tests as hints for the eigentriple identification, it is not necessary to set the threshold value T_0, since we can simply consider the values of the test statistics $T(\mathbb{Y}; 0, \omega_0)$ for the series $\mathbb{Y}$ (the eigenvectors or the elementary reconstructed components) related to each eigentriple.

Let us consider the 'Production' example (Sect. 2.2.1.1, Fig. 2.1b), where a reasonably accurate trend is described by the three leading eigentriples. If we choose $\omega_0 = 1/24$ and $T_0 = 0.9$, then the described procedure identifies ET1–3,6,8,11,12; that is, the trend identified (see Fig. 2.27) is even more accurate than that depicted in Fig. 2.1b. The result is stable with respect to the choice of the threshold and is exactly the same when we apply it to eigenvectors, factor vectors or reconstructed components. The values of the test $T(\cdot\,; 0, 1/24)$ applied to the 12 leading factor vectors are respectively: 0.9999, 0.9314, 0.9929, 0.0016, 0.0008, 0.9383, 0.0053, 0.9908, 0.0243, 0.0148, 0.9373, 0.9970. If we are interested in general tendency, then the test T with $\omega_0 = 1/120$ identifies one leading component only, the same result as in Fig. 2.1a.

For the 'Rosé wine' example, where the trend was extracted by ET1, 12, and 14, the test $T(\cdot; 0, 1/24)$ applied to 16 leading eigenvectors gives 0.9993 for ET1, 0.8684 for ET12, 0.9839 for ET14 and values smaller than 0.02 for all other eigentriples. This outcome perfectly agrees with visual examination.

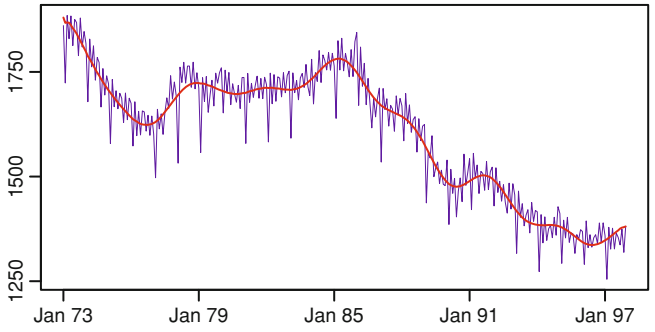

Fig. 2.27 Production: automatically identified refined trend

2.4.5.3 Identification of Harmonics

The method for the identification of the harmonic components is based on the study of the corresponding singular vectors. Ideally, any harmonic component produces two eigenvectors, which are sine and cosine sequences if L and $K = N - L + 1$ are divisible by the period of the harmonic. Also, if $\min(L, K) \to \infty$ then the pair of the corresponding either left or right singular vectors tends to the sine and cosine sequences, correspondingly.

Define for $H, G \in \mathsf{R}^L$

$$\rho(G, H) = \max_{0 \le k \le L/2} \gamma(G, H; k), \text{ where } \gamma(G, H; k) = 0.5(I_g^L(k/L) + I_h^L(k/L))$$

and the quantity I is the same as in (2.22). It is clear that $\rho(G, H) \le 1$ and that for any integer $L\omega$ the equality $\rho(G, H) = 1$ is valid if and only if $h_n = \cos(2\pi\omega n + \varphi)$ and $g_n = \sin(2\pi\omega n + \varphi)$. Also, for arbitrary ω, $\rho(G, H) \to 1$ as $L \to \infty$.

Therefore, the value of $\rho(U_i, U_j)$ (as well as $\rho(V_i, V_j)$) can be used as an indicator of whether the pair of eigenvectors U_i, U_j (or factor vectors V_i, V_j) is produced by a harmonic component.

The case of amplitude-modulated harmonics is slightly more complicated. Let us consider the identification of the exponentially damped sine waves; recall that these waves are naturally generated by SSA. Both eigenvectors (and factor vectors) have the same form (2.20) with the same frequency ω and the exponential rate α. Therefore we generally can apply the $\rho(G, H)$ for their identification. However, the modulation leads to decreasing of $\rho(G, H)$ and this should be accounted for choosing the threshold value.

Let us introduce the test which is a modification of the test suggested in [35] to take into consideration a leakage caused by possible modulation of the harmonics and location of their frequencies between positions in the periodogram grid. Define

$$\tau(G, H) = \max_{0 \le k \le L/2 - m_0} \sum_{j=0}^{m_0 - 1} \gamma(G, H; k + j),$$

where m_0 is some integer.

Note that below we use the result stating that an exponentially damped sinusoid produces asymptotically equal eigenvalues. We therefore consider only adjacent eigenvectors.

Harmonic test τ. *An eigenvector pair* (U_j, U_{j+1}) *is identified as corresponding to some damped sinusoid if the periodograms of* U_j *and* U_{j+1} *are peaked at frequencies differing not more than* m_0/L *and* $\tau(U_j, U_{j+1}) \ge \tau_0$ *for given threshold* $\tau_0 \in [0, 1]$.

Here m_0 should be chosen equal to 0 if the period is known and we can choose L such that L and K are divisible by the period. Otherwise we choose $m_0 = 1$.

Note that the procedure needs special treatment of the components with frequencies 0 and 0.5: the frequency 0 should not be considered as a candidate for periodicity, while the sine wave with frequency 0.5 is the saw-tooth function and produces just one component with frequency 0.5. Also, the procedure can be supplemented with the frequency estimation (see Sect. 2.4.2.4) and the results can be filtered in accordance with the chosen frequency range.

Let us apply the τ-test to the 'Production' example considered in Sects. 2.2.1.1 and 2.2.2.1. This time series has a trend of complex form and we need to set a period-based filter to distinguish between the cyclic components of the trend and the seasonal components. Assume that all possible periods fall into the interval [2,13]. Then the τ-test with thresholds τ_0 from the range 0.86–0.96 identifies the same seasonal components as were chosen in Sect. 2.2.2.1 by visual inspection except for the pair ET19–20 (period 12) with $\tau(U_{19}, U_{20}) = 0.85$. This is explained by the sharp decrease of the harmonic with period 12 and a poor separability of the annual harmonic.

Warning. Above we considered examples with well separable components. However, if the separability is poor, then the automatic procedure typically fails. Therefore, the automatic identification is useful for grouping but can not replace the techniques that improve separability.

2.5 Some Variations of Basic SSA

In some circumstances, a clever modification of Basic SSA or its skillful computer implementation may visibly improve either the accuracy or efficiency of SSA. In this section, we describe several techniques for making modifications to Basic SSA and discuss some computer implementations of SSA. We start this section with a short discussion concerning preliminary preprocessing of time series that can be considered as a part of the SSA processing.

2.5.1 Preprocessing

There are two standard ways of preprocessing, log-transformation and centering. The log-transformation has already been discussed in Sect. 2.3.1.3. It is a very important feature of SSA that even if the main model of the series is multiplicative, SSA can work well without the use of the log-transformation. It is an essential advantage of SSA over many other methods as full multiplicativity is a very strong assumption and generally it is not met in practice. For example, the time series 'War', 'US unemployment' and 'Germany unemployment' are similar to multiplicative time series. However, the log-transformation does not provide constancy of seasonal amplitudes while the main assumption of many conventional methods is similarity of the seasonal components from year to year.

Centering of time series (that is, the subtraction of the general mean) is necessary for the application of methods of analysis of stationary series. Usually, these

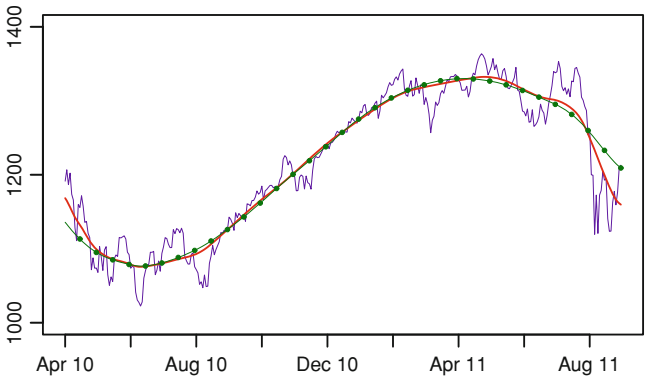

Fig. 2.28 S&P500: trends extracted from the initial series and from the centered series

methods deal with estimation of spectral characteristics of time series. This means that centering has little sense for time series with trends. From the viewpoint of SSA, centering can both increase and decrease the rank of the time series. For example, the trend of 'Fortified wine' (Sect. 2.3.1.2, Fig. 2.16) is very well described by one leading eigentriple with share 94.6 % ($L = 84$), i.e., it is well approximated by an exponential series of rank 1. After centering, the trend is described by two eigentriples ET3 (11.2 %) and ET4 (8.9 %). The accuracy of trend extraction is worse and the extraction of trend is more complicated since the corresponding eigentriples are no longer the leading eigentriples.

Sometimes the centering of the series may be very useful. As an example, consider the series 'S&P500', the free-float capitalization-weighted index of the prices of 500 large-cap common stocks actively traded in the United States. Its trend has complex form. However, in the timeframe of 1.5 year the trend of the centered series can be approximated by a sinusoid. The resultant trends are depicted in Fig. 2.28 together with the initial series. The first trend is extracted from the initial series by ET1–3 (the bold line), the second trend is extracted from the centered series by ET1–2 (the line with black dots), $L = 170$. The former trend is more detailed, while the latter one is more stable.

2.5.2 Centering in SSA

Consider the following extension of Basic SSA. Assume that we have selected the window length L. For $K = N - L + 1$, consider a matrix $\mathbf{A}$ of dimension $L \times K$ and rather than using the trajectory matrix $\mathbf{X}$ of the series $\mathbb{X}$ we shall use the matrix $\mathbf{X}^\star = \mathbf{X} - \mathbf{A}$. Let $\mathbf{S}^\star = \mathbf{X}^\star(\mathbf{X}^\star)^{\mathrm{T}}$, and denote by λ_i and U_i $(i = 1, \ldots, d)$ the nonzero eigenvalues and the corresponding orthonormal eigenvectors of the matrix $\mathbf{S}^\star$. Setting $V_i = (\mathbf{X}^\star)^{\mathrm{T}} U_i / \sqrt{\lambda_i}$ we obtain the decomposition

$$\mathbf{X} = \mathbf{A} + \sum_{i=1}^{d} \mathbf{X}_i^{\star} \tag{2.23}$$

with $\mathbf{X}_i^{\star} = \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$, instead of the standard SVD (2.2). At the grouping stage, the matrix $\mathbf{A}$ will enter one of the resultant matrices as an addend. In particular, it will produce a separate time series component after diagonal averaging.

If the matrix $\mathbf{A}$ is orthogonal to all $\mathbf{X}_i^{\star}$, then the matrix decomposition (2.23) yields the decomposition $||\mathbf{X}||_{\mathrm{F}}^2 = ||\mathbf{A}||_{\mathrm{F}}^2 + \sum_{i=1}^{d} ||\mathbf{X}_i^{\star}||_{\mathrm{F}}^2$ of the squared norms of the corresponding matrices. Then $||\mathbf{A}||_{\mathrm{F}}^2 / ||\mathbf{X}||_{\mathrm{F}}^2$ corresponds to the share of $\mathbf{A}$ in the decomposition.

Here we briefly consider two ways of choosing the matrix $\mathbf{A}$, both of which are thoroughly investigated in [14, Sects. 1.7.1 and 6.3].

Single centering is the row centering of the trajectory matrix. Here $\mathbf{A} = [\mathrm{E}(\mathbf{X}) : \ldots : \mathrm{E}(\mathbf{X})]$, where ith component of the vector $\mathrm{E}(\mathbf{X})$ $(i = 1, \ldots, L)$ is equal to the average of the ith components of the lagged vectors $X_1, \ldots, X_K$. Basic SSA with single centering can have advantage over the standard Basic SSA if the series $\mathbb{X}$ has the form $\mathbb{X} = \mathbb{X}^{(1)} + \mathbb{X}^{(2)}$, where $\mathbb{X}^{(1)}$ is a constant series and $\mathbb{X}^{(2)}$ oscillates around zero.

For *double centering*, the SVD is applied to the matrix computed from the trajectory matrix, by subtracting from each of its elements the corresponding row and column averages and by adding the total matrix average. Basic SSA with double centering can outperform the standard Basic SSA if the series $\mathbb{X}$ can be expressed in the form $\mathbb{X} = \mathbb{X}^{(1)} + \mathbb{X}^{(2)}$, where $\mathbb{X}^{(1)}$ is a linear series (that is, $x_n^{(2)} = an + b$) and $\mathbb{X}^{(2)}$ oscillates around zero. As shown in [14, Sects. 1.7.1 and 6.3], Basic SSA with double centering can have serious advantage over linear regression.

2.5.3 Stationary Series and Toeplitz SSA

If the length N of the series $\mathbb{X}$ is not sufficiently large and the series is assumed stationary, then the usual recommendation is to replace the matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ by some other matrix, which is constructed under the stationarity assumption.

Note first that we can consider the *lag-covariance matrix* $\mathbf{C} = \mathbf{S}/K$ instead of $\mathbf{S}$ for obtaining the SVD of the trajectory matrix $\mathbf{X}$. Indeed, the eigenvectors of the matrices $\mathbf{S}$ and $\mathbf{C}$ are the same.

Denote by $c_{ij} = c_{ij}(N)$ the elements of the lag-covariance matrix $\mathbf{C}$. If the time series is stationary, and $K \to \infty$, then $\lim c_{ij} = R_{\mathbb{X}}(|i - j|)$ as $N \to \infty$, where $R_{\mathbb{X}}(k)$ stands for the lag k term of the time series covariance function. We can therefore define a Toeplitz version of the lag-covariance matrix by putting equal values $\widetilde{c}_{ij}$ at each matrix antidiagonal $|i - j| = k$. The most natural way for defining the values $\widetilde{c}_{ij}$ and the corresponding matrix $\widetilde{\mathbf{C}}$ is to compute.

$$\widetilde{c}_{ij} = \frac{1}{N - |i - j|} \sum_{m=1}^{N-|i-j|} x_m x_{m+|i-j|}, \quad 1 \le i, j \le L. \tag{2.24}$$

If the original series is stationary, the use of *Toeplitz lag-covariance matrix* $\widetilde{\mathbf{C}}$ can be more appropriate than the use of the lag-covariance matrix $\mathbf{C}$. However, Toeplitz SSA is not appropriate for nonstationary series and if the original series has an influential nonstationary component, then Basic SSA seems to work better than Toeplitz SSA. For example, if we are dealing with a pure exponential series, then it is described by a single eigentriple for any window length, while Toeplitz SSA produces L eigentriples for the window length L with harmonic-like eigenvectors. The same effect takes place for the linear series, exponential-cosine series, etc.

A number of papers devoted to SSA analysis of climatic time series (e.g. [11]) consider Toeplitz SSA as the main version of SSA and state that the difference between the Basic and Toeplitz versions of SSA is marginal. This is, however, not true if the series we analyze is non-stationary. It seems that using the Toeplitz version of SSA algorithm is unsafe if the series contains a trend or oscillations with increasing or decreasing amplitudes. Examples of effects observed when Toeplitz SSA is applied to non-stationary time series are presented in [13]. For the study of theoretical properties of Toeplitz SSA, see, for example, [16].

2.5.4 Rotations for Separability: SSA–ICA

The SVD is the key step in SSA; it provides the best matrix approximations to the trajectory matrix $\mathbf{X}$. The SVD often delivers the proper decomposition from the viewpoint of weak separability. However, if several components of the original series are mixed in such a way that their contributions are very similar, then the optimality of the SVD does not help to separate these components and we find ourselves in the situation where we have weak separability of components but lack their strong separability. In this situation, we need to find special rotations which would allow us to separate the components. We will choose these rotations so that they satisfy some additional optimality criterion, which we are going to introduce.

Let us use the idea from the projection pursuit method of multivariate analysis (see [18] for a review). For choosing directions, the projection pursuit uses a criterion based on the form of the distribution of the projection on a given direction. Assuming $L \le K$, we apply the projection pursuit to the trajectory matrix with its rows considered as variables.

Let us start by considering projections on different directions for two vectors taken from subspaces corresponding to different time series components. For simplicity of depiction we rotate the data and consider projections on the x-axis. Figure 2.29c shows projections for different rotations of two sine wave variables. The first picture in a row (the case $\alpha = 0$) corresponds to the proper rotation, the last one (with $\alpha = \pi/4$) shows the worst possible mixture. We can see that the estimated densities

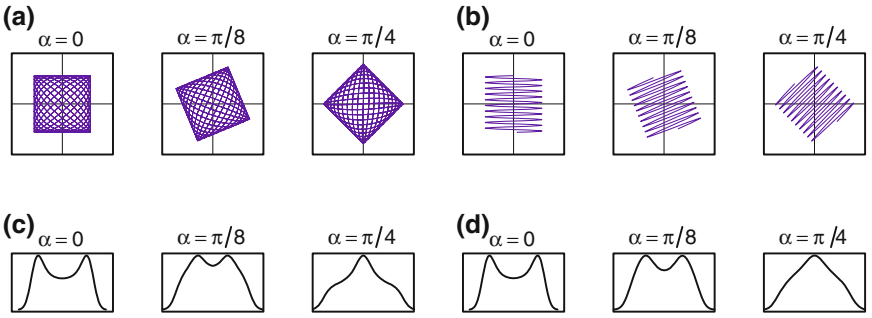

Fig. 2.29 Projection: two sine waves (*left*) and sine wave with linear series (*right*). **a** Scatterplots. **b** Scatterplots. **c** Densities. **d** Densities

are totally different. To check that this result is generic, let us consider similar pictures for a sine wave and a linear function (Fig. 2.29d). The result is very similar. We thus conclude that the idea of projection pursuit may help in solving the problem of separation.

Let us consider the projection pursuit method for cluster analysis where the proper rotation ($\alpha = 0$) corresponds to the maximal distance from the normal distribution. Figure 2.30c shows that the distributions of proper projections and improper projections are similar to the ones depicted in Fig. 2.29c, d.

It is known that there is a method of multivariate analysis, which can be reduced to the projection pursuit method (Fig. 2.30d confirms it). This method is called Independent Component Analysis (ICA); see, for example [19]. The aim of the ICA is finding statistically independent components $\{\eta_i;\ i = 1, \ldots, p\}$ from observations of their linear combinations $\{\xi_i;\ i = 1, \ldots, p\}$. Let us describe the main idea of the ICA. Without loss of generality, we can assume that $\{\xi_i\}$ are pre-whitened.

The mutual information of the random vector $(\eta_1 \ldots, \eta_p)$ can be measured as $I(\eta_1, \ldots, \eta_p) = \sum_{k=1}^{p} H(\eta_k) - H(\eta_1, \ldots, \eta_p)$, where $H(\eta) = \int f(x) \log_2(f(x))dx$ is the differential entropy and $f(x)$ is the density function of η. Therefore, searching for independent components is equivalent to searching for random variables $\{\eta_i\}$, which are linear combinations of $\{\xi_i\}$ and have the minimal value of the mutual information.

It appears that the minimization of the mutual information is equivalent to the maximization of the total negentropy of $\{\eta_i\}$, which is the sum of marginal negentropies $\mathrm{J}(\eta_i) = \mathrm{H}(\nu) - \mathrm{H}(\eta_i)$, $\nu \sim \mathrm{N}(0, 1)$. This means that the ICA works similar to the search for the direction with the distribution of projections that are maximally distant from the normal distribution; that is, to the projection pursuit.

Rather than maximizing negentropies, which requires the estimation of the marginal densities for calculating entropies of η_i's, we can consider maximization of simple functionals like

$$\mathrm{J}(\eta_i) \sim \left[-\mathsf{E}G(\eta_i) + C_\nu\right]^2, \tag{2.25}$$

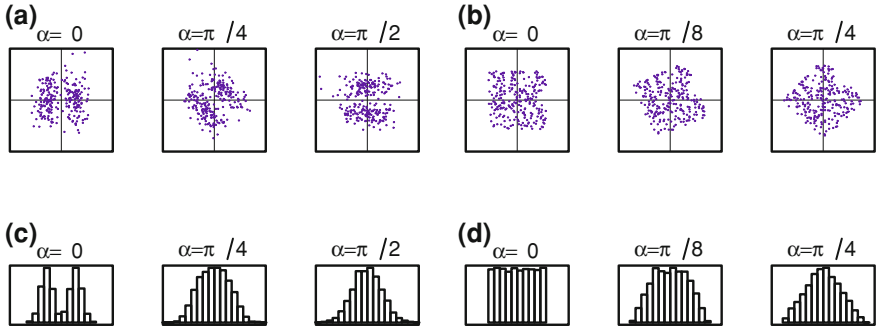

Fig. 2.30 Projection pursuit: two clusters (*left*) and two independent uniformly distributed variables (*right*). **a** Scatterplots. **b** Scatterplots. **c** Histograms. **d** Histograms

where $C_v = \mathsf{E}G(v)$, $G(u) = e^{-u^2/2}$; other choices of G can be considered as well, see [19]. An implementation of the ICA by means of optimizing the functional 2.25 can be found in the R-package *fastICA*, see [25].

Since we observe realizations of p random variables $\mathbf{Y} = [Y_1 : \ldots : Y_p]$, $Y_i \in \mathsf{R}^K$, rather than maximizing (2.25) we should calculate and maximize the following functional of their linear combinations with coefficients $W \in \mathsf{R}^p$:

$$\mathrm{J}(Z) = \Big(-\frac{1}{K}\sum_{i=1}^{K} e^{z_i^2/2} - C_v \Big)^2 \longrightarrow \max_{Z=\mathbf{Y}W,\ ||Z||=1.} \tag{2.26}$$

In applications to blind signal separation, the cooperation between SSA and ICA has been already considered, see [30]. In this application, Basic SSA is used for removal of noise and then the ICA is applied for the extraction of independent components from the mixture.

The theory of ICA is developed for random variables and is not applicable in the deterministic case. Therefore, the application of the ICA to deterministic sources can be formally considered as the projection pursuit which searches for the linear combination of the observed variables (factor vectors in SSA) that maximizes some functional like (2.26). Since the concept of statistical independence is not defined for deterministic vectors we will use the names 'ICA' and 'independent vectors' purely formally and may use quotes while referring to them. It has been established by computer simulations and confirmed by theoretical results that in the examples considered in Fig. 2.30 and some similar ones, the 'ICA' does indeed succeed in separating the time series components, even if the SVD does not provide strong separability.

The 'ICA' has the following important drawback: it does not make ordering of the found components (vectors) like the SVD does. In particular, two vectors corresponding to a sine wave can have arbitrary numbers in the decomposition by the ICA and therefore searching for them is a more difficult task than while applying the SVD. Also, the accuracy of weak separability which the ICA provides is worse than

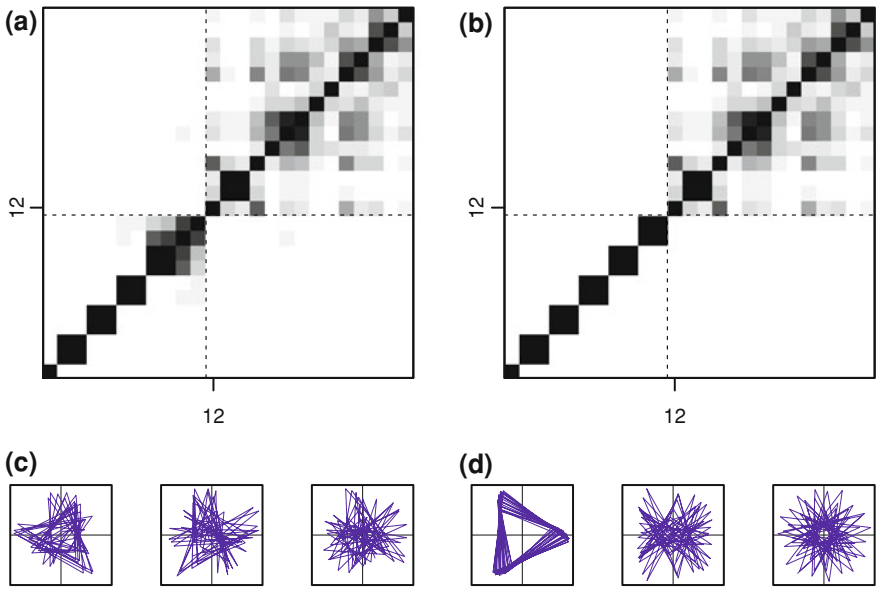

Fig. 2.31 'Fortified wines': SVD (*left*) and ICA for separability ET8-11 (*right*). **a** **w**-correlations. **b** **w**-correlations. **c** Scatterplots ET8–11. **d** Scatterplots ET8–11

that for the SVD. Moreover, the stability of numerical the ICA procedures is worse than for the SVD. Therefore, in SSA, the ICA is worthwhile to consider only as a supplement to the SVD for finding proper rotations in the presence of weak separability but lack of strong separability. By no means the ICA can be recommended as a full replacement of the SVD.

Below we suggest a scheme for building a refined grouping by the SSA–ICA procedure. This scheme could be used as a substitution of the grouping step in Basic SSA. Let us have the expansion $\mathbf{X} = \sum_{j=1}^{d} \sqrt{\lambda_j} U_j V_j^{\mathrm{T}}$ at the SVD step.

Refined grouping by SSA–ICA

1. Make a grouping $\mathbf{X} = \mathbf{X}_{I_1} + \ldots + \mathbf{X}_{I_m}$ as in Basic SSA; this corresponds to weakly separated time series components.
2. Choose a group I consisting of p indices, which is possibly composed of several interpretable components that are mixed.
3. Extract p 'independent' vectors Q_i applying the 'ICA' to $\mathbf{X}_I$. Then $\mathbf{X}_I = \sum_{i=1}^{p} P_i Q_i^{\mathrm{T}}$, where $P_i = \mathbf{X}_I Q_i$.
4. Make k subgroups from the group I by splitting $\mathbf{X}_I = \mathbf{X}_{I,1} + \ldots + \mathbf{X}_{I,k}$.

Example 2.2 Let us provide an example of application of the algorithm of SSA–ICA. Consider the example 'Fortified wines' depicted in Fig. 2.16. For the analysis, we take the first 120 points. The window length L does not provide strong separability for ET8–11 (sine waves with periods 2.4 and 3), see Fig. 2.31a depicting

the **w**-correlation matrix, where the block of four correlated components is clearly seen. 2D-scatterplots of factor vectors are depicted in Fig. 2.31c and demonstrate the absence of structure. Let us apply 'ICA' to the trajectory matrix reconstructed by the eigentriples 8–11. Figures 2.31b, d show that the 'ICA' makes a successful separation of the two sine waves. Let us remark that the resultant components of the 'ICA' needed an additional ordering so that the two sine waves with the same frequency obtain consecutive indices.

2.5.5 Sequential SSA

The hurdle of mixed time series components (formally, the problem of close singular values for weakly separable series components) may sometimes be overcome by the use of what was called in [14] *Sequential SSA* (alternative names for this procedure would be 'Multi-stage SSA' or 'Reiterated SSA').

The Sequential SSA with two stages can be described as follows. First, we extract several time series components by Basic SSA (or any other version of SSA) with certain window length L_1. Then we apply Basic SSA with window length L_2 to the residuals. Having extracted two sets of time series components, we can group them in different ways. For instance, if a rough trend has been extracted at the first stage and other trend components at the second stage, then we have to add them together to obtain the accurate trend. Let us illustrate this on the following example.

Example 2.3 'Germany Unemployment' series: extraction of harmonics
The 'Germany unemployment' series (West Germany, monthly, from April 1950 to December 1980, [31]) serves as an example of complex trends and amplitude-modulated periodicities. The series is depicted in Fig. 2.32.

Selecting large L would mix up the trend and periodic components of the series. For small L the periodic components are not separable from each other. Hence Basic SSA fails to extract (amplitude-modulated) harmonic components of the series.

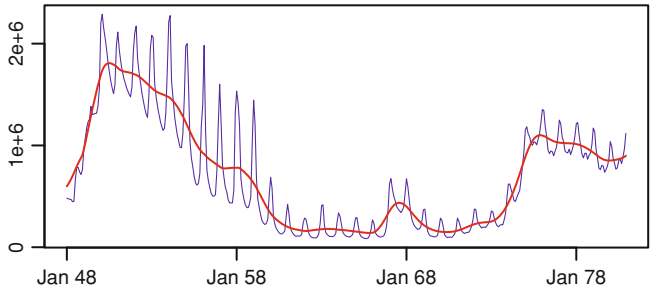

Fig. 2.32 Germany unemployment: the initial series and its trend

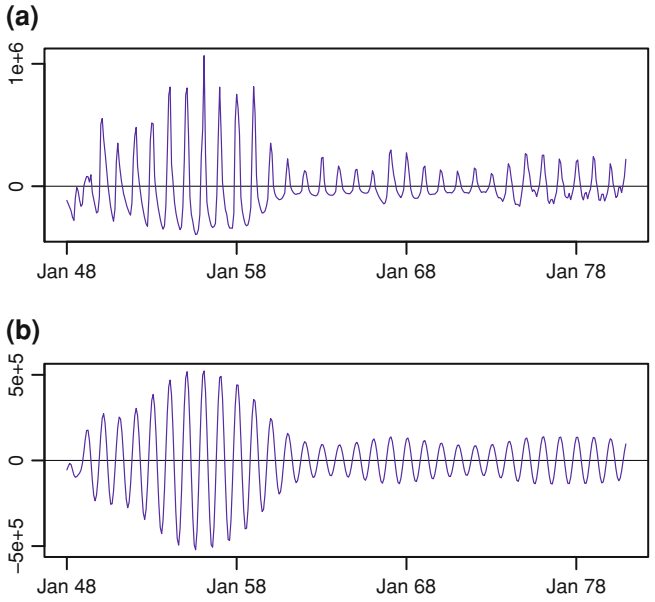

Fig. 2.33 'Germany unemployment': oscillations. **a** Trend residuals. **b** Annual periodicity

The Sequential SSA with two stages is a better method in this case. If we apply Basic SSA with $L = 12$ to the initial series, then the first eigentriple will describe the trend (see Fig. 2.32) which is extracted rather well: the trend component does not include high frequencies, while the residual component practically does not contain low ones (see Fig. 2.33a for the residual series).

The second stage of Sequential SSA is applied to the residual series with $L = 180$. Since the series is amplitude modulated, the main periodogram frequencies (annual $\omega = 1/12$, half-annual $\omega = 1/6$ and 4-months $\omega = 1/4$) are somewhat spread out, and therefore each (amplitude-modulated) harmonic can be described by several (more than 2) eigentriples. The periodogram analysis of the obtained singular vectors shows that the leading 14 eigentriples with share 91.4 % can be related to 3 periodicities: the eigentriples 1, 2, 5–8, 13, 14 describe the annual amplitude-modulated harmonic (Fig. 2.33b), the eigentriples 3, 4, 11–12 are related to half-year periodicity, and the eigentriples 9, 10 describe the 4-months harmonic.

The same technique can be applied to the 'Births' series if we want to obtain better results than those described in Sect. 2.2.2.2. (See Sect. 2.4.3 for a discussion concerning the large window length problem in this example.)

2.5.6 Computer Implementation of SSA

There are many implementations of SSA. They can be classified as follows. First, the implementations differ by the potential areas of application: for example, general purpose SSA, see e.g. [14], and SSA oriented mainly for climatic applications, see e.g. [11]. Second, the software can be free-access and not free-access. One of the main drawbacks of free-access packages is that they generally have no support and that their implementation consists of direct and usually non-efficient use of the main formulas. Third, the software can be interactive (for different systems, Window, Unix or Mac) and non-interactive. Interactive implementations of SSA provide executable programs in some programming language such as special mathematical languages like R and Matlab or high-level programming languages like C++, VBA and others.

We draw special attention to the following four supported software packages:

1. http://gistatgroup.com:
   'Caterpillar'-SSA software (Windows) following the methodology from [14];
2. http://www.atmos.ucla.edu/tcd/ssa/:
   SSA-MTM Toolkit for spectral analysis [11] (Unix) and its commercial extension kSpectra Toolkit (Mac);
3. http://cran.r-project.org/web/packages/Rssa/:
   R-package 'Rssa' [14, 22], a very fast implementation of the main SSA procedures for any platform.
4. The commercial statistical software, SAS, includes Singular Spectrum Analysis to its econometric extension SAS/ETS® Software.

The fastest implementation of SSA can be found in the R-package 'Rssa'. Let us describe the idea of its implementation. Note that the most time-consuming step of SSA is the Singular Value Decomposition (SVD). The SVD in SSA has two specific features. First, SSA as a rule uses only a few leading components. Therefore, we need to use the so-called Partial SVD to compute only a given number of leading eigentriples. Second, the matrix used for decomposition is Hankel. This can be effectively used to speed up the matrix-vector multiplications. The fastest acceleration is reached for the case $L \sim N/2$, which is frequently one of the commonly used window lengths, and long time series. However, even for moderate N the advantage is often very visible.

The acceleration in the 'Rssa' package is achieved by the following means.

- The embedding step is combined with the SVD step; this decreases the storage requirement as we do not need to store the trajectory matrix.
- The 'Rssa' includes the Lanczos-based Partial SVD that generally provides the computational complexity $O(rN^2)$ for calculation of r eigentriples rather than $O(N^3)$ needed for the full SVD.
- The Fast Fourier Transform (FFT) is used for the multiplication of a Hankel matrix by a vector and therefore we have the computational complexity $O(rN \log N)$ of the SVD step.

- Similarly, FFT is used at the Reconstruction stage; this reduces its complexity from $O(rN^2)$ to $O(rN \log N)$.

Let us demonstrate how the Reconstruction stage of Basic SSA can be accelerated. Fix the eigentriple $(\sqrt{\lambda}, U, V)$, where $U \in \mathsf{R}^L$, $V \in \mathsf{R}^K$, $L \le K$, $\lambda \in \mathsf{R}$, and consider the procedure of calculating the related time series component by the diagonal averaging procedure applied to the elementary matrix $\sqrt{\lambda} U V^{\mathrm{T}}$. The output of the algorithm is the elementary time series $\widetilde{\mathbb{Y}} = \left(\tilde{y}_j\right)_{j=1}^{N}$ corresponding to the matrix $\sqrt{\lambda} U V^{\mathrm{T}}$ after hankelization.

Algorithm: Rank 1 Hankelization via Linear Convolution

1. $U' \longleftarrow (u_1, \ldots, u_L, 0, \ldots, 0)^{\mathrm{T}} \in \mathsf{R}^N$
2. $\hat{U} \longleftarrow \mathrm{FFT}_N(U')$
3. $V' \longleftarrow (v_1, \ldots, v_K, 0, \ldots, 0)^{T} \in \mathsf{R}^N$
4. $\hat{V} \longleftarrow \mathrm{FFT}_N(V')$
5. $Y' \longleftarrow \mathrm{IFFT}_N(\hat{V} \odot \hat{U})$
6. $W \longleftarrow (1, 2, \ldots, L, L, \ldots, L, L, L-1, \ldots, 1) \in \mathsf{R}^N$
7. $\mathbb{Y} \longleftarrow \sqrt{\lambda}\left(W \odot Y'\right)$.

Here $(A \odot B)$ denotes element-wise vector multiplication and IFFT is the inverse FFT. The versions of FFT and IFFT which are effective for arbitrary N should be used, see e.g. [10].

2.5.7 Replacing the SVD with Other Procedures

Some variations to the standard SVD procedure have been already mentioned in Sects. 2.5.4 and 2.5.6. These variations include rotations within the eigenspaces, Independent Component Analysis (ICA) and Partial SVD where only few leading eigenvectors of the matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ are computed.

There are three main reasons why it may be worthwhile to replace the SVD operation in Basic SSA with some other operation. These three reasons are: (a) simplicity, (b) improved performance, and (c) different optimality principles for the decomposition.

(a) Simplicity. This is important in the problems where the dimension of the trajectory matrix is very large. In these cases, the SVD may be too costly to perform. The most obvious substitution of the SVD is by Partial SVD, see above. Let us briefly describe (following [28]) another useful substitution of the SVD, which is oriented on solving the problems of the form 'signal plus noise'. Assume that in order to approximate the signal we want to choose the eigentriples with eigenvalues $\lambda \ge a$, for given a. Computation of the signal subspace $\mathcal{X}^{(1)}$ (in the notation introduced at the end of Sect. 2.1.2.3) is equivalent to the computation of the matrix function $f_a(\mathbf{S})$, where $f_a(\lambda)$ is the indicator function $f_a(\lambda) = \mathbf{1}_{[\lambda \ge \mathbf{a}]}$. The function $f_a(\lambda)$ can be approximated by a simple polynomial $P(\lambda)$, for all λ belonging to the spectrum

of $\mathbf{S}$ which is $[\lambda_L, \lambda_1]$. This implies that $f_a(\mathbf{S})$ can be approximated by a matrix polynomial $P(\mathbf{S})$ which yields a simple approximation for the signal subspace $\mathcal{X}^{(1)}$.

Many numerical approximations for the solution of the full or partial SVD problem are also available, see [12]. In cases where the dimension of the matrix $\mathbf{S}$ is exceptionally large, one can use the approximations for the leading eigenvectors used in internet search engines, see e.g. [23].

(b) Improved performance. In some cases (usually when a parametric form of the signal is given), one can slightly correct the SVD (both, eigenvalues and eigenvectors) using the recommendations of SSA perturbation theory, see [29]. As a simple example, in the problems of separating signal from noise, some parts of noise are often found in the SVD components mostly related to the signal, see Fig. 2.22a, b. As a result, it may be worthwhile to make small adjustments to the eigenvalues and eigenvectors to diminish this effect. The simplest version of Basic SSA with constant adjustment in all eigenvalues was suggested in [34]. and is sometimes called the minimum-variance SSA.

(c) Different optimality principles. Here the basis for the decomposition of the series is chosen using some principle which different from the SVD optimality. For example, in ICA discussed in Sect. 2.5.4, the independence of components (rather than the precision of approximation) is considered as the main optimality criteria.

References

1. Alexandrov T (2009) A method of trend extraction using singular spectrum analysis. RevStat 7(1):1–22
2. Allen M, Smith L (1996) Monte Carlo SSA: detecting irregular oscillations in the presence of colored noise. J Clim 9(12):3373–3404
3. Alonso F, Salgado D (2008) Analysis of the structure of vibration signals for tool wear detection. Mech Syst Signal Process 22(3):735–748
4. Alonso F, Castillo J, Pintado P (2005) Application of singular spectrum analysis to the smoothing of raw kinematic signals. J Biomech 38(5):1085–1092
5. Andrews D, Herzberg A (1985) Data. A collection of problems from many fields for the student and research worker. Springer, New York
6. Badeau R, David B, Richard G (2004) Selecting the modeling order for the ESPRIT high resolution method: an alternative approach. In: Proceedings of IEEE ICASSP, vol 2, pp 1025–1028
7. Bilancia M, Campobasso F (2010) Airborne particulate matter and adverse health events: robust estimation of timescale effects. In: Bock HH et al. (eds) Classification as a tool for research, studies in classification, data analysis, and knowledge organization. Springer, Heidelberg, pp 481–489
8. Brillinger D (1975) Time series. Data analysis and theory. Holt, Rinehart and Winston, Inc., New York
9. Clemens J (1994) Whole earth telescope observation of the white dwarf star PG1159-035. In: Weigend A, Gershenfeld N (eds) Time series prediction: forecasting the future and understanding the past. Addison-Wesley, Reading
10. Frigo M, Johnson SG (2005) The design and implementation of FFTW3. Proc IEEE 93(2):216–231

11. Ghil M, Allen RM, Dettinger MD, Ide K, Kondrashov D, Mann ME, Robertson A, Saunders A, Tian Y, Varadi F, Yiou P (2002) Advanced spectral methods for climatic time series. Rev Geophys 40(1):1–41
12. Golub GH, Van Loan CF (1996) Matrix computations, 3rd edn. Johns Hopkins University Press, Baltimore
13. Golyandina N (2010) On the choice of parameters in singular spectrum analysis and related subspace-based methods. Stat Interface 3(3):259–279
14. Golyandina N, Nekrutkin V, Zhigljavsky A (2001) Analysis of time series structure: SSA and related techniques. Chapman & Hall/CRC, New York
15. Golyandina N, Pepelyshev A, Steland A (2012) New approaches to nonparametric density estimation and selection of smoothing parameters. Comput Stat Data Anal 56(7):2206–2218
16. Harris T, Yan H (2010) Filtering and frequency interpretations of singular spectrum analysis. Physica D 239:1958–1967
17. Hipel K, McLeod A (1994) Time series modelling of water resources and environmental systems. Elsevier Science, Amsterdam
18. Huber PJ (1985) Projection pursuit. Ann Stat 13(2):435–475
19. Hyvärinen A, Oja E (2000) Independent component analysis: algorithms and applications. Neural Networks 13(4–5):411–430
20. Janowitz M, Schweizer B (1989) Ordinal and percentile clustering. Math Soc Sci 18:135–186
21. Kendall M, Stuart A (1976) Design and analysis, and time series, the advanced theory of statistics, vol 3, 3rd edn. Charles Griffin, London
22. Korobeynikov A (2010) Computation- and space-efficient implementation of SSA. Stat Interface 3:357–368
23. Langville AN, Meyer CD (2005) A survey of eigenvector methods for web information retrieval. SIAM Rev 47:135–161
24. Lemmerling P, Van Huffel S (2001) Analysis of the structured total least squares problem for Hankel/Toeplitz matrices. Numerical Algorithms 27:89–114
25. Marchini JL, Heaton C, Ripley BD (2010) fastICA: FastICA algorithms to perform ICA and projection pursuit. http://CRAN.R-project.org/package=fastICA, R package version 1.1-13
26. Markovsky I (2011) Low rank approximation: algorithms, implementation, applications. Springer, Heidelberg
27. Markovsky I, Van Huffel S (2007) Overview of total least-squares methods. Signal Process 87:2283–2302
28. Moskvina V, Schmidt KM (2002) Approximate projectors in singular spectrum analysis. SIAM J Matrix Anal Appl 24:932–942
29. Nekrutkin V (2010) Perturbation expansions of signal subspaces for long signals. Stat Interface 3:297–319
30. Pietilä A, El-Segaier M, Vigário R, Pesonen E (2006) Blind source separation of cardiac murmurs from heart recordings. In: Rosca J et al (eds) Independent component analysis and blind signal separation, Lecture Notes in Computer Science, vol 3889. Springer, Heidelberg, pp 470–477
31. Rao TS, Gabr M (1984) An introduction to bispectral analysis and bilinear time series models. Springer, Heidelberg
32. Sauer Y, Yorke J, Casdagli M (1991) Embedology. J Stat Phys 65:579–616
33. Takens F (1981) Detecting strange attractors in turbulence. In: Rand D, Young LS (eds) Dynamical systems and turbulence, Lecture Notes in Mathematics, vol 898. Springer, Berlin, pp 366–381
34. Van Huffel S (1993) Enhanced resolution based on minimum variance estimation and exponential data modeling. Signal Process 33:333–355
35. Vautard R, Yiou P, Ghil M (1992) Singular-Spectrum Analysis: a toolkit for short, noisy chaotic signals. Physica D 58:95–126
36. Wax M, Kailath T (1985) Detection of signals by information theoretic criteria. IEEE Trans Acoust 33:387–392

Chapter 3
SSA for Forecasting, Interpolation, Filtration and Estimation

3.1 SSA Forecasting Algorithms

3.1.1 Main Ideas and Notation

3.1.1.1 Main Ideas

A reasonable forecast of a time series can be performed only if the series has a structure and there are tools to identify and use this structure. Also, we should assume that the structure of the time series is preserved for the future time period over which we are going to forecast (continue) the series. The last assumption cannot be validated using the data to be forecasted. Moreover, the structure of the series can rarely be identified uniquely. Therefore, the situation of different (and even contradictory) forecasts is not impossible. Thus, it is important not only to understand and express the structure but also to assess its stability.

A forecast can be made only if a model is built. The model should be either derived from the data or at least checked against the data. In SSA forecasting, these models can be described through the linear recurrence relations (LRRs). The class of series governed by LRRs is rather wide and important for practical applications. This class contains the series that are linear combinations of products of exponential, polynomial and harmonic series.

Assume that $\mathbb{X}_N = \mathbb{X}_N^{(1)} + \mathbb{X}_N^{(2)}$, where the series $\mathbb{X}_N^{(1)}$ satisfies an LRR of relatively small order and we are interested in the forecasting of $\mathbb{X}_N^{(1)}$. For example, $\mathbb{X}_N^{(1)}$ can be signal, trend or seasonality. The idea of recurrent forecasting is to estimate the underlying LRR and then to perform forecasting by applying the estimated LRR to the last points of the SSA approximation of the series $\mathbb{X}_N^{(1)}$. The main assumption allowing SSA forecasting is that for a certain window length L the series components $\mathbb{X}_N^{(1)}$ and $\mathbb{X}_N^{(2)}$ are approximately strongly separable. In this case, we can reconstruct the series

N. Golyandina and A. Zhigljavsky, *Singular Spectrum Analysis for Time Series*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-34913-3_3,

$\mathbb{X}_N^{(1)}$ with the help of a selected set of the eigentriples and obtain approximations to both the series $\mathbb{X}_N^{(1)}$, its trajectory space and the true LRR.

3.1.1.2 Statement of the Problem, Notation and an Auxiliary Result

Let $\mathbb{X}_N = \mathbb{X}_N^{(1)} + \mathbb{X}_N^{(2)}$ and we intend to forecast $\mathbb{X}_N^{(1)}$. If $\mathbb{X}_N^{(1)}$ is a time series of finite rank $r < L$, then it generates some L-trajectory subspace of dimension r. This subspace reflects the structure of $\mathbb{X}_N^{(1)}$ and hence can be taken as a base for forecasting.

Let us formally describe the forecasting algorithms in a chosen subspace. As we assume that the subspaces are estimated by SSA, we shall refer to the algorithms as the algorithms of SSA forecasting.

Forecasting within the subspace means a continuation of the L-lagged vectors of the forecasted series in such a way that they lie in or very close to the chosen subspace of R^L. We consider three algorithms of forecasting: the recurrent, vector and simultaneous forecasting.

Inputs in the forecasting algorithms:

(a) Time series $\mathbb{X}_N = (x_1, \ldots, x_N)$, $N > 2$.
(b) Window length L, $1 < L < N$.
(c) Linear space $\mathcal{L}_r \subset \mathsf{R}^L$ of dimension $r < L$. We assume that $\mathbf{e}_L \notin \mathcal{L}_r$, where $\mathbf{e}_L = (0, 0, \ldots, 0, 1)^{\mathrm{T}} \in \mathsf{R}^L$; in other terms, $\mathcal{L}_r$ is not a 'vertical' space.
(d) Number M of points to forecast for.

Notation:

(a) $\mathbf{X} = [X_1 : \ldots : X_K]$ (with $K = N - L + 1$) is the trajectory matrix of $\mathbb{X}_N$.
(b) $P_1, \ldots, P_r$ is an orthonormal basis in $\mathcal{L}_r$.
(c) $\widehat{\mathbf{X}} \stackrel{\text{def}}{=} [\widehat{X}_1 : \ldots : \widehat{X}_K] = \sum_{i=1}^{r} P_i P_i^{\mathrm{T}} \mathbf{X}$. The vector $\widehat{X}_i$ is the orthogonal projection of X_i onto the space $\mathcal{L}_r$.
(d) $\widetilde{\mathbf{X}} = \mathscr{H}\widehat{\mathbf{X}} = [\widetilde{X}_1 : \ldots : \widetilde{X}_K]$ is the result of the hankelization of the matrix $\widehat{\mathbf{X}}$. The matrix $\widetilde{\mathbf{X}}$ is the trajectory matrix of some time series $\widetilde{\mathbb{X}}_N = (\widetilde{x}_1, \ldots, \widetilde{x}_N)$.
(e) For any vector $Y \in \mathsf{R}^L$, we denote by $\overline{Y} \in \mathsf{R}^{L-1}$ the vector consisting of the last $L-1$ components of the vector Y, and by $\underline{Y} \in \mathsf{R}^{L-1}$ the vector consisting of the first $L-1$ components of Y.
(f) We set $\nu^2 = \pi_1^2 + \ldots + \pi_r^2$, where π_i is the last component of the vector P_i $(i = 1, \ldots, r)$. Since ν^2 is the squared cosine of the angle between the vector $\mathbf{e}_L$ and the linear space $\mathcal{L}_r$, it can be called the *verticality coefficient* of $\mathcal{L}_r$. Since $\mathbf{e}_L \notin \mathcal{L}_r$, $\nu^2 < 1$.

The following statement is fundamental.

Proposition 3.1 *In the notation above, the last component y_L of any vector $Y = (y_1, \ldots, y_L)^{\mathrm{T}} \in \mathcal{L}_r$ is a linear combination of the first components $y_1, \ldots, y_{L-1}$:*

$$y_L = a_1 y_{L-1} + a_2 y_{L-2} + \ldots + a_{L-1} y_1,$$

where the vector $R = (a_{L-1}, \ldots, a_1)^{\mathrm{T}}$ *can be expressed as*

$$R = \frac{1}{1-\nu^2} \sum_{i=1}^{r} \pi_i \underline{P_i} \tag{3.1}$$

and does not depend on the choice of the basis $P_1, \ldots, P_r$ *in the linear space* $\mathcal{L}_r$.

Proof follows from the fact that the formula (3.1) is a particular case of (3.10) below with $n = L$, $m = r$ and $\mathcal{Q} = \{L\}$.

3.1.2 Formal Description of the Algorithms

3.1.2.1 Recurrent Forecasting

In the above notation, the *recurrent forecasting algorithm* (briefly, *R-forecasting*) can be formulated as follows.

Algorithm (R-forecasting):

1. The time series $\mathbb{Y}_{N+M} = (y_1, \ldots, y_{N+M})$ is defined by

$$y_i = \begin{cases} \widetilde{x}_i & \text{for } i = 1, \ldots, N, \\ \sum\limits_{j=1}^{L-1} a_j y_{i-j} & \text{for } i = N+1, \ldots, N+M. \end{cases} \tag{3.2}$$

2. The numbers $y_{N+1}, \ldots, y_{N+M}$ form the M terms of the recurrent forecast.

Thus, the R-forecasting is performed by the direct use of the LRR with coefficients $\{a_j, j = 1, \ldots, L-1\}$ derived in Proposition 3.1.

Remark 3.1 Let us define the linear operator $\mathcal{P}_{\mathrm{Rec}} : \mathsf{R}^L \mapsto \mathsf{R}^L$ by the formula

$$\mathcal{P}_{\mathrm{Rec}} Y = \begin{pmatrix} \overline{Y} \\ R^{\mathrm{T}} \overline{Y} \end{pmatrix}. \tag{3.3}$$

Set

$$Z_i = \begin{cases} \widetilde{X}_i & \text{for } i = 1, \ldots, K, \\ \mathcal{P}_{\mathrm{Rec}} Z_{i-1} & \text{for } i = K+1, \ldots, K+M. \end{cases} \tag{3.4}$$

It is easily seen that the matrix $\mathbf{Z} = [Z_1 : \ldots : Z_{K+M}]$ is the trajectory matrix of the series $\mathbb{Y}_{N+M}$. Therefore, (3.4) can be regarded as the vector form of (3.2).

3.1.2.2 Vector Forecasting

Let us now describe the *vector forecasting algorithm* (briefly, *V-forecasting*). The idea of vector forecasting is as follows. Let us assume that we can continue the sequence of vectors $\widehat{X}_1, \ldots, \widehat{X}_K$ (which belong to the subspace $\mathcal{L}_r$) for M steps so that:

(a) the continuation vectors Z_m ($K < m \le K + M$) belong to the same subspace $\mathcal{L}_r$;
(b) the matrix $\mathbf{X}_M = [\widehat{X}_1 : \ldots : \widehat{X}_K : Z_{K+1} : \ldots : Z_{K+M}]$ is approximately Hankel.

Then, having obtained the matrix $\mathbf{X}_M$ we can obtain the forecasted series $\mathbb{Y}_{N+M}$ by the diagonal averaging of this matrix.

In addition to the notation introduced above let us bring in some more notation. Consider the matrix

$$\Pi = \underline{\mathbf{V}}\underline{\mathbf{V}}^{\mathrm{T}} + (1 - \nu^2) R R^{\mathrm{T}}, \tag{3.5}$$

where $\underline{\mathbf{V}} = [\underline{P_1} : \ldots : \underline{P_r}]$. The matrix Π is the matrix of the linear operator that performs the orthogonal projection $\mathsf{R}^{L-1} \mapsto \underline{\mathcal{L}_r}$, where $\underline{\mathcal{L}_r} = \mathrm{span}(\underline{P_1}, \ldots, \underline{P_r})$; note that this matrix Π is a particular case of the matrix defined in (3.11) with $m = r$, $n = L$ and $\mathcal{Q} = \{L\}$. Finally, we define the linear operator $\mathcal{P}_{\mathrm{Vec}} : \mathsf{R}^L \mapsto \mathcal{L}_r$ by the formula

$$\mathcal{P}_{\mathrm{Vec}} Y = \begin{pmatrix} \Pi \overline{Y} \\ R^{\mathrm{T}} \overline{Y} \end{pmatrix}. \tag{3.6}$$

Algorithm (V-forecasting):

1. In the notation above, define the vectors Z_i as follows:

$$Z_i = \begin{cases} \widehat{X}_i & \text{for } i = 1, \ldots, K, \\ \mathcal{P}_{\mathrm{Vec}} Z_{i-1} & \text{for } i = K+1, \ldots, K+M+L-1. \end{cases} \tag{3.7}$$

2. By constructing the matrix $\mathbf{Z} = [Z_1 : \ldots : Z_{K+M+L-1}]$ and making its diagonal averaging we obtain the series $y_1, \ldots, y_{N+M+L-1}$.
3. The numbers $y_{N+1}, \ldots, y_{N+M}$ form the M terms of the vector forecast.

Remark 3.2 Note that in order to get M forecast terms, the vector forecasting procedure performs $M+L-1$ steps. The aim is the permanence of the forecast under variations in M: the M-step forecast ought to coincide with the first M values of the forecast for $M+1$ or more steps. In view of the definition of the diagonal averaging, we have to make $L-1$ extra steps.

3.1.2.3 Simultaneous Forecasting

The recurrent forecasting is based on the fact that the last coordinate of any vector in the subspace $\mathcal{L}_r$ is determined by the first $L-1$ coordinates. The idea of the *simultaneous forecasting algorithm* is based on the following relation: under some additional conditions, the last M coordinates of any vector in $\mathcal{L}_r$ can be expressed through its first $L-M$ coordinates. Certainly, $L-M$ should be larger than r and therefore $M < L-r$.

Let $\mathrm{span}(\mathbf{e}_i, i = L-M+1, \dots, L) \cap \mathcal{L}_r = \{\mathbf{0}\}$. For a vector $Y \in \mathcal{L}_r$, denote $Y_1 = (y_1, \dots, y_{L-M})^{\mathrm{T}}$ and $Y_2 = (y_{L-M+1}, \dots, y_L)^{\mathrm{T}}$. Then $Y_2 = \mathbf{R}Y_1$, where the matrix $\mathbf{R}$ is defined in (3.10) with $n = L$, $m = r$ and $\mathcal{Q} = \{L-M+1, \dots, L\}$.

Algorithm (simultaneous forecasting):

1. In the notation above, define the time series $\mathbb{Y}_{N+M} = (y_1, \dots, y_{N+M})$ by

$$\begin{aligned} &y_i = \widetilde{x}_i \quad \text{for} \quad i = 1, \dots, N, \\ &(y_{N+1}, \dots, y_{N+M})^{\mathrm{T}} = \mathbf{R}(y_{N-(L-M)+1}, \dots, y_N)^{\mathrm{T}}. \end{aligned} \tag{3.8}$$

2. The numbers $y_{N+1}, \dots, y_{N+M}$ form the M terms of the simultaneous forecast.

Remark 3.3 The algorithm formulated above is an analogue of the R-forecasting, since $\mathbf{R}$ in (3.8) is applied to the reconstructed series. An analogue of the V-forecasting can also be considered.

3.1.3 SSA Forecasting Algorithms: Similarities and Dissimilarities

If $\mathcal{L}_r$ is spanned by certain eigenvectors obtained from the SVD of the trajectory matrix of the series $\mathbb{X}_N$, then the corresponding forecasting algorithm will be called the *Basic SSA forecasting algorithm*.

Let us return to Basic SSA and assume that our aim is to extract an additive component $\mathbb{X}_N^{(1)}$ from a series $\mathbb{X}_N$. For an appropriate window length L, we obtain the SVD of the trajectory matrix of the series $\mathbb{X}_N$ and select the eigentriples $(\sqrt{\lambda_i}, U_i, V_i)$, $i \in I$, corresponding to $\mathbb{X}_N^{(1)}$. Then we obtain the resultant matrix

$$\mathbf{X}_I = \sum_{i \in I} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$$

and, after the diagonal averaging, we obtain the reconstructed series $\widetilde{\mathbb{X}}_N^{(1)}$ that estimates $\mathbb{X}_N^{(1)}$.

Note that the columns $\widehat{X}_1, \dots, \widehat{X}_K$ of the resultant matrix $\mathbf{X}_I$ belong to the linear space $\mathcal{L}_r = \mathrm{span}(U_i, i \in I)$. If $\mathbb{X}_N^{(1)}$ is strongly separable from $\mathbb{X}_N^{(2)} \stackrel{\mathrm{def}}{=} \mathbb{X}_N - \mathbb{X}_N^{(1)}$,

then $\mathcal{L}_r$ coincides with $\mathcal{X}^{(L,1)}$ (the trajectory space of the series $\mathbb{X}_N^{(1)}$) and $\mathbf{X}_I$ is a Hankel matrix (in this case, $\mathbf{X}_I$ is the trajectory matrix of the series $\mathbb{X}_N^{(1)}$). Then the recurrent, vector and simultaneous forecasts coincide and the resulting procedure could be called the *exact continuation* of $\mathbb{X}_N^{(1)}$. More precisely, in this situation the matrix Π is the identity matrix, and (3.6) coincides with (3.3). Furthermore, the matrix $\mathbf{Z}$ has Hankel structure and the diagonal averaging does not change the matrix elements.

If $\mathbb{X}_N^{(1)}$ and $\mathbb{X}_N^{(2)}$ are approximately strongly separable, then $\mathcal{L}_r$ is close to $\mathcal{X}^{(L,1)}$ and $\mathbf{X}_I$ is approximately a Hankel matrix.

If there is no exact separability, then different modifications of the forecasting algorithms usually give different results. Let us describe the difference between them. Since the recurrent and vector forecasting algorithms are more conventional and have less limitations, we shall concentrate on the recurrent and vector forecasting algorithms only (besides, the simulations show that the simultaneous forecasting is often less accurate).

- In a typical situation, there is no time series such that the linear space $\mathcal{L}_r$ (for $r < L-1$) is its trajectory space [12, Proposition 5.6], and therefore this space cannot be the trajectory space of the series to be forecasted. The R-forecasting method uses $\mathcal{L}_r$ to obtain the LRR of the forecasting series. The V-forecasting procedure tries to perform the L-continuation of the series in $\mathcal{L}_r$: any vector $Z_{i+1} = \mathcal{P}_{\mathrm{Vec}} Z_i$ belongs to $\mathcal{L}_r$, and Z_{i+1} is as close as possible to $\overline{Z_i}$. The last component of Z_{i+1} is obtained from $\overline{Z_{i+1}}$ by the same LRR as used in the R-forecasting.
- Both forecasting methods have two general stages: the diagonal averaging and continuation. For the R-forecasting, the diagonal averaging is used to obtain the reconstructed series, and continuation is performed by applying the LRR. In the V-forecasting, these two stages are used in the reverse order; first, vector continuation in $\mathcal{L}_r$ is performed and then the diagonal averaging gives the forecast.
- If there is no exact separability it is hard to compare the recurrent and vector forecasting methods theoretically. Closeness of the two forecasting results can be used as an argument in favour of the forecasting stability.
- R-forecasting is simpler to interpret in view of the link between LRRs and their characteristic polynomials, see Sect. 3.2. On the other hand, numerical study demonstrates that the V-forecasting is typically more 'conservative' (or less 'radical') when the R-forecasting exhibits either rapid increase or decrease.
- V-forecasting has a larger computational cost than R-forecasting.

Remark 3.4 Forecasting algorithms described in Sect. 3.1 are based on the estimation of the trajectory subspace of the forecasted component. In addition to Basic SSA, there are other methods of estimation of the trajectory subspace. For example, if the subspace is estimated by Toeplitz SSA, we obtain Toeplitz SSA forecasting algorithms. We may wish to use SSA with centering for estimating the subspace; in this case, we arrive at corresponding modifications of SSA forecasting with centering, see [12, Sect. 2.3.3].

3.1.4 Appendix: Vectors in a Subspace

In this section, we formulate two technical results that provide the theoretical ground for both forecasting and filling in methods. For proofs and details, we refer to [11].

Consider the Euclidean space R^n. Define $J_n = \{1, \dots, n\}$ and denote by $\mathcal{Q} = \{i_1, \dots, i_s\} \subset J_n$ an ordered set, $|\mathcal{Q}| = s$. Let $\mathbf{I}_s$ denote the unit $s \times s$ matrix. We define a *restriction of a vector* $X = (x_1, \dots, x_n)^{\mathrm{T}} \in \mathsf{R}^n$ onto a set of indices $\mathcal{Q} = \{i_1, \dots, i_s\}$ as the vector $X\big|_{\mathcal{Q}} = (x_{i_1}, \dots, x_{i_s})^{\mathrm{T}} \in \mathsf{R}^s$. *The restriction of a matrix* onto a set of indices is the matrix consisting of restrictions of its column vectors onto this set.

The restriction of a q-dimensional subspace $\mathcal{L}_q$ onto a set of indices $\mathcal{Q}$ is the space spanned by restrictions of all vectors of $\mathcal{L}_q$ onto this set; the restricted space will be denoted by $\mathcal{L}_q\big|_{\mathcal{Q}}$. It is easy to prove that for any basis $\{H_i\}_{i=1}^q$ of the subspace $\mathcal{L}_q$, the equality $\mathcal{L}_q\big|_{\mathcal{Q}} = \operatorname{span}\big(H_1\big|_{\mathcal{Q}}, \dots, H_q\big|_{\mathcal{Q}}\big)$ holds.

3.1.4.1 Filling in Vector Coordinates in the Subspace

Consider an m-dimensional subspace $\mathcal{L}_m \subset \mathsf{R}^n$ with $m < n$. Denote by $\{P_k\}_{k=1}^m$ an orthonormal basis in $\mathcal{L}_m$ and define the matrix $\mathbf{P} = [P_1 : \dots : P_m]$. Fix an ordered set of indices $\mathcal{Q} = \{i_1, \dots, i_s\}$ with $s = |\mathcal{Q}| \le n - m$.

First, note that the following conditions are equivalent (it follows from [11, Lemma 2.1]): (1) for any $Y \in \mathcal{L}_m\big|_{J_n \setminus \mathcal{Q}}$ there exists an unique vector $X \in \mathcal{L}_m$ such that $X\big|_{J_n \setminus \mathcal{Q}} = Y$, (2) the matrix $\mathbf{I}_s - \mathbf{P}\big|_{\mathcal{Q}}\big(\mathbf{P}\big|_{\mathcal{Q}}\big)^{\mathrm{T}}$ be non-singular, and (3) $\operatorname{span}(\mathbf{e}_i, i \in \mathcal{Q}) \cap \mathcal{L}_m = \{\mathbf{0}_n\}$. Either of these conditions can be considered as a condition of unique filling in of the missing vector components with indices from $\mathcal{Q}$.

Proposition 3.2 *Let the matrix* $\mathbf{I}_s - \mathbf{P}\big|_{\mathcal{Q}}\big(\mathbf{P}\big|_{\mathcal{Q}}\big)^{\mathrm{T}}$ *be non-singular. Then for any vector* $X \in \mathcal{L}_m$ *we have*

$$X\big|_{\mathcal{Q}} = \mathbf{R}\, X\big|_{J_n \setminus \mathcal{Q}}, \tag{3.9}$$

where

$$\mathbf{R} = \big(\mathbf{I}_s - \mathbf{P}\big|_{\mathcal{Q}}\big(\mathbf{P}\big|_{\mathcal{Q}}\big)^{\mathrm{T}}\big)^{-1}\, \mathbf{P}\big|_{\mathcal{Q}}\big(\mathbf{P}\big|_{J_n \setminus \mathcal{Q}}\big)^{\mathrm{T}}. \tag{3.10}$$

3.1.4.2 Projection Operator

Let $Y \in \mathsf{R}^n$ and $Z = Y\big|_{J_n \setminus \mathcal{Q}} \in \mathsf{R}^{n-s}$. Generally, $Z \notin \mathcal{L}_m\big|_{J_n \setminus \mathcal{Q}}$. However, for applying formula (3.9) to obtain the vector from $\mathcal{L}_m$, it is necessary that $Z \in \mathcal{L}_m\big|_{J_n \setminus \mathcal{Q}}$. The orthogonal projector $\mathsf{R}^{n-s} \to \mathcal{L}_m\big|_{J_n \setminus \mathcal{Q}}$ transfers Z to $\mathcal{L}_m\big|_{J_n \setminus \mathcal{Q}}$.

Let us derive the form of the matrix of the projection operator $\Pi_{J_n \setminus \mathcal{Q}}$. Set $\mathbf{V} = \mathbf{P}\big|_{J_n \setminus \mathcal{Q}}$ and $\mathbf{W} = \mathbf{P}\big|_{\mathcal{Q}}$ for the convenience of notation.

Proposition 3.3 *Assume that the matrix* $\mathbf{I}_s - \mathbf{W}\mathbf{W}^{\mathrm{T}}$ *is nonsingular. Then the matrix of the orthogonal projection operator* $\Pi_{J_n \setminus \mathcal{Q}}$ *has the form*

$$\Pi_{J_n \setminus \mathcal{Q}} = \mathbf{V}\mathbf{V}^{\mathrm{T}} + \mathbf{V}\mathbf{W}^{\mathrm{T}}(\mathbf{I}_s - \mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{W}\mathbf{V}^{\mathrm{T}}. \tag{3.11}$$

3.2 LRR and Associated Characteristic Polynomials

3.2.1 Basic Facts

The theory of the linear recurrence relations and associated characteristic polynomials is well known (for example, see [9, Chap. V, Sect. 4]). Here we provide a short survey of the results which are most essential for understanding SSA forecasting.

Definition 3.1 A time series $\mathbb{S}_N = \{s_i\}_{i=1}^N$ is *governed by a linear recurrence relation* (LRR), if there exist $a_1, \ldots, a_t$ such that

$$s_{i+t} = \sum_{k=1}^{t} a_k s_{i+t-k}, \ 1 \le i \le N-t, \ a_t \ne 0, \ t < N. \tag{3.12}$$

The number t is called the order of the LRR, $a_1, \ldots, a_t$ are the coefficients of the LRR. If $t = r$ is the minimal order of an LRR that governs the time series S_N, then the corresponding LRR is called *minimal* and we say that the time series $\mathbb{S}_N$ has *finite-difference dimension* r.

Note that if the minimal LRR governing the signal $\mathbb{S}_N$ has order r with $r < N/2$, then $\mathbb{S}_N$ has rank r (see Sect. 2.3.1.2 for the definition of the series of finite rank).

Definition 3.2 A polynomial $P_t(\mu) = \mu^t - \sum_{k=1}^t a_k \mu^{t-k}$ is called a *characteristic polynomial* of the LRR (3.12).

Let the time series $\mathbb{S}_\infty = (s_1, \ldots, s_n, \ldots)$ satisfy the LRR (3.12) with $a_t \ne 0$ and $i \ge 1$. Consider the characteristic polynomial of the LRR (3.12) and denote its different (complex) roots by $\mu_1, \ldots, \mu_p$ with $1 \le p \le t$. All these roots are nonzero as $a_t \ne 0$. Let the multiplicity of the root μ_m be k_m, where $1 \le m \le p$ and $k_1 + \ldots + k_p = t$. The following well-known result (see e.g. [12, Theorem 5.3] or [13]) provides an explicit form for the series which satisfies the LRR.

Theorem 3.1 *The time series* $\mathbb{S}_\infty = (s_1, \ldots, s_n, \ldots)$ *satisfies the LRR* (3.12) *for all* $i \ge 0$ *if and only if*

$$s_n = \sum_{m=1}^{p} \left(\sum_{j=0}^{k_m - 1} c_{mj} n^j \right) \mu_m^n, \tag{3.13}$$

where the complex coefficients c_{mj} depend on the first t points $s_1, \ldots, s_t$.

For real-valued time series, Theorem 3.1 implies that the class of time series governed by the LRRs consists of sums of products of polynomials, exponentials and sinusoids.

3.2.2 Roots of the Characteristic Polynomials

Let the series $\mathbb{S}_N = (s_1, \ldots, s_N)$ be governed by an LRR (3.12) of order t. Let $\mu_1, \ldots, \mu_p$ be the different (complex) roots of the characteristic polynomial $P_t(\mu)$. As $a_t \neq 0$, these roots are not equal to zero. We also have $k_1 + \ldots + k_p = t$, where k_m are the multiplicities of the roots μ_m $(m = 1, \ldots, p)$.

Denote $s_n(m, j) = n^j \mu_m^n$ for $1 \leq m \leq p$ and $0 \leq j \leq k_m - 1$. Theorem 3.1 tells us that the general solution of the Eq. (3.12) is

$$s_n = \sum_{m=1}^{p} \sum_{j=0}^{k_m - 1} c_{mj} s_n(m, j) \tag{3.14}$$

with certain complex c_{mj}. The coefficients c_{mj} are defined by $s_1, \ldots, s_t$, the first t elements of the series $\mathbb{S}_N$.

Thus, each root μ_m produces a component

$$s_n^{(m)} = \sum_{j=0}^{k_m - 1} c_{mj} s_n(m, j) \tag{3.15}$$

of the series $\mathbb{S}_N$. Let us fix m and consider the mth component in the case $k_m = 1$, which is the main case in practice. Set $\mu_m = \rho e^{\mathrm{i}2\pi\omega}$, $\omega \in (-1/2, 1/2]$, where $\rho > 0$ is the modulus (absolute value) of the root and $2\pi\omega$ is its polar angle.

If ω is either 0 or $1/2$, then μ_m is a real root of the polynomial $P_t(\mu)$ and the series component $s_n^{(m)}$ is real and is equal to $c_{m0}\mu_m^n$. This means that $s_n^{(m)} = A\rho^n$ for positive μ_m and $s_n^{(m)} = A\,(-1)^n \rho^n = A\rho^n \cos(\pi n)$ for negative μ_m. This last case corresponds to the exponentially modulated saw-tooth sequence.

All other values of ω lead to complex μ_m. In this case, P_t has a complex conjugate root $\mu_l = \rho e^{-\mathrm{i}2\pi\omega}$ of the same multiplicity $k_l = 1$. We thus can assume that $0 < \omega < 1/2$ and describe a pair of conjugate roots by the pair of real numbers (ρ, ω) with $\rho > 0$ and $\omega \in (0, 1/2)$.

Adding up the components $s_n^{(m)}$ and $s_n^{(l)}$ corresponding to these conjugate roots we obtain the real series $A\rho^n \cos(2\pi\omega n + \varphi)$, with A and φ expressed in terms of c_{m0} and c_{l0}. The frequency ω can be expressed in the form of the period $T = 1/\omega$ and vice versa.

The asymptotic behaviour of $s_n^{(m)}$ mainly depends on $\rho = |\mu_m|$. Let us consider the simplest case $k_m = 1$ as above. If $\rho < 1$, then $s_n^{(m)}$ rapidly tends to zero and asymptotically has no influence on the whole series (3.14). Alternatively, the root with $\rho > 1$ and $|c_{m0}| \neq 0$ leads to a rapid increase of $|s_n|$ (at least, of a certain subsequence of $\{|s_n|\}$).

Let the series $\mathbb{S}_N$ have finite-difference dimension r. Then the characteristic polynomial of its minimal LRR of order r has r roots. The same series satisfies many other LRRs of dimensions $t > r$. Consider such an LRR (3.12). The characteristic polynomial $P_t(\mu)$ of the LRR (3.12) has t roots with r roots (we call them the *main roots*) coinciding with the roots of the minimal LRR. The other $t-r$ roots are *extraneous*: in view of the uniqueness of the representation (3.15), the coefficients c_{mj} corresponding to these roots are equal to zero. However, the LRR (3.12) governs a wider class of series than the minimal LRR.

Since the roots of the characteristic polynomial specify its coefficients uniquely, they also determine the corresponding LRR. Consequently, by removing the extraneous roots of the characteristic polynomial $P_t(\mu)$, corresponding to the LRR (3.12), we can obtain the polynomial describing the minimal LRR of the series.

Example 3.1 Annual periodicity

Let the series $\mathbb{S}_N$ have zero mean and period 12. Then it can be expressed as a sum of six harmonics:

$$s_n = \sum_{k=1}^{5} c_k \cos(2\pi nk/12 + \varphi_k) + c_6 \cos(\pi n). \tag{3.16}$$

Under the condition $c_k \neq 0$ for $k = 1, \ldots, 6$, the series has finite-difference dimension 11. In other words, the characteristic polynomial of the minimal LRR governing the series (3.16) has 11 roots. All these roots have modulus 1. The real root -1 corresponds to the last term in (3.16). The harmonic term with frequency $\omega_k = k/12$, $k = 1, \ldots, 5$, generates two complex conjugate roots $\exp(\pm \mathrm{i} 2\pi k/12)$, which have polar angles $\pm 2\pi k/12$.

3.2.3 Min-Norm LRR

Consider a time series $\mathbb{S}_N$ of rank r governed by an LRR. Let L be the window length ($r < \min(L, K)$, $K = N - L + 1$), $\mathbf{S}$ be the trajectory matrix of $\mathbb{S}_N$, $\mathbb{S}$ be its trajectory space, $P_1, \ldots, P_r$ form an orthonormal basis of $\mathbb{S}$ and $\mathbb{S}^{\perp}$ be the orthogonal complement to $\mathbb{S}$. Denote $A = (a_{L-1}, \ldots, a_1, -1)^{\mathrm{T}} \in \mathbb{S}^{\perp}$, $a_{L-1} \neq 0$. Then the time series $\mathbb{S}$ satisfies the LRR

$$s_{i+(L-1)} = \sum_{k=1}^{L-1} a_k s_{i+(L-1)-k}, \ 1 \le i \le K. \tag{3.17}$$

Conversely, if a time series is governed by an LRR (3.17), then the LRR coefficients $B = (a_{L-1}, \ldots, a_1)^{\mathrm{T}}$ complemented with -1 yield the vector $\begin{pmatrix} B \\ -1 \end{pmatrix} \in \mathcal{S}^{\perp}$. Note that any LRR that governs the time series can be treated as a forward linear prediction. In addition, if we consider a vector in $\mathcal{S}^{\perp}$ with -1 as the first coordinate, then we obtain the so-called backward linear prediction [26].

For any matrix $\mathbf{A}$, we denote by $\underline{\mathbf{A}}$ the matrix $\mathbf{A}$ with the last row removed and by $\overline{\mathbf{A}}$ the matrix $\mathbf{A}$ without the first row.

From the viewpoint of prediction, the LRR governing a time series of rank r has coefficients derived from the condition $\underline{\mathbf{S}}^{\mathrm{T}} B = (s_L, \ldots, s_N)^{\mathrm{T}}$. This system of linear equations may have several solutions, since the vector $(s_L, \ldots, s_N)^{\mathrm{T}}$ belongs to the column space of the matrix $\underline{\mathbf{S}}^{\mathrm{T}}$. It is well-known that the least-squares solution expressed by the pseudo-inverse to $\underline{\mathbf{S}}^{\mathrm{T}}$ yields the vector B with minimum norm (the solution for the method of total least squares coincides with it). It can be shown that this minimum-norm solution B_{LS} can be expressed as

$$B_{\mathrm{LS}} = (a_{L-1}, \ldots, a_1)^{\mathrm{T}} = \frac{1}{1-\nu^2} \sum_{i=1}^{r} \pi_i \underline{P_i}, \tag{3.18}$$

where π_i are the last coordinates of P_i and $\nu^2 = \sum_{i=1}^{r} \pi_i^2$.

Thus, one of the vectors from $\mathcal{S}^{\perp}$, which equals $A_{\mathrm{LS}} = \begin{pmatrix} B_{\mathrm{LS}} \\ -1 \end{pmatrix}$, has a special significance and the corresponding LRR is called the *min-norm LRR*, which provides the min-norm (forward) prediction. Similarly, we can derive a relation for the min-norm backward prediction.

It is shown in [12, Proposition 5.5] and [16] that the forward min-norm prediction vector A_{LS} is the normalized (so that its last coordinate is equal to -1) projection of the Lth coordinate vector $\mathbf{e}_L$ on the orthogonal complement to the signal subspace. Therefore, the min-norm prediction vector depends on the signal subspace only.

The following property demonstrates the importance of the minimum norm of the LRR coefficients for noise reduction.

Proposition 3.4 *Let $\mathbb{X}_N = \mathbb{S}_N + \mathbb{P}_N$, where $\mathbb{P}_N$ is stationary white noise with zero mean and variance σ^2, X, S be L-lagged vectors of $\mathbb{X}_N$ and $\mathbb{S}_N$ correspondingly and $C \in \mathsf{R}^{L-1}$. Then for $x = C^{\mathrm{T}} S$ and $\widetilde{x} = C^{\mathrm{T}} X$, we have $\mathsf{E}\widetilde{x} = x$ and $\mathsf{D}\widetilde{x} = \|C\|^2 \sigma^2$.*

If $X = X_K$ is the last lagged vector of $\mathbb{S}_N$, then $\widetilde{x} = C^{\mathrm{T}} \overline{X}_K$ can be considered as a forecasting formula applied to a noisy signal and $\|C\|^2$ regulates the variance of this forecast.

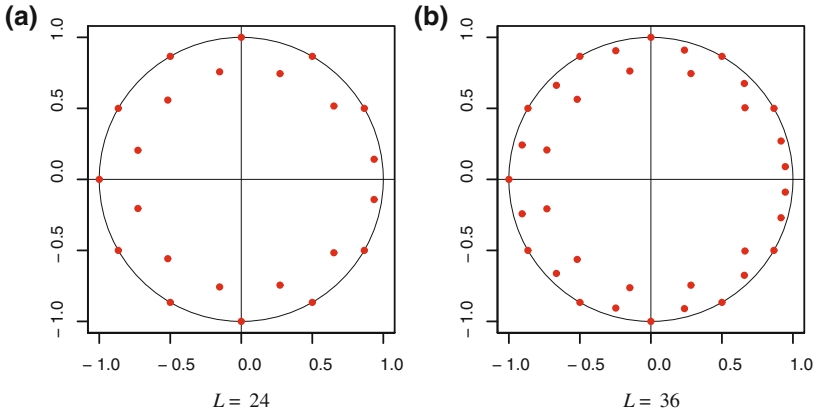

Fig. 3.1 Annual periodicity: main and extraneous roots

The following property of the min-norm LRR, which was derived in [17], is extremely important for forecasting: all extraneous roots of the min-norm LRR lie inside the unit circle of the complex plane. Example 3.2, where the min-norm LRR is used, illustrates it. This property gives us hope that in the case of real-life time series (when both the min-norm LRR and the related initial data are perturbed) the terms related to the extraneous roots in (3.13) only slightly influence the forecast. Moreover, bearing in mind the results concerning the distribution of the extraneous roots (see [21, 27]), we can expect that the extraneous summands compensate (cancel out) each other.

Example 3.2 Annual periodicity and extraneous roots

Let us consider the series (3.16) from Example 3.1 and the min-norm LRR, which is not minimal.

Let N be large enough. If we select certain $L \geq 12$ and take $r = 11$, $\mathcal{L}_r = \mathcal{S}(\mathbb{S}_N)$, then the vector $R = (a_{L-1}, \ldots, a_1)^{\mathrm{T}}$ defined in (3.18) produces the LRR (3.17), which is not minimal but governs the series (3.16).

Let us take $c_i = i - 1$, $\varphi_1 = \ldots = \varphi_5 = 0$ and $L = 24,\ 36$. The roots of the characteristic polynomials of the LRR (3.17) are depicted in Fig. 3.1. We can see that the main 11 roots of the polynomial compose 11 of 12 vertices of a regular dodecagon and lie on the unit circle in the complex plane. Twelve ($L = 24$) and twenty four ($L = 36$) extraneous roots have smaller moduli.

Remark 3.5 Note that the min-norm LRR forms the basis for SSA forecasting methods introduced in Sect. 3.1 (see [12, Sect. 2.1]). In particular, the R-forecasting uses the estimated min-norm LRR for forecasting: compare the formulas for coefficients (3.1) with (3.18).

3.3 Recurrent Forecasting as Approximate Continuation

Exact continuation does not have practical meaning. Indeed, it seems unwise to assume that a real-life time series is governed by some LRR of relatively small dimension. Therefore we need to consider approximate continuation, which is of much greater importance in practice than exact continuation. In this section we consider approximate continuation with the help of recurrent forecasting. However, most discussions are also relevant for other SSA forecasting algorithms.

3.3.1 Approximate Separability and Forecasting Errors

Let $\mathbb{X}_N = \mathbb{X}_N^{(1)} + \mathbb{X}_N^{(2)}$ and suppose that the series $\mathbb{X}_N^{(1)}$ admits a recurrent continuation. Denote by d the dimension of the minimal recurrence relation governing $\mathbb{X}_N^{(1)}$. If $d < \min(L, N - L + 1)$, then $d = \operatorname{rank}_L(\mathbb{X}_N^{(1)})$.

If $\mathbb{X}_N^{(1)}$ and $\mathbb{X}_N^{(2)}$ are strongly separable for some window length L, then the trajectory space of $\mathbb{X}_N^{(1)}$ can be found and we can perform recurrent continuation of the series $\mathbb{X}_N^{(1)}$ by the method described in Sect. 3.1.2.1. We now assume that $\mathbb{X}_N^{(1)}$ and $\mathbb{X}_N^{(2)}$ are approximately strongly separable and discuss the problem of approximate continuation (forecasting) of the series $\mathbb{X}_N^{(1)}$ in the subspace $\mathcal{L}_r$. The choice of $\mathcal{L}_r$ is described in Sect. 3.1.3. If the choice is proper, $r = d$.

The series of forecasts y_n ($n > N$) defined by (3.2) generally does not coincide with the recurrent continuation of the series $\mathbb{X}_N^{(1)}$. The deviation between these two series makes the forecasting error. This error has two origins. The main one is the difference between the linear space $\mathcal{L}_r$ and $\mathcal{X}^{(L,1)}$, the trajectory space of the series $\mathbb{X}_N^{(1)}$ (some inequalities connecting the perturbation of the LRR (3.2) with that of $\mathcal{X}^{(L,1)}$ are derived in [19]). Since the LRR (3.2) is produced by the vector R and the latter is strongly related to the space $\mathcal{L}_r$, the discrepancy between $\mathcal{L}_r$ and $\mathcal{X}^{(L,1)}$ produces an error in the LRR governing the series of forecasts. In particular, the finite-difference dimension of the series of forecasts y_n ($n > N$) is generally larger than r.

The other origin of the forecasting error lies in the initial data used to build the forecast. In the recurrent continuation, the initial data is $x_{N-L+2}^{(1)}, \ldots, x_N^{(1)}$, where $x_n^{(1)}$ is the nth term of the series $\mathbb{X}_N^{(1)}$. In Basic SSA R-forecasting algorithm, the initial data consists of the last $L-1$ terms $y_{N-L+2}, \ldots, y_N$ of the reconstructed series. Since generally $x_n^{(1)} \neq y_n$, the initial data used in LRR is a source of forecasting errors. The splitting of the whole error into two parts is investigated in [10] by simulations. For L close to $N/2$ these parts are comparable, while for small L the contribution of the error caused by the wrong reconstruction is larger.

On the other hand, if the quality of approximate separability of $\mathbb{X}_N^{(1)}$ and $\mathbb{X}_N^{(2)}$ is rather good and we select the proper eigentriples associated with $\mathbb{X}^{(1)}$, then we can expect that the linear spaces $\mathcal{L}_r$ and $\mathcal{X}^{(L,1)}$ are close. Therefore, the coefficients in the LRR (3.2) are expected to be close to those of the LRR governing the recurrent continuation of the series $\mathbb{X}_N^{(1)}$. Similarly, approximate separability implies that the reconstructed series y_n is close to $x_n^{(1)}$ and therefore the error due to the imprecision of the initial data used for forecasting is also small. As a result, in this case we can expect that Basic SSA R-forecasting procedure provides a reasonably accurate approximation to the recurrent continuation of $\mathbb{X}_N^{(1)}$, at least in the first few steps.

Remark 3.6 Since the forecasting procedure contains two generally unrelated parts, namely, estimation of the LRR and estimation of the reconstruction, we can modify these two parts of the algorithm separately. For example, for forecasting a signal, the LRR can be applied to the initial time series if the last points of the reconstruction are expected to be biased. Another modification of the forecasting procedure is considered in [10] and consists in the use of different window lengths to estimate the LRR and to reconstruct the time series.

3.3.2 Approximate Continuation and the Characteristic Polynomials

In this section we continue the discussion about the errors of separability and forecasting. The discrepancy between $\mathcal{L}_r$ and $\mathcal{X}^{(L,1)}$ can be described in terms of the characteristic polynomials.

We have three LRRs: (i) the minimal LRR of dimension r governing the series $\mathbb{X}_N^{(1)}$, (ii) the continuation LRR of dimension $L-1$, which also governs $\mathbb{X}_N^{(1)}$, but produces $L-r-1$ extraneous roots in its characteristic polynomial P_{L-1}, and (iii) the forecasting min-norm LRR governing the series of forecasts y_n ($n > N$).

The characteristic polynomial $P_{L-1}^{(x)}$ of the forecasting LRR and continuation polynomial P_{L-1} have $L-1$ roots. If $\mathcal{L}_r$ and $\mathcal{X}^{(L,1)}$ are close, then the coefficients of the continuation and forecasting recurrence relations must be close too. Therefore, all simple roots of the forecasting characteristic polynomial $P_{L-1}^{(x)}$ must be close to that of the continuation polynomial P_{L-1}. The roots μ_m with multiplicities $k_m > 1$ could be perturbed in a more complex manner.

Example 3.3 Perturbation of the multiple roots

Let us consider the series $\mathbb{X}_N$ with

$$x_n = (A + 0.1\,n) + \sin(2\pi n/10), \quad n = 0, \ldots, 199.$$

Evidently, $\mathbb{X}_N = \mathbb{X}_N^{(1)} + \mathbb{X}_N^{(2)}$ with the linear series $\mathbb{X}_N^{(1)}$ defined by $x_{n+1}^{(1)} = A + 0.1\,n$ and the harmonic series $\mathbb{X}_N^{(2)}$ corresponding to $x_{n+1}^{(2)} = \sin(2\pi n/10)$.

The series $\mathbb{X}_N$ has rank 4 and is governed by the minimal LRR of order 4. Therefore, any LRR governing $\mathbb{X}_N$ produces a characteristic polynomial with four main roots. These main roots do not depend on A; the linear part of the series generates one real root $\mu = 1$ of multiplicity 2, while the harmonic series corresponds to two complex conjugate roots $\rho e^{\pm \mathrm{i} 2\pi\omega}$ with modulus $\rho = 1$ and frequency $\omega = 0.1$.

Our aim is to forecast the series $\mathbb{X}_N^{(1)}$ for $A = 0$ and $A = 50$ with the help of Basic SSA R-forecasting algorithm. In both cases, we take the window length $L = 100$ and choose the eigentriples that correspond to the linear part of the initial time series $\mathbb{X}_N$. (For $A = 0$ we take the two leading eigentriples, while for $A = 50$ the appropriate eigentriples have the ordinal numbers 1 and 4.) Since the series $\mathbb{X}_N^{(1)}$ and $\mathbb{X}_N^{(2)}$ are not exactly separable for any A and any choice of L, we deal with approximate separability.

The forecasting polynomials $P_{L-1}^{(x)}$ with $A = 0$ and $A = 50$ demonstrate different splitting of the double root $\mu = 1$ into two simple ones. For $A = 0$ there appear two complex conjugate roots with $\rho = 1.002$ and $\omega = 0.0008$, while in the case $A = 50$ we obtain two real roots equal to 1.001 and 0.997. All extraneous roots are smaller than 0.986.

This means that for $A = 0$ the linear series $\mathbb{X}_N^{(1)}$ is approximated by a low-frequency harmonic with slightly increasing exponential amplitude. In the case $A = 50$ the approximating series is the sum of two exponentials, one of them is slightly increasing and another one is slightly decreasing.

These discrepancies lead to quite different long-term forecasting results: oscillating for $A = 0$ and exponentially increasing for $A = 50$. For short-term forecasting this difference is not important.

Let us consider the part of the forecasting error caused by errors in the initial data, that is, in the reconstruction of the forecasted series component. If the LRR is not minimal ($L > r + 1$), then the corresponding characteristic polynomial P_{L-1} has $L - 1 - r$ extraneous roots. If there is no reconstruction error, then the extraneous roots do not affect the forecast behavior, since the coefficients c_{mj} in (3.13) for the corresponding summands are equal to zero. However, if one applies the LRR to the perturbed initial terms, then the extraneous roots start to affect the forecasting results. The extraneous roots of the min-norm LRR lie within the unit circle and their effect on the forecasting decreases for long-term forecasting. Unfortunately, the minimal LRR is not appropriate for forecasting as it is very sensitive to errors in the initial data. Hence the presence of extraneous roots should be taken into account.

In the case of approximate separability, the min-norm LRR is found approximately. As a consequence, the extraneous roots can have absolute values larger than 1. The extraneous roots with moduli greater than 1 are the most hazardous, since the extraneous summand μ^n in (3.13), caused by an extraneous root μ with $|\mu| > 1$, grows to infinity. Therefore, it is important to look at the extraneous roots of the LRR used for forecasting.

If the forecasted series component $\mathbb{X}_N^{(1)}$ is the signal, then the main roots can be called signal roots. Note that the scan of extraneous roots should be used both to find a parametric form (3.13) of the signal (then we should identify the signal roots and remove extraneous roots) and also to forecast the signal (then we do not need to know the values of the roots but we would like to have no extraneous roots outside the unit circle).

Since the forecasting LRR is fully determined by the roots of its characteristic polynomial certain manipulations with the polynomial roots can be performed to modify the R-forecasting procedure.

- Let the main roots of the min-norm LRR of order $L-1$ be identified or estimated (e.g. by ESPRIT, see Sect. 3.8.2). For example, for a time series with the signal components which are not decreasing, the estimated main roots typically have maximal moduli, since the extraneous roots lie inside the unit circle. Thereby, we obtain the estimated minimal LRR (which is also the min-norm LRR of order r). However, it follows from the definition of the minimum norm that the norm of coefficients of the minimal LRR is larger than that of the min-norm LRR of order $L-1$ for $L > r+1$. Therefore, the forecast by the minimal LRR is more sensitive to errors in the initial data. Simulations demonstrate that the use of the minimal LRR usually does not gives the most accurate forecast and, moreover, these forecasts are often rather unstable.
- A safe way of correcting the LRR is by adjusting the identified main roots when an additional information about the time series is available. For example, if we know that the forecasted oscillations have stationary periodicities with constant amplitudes, then we know that the root moduli are equal to one and therefore the corresponding roots can be substituted with $\mu' = \mu/\|\mu\|$. If there is a periodicity with known period in the time series, then we can correct the arguments of the corresponding roots (for example, to 1/12, 1/6 and so on for a time series with seasonality).
- If the main roots have been estimated, then the explicit formula for the time series values in the form (3.13) can be obtained (with estimation of c_{mj} by the least squares method) and the forecast can be produced by this explicit formula. However, the explicit forecast needs root estimation, while the R-forecasting does not need root estimation and therefore is more robust.

3.4 Confidence Bounds for the Forecast

There are several conventional ways of estimating the accuracy of a forecast. Most of them can be applied for forecasting of the signal in the signal plus noise model.

1. Theoretical confidence intervals can be constructed if the model of time series is known and there are theoretical results about the distribution of the forecast.
2. Bootstrap confidence intervals can be constructed if the model is estimated in the course of analysis of the time series.

3. The accuracy of forecasting can be tested by removing the last points and then forecasting their values (the so-called *retrospective forecast*). This can be repeated with the cut made at different points.
4. If we are not interested in the retrospective forecast (we really need to forecast the future) and cannot reliably build an SSA model (as well as any other model) then we can use the following approach: we build a large number of SSA forecasts (e.g. using a variety of L and different but reasonable grouping schemes) and compare the forecast values at the horizon we are interested in. If the forecasts are going all over the place then we cannot trust any of them. If however the variability of the constructed forecasts is small then we (at least partly) may trust them. This approach has been applied in [22] for forecasting Earth's temperatures, where it is shown that the forecasts of Earth's temperatures for the next 5–10 years became very stable if forecasts use the records up to 2008 or later. On contrast, if we use temperature records only until 2005 or 2006 then SSA forecasting with different parameters gives totally different forecasts; see [22] for more details and explanations.

If there is a set of possible models, then the model can be chosen by minimizing the forecasting errors. An adjustment taking into account the number of parameters in the models should be made similar to Akaike-like methods or by using degree-of-freedom adjustments.

We do not consider the theoretical approach for estimating accuracy as there are not enough theoretical results which would estimate the precision of SSA forecasts theoretically. Below in this section we consider bootstrap confidence intervals in some detail. Since the construction of bootstrap confidence intervals is very similar to that of the Monte Carlo confidence intervals, we also consider Monte Carlo techniques for the investigation of the precision of reconstruction and forecasting. Note that by constructing bootstrap confidence intervals for forecasting values we also obtain confidence limits for the reconstructed values.

3.4.1 Monte Carlo and Bootstrap Confidence Intervals

According to the main SSA forecasting assumptions, the component $\mathbb{X}_N^{(1)}$ of the series $\mathbb{X}_N$ ought to be governed by an LRR of relatively small dimension, and the residual series $\mathbb{X}_N^{(2)} = \mathbb{X}_N - \mathbb{X}_N^{(1)}$ ought to be approximately strongly separable from $\mathbb{X}_N^{(1)}$ for some window length L. In particular, $\mathbb{X}_N^{(1)}$ is assumed to be a finite subseries of an infinite series, which is a recurrent continuation of $\mathbb{X}_N^{(1)}$. These assumptions hold for a wide class of practical series.

To establish confidence bounds for the forecast, we have to apply even stronger assumptions, related not only to $\mathbb{X}_N^{(1)}$, but to $\mathbb{X}_N^{(2)}$ as well. We assume that $\mathbb{X}_N^{(2)}$ is a finite subseries of an infinite random noise series $\mathbb{X}^{(2)}$ that perturbs the signal $\mathbb{X}^{(1)}$.

We only consider Basic SSA R-forecasting method. All other SSA forecasting procedures can be treated analogously.

Let us consider a method of constructing confidence bounds for the signal $\mathbb{X}^{(1)}$ at the moment of time $N+M$. In the unrealistic situation, when we know both the signal $\mathbb{X}^{(1)}$ and the true model of the noise $\mathbb{X}^{(2)}_N$, a direct Monte Carlo simulation can be used to check statistical properties of the forecast value $\widetilde{x}^{(1)}_{N+M}$ relative to the actual value $x^{(1)}_{N+M}$. Indeed, assuming that the rules for the eigentriple selection are fixed, we can simulate Q independent copies $\mathbb{X}^{(2)}_{N,i}$ of the process $\mathbb{X}^{(2)}_N$ and apply the forecasting procedure to Q independent time series $\mathbb{X}_{N,i} \stackrel{\text{def}}{=} \mathbb{X}^{(1)}_N + \mathbb{X}^{(2)}_{N,i}$. Then the forecasting results will form a sample $\widetilde{x}^{(1)}_{N+M,i}$ $(1 \le i \le Q)$, which should be compared against $x^{(1)}_{N+M}$. In this way, *Monte Carlo confidence bounds* for the forecast can be build.

Since in practice we do not know the signal $\mathbb{X}^{(1)}_N$, we cannot apply this procedure. Let us describe the bootstrap procedure for constructing the confidence bounds for the forecast (for a general methodology of bootstrap, see, for example, [8, Sect. 5]).

For a suitable window length L and the grouping of eigentriples, we have the representation $\mathbb{X}_N = \widetilde{\mathbb{X}}^{(1)}_N + \widetilde{\mathbb{X}}^{(2)}_N$, where $\widetilde{\mathbb{X}}^{(1)}_N$ (the reconstructed series) approximates $\mathbb{X}^{(1)}_N$, and $\widetilde{\mathbb{X}}^{(2)}_N$ is the residual series. Suppose now that we have a (stochastic) model of the residuals $\widetilde{\mathbb{X}}^{(2)}_N$. For instance, we can postulate some model for $\mathbb{X}^{(2)}_N$ and, since $\widetilde{\mathbb{X}}^{(1)}_N \approx \mathbb{X}^{(1)}_N$, apply the same model for $\widetilde{\mathbb{X}}^{(2)}_N$ with the estimated parameters. Then, simulating Q independent copies $\widetilde{\mathbb{X}}^{(2)}_{N,i}$ of the series $\mathbb{X}^{(2)}_N$, we obtain Q series $\mathbb{X}_{N,i} \stackrel{\text{def}}{=} \widetilde{\mathbb{X}}^{(1)}_N + \widetilde{\mathbb{X}}^{(2)}_{N,i}$ and produce Q forecasting results $\widetilde{x}^{(1)}_{N+M,i}$ in the same manner as in the straightforward Monte Carlo simulation.

More precisely, any time series $\mathbb{X}_{N,i}$ produces its own reconstructed series $\widetilde{\mathbb{X}}^{(1)}_{N,i}$ and its own forecasting linear recurrence relation LRR_i for the same window length L and the same set of the eigentriples. Starting at the last $L-1$ terms of the series $\widetilde{\mathbb{X}}^{(1)}_{N,i}$, we perform M steps of forecasting with the help of its LRR_i to obtain $\widetilde{x}^{(1)}_{N+M,i}$.

As soon as the sample $\widetilde{x}^{(1)}_{N+M,i}$ $(1 \le i \le Q)$ of the forecasting results is obtained, we can calculate its (empirical) lower and upper quantiles of some fixed level γ and obtain the corresponding confidence interval for the forecast. This interval will be called the *bootstrap confidence interval.* Simultaneously with the bootstrap confidence intervals for the signal forecasting values, we obtain the bootstrap confidence intervals for the reconstructed values.

The average of the bootstrap forecast sample (*bootstrap average forecast*) estimates the mean value of the forecast, while the mean square deviation of the sample shows the accuracy of the estimate.

The simplest model for $\widetilde{\mathbb{X}}^{(2)}_N$ is the model of Gaussian white noise. The corresponding hypothesis can be checked with the help of the standard tests for randomness and normality.

3.4.2 *Confidence Intervals: Comparison of Forecasting Methods*

The aim of this section is to compare different SSA forecasting procedures using several artificial series and the Monte Carlo confidence intervals.

Let $\mathbb{X}_N = \mathbb{X}_N^{(1)} + \mathbb{X}_N^{(2)}$, where $\mathbb{X}_N^{(2)}$ is Gaussian white noise with standard deviation σ. Assume that the signal $\mathbb{X}_N^{(1)}$ admits a recurrent continuation. We shall perform a forecast of the series $\mathbb{X}_N^{(1)}$ for M steps using different versions of SSA forecasting and appropriate eigentriples associated with $\mathbb{X}_N^{(1)}$. Several effects will be illustrated in the proposed simulation study. First, we shall compare some forecasting methods from the viewpoint of their accuracy. Second, we shall demonstrate the role of the proper choice of the window length.

We will consider two examples. In both of them, $N = 100$, $M = 50$ and the standard deviation of the Gaussian white noise $\mathbb{X}_N^{(2)}$ is $\sigma = 1$. The confidence intervals are obtained in terms of the 2.5 % upper and lower quantiles of the corresponding empirical c.d.f. using the sample size $Q = 10000$.

3.4.2.1 Periodic Signal: Recurrent and Vector Forecasting

Let us consider a periodic signal $\mathbb{X}_N^{(1)}$ of the form

$$x_n^{(1)} = \sin(2\pi n/17) + 0.5\sin(2\pi n/10).$$

The series $\mathbb{X}_N^{(1)}$ has difference dimension 4, and we use four leading eigentriples for its forecasting under the choice $L = 50$. The initial series $\mathbb{X}_N = \mathbb{X}_N^{(1)} + \mathbb{X}_N^{(2)}$ and the signal $\mathbb{X}_N^{(1)}$ (the thick line) are depicted in Fig. 3.2a.

Let us apply the Monte Carlo simulation for Basic SSA recurrent and vector forecasting algorithms. Figure 3.2b shows the confidence Monte Carlo intervals for both methods and the true continuation of the signal $\mathbb{X}_N^{(1)}$ (thick line). Confidence

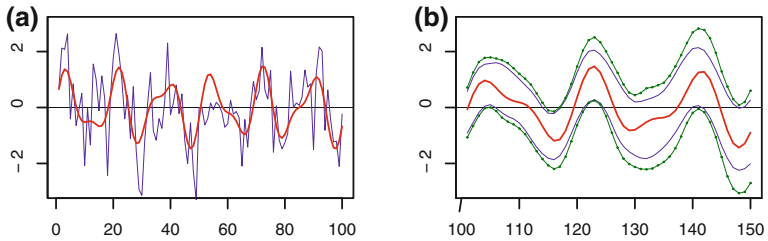

Fig. 3.2 Comparison of recurrent and vector forecasts. **a** Periodic signal and the initial series. **b** Confidence intervals

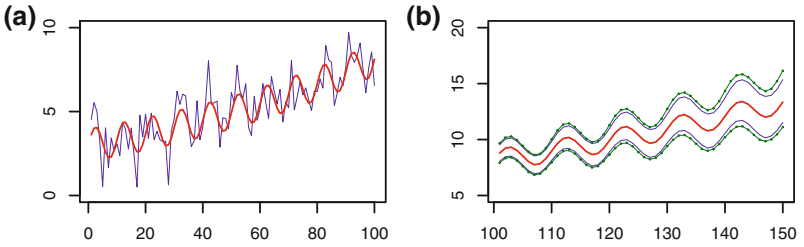

Fig. 3.3 Separability and forecasting. **a** The signal and the initial series. **b** Two confidence intervals

intervals for R-forecasting are marked by dots, while thin solid lines correspond to V-forecasting. We can see that these intervals practically coincide for relatively small numbers of forecasting steps, while V-forecasting has some advantage in the long-term forecasting.

3.4.2.2 Separability and Forecasting

Consider the series $\mathbb{X}_N^{(1)}$ with

$$x_n^{(1)} = 3a^n + \sin(2\pi n/10), \quad a = 1.01.$$

This series is governed by an LRR of dimension 3. Consider Basic SSA R-forecasting for up to 50 points of the signal values $x_{N+j}^{(1)}$ using the series $\mathbb{X}_N = \mathbb{X}_N^{(1)} + \mathbb{X}_N^{(2)}$. We compare two window lengths, $L = 15$ and $L = 50$. The first three eigentriples are chosen for the reconstruction in both choices of L. The series $\mathbb{X}_N$ and the signal $\mathbb{X}_N^{(1)}$ (thick line) are depicted in Fig. 3.3a.

Figure 3.3b shows that the Monte Carlo forecasting confidence intervals for $L = 15$ (thin line marked with dots) are apparently wider than that for $L = 50$. This is not surprising since the choice $L = 50$ corresponds to better separability. This is confirmed by comparing the values of the separability characteristics. In particular, the **w**-correlation (2.18) between the extracted signal and the residual is equal to 0.0083 for $L = 15$ and it equals 0.0016 for $L = 50$. Recall that the exact separability gives zero value for the **w**-correlation.

3.5 Summary and Recommendations on Forecasting Parameters

Let us summarize the material of the previous sections, taking as an example Basic SSA R-forecasting method. Other versions of SSA forecasting can be described and commented on similarly.

1. *Statement of the problem*
   We have a series $\mathbb{X}_N = \mathbb{X}_N^{(1)} + \mathbb{X}_N^{(2)}$ and need to forecast its component $\mathbb{X}_N^{(1)}$.
2. *The main assumptions*
   - The series $\mathbb{X}_N^{(1)}$ admits a recurrent continuation with the help of an LRR of relatively small dimension r.
   - There exists L such that the series $\mathbb{X}_N^{(1)}$ and $\mathbb{X}_N^{(2)}$ are approximately strongly separable for the window length L.
3. *Proper choice of parameters*
   Since we have to select the window length L providing a sufficient quality of separability and to find the eigentriples corresponding to $\mathbb{X}_N^{(1)}$, all the major rules and recommendations for the use of Basic SSA are applicable here. Note that in this case we must separate $\mathbb{X}_N^{(1)}$ from $\mathbb{X}_N^{(2)}$, but we do not need to obtain a detailed decomposition of the series $\mathbb{X}_N$.
4. *Specifics and potential pitfalls*
   The SSA forecasting problem has some specifics in comparison with Basic SSA reconstruction problem:
   - Since SSA forecasting procedure needs an estimation of the LRR, some recommendations concerning the window length can differ. In particular, SSA modifications that use different window lengths for the reconstruction and for building the forecasting formula can be used.
   - In Basic SSA, if we enlarge the set of proper eigentriples by some extra eigentriples with small singular values, then the result of reconstruction will essentially be the same. When dealing with forecasting, such an operation can produce large perturbations since the trajectory space $\mathcal{X}^{(L,1)}$ will be perturbed a lot; its dimension will be enlarged, and therefore the LRR governing the forecast will be modified. In this case, the magnitude of the extra singular values is not important but the location of the extraneous roots of the characteristic polynomials is important.
5. *Characteristics of forecasting*
   Let us mention several characteristics that might be helpful in judging the forecasting quality.
   - *Separability characteristics.* All separability characteristics considered in Sect. 2.3.3 are of importance for forecasting.
   - *Polynomial roots.* The roots of the characteristic polynomial of the forecasting LRR can give an insight into the behaviour of the forecast. These polynomial roots can be useful in answering the following two questions:
     (a) We expect that the forecast has some particular form (for example, we expect it to be increasing). Do the polynomial roots describe such a possibility? For instance, an exponential growth has to be indicated by a single real root (slightly) greater than 1 but if we try to forecast the annual periodicity, then pairs of complex roots with frequencies $\approx k/12$ have to exist.

(b) Although extraneous roots of the true min-norm LRR have moduli smaller than 1, the extraneous roots of the estimated LRR can be larger than 1. Since the polynomial roots with moduli greater than 1 correspond to the series components with increasing envelopes (see Sect. 3.2), large extraneous roots may cause problems even in the short-term forecasting. This is a serious pitfall that always has to be closely monitored.

- *Verticality coefficient.* The verticality coefficient ν^2 is the squared cosine of the angle between the space $\mathcal{L}_r$ and the vector $\mathbf{e}_L$. The condition $\nu^2 < 1$ is necessary for forecasting. The norm of the min-norm LRR (3.18) coefficients is equal to $\nu^2/(1-\nu^2)$. This characteristic reflects the ability of the LRR to decrease the noise level, see Proposition 3.4. If ν^2 is close to 1, then the norm is very large. This typically means that extra eigentriples are taken to describe $\mathbb{X}_N^{(1)}$ (alternatively, the whole approach is inadequate).

6. *The role of the initial data*
Apart from the number M of forecast steps, the formal parameters of Basic SSA R-forecasting algorithm are the window length L and the set I of eigentriples describing $\mathbb{X}_N^{(1)}$. These parameters determine both the forecasting LRR (3.1) and the initial data used in the forecasting formula. Evidently, the forecasting result essentially depends on this data, especially when the forecasting LRR has extraneous roots.
The SSA R-forecasting method uses the last $L-1$ terms $\widetilde{x}_{N-L+2}^{(1)}, \dots, \widetilde{x}_N^{(1)}$ of the reconstructed series $\widetilde{\mathbb{X}}_N^{(1)}$ as the initial data for forecasting. In view of the properties of the diagonal averaging, the last (and the first) terms of the series $\mathbb{X}_N^{(1)}$ are usually reconstructed with poorer precision than the middle ones. This effect may cause essential forecasting errors.
For example, any linear (and nonconstant) series $x_n = an + b$ is governed by the minimal LRR $x_n = 2x_{n-1} - x_{n-2}$, which does not depend on a and b. The parameters a and b used in the forecast are completely determined by the initial data x_1 and x_2. Evidently, errors in this data may essentially modify the forecast. Thus, it is important to check the last points of the reconstructed series (for example, to compare them with the expected future behaviour of the series $\mathbb{X}_N^{(1)}$). Even the use of the last points of the initial series as the initial data for the forecasting formula may improve the forecast.

7. *Reconstructed series and LRRs*
In the situation of strong separability between $\mathbb{X}_N^{(1)}$ and $\mathbb{X}_N^{(2)}$ and proper eigentriple selection, the reconstructed series is governed by the LRR which exactly corresponds to the series $\mathbb{X}_N^{(1)}$. Discrepancies in this correspondence indicate on possible errors: insufficient separability (which can be caused by the bad choice of the forecasting parameters) or general inadequacy of the model. We can suggest the following ways of testing for the presence of these errors and reducing them.

- *Global discrepancies.* Rather than using an LRR for forecasting, we can use it for approximation of either the whole reconstructed series or its subseries. For instance, if we take the first terms of the reconstructed series as the initial data (instead of the last ones) and make $N - L + 1$ steps of the procedure, we can check whether the reconstructed series can be globally approximated with the help of the LRR.
- *Local discrepancies.* The procedure above corresponds to the long-term forecasting. To check the short-term correspondence of the reconstructed series and the forecasting LRR, one can apply a slightly different method.
  This method is called the multistart recurrent continuation. According to it, for a relatively small M we perform M steps of the multi-start recurrent continuation procedure, modifying the initial data from $(\widetilde{x}_1^{(1)}, \dots, \widetilde{x}_{L-1}^{(1)})$ to $(\widetilde{x}_{K-M+1}^{(1)}, \dots, \widetilde{x}_{N-M}^{(1)})$, $K = N-L+1$. The M-step continuation is computed with the help of the forecasting LRR. The results should be compared with $\widetilde{x}_{L+M-1}^{(1)}, \dots, \widetilde{x}_N^{(1)}$. Since both the LRR and the initial data have errors, the local discrepancies for small M are usually more informative than the global ones. Moreover, by using different M we can estimate the maximal number of steps for a reasonable forecast.

Note that small discrepancies is only the necessary condition of accurate forecasting as the forecasting LRR is tested on the same points that were used for the calculation of the forecasting LRR.

8. *Forecasting stability and reliability*
   While the correctness of the forecast cannot be checked using the data only, the reliability of the forecast can be examined. Let us mention several methods for carrying out such an examination.

   - *Different algorithms.* We can try different forecasting algorithms (for example, recurrent and vector) with the same parameters. If their results approximately coincide, we have an argument in favour of the stability of forecasting.
   - *Different window lengths.* If the separability characteristics are stable under small variation in the window length L, we can compare the forecasts for different L.
   - *Forecasting of truncated series.* We can truncate the initial series $\mathbb{X}_N$ by removing the last few terms from it. If the separability conditions are stable under this operation, then we can forecast the truncated terms and compare the result with the initial series $\mathbb{X}_N$ and the reconstructed series $\widetilde{\mathbb{X}}_N^{(1)}$ obtained without truncation. If the forecast is regarded as adequate, then its continuation by the same LRR can be regarded as reliable.

9. *Confidence intervals*
   Confidence intervals discussed in Sect. 3.4 give important additional information about the accuracy and stability of the forecasts.

3.6 Case Study: 'Fortified Wine'

To illustrate SSA forecasting technique, we consider the time series 'Fortified wine' (monthly volumes of fortified wine sales in Australia from January 1984 till June 1994, Fig. 2.16). Naturally, time series forecasting should be based on the preliminary time series investigation. We examine both the initial time series of length 174 and its subseries consisting of the first 120 points. We name the former FORT174 and the latter FORT120.

SSA forecasting should only be applied to a time series governed (may be approximately) by some LRR. Therefore, we start with the study of the series from this point of view.

3.6.1 Linear Recurrence Relation Governing the Time Series

Preliminary analysis shows that the 'FORT174' time series (see Sects. 2.3.1.2 and 2.4.2.2) can be decomposed into a sum of a signal and a noise. For window length $L = 84$, the signal can be reconstructed by means of ET1–11 and the **w**-correlation between the signal component and the noise component is 0.004 which is small enough. Thus, the estimated signal subspace of R^L has dimension 11, the min-norm LRR has dimension $L - 1$ and the reconstructed time series (the signal) can be approximated by a time series governed by this LRR. For the series FORT120 and $L = 60$ the signal also corresponds to ET1–11, the **w**-correlation with the residual is slightly larger (equals 0.005).

Table 3.1 presents the information for 19 leading roots of the characteristic polynomial corresponding to two estimated min-norm LRR. The roots (recall that they are complex numbers) are ordered in the order of decrease of their moduli. The label 'compl.' for the 'Type' column of Table 3.1 notes that this line relates to two conjugate complex roots $\rho_j\, e^{\pm \mathrm{i}2\pi\omega_j}$, $0 < \omega_j < 0.5$. In this case, the period $1/\omega_j$ is listed in the table. The first six rows can be interpreted easily: the rows 1–3 and 5–6 correspond to conjugate complex roots, which produce harmonics with periods 6, 4, 2.4, 12, and 3. Moduli larger than one correspond to harmonics with increasing amplitudes, a modulus smaller than one yield a decreasing amplitude. The forth row of the table corresponds to the real-valued root with modulus 0.997. There are no more signal roots and all other roots are extraneous. All moduli of the extraneous roots are less than one. The column marked 'ET' indicates the correspondence between the eigentriples and the polynomial roots.

The series is decreasing and therefore the roots with modulus larger than 1 are most probably inadequate. Especially, the leading root (ET6–7) has modulus 1.013 for FORT120 which is a possible reason for an unstable forecast. Also, for FORT120 two harmonics are mixed; therefore, two pairs of conjugated roots put into correspondence with four eigentriples ET8–11.

Table 3.1 Time series FORT174 and FORT120: the leading roots of the characteristic polynomial for the min-norm LRR

FORT174, $L = 84$					FORT120, $L = 60$				
N	ET	Modulus	Period	Type	N	ET	Modulus	Period	Type
1	6–7	1.003	5.969	Compl.	1	6–7	1.013	5.990	Compl.
2	8–9	1.000	3.994	Compl.	2	8–11	1.007	2.376	Compl.
3	4–5	0.998	2.389	Compl.	3	4–5	1.000	4.001	Compl.
4	1	0.997	No	Real	4	1	0.997	No	Real
5	2–3	0.994	12.002	Compl.	5	2–3	0.994	12.033	Compl.
6	10–11	0.989	3.028	Compl.	6	8–11	0.982	3.002	Compl.
7		0.976	3.768	Compl.	7		0.968	5.311	Compl.
8		0.975	3.168	Compl.	8		0.966	9.635	Compl.
9		0.975	10.212	Compl.	9		0.966	3.688	Compl.
10		0.975	5.480	Compl.	10		0.965	2.268	Compl.

Let us check whether the time series FORT174 is well fitted by the estimated min-norm LRR. The maximum value of the global discrepancy between the reconstructed signal and its approximation by a time series governed by the used LRR (that is, the error of global approximation) is equal to 132 and it is smaller than 10 % of the time series values. Note that we use the first 83 points as the initial data for the LRR and so the approximation error is calculated starting from the 84th point.

Let us consider the minimal LRR of dimension 11 generated by the estimated signal roots presented in Table 3.1 above the horizontal line. (Recall that there is a one-to-one correspondence between LRRs and the roots of the associated characteristic polynomials.) If we take the points 73–83 as the initial data for this LRR, the series governed by the minimal LRR better approximates the time series (maximum discrepancy is equal to 94). Thus we conclude that the time series is well approximated by the time series governed by the minimal 11-dimensional LRR. Note that since the long-term forecast by the minimal LRR is very sensitive to the initial data, the choice of points 73–83 as the initial data was rather fortunate. The results for local approximation (discrepancy) are similar (magnitudes of errors are smaller while using the minimal LRR).

Since we know the exact period of the time series periodical component (due to its seasonal behavior), we can adjust the LRR by changing the roots so that they correspond to the periods 6, 4, 2.4, 12 and 3. This 11-dimensional formula is called an *adjusted minimal LRR*. The local approximation errors, corresponding to the adjusted minimal LRR, are slightly smaller than for the minimal LRR.

The analytic form of the time series governed by the adjusted minimal LRR is

$$\begin{aligned} y_n = {} & C_1 0.997^n + C_2 0.994^n \sin(2\pi n/12 + \varphi_2) \\ & + C_3 \sin(2\pi n/4 + \varphi_3) + C_4 1.003^n \sin(2\pi n/6 + \varphi_4) \\ & + C_5 0.998^n \sin(2\pi n/2.4 + \varphi_5) + C_6 0.989^n \sin(2\pi n/3 + \varphi_6). \end{aligned}$$

The coefficients C_i and φ_i are determined by the initial data. The terms are ordered by their eigenvalue shares (in the order of decrease). Recall that ordering by roots moduli is generally different from ordering by eigenvalues, since roots moduli are related to the rates of increase/decrease of the time series components and thereby influence a future behavior of the time series governed by the corresponding LRR.

Thus, a preliminary investigation implies that the time series FORT174 and FORT120 well fit to the respective models of the form required, so we can start their forecasting.

3.6.2 Choice of Forecasting Methods and Parameters

Let us demonstrate the approach to forecasting on the 'Fortified wine' example, investigating the accuracy of forecasting the values at the points 121–174 (the test period) on the base of the reconstruction of the points 1–120 (the base period 'FORT120'). The 12-point ahead and 54-point ahead forecasts are considered. Table 3.2 summarizes the errors of forecasts for different forecasting methods. The relative MSD errors of estimation of $\mathbb{Y}$ by $\widetilde{\mathbb{Y}}$ are calculated as

$$\|\widetilde{\mathbb{Y}} - \mathbb{Y}\|_{\mathrm{F}}/\|\mathbb{Y}\|_{\mathrm{F}} \cdot 100\,\%. \tag{3.19}$$

In Table 3.2, the column 'ET' shows the chosen numbers of the leading eigentriples, the column 'rec' gives the reconstruction errors, the columns 'vec12', 'rec12', 'vec54', 'rec54' correspond to vector and recurrent forecasting for the horizons 12 and 54 terms respectively. The suffix '_init' means that the forecasting formula was applied to the initial series rather than to the reconstructed one.

Below we enumerate the main points of the forecasting logic.

1. Note first that only a set of components separated from the residual may be chosen. For the 'FORT120' series the admissible numbers of components are 1, 3, 5, 7, or 11.
2. There is a conflict between the accuracy of reconstruction and stability of forecasting. In Table 3.2 the errors of reconstruction decrease (the column 'rec') while the errors of forecasts decrease in the beginning and increase later. Note that all

Table 3.2 Time series FORT120: relative MSD errors of the reconstruction and forecasts

ET	rec (%)	vec12 (%)	rec12 (%)	rec_init12 (%)	vec54 (%)	rec54 (%)	rec_init54 (%)
1	23.11	23.34	23.46	23.49	23.84	23.73	24.02
3	14.79	15.82	16.19	16.41	17.60	17.78	18.17
5	11.63	15.49	15.58	15.44	15.23	15.23	15.57
7	9.70	14.13	15.65	14.41	15.12	24.98	23.26
11	7.45	16.76	17.48	15.59	21.34	23.30	20.57

considered components are related to the signal and therefore the increase of errors is related to instability.

3. The observed behaviour of the forecasting errors means that the optimal number of the components for forecasting is 7 for 12-term ahead and 5 for 54-term ahead forecasts. This is a natural result since the stability of forecasting is much more important for the long-term forecasting.
4. The vector forecasting method provides more stable forecast of 'FORT120'. This is clearly seen on the long-term forecast.
5. Comparison of the forecasting methods can be performed by means of the confidence intervals: smaller size of the confidence intervals indicates better stability of forecasting. This approach does not help to choose the optimal number of components, since the rough forecast can be the most stable. However, this is a good tool to compare the forecast modifications for a fixed number of components. In particular, the size of the bootstrap confidence interval for ET1–7 is one and half times smaller for the vector forecast than that for the recurrent forecast.
6. Table 3.2 shows that generally the forecasting formula can be applied to the initial time series (the columns with the suffix '_init') instead of the reconstructed one. However, there is no noticeable improvement.
7. The linear recurrence relation can be adjusted by two ways. The first modification is to remove extraneous roots and to adjust the signal roots using the known information. For the 'FORT120' series we know the periods of the seasonal component. The modified LRR is closer to the true LRR. However, the forecast is very unstable and gives the forecasting error several times larger than the min-norm LRR. The reason is that the initial data with error can cardinally change the amplitudes of the true harmonics. Certainly, the minimal LRR should be applied to the reconstructed series.
8. A specific feature of this dataset is that the behaviour of the series is close to multiplicative. However, this time series is not pure multiplicative since the form of the seasonal period differ from year to year (Fig. 2.17). The last conclusions is confirmed by different moduli of the roots. For the initial time series the leading harmonic with period 12 is decreasing and the estimated modulus of the corresponding root is equal to 0.994. Therefore, the decreasing exponential has stable behaviour, regardless of the estimation errors. After the log-transformation of a multiplicative series the root modulus becomes close to 1 and the estimation error can give the modulus of the estimated root larger than 1; that is, the forecast (especially, long-term) could be unstable. The 'FORT120' series demonstrates this effect, since the forecasting error for the log-transformed data is larger than that for the original data.

Figure 3.4 shows the last 4 years of the series 'FORT120' (thin line) and two vector forecasts for 54 points ahead: the stable and accurate forecast based on ET1–5 (boldface line) and the forecast with unstable and less accurate behaviour based on ET1–11 (line with circles). Note that the accuracy of forecasting for 12-points ahead is approximately the same for both forecasts.

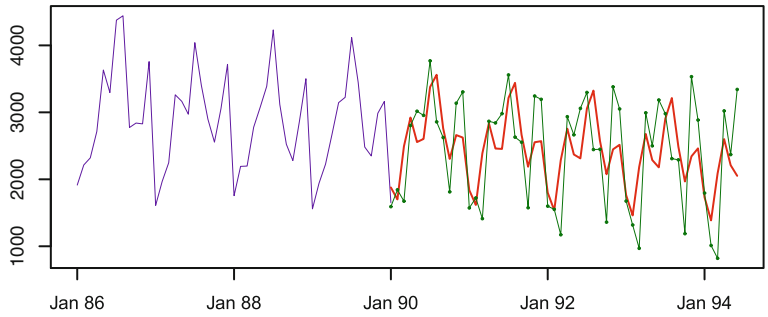

Fig. 3.4 FORT120: two forecasts

Summarizing we make the following conclusions concerning our experience with forecasting the 'Fortified wine' series: (1) it is better to use the original time series rather than its log-transformed version; (2) the best eigentriple group used for forecasting is either ET1–5 for long-term forecast or ET1–7 for short-term forecast; (3) V-forecasting is more accurate than R-forecasting.

The conclusion about the accuracy of the forecasts is made on the base of comparison of the forecasts with series values in the forecasted points. Stability of the forecasts can be checked by means of the confidence intervals and does not need the knowledge of the series values. For two considered forecasts, the size of confidence intervals for ET1–11 is more than twice larger than that for ET1–5, if we take the last forecasted year. Thus, this example demonstrates that the more accurate long-term forecast corresponds to the more stable one.

3.7 Missing Value Imputation

This section is devoted to the extension of SSA forecasting algorithms for the analysis of time series with missing data.

The following three approaches for solving this problem are known. The first approach was suggested in [24]. This approach is suitable for stationary time series only and uses the following simple idea: in the process of the calculation of inner products of vectors with missing components we use only pairs of valid vector components and omit the others.

The second '*Iterative*' approach uses an iterative interpolation. Initially, the places of missing values are filled with arbitrary numbers. Then these numbers are iteratively refined by the successive application of SSA. After each iteration, the values at the places of missing values are taken from the previous iteration but the other values are taken from the initial time series. This approach can be formally applied for almost any location of missing values. Therefore, several artificial gaps can be added and then be used to justify the choice of SSA parameters, namely, the window length and

the number of chosen components. This idea was suggested in [4] for the imputation of missing values in matrices and then was extended to time series in [15]. The iterative approach has the semi-empirical reasoning of convergence. However, even for noiseless signals the gaps cannot be filled in one iteration. Therefore, this method has large computational cost. Also, it does not provide exact imputation and it needs an additional information about the subspace dimension.

The third approach of filling in missing data is an extension of SSA forecasting algorithms. This approach is considered below in this section and is called '*the subspace approach*'. According to this approach we continue the structure of the extracted component to the gaps caused by the missing data [11]. The theory of SSA assumes that the forecasted component is (or is approximated by) a time series of finite rank. The theoretical results concerning the exact reconstruction of missing values are also based on this assumption. Nevertheless, the constructed algorithms are applicable to real-life time series with missing values where they give approximate results.

Note that in a particular case, when the missing values are located at the end of the series, the problem of their filling in coincides with the problem of forecasting.

3.7.1 SSA for Time Series with Missing Data: Algorithm

3.7.1.1 The Layout of the Algorithm

As above, we assume that any application of SSA gives us a decomposition of the observed time series into additive components such as trend, periodic components, and noise. The SSA algorithm for time series with no missing data consists of two stages: decomposition and reconstruction. Each stage, in turn, consists of two steps: Embedding and Singular Value Decomposition are the steps of the first stage, Grouping and Diagonal Averaging are the steps of the second stage. The general structure of the algorithm for the analysis of time series with missing data is the same, but the steps are somewhat different.

Assume that we have the initial time series $\mathbb{X}_N = (x_1, \dots, x_N)$ consisting of N elements, some part of which is unknown. Let us describe the scheme of the algorithm in the case of reconstruction of the first component $\mathbb{X}_N^{(1)}$ of the observed series $\mathbb{X}_N = \mathbb{X}_N^{(1)} + \mathbb{X}_N^{(2)}$. The notation of Sect. 3.1.4 is used.

First Stage: Decomposition

Step 1. Embedding. Let us fix the window length L, $1 < L < N$. The embedding procedure transforms the initial time series into the sequence of L-dimensional lagged vectors $\{X_i\}_{i=1}^{K}$, where $K = N - L + 1$. Some of the lagged vectors may be incomplete, i.e., contain missing components. Let C be the set of indices such that

the lagged vectors X_i with $i \in C$ are complete. Let us collect all complete lagged vectors X_i, $i \in C$, into the matrix $\widetilde{\mathbf{X}}$. Assume that this matrix is non-empty. If there are no missing values, then the matrix $\widetilde{\mathbf{X}}$ coincides with the trajectory matrix of the series $\mathbb{X}_N$.

Step 2. Finding the basis. Let $\widetilde{\mathbf{S}} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}}$. Denote by $\lambda_1 \geq \ldots \geq \lambda_L \geq 0$ the ordered eigenvalues of the matrix $\widetilde{\mathbf{S}}$ and by $U_1, \ldots, U_L$ the orthonormal system of the eigenvectors of the matrix $\widetilde{\mathbf{S}}$ corresponding to these eigenvalues, $d = \max\{i : \lambda_i > 0\}$.

Second Stage: Reconstruction

Step 3a. Choosing the subspace and projection of the complete lagged vectors. Let a set of indices $I_r = \{i_1, \ldots, i_r\} \subset \{1, \ldots, d\}$ be chosen and the subspace $\mathcal{M}_r = \mathrm{span}(U_{i_1}, \ldots, U_{i_r})$ be formed. The choice of the eigenvectors (i.e., their indices) corresponding to $\mathbb{X}_N^{(1)}$ is the same as in Basic SSA. The complete lagged vectors can be projected onto the subspace $\mathcal{M}_r$ in the usual way:

$$\widehat{X}_i = \sum_{k \in I_r} (X_i, U_k) U_k, \quad i \in C.$$

Step 3b. Projection of the incomplete lagged vectors. For each $\mathcal{Q}$-incomplete lagged vector with missing components in the positions from the set $\mathcal{Q}$, the given step consists of two parts:

(α) calculation of $\widehat{X}_i\big|_{J_L \setminus \mathcal{Q}}$, $\quad i \notin C$,
(β) calculation of $\widehat{X}_i\big|_{\mathcal{Q}}$, $\quad i \notin C$.

Since adjacent lagged vectors have common information (the trajectory matrix (2.1) consisting of the lagged vectors is Hankel) there are many possible ways of solving the formulated problems. Some of these ways will be discussed in the following sections. The common information also enables processing of 'empty' vectors with $\mathcal{Q} = J_L = \{1, \ldots, L\}$. Note that Step 3b may change the vectors $\widehat{X}_i$, $i \in C$. The result of Steps 3a and 3b is the matrix $\widehat{\mathbf{X}} = [\widehat{X}_1 : \ldots : \widehat{X}_K]$, which serves as an approximation to the trajectory matrix of the series $\mathbb{X}_N^{(1)}$, under the proper choice of the set I_r.

Step 4. Diagonal averaging. At the last step of the algorithm, the matrix $\widehat{\mathbf{X}}$ is transformed into the new series $\widetilde{\mathbb{X}}_N^{(1)}$ (the reconstructed time series) by means of the diagonal averaging.

3.7.1.2 Clusters of Missing Data

Implementation of Step 3b for projecting the incomplete vectors needs a definition of clusters of missing data and their classification assuming that L is fixed.

A sequence of missing data of a time series is called a *cluster of missing data* if every two adjacent missing values from this sequence are separated by less than L non-missing values and there is no missing data among L neighbours (if they exist) of the left/right element of the cluster. Thus, a group of not less than L successive non-missing values of the series separates clusters of missing data.

A cluster is called *left/right* if its left/right element is located at a distance of less than L from the left/right end of the series. If the left or the right element of the cluster coincides with the end of the series, the cluster is called *extreme*. Neither left nor right cluster is called *inner*. A cluster is called *continuous* if it consists of successive missing data.

Step 3b can be performed independently for each cluster of missing data (for each lagged-vector set).

3.7.1.3 Methods for Step 3b

Different realizations of Step 3b are thoroughly considered in [11]. Here we briefly describe several typical versions and their relation to SSA forecasting methods formulated in Sect. 3.1.

Let the window length L and the indices of the eigentriples corresponding to the chosen time series component be fixed. Propositions 3.1 and 3.3 (where we take $n = L$, $m = r$, $I_r = \{i_1, \ldots, i_r\}$, $\mathbf{P} = [U_{i_1} : \ldots : U_{i_r}]$) provide the base for the methods of filling in.

If the considered cluster is continuous and is not left, then (3.9) with $\mathcal{Q} = \{L\}$ provides the coefficients of recurrence relation that can be applied to the reconstructed points that lie on the left from the missing data cluster (*sequential filling in from the left*). Similarly, setting $\mathcal{Q} = \{1\}$ and applying the backward recurrence relation (3.9) to the reconstructed data taken from the right side, *sequential filling in from the right* can be introduced. Different combinations of the sequential fillings in from the left and from the right (the so-called two-sided methods) can be constructed. For example, their average can be used.

Remark 3.7 Consider a continuous cluster of missing data of length M, which is a right extreme cluster (and assume that there are no other clusters of missing data in the series). If the sequential method described above is applied to this cluster, then the result will coincide with the recurrent forecast for M terms ahead (Sect. 3.1.2.1), where the forecast is constructed on the first $N - M$ points of the time series and the same parameters L and I_r.

In the same manner as we have used for the vector forecasting (Sect. 3.1.2.2), the vector coordinates at the positions of non-missing components can be filled with the

help of the adjacent complete vectors and then projected to $\mathcal{M}|_{J_L \setminus \mathcal{Q}}$ by the projector given by formula (3.11) (‘*Π Projector*’).

Also, in the same manner as the simultaneous forecasting was introduced (see Sect. 3.1.2.3), the vector coordinates at the positions of missing components can be filled in simultaneously, not one by one as in the sequential filling in, since Proposition 3.2 allows filling in several vector coordinates at once (‘*simultaneous filling in*’). This may simplify the imputation of not-continuous clusters of missing data.

3.7.2 Discussion

- As well as for forecasting, the approach above allows filling in missing values in any component of the time series, not necessary in the signal. For example, missing values in the trend can be filled in. Certainly, an approximate separability of the imputed component from the residual is required.
- If the time series component is exactly separated from the residual, it can be filled in exactly.
- The location of missing data is very important for the possibility of imputation by the subspace method, since the number of non-missing values should be large enough for achieving separability. At least, the number of the complete lagged vectors should be larger than the rank of the forecasted time series component.
- If there are many randomly located missing data, then it can be impossible to extract sufficient number of lagged vectors. However, it is possible to estimate the subspace by involving the lagged vectors with a few missing entries, see [11] for details.

3.7.3 Example

To demonstrate the work of the methods of filling in missing data, let us consider the time series FORT120, which was investigated for forecasting in Sect. 3.6.

Let us remove 12 known values, starting with 60th point (i.e., we assume that the values for a year since December 1984 are unknown). For this artificially missing data we estimate the accuracy of their recovery for different versions of the algorithm. Also, to simulate forecast, we add 12 missing data after the last, 120th point of the series. The time series obtained is illustrated in Fig. 3.5a.

The first question is how to choose the window length L. In the case of no missing data, the general recommendation is to choose the window length close to $N/2$ and divisible by the period of expected periodicity (12 months here). The window length $L = 60$ meets these conditions. However, for $L = 60$ all lagged vectors will contain missing data. Hence we have to choose smaller L. The choice of $L = 36$ provides us with 38 complete lagged vectors with no missing data.

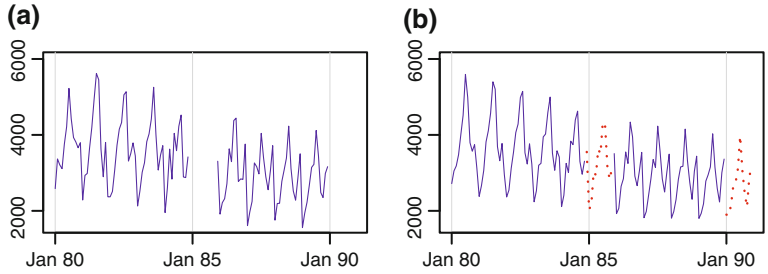

Fig. 3.5 FORT120: Filling in missing data. **a** Initial time series with missing data. **b** Reconstructed time series with filled in data

The analysis of the time series FORT120 in Sect. 3.6 shows that the eigenvectors with indices 1–7 provide the best forecast for 12 points ahead for the choice $L = 60$, while the whole signal is described by the 11 leading eigentriples. The structure of the eigentriples for $L = 36$ is similar and we can use the interpretation of the leading eigentriples found in Sect. 3.6.

The comparison of the filling in results with the values that were artificially removed from the initial time series shows an advantage of the version using 'Π Projector' with simultaneous filling in of the missing data and the choice $r = 11$, $I_r = \{1, 2, \ldots, 11\}$. This differs from the forecasting results obtained in Sect. 3.6. Note that the used method of filling in missing data at the end of the time series was not considered during forecasting. Therefore, the ideas of missing data imputation can extend the number of forecasting methods. On the other hand, the accuracy of imputation of missing data at the middle of the time series can be less sensitive to the precision of reconstruction than to the accuracy of forecasting.

The result of missing data imputation is illustrated in Fig. 3.5b. The reconstructed series is marked by the dotted line in the area of missing data. The relative MSD error (3.19) of reconstruction is approximately equal to 9 % for the missing data and to 6 % for the non-missing terms in the series.

Comparison with the Iterative Method

Let us apply the iterative method to the same FORT120 data with the same missing entries, at the middle of the series and at the end. If we replace the missing data by the average value of all valid series points, then 20 iterations are sufficient for convergence. The results are presented in Table 3.3. The errors are calculated as

Table 3.3 FORT120: MSD errors for iterative and subspace methods of filling in

Method	Middle	End	Total
Subspace $L = 36$	255.9	292.8	275.0
Iterative $L = 36$	221.2	333.0	282.7
Iterative $L = 60$	216.2	419.3	333.6

square root of the average squared deviations. For the missing values in the middle, the iterative method provides slightly smaller errors of reconstruction than the subspace method, while for the end points (that is, for forecasting) the iterative method is not stable with respect to the window length. Note that the choice $L = 60$ is not appropriate for the subspace method.

Simulations performed for noisy model series of finite rank in the form of a sum of several products of exponential and harmonic series confirm that the error of filling in missing data at the middle are similar for both methods, while the subspace method is more stable for forecasting.

3.8 Subspace-Based Methods and Estimation of Signal Parameters

While the problems of reconstruction and forecasting are traditionally included into the scope of problems solved by SSA, the estimation of signal parameters is usually not. At the same time, the estimation of signal parameters is the primary objective for many subspace-based methods of signal processing. In this section we follow [10] to describe the most common subspace-based methods and demonstrate their cohesion with SSA. For simplicity of notation we always assume $L \le K = N - L + 1$.

Let us shortly describe the problem. Consider the signal $\mathbb{S}_N = (s_1, \dots, s_N)$ in the form $s_n = \sum_{j=1}^{r} c_j \mu_j^n$, $n = 1, \dots, N$, where all μ_j are assumed to be different (the more complicated form (3.13) can be considered in a similar manner). The problem is to estimate μ_j observing the noisy signal. The $\mu_j = \rho_j e^{\mathrm{i}2\pi\omega_j}$ are expressed in terms of parameters ρ_j and ω_j which can often be interpreted. In particular, ω_j are the frequencies presented in the signal. Hence an estimator of μ_j provides the information about the structure of the signal which is distinct from the information we get from the coefficients c_j. Note that if the time series is real-valued, then s_n can be written as the sum of modulated sinusoids $A_j \rho_j^n \cos(2\pi\omega_j n + \varphi_j)$.

The idea of subspace-based methods is as follows. Let $r < N/2$. The signal $\mathbb{S}_N$ with $s_n = \sum_{j=1}^{r} c_j \mu_j^n$ has rank r and is governed by linear recurrence relations like $s_n = \sum_{m=1}^{t} a_m s_{n-m}$, $t \ge r$. Then μ_j can be found as the signal roots of the characteristic polynomial of a governing LRR (see Sect. 3.2). Simultaneously, the L-trajectory space ($L > r$) of the signal (the so-called signal subspace) has dimension r and is spanned by the vectors $(1, \mu_j, \dots, \mu_j^{L-1})^{\mathrm{T}}$. The coefficients of the governing LRRs of order $L - 1$ can also be found using the information about the signal subspace. Methods of estimating μ_j based on the estimation of the signal subspace are called *subspace-based methods*.

Since finding signal roots of the characteristic polynomial of the LRR governing the signal is very important for the estimation of the signal parameters, we start with several facts that allow estimation of signal roots as eigenvalues of some matrix.

3.8.1 Basic Facts

The following statement is obvious.

Proposition 3.5 *Roots of a polynomial* $p(\mu) = \mu^M + c_1\mu^{M-1} + \ldots + c_{M-1}\mu + c_M$ *coincide with eigenvalues of its companion matrix* $\mathbf{C}$ *defined by*

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 0 & \ldots & -c_M \\ 1 & 0 & 0 & \ldots & -c_{M-1} \\ 0 & 1 & \ldots & 0 & -c_{M-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & -c_1 \end{pmatrix}.$$

Note that the multiplicities of the roots of the polynomial $p(\mu)$ are equal to the algebraic multiplicities of the eigenvalues of its companion matrix (i.e., to the multiplicities of the roots of the characteristic polynomial of this matrix). However, these multiplicities do not always coincide with the geometric multiplicities which are equal to the dimensions of the eigenspaces corresponding to the eigenvalues.

To derive an analytic form of the signal ($s_n = \sum_{j=1}^{r} c_j \mu_j^n$ or see (3.13) for the general case), we need to find roots of the characteristic polynomial of the LRR which governs the signal. By Proposition 3.5, we have to find either the roots of the characteristic polynomial or the eigenvalues of its companion matrix. The latter does not require the full knowledge of the LRR. Let us demonstrate that for finding the signal roots it is sufficient to know the basis of the signal trajectory space.

Let $\mathbf{C}$ be a full-rank $d \times d$ matrix, $Z \in \mathsf{R}^d$, and $\mathbf{Z}$ be a full-rank $L \times d$ matrix ($L > d$), which can be expressed as

$$\mathbf{Z} = \begin{pmatrix} Z^{\mathrm{T}} \\ Z^{\mathrm{T}}\mathbf{C} \\ \vdots \\ Z^{\mathrm{T}}\mathbf{C}^{L-1} \end{pmatrix}. \tag{3.20}$$

Let us again denote the matrix $\mathbf{Z}$ without the last row by $\underline{\mathbf{Z}}$ and the matrix $\mathbf{Z}$ without its first row by $\overline{\mathbf{Z}}$. It is clear that $\overline{\mathbf{Z}} = \underline{\mathbf{Z}}\mathbf{C}$. We call this property of $\mathbf{Z}$ the *shift property* given by the matrix $\mathbf{C}$.

Proposition 3.6 *Let* $\mathbf{Z}$ *satisfy the shift property given by the matrix* $\mathbf{C}$, $\mathbf{P}$ *be a full-rank* $d \times d$ *matrix, and* $\mathbf{Y} = \mathbf{ZP}$. *Then the matrix* $\mathbf{Y}$ *satisfies the shift property given by the matrix* $\mathbf{D} = \mathbf{P}^{-1}\mathbf{CP}$, *i.e.,* $\overline{\mathbf{Y}} = \underline{\mathbf{Y}}\mathbf{D}$.

The proof of this proposition is straightforward.

Note that the multiplication by a nonsingular matrix $\mathbf{P}$ can be considered as a transformation of the vector coordinates in the column space of the matrix $\mathbf{Z}$.

It is easily seen that the matrices $\mathbf{C}$ and $\mathbf{D} = \mathbf{P}^{-1}\mathbf{C}\mathbf{P}$ have the same eigenvalues; these matrices are called *similar*.

Remark 3.8 Let the matrix $\mathbf{Y}$ satisfy the shift property given by the matrix $\mathbf{D}$. Then $\mathbf{D} = \underline{\mathbf{Y}}^{\dagger}\overline{\mathbf{Y}}$, where $\mathbf{A}^{\dagger}$ denotes the Moore-Penrose pseudoinverse of $\mathbf{A}$.

Proposition 3.7 *Let a time series $\mathbb{S}_N = (s_1, \ldots, s_N)$ satisfy the minimal LRR (3.12) of order d, $L > d$ be the window length, $\mathbf{C}$ be the companion matrix of the characteristic polynomial of this LRR. Then any $L \times d$ matrix $\mathbf{Y}$ with columns forming a basis of the trajectory space of $\mathbb{S}_N$ satisfies the shift property given by some matrix $\mathbf{D}$. Moreover, the eigenvalues of this shift matrix $\mathbf{D}$ coincide with the eigenvalues of the companion matrix $\mathbf{C}$ and hence with the roots of the characteristic polynomial of the LRR.*

Proof Note that for any $1 \le i \le N - d$ we have

$$(s_i, s_{i+1}, \ldots, s_{i+(d-1)})\mathbf{C} = (s_{i+1}, s_{i+2}, \ldots, s_{i+d}).$$

Therefore, (3.20) holds for $Z = (x_1, x_2, \ldots, s_d)^{\mathrm{T}}$. It can be easily proved that for a time series governed by the minimal LRR of order d, any d adjacent columns of the trajectory matrix are linearly independent. Consequently, the matrix $\mathbf{Z} = [S_1 : \ldots : S_d]$ is of full rank and we can apply Proposition 3.6.

Remark 3.9 The SVD of the L-trajectory matrix of a time series provides a basis of its trajectory space. Specifically, the left singular vectors which correspond to the nonzero singular values form such a basis. If we observe a time series of the form 'signal + residual', then the SVD of its L-trajectory matrix provides the basis of the signal subspace under the condition of exact strong separability of the signal and the residual.

3.8.2 ESPRIT

Consider a time series $\mathbb{X}_N = \{x_i\}_{i=1}^N$ with $x_i = s_i + p_i$, where $\mathbb{S}_N = \{s_i\}_{i=1}^N$ is a time series governed by an LRR of order r (that is, signal) and $\mathbb{P}_N = \{p_i\}_{i=1}^N$ is a residual (noise, perturbation). Let $\mathbf{X}$ be the trajectory matrix of $\mathbb{X}_N$. In the case of exact or approximate separability of the signal and the residual, there is a set I of eigenvector numbers in (2.12), which correspond to the signal. If the signal dominates, then $I = \{1, \ldots, r\}$ and the subspace $\mathcal{L}_r = \mathrm{span}\{U_1, \ldots, U_r\}$ can be considered as an estimate of the true signal subspace $\mathcal{S}$. Therefore, we can use $\widetilde{\mathbf{Y}} = \mathbf{U}_r = [U_1 : \ldots : U_r]$ as an estimate of $\mathbf{Y}$ from Proposition 3.7. Then the shift property is fulfilled approximately and $\underline{\mathbf{U}}_r\mathbf{D} \approx \overline{\mathbf{U}}_r$.

The method ESPRIT consists in estimation of the signal roots as the eigenvalues of a matrix $\widehat{\mathbf{D}}$, for which

$$\underline{\mathbf{U}}_r\widehat{\mathbf{D}} \approx \overline{\mathbf{U}}_r. \tag{3.21}$$

By estimating the signal roots ESPRIT provides estimates of the signal parameters.

Let us study the methods of finding the matrix $\widehat{\mathbf{D}}$. The main idea of LS-ESPRIT was introduced in the paper [18] devoted to the problem of estimating frequencies in a sum of sinusoids, in the presence of noise. The method was given the name ESPRIT in [23]; this name was later used in many other papers devoted to the DOA (Direction of Arrival) problem. For time series processing, LS-ESPRIT is also called Hankel SVD (HSVD, [3]). Later the so-called TLS-ESPRIT modification was suggested (see e.g. [28], where the method was called Hankel Total Least Squares (HTLS)). There are a number of papers devoted to the perturbation study of ESPRIT, see e.g. [2], where specific features of ESPRIT in the case of multiple roots are also described.

Remark 3.10 ESPRIT is able to estimate parameters of a separable time series component, not necessary the signal, if the matrix $\mathbf{U}_r$ consists of the corresponding eigenvectors.

3.8.2.1 Least Squares (LS-ESPRIT)

The LS-ESPRIT estimate of the matrix $\mathbf{D}$ is

$$\widehat{\mathbf{D}} = \underline{\mathbf{U}}_r^{\dagger}\overline{\mathbf{U}}_r = (\underline{\mathbf{U}}_r^{\mathrm{T}}\underline{\mathbf{U}}_r)^{-1}\underline{\mathbf{U}}_r^{\mathrm{T}}\overline{\mathbf{U}}_r. \tag{3.22}$$

The eigenvalues of $\widehat{\mathbf{D}}$ do not depend on the choice of the basis of the subspace $\mathcal{L}_r = \mathrm{span}\{U_1, \ldots, U_r\}$.

3.8.2.2 Total Least Squares (TLS-ESPRIT)

As $\mathbf{U}_r$ is known only approximately there are errors in both $\underline{\mathbf{U}}_r$ and $\overline{\mathbf{U}}_r$. Therefore, the solution of the approximate equality $\underline{\mathbf{U}}_r\mathbf{D} \approx \overline{\mathbf{U}}_r$ based on the method of Total Least Squares (TLS) can be more accurate.

Recall that to solve the equation $\mathbf{AX} \approx \mathbf{B}$, TLS minimizes the sum

$$\begin{gathered} ||\widetilde{\mathbf{A}} - \mathbf{A}||_{\mathrm{F}}^2 + ||\widetilde{\mathbf{B}} - \mathbf{B}||_{\mathrm{F}}^2 \longrightarrow \min, \\ \text{with respect to } \widetilde{\mathbf{A}} \text{ and } \widetilde{\mathbf{B}} \text{ such that } \exists \mathbf{Z}: \ \widetilde{\mathbf{A}}\mathbf{Z} = \widetilde{\mathbf{B}}. \end{gathered} \tag{3.23}$$

Set $\mathbf{A} = \underline{\mathbf{U}}_r$, $\mathbf{B} = \overline{\mathbf{U}}_r$ in (3.23). Then the matrix $\mathbf{Z}$ that minimizes (3.23) is called the TLS-estimate of $\mathbf{D}$ (see [7] for explicit formulas).

Let us consider the dependence of the TLS-ESPRIT solution on the choice of the basis of $\mathcal{L}_r$. Numerical experiments show that this dependence takes place if the bases are not orthogonal. However, the TLS-ESPRIT estimate is the same for any orthonormal basis as shown in [10].

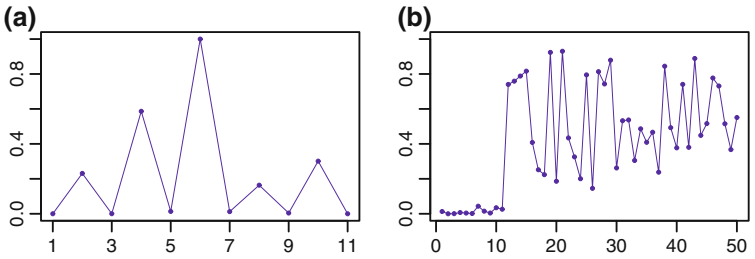

Fig. 3.6 Rank and separability detection by ESTER. **a** 'Eggs'. **b** 'White dwarf'

3.8.2.3 ESPRIT and Rank Estimation

ESPRIT deals with the matrix equation (3.21) that has a solution if the r leading components are exactly separated from the residual (the remaining $L - r$ components). Therefore, some measure of difference between the left-hand and the right-hand sides of the matrix equation can indicate the cut-off points of separability; that is, it can suggest the number of the leading SVD components that are separated from the residual. Therefore, the last cut-off point of separability corresponds to the rank estimation. In [1], the L_2-norm of the difference (we denote it $\rho_2(r)$) is used for estimating the rank of the signal (the ESTER method) provided that there are no separability points within the signal components. An attractive feature of the ESTER-type methods is that they assume only separability of the signal from noise and do not assume parametric forms for the signal and noise.

However, the ESTER-type estimates of rank appear to be unstable for real-world time series. Figure 3.6a shows how the ESTER reflects the points of separability: small values of $\rho_2(r)$ correspond to the cut-off points of separability (compare with Fig. 2.24a). In Fig. 3.6b, where the rank of the signal is estimated to be 11 (see Fig. 2.26), the behavior of $\rho_2(r)$ demonstrates that there are no small values of $\rho_2(r)$ for $r \le 11$.

3.8.3 Overview of Other Subspace-Based Methods

In this subsection, we demonstrate ideas of other subspace-based methods which are different from SSA and ESPRIT-like methods. These methods are applied to time series governed by LRRs and in fact estimate the main (signal) roots of the corresponding characteristic polynomials. The most fundamental subspace-based methods were developed for the cases of a noisy sum of imaginary exponentials (cisoids) and of real sinusoids, for the purpose of the estimating their frequencies, see e.g. [25]. We are mostly interested in the methods that can be applied to any time series of finite rank given in the form (3.13).

We start with a description of general methods in the complex-valued case. Most of these general methods use the correspondence between LRR and vectors from the subspace orthogonal to the signal subspace which was introduced in Sect. 3.2.3.

The first idea is to use the properties of signal and extraneous roots to distinguish from one another. Let us introduce three possible realizations of this idea. As before, $\mathcal{S}$ is the signal subspace, $\mathcal{L}_r$ is its estimate.

Version 1. Consider an LRR that governs the signal (the best choice is the min-norm LRR, see Sect. 3.2.3; however, this is not essential). Then find all the roots μ_m of the characteristic polynomial of this LRR and then find coefficients c_{mj} in (3.13). The coefficients c_{mj} corresponding to the extraneous roots are equal to 0. In the case of a noisy signal, $\widehat{\mu}_m$ are the roots of a polynomial with coefficients taken from a vector that belongs to $\mathcal{L}_r^{\perp}$, and the extraneous roots have small absolute values of the LS estimates $\widehat{c}_{mj}$.

Version 2. Let us consider the forward and backward min-norm predictions. It is known that the corresponding characteristic polynomials have the conjugate extraneous roots and their signal roots are connected by the relation $z' = z^*/\|z\|^2$. Note that the forward prediction given by a vector $A \in \mathcal{S}^{\perp}$ corresponds to the roots of $\langle Z(z), A\rangle = 0$, where $Z(z) = (1, z, \ldots, z^{L-1})^{\mathrm{T}}$ and $\langle \cdot, \cdot \rangle$ is the inner product in the complex Euclidean space. At the same time, the backward prediction given by a vector $B \in \mathcal{S}^{\perp}$ corresponds to the roots of $\langle Z(1/z), B\rangle = 0$. If we consider the roots of the forward and backward min-norm polynomials together, then all the extraneous roots lie inside the unit circle, while one of z' and z is located on or outside the unit circle. This allows us to detect the signal roots. For a noisy signal, A and B are specific vectors taken from $\mathcal{L}_r^{\perp}$: these vectors are the projections onto $\mathcal{L}_r^{\perp}$ of unit vectors $\mathbf{e}_L$ and $\mathbf{e}_1$, correspondingly.

Version 3. Let us take a set of vectors from $\mathcal{S}^{\perp}$. Each vector from $\mathcal{S}^{\perp}$ with nonzero last coordinate generates an LRR. The signal roots of the characteristic polynomials of these LRRs are equal, whereas the extraneous roots are arbitrary. For a noisy signal, the set of vectors is taken from $\mathcal{L}_r^{\perp}$. Then the signal roots correspond to clusters of roots if we consider pooled roots.

A few more methods are developed for estimating frequencies in a noisy sum of undamped sinusoids or complex exponentials. Let for simplicity $s_n = \sum_{k=1}^{r} c_k e^{\mathrm{i}2\pi\omega_k n}$. In this case, the signal roots $e^{\mathrm{i}2\pi\omega_k}$ all have the absolute value 1 and can be parameterized by one parameter (frequency) only. Let $W = W(\omega) = Z(e^{\mathrm{i}2\pi\omega})$. As $W(\omega_k) \in \mathcal{S}$, $\langle W(\omega_k), A\rangle = 0$ for all $A \in \mathcal{S}^{\perp}$. Therefore, if $A \in \mathcal{S}^{\perp}$, then we can consider the square of the cosine of the angle between $W(\omega)$ and A as a measure of their orthogonality. This idea forms the basis for the Min-Norm and MUSIC methods. The modifications of the methods in which the roots are ordered by the absolute value of the deviation of their moduli from the unit circle have names with the prefix 'root-'.

Version 4. Min-Norm. Let $f(\omega) = \cos^2(\widehat{W(\omega), A})$, where A, the projection of $\mathbf{e}_L$ onto $\mathcal{L}_r^{\perp}$, is the vector corresponding to the min-norm forward prediction. The Min-

Norm method consists in searching for the maximums of $1/f(\omega)$; this function can be interpreted as a pseudospectrum.

Version 5. MUSIC. Let $f(\omega) = \cos^2(\widehat{W(\omega), \mathcal{L}_r^{\perp}})$. If we take eigenvectors U_j, $j = r+1, \dots, L$, as a basis of $\mathcal{L}_r^{\perp}$, $\mathbf{U}_{r+1,L} = [U_{r+1} : \dots : U_L]$, then $\mathbf{U}_{r+1,L}\mathbf{U}_{r+1,L}^*$ provides the matrix of projection on $\mathcal{L}_r^{\perp}$ and therefore $f(\omega) = W^*(\mathbf{U}_{r+1,L}\mathbf{U}_{r+1,L}^*)W/\|W\|^2 = \sum_{j=r+1}^{L} f_j(\omega)$, where $f_j(\omega) = \cos^2(\widehat{W(\omega), U_j})$. Thus, the MUSIC method can be considered from the viewpoint of the subspace properties and does not require the computation of roots of the characteristic polynomials. Similar to the Min-Norm method, the MUSIC method consists in searching for the maximums of the pseudospectrum $1/f(\omega)$.

3.8.4 Cadzow Iterations

Cadzow iterations [6] were suggested as a method of signal processing, without any relation to SSA method. However, these two methods are very much related. The Basic SSA with fixed L and fixed grouping $I = \{1, 2, \dots, r\}$ is simply the first iteration of similar Cadzow iterations. This means that Cadzow iterations can be defined as a repeated application of Basic SSA with parameters as above to the series $\widetilde{\mathbb{X}}_I$, see (2.17), obtained by Basic SSA in the previous step; the initial Cadzow iteration is Basic SSA applied to the original series $\mathbb{X}_N$.

The aim of Cadzow iterations is to extract the finite-rank signal $\mathbb{S}_N$ of rank r from an observed noisy signal $\mathbb{X}_N = \mathbb{S}_N + \mathbb{P}_N$. Formally, Cadzow iterations is a method of solving the general HTLS (Hankel matrix low-rank approximation) problem. There is only a partial theoretical proof of convergence of Cadzow iterations but examples demonstrate the convergence. Cadzow iterations, however, do not have to converge to the optimal solution of the HTLS problem (and usually they do not).

Cadzow iterations present an example of the procedure called alternating projections. A short form of M iterations is $\widetilde{\mathbb{S}}_N = \mathsf{T}^{-1}\left(\mathscr{H}\mathsf{P}_r\right)^M \mathsf{T}(\mathbb{X}_N)$. Here T is the one-to-one correspondence between time series and trajectory matrices for the fixed window length L, P_r is the projection of a matrix to the space of $L \times K$ matrices of rank not larger than r, the hankelisation operator $\mathscr{H}$ is also the projection into the space of Hankel matrices in the Frobenious norm. The difficulty in understanding the properties of convergence is caused by the fact that the space of matrices of rank $\le r$ is not convex.

The result of Cadzow iterations is a signal of finite rank $\le r$. However, this does not guarantee that the limiting result is closer to the true signal than SSA result (that is, just one iteration). Among other factors, this depends on how well the true signal can be approximated by the series of rank r and the recommended choice of L: indeed, a usual recommendation in signal processing literature is to choose L which is just slightly larger than r which is unwise from the viewpoint of SSA.

3.9 SSA and Filters

As demonstrated in Sect. 2.3, one of SSA's capabilities is its ability to be a frequency filter. The relation between SSA and filtering was considered in a number of papers, see for example [5, 14]. These results are mostly related to the case where (a) the window length L is small, that is, much less than $N/2$, and (b) Toeplitz SSA is considered and the filter properties are based on the properties of the eigenvectors of Toeplitz matrices (therefore, the time series is assumed to be stationary, see Sect. 2.5.3).

In this section we describe the relation between Basic SSA and filtration in a general form and also consider specific filters generated by Basic SSA.

3.9.1 Linear Filters and Their Characteristics

Let $\mathbf{x} = (\ldots, x_{-1}, \overset{\circ}{x_0}, x_1, x_2, \ldots)$ be an infinite sequence and the symbol 'o' over an element denotes its middle location. Finite series $\mathbb{X}_N = (x_1, \ldots, x_N)$ can be formally presented as a infinite sequence $(\ldots, 0, \ldots, \overset{\circ}{0}, x_1, x_2, \ldots, x_N, 0, \ldots)$. Each linear filter Φ can be expressed as $\big(\Phi(\mathbf{x})\big)_j = \sum\limits_{i=-\infty}^{+\infty} h_i x_{j-i}$. The sequence $\mathbf{h}_\Phi = (\ldots, h_{-1}, \overset{\circ}{h_0}, h_1, \ldots)$ is called *the impulse response*. A filter Φ is called FIR-filter (i.e. with Finite Impulse Response) if $\big(\Phi(\mathbf{x})\big)_j = \sum\limits_{i=-r_1}^{r_2} h_i x_{j-i}$. The filter Φ is called *causal* if $\big(\Phi(\mathbf{x})\big)_j = \sum\limits_{i=0}^{r-1} h_i x_{j-i}$.

The following characteristics are standard for filters: $H_\Phi(z) = \sum_i h_i z^{-i}$ is a *transfer function*, $A_\Phi(\omega) = |H_\Phi(e^{\mathrm{i}2\pi\omega})|$ is a *frequency (amplitude) response* and $\varphi_\Phi(\omega) = \mathrm{Arg} H_\Phi(e^{\mathrm{i}2\pi\omega})$ is a *phase response*. The meaning of the amplitude and phase responses is as follows: for the sequence $\mathbf{x}$ with $(\mathbf{x})_j = \cos(2\pi\omega j)$ we have $\big(\Phi(\mathbf{x})\big)_j = A_\Phi(\omega)\cos(2\pi\omega j + \varphi_\Phi(\omega))$.

An important filter characteristic reflecting its noise reduction capability is the filter *power* $\mathscr{E}\Phi = \|\mathbf{h}\|^2 = \sum_i h_i^2$. The following proposition is analogous to Proposition 3.4.

Proposition 3.8 *Let* $\mathbf{x} = \mathbf{s} + \varepsilon$, *where* $(\varepsilon)_j$ *are i.i.d,* $\mathsf{E}(\varepsilon)_j = 0$, $\mathsf{D}(\varepsilon)_j = \sigma^2$. *Let* $\Phi : \Phi(\mathbf{s}) = \mathbf{s}$ *and denote* $\widetilde{\mathbf{x}} = \Phi(\mathbf{x})$. *Then* $\mathsf{E}(\widetilde{\mathbf{x}})_j = (\mathbf{s})_j$ *and* $\mathsf{D}(\widetilde{\mathbf{x}})_j = \sigma^2 \cdot \mathscr{E}\Phi$.

Also, there is a relation between the filter power and the frequency response. Define $\Delta_a\Phi = \mathrm{meas}\{\omega \in (-0.5, 0.5] \colon A_\Phi(\omega) \geq a\}$. Parseval's identity has the following form for filters:

$$\mathscr{E}\Phi = \sum_j \mathrm{h}_j^2 = \int_{-0.5}^{0.5} A_\Phi(\omega)^2\, d\omega.$$

Therefore,

$$\Delta_a \Phi \le \mathscr{E}\Phi / a^2. \tag{3.24}$$

The inequality (3.24) shows how the support of the frequency response (with threshold a) is related to the filter power.

3.9.2 SSA Reconstruction as a Linear Filter

Let us return to Basic SSA . Let L be the window length and $(\sqrt{\lambda}, U, V)$ be one of the eigentriples generated by the SVD of the trajectory matrix of $\mathbb{X}_N$ (see Sect. 2.1.1 for notation and definitions). Since the reconstruction operation in Basic SSA is the linear operation, it can be written in matrix form.

Let $K = N - L + 1$, $L^* = \min(L, K)$. Define the diagonal $N \times N$ matrix

$$\mathbf{D} = \operatorname{diag}(1, 2, 3, \ldots, L^* - 1, L^*, L^*, \ldots, L^*, L^* - 1, \ldots, 2, 1)$$

and the $K \times N$ matrix

$$\mathbf{W} = \begin{pmatrix} u_1 & u_2 & u_3 & \cdots & u_L & 0 & \cdots & 0 & 0 & 0 \\ 0 & u_1 & u_2 & u_3 & \cdots & u_L & 0 & \cdots & 0 & 0 \\ \vdots & 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & u_1 & u_2 & u_3 & \cdots & u_L & 0 & \vdots \\ 0 & 0 & \cdots & 0 & u_1 & u_2 & u_3 & \cdots & u_L & 0 \\ 0 & 0 & 0 & \cdots & 0 & u_1 & u_2 & u_3 & \cdots & u_L \end{pmatrix}.$$

Proposition 3.9 *The time series component $\widetilde{\mathbb{X}}_N$ reconstructed by the eigentriple $(\sqrt{\lambda}, U, V)$ has the form*

$$\widetilde{\mathbb{X}}_N^{\mathrm{T}} = \mathbf{D}^{-1} \mathbf{W}^{\mathrm{T}} \mathbf{W} \mathbb{X}_N^{\mathrm{T}}.$$

Proof First, note that $\mathbf{W}\mathbb{X}_N^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}} U = \sqrt{\lambda} V \in \mathsf{R}^K$. This yields that the vector $\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbb{X}_N^{\mathrm{T}}$ (of size N) consists of sums along N antidiagonals of $\sqrt{\lambda} U V^{\mathrm{T}}$, the elementary summand of the SVD. Multiplication by $\mathbf{D}^{-1}$ provides the normalization of the sums by the number of summands and therefore by the definition we obtain the elementary reconstructed component. □

Remark 3.11 Let us add the index i to $\widetilde{\mathbb{X}}$ to indicate that it corresponds to the ith eigenvector $U = U_i$. Then, evidently, the reconstructed series $\widetilde{\mathbb{X}}^{(I)}$ by the set of eigentriples $\{(\sqrt{\lambda_i}, U_i, V_i),\ i \in I\}$ is equal to the sum of the reconstructed elementary series $\widetilde{\mathbb{X}}^{(i)}$. Therefore, the matrix form for $\widetilde{\mathbb{X}}^{(I)}$ immediately follows from (3.9).

Proposition 3.9 and Remark 3.11 allow us to describe the reconstruction produced by Basic SSA as an application of a set of linear filters.

Let $L \le K$. Define the linear filters Θ_L, Θ_{L-1}, ..., Θ_1 and Ψ by their impulse characteristics $\mathbf{h}_{\Theta_L}, \ldots, \mathbf{h}_{\Theta_1}$ and $\mathbf{h}_\Psi$:

$$\begin{aligned}
\mathbf{h}_{\Theta_L} &= (\ldots, 0, \overset{\circ}{u_L}, 0, \ldots),\\
\mathbf{h}_{\Theta_{L-1}} &= (\ldots, 0, u_{L-1}, \overset{\circ}{u_L}, 0, \ldots),\\
&\ldots\\
\mathbf{h}_{\Theta_1} &= (\ldots, 0, u_1, \ldots, u_{L-2}, u_{L-1}, \overset{\circ}{u_L}, 0, \ldots);\\
\mathbf{h}_\Psi &= \operatorname{rev} \mathbf{h}_{\Theta_1} = (\ldots, 0, \overset{\circ}{u_L}, u_{L-1}, \ldots, u_1, 0, \ldots).
\end{aligned}$$

Now we can introduce the reconstructing filters Φ_k, $k = 1, \ldots, L$, generated by the vector U:

$$\Phi_k = \Theta_k \circ \Psi/(L - k + 1), \tag{3.25}$$

where '$\circ$' stands for the filter composition, which is equivalent to the convolution '$*$' of the filter impulse characteristics.

Proposition 3.10 *Let $\mathbb{X}_N = (x_1, x_2, \ldots, x_N)$. Then the terms of the elementary reconstructed series $\widetilde{\mathbb{X}}_N$ corresponding to the eigentriple $(\sqrt{\lambda}, U, V)$ have the following representation:*

- $\widetilde{x}_s = (\Phi_{s-K+1}(\mathbb{X}_N))_s \quad$ *for* $K + 1 \le s \le N$;
- $\widetilde{x}_s = (\Phi_1(\mathbb{X}_N))_s \quad$ *for* $L \le s \le K$.

This result is the direct consequence of Proposition 3.9. The set of filters providing the reconstruction of $\widetilde{x}_s$ for $1 \le s < L$ can be built in a similar way.

Let us examine two special filters: Φ_1, which is used for the reconstruction of the middle points of the time series with numbers $L, \ldots, K$, and Φ_L, which is used for the reconstruction of the last point only. The former is called *the MPF* (Middle Point Filter) and the latter is referred as *the LPF* (Last Point Filter). In next two sections we consider them separately.

3.9.3 Middle Point Filter

As above, we assume $L \le K$. According to Proposition 3.10, the MPF Φ_1 acts only at the Lth to $(N-L+1)$th points. This leads to a limited use of the MPF in the case $L \sim N/2$. The explicit formula for the MPF filter $\Phi_1^{(i)}$ corresponding to the eigenvector $U_i = (u_1, \ldots, u_L)^{\mathrm{T}}$ has the following form:

$$\widetilde{x}_s = \sum_{j=-(L-1)}^{L-1} \left(\sum_{k=1}^{L-|j|} u_k u_{k+|j|}/L \right) x_{s-j}, \quad L \le s \le K. \tag{3.26}$$

It is clearly seen that the order of the MPF is equal to $2L-1$. Alternative representation of (3.26) is

$$\widetilde{x}_s = \sum_{j=1}^{L}\sum_{l=1}^{L} u_j u_l x_{s+j-l}/L,\ L \le s \le K. \tag{3.27}$$

Let us enumerate several properties of the MPF $\Phi_1^{(I)}$.

1. The filter $\Phi_1^{(I)}$ is symmetric. Hence the MPF is a zero-phase filter. In particular, the MPF does not change phases of sinusoids.
2. In a particular case of $I = \{i\}$, applying Jensen's inequality, we obtain that the sum of coefficients of the $\Phi_1^{(i)}$ given in (3.27) is not larger than 1:

$$\sum_{j=1}^{L}\sum_{l=1}^{L} u_j u_l/L = \left(\sum_{j=1}^{L} u_j\right)^2 \Big/ L \le \sum_{j=1}^{L} u_j^2 = 1.$$

3. If the matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ is positive, then the leading eigenvector U_1 is positive too (Perron's theorem) and, therefore, the coefficients of the filter $\Phi_1^{(1)}$ are positive. This is true, for example, in the case of positive time series. If the time series is close to a constant (at the timescale of L), then the coordinates of U_1 will be close one to another and the MPT filter $\Phi_1^{(1)}$ will be close to the so-called triangle (Bartlett) filter. This implies, for instance, that the extraction of trend by Sequential SSA with small L (see Sect. 2.5.5) is similar to the application of a weighted moving average procedure with positive nearly triangular weights.
4. Power of the MPF satisfies the following inequalities.

Proposition 3.11 *Let the filter $\Phi_1^{(i)}$ be the MPF generated by eigenvector $U_i = (u_1, \ldots, u_L)^{\mathrm{T}}$. Then $\mathscr{E}\Phi_1^{(i)} \le 1/L$.*

Proof The proof of the proposition results from the following inequality:

$$||\mathbf{h}_\Psi * \operatorname{rev}\mathbf{h}_\Psi|| \le \sum_{j=1}^{L} |u_j| \cdot \|\mathbf{h}_\Psi\| = \sum_{j=1}^{L} |u_j| \cdot \|U\| = \sum_{j=1}^{L} |u_j| \le \sqrt{L}\|U\| = \sqrt{L}.$$

□

Proposition 3.12 *Let $\Phi_1^{(I)}$ be the MPF generated by eigenvectors $\{U_i, i \in I\}$ where $|I| = r$. Then $\mathscr{E}\Phi_1^{(I)} \le r^2/L$.*

Proof By linearity of the grouping operation, $\Phi_1^{(I)} = \sum_{i\in I} \Phi_1^{(i)}$, and therefore, by Proposition 3.11 we have:

$$\mathscr{E}\Phi_1^{(I)} = \left\|\sum_{i\in I} \mathbf{h}_{\Phi_1^{(i)}}\right\|^2 \le \left(\sum_{i\in I} \left\|\mathbf{h}_{\Phi_1^{(i)}}\right\|\right)^2 = \left(\sum_{i\in I} \sqrt{\mathscr{E}\Phi_1^{(i)}}\right)^2 \le r^2/L.$$

□

5. A direct consequence of Proposition 3.12 and inequality (3.24) is the inequality $\Delta_a \Phi_1^{(I)} \le r^2/(a^2 L)$. This means that for any threshold a, the support of filter frequency response tends to 0 as $L \to \infty$. This effect is clearly seen in Fig. 2.22 (Sect. 2.4.3) showing the smoothing effect of Basic SSA.
6. Let us define for $\omega \in (-0.5, 0.5]$

$$g_U(\omega) = \frac{1}{L}\left|\sum_{j=1}^{L} u_j e^{-\mathrm{i}2\pi\omega j}\right|^2. \tag{3.28}$$

The function g_U is closely related to the periodogram Π_u^L introduced in (2.10) of Sect. 2.3.1.1: $g_U(k/L) = L\,\Pi_u^L(k/L)/2$ for $0 < k < N/2$ and $g_U(k/L) = L\,\Pi_u^L(k/L)$ otherwise. It appears that the frequency response of the MPF is almost the same as the periodogram of the vector U.

Proposition 3.13 *Let A_{Φ_1} be the frequency response of the MPF filter Φ_1. Then $g_U(\omega) = A_{\Phi_1}(\omega)$.*

Proof Recall that $\Phi_1 = \Theta_1 \circ \Psi / L$ where $\mathbf{h}_\Psi = (\ldots, 0, \overset{\circ}{u_L}, u_{L-1}, \ldots, u_1, 0, \ldots)$ and $\mathbf{h}_{\Theta_1} = \mathrm{rev}\,\mathbf{h}_\Psi$. Also, from the theory of linear filters [20] we have $A_{\Phi_1 \circ \Psi}(\omega) \equiv A_{\Phi_1}(\omega) A_\Psi(\omega)$. Then

$$A_{\Phi_1}(\omega) = \frac{1}{L}\left|\sum_{j=0}^{L-1} u_{L-j} e^{-\mathrm{i}2\pi\omega j}\right| \cdot \left|\sum_{j=0}^{1-L} u_{L+j} e^{-\mathrm{i}2\pi\omega j}\right| = \frac{1}{L}\left|\sum_{j=1}^{L} u_j e^{-\mathrm{i}2\pi\omega j}\right|^2. \qquad \square$$

7. It follows from Proposition 3.13 that for SSA identification and interpretation of the SVD components, the periodogram analysis of eigenvectors can be very helpful. Also, an automatic identification of components introduced in Sect. 2.4.5 is based on properties of periodograms of eigenvectors and therefore can also be expressed in terms of the frequency response of the MPF.

3.9.4 Last Point Filter and Forecasting

The last-point filter (LPF) is not really a filter as it is used only for the reconstruction of the last point: $\widetilde{x}_N = \sum_{i=0}^{L-1} u_L u_{i+1} x_{N-i}$. The reconstruction by the eigentriples with numbers from the set I has the following form:

$$\widetilde{x}_N^{(I)} = \sum_{k=0}^{L-1} \left(\sum_{i \in I} u_L^{(i)} u_{k+1}^{(i)} \right) x_{N-k}. \tag{3.29}$$

However, it is the only reconstruction filter that is causal. This has two consequences. First, the LPF of a finite-rank series is closely related to the LRR governing and forecasting this time series. Second, the so-called Causal SSA (or last-point SSA) can be constructed by means of the use of the last reconstructed points of the accumulated data. Since in the Causal SSA the LPF is applied many times, the study of properties of LPF is important.

Let the signal $\mathbb{S}_N$ has rank r and is governed by an LRR. Unbiased causal filters of order L and linear recurrence relations of order $L-1$ are closely related. In particular, if the causal filter is given by $s_j = \sum_{k=0}^{L-1} a_{L-k} s_{j-k}$ and $a_L \neq 1$, then this filter generates the LRR of order $L-1$: $s_j = \sum_{k=1}^{L-1} c_{L-k} s_{j-k}$, where $c_k = a_k/(1-a_L)$.

Similar to the minimum-norm LRR, the minimum-power filters can be considered. It appears that the LPF has minimal power among all unbiased filters. This follows from the relation between LRRs and causal filters. Denote by P_r the orthogonal projector on the signal subspace. The LPF has the form $s_N = (\mathrm{P}_r S_K)_L = A^{\mathrm{T}} S_K$, where $A = \mathrm{P}_r \mathbf{e}_L$, while the min-norm LRR is produced by the last-point filter, i.e. $R = \underline{A}/(1-a_L)$.

In the general case, the filter (3.29) can be rewritten as $\widetilde{x}_N = (\mathrm{P}_r X_K)_L$, where X_K is the last L-lagged vector, P_r is the projector on SSA estimate of the signal subspace $\mathrm{span}(U_i, i \in I)$. Formally applying this filter to the whole time series, we obtain the series of length K consisting of the last points of the reconstructed lagged vectors $\widetilde{X}_k$.

Note that if we use other estimates of the signal subspace, then we obtain other versions of the last-point filter.

3.9.5 Causal SSA (Last-Point SSA)

Let $\mathbb{X}_\infty = (x_1, x_2, \ldots)$ be an infinite series, $\mathbb{X}_M = (x_1, \ldots, x_M)$ be its subseries of length M, L be fixed, $\mathcal{L}(M)$ be a subspace of R^L, $\mathrm{P}(M)$ be a projector to $\mathcal{L}(M)$, $A(M) = \mathrm{P}_r(M)\mathbf{e}_L$, $K = K(M) = M - L + 1$.

Introduce the series $\check{\mathbb{X}}_\infty$ as follow. Define $(\check{\mathbb{X}}_\infty)_M = (\mathrm{P}(M) X_{K(M)})_L = A(M)^{\mathrm{T}} X_{K(M)}$, where $X_{K(M)}$ is the last L-lagged vector of $\mathbb{X}_M$. Thus, $(\check{\mathbb{X}}_\infty)_M$ is a linear combination of the last L terms of $\mathbb{X}_M$ with coefficients depending on M; that is, $\check{\mathbb{X}}_\infty$ can be considered as a result of application of a sequence of different causal filters to $\mathbb{X}_\infty$.

If $\mathcal{L}_r(M) = \mathcal{L}(M) = \mathrm{span}(U_1(M), \ldots, U_r(M))$, where $U_1(M), \ldots, U_r(M)$ are the signal eigenvectors produced by SSA with window length L applied to $\mathbb{X}_M$, then this sequence of causal filters is called Causal SSA. In this case, $(\check{\mathbb{X}})_M$ is equal to the last point of SSA reconstruction $\widetilde{\mathbb{X}}_M$, which in turn is equal to the last coordinate of the projection of the last lagged vector of $\mathbb{X}_M$ to $\mathcal{L}_r(M)$.

Given that $\mathcal{L}_r(M)$ is used as an estimate of the signal subspace, M should be large enough to provide a good estimate. Therefore, we need to introduce a starting point M_0 (with $M_0 > L$) in the Causal SSA and consider $M \geq M_0$ only.

Since the result of application of the Causal SSA is a sequence $\breve{\mathbb{X}}_\infty$ built from SSA reconstructions of the last points of the subseries, the Causal SSA can be called Last-point SSA.

Remark 3.12 Let us fix N, L and consider $\mathbb{X}_N$, the corresponding $U_1, \dots, U_r$ and $\mathfrak{L}(M) = \mathrm{span}(U_1, \dots, U_r)$ for any M. Then the series $\breve{\mathbb{X}}_N$ is the result of application of the last-point filter to $\mathbb{X}_N$. Assuming that the estimates of the signal subspace on the base of $\mathbb{X}_M$ are stable for large enough M, we can conclude that the result of the Causal SSA will be close to the result of the last-point filter (LPF).

Note also that in the considered particular case, $\breve{\mathbb{X}}_N$ (more precisely, its last K points; the first $L-1$ points are not defined or can be set to zero) coincides with the last row of the reconstructed matrix $\widehat{\mathbf{X}}$ of the series $\mathbb{X}_N$. That is, $\breve{\mathbb{X}}_N$ is similar to the result of reconstruction before the diagonal averaging is made.

Causality yields the following relation: under the transition from $\mathbb{X}_M$ to $\mathbb{X}_{M+1}$, the first M points of the $\breve{\mathbb{X}}_{M+1}$ coincide with $\breve{\mathbb{X}}_M$. This is generally not true if we consider the reconstructions $\widetilde{\mathbb{X}}_M$ for $\mathbb{X}_M$ obtained by the conventional SSA. The effect $(\widetilde{\mathbb{X}}_M)_j \neq (\widetilde{\mathbb{X}}_{M+1})_j$, $j \le M$, is called 'redrawing'. For real-life time series we usually have redrawing for all j and the amount of redrawing depends on j. Redrawing of only a few last points is usually of interest. Moreover, redrawing of local extremes of the series is more practically important than redrawing of regular series points. Small values of redrawing indicate stability of SSA decompositions and hence stability of time series structure. The amount of redrawing can be assessed by visual examination or measured using the variance of the redrawings at each time moment of interest. These variances can be averaged if needed.

Generally, the reconstruction has no delay (at least, the middle-point filter has zero phase shift). On the other hand, delays in the Causal SSA are very likely. In a sense, a redrawing in SSA is transferred to a delay in the Causal SSA. If $\mathfrak{L}_r$ corresponds to the exactly separated component of the time series $\mathbb{X}$, then SSA has no redrawing and the Causal SSA has no delay. In conditions of good approximate separability the redrawing is almost absent and the delay is small.

Example

Let us demonstrate the described effects on the 'S&P500' example introduced in Sect. 2.5.1. Figure 3.7 shows the result of the Causal SSA with window length $L = 30$, $\mathfrak{L}_r(M) = \mathrm{span}(U_1(M), U_2(M))$ and $M_0 = 200$. The delay is clearly seen. If we consider non-causal (Basic) SSA reconstructions of cumulative subseries, then the redrawing takes place (Fig. 3.8). This redrawing increases in the points of local maximums and minimums.

Finally, let us note that if the direction of change of the time series is of primary concern (rather than the values themselves), then, instead of taking differences of the Causal SSA series, it may be worthwhile considering the time series which consists of differences of the last two points of reconstructions. The result should be expected

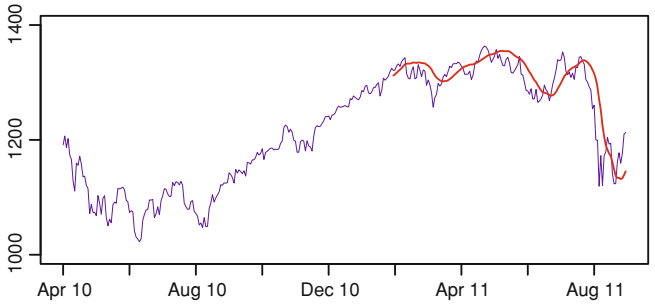

Fig. 3.7 S&P500: causal SSA

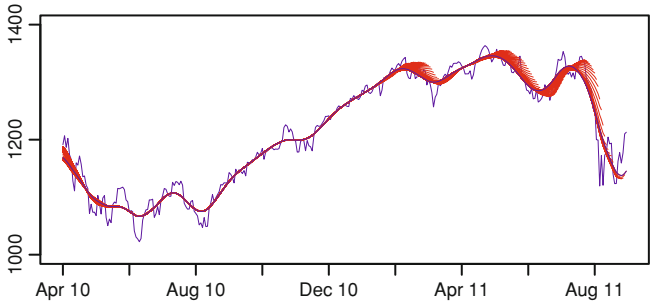

Fig. 3.8 S&P500: non-causal SSA with redrawing

to be more stable since the reconstruction of the next to last point has better accuracy than that of the last point.

References

1. Badeau R, David B, Richard G (2004) Selecting the modeling order for the ESPRIT high resolution method: an alternative approach. In: Proceedings of the IEEE ICASSP, vol 2, pp 1025–1028
2. Badeau R, Richard G, David B (2008) Performance of ESPRIT for estimating mixtures of complex exponentials modulated by polynomials. IEEE Trans Signal Process 56(2):492–504
3. Barkhuijsen H, de Beer R, van Ormondt D (1987) Improved algorithm for noniterative time-domain model fitting to exponentially damped magnetic resonance signals. J Magn Reson 73:553–557
4. Beckers J, Rixen M (2003) EOF calculations and data filling from incomplete oceanographic data sets. Atmos Ocean Technol 20:1839–1856
5. Bozzo E, Carniel R, Fasino D (2010) Relationship between singular spectrum analysis and Fourier analysis: theory and application to the monitoring of volcanic activity. Comput Math Appl 60(3):812–820
6. Cadzow JA (1988) Signal enhancement: a composite property mapping algorithm. IEEE Trans Acoust 36(1):49–62

7. de Groen P (1996) An introduction to total least squares. Nieuw Archief voor Wiskunde 14:237–253
8. Efron B, Tibshirani R (1986) Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. Stat Sci 1(1):54–75
9. Gel'fond A (1971) Calculus of finite differences. Translated from the Russian. International monographs on advanced mathematics and physics. Hindustan Publishing Corp, Delhi
10. Golyandina N (2010) On the choice of parameters in singular spectrum analysis and related subspace-based methods. Stat Interface 3(3):259–279
11. Golyandina N, Osipov E (2007) The "Caterpillar"-SSA method for analysis of time series with missing values. J Stat Plan Inference 137(8):2642–2653
12. Golyandina N, Nekrutkin V, Zhigljavsky A (2001) Analysis of time series structure: SSA and related techniques. Chapman&Hall/CRC, Boca Raton
13. Hall MJ (1998) Combinatorial theory. Wiley, New York
14. Harris T, Yan H (2010) Filtering and frequency interpretations of singular spectrum analysis. Physica D 239:1958–1967
15. Kondrashov D, Ghil M (2006) Spatio-temporal filling of missing points in geophysical data sets. Nonlinear Process Geophys 13(2):151–159
16. Kumaresan R, Tufts DW (1980) Data-adaptive principal component signal processing. In: Proceedings of the IEEE conference on decision and control. Albuquerque, pp 949–954
17. Kumaresan R, Tufts DW (1983) Estimating the angles of arrival of multiple plane waves. IEEE Trans Aerosp Electron Syst AES-19(1):134–139
18. Kung SY, Arun KS, Rao DVB (1983) State-space and singular-value decomposition-based approximation methods for the harmonic retrieval problem. J Opt Soc Am 73(12):1799–1811
19. Nekrutkin V (2010) Perturbation expansions of signal subspaces for long signals. Stat Interface 3:297–319
20. Oppenheim AV, Schafer RW (1975) Digital signal processing. Prentice-Hall, Upper Saddle River
21. Pakula L (1987) Asymptotic zero distribution of orthogonal polynomials in sinusoidal frequency estimation. IEEE Trans Inf Theor 33(4):569–576
22. Pepelyshev A, Zhigljavsky A (2010) Assessing the stability of long-horizon SSA forecasting. Stat Interface 3:321–327
23. Roy R, Kailath T (1989) ESPRIT: estimation of signal parameters via rotational invariance techniques. IEEE Trans Acoust 37:984–995
24. Schoellhamer D (2001) Singular spectrum analysis for time series with missing data. Geophys Res Lett 28(16):3187–3190
25. Stoica P, Moses R (1997) Introduction to spectral analysis. Prentice Hall, Englewood Cliffs
26. Tufts DW, Kumaresan R (1982) Estimation of frequencies of multiple sinusoids: making linear prediction perform like maximum likelihood. Proc IEEE 70(9):975–989
27. Usevich K (2010) On signal and extraneous roots in singular spectrum analysis. Stat Interface 3(3):281–295
28. Van Huffel S, Chen H, Decanniere C, van Hecke P (1994) Algorithm for time-domain NMR data fitting based on total least squares. J Magn Reson Ser A 110:228–237