J Glob Optim
DOI 10.1007/s10898-013-0117-3

A study of singular spectrum analysis with global optimization techniques

Moody T. Chu · Matthew M. Lin · Liqi Wang

Received: 17 November 2012 / Accepted: 4 October 2013

Abstract Singular spectrum analysis has recently become an attractive tool in a broad range of applications. Its main mechanism of alternating between rank reduction and Hankel projection to produce an approximation to a particular component of the original time series, however, deserves further mathematical justification. One paramount question to ask is how good an approximation that such a straightforward apparatus can provide when comparing to the absolute optimal solution. This paper reexamines this issue by exploiting a natural parametrization of a general Hankel matrix via its Vandermonde factorization. Such a formulation makes it possible to recast the notion of singular spectrum analysis as a semi-linear least squares problem over a compact feasible set, whence global optimization techniques can be employed to find the absolute best approximation. This framework might not be immediately suitable for practical application because global optimization is expectedly more expensive, but it does provide a theoretical baseline for comparison. As such, our empirical results indicate that the simpler SSA algorithm usually is amazingly sufficient as a handy tool for constructing exploratory model. The more complicated global methods could be used as an alternative of rigorous affirmative procedure for verifying or assessing the quality of approximation.

Moody T. Chu's research was supported in part by the National Science Foundation under grant DMS-1014666.

Matthew M. Lin's research was supported in part by the National Science Council of Taiwan under grant NSC 101-2115-M-194-007-MY3.

M. T. Chu
Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, USA
e-mail: chu@math.ncsu.edu

M. M. Lin (✉)
Department of Mathematics, National Chung Cheng University, Min-Hsiung, Chia-Yi 621, Taiwan
e-mail: mhlin@ccu.edu.tw

L. Wang
Department of Mathematics, Dalian University of Technology, Dalian, China
e-mail: lizzy@mail.dlut.edu.cn

Published online: 13 October 2013

Keywords Singular spectrum analysis · Time series · Hankel operator · Semi-infinite matrix · Low rank approximation · Vandermonde factorization · Global optimization

1 Introduction

Singular spectrum analysis (SSA) has been attracting considerable interest in recent years as a useful technique for extracting critical information from general time series. This novel method, based on principles of multivariate statistics, accounts for dynamically improving the covariance structure of the underlying time series [29]. But what makes its wide applicability is the non-parametric and model-free nature that enables practitioners to deploy it without a prior knowledge of any underlying structure [20,58]. Its utilization is more as a handy exploratory model-building tool than an affirmative procedure. The development of the SSA is usually attributed to Broomhead and King [8] whose original intention was to identify qualitative information in the analysis of dynamical systems. Actually, similar ideas also came to surface independently at about the same time in other disciplines, such as that in [13] for the dimensional analysis of weather and climate attractors. Over years, the methodologies and applications of the SSA have been refined and broadened significantly in the literature. A quick Google search for the term "singular spectrum analysis" shows up a wide range of applications. Far from being complete, we mention a few representative applications including smoothing, filtration [57], noise reduction [55], trend and modulated harmonics extraction [10,39], change-point detection [41], signal parameter estimation, climatic and meteorological forecasting [2,16,54], economic and financial applications [24,28,29,46], causality discovery [46], phase reconstruction [4,17,21,53]. For an elementary introduction to this subject we suggest the two monographs [11,19]. For a synopsis of more recent advances in this field, we suggest the brief treatise [20]. The goal of this paper is to propose a mathematical framework allowing us to assess the effectiveness of the SSA from a theoretical point of view.

So that this paper is self-contained, we briefly review the basic SSA as follows. There are quite a few variations of the basic SSA nowadays, including the multivariate SSA [21,25,46] and the minimum-variance based SSA [19,23], but we will focus on the basic scheme to convey the idea. Given a finite time series $\mathbf{z} = (z_0, \ldots, z_n)$ of length $n+1$,[1] the basic SSA consists of four basic parts [19]:

1.1 Embedding

Choose an integer $1 \leq p < n$, referred to as the *embedding dimension* or *window length*, and let $q = n - p + 2$. Define *p-lagged vectors* $\mathbf{z}_0, \ldots, \mathbf{z}_{q-1}$ by $\mathbf{z}_i := [z_i, \ldots, z_{i+p-1}]^\top \in \mathbb{C}^p$ and the associated *trajectory matrix* Z by

$$Z := [\mathbf{z}_0, \ldots, \mathbf{z}_{q-1}] = \begin{bmatrix} z_0 & z_1 & \cdots & z_{q-1} \\ z_1 & z_2 & \cdots & z_q \\ \vdots & \vdots & \ddots & \vdots \\ z_{p-1} & z_p & \cdots & z_n \end{bmatrix}. \tag{1}$$

Note that Z is a $p \times q$ Hankel matrix. This process of embedding $\mathbf{z}$ into Z, fundamental in time series analysis, creates a handle for manipulating rank reduction.

[1] For the sake of characterizing the Vandemonde parameters more easily for the Hankel operator, which will be discussed subsequently, we begin the index with 0.

The choice of p affects the quality of the approximation. See discussions in [11,19]. A more recent work [26,27] shows that the optimal value of p is related to the separability between signal and noise components in the original series. It is recommended both theoretically and empirically that the threshold should be somewhere between $\frac{n}{4}$ and $\frac{n}{2}$.

1.2 Rank reduction

Suppose that

$$Z = U\Sigma V^* = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^* \tag{2}$$

is the singular value decomposition (SVD) of the trajectory matrix Z, where nonzero singular values are arranged in descending order $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$ and $r = \operatorname{rank}(Z)$. A fundamental mechanism used for rank reduction is the truncated singular value decomposition. A truncated SVD of Z is a partial sum of the form

$$\widetilde{Z} := \sum_{i=1}^{d} \sigma_i \mathbf{u}_i \mathbf{v}_i^*, \tag{3}$$

where $d \leq r$ is a preselected positive integer serving as the target rank for the analysis.

There is a statistical meaning of the truncated singular value decomposition. With regard to the *lag-covariance* matrix $Z^\top Z$ when columns of Z have been centered, it can be argued that $\widetilde{Z}$ represents the best, unbiased, linear, minimum-variance estimate of Z among all possible d-dimensional sample subspaces in $\mathbb{C}^p$. Because of this, if d is properly chosen, then those components of Z corresponding to "small" singular values are considered as less significant and often are disregarded.

1.3 Group selection

In order to identify possibly disparate physical behaviors of the time series, the above process can further be refined. Without being specific, we simply mention that different rank-one matrices from the SVD of Z can be chosen and collected together, and be processed separately, with the hope that properly partitioned groups would reflect different dynamical components of the original time series. See [19,40] for some practical concerns and applications.

1.4 Signal recovery

It is almost always true that the truncated SVD of a structured matrix cannot preserve the original structure [9,45]. As such, $\widetilde{Z}$ generally is not a Hankel matrix. The nearest Hankel matrix to $\widetilde{Z}$ can be obtained by *diagonally averaging* entries of $\widetilde{Z}$ along the anti-diagonals. In doing so, however, the rank of the resulting Hankel matrix approximation is no longer d. So the above four steps have to be repeated with the hope that a limit point which is a Hankel matrix of rank d is finally determined. This procedure is similar in spirit to the general lift-and-project algorithm developed for structured low rank approximation [9].

In short, the basic idea of the SSA is to decompose the original time series into the sum of a number of interpretable components, each of which should be easily identified as part of a certain modulated signals or, in reciprocity, the random noises. The workings of the

SSA are by repeatedly sifting the time series through the truncated SVD of the embedded trajectory matrix with the hope that a cleaner, more representative time series showing critical traits, such as low-frequency trends, high-frequency variability, cyclic components, or other structures of the original series, can be extracted.

The notion of the SSA is attractive because of its simplicity in implementation and its broadness in applications. The main mechanism whereby the scheme alternates between the rank reduction and the Hankel projection, however, deserves further understanding. We can think of at least three reasons for this investigation. Firstly, since the method is within the general frame of lift-and-project, at best the linear rate of convergence is expected [9]. How would the strength of noise or the inherent intricacy of the original signal affects the convergence? Secondly, the iteration may not achieve a Hankel approximation of rank d because the alternating projection gets stagnant at a local minimizer of the distance between the manifold of matrices of rank d and the subspace of Hankel matrices. What does the limit point of the SSA really represent? Thirdly, there might be multiple local minimizers. How does this local solution affect the interpretation of physical properties being investigated? The purpose of this paper is to study, by means of global optimization techniques on parameters of the Vandermonde factorization, what the absolutely best and low rank Hankel approximation can be achieved for a general time series. In return, we are interested in using this information as a reference point for assessing the quality of the SSA solutions.

This paper is organized as follows. In Sect. 2 we explain the cause and effect of the notion of the "low rank Hankel structure". The SSA usually applies only to finite series which actually is no more than a segment of a long term behavior. So as to take into account all possible intrinsic characteristics, we describe a natural linkage between the low rank Hankel structure and a sinusoidal signal which, in turns, leads to an infinite series whose trajectory matrix is necessarily semi-infinite. We raise the question of how much information is contained in a segment or submatrix of the infinite case. In Sect. 3 we derive the fact that any finite rank Hankel operator, of finite dimension or not, always enjoys a Vandermonde factorization. Related to Sect. 2, this argument also shows that, in a broad sense, all infinite series with finite rank trajectory matrix can be cast as a sinusoidal signal. Equipped with the Vandermonde parameters, we formulate the SSA as a least squares minimization problem over a compact set in Sect. 4.

2 Sinusoidal signal

To motivate the connection of a time series to its trajectory matrix, we first consider a noise-free time-domain signal comprising d components of exponentially decaying sinusoids

$$\mathfrak{s}(t) = \sum_{\ell=1}^{d} a_\ell e^{-\alpha_\ell t} e^{\imath(2\pi \nu_\ell t + \phi_\ell)}, \tag{4}$$

where a_ℓ, α_ℓ, ν_ℓ, and ϕ_ℓ are real numbers denoting the magnitude, the decay rate, the frequency, and the phase angle, respectively. Starting with $t_0 = 0$ and sampling this signal at uniformly spaced nodes $t_0, t_1, \ldots$ with fixed interval length Δt (so $1/\Delta t$ is the so called sampling rate), we obtain an infinite sequence

$$\mathfrak{s}(t_k) = \sum_{\ell=1}^{d} a_\ell e^{-\alpha_\ell k \Delta t} e^{\imath(2\pi \nu_\ell k \Delta t + \phi_\ell)} = \sum_{\ell=1}^{d} a_\ell e^{\imath\phi_\ell} \left(e^{(-\alpha_\ell + \imath 2\pi \nu_\ell)\Delta t}\right)^k.$$

For simplicity, denote $s_k := \mathfrak{s}(t_k)$ and define

$$\beta_\ell := a_\ell e^{\imath\phi_\ell}, \tag{5}$$

$$\lambda_\ell := e^{(-\alpha_\ell + \imath 2\pi\nu_\ell)\Delta t}. \tag{6}$$

Then the time series $\{s_0, s_1, \ldots\}$ enjoys the relationship

$$s_k = \sum_{\ell=1}^{d} \beta_\ell \lambda_\ell^k. \tag{7}$$

We can rewrite the corresponding trajectory matrix S of the series $\{s_0, s_1, \ldots\}$ as

$$\begin{aligned} S &= \sum_{\ell=1}^{d} \beta_\ell \begin{bmatrix} \lambda_\ell^0 & \lambda_\ell^1 & \lambda_\ell^2 & \cdots \\ \lambda_\ell^1 & \lambda_\ell^2 & & \\ \lambda_\ell^2 & & & \\ \vdots & & & \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1^0 & \lambda_2^0 & \cdots & \lambda_d^0 \\ \lambda_1^1 & \lambda_2^1 & & \lambda_d^1 \\ \lambda_1^2 & & & \\ \vdots & & & \vdots \end{bmatrix} \begin{bmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & & \\ \vdots & & \ddots & \\ 0 & & & \beta_d \end{bmatrix} \begin{bmatrix} \lambda_1^0 & \lambda_1^1 & \lambda_1^2 & \cdots \\ \lambda_2^0 & \lambda_2^1 & & \\ \vdots & & & \\ \lambda_d^0 & \lambda_d^1 & & \cdots \end{bmatrix}, \end{aligned} \tag{8}$$

which easily shows that the matrix S is of rank d.

For strictly decaying signals, we should expect that $\alpha_\ell > 0$ and, hence, $|\lambda_\ell| < 1$, implying that S is a bounded operator. The decomposition in the last equation of (8) is known as the *Vandermonde factorization* of S. Note that the trajectory matrix S is naturally a semi-infinite matrix. If we take only finitely many samples, resulting a finite time series as is typically assumed in the application of the SSA, then it means only a leading principal submatrix of size $p \times q$ extracted from S. In this case, the corresponding finite trajectory matrix still enjoys a Vandermonde factorization after appropriately truncating that of (8). In other words, the formation of the basic SSA is a special case of our general theory. Since the set of infinitely many samples $\{s_0, s_1, \ldots\}$ is supposedly carrying more or even complete information, a classical question that immediately arises is how much information of the original signal can be retrieved from a finite number of samples. Keys to our global optimization approach, as will be explained below, are the variables $\beta_\ell, \lambda_\ell,\ \ell = 1, \ldots, d$, which serve as a parameterization of *any* Hankel matrices of rank d.

The converse of the above characterization is worth noting. Given a Vandermonde factorization in the form of (8), then by (5) the polar form of the complex number β_ℓ determines the magnitude a_ℓ and phase angle ϕ_ℓ. In the meantime, by (6) the polar form of the quantity λ_ℓ determines the decay rate α_ℓ and the frequency ν_ℓ up to a scaling by Δt. The dependence on the scaling Δt is due to the fact that whenever a measurement of decay rate and frequency is referred to, the meaning of a unit time ought to be defined first. Once Δt is specified, then a composite signal is completely determined from a given Vandermonde decomposition. In other words, we can go back and forth interchangeably between a bounded low rank Hankel matrix and a sinusoidal signal through the relationship (7).

In practice, the true signal $\mathfrak{s}(t)$ is unknown and often the observed data $\{z_k\}$ are contaminated. Since the samples are taken independently, we may assume the popular AWGN channel model,

$$z_k := s_k + \epsilon_k, \tag{9}$$

i.e., white Gaussian noise ϵ_k with specified signal-to-noise ratio (SNR) is added linearly per sample. Through embedding, the corresponding trajectory matrix Z can be forced to maintain the Hankel structure. But the property of rank d is generally lost. Indeed, Z could easily be of "full" rank. In the SSA, the idea is to approximate Z by a Hankel approximation of rank d, whence the signal information of $\mathfrak{s}(t)$ is partially retrieved. As we have pointed out earlier, the SSA does not even make use of the semi-infinite matrix Z in its entirety. Only a segment of finite length n is fed into the procedure. To what extent this process of alternating projection can recover the essential information is the question we want to investigate in this study.

3 Vandermonde parameterization

In the preceding section, we have used a sinusoidal signal to introduce the Vandermonde factorization of a trajectory matrix. In this section, we want to bring forth the fact that any Hankel operator, finite dimensional or not, always enjoys such a factorization so long as it is of finite rank. So suitable for the SSA, our theory can be applied regardless of wherever the time series might arise. The Vandermonde factorization offers an effective characterization of a Hankel operator in $2d$ parameters, provided the operator is of rank d. Later on, we shall make use of these parameters to set up a global optimization framework.

We mention in passing that the importance of Hankel operators goes beyond its involvement in the SSA. They appear frequently in many other seemingly disparate areas of classical mathematics [1,33,47]. Their applications cover a wide range of disciplines outside mathematics. See, for example, [14,18] and the many references in [48]. It is thus of practical significance to characterize a general low rank Hankel operator in the least possible parameters. To derive Vandermonde parametrization which will be used by global optimization to establish basis of comparison for assessing the effectiveness of the SSA, we now recall several classical results in the literature.

We begin with the characterization of a *bounded* Hankel operators over the space ℓ^2 of square summable (semi-infinite) sequences [6,42,50]. The following theorem states that whether the time sequence $\{h_0, h_1, \ldots\}$ determines a bounded operator H on ℓ^2 is equivalent to whether the sequence itself represents the Fourier coefficients of an essentially bounded function over the unit disk [47, Theorem 2.1].

Theorem 1 *The Hankel matrix H represents a bounded operator over ℓ^2 if and only if there exists a function $\psi \in L^\infty$ on the unit circle such that its*

$$h_m = \frac{1}{2\pi} \int_0^{2\pi} \psi(\theta) e^{-im\theta} d\theta, \quad m = 0, 1, \ldots \tag{10}$$

In this case, the operator norm of H is bounded above by $\|\psi\|_\infty$.

We have learned from the preceding section that the the number of components of a sinusoidal signal is the same as the rank of the corresponding trajectory matrix. What can be said about the converse, which is needed for the rank reduction step in the SSA? Given a time series $\{h_0, h_1, ...\}$, the associated generating function is defined by the formal power series

$$G(z;\{h_n\}) := \sum_{n=0}^{\infty} h_n z^n. \tag{11}$$

The following criterion is a classical result by Kronecker [47, Theorem 4.1].

Theorem 2 *The Hankel matrix H has finite rank if and only if the power series (11) determines a rational function. In this case,*

$$\operatorname{rank}(H) = \deg(zG(z;\{h_n\})), \tag{12}$$

where the degree of a rational function is the maximum of the degrees of its minimal constituent polynomials.

Finding the limit of the generating function $G(z;\{h_n\})$ is not always an easy task. An equivalent but more direct observation of finite rank is through a linear recursive relationship (LRR) among elements of H [15, Chapter XV, Theorem 7].

Theorem 3 *The Hankel matrix H is of finite rank d if and only if there exist constants $\gamma_0, \ldots, \gamma_{d-1}$ such that*

$$h_i = \gamma_{d-1}h_{i-1} + \gamma_{d-2}h_{i-2} + \ldots + \gamma_o h_{i-d}, \quad i = d, d+1, \ldots \tag{13}$$

and d is the least integer having this property.

While Theorem 3 is long known as a necessary and sufficient condition concerning the finite rank of a semi-infinite Hankel matrix, it is interesting to note that the LRR (13) has been redeveloped by practitioners of the SSA from the finite-dimensional point of view as a forecasting scheme [19,24,29]. It is worth noting the interchange between finite and infinite cases as follows.

Assume the scenario that the semi-infinite Hankel matrix H in Theorem 3 is already known to be of rank d. Then it can be argued that the $d \times d$ leading principal submatrix $\widehat{H}$, i.e., the trajectory matrix of the finite series $\{h_0, \ldots, h_{2d-2}\}$, is necessarily nonsingular. The finite difference equation (13) can be written in the form of a linear system

$$\begin{bmatrix} h_0 & h_1 & h_2 & \ldots & h_{d-1} \\ h_1 & h_2 & h_3 & & h_d \\ h_2 & h_3 & & & h_{d+1} \\ \vdots & & & & \\ h_{d-1} & h_d & & & h_{2d-2} \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{d-1} \end{bmatrix} = \begin{bmatrix} h_d \\ h_{d+1} \\ \vdots \\ h_{2d-1} \end{bmatrix}, \tag{14}$$

known as Yule-Walker equations in signal processing, for the coefficients $\gamma_0, \gamma_1, \ldots, \gamma_{d-1}$. We thus see that the parameters $\gamma_0, \ldots, \gamma_{d-1}$ in the LRR and consequently the entire semi-infinite time series are uniquely determined by $\widehat{H}$ and one extra element h_{2d-1}. In other words, to uniquely determine a Hankel matrix of rank d, we need to know exactly $2d$ many elements $\{h_0, h_1, \ldots, h_{2d-1}\}$. Obviously, linear systems similar to (14) can also be formed for the same set of parameters $\gamma_0, \gamma_1, \ldots, \gamma_{d-1}$ by taking up any segment of $2d$ successive elements in the infinite time series, so long as the corresponding trajectory matrix is nonsingular.

Once the parameters $\gamma_0, \ldots, \gamma_{d-1}$ are in hand, we can take one step further. That is, we can express the entry of H at any location in term of roots of the characteristic polynomial

$$p(\lambda) := \lambda^d - \gamma_{d-1}\lambda^{d-1} - \ldots - \gamma_1\lambda - \gamma_0, \tag{15}$$

without the need of the recursive reference to other entries. This formulation leads to a natural parameterization of H which will be used to reassess the rationales behind the SSA.

To fix the ideas, let λ_ℓ, $\ell = 1, \ldots, r$, denote the distinct roots of $p(\lambda)$ in (15) and each of which has multiplicity ρ_ℓ. So $\sum_{\ell=1}^{r} \rho_\ell = d$. A general solution to the difference equation (13) can be formulated as follows. Let

$$\mathscr{V}_{p(\lambda)} := [\mathscr{V}^{(1)}, \ldots, \mathscr{V}^{(r)}] \in \mathbb{C}^{d \times d},$$

be a block matrix whose block $\mathscr{V}^{(\ell)} = [v_{ij}^{(\ell)}]$, $\ell = 1, \ldots, r$, is of size $d \times \rho_\ell$, where

$$v_{ij}^{(\ell)} := c_{ij} \lambda_\ell^{i-j}, \quad i = 0, 1, \ldots, d-1, \quad j = 0, 1, \ldots, \rho_\ell - 1, \tag{16}$$

and $c_{ij} = 0$ if $i < j$; and $c_{ij} = \frac{i!}{(i-j)!\, j!}$ otherwise. A typical block looks like

$$\mathscr{V}^{(\ell)} := \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ \lambda_\ell & 1 & 0 & \ldots & \\ \lambda_\ell^2 & 2\lambda_\ell & 1 & \ldots & \\ \lambda_\ell^3 & 3\lambda_\ell^2 & 3\lambda_\ell & & \\ \vdots & & & \ddots & \\ \lambda_\ell^{\rho_\ell - 1} & & & & 1 \\ \lambda_\ell^{\rho_\ell} & & & & \rho_\ell \lambda_\ell \\ \vdots & \vdots & & & \vdots \\ \lambda_\ell^{d-1} & (d-1)\lambda_\ell^{d-2} & \frac{(d-1)(d-2)}{2} \lambda_\ell^{d-3} & \ldots & \binom{d-1}{\rho_\ell - 1} \lambda_\ell^{d-\rho_\ell} \end{bmatrix} \in \mathbb{C}^{d \times \rho_\ell}.$$

For convenience, note that in the above we start counting indices of the entries from (0, 0). In order to enforce the initial values $h_0, \ldots, h_{d-1}$ on the general solution of (13), we introduce the linear system

$$\mathscr{V}_{p(\lambda)} \begin{bmatrix} \beta_1^{(0)} \\ \vdots \\ \beta_1^{(\rho_1 - 1)} \\ \vdots \\ \beta_r^{(0)} \\ \vdots \\ \beta_r^{(\rho_r - 1)} \end{bmatrix} = \begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ \vdots \\ h_{d-1} \end{bmatrix}, \tag{17}$$

known as the confluent Vandermonde system [34], for the undetermined coefficients $\beta_1^{(0)}, \ldots, \beta_1^{(\rho_1 - 1)}, \ldots, \beta_r^{(0)}, \ldots, \beta_r^{(\rho_r - 1)}$. Trivially, the equations in (17) are equivalent to

$$h_i = \sum_{\ell=1}^{r} \sum_{j=0}^{\rho_\ell - 1} \beta_\ell^{(j)} c_{ij} \lambda_\ell^{i-j}, \quad i = 0, 1, \ldots, d-1.$$

It can further be verified recursively that

$$\begin{aligned} h_i &= \sum_{s=0}^{d-1} \gamma_s h_{i-d+s} = \sum_{s=0}^{d-1} \gamma_s \left(\sum_{\ell=1}^{r} \sum_{j=0}^{\rho_\ell - 1} \beta_\ell^{(j)} c_{i-d+s,j} \lambda_\ell^{i-d+s-j} \right) \\ &= \sum_{\ell=1}^{r} \sum_{j=0}^{\rho_\ell - 1} \beta_\ell^{(j)} \left(\sum_{s=0}^{d-1} \left(c_{i-d+s,j} \gamma_s \lambda_\ell^{i-d+s-j} \right) \right) \\ &= \sum_{\ell=1}^{r} \sum_{j=0}^{\rho_\ell - 1} \beta_\ell^{(j)} c_{ij} \lambda_\ell^{i-j}, \quad i = d, d+1, \ldots \end{aligned} \tag{18}$$

For any given time series $\{h_0, h_1, \ldots\}$, we have thus shown that if the corresponding trajectory matrix H is of rank d, then each h_i may be represented by the parameters $\beta_\ell^{(j)}$ and λ_ℓ, referred to as the *Vandermonde parameters*, according to the formula (18).

In the literature, the so called *Vandermonde factorization* for a semi-infinite Hankel matrix H of rank d is simply a rewriting of the same parametric representation (18). More specifically, let V_∞ denote the *confluent Vandermonde matrix* of size $\infty \times d$ obtained by extending the matrix $\mathscr{V}_{p(\lambda)}$ defined in (16) downward to infinite length. We can rearrange the expression (18) and establish the following identity [5,12].

Theorem 4 *Suppose H is a semi-infinite Hankel matrix of rank d. Then there exists a $d \times d$ block diagonal matrix D_∞ whose ℓ-th block is of size $\rho_\ell \times \rho_\ell$ and is Hankel and upper anti-triangular such that*

$$H = V_\infty D_\infty V_\infty^\top. \tag{19}$$

Assume for simplicity the generic case that all roots of $p(\lambda)$ are distinct. Then the solution to the LRR (13) can be written as

$$\begin{aligned} h_i &= \sum_{s=0}^{d-1} \gamma_s h_{i-d+s} = \sum_{s=0}^{d-1} \gamma_s \left(\sum_{\ell=1}^{d} \beta_\ell^{(0)} \lambda_\ell^{i-d+s} \right) = \sum_{\ell=1}^{d} \beta_\ell^{(0)} \left(\sum_{s=0}^{d-1} \gamma_s \lambda_\ell^{i-d+s} \right) \\ &= \sum_{\ell=1}^{d} \beta_\ell^{(0)} \lambda_\ell^{i-d} \left(\sum_{s=0}^{d-1} \gamma_s \lambda_\ell^{s} \right) = \sum_{\ell=1}^{d} \beta_\ell^{(0)} \lambda_\ell^{i}, \quad i = d, d+1, \ldots, \end{aligned} \tag{20}$$

which, in turn, establishes the very same Vandermonde factorization as that in (8), even though the series $\{h_0, h_1, \ldots\}$ may not geminate from a sinusoidal signal at all. Together with the remarks made in the second to the last paragraph of Section 2, we may say that any time series with finite rank trajectory matrix is equivalent to a sinusoidal signal.

Truncating a semi-infinite series (or trajectory matrix) into a finite series (or trajectory matrix) does preserve all the properties we have discussed thus far. The reality is that we usually begin with a finite series (or trajectory matrix) in practice, such as the SSA application. So can the finite dimensional problem contains enough information about the underlying infinite dimensional problem? The answer, known as the singular extension theory [33, II.9], is not trivial. Padding a given $m \times k$ Hankel matrix with zeros to expand it into a semi-infinite Hankel matrix usually results in a semi-infinite matrix with higher rank. One feasible way of extension is already manifested in the course of our discussion. That is, given any $d \times d$ nonsingular Hankel matrix, we embed this matrix as the leading principal submatrix of a nontrivial semi-infinite Hankel matrix H by specifying an extra value for h_{2d-1}. Only after

having solved the corresponding Eq. (14), we will have created the recursive relationship (13) and the resulting H is of rank d.

4 Least squares approximation

We have argued that almost all time series with a rank-d trajectory matrix can be interpreted as discrete samples of a noise-free sinusoidal signal with d components in the form (4). For contaminated series $\{z_0, z_1, \ldots\}$, however, the rank condition is usually lost. Seeking out the noise-free signal (4) means finding suitable values $\{\beta_1, \ldots, \beta_d\}$ and $\{\lambda_1, \ldots, \lambda_d\}$ for its Vandermonde parameters. In this section, we set up the least squares approximation for this purpose.

Consider the scenario that the SSA is applied to the finite time series $\{z_0, z_1, \ldots, z_{2p-2}\}$ with a square Hankel embedding, that is, suppose that the corresponding trajectory matrix Z is of size precisely $p \times p$. Without knowing any specfic parametrization, the SSA alternates between the best rank d approximation and the best Hankel approximation with respect to the Frobenius matrix norm. In contrast, since we now know that an element in a generic Hankel matrix of rank d should look like (20), the SSA application amounts to solving the minimization problem

$$\min_{\substack{\beta_\ell, \lambda_\ell \in \mathbb{C} \\ |\lambda_\ell| \le 1,\ \ell = 1, \ldots, d}} \sum_{i=0}^{p-1} (i+1) \left| z_i - \sum_{\ell=1}^{d} \beta_\ell \lambda_\ell^i \right|^2 + \sum_{i=p}^{2p-2} (2p-1-i) \left| z_i - \sum_{\ell=1}^{d} \beta_\ell \lambda_\ell^i \right|^2. \tag{21}$$

As $p \to \infty$, the scheme eventually is dealing with the problem

$$\min_{\substack{\beta_\ell, \lambda_\ell \in \mathbb{C} \\ |\lambda_\ell| \le 1,\ \ell = 1, \ldots, d}} \sum_{i=0}^{\infty} (i+1) \left| z_i - \sum_{\ell=1}^{d} \beta_\ell \lambda_\ell^i \right|^2. \tag{22}$$

Noticeable in both formulations is that the term-wise errors are not equally weighted. Does this make any sense? Take the extreme case (22) as a point of argument. We see that the weight is increased at a constant rate even when the signal gradually decays to zero. In a sense we are imposing more penalty on the noise for weaker signals. There are two consequential concerns. One is that the weights grow unboundedly and there is no satisfactory theory to guarantee the convergence of such a weighted infinite series in general. The other is that the weight scheme might have biased the outcome of the optimizers $(\beta_\ell, \lambda_\ell)$ for overly correcting the higher order terms which are composed mostly of noises. On the other hand, we think that the bell-shape weight distribution $1, 2, \ldots, p, p-1, \ldots 1$ in the finite case (21) is more an artifact due to our obstinacy in maintaining the Hankel structure than any other mathematical reason [9,45].

In what follows, we consider to fit the data by means of solving

$$\min_{\substack{\beta_\ell, \lambda_\ell \in \mathbb{C} \\ |\lambda_\ell| \le 1,\ \ell = 1, \ldots, d}} \sum_{i=0}^{2p-2} \left| z_i - \sum_{\ell=1}^{d} \beta_\ell \lambda_\ell^i \right|^2 \tag{23}$$

from the first $2p-1$ samples, that is, we regard all noises in (9) as being uncorrelated, independent, and having equal variance. For the SSA computation, it is preferred to keep p as small as is necessary because the size of the associated trajectory matrix does affect the overhead in the sequence of matrix decomposition. But in our formulation, the size p matters

only to the summation and there are a fixed number of only $2d$ parameters to be estimated. Note that we allow λ_ℓ to range over a closed unit disk. The following arguments assure that the minimum in (23) does exist.

Define

$$\Lambda := \Lambda(\lambda_1, \ldots, \lambda_d) := \begin{bmatrix} I & I & \ldots & I \\ \lambda_1 & \lambda_2 & \ldots & \lambda_d \\ \lambda_1^2 & \lambda_2^2 & & \lambda_d^2 \\ \vdots & & & \\ \lambda_1^{2p-2} & \lambda_2^{2p-2} & & \lambda_d^{2p-2} \end{bmatrix}, \quad \boldsymbol{\beta} := \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{bmatrix}, \quad \mathbf{z} := \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_{2p-2} \end{bmatrix}.$$

Trivially, we can rewrite (23) in vector form

$$\min_{\substack{\beta_\ell, \lambda_\ell \in \mathbb{C} \\ |\lambda_\ell| \le 1,\ \ell = 1, \ldots, d}} \quad \|\Lambda \boldsymbol{\beta} - \mathbf{z}\|^2. \tag{24}$$

This is a nonlinear least squares problem over the complex field. For each given Λ of full column rank, the unique optimal solution for $\boldsymbol{\beta}$ is given by [38]

$$\boldsymbol{\beta} := \mathcal{L}(\Lambda; \mathbf{z}) := (\Lambda^* \Lambda)^{-1} \Lambda^* \mathbf{z} \tag{25}$$

where * stands for the conjugate transpose. The problem is thus reduced to finding λ for the minimization problem

$$\min_{|\lambda_1|, \ldots, |\lambda_d| \le 1} \quad \|\Lambda (\Lambda^* \Lambda)^{-1} \Lambda^* \mathbf{z} - \mathbf{z}\|^2. \tag{26}$$

An example of such an objective function for the case $d = 1$ with randomly generated noise is illustrated in Fig. 1, showing the presence of multiple minimizers. Most importantly, in all cases a minimizer over the compact d-dimensional torus must exist. Since we are dealing with a compact set, it is possible to exploit global optimization techniques to search for the absolute minimizer, which will be delineated later. We are interested in using the global solution to compare the result constructed from the SSA algorithm.

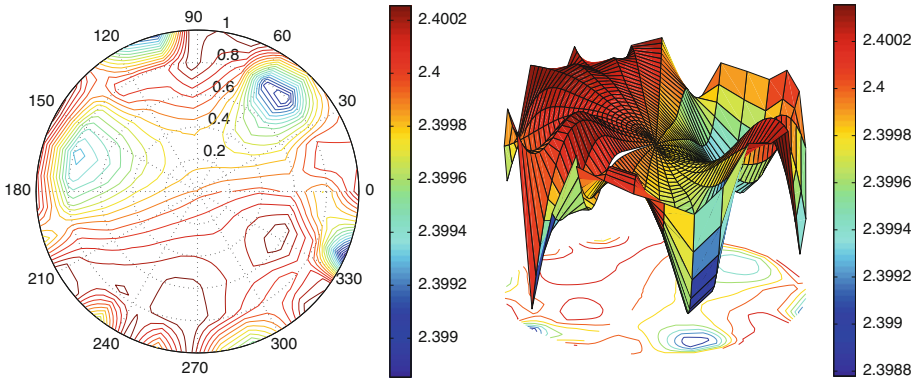

Fig. 1 An example of objective function for (26). (The surface has been rotated with AZ = 124 and EL = 38)

5 Absolute solution versus SSA solution

The feasible set in our least squares formulation for the SSA is compact, so the absolute minimizer always exists. In this section, we report some numerical experiments on employing the global optimization techniques to find the absolute minimizer. We stress again that our goal is not to propose a new method for the SSA. Rather, because our parametrization makes it possible to locate the absolutely best optimizer, we want to employ this more expensive method to yield an absolute baseline for comparison with results obtained from the SSA algorithm. As will be seen, our discovery is rather surprising.

5.1 Global optimization

Any existing global optimization software should suffice to meet our need for building a baseline of comparison. For completion, we briefly review only the general notion of global optimization and describe how we set up the experiments.

The objective of global optimization is to find the globally best solution in the presence of multiple local optima. This is an extremely important tool for almost all fields of applications where optimization is needed. Over the years, many strategies for global optimization have been proposed. Research results are too numerous to count and the research is ongoing. Far from being complete, we mention books [31,44] for introduction and [37,43,49,59] for comprehensive survey on continuous global optimization. Over all, it might be fair to say that one prevailing scheme that allows general structural constraints and makes possible straightforward generalizations and extensions of existent local algorithms is the notion of adaptive partition.

A partial list of practical methods with a view towards efficient computer-based implementations includes multistart framework equipped with sophisticated scatter search algorithm [52]; generic algorithms motivated by the process of natural selection and the survival of the fittest principle [30,56]; simulated annealing based upon the physical analogy of cooling crystal structures that spontaneously attempt to arrive at some stable and, hopefully, global equilibrium [51,36]; and pattern search methods which compare instantiations and explore new regions with the goal of conducting a global search [3,35]. Depending on the applications, there are also other more specialized techniques such as the stochastic global optimization. See, for examples, the discussions in [43,58] and the references contained therein.

It is not the intention of this paper to propose a new global optimization method. Rather, we employ global optimization techniques to help to establish an absolute baseline to evaluate the performance of the SSA. Toward that goal, our effort has been to set up the framework for global optimization calculation. Once this framework is established, any existent global optimization software can serve our purpose. For demonstration, we take advantage of the global optimization toolbox available in MATLAB and choose to use the MULTISTART method with solver FMINCON to carry out our experiments. As the optimization solver makes use of derivatives, exact gradient information has the benefit of improving efficiency and precision. We supply such a calculation of gradient in the Appendix.

We carry out our experiment by setting up test data in the following way. Our design is based on two assumptions. First, as we have advanced in this discourse, it suffices to use noise-free sinusoidal signals as the control group. Second, we assume that the expected rank d is known a priori, which itself is a difficult problem in practice. We generate a sinusoidal signal $\mathfrak{s}(t)$ in the form of (4) with d components as the basis for comparison. This exact

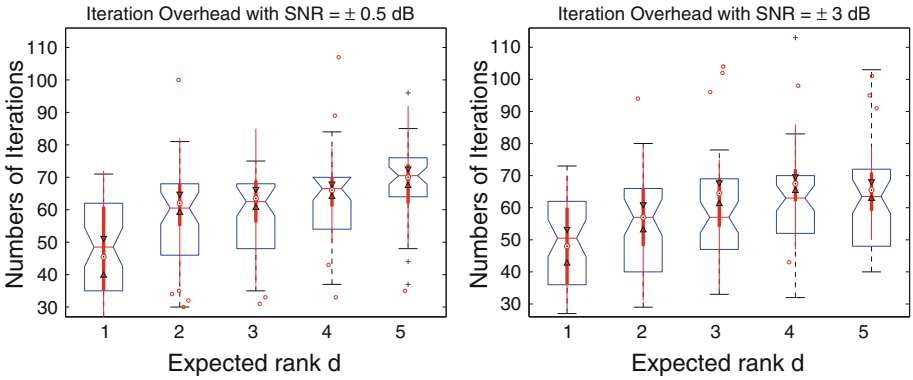

Fig. 2 Overhead of SSA in terms of iterations needed for convergence. SNR > 0 (*red compact box style*); SNR < 0 (*blue empty box style*). (Color figure online)

signal is random in the sense that its parameters $a_\ell \in [0, 6]$, $\alpha_\ell \in [0, 5]$, and $\phi_\ell \in [0, 1]$ are uniformly distributed pseudorandom numbers, while the frequencies ν_ℓ are pseudorandom integers from a uniform discrete distribution over $[0, 400]$. Sampling rate is set at 2^9 and we take $2^{10} - 1$ samples $s_k := \mathfrak{s}(t_k)$ over the time interval $[0, 2)$. Additive white Gaussian noises based on the AWGN channel model with fixed SNR are then added to the signal in the way of (9) as the observed signal. We apply both the SSA algorithm and the MULTISTART method to the observed signals. The goal is to examine the closeness of the reconstructed signals by the algorithms to the original signal.

5.2 Overall effect of error strength

We first experiment with the effect of error strength[2] on the performance of the SSA method. We want to examine the impact on both the overhead and the accuracy.

For simplicity of demonstration, we choose not to perform the group selection and consider only the case with embedding dimension 512, i.e., the trajectory matrix Z is always of size 512×512. With each fixed $d = 1, \ldots, 5$, we randomly generate 100 signals. The alternating projection scheme is applied to each randomly perturbed time series until two consecutive Hankel projections differ by less than 10^{-12} when we regard that convergence has been achieved. Each iteration involves the SVD of a 512×512 Hankel matrix, which is not cheap. Depicted Fig. 2 are the boxplots for numbers of iterations needed by the SSA to attain a rank d Hankel approximation of the original Z subject to four different strengths of noise ± 0.5 and ± 3 dB. It seems to suggest that the strength of noise is not critical for convergence. However, from Fig. 3, it is quite obvious that stronger noise (negative SNR) does affect the quality of approximation.

We caution that, because the manifold of rank d matrices is not convex, the alternating projection scheme, central to the SSA algorithm, may not return at its convergence a rank d Hankel approximation to the given trajectory matrix Z. Even if it does, the limit point may not be the nearest rank d Hankel approximation to Z. See a counterexample in [9]. In contrast, our formulation directly searches for the best fitting signal in the sense of (24).

[2] We measure the strength by the logarithm unit dB. Recall that one decibel is ten times the base-10 logarithm of the ratio of the measured quantity to the reference level. Positive dB means that the signal is stronger than the noise; otherwise, the noise is stronger.

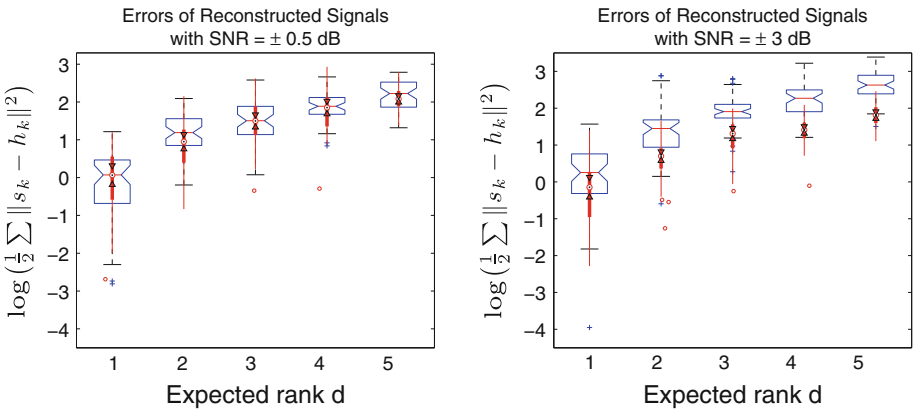

Fig. 3 Errors between the true time series $\{s_k\}$ and the reconstructed time series $\{h_k\}$ by SSA. SNR > 0 (*red compact box style*); SNR < 0 (*blue empty box style*). (Color figure online)

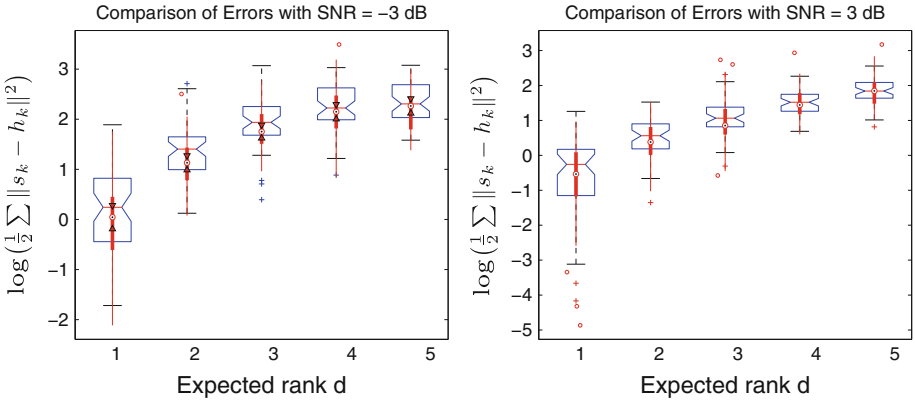

Fig. 4 Contrast of errors between the true time series $\{s_k\}$ and the reconstructed time series $\{h_k\}$ by the global optimization method (*red compact box style*) and by the SSA algorithm (*blue empty box style*). (Color figure online)

In yet another experiment with randomly generate signals, we contrast the errors of the signals reconstructed by the SSA algorithm with those by our global optimization approach for each $d = 1, \ldots, 5$ with AWGN noises at SNR $= \pm 3$ dB in Fig. 4. Recall that this is the primary purpose of this paper. We see that in general the global optimization approach gives rise to better approximation to the original time series $\{s_k\}$ than the SSA algorithm. The advantage of the global optimization approach to the SSA algorithm becomes more obvious when the noise is stronger (e.g., dB $= -3$). For weakly perturbed signals, the SSA performs almost compatibly to that by global optimization approach.

5.3 Sample by sample comparison

The boxplots in Fig. 4 indicate to us the overall trend. It is informative to also check out the errors produced by each independent random test. We take the ratio of the SSA error to the global optimization error sample by sample and plot the histogram of logarithm of these

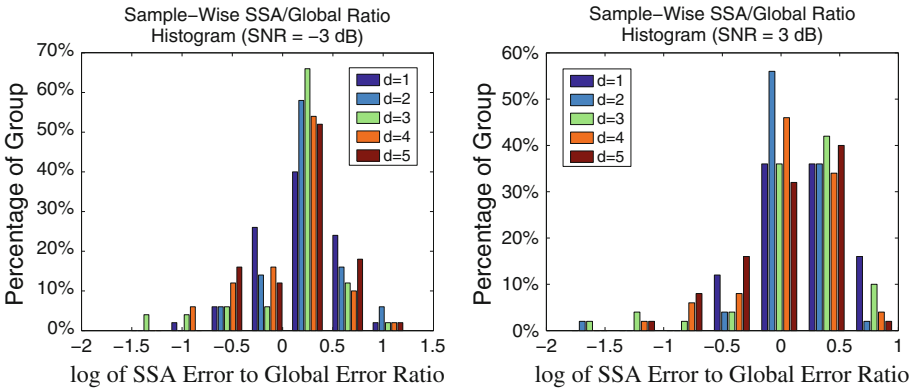

Fig. 5 Histogram of sample-by-sample SSA error to global optimization error ratios

ratios in Fig. 5. For comparison, we plot the histogram for all ranks together, distinguished by colors, while containers in each cluster share the same bin value which is the midpoint of the corresponding cluster range shown in the drawing. A negative logarithm of the ratio indicates that the SSA reconstruction is superior to the global optimization reconstruction.

From the histogram we observe that, except for the case $d = 1$, the global optimization usually has a higher percentage of producing a better reconstruction when the noise is strong. In contrast, when the noise is relatively weak (e.g., dB = 3), the simpler SSA performs amazingly well, comparing to the more complicated global optimization.

5.4 Effect of weight scheme

One might wonder why sometimes the SSA algorithm gives better approximation to $\{s_k\}$. Should not the global optimization does the best? It is worth noting the multiple factors that might cause this to happen. Firstly, the quantity being minimized in the global optimization is the sum $\sum_{i=0}^{2p-2} \left| z_i - \sum_{\ell=1}^{d} \beta_\ell \lambda_\ell^i \right|^2$ (see 23), not the absolute error $\sum_{i=0}^{2p-2} \left| s_i - \sum_{\ell=1}^{d} \beta_\ell \lambda_\ell^i \right|^2$. Because the true time series $\{s_k\}$ is practically not available in reality, the global optimization offers at its best an optimal approximation to the series $\{z_k\}$, not $\{s_k\}$. Secondly, the machinery we employed in the global optimization is the scheme MULTISTART which uses uniformly distributed start points within predefined bounds. In our experiment we allow merely 40 start points. It is possible to improve the objective value, or to confirm the global solution, by more start points. Thirdly, we have mentioned that not necessarily the SSA algorithm will produce the best rank d Hankel approximation at its convergence. Even if it does, the SSA is a specially weighted least squares problem (see 21), whereas our formulation (23) is not. So we are comparing nearness to a "blackbox" $\{s_k\}$ from two different objective functions.

If we really want to consider

$$\min_{\substack{\beta_\ell, \lambda_\ell \in \mathbb{C} \\ |\lambda_\ell| \le 1,\ \ell = 1, \ldots, d}} \quad \sum_{i=0}^{2p-2} \omega_i \left| z_i - \sum_{\ell=1}^{d} \beta_\ell \lambda_\ell^i \right|^2 \tag{27}$$

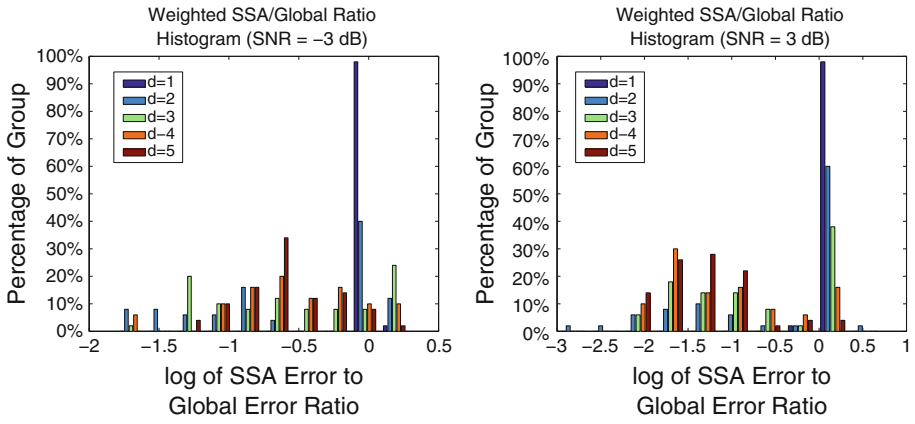

Fig. 6 Histogram of weighted sample-by-sample SSA error to global optimization error ratios

with weights $\omega_i > 0$, the normal Eq. (25) per fixed Λ becomes

$$\Lambda^* W \Lambda \boldsymbol{\beta} = \Lambda^* W \mathbf{z}, \tag{28}$$

with $W := \mathrm{diag}\{\omega_0, \omega_0, \ldots, \omega_{2p-2}, \omega_{2p-2}\}$. Equivalently, with the change of variables by diagonal scaling

$$\Upsilon := W^{\frac{1}{2}} \Lambda, \quad \mathbf{c} := W^{\frac{1}{2}} \mathbf{z}, \tag{29}$$

the weighted least squares problem (27) can be reformulated as

$$\min_{|\lambda_1|,\ldots,|\lambda_d| \le 1} \quad \|\Upsilon(\Upsilon^*\Upsilon)^{-1}\Upsilon^*\mathbf{c} - \mathbf{c}\|^2, \tag{30}$$

which is analogous to (24) and our code can easily be modified accordingly.

If in particular we use the special SSA weight scheme, a rather surprising reversal situation then is observed. As can be seen in Fig. 6, there are many more samples showing a negative value in its logarithm of SSA error to global error ratio than those in Fig. 5, indicating that the SSA generally performs better than the global optimization. The contrast is especially strong at the presence of stronger noise.

6 Forecasting

An important application of the SSA is to predict future values based on the underlying model constructed from existent data. In this regard, our global optimization framework has the advantage of working directly with the Vandermonde parameters β_ℓ's and λ_ℓ's. That is, once these parameters are estimated, the forecasting is a natural consequence as is described in (20). Though it is not the primary goal of this paper, we demonstrate this capacity in this section by comparing forecast values obtained from the global optimization and the basic SSA.

For completion, we briefly review how forecasting by the basic SSA is accomplished. More details can be found in [19,23,29]. Suppose that $\mathbf{u}_1, \ldots, \mathbf{u}_d$ are the left singular vectors at convergence of the SSA applied to the given time series $\{z_0, \ldots, z_n\}$. Ideally, we would prefer to see that a rank-d Hankel matrix $\widetilde{Z}$ has been found. Otherwise, the final reconstructed series

$\{h_0, \ldots, h_n\}$ is obtained by diagonally averaging entries of $\widetilde{Z}$. Let each singular vector $\mathbf{u}_i$ be partitioned as

$$\mathbf{u}_i = \begin{bmatrix} \widetilde{\mathbf{u}}_i \\ \pi_i \end{bmatrix}$$

with $\widetilde{\mathbf{u}}_i$ denoting the first $p-1$ entries and π_i is the last entry of $\mathbf{u}_i$. Assuming $\nu := \pi_1^2 + \cdots + \pi_d^2 < 1$, define

$$\begin{bmatrix} a_{p-1} \\ \vdots \\ a_1 \end{bmatrix} := \sum_{i=1}^{d} \frac{\pi_i \widetilde{\mathbf{u}}_i}{1-\nu^2}. \tag{31}$$

Then it has been proposed [29, Section 3.2] (See also [19, Theorem 5.2]) that the finite difference scheme

$$h_k := \sum_{i=1}^{p-1} a_i h_{k-i}, \quad k = n+1, \ldots \tag{32}$$

be used as a forecasting scheme. In contrast, the Vandermonde parameters naturally give rise to any future values via the closed form (20). If necessary, we can generate the LRR (13) via (14). In that case, note the fundamental difference that the LRR (13) uses only d starting values $\{h_0, \ldots, h_{d-1}\}$ and is of minimal order.

6.1 Trend detection

We use the real-valued signal

$$s(t) = 3(.1)^t + \sin(.2\pi t) \tag{33}$$

to demonstrate the forecasting capacity of both approaches. We sample the signal at the rate 2^9 over interval $[0, 2)$. It can be verified that the exact time series corresponds to the case $d = 3$ with exact Vandermonde parameters $\boldsymbol{\beta} = [3, -.5\imath, .5\imath]$ and $\boldsymbol{\lambda} = [.1, e^{\frac{.2\pi\imath}{2^9}}, e^{-\frac{.2\pi\imath}{2^9}}]$. We then add real-valued AWGN noises with specified SNR to create artificial observed data $\{z_k\}$. Out of the 1023 samples, we use the first $n = 511$ to construct the basic model which then is used to predict the remaining 512 future values. For the SSA, the window length is set at $p = \lceil \frac{n}{2} \rceil = 256$. For the global optimization, 40 initial values are taken. In the latter, we must point out that the optimization takes place over a complex domain, so the computed Vandermonde parameters and the reconstructed signal could be complex-valued. To retrieve real-valued approximation, we simply project the signal to the real line.

Two possible scenarios are depicted in Fig. 7. Note that in both cases, we have used only the information over the interval [0,1) to construct the model. As is expected, the Vandermonde parameters obtained from global optimization are fairly consistent in predicting values over the interval [1,2) by using (20). The trend of the original signal is reasonably identifiable, though in the picture on the right of Fig. 7 the prediction begins to deviate. We notice that in both scenarios the SSA results suffer from considerable fluctuations around the true curve. Conceivably, its forecasting of values over the interval [1,2) will not be as accurate as that from the global optimization. Instead of displaying the forecast values by the SSA, we report the following measurement.

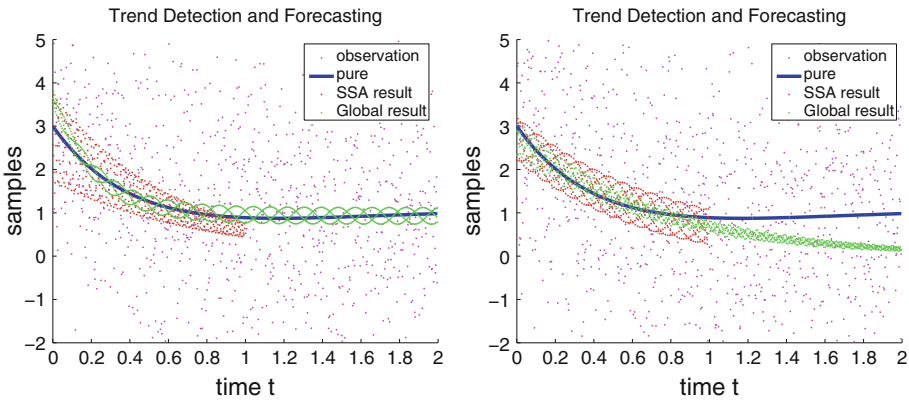

Fig. 7 Detecting trends and making forecasting under strong noise SNR $= -3$ by SSR (*red*) and global optimization (*green*). (Color figure online)

To gauge the effectiveness of the forecasting in general, we employ the notion of root-mean-square errors (RMSE) defined by [32]

$$\text{RMSE} := \sqrt{\frac{1}{512}\sum_{k=512}^{1023}(z_k - h_k)^2}, \tag{34}$$

where z_k and h_k stand for the observation and the forecasting values at time t_k, respectively. We repeat the experiment 50 times for each SNR $= \pm 3$ by generating different AGWN noises and collect the resulting RMSE for each test. The results are depicted in Fig. 8. The figure on the left indicates that the average RMSE by the global optimization is about 66 % of that by the SSA under weak noise (SNR $= 3$dB); whereas the ratio is about 80 % under strong noise (SNR $= -3$dB). Also indicated is that the RMSE for SSA has lots of outliers, making the mean calculation biased toward the higher end. The figure on the right is the histogram of sample to sample future SSA error to the global optimization error. A positive logarithm of this ratio indicates that the global optimization forecasting is more accurate the the SSA forecasting. Together with phenomenon observed in Fig. 7, we think the forecasting capacity of the global optimization should be quite convincing.

6.2 Real data application

Thus far, all our experiments have been done under the "controlled" condition that the rank of the underlying signal is precisely known a priori. These experiments are designed to provide an absolute base to evaluate the performance of the basic SSA. In real applications, the determination of a suitable rank is not easy and plays a critical role in the effectiveness of the reconstructed model. We realize that for each individual real data set, a lot endeavors need to be taken to select as appropriate rank, to fine tune the parameters, and to interpret the resulting model. We notice that most of the many research articles have had to deal with one individual case each time. In the section, we consider two real data sets and demonstrate how the global optimization method performs under these uncertainties.

Depicted by the blue curve in Fig. 9 are the daily closing gold prices for 97 successive trading days recorded by Hipel and Mcleod in 1994 [19]. Needless to say, many

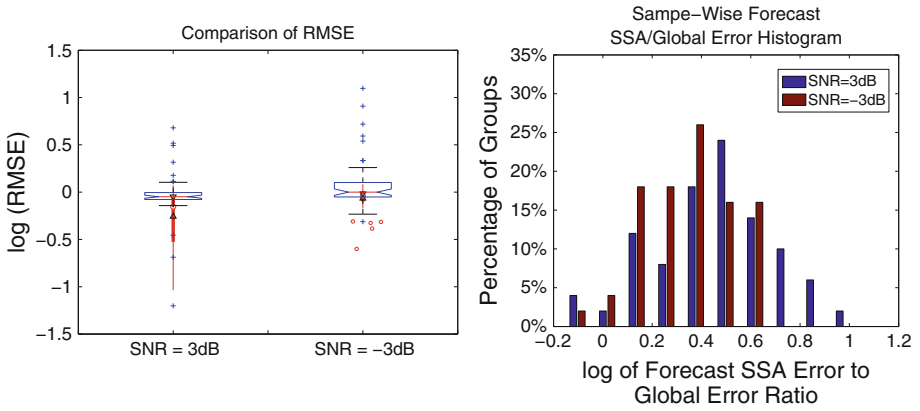

Fig. 8 Comparisons of the RMSE and forecast errors between the global optimization method (*red compact box style*) and the SSA algorithm (*blue empty box style*) with respect to the signal (33). (Color figure online)

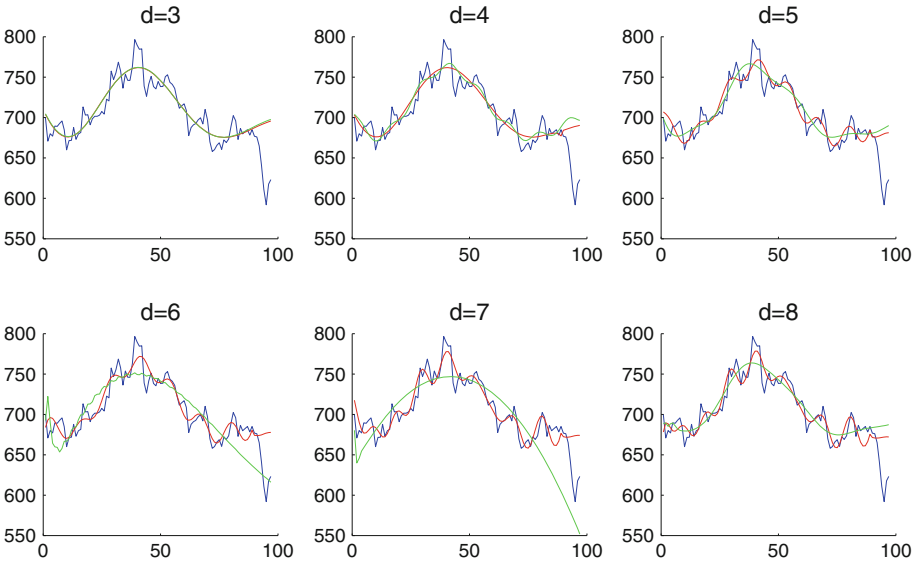

Fig. 9 Approximation of 97-day gold price (*blue*) by using the first 89 daily prices via SSA (*red*) and global optimization (*green*). Prices for day 90 to 97 are by forecasting. (Color figure online)

factors would have affected some abrupt changes of gold price in the series. The fundamental question is whether the gold price can be forecast based on past values. For our experiment, we take the first 89-day prices and wish to compare the forecast the last 8-day prices. We take the suggestion in [19] by setting the window length at $p = 45$. Also suggested is the rank at about $d = 5$ or 6, but we choose to try all ranks from 3 to 8. For the global optimization, we try 50 starting values. Approximations by both methods at $d = 3$ are nearly identical. The SSA approximations for $d = 4$ and $d = 5$ is distinguishable, but for higher ranks the SSA approximations are almost identical. It is worth noting that for $d \geq 5$, the SSA approximations fluctuate almost at the same rhythm as the observed daily prices. This behavior is critically important and show the amazing ability

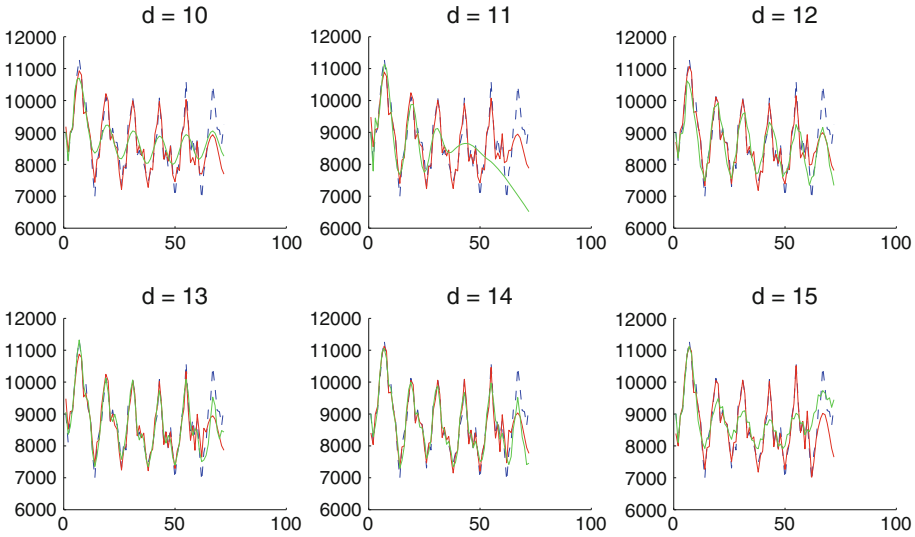

Fig. 10 Approximation of 1973–1978 monthly accidental deaths (*blue*) in the USA by the first 63 monthly data via SSA (*red*) and global optimization (*green*). Data after the 64th month are by forecasting. (Color figure online)

of the SSA method. In contrast, the global optimization method is able to provide smoothing of the otherwise fluctuant prices, but it fails to pick up the fluctuant point. On the other hand, the SSA method fails to predict the deep dive of the gold price at the last 8 days, while on two occasions ($d = 6, 7$) the global optimization predicts a continuing down slide of the gold price. We have stressed the determination of a suitable rank d is itself a difficult problem. This experiment strongly suggests that in real applications a lot more factors, some might be even unknown, will affect the final construction of the model.

Similarly, depicted in Fig. 10 are the monthly accidental deaths in the USA from 1973 to 1978. This data set has been used by many studies for testing the efficiency of algorithms [7,22,23]. We set the window length at $p = 32$ and use the first 63 monthly data to construct the model. We try all ranks from $d = 10$ to 15. Except for the case $d = 11$ where the model from the global optimization gives the wrong prediction, both methods perform reasonably well, whereas we think that the simple and fast SSA method is superior to more expensive global optimization approach.

It is important to point out one possible explanation on why the SSA method can outperform the global optimization method in both experiments above. We notice that the data stay far away from the unit disk and change rapidly, whereas in our setup for the global optimization we have restricted λ_ℓ (See 6) to be within the unit disk. As such, the global optimization method is more suitable if the data inherit some asymptotic behavior when the powers of λ_ℓ die out. When the data simply fluctuate as we have seen in the two real data sets, we rely on β_ℓ to compensate the decline of the powers of λ_ℓ. We may also need more terms in the summation (7), which might then bring in more noise than the lower rank approximation.

7 Conclusion

Alternating projection between two manifolds with the aim of seeking out a point of intersection, or a pair of points yielding minimum distance, has long been employed across the fields

in a variety of applications. Starting with a Hankel matrix embedded with a contaminated time series, the SSA exploits this notion to find, in particular, a low rank Hankel approximation. The idea is simple enough for quick implementation and has gained popularity among practitioners.

This paper investigates the effectuality of the SSA algorithm by comparing its reconstructed time series with the absolute best approximation obtained from global optimization techniques. The framework is built upon Vandermonde factorization of Hankel operators, which provides natural parameters over compact feasible set and, hence, guarantees the existence of a global solution.

Setting the stopping criterion at the nearness of two successive iterates within a uniformly specified bound, we find that the convergence behavior of the SSA algorithm is generally insensitive to the strength of the added noise. That is, so long as the size of the embedding is fixed, it usually takes about the same order of iterations for the SSA to converge, regardless of the SNR. On the other hand, as is expected, stronger noise does degrade the quality of the final approximation to the original, uncontaminated signal.

What is most interesting is the empirical evidence that, despite of its simplicity, the SSA does indeed perform remarkably well when comparing to the results from the more complicated global methods. This is especially so when the SNR is relatively high. Since the global methods, if successful, produce the best possible approximation to the Vandermonde parameters from the contaminated data, such a comparison justifies therefore that, for exploratory model-building purpose, the simpler SSA algorithm might be sufficient as a handy tool. For more rigorous affirmative procedure, however, we must stress that the SSA does not always give rise to an optimal solution.

8 Appendix

In this section we explain how the gradient information can be derived for the global optimization code MULTISTART.

Let the objective function (26) be rewritten in the form

$$f(\lambda_1, \ldots, \lambda_d) = \frac{1}{2}\langle \Lambda\mathcal{L}(\Lambda; \mathbf{z}) - \mathbf{z}, \Lambda\mathcal{L}(\Lambda; \mathbf{z}) - \mathbf{z}\rangle,$$

with inner product $\langle \mathbf{p}, \mathbf{q}\rangle := \sum_i p_i \overline{q}_i$ for complex vectors. Employing polar coordinates (ρ_i, θ_i) when $\lambda_i = \rho_i e^{\imath\theta_i}$ as variables in our global method, we identify $\Lambda = \Lambda(\lambda_1, \ldots, \lambda_d) = \Lambda(\rho_1, \ldots, \rho_d, \theta_1, \ldots, \theta_d)$. We now calculate the gradient of f with respect to the real variables ρ_i and θ_i.

Using the fact that $\Lambda^*\,(\Lambda\mathcal{L}(\Lambda; \mathbf{z}) - \mathbf{z}) = 0$, the action of the Fréchet derivative of f with respect to Λ at a complex matrix H is given by

$$\frac{\partial f}{\partial \Lambda}.H = \Re\left(\langle H, (\Lambda\mathcal{L}(\Lambda; \mathbf{z}) - \mathbf{z})\,\mathcal{L}(\Lambda; \mathbf{z})^*\rangle\right),$$

where $\Re$ stands for the real part of a complex-valued quantity and the same notation $\langle\cdot,\cdot\rangle$ denotes the generalization to the Frobenius inner product of complex matrices. On the other hand, the action of the Fréchet derivative of Λ with respect to variables $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_d]$ and $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_d]$ at vectors $\mathbf{h}, \mathbf{k} \in \mathbb{R}^d$ can be expressed as

$$\frac{\partial \Lambda}{\partial \boldsymbol{\rho}}.\mathbf{h} = \underbrace{\operatorname{diag}\{0, 1, 2, \ldots, 2n-2\}}_{\Xi} \Lambda \operatorname{diag}\{\boldsymbol{\rho}\}^{-1} \operatorname{diag}\{\mathbf{h}\},$$

$$\frac{\partial \Lambda}{\partial \boldsymbol{\theta}}.\mathbf{k} = \imath\, \Xi \Lambda \operatorname{diag}\{\mathbf{k}\},$$

respectively. By the chain rule, we obtain the actions of the gradient of f as follows:

$$\begin{aligned}
\frac{\partial f}{\partial \boldsymbol{\rho}}.\mathbf{h} = \frac{\partial f}{\partial \Lambda}.\left(\frac{\partial \Lambda}{\partial \boldsymbol{\rho}}.\mathbf{h}\right) &= \Re\left(\langle \Xi \Lambda \operatorname{diag}\{\boldsymbol{\rho}\}^{-1} \operatorname{diag}\{\mathbf{h}\},\, (\Lambda \mathcal{L}(\Lambda; \mathbf{z}) - \mathbf{z})\, \mathcal{L}(\Lambda; \mathbf{z})^* \rangle\right), \\
&= \Re\left(\langle \operatorname{diag}\{\mathbf{h}\},\, \operatorname{diag}\{\boldsymbol{\rho}\}^{-1}\, \Lambda^* \Xi\, (\Lambda \mathcal{L}(\Lambda; \mathbf{z}) - \mathbf{z})\, \mathcal{L}(\Lambda; \mathbf{z})^* \rangle\right), \\
\frac{\partial f}{\partial \boldsymbol{\theta}}.\mathbf{k} = \frac{\partial f}{\partial \Lambda}.\left(\frac{\partial \Lambda}{\partial \boldsymbol{\theta}}.\mathbf{k}\right) &= \Re\left(\langle \imath\, \Xi \Lambda \operatorname{diag}\{\mathbf{k}\},\, (\Lambda \mathcal{L}(\Lambda; \mathbf{z}) - \mathbf{z})\, \mathcal{L}(\Lambda; \mathbf{z})^* \rangle\right) \\
&= \Re\left(\langle \operatorname{diag}\{\mathbf{k}\},\, -\imath\, \Lambda^* \Xi\, (\Lambda \mathcal{L}(\Lambda; \mathbf{z}) - \mathbf{z})\, \mathcal{L}(\Lambda; \mathbf{z})^* \rangle\right)
\end{aligned}$$

By the Riesz representation theorem, we see that the gradient of f can be expressed as

$$\begin{aligned}
\frac{\partial f}{\partial \boldsymbol{\rho}} &= \Re\left(\operatorname{diag}\left\{\operatorname{diag}\{\boldsymbol{\rho}\}^{-1}\, \Lambda^* \Xi\, (\Lambda \mathcal{L}(\Lambda; \mathbf{z}) - \mathbf{z})\, \mathcal{L}(\Lambda; \mathbf{z})^*\right\}\right), \\
\frac{\partial f}{\partial \boldsymbol{\theta}} &= \Re\left(\operatorname{diag}\left\{-\imath\, \Lambda^* \Xi\, (\Lambda \mathcal{L}(\Lambda; \mathbf{z}) - \mathbf{z})\, \mathcal{L}(\Lambda; \mathbf{z})^*\right\}\right).
\end{aligned}$$

References

1. Adamjan, V.M., Arov, D.Z., Kreĭn, M.G.: Infinite Hankel matrices and generalized Carathéodory-Fejér and I. Schur problems. Funkcional. Anal. i Priložen. **2**(4), 1–17 (1968)
2. Allen, M.R., Smith, L.A.: Monte Carlo SSA: Detecting irregular oscillations in the presence of coloured noise. J. Clim. **9**, 3372–3404 (1996)
3. Audet, C., Dennis Jr, J.E., Le Digabel, S.: Globalization strategies for mesh adaptive direct search. Comput. Optim. Appl. **46**(2), 193–215 (2010)
4. Auvergne, M.: Singular value analysis applied to phase space reconstruction of pulsating stars. Astron Astrophys **204**, 341–348 (1988)
5. Boley, D.L., Luk, F.T., Vandevoorde, D.: Vandermonde factorization of a Hankel matrix. Scientific computing (Hong Kong, 1997), pp. 27–39. Springer, Singapore (1997)
6. Bowden, C.M.: Boundedness of linear operators in the space ℓ^2. Int. J. Quantum Chem. **2**, 363–371 (1968)
7. Brockwell, P.J., Davis, R.A.: Introduction to time series and forecasting, second edn. Springer Texts in Statistics. Springer, New York (2002). doi:10.1007/b97391. With 1 CD-ROM (Windows)
8. Broomhead, D.S., King, G.P.: Extracting qualitative dynamics from experimental data. Physica D **20**(2–3), 217–236 (1986)
9. Chu, M.T., Funderlic, R.E., Plemmons, R.J.: Structured low rank approximation. Linear Algebra Appl. **366**, 157–172 (2003). Special issue on structured matrices: analysis, algorithms and applications (Cortona, 2000)
10. Danilov, D., Zhigljavsky, A.E.: Principal components of time series: the 'caterpillar' method. University of St. Petersburg Press. (In Russian) (1997)
11. Elsner, J., Tsonis, A.: Singular Spectrum Analysis: A New Tool in Time Series Analysis. Language of Science. Plenum Press, NY (1996)
12. Feldmann, S., Heinig, G.: Vandermonde factorization and canonical representations of block Hankel matrices. In: Proceedings of the Fourth Conference of the International Linear Algebra Society (Rotterdam, 1994), vol. 241/243, pp. 247–278 (1996)
13. Fraedrich, K.: Estimating the dimensions of weather and climate attractors. J. Atmos. Sci. **43**(5), 419–432 (1986)
14. Francis, B.A.: A course in H_∞ control theory, Lecture Notes in Control and Information Sciences, vol. 88. Springer, Berlin (1987)
15. Gantmacher, F.R.: The theory of matrices. Vols. 1, 2. Translated by K.A. Hirsch. Chelsea Publishing Co., New York (1959)

16. Ghil, M., Allen, R.M., Dettinger, M.D., Ide, K., Kondrashov, D., Man, M.E.: Advanced spectral methods for climatic time series. Rev. Geophys. **40**(1), 3-1–3-41 (2002)
17. Ghil, M., Taricco, C.: Advanced spectral analysis methods. In: Castagnoli G.C., Provenzale A. (eds.) Past and Present Variability of the Solar-Terrestrial System: Measurement, Data Analysis and Theoretical Models, International School of Physics Enrico Fermi, vol. 133, pp. 137–159. IOS Press, NY (1997)
18. Glover, K.: All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds. Int. J. Control **39**(6), 1115–1193 (1984)
19. Golyandina, N., Nekrutkin, V., Zhigljavsky, A.: Analysis of Time Series Structure: SSA and Related Techniques, Monographs on Statistics and Applied Probability, vol. 90. Chapman & Hall/CRC, Boca Raton, FL (2001)
20. Golyandina, N., Zhigljavsky, A.: Singular Spectrum Analysis for Time Series. Springer Briefs in Statistics. Springer, Berlin (2013)
21. Groth, A., Ghil, M.: Multivariate singular spectrum analysis and the road to phase synchronization. Phys. Rev. E **84**(036206), 1–10 (2011)
22. Hassani, H.: Singular spectrum analysis: Methodology and comparison. J. Data Sci. **5**(2), 239–257 (2007)
23. Hassani, H.: Singular spectrum analysis based on the minimum variance estimator. Nonlinear Anal. Real World Appl. **11**(3), 2065–2077 (2010)
24. Hassani, H., Heravi, S., Zhigljavsky, A.: Forecasting european industrial production with singular spectrum analysis. Int. J. Forecast. **25**(1), 103–118 (2009)
25. Hassani, H., Mahmoudvand, R.: Multivariate singular spectrum analysis: A general view and new vector forecasting approach. Int. J. Energy Stat. **01**(01), 55–83 (2013)
26. Hassani, H., Mahmoudvand, R., Zokaei, M.: Separability and window length in singular spectrum analysis. C. R. Math. Acad. Sci. Paris **349**(17–18), 987–990 (2011)
27. Hassani, H., Mahmoudvand, R., Zokaei, M., Ghodsi, M.: On the separability between signal and noise in singular spectrum analysis. Fluct. Noise Lett. **11**(02), 1250,014 (2012)
28. Hassani, H., Soofi, A.S., Zhigljavsky, A.A.: Predicting daily exchange rate with singular spectrum analysis. Nonlinear Anal. Real World Appl. **11**(3), 2023–2034 (2010)
29. Hassani, H., Thomakos, D.: A review on singular spectrum analysis for economic and financial time series. Stat. Interface **3**(3), 377–397 (2010)
30. Haupt, R.L., Haupt, S.E.: Practical Genetic Algorithms, 2nd edn. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ (2004). With 1 CD-ROM (Windows)
31. Hendrix, E.M.T., G.-Tóth, B.: Introduction to Nonlinear and Global Optimization, Springer Optimization and Its Applications, vol. 37. Springer, New York (2010)
32. Hyndman, R.J., Koehler, A.B.: Another look at measures of forecast accuracy. Int. J. Forecast. **22**(4), 679–688 (2006)
33. Iohvidov, I.S.: Hankel and Toeplitz matrices and forms. Birkhäuser Boston, Mass. (1982). Algebraic theory, Translated from the Russian by G. Philip A. Thijsse, With an introduction by I. Gohberg
34. Kalman, D.: The generalized Vandermonde matrix. Math. Mag. **57**(1), 15–21 (1984)
35. Kolda, T.G., Lewis, R.M., Torczon, V.: Optimization by direct search: new perspectives on some classical and modern methods. SIAM Rev. **45**(3), 385–482 (2003)
36. van Laarhoven, P.J.M., Aarts, E.H.L.: Simulated Annealing: Theory and Applications, Mathematics and its Applications, vol. 37. D. Reidel Publishing Co., Dordrecht (1987)
37. Liberti, L., Maculan, N. (eds.): Global Optimization, Nonconvex Optimization and its Applications: From Theory to implementation, vol. 84. Springer, New York (2006)
38. Miller, K.S.: Complex linear least squares. SIAM Rev. **15**, 706–726 (1973)
39. Mineva, A., Popivanov, D.: Method for single-trial readiness potential identification, based on singular spectrum analysis. J. Neurosci. Methods **68**(1), 91–99 (1996)
40. Moskvina, V., Schmidt, K.M.: Approximate projectors in singular spectrum analysis. SIAM J. Matrix Anal. Appl **24**(4), 932–942 (2003). electronic
41. Moskvina, V., Zhigljavsky, A.: An algorithm based on singular spectrum analysis for change-point detection. Commun. Stat. Simul. Comput. **32**(2), 319–352 (2003)
42. Nehari, Z.: On bounded bilinear forms. Ann. of Math. (2) **65**, 153–162 (1957)
43. Pardalos, P.M., Coleman, T.F. (eds.) Lectures on global optimization, Fields Institute Communications, vol. 55. American Mathematical Society, Providence, RI (2009). In: Papers from the Workshop on Global Optimization: Methods and Applications held in Toronto, ON, May 11–12, 2007
44. Pardalos, P.M., Romeijn, H.E. (eds.): Handbook of global optimization. Vol. 2, Nonconvex Optimization and its Applications, Vol. 62. Kluwer Academic Publishers, Dordrecht (2002)
45. Park, H., Zhang, L., Rosen, J.B.: Low rank approximation of a Hankel matrix by structured total least norm. BIT **39**(4), 757–779 (1999)

46. Patterson, K., Hassani, H., Heravi, S., Zhigljavsky, A.: Multivariate singular spectrum analysis for forecasting revisions to real-time data. J. Appl. Stat. **38**(10), 2183–2211 (2011)
47. Peller, V.V.: An excursion into the theory of Hankel operators. In: Holomorphic spaces (Berkeley, CA, 1995), Math. Sci. Res. Inst. Publ., vol. 33, pp. 65–120. Cambridge University Press, Cambridge (1998)
48. Peller, V.V.: Hankel operators and their applications. Springer Monographs in Mathematics. Springer, New York (2003)
49. Pintér, J.: Global Optimization in Action, Nonconvex Optimization and its Applications, vol. 6. Kluwer Academic Publishers, Dordrecht (1996). Continuous and Lipschitz Optimization: Algorithms, Implementations and Applications
50. Power, S.C.: Hankel operators on Hilbert space, Research Notes in Mathematics, vol. 64. Pitman (Advanced Publishing Program), Boston, Mass (1982)
51. Salamon, P., Sibani, P., Frost, R.: Facts, Conjectures, and Improvements for Simulated Annealing. SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2002)
52. Ugray, Z., Lasdon, L., Plummer, J., Glover, F., Kelly, J., Martí, R.: Scatter search and local NLP solvers: A multistart framework for global optimization. INFORMS J. Comput. **19**(3), 328–340 (2007)
53. Varadi, F., Ulrich, R.K., Bertello, L., Henney, C.J.: Searching for signal in noise by random-lag singular spectrum analysis. Astrophys. J. **526**, 1052–1061 (1999)
54. Vautard, R., Ghil, M.: Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. Physica D Nonlinear Phenom. **35**(3), 395–424 (1989)
55. Vautard, R., Yiou, P., Ghil, M.: Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. Physica D: Nonlinear Phenom. **58**(1–4), 95–126 (1992)
56. Vose, M.D.: The Simple Genetic Algorithm. Complex Adaptive Systems. MIT Press, Cambridge, MA (1999). Foundations and theory, A Bradford Book
57. Yiou, P., Sornette, D., Ghil, M.: Data-adaptive wavelets and multi-scale singular-spectrum analysis. Physica D **142**(3–4), 254–290 (2000)
58. Zhigljavsky, A.: Singular spectrum analysis for time series. In: Lovric M. (ed.) International Encyclopedia of Statistical Science, pp. 1335–1337. Springer (2011). See also http://www.cf.ac.uk/maths/subsites/zhigljavskyaa/r_ssa.html
59. Zhigljavsky, A., Žilinskas, A.: Stochastic Global Optimization, Springer Optimization and its Applications, vol. 9. Springer, New York (2008)