APPLIED AND COMPUTATIONAL MATHEMATICS

Original article
DOI: https://doi.org/10.18721/JPM.18114

A SINGULAR SPECTRUM ANALYSIS: PROBLEMS AND SOLUTIONS

Yu. A. Pichugin✉, *N. Yu. Pichugina*

Saint-Petersburg State University of Aerospace Instrumentation, St. Petersburg, Russia

✉ yury-pichugin@mail.ru

Abstract. The paper has examined various aspects of the singular spectrum analysis (SSA) method. The original purpose of the method is to optimize harmonic (spectral) analysis of time series. A theoretical justification based on integral equations was given for the SSA. A solution to the problem of large dimensionality of the autocovariance matrix (AM) was proposed. A method for estimating the periods of AM eigenvectors considered as basis time functions (BTF) was put forward. The problems of repetitive frequencies and frequency overlaps in the BTF structure were discussed. Those overlaps could have lower or higher frequences than the main BTF one. A possibility of distortion of the BTF period estimate due to relatively high-frequency superpositions was noted. To solve this problem, a modification of SSA (SSA-HC) has been proposed, that allows for a more correct and effective solution to the main problem. The solutions to problems were supported by numerical examples.

Keywords: singular spectrum analysis, time series, covariance estimation, dimensionality problem, repeated frequencies

Citation: Pichugin Yu. A., Pichugina N. Yu., A singular spectrum analysis: problems and solutions, St. Petersburg State Polytechnical University Journal. Physics and Mathematics. 18 (1) (2025) 163–181. DOI: https://doi.org/10.18721/JPM.18114

Научная статья
УДК 519.2-519.6
DOI: https://doi.org/10.18721/JPM.18114

АНАЛИЗ СИНГУЛЯРНОГО СПЕКТРА: ПРОБЛЕМЫ И РЕШЕНИЯ

Ю. А. Пичугин✉, *Н. Ю. Пичугина*

Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, Россия

✉ yury-pichugin@mail.ru

Аннотация. Статья посвящена различным аспектам метода анализа сингулярного спектра (SSA). Изначальная цель метода – оптимизировать гармонический спектральный анализ временных рядов. Дано теоретическое обоснование SSA на основе интегральных уравнений. Рассмотрен вопрос адекватного оценивания автоковариаций исследуемого ряда. Предложено решение проблемы большой размерности автоковариационной матрицы (АМ). Выдвинут метод оценки периодов собственных векторов АМ, рассматриваемых как базисные функции времени (BTF). Рассмотрены проблемы повторяющихся частот и частотных наложений в структуре BTF. Эти наложения могут иметь как более низкую, так и более высокую частоту, по сравнению с основной частотой BTF. Отмечена возможность искажения оценки периода BTF из-за относительно высокочастотных наложений. Для решения проблемы предложена модификация SSA (SSA-HC), позволяющая результативнее и корректнее решать основную задачу. Решения проблем подкреплены численными примерами.

Ключевые слова: анализ сингулярного спектра, временной ряд, оценивание ковариаций, проблема размерности, повторяющиеся частоты

Ссылка для цитирования: Пичугин Ю. А., Пичугина Н. Ю. Анализ сингулярного спектра: проблемы и решения // Научно-технические ведомости СПбГПУ. Физико-математические науки. 2025. Т. 18. № 1. С. 163–181. DOI: https://doi.org/10.18721/JPM.18114

Introduction

As soon as it emerged, Singular Spectrum Analysis (SSA) immediately gained great popularity in studies of various time series describing diverse processes (natural, economic, social, etc.). SSA is an application of the principal component method (PCA)) to analysis of time series, consisting of the following stages.

Stage I The matrix of autocorrelations (or autocovariances) is estimated. The first phase of the estimation can be carried out by different methods: either classical (see, for example, monograph [1]) or the caterpillar method (see, for example, monograph [2]). The term *caterpillar method* is used for simplicity here and below. In the case of estimation used in classical time series analysis, the dimension of the autocorrelation matrix is $N \times N$, where N is the length of the initial series. In the caterpillar method, a certain segment of a fixed-length series (the caterpillar) is taken as the implementation of a random vector, and the total set of implementations that make up the sample matrix is obtained by stepwise shifts.

Stage II. Eigenvalues and eigenvectors of the autocorrelation (or autocovariance) matrix are calculated. Due to symmetry and positive definiteness of the autocorrelation and autocovariance matrices, their eigenvalues coincide with the singular values. This is why the method came to be known as singular spectrum analysis.

In this article, we use the abbreviation SSA and continue to use the term *eigenvalues.*

Stage III. The eigenvectors corresponding to the main significant part of the spectrum (the latter is determined by testing the spectrum or subjectively) are used as a basis for smoothing the initial series with a minimum RMS deviation, which is provided by PCA. Here, the expansion coefficients based on the eigenvectors of the autocovariance (autocorrelation) matrix are the principal components.

Another approach to the third and final stage is determining the periods of the eigenvectors (see below), when the principal components basis of the time series becomes the main object for analysis. In this case, it is required to find the periods that are carriers of a certain largest share of the total variance of the series. This share is determined, as noted above, based on statistical tests and criteria that evaluate a significant part of the spectrum, or subjectively. The latter makes SSA a very effective alternative to standard spectral analysis, given its apparent redundancy. Recall that spectral analysis deals with estimation of all possible frequencies (periodicities) contained in a time series, without considering problems of optimal choice. This circumstance was the reason for involving algebraic methods, such as the PCA.

Thus, SSA is essentially the principal component method, solving the problem of optimizing harmonic spectral analysis. The term *principal components of time series* can be found in some works using the caterpillar method [2, 3], which, in our opinion, better reflects the essence of the method.

Notably, the caterpillar SSA method is much more popular [2–5] than the one based on classical estimation of autocovariances. While the recommended dimension of the autocorrelation matrix in the caterpillar method does not exceed half the length of the initial series and can vary (which is undoubtedly attractive), the above-mentioned circumstance is likely due to the fact that the stationarity condition for the given series is not strictly necessary for using this method. In the classical SSA method, stationarity is an indispensable condition, and the problem of high dimensionality, as discussed below, is solved using the appropriate calculation methods.

It should be specifically noted that the interest in using certain research methods is currently largely determined by the availability of software tools that implement these methods.

The objective of this paper was to consider all the problems arising in the implementation of classical SSA, proposing approaches to solving them directly in numerical algorithms.

This should open up opportunities to create appropriate software products for using SSA in various studies.

The computational aspects of the method are mainly considered, but the general idea may undoubtedly be of interest to readers getting acquainted with SSA for the first time.

Theoretical foundations of SSA

Consider the following homogeneous integral equation:

$$\int_{-\infty}^{\infty} e^{-p|x-y|} f(y)dy = \lambda f(x), \tag{1}$$

where the parameter $p > 0$; λ is some unknown number; $f(x)$ is an unknown function that is a solution to this problem.

Eq. (1) is a particular case of the homogeneous Fredholm equation. The coefficient λ is customarily written before the integral, but here we should emphasize that λ is the eigenvalue of the integral operator on the left-hand side of Eq. (1). Eq. (1) in this form can be corresponded with a discrete matrix equivalent:

$$\mathbf{RX} = \lambda \mathbf{X}, \tag{2}$$

where $\mathbf{R}$ is a matrix with elements $r_{ij} = R(i, j) = e^{-p|i-j|}$; $\mathbf{X}$ is an unknown nonzero vector; λ is an unknown eigenvalue of the matrix $\mathbf{R}$ (as in Eq. (1)).

The vector $\mathbf{X}$ satisfying Eq. (2) is an eigenvector of the matrix $\mathbf{R}$, while the function $f(x)$ satisfying Eq. (1) is an eigenfunction of an integral operator with a correlation kernel.

$$R(x, y) = e^{-p|x-y|}$$

(see below).

It is easy to prove that a function of the form

$$f_1(x) = \sin qx \text{ or } f_2(x) = \cos qx,$$

where $q > 0$ is the solution of Eq. (1).

Indeed,

$$\int_{-\infty}^{\infty} e^{-p|x-y|}\sin qy dy = \int_{0}^{\infty} e^{-pt}\sin q(x+t)dt +$$

$$+\int_{0}^{\infty} e^{-pt}\sin q(x-t)dt = 2\sin qx\int_{0}^{\infty} e^{-pt}\cos qt dt,$$

where $t = |x - y|$.

Calculating the last integral or considering the table of Laplace transforms, we obtain:

$$\int_{0}^{\infty} e^{-pt}\cos qt dt = \frac{p}{p^2+q^2}.$$

Ultimately, we have an equation of the form

$$\int_{-\infty}^{\infty} e^{-p|x-y|}\sin qy dy = \frac{2p}{p^2+q^2}\sin qx.$$

Similarly, we obtain

$$\int_{-\infty}^{\infty} e^{-p|x-y|}\cos qy dy = \frac{2p}{p^2+q^2}\cos qx,$$

that is, the following expression holds true in both cases:

$$\lambda = \frac{2p}{p^2+q^2}. \tag{3}$$

It follows from Eq. (3) that as the frequency q increases, the eigenvalue λ corresponding to this frequency decreases. In practical problems, this is the corresponding proportion of the total variance (see below).

It follows from the linearity of the definite integral that functions of the form

$$f(x) = \alpha\sin qx + \beta\cos qx = A\sin(qx+\varphi) \tag{4}$$

can be solutions of Eq. (1), i.e., eigenfunctions of the integral operator on the left-hand side of this equation, for any values of the coefficients α and β, and, accordingly, for any values of amplitude $A = \sqrt{\alpha^2+\beta^2}$ and phase $\varphi = \arcsin\left(\beta\Big/\sqrt{\alpha^2+\beta^2}\right)$.

As a result, it can be seen that solutions (2), i.e., the eigenvectors of the matrix **R**, turn out to be discrete equivalents of harmonics, but the total number of these solutions does not exceed the dimension of the matrix **R**

It is known from works on time series analysis (see monograph [1]) that the function of discrete arguments

$$R(i,j) = e^{-p|i-j|}. \tag{5}$$

with the appropriately selected value of the parameter p is an autocorrelation function of a stationary time series belonging to a class of discrete processes that can be described by a first-order autoregression model.

Importantly, the eigenvectors of the sample autocorrelation matrix of the stationary series should be discrete equivalents of harmonics or be sufficiently close to harmonics. The deviations of the sample autocorrelation function from the ideal case expressed by Eq. (5) may potentially lead to deviations of eigenvectors of the autocorrelation matrix from discrete equivalents of harmonics (this statement is illustrated below, with a method proposed for solving this problem).

It should be borne in mind that in the case of stationary time series, the autocovariance function $C(i,j)$ differs from the autocorrelation function only by a factor σ^2, i.e., follows the expression

$$C(i,j) = \sigma^2 R(i,j) \tag{6}$$

(σ^2 is the variance of time series terms), therefore, the covariance matrix **C** differs from the autocorrelation matrix **R** only by the same factor, i.e., $\mathbf{C} = \sigma^2\mathbf{R}$.

It follows that these matrices have the same set of normalized eigenvectors with the accuracy up to the sign.

It is known from linear algebra that the eigenvectors of a real symmetric matrix are mutually orthogonal and, provided that they have different eigenvalues, form the basis of the space R^N. On the other hand, each ith value of each of the normalized eigenvectors of the autocovariance matrix **C** corresponds to a certain time instant τ_i, therefore, the eigenvectors of the matrix **C** are regarded as discrete basis time functions (BTFs):

$$\psi_j(\tau),\ j = 1, 2, \ldots, N$$

(the term 'discrete' is omitted from now on).

Note that each BTF serves as a carrier of the corresponding fraction of the total variance ($\sigma^2 N$) of a given time series. Another name has been established in the scientific literature: empirical orthogonal functions (EOFs). However, as we establish below, it is preferable to dismiss the indispensable requirement for strict mutual orthogonality, since this has a significant positive effect.

Estimation of covariances

We assume that the initial data are a time series $\{y_i\}_{i=1}^{N}$, or, in vector-matrix form,

$$\mathbf{Y} = (y_1, y_2, ..., y_N)^T$$

(T is the transpose operator).

Here it is necessary to calculate estimates of covariances making up the matrix **C**. Several studies (see, for example, monograph [6]) propose to use an unbiased estimate of the form

$$\hat{c}_{ij} = \hat{c}(k) = \frac{1}{N-k}\sum_{t=1}^{N-k}\left(y_t - \bar{y}\right)\left(y_{t+k} - \bar{y}\right), \tag{7}$$

where $k = |i - j|$; $\bar{y} = \frac{1}{N}\sum_{t=1}^{N} y_t$; $i, j = 1, 2, ..., N$.

However, the above-mentioned (and very popular) monograph on time series analysis [1] suggests using biased estimates of the form

$$\tilde{c}_{ij} = \tilde{c}(k) = \frac{1}{N}\sum_{t=1}^{N-k}\left(y_t - \bar{y}\right)\left(y_{t+k} - \bar{y}\right). \tag{8}$$

As reasoning, Jenkins and Watts considered the continuous process of first-order autoregression $Y(\tau)$, where τ is continuous time. The autocovariance function of this process is a continuous equivalent of expression (6) taking into account Eq. (5) as a function of the magnitude of the shift u, and estimates similar to estimates (7) and (8) calculated in the time interval $[0, T]$, respectively, have the following form:

$$\hat{c}(u) = \frac{1}{T-u}\int_0^{T-u}\left(Y(\tau) - \bar{Y}\right)\left(Y(\tau+u) - \bar{Y}\right)d\tau, \tag{7a}$$

$$\tilde{c}(u) = \frac{1}{T}\int_0^{T-u}\left(Y(\tau) - \bar{Y}\right)\left(Y(\tau+u) - \bar{Y}\right)d\tau, \tag{8a}$$

where $0 \le u \le T$, $\bar{Y} = \frac{1}{T}\int_0^T Y(\tau)d\tau$.

Notably, both positive and negative values of the delay u were considered in the right-hand-sides of Eqs. (7a) and (8a) in monograph [1]. Therefore, the authors used the absolute value of $|u|$ instead of u in the right-hand sides of these formulas.

Calculating the variances of estimates (7a) and (8a), Jenkins and Watts found that the variance of the estimate $\tilde{c}(u)$ (see Eq. (8a)) tends to zero at $u \to$ T, while the variance of the estimate $\hat{c}(u)$ (see Eq. (7a)) not only does not decrease but it also tends to infinity in the case of an unbounded interval.

In view of this, using estimate (7) in the discrete case naturally leads to modulo values that are overestimated compared to their expected value.

Another important aspect is that the covariance matrix of a random vector is always positive definite, or, in the case when the distribution is degenerate and there are zeros in the matrix spectrum, it is semi-definite. The same is true for the estimates of these matrices, and for the estimates

obtained using the caterpillar method in SSA. It was found in our work [7] that the application of $\hat{c}_{ij}$ estimates (7) to the data of climatic temperature series led to the loss of positive definiteness of autocovariance matrices. Negative values appeared in the spectra of these matrices. Therefore, we have another strong argument for using estimates $\tilde{c}_{ij}$ (8).

Calculation of eigenvalues and eigenvectors of autocovariance matrix and solution of high-dimensional problem

After estimating the autocovariances, we are no longer dealing with the initial series, but with its autocovariance matrix. A factor considerably reducing interest in SSA is the prospect of rotating (resulting in a diagonal form) a matrix with the dimension $N \times N$ for calculating its spectrum and eigenvectors (see above). However, using the von Mises iteration method [8], as we establish in this section, completely solves the problem of high dimensionality and allows SSA to be applied to analysis of time series of arbitrary length. This is ensured by the fact that all information about the autocovariance matrix is contained in its first row.

Before explaining the essence of the algorithm and the method for solving the high-dimensional problem, we recall that the Euclidean norm of the vector $\mathbf{X} = (x_1, x_2, ..., x_N)^T$ is determined by the equality

$$\|\mathbf{X}\| = \sqrt{\sum_{i=1}^{N} x_i^2}.$$

To avoid additional notations for variables, we use the symbol := (assignment of values). We predefine the calculation accuracy $\varepsilon > 0$, and the number of calculated eigenvectors m is assumed to be zero ($m = 0$). The subscript is assigned to indicate the number of the eigenvector or the number of the component, and the superscript (0, n, $n + 1$, see below) is to indicate the iteration number in the cyclic sequence of the algorithm.

Algorithm 1. *Step 0* (preliminary sequence of the algorithm (n:= 0).

Assign arbitrary values to the components of the vector $\mathbf{X}^0$; Normalize

$$\mathbf{X}^0 := \left\|\mathbf{X}^0\right\|^{-1} \mathbf{X}^0. \tag{9}$$

If all components of $\mathbf{X}^0$ are assigned $1/\sqrt{N}$ values, then normalization (9) is not required.

Step 1 (cyclic sequence of the algorithm, $n \geq 0$).

Multiplication by the matrix $\tilde{\mathbf{C}}$:

$$\mathbf{X}^{n+1} := \tilde{\mathbf{C}}\mathbf{X}^n. \tag{10}$$

Step 2. Calculate an approximation to the eigenvalue:

$$l^n := \left(\tilde{\mathbf{C}}\mathbf{X}^n\right)^T \mathbf{X}^n = \left(\mathbf{X}^{n+1}\right)^T \mathbf{X}^n. \tag{11}$$

The value of l^n calculated by Eq. (11) is known as the Rayleigh quotient. Here, division by $(\mathbf{X}^n)^T\mathbf{X}^n$ is absent, because due to normalization, the equality $(\mathbf{X}^n)^T\mathbf{X}^n = 1$ holds true.

Step 3. Normalize the new value of the eigenvector approximation:

$$\mathbf{X}^{n+1} := \left\|\mathbf{X}^{n+1}\right\|^{-1} \mathbf{X}^{n+1}. \tag{12}$$

Step 4. Calculate the cosine of the angle φ_n between successive approximations of the eigenvector:

$$\cos\varphi_n := \left(\mathbf{X}^{n+1}\right)^T \mathbf{X}^n. \tag{13}$$

Step 5. Verify the condition for completing the cyclic sequence:

$$\cos\varphi_n \geq 1-\varepsilon. \tag{14}$$

If condition (14) is not satisfied, we return to step 1 (calculation of Eq. (10)). If condition (14) is satisfied, we increase the number of eigenvectors $m := m + 1$ and complete the calculation of the eigenvector and eigenvalue $\mathbf{X}_m := \mathbf{X}^{n+1}$, $\lambda_m := l^n$.

End of Algorithm 1.

Apparently, we encounter high dimensionality mainly in the operation described by Eq. (10), where the approximation of the eigenvector is multiplied by the matrix $\tilde{\mathbf{C}}$ of dimension $N \times N$. However, as noted above, all the elements $\tilde{c}_{ij}$ of the matrix $\tilde{\mathbf{C}}$ are elements of the set of covariance estimates (see Eq. (8)). In particular, the set of elements in the first row of the matrix $\tilde{\mathbf{C}}$ is

$$\{\tilde{c}_{1j} = \tilde{c}\left(j-1\right), j=1,2,\ldots,N\},$$

and all the other rows consist of the same elements.

For this reason, there is no need to store the entire matrix $\tilde{\mathbf{C}}$ in RAM, and the components of the vector $\mathbf{X}^{n+1}$ in Eq. (10) can be calculated as follows [9]:

$$x_1^{n+1} := \sum_{j=1}^{N} \tilde{c}(j-1)x_j^n;\ x_N^{n+1} := \sum_{j=1}^{N} \tilde{c}(N-j)x_j^n;$$

$$x_i^{n+1} := \sum_{j=1}^{i} \tilde{c}(i-j)x_j^n + \sum_{j=i+1}^{N} \tilde{c}(j-i)x_j^n \ (i=2,3,\ldots,N-1).$$

Here, the subscript is the number of the corresponding vector component.

If we use the absolute value of the difference between i and j, i.e. $|i-j|$, a single formula is sufficient, convenient for software implementation:

$$x_i^{n+1} := \sum_{j=1}^{N} \tilde{c}(|i-j|)x_j^n \quad (i=1,2,\ldots,N). \tag{10a}$$

Analyzing Algorithm 1 using Eq. (10a), we can conclude that it completely solves the high dimensional problem.

Remark 1. It is easy to show (and this is a crucial point) that Algorithm 1 gives the maximum eigenvalue as λ_1.

Applying Algorithm 1 to calculate the following eigenvalues and eigenvectors of the autocovariance matrix $\tilde{\mathbf{C}}$ requires performing Gram–Schmidt orthogonalization. So, if m normalized eigenvectors $\{\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_m\}$ are calculated and the next vector needs to be calculated, then the following operations must be performed:

$$\mathbf{X}^0 := \mathbf{X}^0 - \sum_{j=1}^{m} (\mathbf{X}_j^T \mathbf{X}^0)\mathbf{X}_j, \tag{15}$$

$$\mathbf{X}^{n+1} := \mathbf{X}^{n+1} - \sum_{j=1}^{m} (\mathbf{X}_j^T \mathbf{X}^{n+1})\mathbf{X}_j. \tag{16}$$

The approximation of the vector $\mathbf{X}^0$(15) should be orthogonalized immediately before normalization (9), and the vector $\mathbf{X}^{n+1}$ (16) should be orthogonalized before normalization (12). In the complete absence of computational error, orthogonalization (15) is sufficient. However, computational errors are unavoidable, at least due to rounding. Therefore, orthogonalization (16) prevents these errors to divert successive approximations of the next eigenvector into the space of eigenvectors already calculated.

The problem of high dimensionality arises again due to the set of vectors $\{\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_m\}$, which, as m grows from zero to N, makes it necessary to store a large amount of information ($m \times N$ numbers).

Since RAM is sufficiently large in modern computers, this becomes less of an issue. In the past, calculated eigenvectors could be immediately written to the computer's memory (to disk) and extracted (read) from the computer's memory (from disk) sequentially to implement the orthogonalization procedure (see Eqs. (15) and (16)), which is feasible in the software intended for solving this problem.

Thus, the problem of high dimensionality in SSA of relatively long time series is completely solved.

Estimation of period of basis time functions (BTFs)

Technically, estimation of the BTF period is not difficult. Let $\mathbf{X} = (x_1,x_2,...,x_N)^T$ be one of the calculated eigenvectors of the matrix $\tilde{C}$, i.e., a BTF (see above). The set of BTF zeros is denoted as $\{z_1,z_2,...,z_p\}$. To calculate the BTF period, we use the following simple algorithm.

Algorithm 2. *Step 0.* We first assign the value 0 to the number of zeros, i.e., $p:= 0$.

Step 1. Let the number j (the number of the component of the vector $\mathbf{X}$) run through the values from 1 to N.

If $x_j = 0$ or $x_j x_{j+1} < 0$, then $p:= p + 1$, $z_p:= j$.

Step 2. We assign the value z_1 ($j:= z_1$) to the number j.

If $|x_j| \le |x_{j+1}|$, then $z_1:= j + |x_j| \big/ \left(|x_j| + |x_{j+1}|\right)$.

On the other hand, if $|x_j| > |x_{j+1}|$, then $z_1:= j + 1 - |x_{j+1}| \big/ \left(|x_j| + |x_{j+1}|\right)$.

Step 3. We repeat step 2, replacing z_1 with z_p.

Step 4. We calculate the period by the formula

$$T = \frac{2}{p-1}\sum_{i=1}^{p-1}\left(z_{i+1} - z_i\right) = \frac{2\left(z_p - z_1\right)}{p-1}. \tag{17}$$

End of Algorithm 2.

It is easy to see that step 1 of Algorithm 2 defines a set of points preceding the zeros of a BTF (continuous, i.e., imaginary), among which zeros may be also found with a very low probability. As follows from step 4, the period is calculated using only the first and last elements of this ordered set (z_1 and z_p). Therefore, these values are refined in the second and third steps.

The estimation of the period (17) in our work [10] was used without the second and third steps of Algorithm 2 (without refining the z_1 and z_p values). The idea of this approach is that the errors in determining the z_1 and z_p values obtained in step 1 and appearing in Eq. (17) are usually significantly lower than the difference (see Eq. (17)) between these values. However, since the sine wave is close to a straight line in the vicinity of zero, the z_1 and z_p values can be refined, which is carried out in the second and third steps of Algorithm 2.

In addition to the period estimate (17), [10] also used an estimate based not on zeros, but on extrema. The same as in Algorithm 2, first the number of extrema was assumed to be q: = 0 and then the set of extremum points $\{u_1,u_2,...,u_q\}$ was calculated by the rule:

$$(j: = 1, 2, \ldots, N);\ \text{if } |x_{j-1}| \le |x_j| \ge |x_{j+1}|, \text{ then } q: = q + 1,\ u_q: = j.$$

The period value calculated using a similar formula:

$$T' = \frac{2\left(u_q - u_1\right)}{q-1}. \tag{17a}$$

The u_1 and u_q values in this formula do not need to be refined due to the auxiliary role of the estimate T': the explicit inequality $T > T'$ can serve as an indicator of the overlap of relatively high frequency in the BTF structure, as discussed in the next section.

It should be noted here that at this stage we completed the description of SSA, which includes estimates of autocovariances, calculation of eigenvectors of the autocovariance matrix and estimates of periods of eigenvectors considered as BTF.

Next, we consider the problems whose solution leads to a significant modification of the SSA.

Problem of repetitive frequencies and frequency overlaps

The method described above was first applied in our earlier paper [10]. Fig. 1 from that paper graphically represents the first four BTFs of the series of average annual values of surface air temperature in St. Petersburg/Petrograd/Leningrad/St. Petersburg for 120 years (1881–2000). Evidently, these BTFs only resemble harmonics, since they do not have a constant amplitude and period.

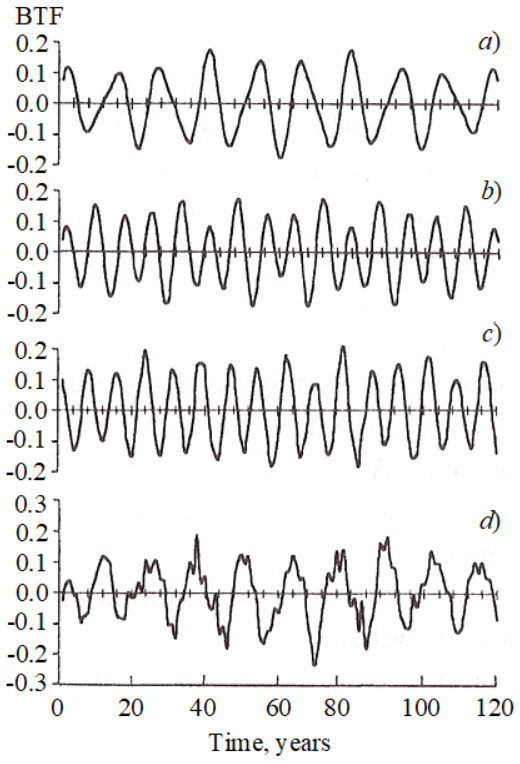

Fig. 1. First 4 basis time functions of average annual temperature in St. Petersburg/Petrograd/Leningrad/St. Petersburg for 120 years (1881–2000), obtained by the SSA method (Example 1, see [10])

Since this phenomenon often occurs in real observational data, climatologists avoid the term *periodicity*, replacing it with the term *cyclicity*. Ref. [10] did not go into the mathematical details of the method, i.e., did not elaborate on the technique for calculating eigenvectors (BFT) and did not contain Eq. (10a), but it did detect two problems.

The first (less significant) problem is that we can obtain the same frequency (the same period) twice due to the mutual orthogonality of the harmonics, which are shifted relative to each other by a quarter of the period as functions of $\sin\tau$ and $\cos\tau$. This problem is solved relatively simply by summing the eigenvalues (i.e., variances) of this frequency. For example, the periods of the second and third BTF are sufficiently close:

$$\mathbf{X}_2\,(T = 7.86 \text{ years}, T' = 7.80 \text{ years}),$$

$$\mathbf{X}_3\,(T = 7.73 \text{ years}, T' = 7.79 \text{ years}).$$

We explain why small discrepancies in the period can be ignored and such cases can be considered repetitions in the section "Numerical examples...".

The second problem is the presence of two types of frequency overlaps in the BTF structure: relatively low-frequency and relatively high-frequency ones. Fig. 1, *a*–*c* shows overlaps of relatively low frequencies, which lead to instability of the amplitude, and, generally speaking, do not complicate the estimation of the period ($T \approx T'$), which was proved above in the example for $\mathbf{X}_2$ and $\mathbf{X}_3$. An overlap of relatively high frequencies ($T > T'$: $T = 10.17$ years, $T' = 3.90$ years) is observed in Fig. 1,*d*, showing the vector $\mathbf{X}_4$. This gives the ramps of the harmonics a serrated appearance. It is easy to see that if the serrations are located near the intersection with the axis, additional zeros may appear, distorting the estimate (17), i.e., the estimate of the main period T will turn out to be erroneous. The term 'noise' is ill-suited for describing this situation for two reasons. Firstly, we need to combine both cases, taking into account that relatively low-frequency overlaps are not commonly called noise, and secondly, noise describes a parasitic component of the initial series, whereas we are dealing with BTF.

To solve this problem, it was proposed in [10] to apply this method (Algorithm 1) to the BTFs themselves, thus isolating the main oscillation. This approach is clearly difficult to implement in a single software product, as it necessitates a ramified algorithm. For this reason, another solution to this problem is proposed here.

Solution of problems of repetitive frequencies and frequency overlaps. Head of Cabbage method

After estimating the covariances, we calculate only the first BTF($\mathbf{X}_1$) and the estimate of the period T ($T_1 := T$). Next, we subtract this periodic component from the initial centered series $\mathbf{Y}_0$ ($\mathbf{Y}_0 = \mathbf{Y} - \bar{y}\mathbf{1}$, all elements of vector $\mathbf{1}$ are equal to unity) using a formula similar to the orthogonalization formulas (see Eqs. (15) and (16)):

$$\mathbf{Y}_1 = \mathbf{Y}_0 - \alpha \mathbf{X}_s\left(T_1\right) - \beta \mathbf{X}_c\left(T_1\right), \tag{18}$$

where the components of the vectors $\mathbf{X}_s(T_1)$ and $\mathbf{X}_c(T_1)$, respectively, are expressed as

$$x_{sj} = \sin\left(\frac{2\pi\left(j-1\right)}{T_1}\right),\ x_{cj} = \cos\left(\frac{2\pi\left(j-1\right)}{T_1}\right),\ j=1,2,\ldots,N,$$

In general, the vectors $\mathbf{X}_s(T_1)$ and $\mathbf{X}_c(T_1)$ are not strictly orthogonal to each other. The latter occurs only in the case when the half-cycle $T_1/2$ divides the length of the series N without a remainder. Therefore, we compose a matrix $\mathbf{X}(T_1)$ of dimension $N \times 2$ from the columns (vectors) $\mathbf{X}_s(T_1)$ and $\mathbf{X}_c(T_1)$, taking the following form:

$$\mathbf{X}\left(T_1\right) = \left(\mathbf{X}_s\left(T_1\right), \mathbf{X}_c\left(T_1\right)\right).$$

Then the values of the coefficients α and β in expression (18) are obtained by the formula

$$\left(\alpha, \beta\right)^T = \left(\mathbf{X}^T\left(T_1\right)\mathbf{X}\left(T_1\right)\right)^{-1}\mathbf{X}^T\left(T_1\right)\mathbf{Y}_0.$$

The variance explained by this periodicity T_1 is estimated as the difference between the values

$$\Delta\mathrm{var}\left(\mathbf{Y}_1\right) = \mathrm{var}\left(\mathbf{Y}_0\right) - \mathrm{var}\left(\mathbf{Y}_1\right) = \left(\tilde{n}_0\left(0\right) - \tilde{n}_1\left(0\right)\right)N, \tag{19}$$

where $\tilde{c}_0(0)$ is the estimate of the variance σ^2 of the initial terms in the series; $\tilde{c}_1(0)$ is the variance estimate of the terms in the series obtained after applying Eq. (18).

In other words, this periodicity T_1 is the carrier of the variance $\Delta\mathrm{var}(\mathbf{Y}_1)$. Apparently, the inequality $\Delta\mathrm{var}(\mathbf{Y}_1) \geq \lambda_1$ must be hold true for the value obtained by Eq. (19), due to the fact that performing procedure (18) solves the problem of repetitive frequencies. This procedure is undoubtedly more effective than directly subtracting the projection $\mathbf{Y}_0$ on $\mathbf{X}_1$, which, taking into account the normalization of $\mathbf{X}_1$, has the form

$$\mathbf{Y}_1 = \mathbf{Y}_0 - \left(\mathbf{Y}_0^T\mathbf{X}_1\right)\mathbf{X}_1 = \mathbf{Y}_0 - \sqrt{\lambda_1}\mathbf{X}_1. \tag{20}$$

In the latter case,

$$\mathbf{Y}_1^T\mathbf{Y}_1 = \mathbf{Y}_0^T\mathbf{Y}_0 - \lambda_1 = \tilde{c}\left(0\right)N - \lambda_1$$

(see Eq. (23) below).

The advantage of the new approach is easier to explain in terms of linear algebra. We subtract the orthogonal projection of the vector $\mathbf{Y}_0$ onto the linear space generated by the vector $\mathbf{X}_1$ in Eqs. (20). However, if we ignore frequency overlaps, then, due to the fact that T_1 is a period of $\mathbf{X}_1$, we can reproduce $\mathbf{X}_1$ as a linear combination of vectors $\mathbf{X}_s(T_1)$ and $\mathbf{X}_c(T_1)$, i.e.,

$$\mathbf{X}_1 \in \Lambda\left(\left\{\mathbf{X}_s\left(T_1\right), \mathbf{X}_c\left(T_1\right)\right\}\right),$$

or

$$\Lambda\left(\left\{\mathbf{X}_1\right\}\right) \subset \Lambda\left(\left\{\mathbf{X}_s\left(T_1\right), \mathbf{X}_c\left(T_1\right)\right\}\right), \tag{21}$$

where $\Lambda(\{*\})$ is the linear shell of the set $\{*\}$.

It follows that the orthogonal projection of the centered series

$$\left(\text{vector } \mathbf{Y}_0 \text{ on } \Lambda\left(\left\{\mathbf{X}_s\left(T_1\right), \mathbf{X}_c\left(T_1\right)\right\}\right)\right.,$$

subtracted in Eq. (18) cannot be less than the orthogonal projection of $\mathbf{Y}_0$ onto the space generated by $\mathbf{X}_1$, i.e., $\Delta\mathrm{var}(\mathbf{Y}_1)$ cannot be less than $_1$ ($\Delta\mathrm{var}(\mathbf{Y}_1) \geq \lambda_1$).

In addition, Eq. (18) is more correct than Eq. (20), since harmonics are subtracted rather than their approximate similarity. We emphasize this major point more than once below.

Next, we apply a certain sequence of operations to the series $\mathbf{Y}_1$ obtained by Eq. (18), i.e., the following algorithm.

Algorithm 3 (modified SSA).

Step 1. Estimate autocovariances (using Eq. (8)).

Step 2. Calculate only the first BTF (by Algorithm 1).

Step 3. Estimate the BTF period (by Algorithm 2);

Step 4. Extract the obtained periodicity (using Eq. (18)) and estimate the proportion of its variance (Eq. (19)).

End of Algorithm 3.

Continuing this process (repeating Algorithm 3), we obtain a sequence of values of the period T_j and the variance Δvar(yj), $\Delta\text{var}(\mathbf{Y}_j)$, $j = 1, 2, \ldots$, by separating the initial time series by periods (frequencies), like peeling leaves from a head of cabbage.

The accumulated variance of the extracted periodicities at the jth step follows the expression (see Eq. (19) for variance):

$$\text{Total var}_j = \sum_{l=1}^{j} \Delta\text{var}\left(\mathbf{Y}_l\right) = \left(\tilde{c}_0\left(0\right) - \tilde{c}_j\left(0\right)\right) N.$$

Next, we will numerical examples to illustrate the effect of the modified SSA method, supplemented by the Head-of-Cabbage method (SSA–HC).

Numerical examples and further interpretation of the SSA–HC method

Example 1. As a first example, let us consider the same series of average annual values of the surface temperature in St. Petersburg/Petrograd/Leningrad/St. Petersburg for 120 years (1881–2000). However, in contrast to [10], where the initial data were provided directly by climatologists of the Voeikov Main Geophysical Observatory (St. Petersburg, Russia), this numerical example uses information from the website [11], where rounded climatic data are presented (to one tenth of a degree).

Table 1 shows the first fifteen steps of applying SSA–HC to analyze the above-mentioned annual average values of surface temperature. In addition, the table includes the eigenvalues $\lambda_1(j)$, corresponding to the eigenvectors $\mathbf{X}_1(j)$ (BTF), calculated in the second step of Algorithm 3 and in the jth step (iteration) of the SSA–HC algorithm. These values are given as percentages of the total variance of the initial time series $\mathbf{Y}_0$, i.e. according by the formula

$$\lambda_1\left(j\right) := \frac{\lambda_1\left(j\right)\cdot 100\,\%}{\text{tr}\tilde{\mathbf{C}}_0} = \frac{\lambda_1\left(j\right)\cdot 100\,\%}{\tilde{c}_0\left(0\right)\cdot N},$$

where $\text{tr}\tilde{\mathbf{C}}_0$ is the trace of the autocovariance matrix of the initial time series ($\mathbf{Y}_0$); $\tilde{c}_0(0)$ is its diagonal element equal to the variance of the series term; $N = 120$.

All other values related to variance are also transformed. Summing up the 15 values from the third column of Table 1, we obtain a value of 33.888%, which is slightly higher than half the value of Total var_{15} (66.733%, see Table 1), and this is explained by the relation

$$\Lambda\left(\left\{\mathbf{X}_1\left(j\right)\right\}_{j=1}^{k}\right) \subset \Lambda\left(\left\{\mathbf{X}_s\left(T_j\right), \mathbf{X}_c\left(T_j\right)\right\}_{j=1}^{k}\right), \tag{22}$$

here j is the step of SSA–HC, $k = 15$.

This relation (22) follows from expression (21), which holds true for all $j \geq 2$.

Naturally, the graph of the first BTF (Fig. 2,*a*) did not change when the SSA–HC method was used (see Fig. 1,*a*), i.e., the rounding of the data from [11] did not have a significant effect.

We should note an important aspect. Only overlaps of relatively low-frequency oscillations are observed in the structure of the BTFs $\mathbf{X}_1(j)$ obtained by the SSA–HC method and shown in Fig. 2, as well as others not shown there. Unlike overlaps of relatively high-frequency oscillations, the overlaps of relatively low-frequency oscillations do not interfere with adequate estimation of the period used in SSA–HC (see Eq. (18)).

Table 1

Application of SSA–HC method to time series of average annual values of surface temperature in St. Petersburg/ Petrograd/Leningrad/St. Petersburg over 120 years (Example 1)

J	T_j	$\lambda_1(j)$	$\Delta\text{var}(\mathbf{Y}_j)$	Total var_j
	years	%		
1	**13.012**	3.915	7.912	7.912
2	**7.752**	3.641	8.163	16.075
3	2.249	3.304	7.556	23.631
4	4.963	2.761	4.939	28.571
5	2.673	2.652	5.270	33.841
6	5.761	2.571	4.950	38.791
7	**2.364**	2.469	4.788	43.578
8	3.401	2.256	3.865	47.443
9	**4.729**	2.092	3.529	50.973
10	3.687	2.061	3.215	54.187
11	11.167	2.006	2.794	56.982
12	2.052	1.930	2.384	59.366
13	20.599	1.816	2.714	62.080
14	8.799	1.794	2.336	64.416
15	4.416	1.619	2.318	66.733

Notations: SSA–HC is singular spectrum analysis modified by Head-of-Cabbage method; j is the SSA–HC step (the number of iterations of Algorithm is 3); T_j is the period of $\mathbf{X}_1(j)$, i.e., the BTF obtained at the jth step; $\lambda_1(j)$ is the eigenvalue corresponding to the vector $\mathbf{X}_1(j)$ (given as a percentage of the total variance of the initial time series $\mathbf{Y}_0$); $\Delta\text{var}(\mathbf{Y}_j)$ is the variance characterized by the period T_j; Total var_j is the accumulated variance of the periodicities extracted in j steps.

Notes. 1. The time series covers the period from 1881 to 2000 ($N = 120$). 2. The first 15 steps of applying the SSA–HC method are presented. 3. The values of T_j repeating the results from [10] are highlighted in bold.

To compare the SSA–HC and SSA methods, Table 2 shows the results of applying SSA to the same series of temperature data (see Table 1) [11].

The fundamental difference between Tables 2 and 1 is that the values of $\lambda_1(j)$ in Table 1 were given only to illustrate Eq. (22) implicitly indicating better efficiency of the SSA–HC method compared with SSA. In the case when SSA is applied, the role of $\Delta\text{var}(\mathbf{Y}_j)$ is played by the value of λ_j itself as the variance determined by the period of fluctuations of the temperature T_j. Therefore, the values of the accumulated variance Total var_j are obtained by summing the values of λ_i ($i = 2, 1, \ldots, j$). At the 15th step of applying SSA, the accumulated variance is 49.387% of the total variance of the series, which is 17.346% less than in the case of SSA–HC (see Table 1).

However, we should note another interesting aspect. Analyzing the data in Table 2, we can see that with standard SSA, the share of the variance $\lambda_1 = 3.915\%$ is explained by the period of 13 years (the exact value is 13.012 years). We can see below in Table 2 that almost the same period of 13 years (this value is 12.890 years in Table 2) explains the variance $\lambda_4 = 3.873\%$. The value of 7.787% is obtained by summation. The value $\text{var}(\mathbf{Y}_1) = 7.912\%$ is also explained by the period of 13 years in Table 1 presenting the results of applying SSA–HC. The small relative discrepancy between these variances, equal to 1.605%, can only be explained by the fact that relatively low-frequency overlaps (see Fig. 2,*a*) affecting the amplitude, which are not excluded in SSA, have a negative effect:

$$7.912\% \approx 7.787\%, \text{ however, } 7.912\% > 7.787\%.$$

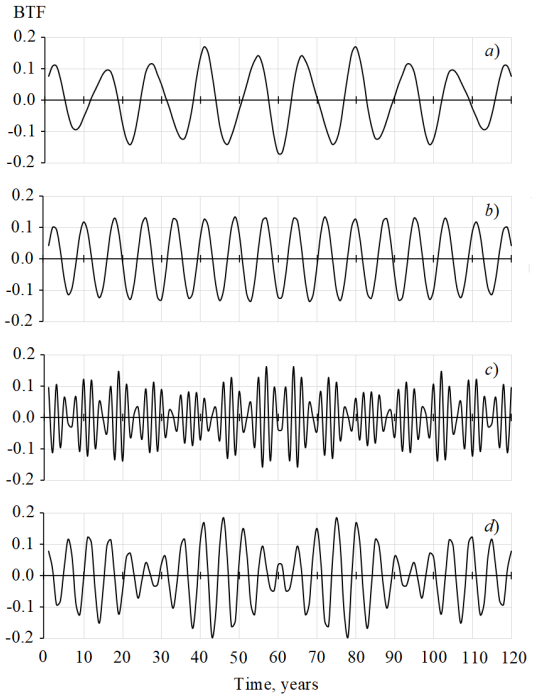

Fig. 2. Graphs of 4 BTFs, similar to those shown in Fig. 1, but obtained by the SSA–HC method

Recall that in SSA–HC we extract 'pure' harmonics from a series, rather than their approximate similarity, and estimate the variance explained by them.

Notice the absence of such close values to the period of 13.012 years (repetitions) in Table 1.

Here, as promised above, we will explain why the close values of the period in Table 2 can be considered a repetition, for example, 13.012 and 12.890 years (relative difference of 0.946%). The series considered is a discrete representation of a continuous process whose spectrum is also continuous. This is implicitly reflected in Algorithm 2. By extracting a harmonic with the period T from a series, which is in fact done in SSA–HC, we bring the variance of the periodicities from the entire interval $(T - \delta;\ T + \delta)$ substantially close to zero for a certain number $\delta > 0$ such that the ratio δ/T is small. Therefore, the values of the period, which are close to 13.012 years, may occur in SSA-HC with only a negligibly small variance if Table 2 is continued. Apparently, the mutual orthogonality of the BTFs with close values of the periods obtained by standard SSA occurs not due to this small discrepancy in the period magnitude, but due to a mutual phase shift of about a quarter of the period, as is the case with the functions $\sin\tau$ and $\cos\tau$.

Table 2

Application of SSA method to time series of average annual values of surface temperature in St. Petersburg/ Petrograd/Leningrad/St. Petersburg for 120 years (Example 1)

j	T_j	$\lambda_1(j)$	Total var$_j$
	years	%	
1	13.012	3.915	3.915
2	7.786	3.741	7.656
3	2.249	3.292	10.948
4	12.890	3.873	14.821
5	7.754	3.744	18.565
6	2.256	3.241	21.806
7	2.995	3.096	24.902
8	2.801	3.107	28.009
9	2.473	3.139	31.148
10	3.886	3.106	34.254
11	3.231	3.006	37.261
12	4.015	3.247	40.508
13	3.790	3.056	43.564
14	3.097	2.980	46.545
15	3.950	2.842	49.387

Notes. 1. The notations are identical to those used in Table 1. 2. The time series covers the period from 1881 to 2000 ($N = 120$). 3. The first 15 steps of applying the SSA method are presented. 4. The initial data is taken from website [11].

A similar situation occurs with close values for the period of 7.786 and 7.754 years (see Table 2, the relative difference is 0.413%). Summing the variances explained by these periods, we obtain a value lower than the estimated periodicity variance of 7,752 years (see Table 1) obtained by the SSA–HC method:

$$3.741\% + 3.744\% = 7.485\% < 8.163\%.$$

These results raise some doubts about the optimality of standard SSA.

Before considering the following example, let us explain in more detail the essence of the difference between the SSA–HC method and SSA.

It is easy to verify that the centered initial time series in vector-matrix form in SSA can be represented by the formula

$$\mathbf{Y}_0 = \hat{\mathbf{Y}}_k + \boldsymbol{\varepsilon} = \sum_{j=1}^{k} \sqrt{\lambda_j}\mathbf{X}_j + \boldsymbol{\varepsilon}, \tag{23}$$

where $\left\{\lambda_j, \mathbf{X}_j\right\}_{j=1}^{k}$ are the eigenvalues and normalized eigenvectors of the autocovariance matrix of the initial series ($k < N$); $\boldsymbol{\varepsilon}$ is the error vector of representation (23).

In terms of time functions, Eq. (23) takes the form

$$y_0(\tau_i) = \sum_{j=1}^{k} \sqrt{\lambda_j}\psi_j(\tau_i) + \varepsilon(\tau_i),\ i = 1,2,\ldots,N, \tag{23a}$$

where $\left\{\psi_j(\tau)\right\}_{j=1}^{k}$ are the functions $\left\{\mathbf{X}_j\right\}_{j=1}^{k}$ considered as BTFs (see the section "Theoretical foundations of SSA).

In this case, $\hat{\mathbf{Y}}_k$ is an orthogonal projection of $\mathbf{Y}_0$ onto the space generated by mutually orthogonal eigenvectors of the matrix $\tilde{\mathbf{C}}_0$, i.e., onto $\Lambda\left(\left\{\mathbf{X}_j\right\}_{j=1}^{k}\right)$. The orthogonality of the projection follows from the fact that the PCA method (PCA) was formulated based on the ordinary least squares method (OLS), see our paper [12]. Therefore, it should provide minimization of the residual sum $\sum_{i=1}^{N}\varepsilon^2(\tau_i)$. However, if PCA is applied to stationary time series, i.e, in the case of SSA, the situation is deferent due to the possible repetition of frequencies (periods), which was illustrated by the example of the temperature series.

In SSA–HC, instead of Eq. (23), we have a representation of the form

$$\mathbf{Y}_0 = \tilde{\mathbf{Y}}_k + \boldsymbol{\delta} = \sum_{j=1}^{k}\left\{\alpha_j\mathbf{X}_s(T_j) + \beta_j\,\mathbf{X}_c(T_j)\right\} + \boldsymbol{\delta}, \tag{24}$$

or, in terms of time functions,

$$y_0(\tau_i) = \sum_{j=1}^{k} A_j \sin\left(2\pi\tau_i/T_j + \varphi_j\right) + \delta(\tau_i), \tag{24a}$$

where the amplitude A_j and phase φ_j are calculated based on the values of α_j и β_j, as in Eq. (4), and the period T_j is determined from the form of the vector $\mathbf{X}_1(j)$, considered as a BTF, following Algorithm 2.

Thus, using the SSA–HC method, we apply PCA method (i.e., essentially the OLS method) step-by-step to find the oscillation mode of the maximum variance $\mathbf{X}_1(j)$ among the remaining possible ones (see Algorithm 1 and Remark 1). Next, after determining the period T_j of this BTF ($\mathbf{X}_1(j)$, see Algorithm 2), we perform the transformation (18), i.e., proceed to harmonic expansion (see Algorithm 3).

The projection $\tilde{\mathbf{Y}}_k$ onto $\Lambda\left(\left\{\mathbf{X}_s\left(T_j\right), \mathbf{X}_c\left(T_j\right)\right\}_{j=1}^{k}\right)$ is not strictly orthogonal because it was obtained step-by-step. This means that the combination of sequential SSA steps with harmonic analysis in the SSA–HC method leads to rejection of strict requirements for orthogonality. To prove the greater efficiency of SSA–HC compared with SSA, we cannot use Eq. (22) as a basis, since

$$\Lambda\left(\left\{\mathbf{X}_j\right\}_{j=1}^{k}\right) \neq \Lambda\left(\left\{\mathbf{X}_1\left(j\right)\right\}_{j=1}^{k}\right),$$

since $\left\{\mathbf{X}_j\right\}_{j=1}^{k}$ is the set of eigenvectors $\tilde{\mathbf{C}}_0$, and $\left\{\mathbf{X}_1\left(j\right)\right\}_{j=1}^{k}$ is the set of the first eigenvectors (see Algorithm 1 and Remark 1) of the sequence of matricesц $\tilde{\mathbf{C}}_{j-1}$ (j = 1, 2, ..., k).

Nevertheless, we can expect SSA–HC to be more efficient than SSA (a comparison of the data in Tables 1 and 2 confirms this), which, provided that k is fixed, can be expressed by the inequality

$$\sum_{i=1}^{N} \delta^2\left(\tau_i\right) \leq \sum_{i=1}^{N} \varepsilon^2\left(\tau_i\right),$$

or

$$\left\|\tilde{\mathbf{Y}}_k\right\| \geq \left\|\hat{\mathbf{Y}}_k\right\|. \tag{25}$$

by virtue of the fact that $\tilde{\mathbf{Y}}_k$ is a projection (even though it is not strictly orthogonal) onto a space of significantly higher dimension, i.e.,

$$\dim\Lambda\left(\left\{\mathbf{X}_s\left(T_j\right), \mathbf{X}_c\left(T_j\right)\right\}_{j=1}^{k}\right) > \dim\Lambda\left(\left\{\mathbf{X}_j\right\}_{j=1}^{k}\right).$$

Squaring inequality (25), we obtain the corresponding inequality for the total variances of these components of the initial series.

Once again, we emphasize that the representation of time series (24) is not only more efficient than (23), but also more correct with respect to harmonic spectral analysis, since in (24) we represent the series by harmonics rather than by their approximate similarity.

Example 2. As a second example, let us consider a series (N = 488) of pressure values obtained at the bottom (behind the ledge) of the experimental flume at the Hydraulic Laboratory of the Faculty of Civil Engineering at St. Petersburg State Polytechnic University (now Peter the Great St. Petersburg Polytechnic University). The details of the experiment can be found in [13]. The values were kindly provided by the author of paper [13]. They are given in arbitrary units (in millimeters of water). Example 2 uses values with a step of 0.32 s. Table 3 shows the first 15 steps of applying the SSA–HC method to this series, where all values are presented as percentages of the variance of the initial series.

Here, as follows from the fifth column in Table 3, the 15 steps of applying the SSA–HC method gain 44.522% of the total variance (see Total var_{15}). However, since the length of the initial series is N = 488, this value demonstrates that the relative variance accumulation rate is 2.7 times higher than in the case of a series of annual average values of surface temperature (see Example 1).

Fig. 3 shows the first 4 BTFs calculated by the SSA–HC method for a series of pressures at the bottom of the experimental flume. Recall that these BTFs are eigenvectors of a sequence of autocovariance matrices.

Remark 2. It should be noted that the average value of the remaining series may deviate from zero in both positive and negative directions for subtractionn of harmonics (see Eq. (18)). In addition, a slight trend may appear. In view of this, it is necessary to perform additional centering of the series after procedure (18) and subsequent extraction of the linear trend. We should note here that the extraction of the linear trend should also be carried out during the initial centering. When the trend is extracted, the variance value will naturally change. For example, the initial series of

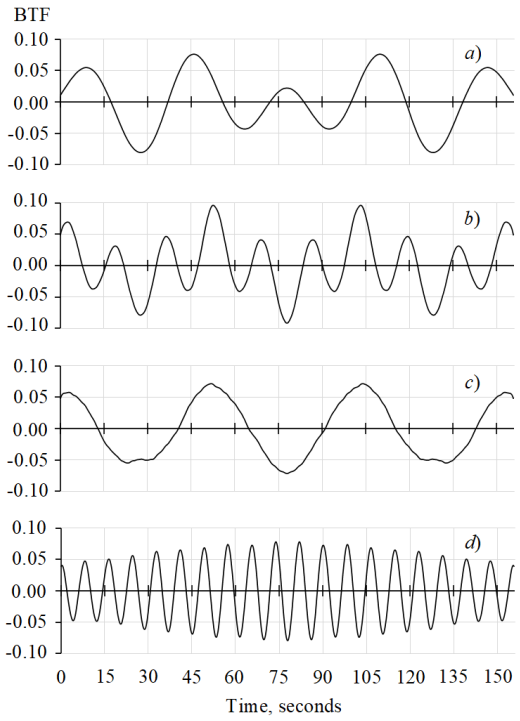

Fig. 3. First 4 BTFs for series of pressures at the bottom of experimental flume, obtained by the SSA–HC method
The initial data were obtained in [13]

pressure values at the bottom of the experimental flume contained a linear trend with a coefficient of $5.625 \cdot 10^{-3}$ arb. units·s^{-1}, and extraction of this trend reduced the variance by 0.03%. Extracting a harmonic with a period of 51.987 s (see Table 3) gave a mean value of $4.975 \cdot 10^{-3}$ arb. units and the linear trend coefficient of $-1.111 \cdot 10^{-2}$ arb. units·s^{-1}, while recentering and extraction of the trend reduced the variance by only 0.116%. In other cases, such corrections reduced the variance by less than a thousandth of a percent.

Fig. 4 shows the initial fragments of empirical autocorrelation functions

$$r(\tau) = c(\tau)/\sigma^2 = c(\tau)/c(0)$$

(see sections "Theoretical basis of SSA" and "Estimation of covariances") of the initial series for water pressure and the series obtained after 15 steps of applying the SSA–HC method. Evidently, the autocorrelation of the time series noticeably decreased with the extraction of 15 harmonics of the corresponding periods estimated in the SSA–HC process.

The data in Tables 1–3 serve for illustrative purposes; if necessary, they can be extended to the required level of accumulated variance.

Table 3

Application of SSA–HC method to series of water pressure values at the bottom of experimental flume (Example 2)

j	T_j	$\lambda_1(j)$	$\Delta\mathrm{var}(\mathbf{Y}_j)$	Total var$_j$
	seconds	%		
1	34.599	2.635	6.303	6.303
2	16.563	1.912	4.815	11.118
3	51.987	1.793	5.364	16.482
4	8.210	1.548	4.017	20.499
5	2.708	1.091	2.764	23.262
6	24.725	0.952	2.554	25.816
7	2.565	0.938	2.259	28.076
8	5.570	0.905	2.188	30.263
9	6.102	0.887	2.432	32.695
10	1.628	0.813	1.974	34.670
11	4.124	0.787	2.136	36.806
12	3.859	0.836	1.987	38.793
13	8.581	0.745	2.262	41.056
14	2.443	0.723	1.856	42.912
15	3.007	0.677	1.611	44.522

Notes. 1. Notations are identical to those used in Tables 1 and 2. 2. The first 15 steps of applying the SSA–HC method are presented. 3. The experimental data were kindly provided by the author of [13].

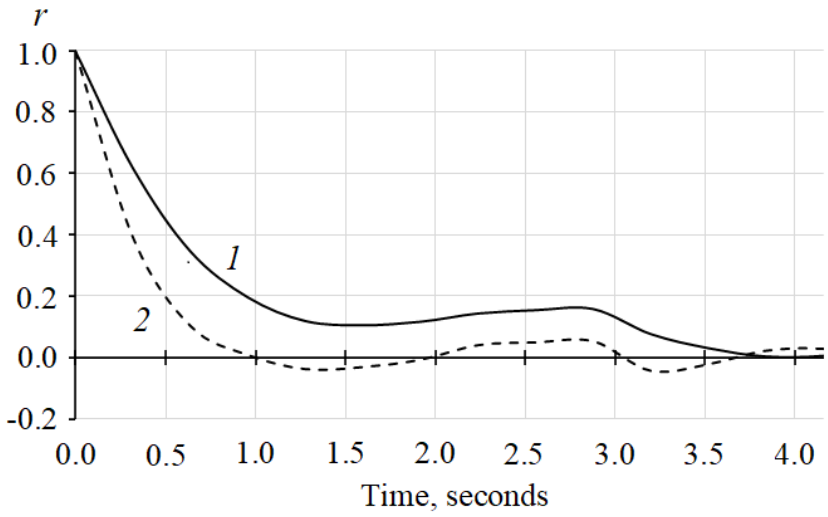

Fig. 4. Initial fragments of empirical autocorrelation functions of initial series of water pressure values (curve *1*) and series obtained after 15 steps of applying SSA–HC method (curve *2*)
The initial data were obtained in [13]

Conclusion

Let us summarize the most important findings of our study.

The application of SSA or SSA–HC methods to analysis of time series (with the exception of caterpillar SSA) requires a biased estimate of autocovariances (8).

Using von Mises iterations with Eq. (10a) solves the problem of high-dimensionality in calculations of eigenvectors of the autocovariance matrix, since this allows applying SSA and SSA–HC to time series of arbitrary length.

The estimate of the BTF period can be improved by refining the first and last zero values.

Both the SSA method and the SSA–HC method solve the problem of optimizing harmonic spectral analysis, but SA–HC provides greater efficiency compared to SSA, solving the problems of repetitive frequencies and frequency overlaps in the BTF structure, which means that it is more accurate.

All the algorithms considered are easily implemented in software products.

Acknowledgment

The authors express their deep gratitude to A.A. Kaverin (Candidate of Physico-Mathematical Sciences, who was junior researcher at the Faculty of Civil Engineering at St. Petersburg State Polytechnic University at the time of the experiment, going on to become Chief Engineer at Triplus Engineering LLC) for the experimental data provided, serving as the second numerical example of SSA–HC application in this study. The authors also express their deep gratitude to M. I. Meshcheryakov (software developer at Technocontent LLC) for his help in implementing SSA–HC in the Python programming language.

REFERENCES

1. **Jenkins G. M., Watts D. G.,** Spectral analysis and its applications, Holden-Day, San Francisco, USA, 1969.

2. **Danilov D. L., Zhiglyavskii A. A. (Eds.),** The principal components of time series: The "Caterpillar method", SPbSU Publishing, St. Petersburg, 1997 (in Russian).

3. **Pichugin Y. A.,** Sampling principal components of moving interval in the analysis of time series of meteorological data, Russian Meteorology and Hydrology. (8) (1999) 22–27.

4. **Awichi R., Müller W. G.,** Improving SSA predictions by inverse distance weighting, REVSTAT Stat. J. 11 (1) (2013) 105–119.

5. **Leles M. C. R., Sansro J. P. H., Mozelli L. A., Guimarres H. N.,** A new algorithm in singular spectrum analysis framework: The Overlap-SSA (ov-SSA), SoftwareX. 8 (July–Dec) (2018) 26–32.

6. **Oppenheim A. V., Schafer R. W.,** Digital signal processing, Pearson, London, 1975.

7. **Pichugin Yu. A.,** Estimation of statistical significance and space heterogeneity of surface air temperature, Proceedings of the Russian Geographical Society. 135 (6) (2003) 78 – 84 (in Russian).

8. **Wilkinson J. H., Reinsch C.,** Handbook for automatic computation, Vol. 2: Linear algebra (Grundlehren Der Mathematischen Wissenschaften, Vol. 186), Springer-Verlag, Berlin-Heidelberg, Germany, 1971.

9. **Pichugin Yu. A.,** Computational features of singular spectrum analysis, Abstracts of the II Int. Forum "Computational and Mathematical Methods and Modelling in High-Tech Manufacturing", Nov. 9, 2022, St. Petersburg State University of Aerospace Instrumentation. (2022) 259–260 (in Russian).

10. **Pichugin Yu. A.,** Iterative singular-spectrum analysis in estimating natural cyclicities in meteorological observation data, Russian Meteorology and Hydrology. (10) (2001) 24–28.

11. Pogoda i klimat. Letopis pogody v Sankt-Peterburge (po online dannym i literaturnym istochnikam) [The weather and climate. The weather annual chronicle in Saint Petersburg (according to online data and literary sources)]. URL: http://www.pogodaiklimat.ru/history/26063.htm (The access data is 15.07.2024).

12. **Pichugin Yu. A.,** Notes on using the principal components in the mathematical simulation, St. Petersburg State Polytechnical University Journal. Physics and Mathematics. 11 (3) (2018) 74–89 (in Russian).

13. **Kaverin A. A.,** Rezultaty eksperimentalnykh issledovaniy granits smeny rezhimov techeniya za ustupom [Results of experimental studies of the boundaries of a flow regime change behind a bench], Magazine of Civil Engineering. (2) (2013) 62–66 (in Russian).

СПИСОК ЛИТЕРАТУРЫ

1. **Дженкинс Г., Ваттс Д.** Спектральный анализ и его приложения. Пер. с англ. В 2-х вып. Выпуск 1. М.: Мир, 1971. 320 с.

2. Главные компоненты временных рядов: метод «Гусеница». Под ред. Д. Л. Данилова, А. А. Жиглявского. СПб.: Изд-во СПбГУ, 1997. 308 с.

3. **Пичугин Ю. А.** Выборочные главные компоненты скользящего отрезка в анализе временных рядов метеорологических данных // Метеорология и гидрология. 1999. № 8. С. 31–36.

4. **Awichi R., Müller W. G.** Improving SSA predictions by inverse distance weighting // REVSTAT – Statistical Journal. 2013. Vol. 11. No. 1. Pp. 105–119.

5. **Leles M. C. R., Sansro J. P. H., Mozelli L. A., Guimarres H. N.** A new algorithm in singular spectrum analysis framework: The Overlap-SSA (ov-SSA) // SoftwareX. 2018. Vol. 8. July–December. Pp. 26–32.

6. **Оппенгейм А. В., Шафер Р. В.** Цифровая обработка сигналов: Пер. с англ. М.: Связь, 1979. 416 с.

7. **Пичугин Ю. А.** Оценка статистической значимости и пространственной неоднородности линейных трендов приземной температуры воздуха // Известия Русского географического общества. 2003. Т. 135. № 6. С. 78–84.

8. **Уилкинсон Дж. Х., Райнш С.** Справочник алгоритмов на языке АЛГОЛ. Линейная алгебра. М.: Машиностроение, 1976. 390 с.

9. **Пичугин Ю. А.** Вычислительные особенности анализа сингулярного спектра // Сборник тезисов докладов II Международного форума «Математические методы и модели в высокотехнологическом производстве». 9 ноября 2022 г. Санкт-Петербург: ГУАП, 2022. С. 259–260.

10. **Пичугин Ю. А.** Итерационный анализ сингулярного спектра в оценке естественных цикличностей данных метеорологических наблюдений // Метеорология и гидрология. 2001. № 10. С. 34–39.

11. Погода и климат. Летопись погоды в Санкт-Петербурге (по online данным и литературным источникам). URL: http://www.pogodaiklimat.ru/history/26063.htm. Дата обращения 15. 07. 2024.

12. **Пичугин Ю. А.** Замечания к использованию главных компонент в математическом моделировании // Научно-технические ведомости СПбГПУ. Физико-математические науки. 2018. Т. 11. № 3. С. 74–89.

13. **Каверин А. А.** Результаты экспериментальных исследований границ смены режимов течения за уступом // Инженерно-строительный журнал. 2013. № 2. С. 62–66.

THE AUTHORS

PICHUGIN Yury A.
Saint-Petersburg State University of Aerospace Instrumentation
61 Bolshaya Morskaya St., St. Petersburg, 190000, Russia
yury-pichugin@mail.ru
ORCID: 0000-0003-0646-364X

PICHUGINA Nika Yu.
Saint-Petersburg State University of Aerospace Instrumentation
61 Bolshaya Morskaya St., St. Petersburg, 190000, Russia
nike.pichugina@gmail.com
ORCID: 0000-0002-9224-8761

СВЕДЕНИЯ ОБ АВТОРАХ

ПИЧУГИН Юрий Александрович – *доктор физико-математических наук, профессор Института инноватики и базовой магистерской подготовки Санкт-Петербургского государственного университета аэрокосмического приборостроения.*
190000, Россия, г. Санкт-Петербург, Большая Морская ул., 61.
yury-pichugin@mail.ru
ORCID: 0000-0003-0646-364X

ПИЧУГИНА Ника Юрьевна – *ассистент Института инноватики и базовой магистерской подготовки Санкт-Петербургского государственного университета аэрокосмического приборостроения.*
190000, Россия, г. Санкт-Петербург, Большая Морская ул., 61.
nike.pichugina@gmail.com
ORCID: 0000-0002-9224-8761

Received 28.08.2024. Approved after reviewing 06.11.2024. Accepted 06.11.2024.
Статья поступила в редакцию 28.08.2024. Одобрена после рецензирования 06.11.2024. Принята 06.11.2024.