The "Caterpillar"-SSA method for analysis of time series with missing values

N. Golyandina [a,*], E. Osipov [a]

[a] *Math. Department, St.Petersburg University, Universitetsky av. 28, 198504, St.Petersburg, Petrodvorets, Russia*

Abstract

The paper concerns the problem of applying Singular Spectrum Analysis to time series with missing data. A method of filling in the missing data is proposed and is applied to time series of finite rank. Conditions of exact reconstruction of missing data are constructed and versions of the algorithm applicable to real-life time series are presented. The proposed algorithms result in extraction of additive components of time series such as trends and periodic components, with simultaneous filling in the missing data. An example is presented.

Key words: time series analysis, missing data, Singular Spectrum Analysis, time series of finite rank.
1991 MSC: 37M10, 62-07, 60G35

1 Introduction

The paper is devoted to adaptation of Singular Spectrum Analysis for analysis of time series with missing data.

Ideas of the SSA method originated in many areas of mathematics and its applications: signal processing, dynamic system analysis, stationary series analysis, signal detection in the presence of red noise, analysis of series governed by linear recurrent formulas, Principle Component Analysis, and the others (see Elsner and Tsonis (1996) and Golyandina et al (2001) for references). Thus very similar algorithms with different names and various methodologies of application based on different theoretical results have been obtained. By the same reason, the different SSA-based methods of forecasting and filling in

* Corresponding author. Email address nina(at)NG1174.spb.edu

missing data differ from each other. For example, the analysis of stationary series by SSA results in the method of filling in described in Schoellhamer (2001); solving the problem of signal detection in the presence of red noise leads to the method outlined in Kondrashov et al (2005).

In this paper we apply and extend the approach described in details in Golyandina et al (2001). In order to emphasize its distinctive features, this approach will be called "Caterpillar"-SSA (the name "Caterpillar" was given in Russia, Danilov and Zhigljavsky (1997)). More information on this approach (papers, examples, software) is available at http://www.gistatgroup.com/cat/.

The idea of filling in missing data is to a great extent similar to the idea of forecasting and, in the framework of the considered approach, consists in continuation of the structure of the extracted component to the gaps caused by the missing data. In the "Caterpillar"-SSA method, it is assumed that the forecasted component is (or is approximated by) a time series of finite rank (see Golyandina et al (2001)) or, what is almost the same, a time series governed by some linear recurrent formula. Thus, it is not surprising that the theoretical results related to exact reconstruction of missing values can be applied only to time series of finite rank.

Similar to the Basic "Caterpillar"-SSA method, the proposed modifications for analysis of time series with missing data give exact results under rather restrictive assumptions. Nevertheless, the constructed algorithms are applicable to real-life time series with missing values; they give approximate results in this case.

Note that in the specific case when missing values are located at the end of the series, the problem of their filling in coincides with the problem of forecasting. Therefore, the developed methods are able both to fill in the missing values and to solve the problem of forecasting. This allows one to look at the problem of forecasting deeper and provides new ways of its solution.

In Section 2 we present preliminary results, which are not directly related to the analysis of time series with missing values. The first two subsections contain results concerning properties of linear subspaces, vectors from these subspaces and also vectors' restrictions onto some fixed set of indices. In the last subsection we introduce the main concepts of the Basic "Caterpillar"-SSA method for the analysis of time series with no missing data.

Section 3 is devoted to application of the results obtained in the previous section to the trajectory subspace and the lagged vectors produced by a series of finite rank with missing data (note that lagged vectors are, in effect, subseries of the observed time series). Thus we obtain conditions and formulas for recovering the missing components of the considered lagged vector and therefore for filling in the missing data of the time series.

In Section 4 we introduce a number of modifications of the "Caterpillar"-SSA method based on the obtained formulas; these modifications allow both extracting the components of real-life time series and filling in the missing data. An example demonstrating the work of the proposed algorithms is presented.

2 Preliminary results

2.1 Recovery of vector's components in a subspace

Let us give necessary notation. Consider the Euclidean space $\mathbb{R}^n$. Define $\mathcal{I} = \{1, \dots, n\}$ and denote by $\mathcal{S} = \{i_1, \dots, i_s\} \subset \mathcal{I}$ an ordered set, $|\mathcal{S}| = s$. Let $\mathbf{E}_s$ denote the unit $s \times s$ matrix.

By definition, *restriction of a vector* $X = (x_1, \dots, x_n)^{\mathrm{T}} \in \mathbb{R}^n$ onto a set of indices $\mathcal{S}$ is the vector $X\Big|_{\mathcal{S}} = (x_{i_1}, \dots, x_{i_{|\mathcal{S}|}})^{\mathrm{T}} \in \mathbb{R}^{|\mathcal{S}|}$.

Restriction of a matrix onto a set of indices is the matrix consisting of restrictions of its column vectors onto this set.

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

Consider an m-dimensional subspace $\mathfrak{D}_m \in \mathbb{R}^n$. Denote by $\{R_k\}_{k=1}^m$ an orthonormal basis of the $\mathfrak{D}_m$ and define the matrix $\mathbf{R} = [R_1 : \ldots : R_m]$. Fix an ordered set of indices $\mathcal{P}$.

Proposition 2.1 *Let the matrix* $\mathbf{E}_{|\mathcal{P}|} - \mathbf{R}\Big|_{\mathcal{P}}\Big(\mathbf{R}\Big|_{\mathcal{P}}\Big)^{\mathrm{T}}$ *be non-singular. Then for any vector* $X \in \mathfrak{D}_m$ *the following formula expressing* $X\Big|_{\mathcal{P}}$ *in terms of* $X\Big|_{\mathcal{I}\setminus\mathcal{P}}$ *holds:*

$$X\Big|_{\mathcal{P}} = \Big(\mathbf{E}_{|\mathcal{P}|} - \mathbf{R}\Big|_{\mathcal{P}}\Big(\mathbf{R}\Big|_{\mathcal{P}}\Big)^{\mathrm{T}}\Big)^{-1}\mathbf{R}\Big|_{\mathcal{P}}\Big(\mathbf{R}\Big|_{\mathcal{I}\setminus\mathcal{P}}\Big)^{\mathrm{T}}X\Big|_{\mathcal{I}\setminus\mathcal{P}}. \tag{1}$$

Proof. For simplicity of notation, let $\mathcal{P} = \{1, \dots, |\mathcal{P}|\}$. Denote $X_1 = X\Big|_{\mathcal{P}}$, $X_2 = X\Big|_{\mathcal{I}\setminus\mathcal{P}}$, $\mathbf{R}_1 = \mathbf{R}\Big|_{\mathcal{P}}$, $\mathbf{R}_2 = \mathbf{R}\Big|_{\mathcal{I}\setminus\mathcal{P}}$. Since $\mathbf{R}\mathbf{R}^{\mathrm{T}}X = X$ for $X \in \mathfrak{D}_m$ and

$$\mathbf{R}\mathbf{R}^{\mathrm{T}} = \begin{pmatrix} \mathbf{R}_1\mathbf{R}_1^{\mathrm{T}} & \mathbf{R}_1\mathbf{R}_2^{\mathrm{T}} \\ \mathbf{R}_2\mathbf{R}_1^{\mathrm{T}} & \mathbf{R}_2\mathbf{R}_2^{\mathrm{T}} \end{pmatrix},$$

we have $X_1 = \mathbf{R}_1\mathbf{R}_1^{\mathrm{T}}X_1 + \mathbf{R}_1\mathbf{R}_2^{\mathrm{T}}X_2$. Turning back to the original notation, we come to the equality

$$\Bigl(\mathbf{E}_{|\mathcal{P}|} - \mathbf{R}\Big|_{\mathcal{P}}\Bigl(\mathbf{R}\Big|_{\mathcal{P}}\Bigr)^{\mathrm{T}}\Bigr)X\Big|_{\mathcal{P}} = \mathbf{R}\Big|_{\mathcal{P}}\Bigl(\mathbf{R}\Big|_{\mathcal{I}\setminus\mathcal{P}}\Bigr)^{\mathrm{T}}X\Big|_{\mathcal{I}\setminus\mathcal{P}}.$$

This completes the proof. □

Lemma 2.1 *The following conditions are equivalent:*

1) $\dim \mathfrak{D}_m\Big|_{\mathcal{I}\setminus\mathcal{P}} = \dim \mathfrak{D}_m$*;*

2) the vectors $\Bigl\{R_k\Big|_{\mathcal{I}\setminus\mathcal{P}}\Bigr\}_{k=1}^{m}$ *are linearly independent and form a basis of the subspace* $\mathfrak{D}_m\Big|_{\mathcal{I}\setminus\mathcal{P}}$*;*

3) $Y\Big|_{\mathcal{I}\setminus\mathcal{P}} \neq 0_{|\mathcal{I}\setminus\mathcal{P}|}$ *for any nonzero vector* $Y \in \mathfrak{D}_m$*;*

4) $\mathrm{span}(e_i,\ i \in \mathcal{P}) \cap \mathfrak{D}_m = \{0_n\}$*;*

5) the matrix $\Bigl(\mathbf{E}_{|\mathcal{P}|} - \mathbf{R}\Big|_{\mathcal{P}}\Bigl(\mathbf{R}\Big|_{\mathcal{P}}\Bigr)^{\mathrm{T}}\Bigr)^{-1}$ *exists.*

Proof. *Equivalence of the conditions 1) – 4)* is evident.

Equivalence of 4) and 5). Proposition 2.1 implies the equivalence of 5) to the following assertion: for any vector $V \in \mathfrak{D}_m\Big|_{\mathcal{I}\setminus\mathcal{P}}$ there exists an unique vector $G \in \mathfrak{D}_m$ such that $V = G\Big|_{\mathcal{I}\setminus\mathcal{P}}$. Let us prove the equivalence of 4) to the same assertion.

Since the vectors $\Bigl\{R_k\Big|_{\mathcal{I}\setminus\mathcal{P}}\Bigr\}_{k=1}^{m}$ span the space $\mathfrak{D}_m\Big|_{\mathcal{I}\setminus\mathcal{P}}$, the vector V can be expressed as $V = \sum\limits_{k=1}^{m} a_k R_k\Big|_{\mathcal{I}\setminus\mathcal{P}}$. Then the required vector is $G = \sum\limits_{k=1}^{m} a_k R_k$. Suppose that there are two different vectors G_1 and G_2 such that $V = G_1\Big|_{\mathcal{I}\setminus\mathcal{P}} = G_2\Big|_{\mathcal{I}\setminus\mathcal{P}}$. Consider their difference $G_1 - G_2 = \sum_{i\in\mathcal{P}} \alpha_i e_i \in \mathfrak{D}_m$. This difference is not equal to the zero vector if and only if $\mathrm{span}(e_i,\ i \in \mathcal{P}) \cap \mathfrak{D}_m = \{0_n\}$. The lemma is proved. □

Remark 2.1 It follows from the item 4) of Lemma 2.1 that $n - m \geq |\mathcal{P}|$. This constraint on the number of vector's missing components is a necessary

condition for applying the formula (1).

Let us consider two special cases, when the first ($\mathcal{P} = \{1\}$) or the last ($\mathcal{P} = \{n\}$) coordinate is expressed through the rest.

Corollary 2.1 *Denote* $\nu^2 = {\pi_1}^2 + \ldots + {\pi_m}^2$, *where* π_i *is the* n*-th component of the vector* R_i, *and* $\{R_i^{\triangledown}\}_{i=1}^m$ *are the vectors* $\{R_i\}_{i=1}^m$ *without the last components (their dimension is equal to* $n-1$*). Suppose that* $e_n \notin \mathfrak{D}_m$ *and* $X = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathfrak{D}_m$. *Then* $\nu^2 < 1$ *and* $x_n = \sum\limits_{k=1}^{n-1} a_k x_{n-k}$, *where*

$$(a_{n-1}, \ldots, a_1)^{\mathrm{T}} = \frac{1}{1-\nu^2} \sum_{i=1}^{m} \pi_i R_i^{\triangledown}. \tag{2}$$

Corollary 2.2 *Denote* $\mu^2 = {\rho_1}^2 + \ldots + {\rho_m}^2$, *where* ρ_i *is the first coordinate of the vector* R_i, *and* $\{R_i^{\triangle}\}_{i=1}^m$ *are the vectors* $\{R_i\}_{i=1}^m$ *without the first components (their dimension is equal to* $n-1$*). Suppose that* $e_1 \notin \mathfrak{D}_m$ *and* $X = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathfrak{D}_m$. *Then* $\mu^2 < 1$ *and* $x_1 = \sum\limits_{k=2}^{n} a_k x_k$, *where*

$$(a_2, \ldots, a_n)^{\mathrm{T}} = \frac{1}{1-\mu^2} \sum_{i=1}^{m} \rho_i R_i^{\triangle}. \tag{3}$$

2.2 Projection operator

Consider the subspaces $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$ of $\mathbb{R}^n$, of dimensions m and $\widetilde{m} \le n-m$ correspondingly. Let us first find the matrix of the operator $\mathbf{\Pi}^{(1)}_{\mathcal{I}\setminus\mathcal{P}}$ corresponding to the orthogonal projection $\mathbb{R}^{|\mathcal{I}\setminus\mathcal{P}|} \to \mathcal{L}^{(1)}\Big|_{\mathcal{I}\setminus\mathcal{P}}$. Denote by $\{R_k\}_{k=1}^m$ an orthonormal basis of the space $\mathcal{L}^{(1)}$, $\mathbf{R} = [R_1 : \ldots : R_m]$. Set $\mathbf{V} = \mathbf{R}\Big|_{\mathcal{I}\setminus\mathcal{P}}$ and $\mathbf{W} = \mathbf{R}\Big|_{\mathcal{P}}$ for convenience of notation.

Proposition 2.2 *Assume that the matrix* $\mathbf{E}_{|\mathcal{P}|} - \mathbf{W}\mathbf{W}^{\mathrm{T}}$ *is nonsingular. Then the matrix* $\Pi^{(1)}_{\mathcal{I}\setminus\mathcal{P}}$ *of the orthogonal projection operator* $\mathbf{\Pi}^{(1)}_{\mathcal{I}\setminus\mathcal{P}}$ *has the form*

$$\Pi^{(1)}_{\mathcal{I}\setminus\mathcal{P}} = \mathbf{V}\mathbf{V}^{\mathrm{T}} + \mathbf{V}\mathbf{W}^{\mathrm{T}}(\mathbf{E}_{|\mathcal{P}|} - \mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{W}\mathbf{V}^{\mathrm{T}}. \tag{4}$$

Proof. Introduce the matrix $\mathbf{A} = \mathbf{V}^{\mathrm{T}}\mathbf{V}$ ($\mathbf{A}$ and $\mathbf{W}^{\mathrm{T}}\mathbf{W}$ are $m \times m$ matrices). According to Lemma 2.1, the matrix $\mathbf{A}$ is nonsingular as a Gram matrix of a linearly independent set of vectors; therefore $\mathbf{A}$ is reversible. It is known that

in this case the operator of orthogonal projection onto the space spanned by the columns of the matrix $\mathbf{V}$ has the form $\Pi^{(1)}_{\mathcal{I}\setminus\mathcal{P}} = \mathbf{V}\mathbf{A}^{-1}\mathbf{V}^{\mathrm{T}}$. The challenge is to find the explicit formula of the matrix $\mathbf{A}^{-1}$.

Since the vectors $\{R_k\}_{k=1}^m$ constitute an orthonormal basis of the space $\mathcal{L}^{(1)}$, we have $\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{E}_m$. On the other hand,

$$\mathbf{R}^{\mathrm{T}}\mathbf{R} = \Big(\mathbf{R}\Big|_{\mathcal{I}\setminus\mathcal{P}}\Big)^{\mathrm{T}}\mathbf{R}\Big|_{\mathcal{I}\setminus\mathcal{P}} + \Big(\mathbf{R}\Big|_{\mathcal{P}}\Big)^{\mathrm{T}}\mathbf{R}\Big|_{\mathcal{P}} = \mathbf{A} + \mathbf{W}^{\mathrm{T}}\mathbf{W}.$$

Thus, $\mathbf{A} = \mathbf{E}_m - \mathbf{W}^{\mathrm{T}}\mathbf{W}$. Straightforward demonstration of

$$(\mathbf{E}_m - \mathbf{W}^{\mathrm{T}}\mathbf{W})(\mathbf{E}_m + \mathbf{W}^{\mathrm{T}}(\mathbf{E}_{|\mathcal{P}|} - \mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{W}) = \mathbf{E}_m$$

completes the proof. □

Denote by $\mathbf{\Pi}^{(1)}$ the operator of orthogonal projection of $\mathbb{R}^n$ onto $\mathcal{L}^{(1)}$. The evident conditions of permutability of the projection and the restriction procedures are given in the following proposition.

Proposition 2.3 *If* $\mathcal{L}^{(1)} \perp \mathcal{L}^{(2)}$ *and* $\mathcal{L}^{(1)}\Big|_{\mathcal{I}\setminus\mathcal{P}} \perp \mathcal{L}^{(2)}\Big|_{\mathcal{I}\setminus\mathcal{P}}$, *then*

$$\mathbf{\Pi}^{(1)}_{\mathcal{I}\setminus\mathcal{P}}\Big(X\Big|_{\mathcal{I}\setminus\mathcal{P}}\Big) = \Big(\mathbf{\Pi}^{(1)}X\Big)\Big|_{\mathcal{I}\setminus\mathcal{P}}$$

for any $X \in \mathcal{L}^{(1)} \oplus \mathcal{L}^{(2)}$.

2.3 Basic concepts of the SSA method

Consider a real-valued time series $F_N = (f_0, \ldots, f_{N-1})$ of length N. Following Golyandina et al (2001), let us outline notions of the SSA method.

Fix a positive integer L, $1 < L < N$, which is called a *window length*. The *embedding procedure* maps the original time series into a sequence of L-dimensional lagged vectors $\{X_i\}_{i=1}^K$, $K = N - L + 1$, by the formula

$$X_i = (f_{i-1}, \ldots, f_{i+L-2})^{\mathrm{T}}, \quad 1 \le i \le K.$$

The *L-trajectory matrix* (or, simply, the *trajectory matrix*) of the series F_N is formed of the lagged vectors: $\mathbf{X} = [X_1 : \ldots : X_K]$.

In other words, the trajectory matrix is

$$\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} f_0 & f_1 & f_2 & \dots & f_{K-1} \\ f_1 & f_2 & f_3 & \dots & f_K \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \dots & f_{N-1} \end{pmatrix}. \tag{5}$$

Obviously, $x_{ij} = f_{i+j-2}$ and the matrix $\mathbf{X}$ has equal values on "diagonals" $i + j =$ Const. Matrices of this type are called *Hankel* matrices.

An arbitrary matrix can be approximated by a Hankel matrix in the space of matrices with the Frobenius norm. This is done by means of the hankelization procedure, which consists in the replacement of values on "diagonals" $i + j =$ Const by their arithmetic average. The transition from a matrix to a time series by means of the hankelization procedure and the subsequent use of the one-to-one correspondence between Hankel matrices and time series (diagonal $\longleftrightarrow$ value of the series, see (5)) is called the *diagonal averaging* procedure.

The linear space $\mathcal{L}^{(L)}$ spanned by the lagged vectors is called an *L-trajectory space of columns* (or, simply, a *trajectory space*) of the series F_N: $\mathcal{L}^{(L)} = \mathcal{L}^{(L)}(F_N) \stackrel{\text{def}}{=} \text{span}(X_1, \dots, X_K)$.

The linear space spanned by the row vectors of the L-trajectory matrix of the series is called an *L-trajectory space of rows* (or, simply a *trajectory space of rows*).

Definition 2.1 Let $0 \le d \le L$. The *series F_N is said to have L-rank d*, if $\dim\mathcal{L}^{(L)} = d$ (briefly, $\text{rank}_L(F_N) = d$). For zero series define $\dim\mathcal{L}^{(L)} = 0$. If $\text{rank}_L(F_N) = d < N/2$ for any L such that $d \le \min(L, K)$, then the *series F_N has rank d* ($\text{rank}(F_N) = d$). If such d exists, the time series F_N is called a *series of finite rank*.

The trajectory space of a series of finite rank d is denoted by $\mathcal{L}_d$.

The eigenvectors U_i, $i = 1, \dots, d = \text{rank}\,\mathbf{X}$, corresponding to d non-zero eigenvalues of the matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ form an orthonormal basis of the trajectory space. The vectors $U_1, \dots, U_d$ are the left singular vectors of the Singular Value Decomposition of the matrix $\mathbf{X}$.

The following proposition shows that the class of time series of finite rank is rather wide.

Proposition 2.4 *A series, which can be represented as a linear combination of products of polynomials, exponents, and cosines, is a series of finite rank.*

The following concept is very important in the SSA theory (its asymptotic analogue is the basis for applying SSA to the analysis of real-life time series).

Definition 2.2 Two series are called *separable* if the trajectory spaces of columns are orthogonal and so are the trajectory spaces of rows.

If two series $F_N^{(1)}$ and $F_N^{(2)}$ are separable, then the basis U_i, $i = 1, \ldots, d$, of the trajectory space of the series $F_N = F_N^{(1)} + F_N^{(2)}$ can be generally split into two parts, one of the parts is the basis of the trajectory space of $F_N^{(1)}$ and the other is the basis of the trajectory space of $F_N^{(2)}$. Thus if the series $F_N^{(1)}$ and $F_N^{(2)}$ are separable, then the SSA method is capable of extracting summands from the observed sum $F_N = F_N^{(1)} + F_N^{(2)}$: the projections of the lagged vectors of the series F_N onto the found trajectory subspaces are the lagged vectors of the series $F_N^{(1)}$ and $F_N^{(2)}$ correspondingly.

3 Lagged vectors and trajectory spaces of time series of finite rank with missing data

3.1 Recovery of missing components of lagged vectors

Consider a time series F_N of finite rank d and fix a window length L, $d < \min(L, K)$, $K = N - L + 1$. Suppose that some data are missing in the observed series F_N (i.e. some data are considered to be unknown), but its L-trajectory space $\mathcal{L}_d \subset \mathbb{R}^L$ is known.

Let us take an incomplete L-lagged vector, which contains both missing and non-missing data of F_N. Denote this vector by X and the ordered set of its missing components' indices by $\mathcal{P}$. From now on, $\mathcal{I}$ stands for the set $\{1, \ldots, L\}$ and the set of indices $\mathcal{I} \setminus \mathcal{P}$ indicates the set of non-missing components of the vector X. The restrictions $X\big|_{\mathcal{P}}$ and $X\big|_{\mathcal{I}\setminus\mathcal{P}}$ are thereby the vectors, which consist of missing and non-missing components of the vector X correspondingly.

To solve the problem of reconstructing the missing components $X\big|_{\mathcal{P}}$ of the vector X by means of the non-missing components $X\big|_{\mathcal{I}\setminus\mathcal{P}}$, we apply the theory of Subsection 2.1 (with $m = d$, $\mathfrak{D}_m = \mathcal{L}_d$) to the vector X. Suppose that the trajectory space $\mathcal{L}_d$ meets the conditions of Proposition 2.1. Then the formula (1) solves the problem of finding $X\big|_{\mathcal{P}}$ in terms of $X\big|_{\mathcal{I}\setminus\mathcal{P}}$ and therefore it restores the missing values of the series F_N that belong to the vector X.

Let us turn to the case when the initial series with missing data is a sum of two separable series of finite rank $F_N^{(1)}$ and $F_N^{(2)}$, i.e. $F_N = F_N^{(1)} + F_N^{(2)}$. Assume that

the trajectory spaces $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$ of the series $F_N^{(1)}$ and $F_N^{(2)}$ correspondingly are given. Since the series are separable, their trajectory spaces are orthogonal: $\mathcal{L}^{(1)} \perp \mathcal{L}^{(2)}$. Again, let X be a lagged vector of the series F_N with indices of missing data from $\mathcal{P}$ and $X = X^{(1)} + X^{(2)}$, where $X^{(i)}$, $i = 1, 2$, are the corresponding lagged vectors of $F_N^{(i)}$. Let us solve the problem of finding the lagged vector $X^{(1)}$ of the first series in terms of $X\Big|_{\mathcal{I}\setminus\mathcal{P}}$.

The problem splits into two: finding $X^{(1)}\Big|_{\mathcal{I}\setminus\mathcal{P}}$ and finding $X^{(1)}\Big|_{\mathcal{P}}$. If the first problem is solved, we can apply the theory of Subsection 2.1 with $m = \operatorname{rank} F_N^{(1)}$ and $\mathfrak{D}_m = \mathcal{L}^{(1)}$ to find $X^{(1)}\Big|_{\mathcal{P}}$ in terms of $X^{(1)}\Big|_{\mathcal{I}\setminus\mathcal{P}}$. The propositions of Section 2.2 provide us solution to the problem of finding $X^{(1)}\Big|_{\mathcal{I}\setminus\mathcal{P}}$ if the condition of orthogonality of the restricted trajectory spaces $\mathcal{L}^{(1)}\Big|_{\mathcal{I}\setminus\mathcal{P}}$ and $\mathcal{L}^{(2)}\Big|_{\mathcal{I}\setminus\mathcal{P}}$ holds. Since the subspaces $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$ are orthogonal, the lagged vector $X^{(1)}$ of the series $F_N^{(1)}$ is equal to $\mathbf{\Pi}^{(1)}X$; therefore the corresponding terms of the series $F_N^{(1)}$ can be obtained by orthogonal projection of the vector $X\Big|_{\mathcal{I}\setminus\mathcal{P}}$ onto $\mathcal{L}^{(1)}\Big|_{\mathcal{I}\setminus\mathcal{P}}$.

Thus, $X^{(1)}\Big|_{\mathcal{I}\setminus\mathcal{P}} = \Pi^{(1)}_{\mathcal{I}\setminus\mathcal{P}} X\Big|_{\mathcal{I}\setminus\mathcal{P}}$, where $\Pi^{(1)}_{\mathcal{I}\setminus\mathcal{P}}$ is defined by formula (4) with the matrix $\mathbf{R}$ whose columns are the vectors of the orthonormal basis of $\mathcal{L}^{(1)}$, and $X^{(1)}\Big|_{\mathcal{P}}$ is expressed in terms of $X^{(1)}\Big|_{\mathcal{I}\setminus\mathcal{P}}$ by the formula (1) with the same matrix $\mathbf{R}$. Thereby, the values of the series $F_N^{(1)}$ belonging to its lagged vector $X^{(1)}$ have been found, including those which are located on the places of missing data of the series F_N.

3.2 Finding trajectory spaces of the initial time series and of its additive components

Consider a time series F_N with $\operatorname{rank}_L(F_N) = d$ and its L-lagged vectors $\{X_i\}_{i=1}^K$. At first, let us obtain conditions of possibility to find the basis of the trajectory space $\mathcal{L}_d = \operatorname{span}(X_i, i = 1, \ldots, K)$ using only non-missing values of the observed series. Denote by $\mathcal{C} \subset \{1, \ldots, K\}$ the set of numbers of the complete lagged vectors with no missing entries. Assume that $\mathcal{C} \neq \emptyset$ and consider the matrix $\widetilde{\mathbf{X}}$ consisting of lagged vectors X_i, $i \in \mathcal{C}$, as its columns. Let $\widetilde{\mathcal{L}}_d = \operatorname{span}(X_i, i \in \mathcal{C})$. It is easy to prove the following proposition.

Proposition 3.1 *The set X_i, $i \in \mathcal{C}$, contains at least d linearly independent vectors if and only if $\widetilde{\mathcal{L}}_d = \mathcal{L}_d$. As this takes place, the eigenvectors $U_1, \ldots, U_d$ of the matrix $\widetilde{\mathbf{S}} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}}$ corresponding to d nonzero eigenvalues of $\widetilde{\mathbf{S}}$ form an orthonormal basis of the subspace $\mathcal{L}_d$.*

Let us formulate a more constructive sufficient condition.

Proposition 3.2 *If the time series F_N has L-rank d, $e_1 \notin \mathcal{L}_d$, $e_L \notin \mathcal{L}_d$ and the series contains at least $L+d-1$ successive non-missing values, then $\widetilde{\mathcal{L}}_d = \mathcal{L}_d$.*

Proof. In order to prove the proposition, let us show that there are d linearly independent vectors among complete lagged vectors. We conclude from the conditions of this proposition that there exists such a number k that lagged vectors $\{X_{k+i}\}_{i=1}^{d}$ do not contain missing data.

Since $e_1 \notin \mathcal{L}_d$ and $e_L \notin \mathcal{L}_d$, the time series F_N can be continued to the infinite in both directions time series F of L-rank d (the proof is based on Corollary 2.1 and Corollary 2.2 with $n = L$, $m = d$, $\mathfrak{D}_m = \mathcal{L}_d$). It follows that F is a time series of finite rank d and, in particular, $\mathrm{rank}_{d+1} F_N = d$, $e_1, e_{d+1} \notin \mathcal{L}^{(d+1)}(F_N)$. Therefore, due to Corollary 2.1 with $n = d+1$, $m = d$, $\mathfrak{D}_m = \mathcal{L}^{(d+1)}(F_N)$,

$$X_{i+d} = \sum_{j=1}^{d} a_j \, X_{i+d-j}, \qquad 1 \le i \le K-d.$$

Hence all lagged vectors with indices exceeding $k+d$ can be expressed as a linear combination of vectors $X_{k+1}, \ldots, X_{k+d}$.

Analogously, taking into account Corollary 2.2, we find that lagged vectors with indices from 1 to k can be written as linear combinations of vectors $X_{k+1}, \ldots, X_{k+d}$. Therefore, all the lagged vectors are expressed through the set of vectors $\{X_{k+i}\}_{i=1}^{d}$. Since dimension of the trajectory space is equal to d, these vectors are independent. □

We now turn to the case when the observed series is a sum of two separable series: $F_N = F_N^{(1)} + F_N^{(2)}$. Assume that the conditions of Proposition 3.1 are met and a basis of the trajectory space $\mathcal{L}_d$ of the series F_N has been found. Let us formulate the problem of finding the trajectory spaces $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$ of the series $F_N^{(1)}$ and $F_N^{(2)}$ using $\widetilde{\mathbf{X}} = \widetilde{\mathbf{X}}^{(1)} + \widetilde{\mathbf{X}}^{(2)}$.

If the row spaces of the trajectory matrices $\widetilde{\mathbf{X}}^{(1)}$ and $\widetilde{\mathbf{X}}^{(2)}$ are orthogonal and so are the column spaces, then the eigenvectors $U_1, \ldots, U_d$ of $\widetilde{\mathbf{S}}$ forming the basis of $\mathcal{L}_d$ can be generally split into two groups constituting bases of the spaces $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$. The trajectory spaces of columns are orthogonal due to the separability of the series $F_N^{(1)}$ and $F_N^{(2)}$. It is easy to show that the following conditions are necessary and sufficient for orthogonality of the row trajectory spaces:

$$\sum_{k \in \mathcal{C}} f_{i+k-1}^{(1)} f_{j+k-1}^{(2)} = 0, \quad i, j = 0, \ldots, L-1. \tag{6}$$

Thus, (6) and the conditions of Proposition 3.1 are sufficient to find bases of the trajectory spaces $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$.

Sum of a constant and a harmonic series of period T can be given as an example. It is known that if L and K are divisible by T, then the series are separable. If T successive values are missing, and N and L are large enough, then it is still possible to find bases of the spaces $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$.

3.3 Conditions for orthogonality of the restricted subspaces

Properties of the operator of orthogonal projection $\mathbf{\Pi}^{(1)}_{\mathcal{I}\setminus\mathcal{P}}$ were obtained in Proposition 2.3 of Section 2.2 (and were used in Section 3.1). In the mentioned proposition, the vector X and the set $\mathcal{P}$ of indices of vector's missing components were considered under the condition that the spaces $\mathcal{L}^{(1)}\big|_{\mathcal{I}\setminus\mathcal{P}}$ and $\mathcal{L}^{(2)}\big|_{\mathcal{I}\setminus\mathcal{P}}$ are orthogonal.

Let $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$ be trajectory spaces of the series $F^{(1)}_N$ and $F^{(2)}_N$. The following equalities give us the necessary and sufficient condition of orthogonality $\mathcal{L}^{(1)}\big|_{\mathcal{I}\setminus\mathcal{P}} \perp \mathcal{L}^{(2)}\big|_{\mathcal{I}\setminus\mathcal{P}}$:

$$\sum_{k\in\mathcal{I}\setminus\mathcal{P}} f^{(1)}_{i+k-1} f^{(2)}_{j+k-1} = 0, \quad i,j = 0,\dots,K-1. \tag{7}$$

Remark 3.1 If the series $F^{(1)}_N$ and $F^{(2)}_N$ are separable, then the condition (7) is equivalent to

$$\sum_{k\in\mathcal{P}} f^{(1)}_{i+k-1} f^{(2)}_{j+k-1} = 0, \quad i,j = 0,\dots,K-1. \tag{8}$$

Sum of a harmonic series of period T and a constant series, where L, K and the number of the successive missing data are divisible by T, is an example when the condition (8) is satisfied. Also, L should be greater than the number of missing values of the time series, and $\mathcal{P}$ specifies the positions of missing data in a lagged vector containing all the missing data.

4 "Caterpillar"-SSA for time series with missing data

4.1 Scheme of the algorithm

The result of the "Caterpillar"-SSA algorithms is the decomposition of the observed time series into additive components such as a trend, periodic components and noise. The algorithm for time series with no missing data consists of two stages: decomposition and reconstruction. Each stage, in its turn, consists of two steps: Embedding and Singular Value Decomposition are the steps of the first stage, Grouping and Diagonal Averaging are the steps of the second stage. General structure of the algorithm for the analysis of time series with missing data is the same, but the steps are somewhat different.

Thus, we have the initial time series $F_N = (f_0, \ldots, f_{N-1})$ consisting of N elements, some part of which is unknown. Let us describe the scheme of the algorithm in the case of reconstruction of the first component $F_N^{(1)}$ of the series, based on the observed sum of two time series: $F_N = F_N^{(1)} + F_N^{(2)}$.

4.1.1 First stage: decomposition

S t e p 1. Embedding
Let us fix the window length L, $1 < L < N$. The embedding procedure transforms the initial time series into the sequence of L-dimensional lagged vectors $\{X_i\}_{i=1}^K$, where $K = N - L + 1$. A part of the lagged vectors may be incomplete, i.e. contain missing components. Matrix $\widetilde{\mathbf{X}}$ is formed of the complete lagged vectors X_i, $i \in \mathcal{C}$, (assume they form a non-empty set) with no missing data; this matrix in the case of absence of missing data coincides with the trajectory matrix of F_N.

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

4.1.2 Second stage: reconstruction

S t e p 3a. Choosing the subspace and projection of the complete lagged vectors
Let a set of indices $I_r = \{i_1, \ldots, i_r\} \subset \{1, \ldots, d\}$ be chosen and the subspace $\mathcal{M}_r = \mathrm{span}(U_{i_1}, \ldots, U_{i_r})$ be formed. The choice of the eigenvectors (i.e., their indices) corresponding to $F_N^{(1)}$ is conducted in a similar manner as it has been

done at the grouping stage in the Basic SSA algorithm (see Golyandina et al (2001) for details). One of the attributes of the eigenvectors to be chosen is the resemblance of their form to the form of $F_N^{(1)}$. The complete lagged vectors can be projected onto the subspace $\mathcal{M}_r$ in the regular way:

$$\widehat{X}_i = \sum_{k \in I_r} (X_i, U_k) U_k, \quad i \in \mathcal{C}. \tag{9}$$

S t e p 3b. Projection of the incomplete lagged vectors
For each $\mathcal{P}$-incomplete lagged vector with missing components in the positions from $\mathcal{P}$ the given step consists of two parts:

(α) calculation of $\widehat{X}_i\big|_{\mathcal{I}\setminus\mathcal{P}}$, $\quad i \notin \mathcal{C}$,

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

Since adjacent lagged vectors have a common information (see (5)), there are many possible ways of solving the formulated problems. Some of these ways will be discussed in the following sections. The common information also enables processing vectors with $\mathcal{P} = \mathcal{I}$. Note that the step 3b may change the vectors $\widehat{X}_i$, $i \in \mathcal{C}$. The result of the steps 3a and 3b is the matrix $\widehat{\mathbf{X}} = [\widehat{X}_1 \ldots : \widehat{X}_K]$, which serves as an approximation to the trajectory matrix of the series $F_N^{(1)}$, under the proper choice of the set I_r.

S t e p 4. Diagonal averaging
At the last step of the algorithm, matrix $\widehat{\mathbf{X}}$ is transformed into the new series $\widetilde{F}_N^{(1)}$ (so called the reconstructed time series) by means of the "diagonal averaging" procedure.

Remark 4.1
If the missing data are located on the right end of the series (up to its edge), the reconstruction of the series $F_N^{(1)}$ with filling in the missing data is in fact the forecasting of the series $F_N^{(1)}$.

4.2 Clusters of missing data

In order to suggest variants of the step 3b for projecting the incomplete vectors, let us give definition of clusters of missing data and their classification under the fixed window length L.

Definition 4.1 A sequence of missing data of a time series is called a *cluster*

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

Definition 4.2 A cluster is called *right/left* if its right/left element is located at the distance less than L from the right/left end of the series. If the left or the right element of the cluster coincides with the end of the series, the cluster is called *extreme*. Neither right nor left cluster is called *inner*. A cluster is called *continuous* if it consists of successive missing data.

Definition 4.3 A set of successive incomplete lagged vectors is called *a lagged-vectors set of a cluster of missing data*, if it consists of all vectors that include missing data from the considered cluster.

Since sets of the lagged vectors of different clusters don't contain any information about each other and are independent from this point of view, we will further consider one cluster of missing data and the corresponding set of the lagged vectors X_i, $l \le i \le p$. The step 3b can be performed independently for each cluster of missing data (for each lagged-vectors set).

4.3 Methods for reconstructing values at the positions of the non-missing components

Here we describe possible solutions to the problem (α) of calculation of $\widehat{X}_i\big|_{\mathcal{I}\setminus\mathcal{P}}$, $l \le i \le p$.

The method when the operator of orthogonal projection onto $\mathcal{M}_r\big|_{\mathcal{I}\setminus\mathcal{P}}$ is applied to $X_i\big|_{\mathcal{I}\setminus\mathcal{P}}$ for each $\mathcal{P}$-incomplete lagged vectors of the cluster of missing data is called *"Π Projector"*. Proposition 2.2 with $m = r$, $\mathcal{L}^{(1)} = \mathcal{M}_r$, and the matrix $\mathbf{R}$, whose columns are the eigenvectors $U_i, i \in I_r$, provides conditions of applying this method and the formula for calculation of $\Pi^{(1)}_{\mathcal{I}\setminus\mathcal{P}}$ in $\widehat{X}_i\big|_{\mathcal{I}\setminus\mathcal{P}} = \Pi^{(1)}_{\mathcal{I}\setminus\mathcal{P}} X_i\big|_{\mathcal{I}\setminus\mathcal{P}}$.

One more method can be applied to continuous (extreme or inner) clusters. This method is based on the fact that values of the diagonal entries of the trajectory matrices (recall that $\widehat{\mathbf{X}}$ is an approximation of the trajectory matrix of the series $F^{(1)}_N$) with indices (i, j), $i + j = \text{Const}$, are equal. Let us consider an inner cluster at first. The positions of the non-missing components $\widehat{X}_q\big|_{\mathcal{I}\setminus\mathcal{P}}$ in the corresponding set of vectors $\widehat{X}_q$, $q = l, \ldots, p$, form two "triangles"

(right and left). Denoting by $\widehat{x}_{i,j}$ the i-th element of the vector $\widehat{X}_j$, we will describe the method by the example of filling in the left "triangle", which will be expressed as the set of the s-th "diagonals": $\{(i,j) : i+j = s,\ l+1 \le s \le l+L-1,\ l \le j \le l+L-2\}$. Note that "triangle" and "diagonal" are interpreted as parts of the matrix $[\widehat{X}_l : \ldots : \widehat{X}_p]$.

Since we deal with the inner cluster, there are l_0, $l_0 \ge 1$, left adjacent vectors $\widehat{X}_{l-m}$, $m = 1, ..., l_0$, which are complete and all their components have already been calculated at the step 3a. If we consider the matrix $[\widehat{X}_{l-l_0} : \ldots : \widehat{X}_p]$, then the method called *"Components of adjacent vectors"* consists in the replacement of all the components of the s-th diagonal $\widehat{x}_{i,j}$ with $i+j = s$ and $l - l_0 \le j \le l+L-2$ by the average value of $\widehat{x}_{i,j}$ with $i+j = s$, $l - l_0 \le j \le l-1$. Note that the method also changes some components of vectors $\widehat{X}_{l-l_0}, \ldots : \widehat{X}_{l-1}$.

A fully similar procedure is carried out with the right "triangle". For extreme clusters, only one of "triangles" needs to be filled.

4.4 Methods for reconstructing values at the positions of missing components

Here we propose several solutions to the problem (β) of calculation of $\widehat{X}_i\big|_{\mathcal{P}}$, $l \le i \le p$, if vectors $\widehat{X}_i\big|_{\mathcal{I}\setminus\mathcal{P}}$ have already been obtained.

For the first method, we will require each vector of the lagged-vectors set to satisfy the conditions of Proposition 2.1. If all the conditions are met, then we can reconstruct missing components of each vector by the formula (1), where $m = r$, $\mathfrak{D}_m = \mathcal{M}_r$ and $\mathbf{R}$ is a matrix whose columns are the eigenvectors $U_i, i \in I_r$. Such a method of reconstructing missing data will be called a *"simultaneous filling in"*. Note that conditions for applying the simultaneous filling in are restrictive enough. In particular, it is impossible to apply the method when there are no non-missing components just in one vector from the lagged-vectors set.

Other methods are based on the fact that values with indices (i,j) on the diagonals $i+j = \text{Const}$ of the trajectory matrices are equal. Therefore, we can reconstruct missing components of one of the lagged vectors and use the obtained values for filling in missing entries of the adjacent vectors.

Consider the lagged-vectors set $X_l, \ldots, X_p$ of the considered cluster of missing values. We assume that the vectors $\widehat{X}_q\big|_{\mathcal{I}\setminus\mathcal{P}}$, $q = l, \ldots, p$, have already been calculated by one of the methods described in Subsection 4.3. Let us describe several variants of the methods of sequential filling in.

1) Let the considered cluster be not a left cluster. In this case the left vector X_l contains only one missing entry, which is located on the place of the last coordinate, so its set of indices of missing entries is $\mathcal{P} = \{L\}$. Therefore Corollary 2.1 can be applied. If $e_L \notin \mathcal{M}_r$, then we can obtain the last value y of the vector $\widehat{X_l}$ as a linear combination of the previous values with coefficients given by the formula (2). Further this value y is used to fill in the $(L-j)$-components of the vectors $\widehat{X}_{l+j}$, $j = 1, \ldots, L-1$. After this procedure the adjacent vector $\widehat{X}_{l+1}$ will have only one missing entry on the place of the last coordinate (if the cluster is continuous). Therefore the whole procedure may be repeated once more. And so on. Such a method will be called a *sequential filling in from the left*. Note that if the cluster of missing data is not continuous, the filling procedure need not be applied to some of the vectors X_q, $l < q < p$.

2) The *sequential filling in from the right* is fully analogous to the sequential filling in from the left and comes down to filling in the first coordinate by means of the formula (3) given in Corollary 2.2. Therefore, the applicable conditions of this method are: the cluster of missing data is not a right cluster and $e_1 \notin \mathcal{M}_r$.

3) Different combination of the sequential filling in from the left and from the right (so called two-sided methods) can be considered.

Note that the conditions of sequential filling in are less restrictive in comparison with simultaneous imputation; however, errors can accumulate.

Remark 4.1 Consider a continuous cluster of missing data of length m, which is a right extreme cluster (and there are no other clusters of missing data in the series). If the methods "Components of adjacent vectors" and "Sequential filling in from the left" are applied to this cluster, the result will coincide with the recurrent forecast (see Golyandina et al (2001)) for m terms ahead; the forecasted time series component is extracted by Basic SSA from the times series consisting of the first $N - m$ points.

4.5 A formal modification of the "Caterpillar"-SSA algorithm for series with missing data

The above-described methods of filling in use a detected structure of the time series. As an addition to them, a formal version of filling in missing data can be proposed. In this modification, the inner product of vectors is replaced by a similar operation, which can be applied to vectors with missing components.

4.5.1 "" operation*

Let us introduce a "*" operation as a formal substitution of the inner product of vectors if they include missing entries. Consider two vectors $A = (a_1, \ldots, a_n)^{\mathrm{T}}$ and $B = (b_1, \ldots, b_n)^{\mathrm{T}}$ and denote the sets of indices of their missing components as $\mathcal{A}$ and $\mathcal{B}$ correspondingly. Suppose that $|\mathcal{A} \cup \mathcal{B}| < n$. Let us define the "*" operation by the formula

$$\Big(A, B\Big)^* = A^{\mathrm{T}} * B = \frac{n}{n - |\mathcal{A} \cup \mathcal{B}|} \sum_{k: k \notin \mathcal{A} \cup \mathcal{B}} a_k b_k.$$

It is clear that if we multiply complete vectors with no missing components, the result of the "*" operation coincides with the result of the inner product in $\mathbb{R}^n$. For matrices $\mathbf{A}^{\mathrm{T}} = [A_1 : \ldots : A_k]$ and $\mathbf{B} = [B_1 : \ldots : B_l]$, $A_i, B_i \in \mathbb{R}^n$, the "*" operation is defined as $\mathbf{A} * \mathbf{B} = \{(A_i, B_j)^*\}_{i=1\,j=1}^{k\quad l}$. If all vectors of both matrices don't contain missing entries, $\mathbf{A} * \mathbf{B} = \mathbf{AB}$.

4.5.2 Modifications using the "" operation*

Let us propose the following modifications.

1. In place of the matrix $\widetilde{\mathbf{S}}$ at the second step of the algorithm described in Section 4.1, let us take the matrix formally calculated by the other formula: $\widetilde{\mathbf{S}} = \mathbf{X} * \mathbf{X}^{\mathrm{T}}$, where the matrix $\mathbf{X}$ is the trajectory matrix of the series F_N consisting of all the lagged vectors, both complete and incomplete.

We can generalize the above proposed method as follows: consider a value τ, $0 \le \tau \le L$, which is called a *threshold of missing components*. Then form the matrix $\widetilde{\mathbf{X}}_{(\tau)}$ from the lagged vectors, which contain not more than τ missing components, and put $\widetilde{\mathbf{S}} = \widetilde{\mathbf{S}}_{(\tau)} = \widetilde{\mathbf{X}}_{(\tau)} * \widetilde{\mathbf{X}}_{(\tau)}^{\mathrm{T}}$. Note that the matrix $\widetilde{\mathbf{X}}_{(0)}$ coincides with $\widetilde{\mathbf{X}}$ consisting of the complete vectors with no missing data, and that $\widetilde{\mathbf{X}}_{(L)} = \mathbf{X}$.

Remark 4.2 The known Toeplitz modification of Basic SSA is used for stationary time series. Toeplitz SSA can be adjusted to time series with missing values in the analogous manner, by substitution of "*" for the inner product: the matrix $\widetilde{\mathbf{S}}$ is equal to $K\widetilde{\mathbf{C}}$, where $\widetilde{\mathbf{C}} = \{\widetilde{c}_{ij}\}$ and for $F^{(k,g)} = (f_k, \ldots, f_g)^{\mathrm{T}}$

$$\widetilde{c}_{ij} = \frac{1}{N - |i-j|} \ \Big(F^{(0, N-|i-j|-1)}, F^{(|i-j|, N-1)}\Big)^*, \quad 1 \le i, j \le L.$$

This way of constructing the matrix $\widetilde{\mathbf{S}}$ can be called *Toeplitz*.

2. At the step 3, all the reconstructed vectors can be calculated like projections (compare with (9)):

$$\widehat{X}_i = \sum_{k \in I_r} (X_i, U_k)^* \, U_k, \quad i = 1, \ldots, K. \tag{10}$$

Here the "*"-product $(X_i, U_k)^*$ has a meaning of the k-th principal component of the vector X_i. Therefore, this method can be called *Projection by means of principal components*.

Remark that simultaneous application of the "Toeplitz" and "Projection by means of principal components" modifications for the analysis of stationary series with missing data was suggested in Schoellhamer (2001). For non-stationary or short series these modifications provide poor results.

4.6 Example

To demonstrate the work of the methods of filling in missing data, let us consider the famous time series of length 144 representing monthly numbers of passengers (in thousands) on the international airlines, since January, 1949 (the data was published for the first time in Brown (1963)).

Let us remove 12 known values, starting with the 68-th point (i.e. we consider that values are unknown for a year since August, 1954). For such kind of artificially missing data we can estimate the accuracy of their recovering for different versions of the algorithm. Also, to simulate forecast, we add 12 missing data after the last, 144-th point of the series. The time series obtained is illustrated in Fig. 1.

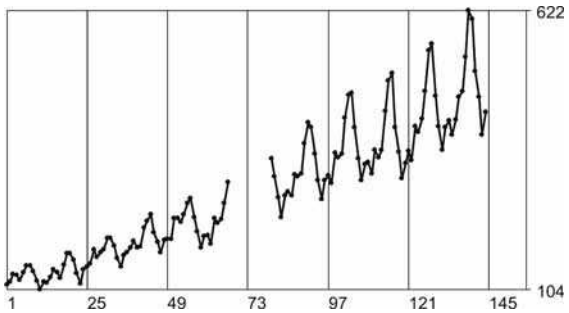

Fig. 1. The initial time series with missing data

The first question is how to choose the window length L. In the case of no missing data, the general recommendation is to choose the window length close to the half of the series length and divisible by the period of expected periodicity (12 months here). The window length equal to 72 meets these

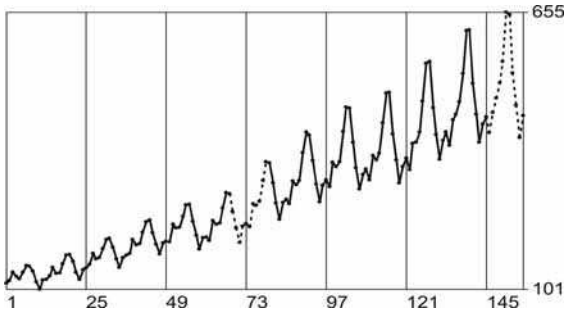

Fig. 2. The reconstructed time series with filled in data

conditions. However, under such choice of L all the lagged vectors will contain missing data. Let us choose a smaller length in order to avoid using the "*" operation. The choice of $L = 36$ provides us 62 complete lagged vectors with no missing data.

The analysis of the eigenvectors $U_1, \ldots, U_L$ (for real-life series it is common that $d = L$) shows that the eigenvectors with indices 1, 6, and 9 correspond to a trend and the eigenvectors with indices 2–5, 7–8, and 10–13 correspond to a seasonal component. All the rest eigenvectors may be classified as produced by noise. Therefore, we will choose $r = 13$ and $I_r = \{1, 2, \ldots, 13\}$ in order to reconstruct the deterministic component of the series (a signal).

All the eight variants of the step 3b are applicable to the first continuous inner cluster of missing data: two variants of (α) ("Π Projector" and "Components of adjacent vectors"), and four variants of implementing (β) (simultaneous filling in and three types of sequential filling in). A comparison of the reconstruction results with the values that were artificially removed from the initial time series shows an advantage of the variant "Π Projector" with simultaneous filling in the missing data. Reconstruction error therewith is approximately equal to 6 for the missing data and is equal to 4.75 (this is not much less) for other terms.

We apply the same method to fill in the second cluster of missing data. The result is illustrated in Fig. 2. The reconstructed series is marked by the dotted line in the area of missing data.

References

Brown, R.G., 1963. Smoothing, forecasting and prediction of discrete time series, New Jersey, Prentice-Hall.

Danilov, D. and Zhigljavsky, A. (Eds.), 1997. Principal Components of Time

Series: the “Caterpillar” method, SPbU Press, Saint-Petersburg (in Russian).

Elsner, J., Tsonis, A., 1996. Singular Spectrum Analysis. A New Tool in Time Series Analysis. New York: Plenum Press.

Golyandina, N., Nekrutkin, V., and Zhigljavsky, A., 2001. Analysis of Time Series Structure: SSA and Related Techniques. London: Chapman & Hall/CRC.

Kondrashov, D., Feliks, Y., and Ghil, M., 2005. Oscillatory modes of extended Nile River records (A.D. 622–1922), Geophys. Res. Lett. 32, No.10, L10702, doi:10.1029/2004GL022156.

Schoellhamer, D.H., 2001. Singular spectrum analysis for time series with missing data. Geophys. Res. Lett. 28, No.16, 3187–3190.