

Н.Э. Голяндина, Е.В. Осипов

## МЕТОД «ГУСЕНИЦА»-SSA ДЛЯ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ С ПРОПУСКАМИ <sup>1</sup>

### 1 Введение

Данная работа посвящена адаптации появившегося в последние десятилетия 20-го века (см., например, книги [3, 5, 6], содержащие большое число ссылок на литературу) метода «Гусеница»-SSA к анализу временных рядов с пропущенными наблюдениями. Такие попытки уже делались (см. [7], где изложение очень краткое и не раскрывает деталей, и [8], где метод заполнения годится только для стационарных рядов), но мы при заполнении будем использовать другие идеи, следуя [2, 3, 5].

Одной из отличительных особенностей метода SSA в применении к анализу временных рядов без пропусков является то, что с его помощью возможно проводить исследование структуры ряда (выявление трендовой, гармонических и шумовых составляющих) без предположений о модели ряда. Однако прогноз выделенных методом составляющих ряда возможен только в рамках некоторой, достаточно широкой, но все-таки модели этих составляющих. Предполагается, что прогнозируемая составляющая является рядом конечного ранга (см. [5]) (или, что почти то же самое, временным рядом, управляемым некоторой линейной рекуррентной формулой). Идея заполнения пропусков в рамках метода «Гусеница»-SSA в большой степени аналогична идее прогноза и состоит в продолжении выделенной методом составляющей ряда и ее структуры на места пропущенных наблюдений. Соответственно и теоретические результаты относительно условий и способов точного восстановления пропущенных значений в составляющих наблюдаемого ряда с пропусками относятся к рядам конечного ранга.

Также, как и в базовом методе «Гусеница»-SSA, построенные в работе модификации для исследования временных рядов с пропусками, дающие точные результаты в достаточно жестких предположениях, оказываются применимыми и к реальным рядам с пропусками, приводя в этом случае к приближенным результатам.

Отметим, что разработанные методы заполнения пропусков одновременно решают задачу прогноза, если группа пропусков расположена в конце ряда. Это позволяет более широко взглянуть на задачу прогнозирования и предоставляет новые варианты ее решения.

В разделе 2 работы приводятся необходимые для дальнейшего изложения результаты, не связанные напрямую с анализом временных рядов с пропусками. Первые два подраздела содержат результаты относительно свойств линейных подпространств, векторов из них, а также их сужений на фиксированное множество индексов. В последнем подразделе приведены основные понятия базового метода «Гусеница»-SSA для анализа временных рядов без пропусков.

Раздел 3 посвящен применению результатов первого раздела к подпространствам, порожденным наблюдаемым временным рядом конечного ранга с пропусками, и так называемым векторам вложения (отрезкам временного ряда заданной длины). Тем самым мы получаем условия и формулы для заполнения входящих в рассматриваемый вектор вложения пропусков.

В разделе 4, основываясь на полученных формулах, мы предлагаем набор модификаций метода «Гусеница»-SSA, позволяющих выделить составляющие из реальных временных рядов с одновременным заполнением в них пропусков. Там же приведен пример, демонстрирующий работу предлагаемых алгоритмов.

---

<sup>1</sup>©Н.Э. Голяндина, Е.В. Осипов

## 2 Общие результаты

### 2.1 Заполнение компонент вектора в подпространстве

Введем необходимые обозначения. Рассмотрим евклидово пространство  $\mathbb{R}^n$ . Единичную матрицу размера  $n \times n$  будем обозначать через  $\mathbf{E}_n$ . Через  $\mathcal{I}$  обозначим множество  $\{1, \dots, n\}$  и будем рассматривать некоторое его упорядоченное подмножество  $\mathcal{S} = \{i_1, \dots, i_s\} \subset \mathcal{I}$ , где  $|\mathcal{S}| = s$ .

Под *сужением* вектора  $X = (x_1, \dots, x_n)^T \in \mathbb{R}^n$  на множество индексов  $\mathcal{S}$  будем понимать вектор  $X|_{\mathcal{S}} = (x_{i_1}, \dots, x_{i_s})^T \in \mathbb{R}^{|\mathcal{S}|}$ .

*Сужением матрицы* на некоторое множество индексов будем называть матрицу, составленную из векторов–столбцов, суженных на это множество.

*Сужением  $q$ -мерного подпространства  $\mathfrak{G}_q$*  на некоторое множество индексов  $\mathcal{S}$  будем называть пространство, натянутое на все вектора пространства  $\mathfrak{G}_q$ , суженные на это множество индексов, и обозначать его через  $\mathfrak{G}_q|_{\mathcal{S}}$ . Легко доказать, что для любого базиса  $\{H_i\}_{i=1}^q$  подпространства  $\mathfrak{G}_q$  имеет место равенство  $\mathfrak{G}_q|_{\mathcal{S}} = \text{span}(H_1|_{\mathcal{S}}, \dots, H_q|_{\mathcal{S}})$ .

Рассмотрим  $m$ -мерное подпространство  $\mathfrak{D}_m \subset \mathbb{R}^n$ . Некоторый ортонормированный базис этого пространства обозначим через  $\{R_k\}_{k=1}^m$ . Также введем матрицу  $\mathbf{R} = [R_1 : \dots : R_m]$ . Зафиксируем некоторое упорядоченное множество индексов  $\mathcal{P}$ .

**Предложение 2.1** Пусть матрица  $\mathbf{E}_{|\mathcal{P}|} - \mathbf{R}|_{\mathcal{P}}(\mathbf{R}|_{\mathcal{P}})^T$  невырожденная. Тогда для произвольного вектора  $X \in \mathfrak{D}_m$  имеет место формула, выражающая вектор  $X|_{\mathcal{P}}$  через  $X|_{\mathcal{I} \setminus \mathcal{P}}$ :

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

**Доказательство.** Не умаляя общности, будем считать, что  $\mathcal{P} = \{1, \dots, |\mathcal{P}|\}$ . Обозначим  $X_1 = X|_{\mathcal{P}}$ ,  $X_2 = X|_{\mathcal{I} \setminus \mathcal{P}}$ ,  $\mathbf{R}_1 = \mathbf{R}|_{\mathcal{P}}$  и  $\mathbf{R}_2 = \mathbf{R}|_{\mathcal{I} \setminus \mathcal{P}}$ . Поскольку равенство  $\mathbf{R}\mathbf{R}^T X = X$  имеет место для любого  $X \in \mathfrak{D}_m$  и

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

то  $X_1 = \mathbf{R}_1\mathbf{R}_1^T X_1 + \mathbf{R}_1\mathbf{R}_2^T X_2$ . Возвращаясь к исходным обозначениям, получаем завершающее доказательство равенство

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

□

**Лемма 2.1** *Равносильны следующие условия:*

- 1)  $\dim \mathfrak{D}_m|_{\mathcal{I} \setminus \mathcal{P}} = \dim \mathfrak{D}_m$ ;
- 2) набор  $\{R_k|_{\mathcal{I} \setminus \mathcal{P}}\}_{k=1}^m$  линейно независим и является базисом подпространства  $\mathfrak{D}_m|_{\mathcal{I} \setminus \mathcal{P}}$ ;
- 3) для любого ненулевого вектора  $Y \in \mathfrak{D}_m$  выполняется  $Y|_{\mathcal{I} \setminus \mathcal{P}} \neq 0|_{\mathcal{I} \setminus \mathcal{P}}$ ;
- 4)  $\text{span}(e_i, i \in \mathcal{P}) \cap \mathfrak{D}_m = \{0_n\}$ ;
- 5) матрица  $(\mathbf{E}_{|\mathcal{P}|} - \mathbf{R}|_{\mathcal{P}}(\mathbf{R}|_{\mathcal{P}})^T)^{-1}$  существует.

**Доказательство.**

*Равносильность условий 1) – 4)* очевидна.

*Равносильность 4) и 5).* Из предложения 2.1 следует, что 5) эквивалентно следующему утверждению: для любого вектора  $V \in \mathfrak{D}_m|_{\mathcal{I} \setminus \mathcal{P}}$  существует единственный вектор  $G \in \mathfrak{D}_m$  такой, что  $V = G|_{\mathcal{I} \setminus \mathcal{P}}$ . Покажем, что 4) тоже эквивалентно этому утверждению.

Поскольку вектора  $\{R_k|_{\mathcal{I} \setminus \mathcal{P}}\}_{k=1}^m$  являются базисом пространства  $\mathfrak{D}_m|_{\mathcal{I} \setminus \mathcal{P}}$ , то вектор  $V$  может быть записан в виде  $V = \sum_{k=1}^m a_k R_k|_{\mathcal{I} \setminus \mathcal{P}}$ . Тогда в качестве нужного вектора  $G$  возьмем вектор  $G = \sum_{k=1}^m a_k R_k$ . Предположим, что существует два разных вектора  $G_1$  и  $G_2$  такие, что  $V = G_1|_{\mathcal{I} \setminus \mathcal{P}} = G_2|_{\mathcal{I} \setminus \mathcal{P}}$ . Тогда рассмотрим их разность  $G_1 - G_2 = \sum_{i \in \mathcal{P}} \alpha_i e_i \in \mathfrak{D}_m$ . Но эта разность не обращается в нуль тогда и только тогда, когда  $\text{span}(e_i, i \in \mathcal{P}) \cap \mathfrak{D}_m = \{0_n\}$ , что и завершает доказательство.  $\square$

**Замечание 2.1** Неравенство пункта 4) леммы 2.1 можно переписать следующим образом:  $n - m \geq |\mathcal{P}|$ . Это ограничение на количество пропущенных компонент в векторе является необходимым условием для применения формулы (1).

Сформулируем два частных случая, дающих выражение первой ( $\mathcal{P} = \{1\}$ ) и последней ( $\mathcal{P} = \{n\}$ ) координаты через остальные.

**Следствие 2.1** Обозначим  $\nu^2 := \pi_1^2 + \dots + \pi_m^2$ , где  $\pi_i$  —  $n$ -я компонента вектора  $R_i$ , и  $\{R_i^\nabla\}_{i=1}^m$  — вектора  $\{R_i\}_{i=1}^m$  без последних компонент (размерности  $n - 1$ ). Пусть  $e_n \notin \mathfrak{D}_m$  и  $X = (x_1, \dots, x_n)^\top \in \mathfrak{D}_m$ . Тогда  $\nu^2 < 1$  и  $x_n = \sum_{k=1}^{n-1} a_k x_{n-k}$ , где

$$(a_{n-1}, \dots, a_1)^\top = \frac{1}{1 - \nu^2} \sum_{i=1}^m \pi_i R_i^\nabla. \quad (2)$$

**Следствие 2.2** Обозначим  $\mu^2 := \rho_1^2 + \dots + \rho_m^2$ , где  $\rho_i$  — 1-я компонента вектора  $R_i$ , и  $\{R_i^\Delta\}_{i=1}^m$  — вектора  $\{R_i\}_{i=1}^m$  без первых компонент (размерности  $n - 1$ ). Пусть  $e_1 \notin \mathfrak{D}_m$  и  $X = (x_1, \dots, x_n)^\top \in \mathfrak{D}_m$ . Тогда  $\mu^2 < 1$  и  $x_1 = \sum_{k=2}^n a_k x_k$ , где

$$(a_2, \dots, a_n)^\top = \frac{1}{1 - \mu^2} \sum_{i=1}^m \rho_i R_i^\Delta. \quad (3)$$

## 2.2 Оператор проектирования

Рассмотрим два подпространства  $\mathcal{L}^{(1)}$  и  $\mathcal{L}^{(2)}$  пространства  $\mathbb{R}^n$  размерностей  $m$  и  $\tilde{m} \leq n - m$  соответственно. Сначала найдем явный вид матрицы оператора  $\Pi_{\mathcal{I} \setminus \mathcal{P}}^{(1)}$  ортогонального проектирования  $\mathbb{R}^{|\mathcal{I} \setminus \mathcal{P}|}$  на  $\mathcal{L}^{(1)}|_{\mathcal{I} \setminus \mathcal{P}}$ . Обозначим через  $\{R_k\}_{k=1}^m$  некоторый ортонормированный базис пространства  $\mathcal{L}^{(1)}$ ,  $\mathbf{R} = [R_1 : \dots : R_m]$ . Для более компактной записи формулировки предложения положим  $\mathbf{V} = \mathbf{R}|_{\mathcal{I} \setminus \mathcal{P}}$  и  $\mathbf{W} = \mathbf{R}|_{\mathcal{P}}$ .

**Предложение 2.2** Предположим, что матрица  $\mathbf{E}_{|\mathcal{P}|} - \mathbf{W}\mathbf{W}^\top$  невырождена. Тогда матрица  $\Pi_{\mathcal{I} \setminus \mathcal{P}}^{(1)}$  оператора ортогонального проектирования  $\Pi_{\mathcal{I} \setminus \mathcal{P}}^{(1)}$  вычисляется по формуле:

$$\Pi_{\mathcal{I} \setminus \mathcal{P}}^{(1)} = \mathbf{V}\mathbf{V}^\top + \mathbf{V}\mathbf{W}^\top (\mathbf{E}_{|\mathcal{P}|} - \mathbf{W}\mathbf{W}^\top)^{-1} \mathbf{W}\mathbf{V}^\top. \quad (4)$$

**Доказательство.** Для удобства записи обозначим  $\mathbf{A} = \mathbf{V}^\top \mathbf{V}$  (матрицы  $\mathbf{A}$  и  $\mathbf{W}^\top \mathbf{W}$  имеют размер  $m \times m$ ). В соответствии с леммой 2.1 матрица  $\mathbf{A}$  является невырожденной как матрица Грама линейно независимого набора векторов и, следовательно, обратимой. Известно, что

матрица оператора ортогонального проектирования на векторное пространство, натянутое на столбцы матрицы  $\mathbf{V}$ , имеет вид  $\Pi_{\mathcal{I} \setminus \mathcal{P}}^{(1)} = \mathbf{V} \mathbf{A}^{-1} \mathbf{V}^T$ . Таким образом, остается найти вид матрицы  $\mathbf{A}^{-1}$ .

Так как вектора  $\{R_k\}_{k=1}^m$  составляют ортонормированный базис  $\mathcal{L}^{(1)}$ , то  $\mathbf{R}^T \mathbf{R} = \mathbf{E}_m$ . С другой стороны,

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

Отсюда имеем  $\mathbf{A} = \mathbf{E}_m - \mathbf{W}^T \mathbf{W}$ . Покажем, что  $\mathbf{A}^{-1} = \mathbf{E}_m + \mathbf{W}^T (\mathbf{E}_{|\mathcal{P}|} - \mathbf{W} \mathbf{W}^T)^{-1} \mathbf{W}$ . Действительно,

$$\begin{aligned} \mathbf{A} (\mathbf{E}_m + \mathbf{W}^T (\mathbf{E}_{|\mathcal{P}|} - \mathbf{W} \mathbf{W}^T)^{-1} \mathbf{W}) &= \\ = (\mathbf{E}_m - \mathbf{W}^T \mathbf{W}) (\mathbf{E}_m + \mathbf{W}^T (\mathbf{E}_{|\mathcal{P}|} - \mathbf{W} \mathbf{W}^T)^{-1} \mathbf{W}) &= \mathbf{E}_m. \end{aligned}$$

Используя полученное выражение матрицы  $\mathbf{A}^{-1}$ , получаем явный вид оператора  $\Pi_{\mathcal{I} \setminus \mathcal{P}}^{(1)}$ :

$$\Pi_{\mathcal{I} \setminus \mathcal{P}}^{(1)} = \mathbf{V} \mathbf{A}^{-1} \mathbf{V}^T = \mathbf{V} \mathbf{V}^T + \mathbf{V} \mathbf{W}^T (\mathbf{E}_{|\mathcal{P}|} - \mathbf{W} \mathbf{W}^T)^{-1} \mathbf{W} \mathbf{V}^T.$$

□

Введем еще один оператор  $\Pi^{(1)}$  ортогонального проектирования  $\mathbb{R}^n$  на  $\mathcal{L}^{(1)}$ . Следующее утверждение дает условия, при которых оператор проектирования и операция сужения в некотором смысле перестановочны.

**Предложение 2.3** Если  $\mathcal{L}^{(1)} \perp \mathcal{L}^{(2)}$  и  $\mathcal{L}^{(1)}|_{\mathcal{I} \setminus \mathcal{P}} \perp \mathcal{L}^{(2)}|_{\mathcal{I} \setminus \mathcal{P}}$ , то

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

для любого  $X \in \mathcal{L}^{(1)} \oplus \mathcal{L}^{(2)}$

**Доказательство.**

Рассмотрим произвольный вектор  $X \in \mathcal{L}^{(1)} \oplus \mathcal{L}^{(2)}$  и запишем его в виде  $X = \Pi^{(1)} X + \Pi^{(2)} X$ , где  $\Pi^{(2)}$  — оператор ортогонального проектирования  $\mathbb{R}^n$  на  $\mathcal{L}^{(2)}$ . Сузим компоненты вектора  $X$  на индексы  $\mathcal{I} \setminus \mathcal{P}$ :

$$X|_{\mathcal{I} \setminus \mathcal{P}} = (\Pi^{(1)} X)|_{\mathcal{I} \setminus \mathcal{P}} + (\Pi^{(2)} X)|_{\mathcal{I} \setminus \mathcal{P}}.$$

По определению операции сужения пространств  $(\Pi^{(1)} X)|_{\mathcal{I} \setminus \mathcal{P}} \in \mathcal{L}^{(1)}|_{\mathcal{I} \setminus \mathcal{P}}$  и  $(\Pi^{(2)} X)|_{\mathcal{I} \setminus \mathcal{P}} \in \mathcal{L}^{(2)}|_{\mathcal{I} \setminus \mathcal{P}}$ . Использование условия  $\mathcal{L}^{(1)}|_{\mathcal{I} \setminus \mathcal{P}} \perp \mathcal{L}^{(2)}|_{\mathcal{I} \setminus \mathcal{P}}$  завершает доказательство. □

## 2.3 Понятия метода SSA

Рассмотрим вещественный временной ряд  $F_N = (f_0, \dots, f_{N-1})$ , состоящий из  $N$  элементов. Введем необходимые в дальнейшем понятия метода “Гусеница”-SSA, следуя [1, 2, 5].

Зафиксируем некоторое целое число  $L$ ,  $1 < L < N$ , называемое *длиной окна*. Процедура вложения переводит исходный временной ряд в последовательность  $L$ -мерных векторов вложения  $\{X_i\}_{i=1}^K$ , где  $K = N - L + 1$ , по формуле

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

Из этих векторов вложения, в свою очередь, образуется  $L$ -траекторная матрица  $\mathbf{X}$  (или просто *траекторная матрица*) ряда  $F_N$ :

$$\mathbf{X} = [X_1 : \dots : X_K].$$

При записи через элементы исходного ряда  $F_N$  траекторная матрица ряда будет иметь вид

$$\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} f_0 & f_1 & f_2 & \dots & f_{K-1} \\ f_1 & f_2 & f_3 & \dots & f_K \\ f_2 & f_3 & f_4 & \dots & f_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \dots & f_{N-1} \end{pmatrix}. \quad (5)$$

Очевидно, что  $x_{ij} = f_{i+j-2}$  и матрица  $\mathbf{X}$  имеет одинаковые элементы на всех «диагоналях» с  $i + j = \text{Const}$ . Матрицы такого вида называют *ганкелевыми*.

Произвольная матрица аппроксимируется ганкелевой матрицей в пространстве матриц с нормой Фробениуса с помощью операции ганкелизации, которая заключается в замене значений на диагоналях  $i + j = \text{Const}$  на среднее арифметическое этих значений. Существует естественное взаимно-однозначное соответствие между ганкелевыми матрицами и временными рядами. Переход от произвольной матрицы к временному ряду с помощью ганкелизации матрицы и затем использование этого взаимно-однозначного соответствия (диагональ  $\leftrightarrow$  значение ряда, см. (5)) называется операцией *диагонального усреднения*.

**Определение 2.1**  $L$ -траекторным пространством столбцов (или просто *траекторным пространством*)  $\mathcal{L}^{(L)}(F_N)$  ряда  $F_N$  при фиксированной длине окна  $L$  будем называть линейную оболочку всех векторов вложения, иначе говоря:  $\mathcal{L}^{(L)} = \mathcal{L}^{(L)}(F_N) \stackrel{\text{def}}{=} \text{span}(X_1, \dots, X_K)$ .  $L$ -траекторным пространством строк (или просто *траекторным пространством строк*)  $\mathcal{L}^{(K)}(F_N)$  ряда  $F_N$  при фиксированной длине окна  $L$  будем называть линейную оболочку строк  $L$ -траекторной матрицы ряда.

**Определение 2.2** Пусть  $0 \leq d \leq L$ . Если  $\dim \mathcal{L}^{(L)} = d$ , то мы будем говорить, что *ряд  $F_N$  имеет  $L$ -ранг  $d$* , и записывать это как  $\text{rank}_L(F_N) = d$ . Для нулевого ряда положим  $\dim \mathcal{L}^{(L)} = 0$ . Если равенство  $\text{rank}_L(F_N) = d < N/2$  имеет место для любого  $L$ ,  $d \leq \min(L, K)$ , то будем говорить, что *ряд  $F_N$  имеет ранг  $d$*  ( $\text{rank}(F_N) = d$ ). Если такое  $d$  существует, то ряд  $F_N$  будем называть *рядом конечного ранга*.

Для ряда конечного ранга  $d$  через  $\mathcal{L}_d$  будем обозначать его траекторное пространство (размерности  $d$ ).

Ортонормированный базис траекторного пространства состоит из собственных векторов  $U_i$ ,  $i = 1, \dots, d = \text{rank } \mathbf{X}$ , матрицы  $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ , соответствующих  $d$  ненулевым собственным числам  $\mathbf{S}$ . Вектора  $U_1, \dots, U_d$  являются левыми сингулярными векторами сингулярного разложения матрицы  $\mathbf{X}$ .

Приведенное ниже утверждение показывает, что множество рядов конечного ранга достаточно большое.

**Предложение 2.4** *Любой ряд, являющийся линейной комбинацией произведений полиномов, экспонент и косинусов, является рядом конечного ранга.*

Следующее понятие является очень важным в теории SSA (а его асимптотический аналог составляет основу применения SSA к анализу реальных временных рядов).

**Определение 2.3** Если траекторные пространства столбцов и траекторные пространства строк двух рядов ортогональны, то ряды называются *разделимыми*.

Если два ряда  $F_N^{(1)}$  и  $F_N^{(2)}$  разделимы, то базис  $U_i, i = 1, \dots, d$ , траекторного пространства ряда  $F_N = F_N^{(1)} + F_N^{(2)}$  можно (с точностью до проблемы, связанной с совпадающими собственными числами, т.е. отсутствием сильной разделимости) разбить на две части, одна из которых составляет базис траекторного пространства первого ряда  $F_N^{(1)}$ , а вторая — базис траекторного пространства ряда  $F_N^{(2)}$ . Тем самым, в случае (сильно) разделимых рядов  $F_N^{(1)}$  и  $F_N^{(2)}$  метод SSA позволяет по наблюдаемой сумме рядов  $F_N = F_N^{(1)} + F_N^{(2)}$  найти слагаемые, спроектировав вектора вложения ряда  $F_N$  на найденные подпространства и тем самым получив вектора вложения рядов  $F_N^{(1)}$  и  $F_N^{(2)}$ .

### 3 Вектора вложения и траекторные пространства для рядов конечного ранга с пропусками

#### 3.1 Заполнение пропусков в векторе вложения

Пусть  $F_N$  является рядом конечного ранга  $d$  и  $L$  — некоторая длина окна,  $d < \min(L, K)$ ,  $K = N - L + 1$ . Предположим, что в наблюдаемом ряде  $F_N$  часть значений пропущена (т.е. часть значений считается неизвестной), но его  $L$ -траекторное пространство  $\mathcal{L}_d \subset \mathbb{R}^L$  известно.

Возьмем один из векторов вложения ряда  $F_N$ , содержащий как пропущенные, так и известные компоненты ряда, и обозначим его через  $X$ , а упорядоченное множество индексов его пропущенных компонент через  $\mathcal{P}$ . Начиная с этого момента, под множеством  $\mathcal{I}$  всегда будем понимать множество  $\{1, \dots, L\}$ . Тогда множество индексов  $\mathcal{I} \setminus \mathcal{P}$  представляет собой множество известных компонент вектора  $X$ , а сужения  $X|_{\mathcal{P}}$  и  $X|_{\mathcal{I} \setminus \mathcal{P}}$  — вектора, состоящие из пропущенных и известных компонент вектора  $X$  соответственно. Будем решать задачу восстановления пропущенных компонент  $X|_{\mathcal{P}}$  вектора  $X$  по известным  $X|_{\mathcal{I} \setminus \mathcal{P}}$ .

Применим теорию раздела 2 (с  $m = d$ ,  $\mathfrak{D}_m = \mathcal{L}_d$ ) к вектору  $X$ . Предположим, что траекторное пространство  $\mathcal{L}_d$  удовлетворяет условиям применимости предложения 2.1. Тогда формула (1) решает задачу нахождения  $X|_{\mathcal{P}}$  по  $X|_{\mathcal{I} \setminus \mathcal{P}}$  и, следовательно, восстанавливает пропущенные значения ряда  $F_N$ , входящие в вектор вложения  $X$ .

Покажем, что условие на число пропущенных наблюдений (замечание 2.1), говорящее о том, что число известных наблюдений в векторе должно быть не меньше ранга ряда, является только необходимым.

**Пример 3.1** Рассмотрим ряд  $f_n = a + b(-1)^n$ ,  $n = 0, \dots, 4$ , длины 5, представляющий собой сумму константного ряда и гармонического ряда с периодом 2 (так называемая "пила"). Ранг  $d$  такого ряда равен 2. Зафиксируем длину окна  $L = 3$  и предположим, что значения ряда  $f_1$  и  $f_3$  пропущены. При данной длине окна ряд будет иметь три вектора вложения —  $X_1, X_2$  и  $X_3$ . Заметим, что вектор  $X_2$  содержит только одну известную компоненту, и поэтому восстановить пропуски в этом векторе не удастся, так как нарушается необходимое условие из замечания 2.1. Остальные два вектора вложения содержат по две известных компоненты и необходимое условие восстановления не нарушается. Однако восстановить пропуски этих векторов невозможно, так как  $X_1|_{\{1,3\}} = X_3|_{\{1,3\}} = (a + b, a + b)^T$  и  $\dim \mathcal{L}_d|_{\{1,3\}} = \dim \text{span}(X_1|_{\{1,3\}}, X_3|_{\{1,3\}}) = 1$ . Получаем нарушение условия  $\dim \mathcal{L}_d|_{\{1,3\}} = \dim \mathcal{L}_d$ , что, в силу леммы 2.1, означает нарушение условия применимости предложения 2.1.

Теперь усложним ситуацию: предположим, что исходный ряд с пропусками является суммой двух разделимых рядов конечного ранга  $F_N^{(1)}$  и  $F_N^{(2)}$ , т.е.  $F_N = F_N^{(1)} + F_N^{(2)}$ . Предположим также, что нам известны траекторные пространства (пространства столбцов)  $\mathcal{L}^{(1)}$  и  $\mathcal{L}^{(2)}$  рядов  $F_N^{(1)}$  и  $F_N^{(2)}$ , соответственно. В силу разделимости рядов, эти пространства ортогональны:  $\mathcal{L}^{(1)} \perp \mathcal{L}^{(2)}$ . Пусть опять  $X$  — вектор вложения ряда  $F_N$  с пропусками на местах из  $\mathcal{P}$ ,

$X = X^{(1)} + X^{(2)}$ , где  $X^{(i)}$ ,  $i = 1, 2$  — соответствующие вектора вложения  $F_N^{(i)}$ . Будем решать задачу нахождения вектора вложения первого ряда  $X^{(1)}$  по  $X|_{\mathcal{I} \setminus \mathcal{P}}$ .

Задача нахождения  $X^{(1)}$  разбивается на задачи поиска  $X^{(1)}|_{\mathcal{I} \setminus \mathcal{P}}$  и  $X^{(1)}|_{\mathcal{P}}$ . Если первая задача решена, то нахождение  $X^{(1)}|_{\mathcal{P}}$  по  $X^{(1)}|_{\mathcal{I} \setminus \mathcal{P}}$  происходит по формуле (1), где  $\mathbf{R}$  — матрица, столбцами которой являются вектора ортонормированного базиса пространства  $\mathcal{L}^{(1)}$ . Утверждения раздела 2.2 дают нам способ решения задачи нахождения  $X^{(1)}|_{\mathcal{I} \setminus \mathcal{P}}$ , если выполнено условие ортогональности суженных траекторных пространств  $\mathcal{L}^{(1)}|_{\mathcal{I} \setminus \mathcal{P}}$  и  $\mathcal{L}^{(2)}|_{\mathcal{I} \setminus \mathcal{P}}$ . Тогда проектирование компонент вектора  $X|_{\mathcal{I} \setminus \mathcal{P}}$  на  $\mathcal{L}^{(1)}|_{\mathcal{I} \setminus \mathcal{P}}$  дает компоненты ряда  $F_N^{(1)}$ , так как соответствующий вектор вложения  $X^{(1)}$  ряда  $F_N^{(1)}$  в силу ортогональности  $\mathcal{L}^{(1)}$  и  $\mathcal{L}^{(2)}$  равен  $\mathbf{\Pi}^{(1)}X$ .

Таким образом,  $X^{(1)}|_{\mathcal{I} \setminus \mathcal{P}}$  строится по  $X|_{\mathcal{I} \setminus \mathcal{P}}$  по формуле (4), где  $\mathbf{R}$  — матрица, столбцами которой являются вектора ортонормированного базиса пространства  $\mathcal{L}^{(1)}$ , а  $X^{(1)}|_{\mathcal{P}}$  строится по  $X^{(1)}|_{\mathcal{I} \setminus \mathcal{P}}$  по формуле (1) с той же матрицей  $\mathbf{R}$ . Тем самым, будут найдены все значения ряда  $F_N^{(1)}$ , входящие в его вектор вложения  $X^{(1)}$ , в том числе стоящие на местах пропущенных значений ряда  $F_N$ .

### 3.2 Нахождение траекторных пространств исходного ряда и его составляющих

Рассмотрим временной ряд  $F_N$ , имеющий  $L$ -ранг  $d$ , и его вектора вложения  $\{X_i\}_{i=1}^K$ . Получим сначала условия, при которых можно найти базис траекторного пространства столбцов  $\mathcal{L}_d = \text{span}(X_i, i = 1, \dots, d)$  этого ряда по наблюдаемому ряду с пропусками.

Обозначим через  $\mathcal{C} \subset \{1, \dots, K\}$  множество номеров столбцов  $L$ -траекторной матрицы ряда  $\mathbf{X}$ , не имеющих пропущенных компонент. Предположим, что  $\mathcal{C} \neq \emptyset$  и рассмотрим матрицу  $\tilde{\mathbf{X}}$ , составленную из векторов вложения без пропущенных значений.

Пусть  $\tilde{\mathcal{L}}_d = \text{span}(X_i, i \in \mathcal{C})$ . Нетрудно доказать следующее утверждение.

**Предложение 3.1** *Среди векторов вложения, не имеющих пропущенных значений, найдется  $d$  линейно независимых тогда и только тогда, когда  $\tilde{\mathcal{L}}_d = \mathcal{L}_d$ . При этом ортонормированным базисом подпространства  $\mathcal{L}_d$  являются собственные вектора  $U_1, \dots, U_d$  матрицы  $\tilde{\mathbf{S}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$ , соответствующие  $d$  ненулевым собственным числам  $\tilde{\mathbf{S}}$ .*

Сформулируем более конструктивное достаточное условие.

**Предложение 3.2** *Если ряд  $F_N$  имеет ранг  $d$ ,  $e_1 \notin \mathcal{L}_d$ ,  $e_L \notin \mathcal{L}_d$  и ряд имеет  $L + d - 1$  подряд идущих известных значений, то  $\tilde{\mathcal{L}}_d = \mathcal{L}_d$ .*

**Доказательство.**

Для того чтобы доказать предложение, покажем, что среди векторов вложения, не содержащих пропуска, найдется  $d$  линейно независимых. Из условия предложения следует, что найдется такой индекс  $k$ , что вектора вложения  $\{X_{k+i}\}_{i=1}^d$  не содержат пропущенных значений. Из условий  $e_1 \notin \mathcal{L}_d$  и  $e_L \notin \mathcal{L}_d$  следует, что временной ряд  $F_N$  можно продолжить до бесконечного в обе стороны временного ряда  $F$   $L$ -ранга  $d$  (доказательство основано на следствии 2.1 и следствии 2.2 при  $n = L$ ,  $m = d$  и  $\mathfrak{D}_m = \mathcal{L}_d$ ). Следовательно, ряд  $F_N$  имеет конечный ранг  $d$  и, в частности,  $\text{rank}_{d+1} F_N = d$ ,  $e_1, e_{d+1} \notin \mathcal{L}^{(d+1)}(F_N)$ . Отсюда, используя следствие 2.1 с  $n = d + 1$ ,  $m = d$ ,  $\mathfrak{D}_m = \mathcal{L}^{(d+1)}(F_N)$ , получаем, что

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

Следовательно, все вектора вложения с индексами большими, чем  $k + d$ , можно записать в виде линейной комбинации векторов  $X_{k+1}, \dots, X_{k+d}$ .

Аналогично, в силу следствия 2.2, получаем, что вектора вложения с индексами от 1 до  $k$  можно записать в виде линейной комбинации векторов  $X_{k+1}, \dots, X_{k+d}$ . Таким образом получили, что все вектора вложения выражаются через набор векторов  $\{X_{k+i}\}_{i=1}^d$ . Поскольку размерность траекторного пространства равна  $d$ , то эти вектора линейно независимы.  $\square$

Теперь рассмотрим случай, когда наблюдается ряд, являющийся суммой двух разделимых рядов:  $F_N = F_N^{(1)} + F_N^{(2)}$ . Пусть условия предложения 3.1 выполнены, базис траекторного пространства  $\mathcal{L}_d$  ряда  $F_N$  найден, и поставим задачу нахождения траекторных пространств  $\mathcal{L}^{(1)}$  и  $\mathcal{L}^{(2)}$  рядов  $F_N^{(1)}$  и  $F_N^{(2)}$ , используя  $\tilde{\mathbf{X}} = \tilde{\mathbf{X}}^{(1)} + \tilde{\mathbf{X}}^{(2)}$ .

Если пространства строк и столбцов матриц  $\tilde{\mathbf{X}}^{(1)}$  и  $\tilde{\mathbf{X}}^{(2)}$  ортогональны, то собственные вектора  $U_1, \dots, U_d$  матрицы  $\tilde{\mathbf{S}}$ , являющиеся базисом пространства  $\mathcal{L}_d$ , также, как и в случае сильной разделимости, делятся на две части, составляющие базисы траекторных пространств  $\mathcal{L}^{(1)}$  и  $\mathcal{L}^{(2)}$ . Траекторные пространства столбцов ортогональны в силу разделимости рядов  $F_N^{(1)}$  и  $F_N^{(2)}$ . Несложно показать, что для ортогональности пространств строк следующие условия являются необходимыми и достаточными:

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

Таким образом, для возможности нахождения базисов траекторных пространств  $\mathcal{L}^{(1)}$  и  $\mathcal{L}^{(2)}$  достаточно выполнения условий предложения 3.1, а также условия (6).

В качестве примера можно привести исходный ряд в виде суммы константного ряда и гармонического ряда с периодом  $T$ . Известно, что если  $L$  и  $K$  кратны  $T$ , то ряды будут разделимы. Если пропущенным окажется полный период подряд идущих значений и  $N$  и  $L$  достаточно большие, то все равно базисы пространств  $\mathcal{L}^{(1)}$  и  $\mathcal{L}^{(2)}$  могут быть найдены.

### 3.3 Условия ортогональности суженных подпространств

В разделе 2.2 было получено (а в разделе 3.1 — использовано) предложение 2.3 о свойствах оператора ортогонального проектирования  $\Pi_{\mathcal{I} \setminus \mathcal{P}}^{(1)}$ , в котором рассматривался вектор  $X$  и его множество индексов пропущенных компонент  $\mathcal{P}$ , при условии ортогональности пространств  $\mathcal{L}^{(1)}|_{\mathcal{I} \setminus \mathcal{P}}$  и  $\mathcal{L}^{(2)}|_{\mathcal{I} \setminus \mathcal{P}}$ . Пусть  $\mathcal{L}^{(1)}$  и  $\mathcal{L}^{(2)}$  — траекторные пространства рядов  $F_N^{(1)}$  и  $F_N^{(2)}$ . Тогда следующие уравнения дают необходимое и достаточное условие ортогональности  $\mathcal{L}^{(1)}|_{\mathcal{I} \setminus \mathcal{P}} \perp \mathcal{L}^{(2)}|_{\mathcal{I} \setminus \mathcal{P}}$ :

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

**Замечание 3.1** Если ряды  $F_N^{(1)}$  и  $F_N^{(2)}$  разделимы, то (7) равносильно условию

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

Примером выполнения условия (8) опять служит исходный ряд в виде суммы константного ряда и гармонического ряда с периодом  $T$ , а также  $L$ ,  $K$  и число подряд идущих пропущенных наблюдений, кратные  $T$ . Однако также необходимо потребовать, чтобы  $L$  было больше числа пропущенных значений и  $\mathcal{P}$  задавало позиции пропусков в векторе вложения, содержащем все пропуски.



## 4 «Гусеница»-SSA для временных рядов с пропущенными значениями

### 4.1 Схема алгоритма

Результатом базового алгоритма метода «Гусеница»-SSA является разложение наблюдаемого временного ряда на аддитивные составляющие. Алгоритм метода «Гусеница»-SSA для рядов без пропусков [1, 5] состоит из двух этапов, разложения и восстановления. В свою очередь, каждый этап состоит из двух шагов: Вложение и Сингулярное разложение — шаги первого этапа, и Группировка и Диагональное усреднение — шаги второго этапа. Для анализа рядов с пропусками общая структура алгоритма та же, но шаги этапов будут несколько другими.

Итак, у нас есть исходный временной ряд  $F_N = (f_0, \dots, f_{N-1})$ , состоящий из  $N$  элементов, часть которых неизвестна. Опишем схему алгоритма для случая восстановления первой составляющей ряда  $F_N^{(1)}$  на основе суммы двух:  $F_N = F_N^{(1)} + F_N^{(2)}$ .

#### 4.1.1 Первый этап: разложение

##### Шаг 1. Вложение

Зафиксируем длину окна  $L : 1 < L < N$ . Процедура вложения переводит исходный временной ряд в последовательность  $L$ -мерных векторов вложения  $\{X_i\}_{i=1}^K$ , где  $K = N - L + 1$ . Часть векторов вложения может иметь пропуски. Из векторов вложения без пропусков  $X_i, i \in \mathcal{C}$  (предполагаем, что их не пустое множество) образуем матрицу  $\tilde{\mathbf{X}}$ , которая при отсутствии пропусков совпадает с траекторной матрицей ряда  $F_N$ .

##### Шаг 2. Нахождение базиса

Пусть  $\tilde{\mathbf{S}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$ . Обозначим  $\lambda_1, \dots, \lambda_L$  *собственные числа* матрицы  $\tilde{\mathbf{S}}$ , взятые в невозрастающем порядке ( $\lambda_1 \geq \dots \geq \lambda_L \geq 0$ ) и  $U_1, \dots, U_L$  — ортонормированную систему *собственных векторов* матрицы  $\tilde{\mathbf{S}}$ , соответствующих собственным числам,  $d = \max\{i : \lambda_i > 0\}$ .

#### 4.1.2 Второй этап: восстановление

##### Шаг 3а. Выбор подпространства и проекция векторов вложения без пропусков

Выбирается набор номеров  $I_r = \{i_1, \dots, i_r\} \subset \{1, \dots, d\}$ , с помощью которых образуется подпространство  $\mathcal{M}_r = \text{span}(U_{i_1}, \dots, U_{i_r})$ , соответствующее выделяемой компоненте. Выбор собственных векторов, соответствующих  $F_N^{(1)}$ , происходит аналогично тому, как это делается на этапе группировки в базовом алгоритме SSA и мы на этом не останавливаемся. Одним из признаков нужного собственного вектора является то, что его форма подобна форме составляющей ряда  $F_N^{(1)}$ . Затем происходит проектирование векторов вложения без пропусков на выбранное подпространство  $\mathcal{M}_r$ :

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

##### Шаг 3б. Проекция векторов вложения с пропусками

Данный шаг для каждого вектора вложения с пропусками на местах из  $\mathcal{P}$  (множество  $\mathcal{P}$  свое для каждого вектора) состоит из двух частей:

( $\alpha$ ) вычисления  $\hat{X}_i|_{\mathcal{I} \setminus \mathcal{P}}$ ,

( $\beta$ ) вычисления  $\hat{X}_i|_{\mathcal{P}}$ .

Так как соседние вектора вложения имеют общую информацию, то ее использование приводит к большому числу возможных способов решения поставленной задачи, в том числе и для векторов с  $\mathcal{P} = \mathcal{I}$ . Некоторые из этих способов будут описаны в следующем разделе. Заметим, что шаг 3б может привести и к изменению векторов  $\hat{X}_i, i \in \mathcal{C}$ .

Результатом шагов 3а и 3б является матрица  $\widehat{\mathbf{X}} = [\widehat{X}_1 \dots : \widehat{X}_K]$ , служащая аппроксимацией траекторной матрицы ряда  $F_N^{(1)}$  при правильном выборе множества  $I_r$ .

#### Шаг 4. Диагональное усреднение

На последнем шаге базового алгоритма матрица  $\widehat{\mathbf{X}}$  переводится в новый ряд  $\widetilde{F}_N^{(1)}$  (восстановленный ряд) с помощью операции диагонального усреднения.

**Замечание 4.1** Если пропуски находятся в конце ряда, то выделение ряда  $F_N^{(1)}$  с заполнением пропусков — это фактически прогноз этой составляющей ряда.

## 4.2 Группы пропусков

Для того чтобы предложить варианты шага 3б для заполнения векторов с пропусками, нам понадобится ввести разбиение всего множества пропусков на группы. Введем понятие группы пропусков при фиксированной длине окна  $L$ .

**Определение 4.1** *Группой пропущенных значений* временного ряда будем называть последовательность его пропущенных значений, удовлетворяющих следующим свойствам:

- 1) между любыми двумя пропущенными значениями последовательности менее  $L$  непропущенных значений;
- 2) у самого правого/левого элемента группы среди его  $L$  соседей справа/слева (если они найдутся) нет ни одного пропуска.

Таким образом, не менее чем  $L$  подряд идущих известных значений ряда разграничивают группы пропущенных значений.

**Определение 4.2** Группу, у которой самый правый/левый элемент лежит на расстояние меньше, чем  $L$ , от последнего элемента ряда, назовем *правой/левой*. Если правый или левый элемент группы совпадает с концом ряда, то группу будем называть *крайней*. Группу будем называть *внутренней*, если она не правая и не левая. Группу будем называть *сплошной*, если она состоит из подряд идущих пропущенных значений.

**Определение 4.3** *Набором векторов вложения группы пропусков* будем называть набор подряд идущих векторов вложения, часть компонент которых соответствует этим пропускам.

Далее будем рассматривать одну группу пропусков и соответствующий ей набор векторов вложения  $X_1, \dots, X_p$ , так как разные наборы векторов вложения не содержат информации друг о друге и в этом смысле независимы между собой. Поэтому шаг 3б можно применять независимо к каждой группе пропусков (каждому набору векторов).

## 4.3 Способы заполнения позиций известных компонент в наборе векторов вложения группы пропущенных значений

Предложим следующие способы решения задачи  $(\alpha)$  вычисления  $\widehat{X}_i|_{I \setminus \mathcal{P}}$ ,  $l \leq i \leq p$ .

Способ получения позиций известных значений за счет непосредственного применения оператора  $\mathbf{\Pi}_{I \setminus \mathcal{P}}^{(1)}$  для всех векторов вложения группы пропусков (естественно, множество  $\mathcal{P}$  для каждого вектора свое) по формуле  $\widehat{X}_i|_{I \setminus \mathcal{P}} = \mathbf{\Pi}_{I \setminus \mathcal{P}}^{(1)}(X_i|_{I \setminus \mathcal{P}})$  назовем «Проектор ПИ». Условия применимости этого способа и формулу для вычисления дает утверждение 2.2, где  $m = r$ ,  $\mathcal{L}^{(1)} = \mathcal{M}_r$  и  $\mathbf{R}$  — матрица, столбцами которой являются собственные вектора  $U_i$ ,  $i \in I_r$ .

Предложим альтернативный вариант, применимый для сплошных групп пропусков, крайних или внутренних. Он основан на том, что в траекторных матрицах (а  $\widehat{\mathbf{X}}$  по смыслу является

аппроксимацией траекторной матрицы ряда  $F_N^{(1)}$  значения на диагоналях с индексами  $(i, j)$ ,  $i + j = \text{Const}$ , одинаковые.

Сначала рассмотрим внутреннюю группу. В соответствующем наборе векторов  $\widehat{X}_q$ ,  $q = l, \dots, p$ , места известных компонент  $\widehat{X}_q|_{\mathcal{I} \setminus \mathcal{P}}$  образуют два “треугольника”, правый и левый. Обозначая  $\widehat{x}_{i,j}$   $i$ -й элемент вектора  $\widehat{X}_j$ , опишем метод на примере заполнения левого “треугольника”, который запишем как набор  $s$ -х “диагоналей”:  $\{(i, j) : i + j = s, l + 1 \leq s \leq l + L - 1, l \leq j \leq l + L - 2\}$ . Заметим, что “треугольник” и “диагональ” понимаются как части матрицы  $[\widehat{X}_l : \dots : \widehat{X}_p]$ .

Поскольку группа внутренняя, соседние слева вектора  $\widehat{X}_{l-m}$ ,  $m = 1, \dots, l_0$  при некотором  $l_0 \geq 1$ , не имеют пропусков и все их компоненты уже вычислены на шаге 3а. Если мы рассмотрим матрицу  $[\widehat{X}_{l-l_0} : \dots : \widehat{X}_p]$ , то метод «Компоненты соседних векторов» состоит в замене всех компонент  $s$ -й диагонали  $\widehat{x}_{i,j}$  с  $i + j = s$  и  $l - l_0 \leq j \leq l + L - 2$  на среднее арифметическое уже полученных на шаге 3а значений  $\widehat{x}_{i,j}$  с  $i + j = s$ ,  $l - l_0 \leq j \leq l - 1$ . Заметим, что метод меняет также некоторые компоненты векторов  $\widehat{X}_{l-l_0}, \dots, \widehat{X}_{l-1}$ .

Полностью аналогичная процедура делается с правым “треугольником”.

Если группа крайняя, то в заполнении нуждается только либо правый, либо левый “треугольник”.

#### 4.4 Способы восстановления позиций пропущенных компонент в наборе векторов вложения группы пропущенных значений

Предложим следующие способы решения задачи  $(\beta)$  вычисления  $\widehat{X}_i|_{\mathcal{P}}$  при условии, что  $\widehat{X}_i|_{\mathcal{I} \setminus \mathcal{P}}$  уже получены.

Для первого способа потребуем одновременного выполнения условий предложения 2.1 для каждого вектора набора и, если все условия выполнены, восстановим пропущенные значения каждого вектора по полученной выше формуле (1). Такой метод восстановления пропущенных значений набора векторов вложения с помощью формулы (1), где  $m = r$ ,  $\mathfrak{D}_m = \mathcal{M}_r$  и  $\mathbf{R}$  — матрица, столбцами которой являются собственные вектора  $U_i$ ,  $i \in I_r$ , будем называть *одновременным методом восстановления*.

Заметим, что одновременное выполнение условий предложения 2.1 для разных  $\mathcal{P}$  является очень жестким условием применимости такого способа восстановления. В частности, когда хотя бы у одного из векторов набора количество известных компонент меньше ранга ряда (или вектор вовсе состоит только из пропусков), решить задачу предложенным способом невозможно.

Предложим другой метод восстановления пропущенных значений, основанный на том, что в траекторных матрицах значения на диагоналях с индексами  $(i, j)$ ,  $i + j = \text{Const}$ , одинаковые. Поэтому можно восстановить пропуски в одном из векторов вложения, а пропуски в соседних векторах заполнить значениями, полученными по уже восстановленному вектору.

Рассмотрим набор векторов вложения  $X_l, \dots, X_p$ . Будем считать, что  $\widehat{X}_q|_{\mathcal{I} \setminus \mathcal{P}}$ ,  $q = l, \dots, p$ , уже сосчитаны одним из способов, описанных в подразделе 4.3.

Предложим несколько конкретных способов последовательного заполнения.

1) Потребуем, чтобы группа пропусков не являлась левой. В этом случае вектор  $X_l$  имеет единственный пропуск на месте своей последней координаты и, следовательно, его множество пропущенных значений  $\mathcal{P} = \{L\}$ . Поэтому можно воспользоваться следствием 2.1. Если  $e_L \notin \mathcal{M}_r$ , то мы можем получить последнее значение  $y$  вектора  $\widehat{X}_l$  как линейную комбинацию его предыдущих значений с коэффициентами, задаваемыми формулой (2). Далее используем это значение  $y$  для заполнения координат векторов  $\widehat{X}_{l+j}$  с номерами  $L - j$ ,  $j = 1, \dots, L - 1$ . После такой операции следующий вектор  $\widehat{X}_{l+1}$  будет иметь (если группа сплошная) только одно пропущенное значение на месте последней координаты. Поэтому всю операцию можно

повторить заново. И т.д. Такой способ будем называть *последовательным способом восстановления слева*. Заметим, что если группа пропусков не сплошная, то для части векторов  $X_q$ ,  $l < q < p$ , не придется применять процедуру заполнения.

2) Для *последовательного способа восстановления справа* все полностью аналогично, и сводится к заполнению первой координаты по формуле (3), задаваемой в следствии 2.2. Условиями возможности такого заполнения будут то, что группа пропусков не является правой, и  $e_1 \notin M_r$ .

3) Под *двусторонним последовательным восстановлением до середины* подразумевается использование последовательного способа восстановления слева для получения левой половины значений группы пропусков и последовательного способа восстановления справа для получения правой половины пропущенных значений.

4) Под *двусторонним последовательным восстановлением с усреднением* подразумевается использование сначала последовательного способа восстановления слева для получения значений группы пропусков, а затем последовательного способа восстановления справа. В качестве результата восстановления точек берется среднее арифметическое результатов восстановления с помощью этих двух способов. Двусторонние методы применимы только к внутренним группам.

Преимуществом последовательных методов восстановления по сравнению с одновременным являются слабые условия применимости, а недостатком — то, что ошибка восстановления может накапливаться.

**Замечание 4.2** Если сплошная группа пропусков длины  $t$  является крайней правой (других групп пропусков нет) и применяются метод заполнения «Компоненты соседних векторов» и «Последовательный способ восстановления слева», то результат будет совпадать с базовым методом SSA, примененным к первым  $N - t$  точкам ряда с последующим рекуррентным прогнозом (см. [2, 5]) выделенной компоненты ряда на  $t$  точек вперед.

#### 4.5 Формальная модификация алгоритма «Гусеница»-SSA для рядов, имеющих пропущенные значения

Кроме приведенных выше способов анализа рядов с пропусками, основанных на свойствах выделяемого ряда, можно предложить также формальный вариант заполнения пропусков. Он состоит в замене скалярного произведения векторов на аналогичную формальную процедуру, применимую к векторам с пропусками.

##### 4.5.1 Операция “ \* ”

Введем операцию “ \* ”, которая будет численно заменять обычное скалярное произведение для векторов с пропусками. Рассмотрим два вектора  $A = (a_1, \dots, a_n)^T$  и  $B = (b_1, \dots, b_n)^T$  и их множества индексов пропущенных компонент  $\mathcal{A}$  и  $\mathcal{B}$  соответственно. Предположим, что  $|\mathcal{A} \cup \mathcal{B}| < n$ . Тогда зададим операцию “ \* ” формулой:

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

(Здесь запись  $(A, B)^*$  введена для удобства, это не скалярное произведение.) Ясно, что результат операции “ \* ” численно совпадает со стандартным скалярным произведением в  $\mathbb{R}^n$ , если умножать вектора без пропусков. Для матриц  $\mathbf{A} = [A_1 : \dots : A_k]$  и  $\mathbf{B} = [B_1 : \dots : B_l]$ ,  $A_i, B_j \in \mathbb{R}^n$ , определим операцию “ \* ” по формуле  $\mathbf{A} * \mathbf{B} = \{(A_i, B_j)^*\}_{i=1, j=1}^{k, l}$ . Если все вектора обеих матриц не содержат пропущенных компонент, то  $\mathbf{A} * \mathbf{B} = \mathbf{AB}$ .

#### 4.5.2 Модификации с использованием операции “ \* ”

Предложим следующие модификации.

1. В качестве матрицы  $\tilde{\mathbf{S}}$  на шаге 2 можно взять матрицу, вычисляемую по формуле  $\tilde{\mathbf{S}} = \mathbf{X} * \mathbf{X}^T$ , где  $\mathbf{X}$  — траекторная матрица ряда  $F_N$ , содержащая пропуски.

Предложенный выше способ обобщим следующим образом: рассмотрим величину  $\tau$ ,  $0 \leq \tau \leq L$ , которую назовем *порогом количества пропущенных компонент*. Далее образуем матрицу  $\tilde{\mathbf{X}}_{(\tau)}$ , состоящую из векторов вложения, содержащих не более  $\tau$  пропущенных компонент, и  $\tilde{\mathbf{S}} = \tilde{\mathbf{X}}_{(\tau)} * \tilde{\mathbf{X}}_{(\tau)}^T$ . Заметим, что матрица  $\tilde{\mathbf{X}}_{(0)}$  совпадает с матрицей  $\tilde{\mathbf{X}}$ , составленной из векторов без пропусков, а  $\tilde{\mathbf{X}}_{(L)} = \mathbf{X}$ .

**Замечание 4.3** Для стационарных рядов известна модификация базового метода SSA под названием Теплицев вариант SSA. С учетом введения операции “ \* ” для обобщения этой модификации на случай рядов с пропусками матрица  $\tilde{\mathbf{S}}$  будет иметь вид  $\tilde{\mathbf{S}} = K\tilde{\mathbf{C}}$ , где  $\tilde{\mathbf{C}} = \{\tilde{c}_{ij}\}$  и для  $F^{(k,g)} = (f_k, \dots, f_g)^T$

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

(здесь  $N - |i - j|$  — общее число пар, которые вообще могут войти в произведение, т.е. размерность умножаемых векторов).

Данный способ построения матрицы  $\tilde{\mathbf{S}}$  назовем «Теплицевым» (в отличие от основного, который по аналогии с шагом 2 базового алгоритма в случае, когда пропусков нет, можно назвать «SVD» — сингулярное разложение).

2. В качестве проекций векторов вложения на шаге 3, вне зависимости от того, имеют ли они пропуски, можно взять

$$\hat{X}_i = \sum_{k \in I_r} (X_i, U_k) * U_k, \quad i \in \mathcal{C}. \quad (10)$$

Здесь  $(X_i, U_k) *$  имеет смысл приближения  $k$ -й главной компоненты вектора  $X_i$ . Поэтому такой способ назовем «Главные компоненты».

Заметим, что одновременное использование модификаций “Теплицев” и “Главные компоненты” было предложено для анализа стационарных временных рядов в работе [8]. Для нестационарных или коротких рядов эти модификации дают неудовлетворительные результаты.

#### 4.6 Пример

Для демонстрации работы методов заполнения пропусков рассмотрим классический временной ряд длины 144, представляющий собой количество перевезенных пассажиров на международных авиалиниях, начиная с января 1949 года (данные впервые опубликованы в [4]). Каждая точка соответствует объему пассажироперевозок за 1 месяц, числа даны в тысячах человек. Поставим 12 искусственных пропусков, начиная с 68-й точки (т.е. будем считать значения неизвестными в течение года с августа 1954 года). Восстановленные значения этих пропусков можно будет сравнить с известными значениями. Также добавим к ряду 12 пропусков в конце. Их заполнение будет соответствовать прогнозу ряда. Полученный временной ряд изображен на рис. 1.

Первый вопрос, какую выбрать длину окна  $L$ . Общая рекомендация для анализа ряда без пропусков была бы такая: взять длину окна, близкую к половине длины ряда и кратную предполагаемому периоду 12 месяцев. Таким критериям соответствует длина окна, равная 72. Однако при таком выборе  $L$  все вектора вложения будут содержать пропуски. Чтобы обойтись

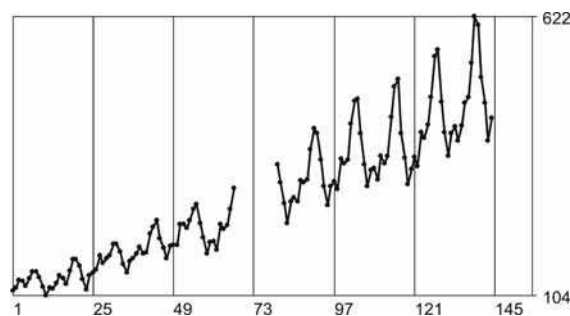


Рис. 1: Исходный временной ряд с пропусками

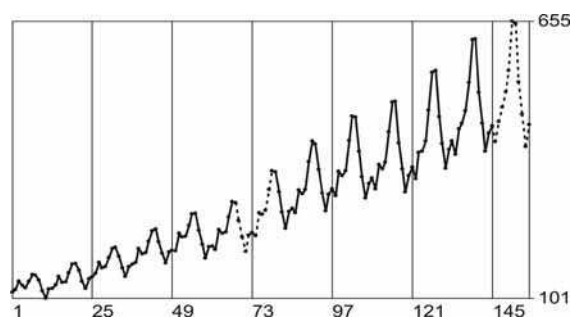


Рис. 2: Восстановленный ряд с заполненными пропусками

без использования операции “ \* ”, выберем длину окна поменьше. Выбор  $L = 36$  даст нам 62 вектора вложения без пропусков.

Анализ собственных векторов  $U_1, \dots, U_L$  (для реальных рядов, как правило,  $d = L$ ) показывает, что собственные вектора с номерами 1, 6 и 9 соответствуют тренду, а собственные вектора с номерами 2–5, 7–8 и 10–13 соответствуют сезонной компоненте ряда. Остальные собственные вектора можно отнести к шуму. Поэтому, чтобы восстановить детерминированную составляющую ряда, выберем  $r = 13$  и  $I_r = \{1, 2, \dots, 13\}$ .

К первой сплошной внутренней группе пропусков применимы все 10 вариантов выполнения шага 3b: два варианта для действия ( $\alpha$ ) («Проектор ПИ» и «Компоненты соседних векторов»), и для каждого из них 5 вариантов для действия ( $\beta$ ) (одновременное заполнение пропущенных значений и четыре разновидности последовательного заполнения). Сравнение результатов восстановления с известными значениями показывает, что наилучшее восстановление произошло для варианта «Проектор ПИ» с одновременным заполнением пропусков. При этом средне-квадратическое отклонение в пропущенных значениях равняется примерно 6, а в остальных точках — 4.75, что ненамного меньше.

Применим эти же варианты и для заполнения второй группы. Результат изображен на рис. 2. На графике пунктирной линией обозначен восстановленный ряд в области восстановления пропусков и прогнозирования.

## Список литературы

- [1] Голяндина Н.Э. Метод «Гусеница»-SSA: анализ временных рядов: Учеб. пособие. СПб: Изд-во С.-Петербур. Ун-та, 2004. 76 с.

- [2] *Голяндина Н.Э.* Метод «Гусеница»-SSA: прогноз временных рядов: Учеб. пособие. СПб: Изд-во С.-Петербур. Ун-та, 2004. 52 с.
- [3] *Главные компоненты временных рядов: метод «Гусеница»* / Под ред. Д.Л.Данилова, А.А.Жиглявского. СПб: Пресском, 1997. 308 с.
- [4] *Brown R.G.* Smoothing, forecasting and prediction of discrete time series, New Jersey, Prentice-Hall, 1963.
- [5] *Golyandina N., Nekrutkin V., and Zhigljavsky A.* Analysis of Time Series Structure: SSA and Related Techniques. London: Chapman & Hall/CRC, 2001. 305 p.
- [6] *Elsner, J., Tsonis, A., 1996.* Singular Spectrum Analysis. A New Tool in Time Series Analysis. New York: Plenum Press.
- [7] *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.
- [8] *Schoellhamer D.H.* Singular spectrum analysis for time series with missing data. *Geophysical Research Letters*, v. 28, no. 16, 2001. P. 3187–3190.