

Модификации метода анализа сингулярного спектра для анализа временных рядов: Circulant SSA и Generalized SSA

Погребников Николай Вадимович, гр. 21.Б04-мм

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель: д. ф.-м. н., доц. Голяндина Н.Э.

Санкт-Петербург, 2024

Введение

Временные ряды представляют собой последовательность данных, собранных или измеренных в хронологическом порядке. Понимание эволюции явлений во времени является критическим для выявления тенденций, циклов и аномалий. В этих целях были созданы методы разложения временных рядов на сумму интерпретируемых компонент такие как **SSA** [3] и его модификации: **GSSA** [4], **CiSSA** [1].

Целью работы является описание модификаций в контексте теории **SSA** и на этой основе сравнение методов по теоретическим свойствам и численно.

Метод SSA. Алгоритм: разложение

Для временного ряда $X = (x_1, \dots, x_N)$ выбирается длина окна L , $1 < L < N$ и определяется $K = N - L + 1$. Строится L -траекторная матрица \mathbf{X} , состоящая из столбцов вида $\mathbf{X}_i = (x_{i-1}, \dots, x_{i+L-2})^T$, $1 \leq i \leq K$. Пусть $\mathbf{S} = \mathbf{X}\mathbf{X}^T$, $\lambda_1, \dots, \lambda_L$ — собственные числа матрицы \mathbf{S} , взятые в убывающем порядке.

Определение 1

Сингулярным разложением называется представление матрицы в виде:

$$\mathbf{X} = \mathbf{X}_1 + \dots + \mathbf{X}_d = \sum_{i=1}^d \sqrt{\lambda_i} U_i V_i^T, \text{ где} \quad (1)$$

U_1, \dots, U_L — ортонормированная система собственных векторов матрицы \mathbf{S} , $d = \max \{i : \lambda_i > 0\}$ и $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$.

Набор $(\sqrt{\lambda_i}, U_i, V_i^T)$ называется i -й собственной тройкой.

Метод SSA. Алгоритм: восстановление

На основе разложения (1) производится процедура группировки, которая делит все множество индексов $\{1, \dots, d\}$ на m непересекающихся подмножеств I_1, \dots, I_m . Пусть $I = \{i_1, \dots, i_p\}$, тогда $\mathbf{X}_I = \mathbf{X}_{i_1} + \dots + \mathbf{X}_{i_p}$. Такие матрицы вычисляются для каждого $I = I_1, \dots, I_m$.

В результате получаются матрицы $\mathbf{X}_{I_1}, \dots, \mathbf{X}_{I_m}$, для каждой из которых проводится операция диагонального усреднения, составляющая ряды длины N : X_1, \dots, X_m .

При этом, $X_1 + \dots + X_m = X$.

Алгоритм **GSSA** сильно схож с базовым **SSA**. Пусть $N > 2$, вещественнозначный временной ряд $X = (x_1, \dots, x_N)$ длины N . Фиксируется параметр $\alpha \geq 0$, отвечающий за веса:

$$w^{(\alpha)} = (w_1, w_2, \dots, w_L) = \left(\left| \sin \left(\frac{\pi n}{L+1} \right) \right| \right)^\alpha, \quad n = 1, 2, \dots, L.$$

Для временного ряда $X = (x_1, \dots, x_N)$ выбирается длина окна L , $1 < L < N$ и определяется $K = N - L + 1$. Строится L -траекторная матрица $\mathbf{X}^{(\alpha)}$, состоящая из столбцов вида $X_i^{(\alpha)} = (w_1 x_{i-1}, \dots, w_L x_{i+L-2})^T$, $1 \leq i \leq K$.

Остальные действия те же самые, что и в **SSA**.

Замечание 1

При $\alpha = 0$, **GSSA** — в точности базовый алгоритм **SSA**.

Сравнение SSA и GSSA. Линейные фильтры 1

Определение 2

Пусть бесконечный временной ряд $X = (\dots, x_{-1}, x_0, x_1, \dots)$. Линейный конечный фильтр — это оператор Φ , который преобразует временной ряд X в новый по следующему правилу:

$$y_j = \sum_{i=-r_1}^{r_2} h_i x_{j-i}; \quad r_1, r_2 < \infty.$$

Связанные определения:

- h_i — импульсная характеристика фильтра;
- $H_\Phi(z) = \sum_{i=-r_1}^{r_2} h_i z^{-i}$ — передаточная функция;
- $A_\Phi(\omega) = |H_\Phi(e^{i2\pi\omega})|$ — АЧХ;
- $\phi_\Phi(\omega) = \text{Arg}(H_\Phi(e^{i2\pi\omega}))$ — ФЧХ.

Пример. При применении фильтра Φ на $X_{\cos} = \cos 2\pi\omega n$, получается ряд $y_j = A_\Phi(\omega) \cos(2\pi\omega j + \phi_\Phi(\omega))$.

Сравнение SSA и GSSA. Линейные фильтры 2

Пусть $X = (x_1, \dots, x_N)$ — временной ряд длины N , $(\sqrt{\lambda}, U, V)$ — одна из собственных троек разложения методом **SSA**.

$U = (u_1, \dots, u_L)$.

Тогда компонента временного ряда \tilde{X} , восстановленная с использованием собственной тройки $(\sqrt{\lambda}, U, V)$, для средних точек (индексы от L до K) имеет вид:

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

Таким образом, имеется представление алгоритма **SSA** через линейные фильтры.

Аналогичное представления для **GSSA**:

$$\tilde{x}_s = \sum_{j=-(L-1)}^{L-1} \left(\sum_{k=1}^{L-|j|} u_k^{(\alpha)} u_{k+|j|}^{(\alpha)} w_k / \sum_{i=1}^L w_i \right) x_{s-j}, \quad L \leq s \leq K.$$

Сравнение SSA и GSSA. Пример

$X = X_{\sin} + X_{\cos} = \sin\left(\frac{2\pi}{12}n\right) + \frac{1}{2} \cos\left(\frac{2\pi}{19}n\right)$. $N = 96 \cdot 2 - 1$, $L = 48$.

Фильтры для различных α

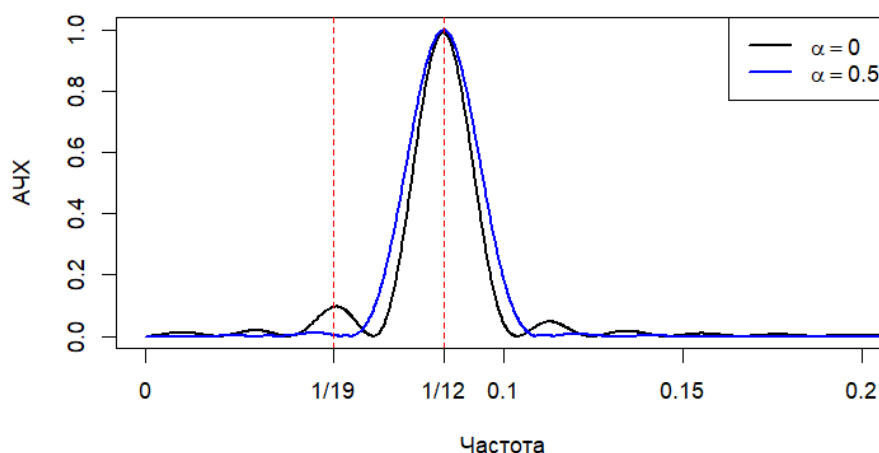


Рис. 1: Ряд $X = X_{\sin} + X_{\cos}$. АЧХ фильтров, отвечающих за $X_{\sin} = \sin\left(\frac{2\pi}{12}n\right)$, при разных α

Сравнение SSA и GSSA. Пример, продолжение

Метод/Ошибка	X_{\sin}	X_{\cos}	X
SSA	5.15e-03	5.15e-03	6.01e-30
GSSA, $\alpha = \frac{1}{2}$	3.68e-04	3.68e-04	9.53e-30

Таблица 1: MSE разложений ряда $X = X_{\sin} + X_{\cos}$ для **SSA** и **GSSA** с $\alpha = \frac{1}{2}$

Добавим к X шумовую компоненту:

$$X = X_{\sin} + X_{\cos} + X_{\text{noise}} = \sin\left(\frac{2\pi}{12}x\right) + \frac{1}{2} \cos\left(\frac{2\pi}{19}x\right) + \varepsilon_n, \text{ где } \varepsilon_n \sim N(0, 0.1^2).$$

Метод	X_{\sin}	X_{\cos}	X
SSA	5.68e-03	5.44e-03	7.48e-04
GSSA, $\alpha = \frac{1}{2}$	1.21e-03	1.25e-03	1.04e-03

Таблица 2: MSE разложений ряда $X = X_{\sin} + X_{\cos} + X_{\text{noise}}$ для **SSA** и **GSSA** с $\alpha = \frac{1}{2}$

Сравнение SSA и GSSA. Выводы

По теоретическим результатам и примерам можно сделать понять, что **GSSA** позволяет улучшить разделимость периодических компонент ряда. Однако, вместе с тем, разложение будет захватывать больше шума в сравнении с базовым **SSA**.

Метод CiSSA. Алгоритм: разложение

Как и в **SSA** считается \mathbf{X} , по которой строится $\hat{\mathbf{C}}_L$:

$$\hat{c}_m = \frac{L-m}{L}\hat{\gamma}_m + \frac{m}{L}\hat{\gamma}_{L-m}, \quad \hat{\gamma}_m = \frac{1}{N-m} \sum_{t=1}^{N-m} x_t x_{t+m}, \quad m = 0 : L-1.$$

$$\hat{\mathbf{C}}_L = \begin{pmatrix} \hat{c}_1 & \hat{c}_2 & \dots & \hat{c}_L \\ \hat{c}_2 & \hat{c}_1 & \dots & \hat{c}_{L-1} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{c}_L & \hat{c}_{L-1} & \dots & \hat{c}_1 \end{pmatrix}.$$

Собственные числа и вектора матрицы $\hat{\mathbf{C}}_L$, задаются по формулам:

$$U_k = L^{-1/2}(u_{k,1}, \dots, u_{k,L}), \quad \text{где } u_{k,j} = \exp\left(-i2\pi(j-1)\frac{k-1}{L}\right),$$

$$\lambda_{L,k} = \sum_{m=0}^{L-1} \hat{c}_m \exp\left(i2\pi m \frac{k-1}{L}\right), \quad k = 1 : L.$$

Метод CiSSA. Алгоритм: восстановление

Для каждой частоты $w_k = \frac{k-1}{L}$, $k = 2 : \lfloor \frac{L+1}{2} \rfloor$, есть два собственных вектора: U_k и U_{L+2-k} . За частоту w_1 отвечает один собственный вектор — U_1 . Если L — четное, то частоте $w_{\frac{L}{2}+1}$ будет соответствовать один вектор $U_{\frac{L}{2}+1}$.

Следовательно, индексы разбиваются на элементарную группировку следующим образом:

$$B_1 = \{1\}; \quad B_k = \{k, L+2-k\}, \quad \text{для } k = 2 : \lfloor \frac{L+1}{2} \rfloor;$$

$$B_{\frac{L}{2}+1} = \left\{ \frac{L}{2} + 1 \right\}, \quad \text{если } L \mid 2.$$

$$\mathbf{X}_{B_k} = \mathbf{X}_k + \mathbf{X}_{L+2-k} = U_k U_k^H \mathbf{X} + U_{L+2-k} U_{L+2-k}^H \mathbf{X},$$

где U^H — это комплексное сопряжение и транспонирование вектора U . Далее идет группировка по диапазонам интересующих частот, после чего следует диагональное усреднение.

Определение 3

Разложение

$$x_n = c_0 + \sum_{k=1}^{\lfloor \frac{N+1}{2} \rfloor} (c_k \cos(2\pi nk/N) + s_k \sin(2\pi nk/N)), \quad (2)$$

где $1 \leq n \leq N$ и $s_{N/2} = 0$ для четного N , называется разложением Фурье ряда X .

Замечание 2

$U_k U_k^H + U_{L+2-k} U_{L+2-k}^H$ является оператором проектирования на подпространство, которое порождено синусами и косинусами с частотой $w_k = \frac{k-1}{L}$. То есть, воспроизводится разложение Фурье для K векторов матрицы X . Затем вычисляется диагональное усреднение.

Метод CiSSA. Свойства: нестационарный ряд

Для использования на нестационарных временных рядах, нужно выполнить расширения ряда (экстраполировать) [1].

Расширение временного ряда IP values

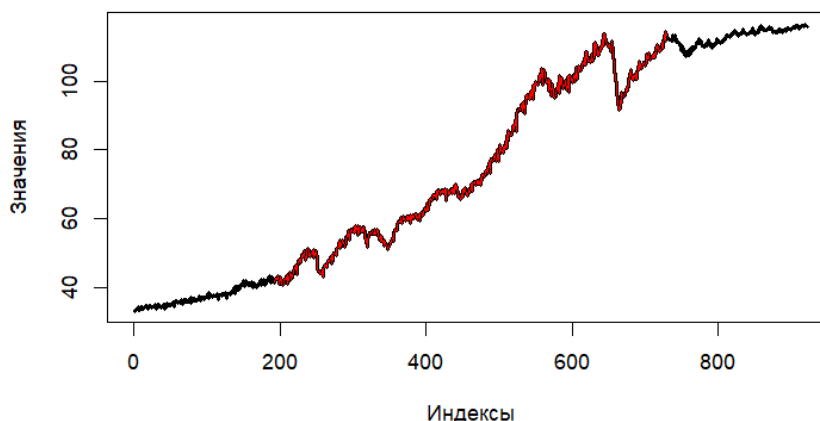


Рис. 2: Красный — настоящий ряд, черный — расширенный

Так, алгоритм лучше выделяет нелинейную составляющую.

Определение 4

Есть метод разделения ряда на компоненты с параметрами Θ , ряд $X = X^{(1)} + X^{(2)}$. \exists набор параметров $\hat{\Theta}$, L , N , что при разделении ряда на компоненты этим методом, $\hat{X}^{(1)}$ является оценкой $X^{(1)}$, при этом, $\text{MSE}(X^{(1)}, \hat{X}^{(1)}) = 0$. Тогда ряды $X^{(1)}$ и $X^{(2)}$ точно разделимы данным методом.

Определение 5

Есть метод разделения ряда на компоненты с параметрами Θ , ряд $X = X^{(1)} + X^{(2)}$. \exists набор параметров $\hat{\Theta}$ и $L = L(N)$, $N \rightarrow \infty$, что при разделении ряда на компоненты этим методом, $\hat{X}^{(1)}$ является оценкой $X^{(1)}$, при этом, $\text{MSE}(X^{(1)}, \hat{X}^{(1)}) \rightarrow 0$. Тогда ряды $X^{(1)}$ и $X^{(2)}$ называются асимптотически $L(N)$ -разделимыми данным методом.

Фиксируем временной ряд $X = X_1 + X_2 =$
 $= A_1 \exp(\alpha_1 n) \cos(2\pi w_1 n + \varphi_1) + A_2 \exp(\alpha_2 n) \cos(2\pi w_2 n + \varphi_2)$.

Условия точной разделимости X для разложения Фурье:

$Nw_1, Nw_2 \in \mathbb{N}$, $w_1 \neq w_2$, $\alpha_1 = \alpha_2 = 0$.

Условия точной разделимости X для **CiSSA**:

$Lw_1, Lw_2 \in \mathbb{N}$, $w_1 \neq w_2$, $\alpha_1 = \alpha_2 = 0$.

Условия точной разделимости X для **SSA**:

$Lw_1, Lw_2, Kw_1, Kw_2 \in \mathbb{N}$, $w_1 \neq w_2$, $A_1 \neq A_2$, $\alpha_1 \neq \alpha_2$.

Таким образом, условия на разделение косинусов, слабее у методов **CiSSA** и Фурье, чем у **SSA**. Однако **SSA** может точно отличать друг от друга больше классов функций.

Сравнение SSA, Фурье, CiSSA. Асимптотическая разделимость

Асимптотически разделимы в методе **SSA** полиномы, гармонические функции, не удовлетворяющие условиям точной разделимости, экспоненты [3].

Замечание 3

*Для **SSA** существуют алгоритмы улучшения разделимости, например, EOSSA и FOSSA [2]. По заданному набору компонент, они позволяют более точно отделять компоненты.*

В алгоритме разложения **CiSSA** (Фурье) увеличение длины окна $L(N)$ изменяет сетку частот. Это означает, что даже если не удастся подобрать такое $L(N)$, при котором косинус будет точно отделим, его постепенное увеличение позволит приблизить частоты сетки к частоте компоненты. В итоге, можно снизить ошибку выделения нужной компоненты, учитывая соседние частоты.

17/21 Погребников Николай Вадимович, гр. 21.Б04-мм Модификации метода SSA

Сравнение SSA, Фурье, CiSSA. Выделение тренда

Любая непериодическая компонента будет отвечать частотам, близким к нулю. Из-за этого алгоритмы **CiSSA** и разложение Фурье не смогут отличить друг от друга две непериодики.

Пример. Рассмотрим ряд

$$X = X_c + X_e + X_{\sin} + X_{\cos} = 1 + e^{\frac{x}{100}} + \sin \frac{2\pi}{12}x + \frac{1}{2} \cos \frac{2\pi}{3}x.$$

Метод	Параметры	MSE($X_c + X_e$)	MSE(X_{\sin})	MSE(X_{\cos})	MSE(X)
SSA	$L = 96, K = 96$	5.0e-03	8.9e-07	5.2e-05	4.4e-03
SSA EOSSA, $r = 7$	$L = 96, K = 96$	1.7e-28	1.6e-29	8.7e-30	1.6e-28
Fourier	$N = 96 \cdot 2$	1.1e-01	6.1e-04	6.8e-03	1.1e-01
Fourier extended	$N = 96 \cdot 2$	1.4e-03	1.3e-03	8.4e-03	9.6e-03
CiSSA	$L = 96$	5.3e-02	1.6e-05	4.9e-04	4.4e-02
CiSSA extended	$L = 96$	5.0e-04	2.1e-04	1.1e-03	6.0e-04

Таблица 3: MSE разложений ряда $X = X_c + X_e + X_{\sin} + X_{\cos}$

По таблице 3 видно, что расширение ряда негативно повлияло на выделение периодики и положительно на трендовую составляющую (непериодику).

18/21 Погребников Николай Вадимович, гр. 21.Б04-мм Модификации метода SSA

Сравнение SSA, Фурье, CiSSA. Выводы 1

Метод/Условие	cos, $Lw \in \mathbb{N},$ $Kw \in \mathbb{N}$	cos, $Lw \in \mathbb{N},$ $Kw \notin \mathbb{N}$	cos, $Lw \notin \mathbb{N},$ $Kw \notin \mathbb{N}$	X_{np1}	X_{np}	group
SSA	+	→	→	→	→	—
SSA EOSSA	+	→	→	→	→	+
CiSSA	+	+	→	—	—	+
CiSSA extended	+	+	→	→	—	+

Таблица 4: Преимущества и недостатки методов **SSA**, **CiSSA**

Метод/Условие	cos, $Nw \in \mathbb{N}$	cos, $Nw \notin \mathbb{N}$	X_{np1}	X_{np}	group
Fourier	+	→	—	—	+
Fourier extended	+	→	→	—	+

Таблица 5: Преимущества и недостатки методов Fourier

Сравнение SSA, Фурье, CiSSA. Выводы 2

По полученным результатам, можно следующие выводы:

- ❶ Алгоритм **CiSSA** работает лучше разложения Фурье;
- ❷ Если понятно, что ряд состоит только из периодических компонент, стоит использовать **CiSSA** без процедуры расширения, поскольку она делает ошибки разделений периодики больше. И напротив, если есть непериодичность, лучше расширять ряд;
- ❸ Если данные зашумлены или имеется непериодичность, алгоритм **SSA** с улучшением делимости справляется в среднеквадратичном лучше **CiSSA** с расширением ряда или без.

- [1] Juan Bogalo, Pilar Poncela, and Eva Senra. Circulant singular spectrum analysis: A new automated procedure for signal extraction. *Signal Processing*, 177, 2020.
- [2] Nina Golyandina, Pavel Dudnik, and Alex Shlemov. Intelligent identification of trend components in singular spectrum analysis. *Algorithms*, 16(7):353, 2023.
- [3] Nina Golyandina, Vladimir Nekrutkin, and Anatoly Zhigljavsky. *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman and Hall/CRC, 2001.
- [4] Jialiang Gu, Kevin Hung, Bingo Wing-Kuen Ling, Daniel Hung-Kay Chow, Yang Zhou, Yaru Fu, and Sio Hang Pun. Generalized singular spectrum analysis for the decomposition and analysis of non-stationary signals. *Journal of the Franklin Institute*, Accepted/In Press, 2024.