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

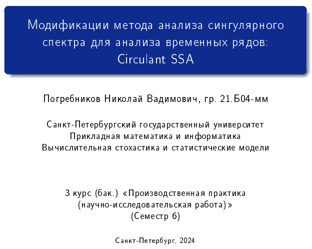
Погребников Николай Вадимович, гр. 21.Б04-мм

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

3 курс (бак.) «Производственная практика
(научно-исследовательская работа)»
(Семестр 6)

Санкт-Петербург, 2024

Модификации метода **SSA**



Научный руководитель д. ф.-м. н., доц. Голяндина Нина Эдуардовна, кафедра статистического моделирования

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** [Golyandina et al., 2001] и его модификация **CiSSA** [Bogalo et al., 2020].

Перед началом исследования были поставлены следующие цели:

1. Ознакомиться с алгоритмом **CiSSA**;
2. Реализовать алгоритм **CiSSA** на языке R;
3. Сравнить алгоритмы **SSA**, разложение Фурье и **CiSSA**.

Модификации метода **SSA**

└─Введение

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** [Golyandina et al., 2001] и его модификация **CiSSA** [Bogalo et al., 2020].

Перед началом исследования были поставлены следующие цели:

1. Ознакомиться с алгоритмом **CiSSA**;
2. Реализовать алгоритм **CiSSA** на языке R;
3. Сравнить алгоритмы **SSA**, разложение Фурье и **CiSSA**.

Сингулярный спектральный анализ (**SSA** [Golyandina et al., 2001]) — метод, целью которого является разложение оригинального ряда на сумму небольшого числа интерпретируемых компонент, таких как медленно изменяющаяся тенденция (тренд), колебательные компоненты (сезонность) и “структурный” шум. В данном исследовании рассматривается математическая составляющая вариации алгоритма **SSA** — circulant singular spectrum analysis (**CiSSA**), предложенная в статье [Bogalo et al., 2020], а также сравнение базового метода и циркулярного, применение их на языке R.

Метод SSA. Алгоритм: разложение

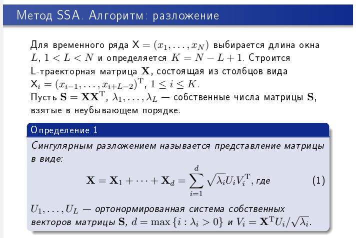
Для временного ряда $\mathsf{X} = (x_1, \ldots, x_N)$ выбирается длина окна L, $1 < L < N$ и определяется $K = N - L + 1$. Строится L-траекторная матрица $\mathbf{X}$, состоящая из столбцов вида $\mathsf{X}_i = (x_{i-1}, \ldots, x_{i+L-2})^{\mathrm{T}}$, $1 \le i \le K$.
Пусть $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$,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

Определение 1

Сингулярным разложением называется представление матрицы в виде:

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

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



Полезным свойством является то, что матрица $\mathbf{X}$ имеет одинаковые элементы на антидиагоналях. Таким образом, L-траекторная матрица является ганкелевой.
Набор $(\sqrt{\lambda_i}, U_i, V_i^{\mathrm{T}})$ называется i-й собственной тройкой разложения $\mathbf{X}$.

Метод SSA. Алгоритм: восстановление

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

В результате получаются матрицы $\mathbf{X}_{\mathbf{I_1}}, \dots, \mathbf{X}_{\mathbf{I_m}}$, для каждой из которых проводится операция диагонального усреднения, составляющая ряды длины N: $\mathsf{X}_1, \dots, \mathsf{X}_m$.
При этом, $\mathsf{X}_1 + \dots + \mathsf{X}_m = \mathsf{X}$.

Модификации метода **SSA**

└─Метод SSA. Алгоритм: восстановление

Метод SSA. Алгоритм: восстановление

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

В результате получаются матрицы $\mathbf{X}_{\mathbf{I_1}}, \dots, \mathbf{X}_{\mathbf{I_m}}$, для каждой из которых проводится операция диагонального усреднения, составляющая ряды длины N: $\mathsf{X}_1, \dots, \mathsf{X}_m$.
При этом, $\mathsf{X}_1 + \dots + \mathsf{X}_m = \mathsf{X}$.

Диагональное усреднение для каждой антидиагонали усредняет значения элементов матрицы.
Применяя данную операцию к матрицам $\mathbf{X}_{\mathbf{I_1}}, \dots, \mathbf{X}_{\mathbf{I_m}}$, получаются m новых рядов: $\mathsf{X}_1, \dots, \mathsf{X}_m$. При этом, $\mathsf{X}_1 + \dots + \mathsf{X}_m = \mathsf{X}$.

Метод SSA. Свойства: точная разделимость

Пусть временной ряд $\mathsf{X} = \mathsf{X}^{(1)} + \mathsf{X}^{(2)}$ и задачей является нахождение этих слагаемых.
Будем говорить, что ряд X точно разделим на $\mathsf{X}^{(1)}$ и $\mathsf{X}^{(2)}$, если существует такое сингулярное разложение траекторной матрицы $\mathbf{X}$ ряда X, что его можно разбить на две части, являющиеся сингулярными разложениями траекторных матриц рядов $\mathsf{X}^{(1)}, \mathsf{X}^{(2)}$ [Golyandina et al., 2001].

Метод SSA. Свойства: точная разделимость

Пусть временной ряд $\mathsf{X} = \mathsf{X}^{(1)} + \mathsf{X}^{(2)}$ и задачей является нахождение этих слагаемых.
Будем говорить, что ряд X точно разделим на $\mathsf{X}^{(1)}$ и $\mathsf{X}^{(2)}$, если существует такое сингулярное разложение траекторной матрицы $\mathbf{X}$ ряда X, что его можно разбить на две части, являющиеся сингулярными разложениями траекторных матриц рядов $\mathsf{X}^{(1)}, \mathsf{X}^{(2)}$ [Golyandina et al., 2001].

Условия точной разделимости выводятся из понятий слабо L-разделимых рядов и сильно L-разделимых рядов [Golyandina et al., 2001]. Стоит отметить, что точная разделимость для $\cos$ достигается, если $Lw \in \mathbb{N}$, $Kw \in \mathbb{N}$, где w — частота.

Однако условия точной разделимости достаточно жесткие и вряд ли выполнимы в реальных задачах. Тогда появляется такое понятие, как асимптотическая разделимость.

Метод SSA. Свойства: асимптотическая разделимость

$$\rho_{i,j}^{(M)} = \frac{\left(\mathsf{X}_{i,i+M-1}^{(1)}, \mathsf{X}_{j,j+M-1}^{(2)}\right)}{\left\|\mathsf{X}_{i,i+M-1}^{(1)}\right\| \left\|\mathsf{X}_{j,j+M-1}^{(2)}\right\|}.$$

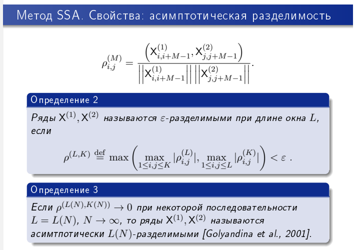
Определение 2

Ряды $\mathsf{X}^{(1)}, \mathsf{X}^{(2)}$ называются ε-разделимыми при длине окна L, если

$$\rho^{(L,K)} \stackrel{\mathrm{def}}{=} \max\left(\max_{1\le i,j\le K} |\rho_{i,j}^{(L)}|, \max_{1\le i,j\le L} |\rho_{i,j}^{(K)}|\right) < \varepsilon\ .$$

Определение 3

Если $\rho^{(L(N),K(N))} \to 0$ при некоторой последовательности $L = L(N)$, $N \to \infty$, то ряды $\mathsf{X}^{(1)}, \mathsf{X}^{(2)}$ называются асимтпотически $L(N)$-разделимыми [Golyandina et al., 2001].



Для любого ряда X длины N определим $\mathsf{X}_{i,j} = (x_{i-1}, \cdots, x_{j-1}),\ 1 \le i \le j < N$. Пусть $\mathsf{X}^{(1)} = (x_0^{(1)}, \ldots, x_{N-1}^{(1)}), \mathsf{X}^{(2)} = (x_0^{(2)}, \ldots, x_{N-1}^{(2)})$. Тогда определим коэффициент корреляции.

Метод SSA. Свойства: асимптотическая разделимость

Замечание 1

*Для **SSA** существуют алгоритмы улучшения разделимости [Golyandina et al., 2023]. Они позволяют более точно отделять временные ряды друг от друга. В данной работе будут использоваться методы EOSSA и FOSSA.*

Метод SSA. Свойства: асимптотическая разделимость

Замечание 1

Для SSA существуют алгоритмы улучшения разделимости [Golyandina et al., 2023]. Они позволяют более точно отделять временные ряды друг от друга. В данной работе будут использоваться методы EOSSA и FOSSA.

Для нас важно, что благодаря применению улучшения разделимости мы можем делать автоматическую группировку по заданным частотам в базовом алгоритме **SSA**.

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

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

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

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

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

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

Метод CiSSA. Алгоритм: разложение

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

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

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

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

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

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

Модификация **SSA** на основе циркулярной матрицы [Bogalo et al., 2020]. Авторы метода называют её автоматизированной. Причем автоматизированная в том смысле, что компоненты ряда группируются по частотам самим алгоритмом.

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_0 отвечает один собственный вектор — U_0. 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:

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

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

$\mathbf{X}_{B_k} = \mathbf{X}_k + \mathbf{X}_{L+2-k} = U_k U_k^H \mathbf{X} + U_{L+2-k} U_{L+2-k}^H \mathbf{X}$,
где U^H — это комплексное сопряжение и транспонирование вектора U. Далее идет группировка по диапазонам интересующих частот, после чего следует диагональное усреднение.

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_0 отвечает один собственный вектор — U_0. 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:

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

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

$\mathbf{X}_{B_k} = \mathbf{X}_k + \mathbf{X}_{L+2-k} = U_k U_k^H \mathbf{X} + U_{L+2-k} U_{L+2-k}^H \mathbf{X}$,
где U^H — это комплексное сопряжение и транспонирование вектора U. Далее идет группировка по диапазонам интересующих частот, после чего следует диагональное усреднение.

Группировка будет производиться на непересекающиеся подгруппы по частотам от 0 до 0.5, поскольку частоты выше 0.5 представляют собой зеркальное отражение частот ниже 0.5. Именно поэтому объединяются матрицы $\mathbf{X}_{B_k} = \mathbf{X}_k + \mathbf{X}_{L+2-k}$.

Метод CiSSA. Свойства: связь с разложением Фурье

Определение 4

Разложение

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

где $1 \le n \le N$ 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 $w_k = \frac{k-1}{L}$. Это пространство соответствует компонентам синусоидальной структуры временного ряда, связанных с конкретной частотой, выделяемой методом.

Метод CiSSA. Свойства: связь с разложением Фурье

Определение 4

Разложение

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

где $1 \le n \le N$ 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 $w_k = \frac{k-1}{L}$. Это пространство соответствует компонентам синусоидальной структуры временного ряда, связанных с конкретной частотой, выделяемой методом.

По замечанию 2 видно, что при вычислении $\mathbf{X}_{B_k} = \mathbf{X}_k + \mathbf{X}_{L+2-k} = U_k U_k^H \mathbf{X} + U_{L+2-k} U_{L+2-k}^H \mathbf{X}$, воспроизводится разложение Фурье для K векторов матрицы X. Затем вычисляется диагональное усреднение $*X_{B_k}$.

Метод CiSSA. Свойства: разделимость

Точная разделимость. Поскольку данный метод является аналогом разложения Фурье, то в смысле сильной разделимости можно точно разделить ряд, в котором одной из компонентов является $\cos(2\pi w + \varphi)$ с частотой w такой, что $Lw = k \in \mathbb{N}$, или константа.

Асимптотическая разделимость.

Определение 5

Пусть $\mathsf{X} = \mathsf{X}^{(1)} + \mathsf{X}^{(2)}$. *Существуют такие диапазоны частот* I_1 *и* I_2 *и последовательность* $L = L(N)$, $N \to \infty$, *что при них* $\mathrm{MSE}\left(\mathsf{X}^{(1)}, \mathsf{X}^{(1)}_{\textbf{CiSSA}}\right) \to 0$ *и* $\mathrm{MSE}\left(\mathsf{X}^{(2)}, \mathsf{X}^{(2)}_{\textbf{CiSSA}}\right) \to 0$, *где* MSE *— среднеквадратическая ошибка,* $\mathsf{X}^{(1)}_{\textbf{CiSSA}}$ *и* $\mathsf{X}^{(2)}_{\textbf{CiSSA}}$ *компоненты ряда, полученные алгоритмом* ***CiSSA*** *для частот* I_1 *и* I_2, *то ряды* $\mathsf{X}^{(1)}$ *и* $\mathsf{X}^{(2)}$ *называются* ***CiSSA****-асимтпотически* $L(N)$*-разделимыми.*

Метод CiSSA. Свойства: разделимость

Точная разделимость. Поскольку данный метод является аналогом разложения Фурье, то в смысле сильной разделимости можно точно разделить ряд, в котором одной из компонентов является $\cos(2\pi w + \varphi)$ с частотой w такой, что $Lw = k \in \mathbb{N}$, или константа.

Асимптотическая разделимость.

Определение 5

Пусть $\mathsf{X} = \mathsf{X}^{(1)} + \mathsf{X}^{(2)}$. *Существуют такие диапазоны частот* I_1 *и* I_2 *и последовательность* $L = L(N)$, $N \to \infty$, *что при них* $\mathrm{MSE}\left(\mathsf{X}^{(1)}, \mathsf{X}^{(1)}_{\textbf{CiSSA}}\right) \to 0$ *и* $\mathrm{MSE}\left(\mathsf{X}^{(2)}, \mathsf{X}^{(2)}_{\textbf{CiSSA}}\right) \to 0$, *где* MSE *— среднеквадратическая ошибка,* $\mathsf{X}^{(1)}_{\textbf{CiSSA}}$ *и* $\mathsf{X}^{(2)}_{\textbf{CiSSA}}$ *компоненты ряда, полученные алгоритмом* ***CiSSA*** *для частот* I_1 *и* I_2, *то ряды* $\mathsf{X}^{(1)}$ *и* $\mathsf{X}^{(2)}$ *называются* ***CiSSA****-асимтпотически* $L(N)$*-разделимыми.*

Асимптотическая разделимость в данном случае будет означать, что при увеличении L разбиение сетки будет увеличиваться, а значит, и частоты в сетке начнут сближаться к истинным частотам периодических компонентов (либо становиться равными им), что будет снижать ошибку вычислений.

Метод CiSSA. Свойства: эквивалентность методов

Определение 6

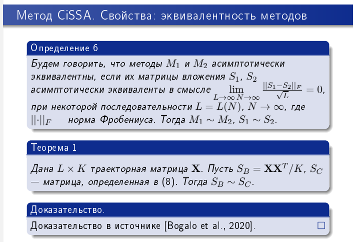
Будем говорить, что методы M_1 и M_2 асимптотически эквивалентны, если их матрицы вложения S_1, S_2 асимптотически эквиваленты в смысле $\lim\limits_{L\to\infty\, N\to\infty} \frac{||S_1 - S_2||_F}{\sqrt{L}} = 0$, *при некоторой последовательности $L = L(N)$, $N \to \infty$, где $||\cdot||_F$ — норма Фробениуса. Тогда $M_1 \sim M_2$, $S_1 \sim S_2$.*

Теорема 1

Дана $L \times K$ траекторная матрица $\mathbf{X}$. Пусть $S_B = \mathbf{X}\mathbf{X}^T/K$, S_C — матрица, определенная в (8). *Тогда $S_B \sim S_C$.*

Доказательство.

Доказательство в источнике [Bogalo et al., 2020]. □



В статье [Bogalo et al., 2020] говорится, что асимптотически методы **SSA** и **CiSSA** эквивалентны и в доказательство приводится теорема.

Метод CiSSA. Свойства: применимость к нестационарным рядам

Алгоритм **CiSSA**, описанный ранее, изначально применим только к стационарным временным рядам. Однако, как утверждается авторами статьи [Bogalo et al., 2020], для использования на нестационарных временных рядах, нужно выполнить расширение ряда. Эта процедура позволяет предсказать значения временного ряда за его пределами (экстраполяция) как в правом, так и в левом направлениях на заданное число шагов H. Таким образом, трендовая (нелинейная) компонента ряда будет выделяться заметно лучше.

Метод CiSSA. Свойства: применимость к нестационарным рядам

Алгоритм **CiSSA**, описанный ранее, изначально применим только к стационарным временным рядам. Однако, как утверждается авторами статьи [Bogalo et al., 2020], для использования на нестационарных временных рядах, нужно выполнить расширение ряда. Эта процедура позволяет предсказать значения временного ряда за его пределами (экстраполяция) как в правом, так и в левом направлениях на заданное число шагов H. Таким образом, трендовая (нелинейная) компонента ряда будет выделяться заметно лучше.

Формальное определение стационарности ряда можно увидеть в отчёте данной работы [Pogrebnikov, 2024]. Стационарный ряд — это такой временной ряд, в котором изменения происходят вокруг некоторого среднего значения, и это среднее остаётся более-менее постоянным на протяжении всего ряда.

Сама процедура расширения ряда X производится с использованием авторегрессионной (AR) модели.

Сравнение алгоритмов. SSA, разложение Фурье, CiSSA

Для начала будем рассматривать разделимость рядов без шума, затем с шумом. В сравнении будут присутствовать пять различных методов: базовый **SSA**, **SSA** с использованием EOSSA для улучшения разделимости, разложения Фурье, базового **CiSSA** и **CiSSA** с расширением ряда. Для наглядного отображения преимуществ каждого из этих методов составлена таблица 1.

Метод/Условие	cos, $Lw = k \in \mathbb{N}$, $Kw = k \in \mathbb{N}$	cos, $Lw = k \in \mathbb{N}$, $Kw = k \notin \mathbb{N}$	cos, $Lw = k \notin \mathbb{N}$, $Kw = k \notin \mathbb{N}$	$\mathrm{X}_{\mathrm{np1}}$	X_{np}	group
SSA	$+$	$\to$	$\to$	$\to$	$\to$	$-$
SSA EOSSA	$+$	$\to$	$\to$	$\to$	$\to$	$+$
Fourier	$+$	$+$	$\to$	$-$	$-$	$+$
CiSSA	$+$	$+$	$\to$	$-$	$-$	$+$
CiSSA extended	$+$	$+$	$\to$	$\to$	$-$	$+$

Таблица 1: Преимущества и недостатки ряти методов

На пересечении строк и столбцов указан знак, показывающий, достигается ли разделение компоненты: плюс (+) обозначает точное выполнение, знак стремления указывает на асимптотическое выполнение, а минус (–) — на отсутствие разделимости. Для разложения Фурье подразумевается, что $L = N$.

Обозначения:

- $\cos$ — в ряде присутствуют только периодические компоненты вида $\cos(2\pi\omega x + \varphi)$;
- $\mathrm{X}_{\mathrm{np1}}$ — одна непериодическая компонента в ряде, остальные имеют период;
- X_{np} — несколько непериодических компонент в ряде, остальные имеют период, интересует разделение между непериодическими компонентами;
- group — автоматическая группировка по заданным частотам.

Сравнение алгоритмов. Пример 1

$X = X_{\sin} + X_{\cos} = \sin\frac{2\pi}{12}x + \frac{1}{2}\cos\frac{2\pi}{3}x$, $L = 96$, $N = 96 \cdot 2$ для разложения Фурье и $N = 96 \cdot 2 - 1$ для остальных, чтобы выполнялись условия выполнения разделимости частот. Сравним результаты по среднеквадратичной ошибке:

Метод/Компонента	$X_{\sin}$	$X_{\cos}$
SSA	6.8e-30	1.5e-29
SSA EOSSA	1.5e-29	7.5e-30
Fourier	1.7e-28	3.5e-28
CiSSA	1.9e-29	5.3e-30
CiSSA extended	2.0e-04	8.6e-04

Таблица 2: MSE разложений ряда $X = X_{\sin} + X_{\cos}$ пяти методов

Сравнение алгоритмов. Пример 1

$X = X_{\sin} + X_{\cos} = \sin\frac{2\pi}{12}x + \frac{1}{2}\cos\frac{2\pi}{3}x$, $L = 96$, $N = 96 \cdot 2$ для разложения Фурье и $N = 96 \cdot 2 - 1$ для остальных, чтобы выполнялись условия выполнения разделимости частот. Сравним результаты по среднеквадратичной ошибке:

Метод/Компонента	$X_{\sin}$	$X_{\cos}$
SSA	6.8e-30	1.5e-29
SSA EOSSA	1.5e-29	7.5e-30
Fourier	1.7e-28	3.5e-28
CiSSA	1.9e-29	5.3e-30
CiSSA extended	2.0e-04	8.6e-04

Таблица 2: MSE разложений ряда $X = X_{\sin} + X_{\cos}$ пяти методов

Таблица 2 показывает, что первые четыре разложения сделали правильное (с точностью до вычислений с помощью компьютера) разделение компонент ряда. Однако расширение в методе **CiSSA** ухудшило разделимость периодических частей.

Сравнение алгоритмов. Пример 2

$\mathrm{X} = \mathrm{X}_{\sin} + \mathrm{X}_{\cos} + \mathrm{X}_{\mathrm{noise}} = \sin\frac{2\pi}{12}x + \frac{1}{2}\cos\frac{2\pi}{3}x + \varepsilon_n$, где $\varepsilon_n \sim \mathrm{N}(0, 0.1)$, $L = 96$, $N = 96 \cdot 2$ для разложения Фурье и $N = 96 \cdot 2 - 1$ для остальных.

Метод/Компонента	$\mathrm{X}_{\sin}$	$\mathrm{X}_{\cos}$
SSA	2.9e-04	3.1e-04
SSA EOSSA	2.9e-04	3.1e-04
Fourier	1.0e-04	1.1e-04
CiSSA	1.6e-04	1.8e-04
CiSSA extended	6.6e-04	1.9e-03

Таблица 3: MSE разложений ряда $\mathrm{X} = \mathrm{X}_{\sin} + \mathrm{X}_{\cos} + \mathrm{X}_{\mathrm{noise}}$ пяти методов

Сравнение алгоритмов. Пример 2

$\mathrm{X} = \mathrm{X}_{\sin} + \mathrm{X}_{\cos} + \mathrm{X}_{\mathrm{noise}} = \sin\frac{2\pi}{12}x + \frac{1}{2}\cos\frac{2\pi}{3}x + \varepsilon_n$, где $\varepsilon_n \sim \mathrm{N}(0, 0.1)$, $L = 96$, $N = 96 \cdot 2$ для разложения Фурье и $N = 96 \cdot 2 - 1$ для остальных.

Метод/Компонента	$\mathrm{X}_{\sin}$	$\mathrm{X}_{\cos}$
SSA	2.9e-04	3.1e-04
SSA EOSSA	2.9e-04	3.1e-04
Fourier	1.0e-04	1.1e-04
CiSSA	1.6e-04	1.8e-04
CiSSA extended	6.6e-04	1.9e-03

Таблица 3: MSE разложений ряда $\mathrm{X} = \mathrm{X}_{\sin} + \mathrm{X}_{\cos} + \mathrm{X}_{\mathrm{noise}}$ пяти методов

Проводилось 100 тестов, в таблице 3 указаны средние значения ошибки для одних и тех же реализаций шума.

Был проведен парный t-критерий для зависимых выборок с целью проверки гипотезы о равенстве средних значений ошибки для каждой компоненты, попарно для всех методов. В качестве нулевой гипотезы (H_0) предполагалось, что средние значения двух сравниваемых выборок равны. Критический уровень значимости был установлен на уровне $\alpha = 0.05$. Результаты анализа показали, что во всех случаях p-значение оказались меньше 0.05, что позволяет отвергнуть нулевую гипотезу.

Сравнение алгоритмов. Пример 3

$\mathsf{X} = \mathsf{X}_{\sin} + \mathsf{X}_{\cos} + \mathsf{X}_c + \mathsf{X}_e = \sin\frac{2\pi}{12}x + \frac{1}{2}\cos\frac{2\pi}{3}x + 1 + e^{\frac{x}{100}}$,
$L = 96$, $N = 96 \cdot 2$ для разложения Фурье и $N = 96 \cdot 2 - 1$

Метод/Компонента	$\mathsf{X}_c + \mathsf{X}_e$	$\mathsf{X}_{\sin}$	$\mathsf{X}_{\cos}$
SSA	5.0e-03	8.9e-07	5.2e-05
SSA EOSSA	1.7e-28	1.6e-29	8.7e-30
Fourier	1.1e-01	6.1e-04	6.8e-03
CiSSA	5.3e-02	1.6e-05	4.9e-04
CiSSA extended	5.0e-04	2.1e-04	1.1e-03

Таблица 4: MSE разложений ряда $\mathsf{X} = \mathsf{X}_{\sin} + \mathsf{X}_{\cos} + \mathsf{X}_c + \mathsf{X}_e$ четырех методов

Сравнение алгоритмов. Пример 3

$\mathsf{X} = \mathsf{X}_{\sin} + \mathsf{X}_{\cos} + \mathsf{X}_c + \mathsf{X}_e = \sin\frac{2\pi}{12}x + \frac{1}{2}\cos\frac{2\pi}{3}x + 1 + e^{\frac{x}{100}}$,
$L = 96$, $N = 96 \cdot 2$ для разложения Фурье и $N = 96 \cdot 2 - 1$

Метод/Компонента | $\mathsf{X}_c + \mathsf{X}_e$ | $\mathsf{X}_{\sin}$ | $\mathsf{X}_{\cos}$
SSA 5.0e-03 8.9e-07 5.2e-05
SSA EOSSA 1.7e-28 1.6e-29 8.7e-30
Fourier 1.1e-01 6.1e-04 6.8e-03
CiSSA 5.3e-02 1.6e-05 4.9e-04
CiSSA extended 5.0e-04 2.1e-04 1.1e-03

Таблица 4: MSE разложений ряда $\mathsf{X} = \mathsf{X}_{\sin} + \mathsf{X}_{\cos} + \mathsf{X}_c + \mathsf{X}_e$ четырех методов

Непериодические компоненты будут отвечать низким частотам. Проблема лишь в том, что с помощью методов разложения Фурье **CiSSA** невозможно различить между собой две непериодические компоненты, поскольку группировка работает по частотам, элементы разложения неизбежно смешаются между собой. Будем искать экспоненту и константу по низким частотам, назовем это трендовой составляющей ряда. По таблице 1 лучше всех должен справиться **SSA** с улучшением разделимости EOSSA. Хуже всех — разложение Фурье, поскольку он никаким образом не сможет вычленить из ряда экспоненту.

Результаты таблицы 4 повторяют вышеизложенные рассуждения. Также заметно, что периодические компоненты лучше выделились с помощью **CiSSA** без процедуры расширения ряда в сравнении с **CiSSA** с расширением.

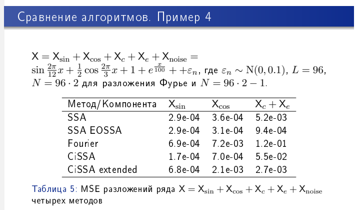
$\mathsf{X} = \mathsf{X}_{\sin} + \mathsf{X}_{\cos} + \mathsf{X}_c + \mathsf{X}_e + \mathsf{X}_{\text{noise}} =$
$\sin\frac{2\pi}{12}x + \frac{1}{2}\cos\frac{2\pi}{3}x + 1 + e^{\frac{x}{100}} + +\varepsilon_n$, где $\varepsilon_n \sim \mathrm{N}(0, 0.1)$, $L = 96$, $N = 96 \cdot 2$ для разложения Фурье и $N = 96 \cdot 2 - 1$.

Метод/Компонента	$\mathsf{X}_{\sin}$	$\mathsf{X}_{\cos}$	$\mathsf{X}_c + \mathsf{X}_e$
SSA	2.9e-04	3.6e-04	5.2e-03
SSA EOSSA	2.9e-04	3.1e-04	9.4e-04
Fourier	6.9e-04	7.2e-03	1.2e-01
CiSSA	1.7e-04	7.0e-04	5.5e-02
CiSSA extended	6.8e-04	2.1e-03	2.7e-03

Таблица 5: MSE разложений ряда $\mathsf{X} = \mathsf{X}_{\sin} + \mathsf{X}_{\cos} + \mathsf{X}_c + \mathsf{X}_e + \mathsf{X}_{\text{noise}}$ четырех методов

Модификации метода **SSA**

└Сравнение алгоритмов. Пример 4



Как видно из таблицы 5, разделения ухудшились, однако **SSA** с улучшением разделимости EOSSA отработал лучше всех. Также был проведен был проведён двухвыборочный t-критерий для зависимых выборок с целью проверки гипотезы о равенстве средних значений ошибки для каждой компоненты, попарно для всех методов. В качестве нулевой гипотезы (H_0) предполагалось, что средние значения двух сравниваемых выборок равны. Критический уровень значимости был установлен на уровне $\alpha = 0.05$. Результаты анализа показали, что во всех случаях p-значение оказалось меньше 0.05, что позволяет отвергнуть нулевую гипотезу.

Сравнение алгоритмов. Собственные пространства

Каждый алгоритм после группировки порождает построенными матрицами собственные подпространства. В случае базового **SSA** алгоритма базис подпространств является адаптивным, то есть зависящим от X, L, N. Таким образом, **SSA** может отличить, например, произведение полиномов, экспонент и косинусов друг от друга.

В случае **CiSSA** базис зависит только от L, N. Если зафиксировать данные параметры, и менять X, базис никак не поменяется.

Сравнение алгоритмов. Собственные пространства

Каждый алгоритм после группировки порождает построенными матрицами собственные подпространства. В случае базового **SSA** алгоритма базис подпространств является адаптивным, то есть зависящим от X, L, N. Таким образом, **SSA** может отличить, например, произведение полиномов, экспонент и косинусов друг от друга.
В случае **CiSSA** базис зависит только от L, N. Если зафиксировать данные параметры, и менять X, базис никак не поменяется.

От собственных подпространств зависит то, какие компоненты временного ряда будут разделимы между собой. Это особенно важно, так как правильный выбор и адаптивность базиса определяют точность разделения сигналов и шумов в ряде. В **SSA** адаптивный базис позволяет эффективно выделять разнородные компоненты, такие как тренды, колебательные и стохастические элементы, даже если они сложно различимы. В то же время в **CiSSA** базис остаётся фиксированным, что может упрощать анализ при постоянных параметрах.

Сравнение алгоритмов. Реальные данные

Теперь рассмотрим реальные данные — месячные ряды промышленного производства (Industrial Production, IP), index $2010 = 100$, в США. Размер выборки составляет $N = 537$. Применим как **CiSSA**, так и **SSA** с автоматическим определением частот и улучшением разделимости по следующим группам:

1. Трендовой составляющей должны отвечать низкие частоты, поэтому диапазон: $[0, \frac{1}{192}]$;
2. Циклы бизнеса по диапазонам: $[\frac{2}{192}, \frac{10}{192}]$;
3. Сезонность по частотам $\omega_k = 1/12, 1/6, 1/4, 1/3, 5/12, 1/2$;

На основе предыдущих требований взято $L = 192$.

Модификации метода **SSA**

└─Сравнение алгоритмов. Реальные данные

Сравнение алгоритмов. Реальные данные

Теперь рассмотрим реальные данные — месячные ряды промышленного производства (Industrial Production, IP), index $2010 = 100$, в США. Размер выборки составляет $N = 537$. Применим как **CiSSA**, так и **SSA** с автоматическим определением частот и улучшением разделимости по следующим группам:

1. Трендовой составляющей должны отвечать низкие частоты, поэтому диапазон: $[0, \frac{1}{192}]$;
2. Циклы бизнеса по диапазонам: $[\frac{2}{192}, \frac{10}{192}]$;
3. Сезонность по частотам $\omega_k = 1/12, 1/6, 1/4, 1/3, 5/12, 1/2$;

На основе предыдущих требований взято $L = 192$.

Данные промышленного производства полезны, поскольку оно указывается в определении рецессии Национальным бюро экономических исследований (NBER), как один из четырех ежемесячных рядов индикаторов, которые необходимо проверять при анализе делового цикла. Эти показатели демонстрируют различные тенденции, сезонность и цикличность (периодические компоненты, которые соответствуют циклам бизнеса). Эти диапазононы частот возникли не случайно. Тренд ассоциируется с частотами, близкими к нулю, что позволяет отразить постоянные изменения с низкой частотой. Циклические компоненты (цикл бизнеса) — это частоты, связанные с деловым циклом, характеризуют циклические колебания, которые, как правило, находятся в диапазоне от полутора до восьми лет. Сезонные компоненты связаны с регулярными колебаниями, такими как месячная или квартальная сезонность.

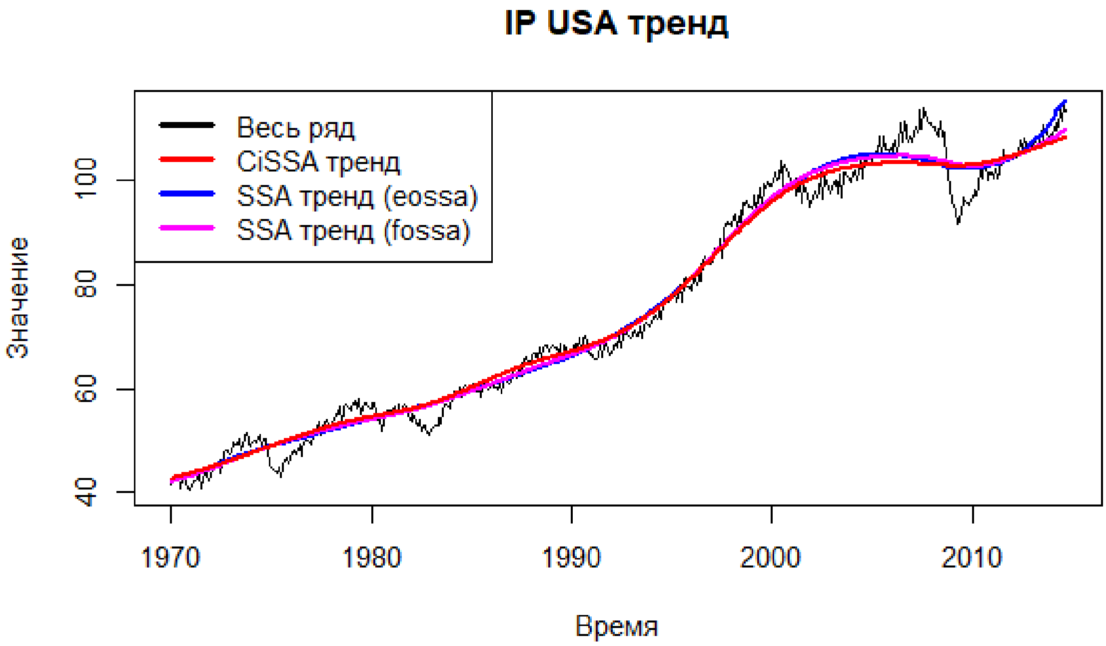
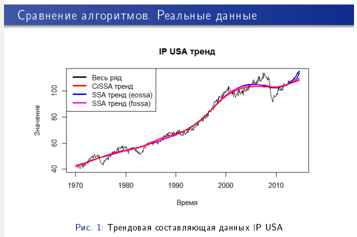


Рис. 1: Трендовая составляющая данных IP USA

Модификации метода **SSA**

└─Сравнение алгоритмов. Реальные данные



При применении FOSSA улучшения разделимости алгоритм **SSA** выделяет тренд довольно похоже с **CiSSA**. Весь график **SSA** тренд EOSSA выглядит более изогнутым при визуальном сравнении с остальными.

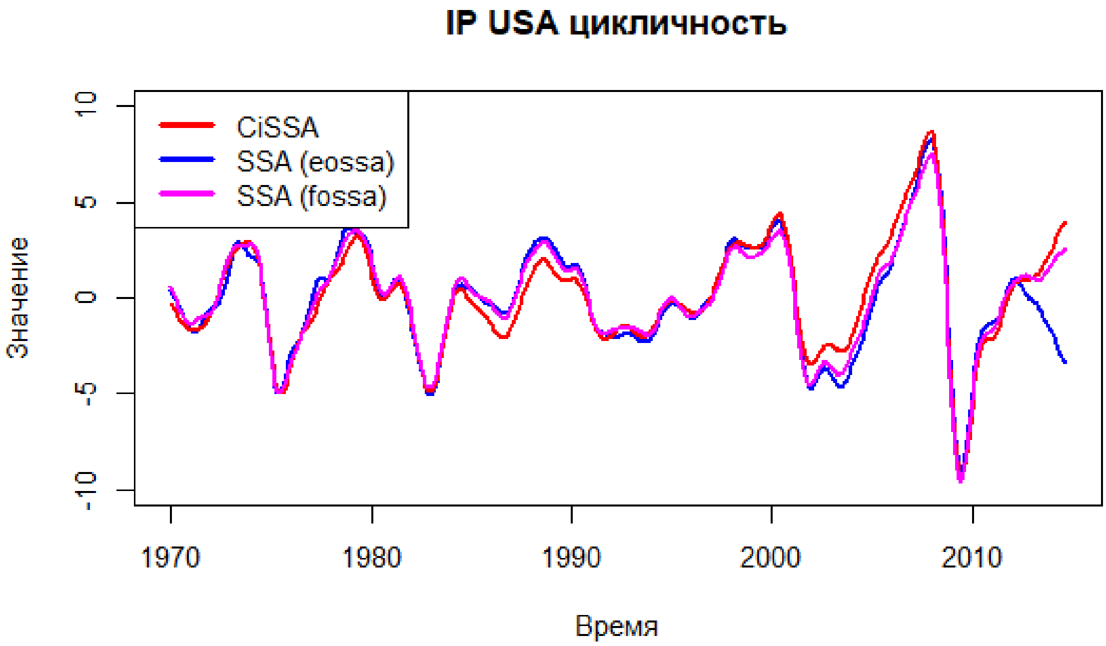
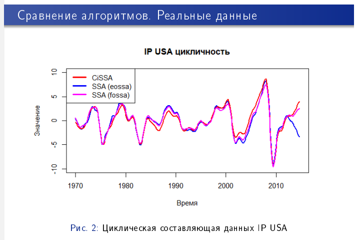


Рис. 2: Циклическая составляющая данных IP USA



Аналогичная тренду ситуация происходит с цикличностью. В случае EOSSA правый хвост (значения ряда после 2010-ого года) смешался между цикличностью и трендом.

Сравнение алгоритмов. Реальные данные

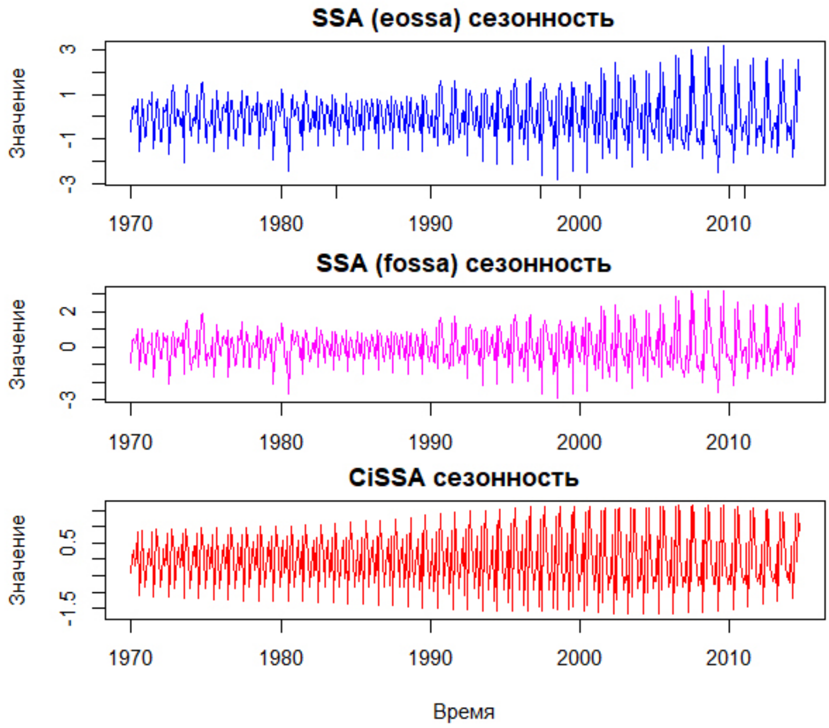
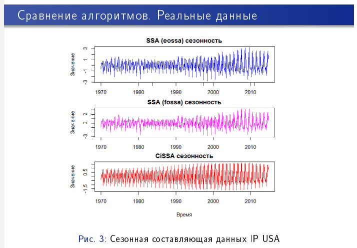


Рис. 3: Сезонная составляющая данных IP USA



Поскольку в базовом **SSA** адаптивный базис, сезонность является менее систематичной, разброс значений выше по сравнению с **CiSSA**. Таким образом, получились довольно похожие результаты в выделении тренда и цикличности при использовании **SSA** с FOSSA и **CiSSA**. Несколько иные результаты при **SSA** с EOSSA. Сезонная составляющая в силу неадаптивного базиса более строго выглядит для метода **CiSSA**.

Заключение

По полученным результатам, можно следующие выводы:

1. Алгоритм **CiSSA** работает лучше разложения Фурье;
2. 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;
3. Если данные зашумлены или имеется непериодичность, алгоритм **SSA** с улучшением разделимости справляется в среднеквадратичном лучше **CiSSA** с расширением ряда или без.

Заключение

По полученным результатам, можно следующие выводы:

1. Алгоритм **CiSSA** работает лучше разложения Фурье;
2. 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;
3. Если данные зашумлены или имеется непериодичность, алгоритм **SSA** с улучшением разделимости справляется в среднеквадратичном лучше **CiSSA** с расширением ряда или без.

В данной работе исследован алгоритм **CiSSA**, сравнены методы **CiSSA** и **SSA**, и полученные знания были проверены на реальных и смоделированных примерах с помощью языка R. Оба алгоритма справляются с поставленными задачами, существенным различием является то, что алгоритм **SSA** является более гибким: в нем адаптивный базис, есть дополнительные алгоритмы, которые довольно похоже приближают этот алгоритм к **CiSSA**, а также методы для автоматического выбора компонентов по частотам. Метод **CiSSA** является простым в использовании.

Дальнейшими действиями является рассмотрение других модификаций метода **SSA**.

Все вычисления, а также код **CiSSA** можно найти в github репозитории [Pogrebnikov, 2024].

Список литературы

Juan Bogalo, Pilar Poncela, and Eva Senra. Circulant singular spectrum analysis: A new automated procedure for signal extraction. *Signal Processing*, 177, 2020. ISSN 0165-1684. doi: 10.1016/j.sigpro.2020.107750.

Nina Golyandina, Vladimir Nekrutkin, and Anatoly Zhigljavsky. *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman and Hall/CRC, 2001.

Nina Golyandina, Pavel Dudnik, and Alex Shlemov. Intelligent identification of trend components in singular spectrum analysis. *Algorithms*, 16(7):353, 2023. doi: 10.3390/a16070353.

Nikolay Pogrebnikov. SPbSU CISSA coursework: Time series analysis. https://github.com/xSICHx/spbu_cissa_coursework, 2024.

Список литературы

Juan Bogalo, Pilar Poncela, and Eva Senra. Circulant singular spectrum analysis: A new automated procedure for signal extraction. *Signal Processing*, 177, 2020. ISSN 0165-1684. doi: 10.1016/j.sigpro.2020.107750.

Nina Golyandina, Vladimir Nekrutkin, and Anatoly Zhigljavsky. *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman and Hall/CRC, 2001.

Nina Golyandina, Pavel Dudnik, and Alex Shlemov. Intelligent identification of trend components in singular spectrum analysis. *Algorithms*, 16(7):353, 2023. doi: 10.3390/a16070353.

Nikolay Pogrebnikov. SPbSU CISSA coursework: Time series analysis. https://github.com/xSICHx/spbu_cissa_coursework, 2024.

На данном слайде представлен список основных источников, используемых в моей работе. Спасибо за внимание.