Санкт-Петербургский государственный университет

Прикладная математика и информатика

Отчет по учебной практике 3 (научно-исследовательской работе)

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

Выполнил:

Погребников Николай Вадимович

группа 21.Б04-мм

Научный руководитель:

д. ф.-м. н., доц.

Голяндина Нина Эдуардовна

Кафедра Статистического Моделирования

Санкт-Петербург

2024

Содержание

1 Введение

Временные ряды представляют собой последовательность данных, собранных или измеренных в хронологическом порядке. Они играют ключевую роль в анализе и прогнозировании в различных областях, таких как экономика, финансы, климатология, медицина. Понимание эволюции явлений во времени является критическим для выявления тенденций, циклов и аномалий.

Сингулярный спектральный анализ (**SSA** [3]) — метод, целью которого является разложение оригинального ряда на сумму небольшого числа интерпретируемых компонентов, таких как медленно изменяющаяся тенденция (тренд), колебательные компоненты (сезонность) и "структурный" шум. Основной концепцией при изучении свойств методов **SSA** является "разделимость", которая характеризует, насколько хорошо разные компоненты могут быть отделены друг от друга. В данном исследовании рассматривается математическая составляющая вариации алгоритма **SSA** — circulant singular spectrum analysis (**CiSSA**), предложенная в статье [1], а также сравнение базового метода и циркулярного, применение их на языке R.

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**.

Далее кратко опишем структуру работы. В разделе 2 рассматривается базовый метод **SSA** и его ключевые свойства. В следующем разделе 3 представлен метод **CiSSA**, также с описанием его основных характеристик. Раздел 4 посвящён сравнению методов SSA, разложения Фурье и **CiSSA** на модельных и реальных примерах. В заключительной секции 5 подведены основные итоги исследования.

2 Базовый метод SSA

Рассмотрим базовый метод сингулярного спектрального анализа [3].

2.1 Алгоритм метода SSA

Пусть $N > 2$, вещественнозначный временной ряд $\mathsf{X} = (x_1, \ldots, x_N)$ длины N. Базовый алгоритм состоит **SSA** из четырех шагов.

2.1.1 Вложение

L — некоторое целое число (длина окна), $1 < L < N$. Строится L-траекторная матрица $\mathbf{X}$, состоящая из $K = N - L + 1$ векторов вложения:

$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & x_3 & \ldots & x_K \\ x_2 & x_3 & x_4 & \ldots & x_{K+1} \\ x_3 & x_4 & x_5 & \ldots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \ldots & x_N \end{pmatrix}. \tag{1}$$

Полезным свойством является то, что матрица $\mathbf{X}$ имеет одинаковые элементы на антидиагоналях. Таким образом, L-траекторная матрица является ганкелевой.

2.1.2 Сингулярное разложение (SVD)

Результатом этого шага является сингулярное разложение (Singular Value Decomposition, **SVD**) траекторной матрицы ряда.

Пусть $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$, $\lambda_1, \ldots, \lambda_L$ — собственные числа матрицы $\mathbf{S}$, взятые в неубывающем порядке и $U_1, \ldots, U_L$ — ортонормированная система собственных векторов соответствующих собственным числам матрицы $\mathbf{S}$.

Определим $d = \max\{i : \lambda_i > 0\}$ и $V_i = \mathbf{X}^{\mathrm{T}} U_i / \sqrt{\lambda_i}$. Тогда сингулярным разложением называется представление матрицы в виде:

$$\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_d = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}. \tag{2}$$

Набор $(\sqrt{\lambda_i}, U_i, V_i^{\mathrm{T}})$ называется i-й собственной тройкой разложения (2).

2.1.3 Группировка

На основе разложения (2) производится процедура группировки, которая делит все множество индексов $\{1, \ldots, d\}$ на m непересекающихся подмножеств $I_1, \ldots, I_d$.

Пусть $I = \{i_1, \ldots, i_p\}$, тогда $\mathbf{X}_I = \mathbf{X}_{\mathbf{i_1}} + \cdots + \mathbf{X}_{\mathbf{i_p}}$. Такие матрицы вычисляются для каждого $I = I_1, \ldots, I_m$. В результате получаются матрицы $\mathbf{X}_{\mathbf{I_1}}, \ldots, \mathbf{X}_{\mathbf{I_m}}$. Тем самым разложение (2) может быть записано в сгруппированном виде:

$$\mathbf{X} = \mathbf{X}_{I_1} + \cdots + \mathbf{X}_{I_m}.$$

2.1.4 Диагональное усреднение

Пусть $\mathbf{Y}$ — матрица размерности $L \times K$. $L^* = \min(L, K)$, $K^* = \max(L, K)$ Диагональное усреднение переводит матрицу $\mathbf{Y}$ в временной ряд $g_0, \ldots, g_{N-1}$:

$$g_k = \begin{cases} \frac{1}{k+1} \sum\limits_{m=1}^{k+1} y^*_{m,k-m+2} & \text{для } 0 \le k < L^* - 1, \\ \frac{1}{L^*} \sum\limits_{m=1}^{L^*} y^*_{m,k-m+2} & \text{для } L^* - 1 \le k < K^*, \\ \frac{1}{N-k} \sum\limits_{m=k-K^*+2}^{N-K^*+1} y^*_{m,k-m+2} & \text{для } K^* \le k < N. \end{cases}$$

Применяя данную операцию к матрицам $\mathbf{X_{I_1}}, \ldots, \mathbf{X_{I_m}}$, получаются m новых рядов: $\mathsf{X}_1, \ldots, \mathsf{X}_m$. При этом, $\mathsf{X}_1 + \cdots + \mathsf{X}_m = \mathsf{X}$.

2.2 Свойства SSA

2.2.1 Точная разделимость

Пусть временной ряд $\mathsf{X} = \mathsf{X}^{(1)} + \mathsf{X}^{(2)}$ и задачей является нахождение этих слагаемых. В результате базового алгоритма **SSA** также получаем 2 ряда. Возникает вопрос: в каких случаях мы можем так выбрать параметр алгоритма L при $m = 2$ и так сгруппировать собственные тройки, чтобы получить исходные 2 ряда без смешиваний? При выборе длины окна L, каждый из рядов $\mathsf{X}^{(1)}$, $\mathsf{X}^{(2)}$, X порождает траекторную матрицу $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}$.

Определение 1. *Будем говорить, что ряды $\mathsf{X}^{(1)}$ и $\mathsf{X}^{(2)}$ слабо L-разделимы, если пространства, порождаемые строками $\mathbf{X}^{(1)}$ и $\mathbf{X}^{(2)}$ соответственно, ортогональны. То же самое должно выполняться для столбцов [3].*

Если выполняется условие слабой L-разделимости, тогда существует такое сингулярное разложение траекторной матрицы $\mathbf{X}$ ряда X, что его можно разбить на две части, являющиеся сингулярными разложениями траекторных матриц рядов $\mathsf{X}^{(1)}, \mathsf{X}^{(2)}$ [3].

Определение 2. *Будем говорить, что ряды $\mathsf{X}^{(1)}, \mathsf{X}^{(2)}$ сильно L-разделимы, если они слабо L-разделимы и после процедуры* **SVD** *собственные числа рядов различны [3].*

Если выполняется условие сильной L-разделимости, тогда любое сингулярное разложение траекторной матрицы $\mathbf{X}$ ряда X можно разбить на две части, являющиеся сингулярными разложениями траекторных матриц рядов $\mathsf{X}^{(1)}, \mathsf{X}^{(2)}$ [3].

Рассмотрим таблицу, в которой знаком + отмечены пары рядов, для которых существуют параметры функций и параметры метода L и $K = N - L + 1$, при которых они разделимы (точно разделимы). Данная таблица 1 и условия разделимости с доказательствами взяты из книги [3].

Таблица 1: Точная разделимость

	const	cos	exp	exp cos	ak+b
const	-	+	-	-	-
cos	+	+	-	-	-
exp	-	-	-	+	-
exp cos	-	-	+	+	-
ak+b	-	-	-	-	-

Стоит отметить, что точная разделимость для cos достигается, если $Lw \in \mathbb{N}$, $Kw \in \mathbb{N}$, где w — частота [3].

Однако, по таблице 1 видно, что условия точной разделимости достаточно жесткие и вряд ли выполнимы в реальных задачах. Тогда появляется такое понятие, как асимптотическая разделимость.

2.2.2 Асимптотическая разделимость

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 следующим образом:

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

Определение 3. *Ряды $F^{(1)}, F^{(2)}$ называются ε-разделимыми при длине окна L, если*

$$\rho^{(L,K)} \stackrel{\text{def}}{=} \max\left(\max_{1\le i,j\le K} |\rho_{i,j}^{(L)}|, \max_{1\le i,j\le L} |\rho_{i,j}^{(K)}|\right) < \varepsilon \text{ [3].}$$

Определение 4. *Если $\rho^{(\mathsf{X}(N),K(N))} \to 0$ при некоторой последовательности $L = L(N)$, $N \to \infty$, то ряды $\mathsf{X}^{(1)}, \mathsf{X}^{(2)}$ называются асимптотически $L(N)$-разделимыми [3].*

Как можно заметить по таблице 2, для гораздо большего класса функций асимптотическая разделимость имеет место [3].

Таблица 2: Асимптотическая разделимость

	const	cos	exp	exp cos	ak+b
const	-	+	+	+	-
cos	+	+	+	+	+
exp	+	+	+	+	+
exp cos	+	+	+	+	+
ak+b	+	+	+	+	-

2.2.3 Алгоритмы улучшения разделимости

Для **SSA** существуют алгоритмы улучшения разделимости компонентов ряда. Они позволяют более точно отделять временные подряды друг от друга. В данной работе будут использоваться методы EOSSA и FOSSA. Подробнее про них можно почитать в [2]. Для нас важно, что благодаря применению улучшения разделимости мы можем делать автоматическую группировку по заданным частотам в базовом алгоритме **SSA**.

3 Метод Circulant singular spectrum analysis (CiSSA)

В этом разделе описана модификация **SSA** на основе циркулярной матрицы [1]. Авторы метода называют её автоматизированной. Причем автоматизированная в том смысле, что компоненты ряда группируются по частотам самим алгоритмом. Сначала будет рассмотрен метод только для стационарного случая, затем показана его применимость при использовании нестационарного ряда.

Стационарность подразумевает неизменность статистических свойств ряда во времени. Однако определим это понятие формально [3].

Определение 5. *Пусть* $\mathsf{X} = (x_1, \dots, x_n, \dots)$ *— временной ряд. Ряд* X *называется стационарным, если существует функция* $R_{\mathsf{X}}(k)$ $(-\infty < k < +\infty)$ *такая, что для любых* $k, l \geq 1$

$$R_{\mathsf{X}}^{(N)}(k,l) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{m=1}^{N} x_{k+m}x_{l+m} \xrightarrow{N\to\infty} R_{\mathsf{X}}(k-l). \tag{3}$$

Если (3) *выполняется, тогда* R_{X} *называется ковариационной функцией стационарного ряда* X.

Теорема 1. *Пусть* R_{X} *— ковариационная функция стационарного ряда* X. *Тогда существует конечная мера* m_{X}, *определенная на борелевских подмножествах* $(-1/2, 1/2]$, *такая, что*

$$R_{\mathsf{X}}(k) = \int_{(-\frac{1}{2},\frac{1}{2}]} e^{i2\pi k\omega} m_{\mathsf{X}}(d\omega).$$

Мера m_{X} *называется спектральной мерой ряда* X.

Доказательство. Доказательство в [3]. □

3.1 Алгоритм метода CiSSA

Данный алгоритм состоит также из четырех основных шагов.

Зафиксируем стационарный временной ряд X состоящий из N элементов и выберем длину окна L.

3.1.1 Вложение

Такой же, как и в **SSA**. Считаем матрицу $\mathbf{X}$, заданную в (1).

3.1.2 Разложение

Будем рассматривать временной ряд как выборку после эксперимента, а не как случайную величину. Соответственно, все формулы будут выборочными.

Определим автоковарицации:

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

На основе $\hat{\gamma}_m$ определим матрицу:

$$\hat{\gamma}_L = \begin{pmatrix} \hat{\gamma}_1 & \hat{\gamma}_2 & \dots & \hat{\gamma}_L \\ \hat{\gamma}_2 & \hat{\gamma}_1 & \dots & \hat{\gamma}_{L-1} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\gamma}_L & \hat{\gamma}_{L-1} & \dots & \hat{\gamma}_1 \end{pmatrix}. \tag{4}$$

Данная матрица $L \times L$ называется Теплицевой и используется в методе Toeplitz SSA (подробнее про данный метод можно прочитать в книге [3]). На ее основе составим циркулярную матрицу для алгоритма Circulant SSA [1]:

$$\hat{\mathrm{C}}_L = \begin{pmatrix} \hat{c}_1 & \hat{c}_2 & \dots & \hat{c}_L \\ \hat{c}_2 & \hat{c}_1 & \dots & \hat{c}_{L-1} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{c}_L & \hat{c}_{L-1} & \dots & \hat{c}_1 \end{pmatrix}, \tag{5}$$

где $\hat{c}_m = \frac{L-m}{L}\hat{\gamma}_m + \frac{m}{L}\hat{\gamma}_{L-m}$, $m = 0 : L-1$. Собственные числа матрицы $\hat{\mathrm{C}}_L$, определенной в (5) задаются по формуле:

$$\lambda_{L,k} = \sum_{m=0}^{L-1} \hat{c}_m \exp\left(i2\pi m \frac{k-1}{L}\right),\ k = 1 : L, \text{ причем } \lambda_{L,k} = \lambda_{L,L+2-k},$$

а собственные вектора, связанные с $\lambda_{L,k}$ вычисляются следующим образом:

$$U_k = L^{-1/2}(u_{k,1}. \cdot \cdot, u_{k,L}), \text{ где } u_{k,j} = \exp\left(-\mathrm{i}2\pi(j-1)\frac{k-1}{L}\right), \text{ причем } U_k = U^*_{L+2-k},$$

где U^* — комплексное сопряжение вектора U.

Элементарное разложение

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 же L — четное, то частоте $w_{\frac{L}{2}+1}$ будет соответствовать один вектор $U_{\frac{L}{2}+1}$.

Следовательно, индексы группируются следующим образом:

$$B_1 = \{1\};\ B_k = \{k, L+2-k\}, \text{ для } k = 2 : \lfloor \frac{L+1}{2} \rfloor;\ B_{\frac{L}{2}+1} = \left\{\frac{L}{2}+1\right\}, \text{ если } L \mod 2 = 0.$$

Разложение $\mathbf{X}_{B_k} = \mathbf{X}_k + \mathbf{X}_{L+2-k} = U_k U_k^H \mathbf{X} + U_{L+2-k} U_{L+2-k}^H \mathbf{X}$, где U^H — это комплексное сопряжение и транспонирование вектора U.

3.1.3 Группировка

Такой же шаг, как и в базовом **SSA**. Однако 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 $\mathbf{X}_{B_k} = \mathbf{X}_k + \mathbf{X}_{L+2-k}$.

3.1.4 Диагональное усреднение

Такой же шаг, как и в базовом **SSA**.

Замечание 1. *$U_k U_k^H + U_{L+2-k} U_{L+2-k}^H$ 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.*

Замечание 2. *В 3.2.2 рассмотрена связь между матрицей $\mathbf{X}_{B_k}$ и разложениями Фурье для векторов вложения.*

Нестационарный случай

Для применения данного алгоритма на нестационарных временных рядах, нужно применить процедуру расширения ряда. Как утверждается в статье [1], после расширения, **CiSSA** можно применить к нестационарному ряду. Сама процедура расширения ряда X производится с использованием авторегрессионной (AR) модели. 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. В ходе работы алгоритм выполняет следующие шаги:

1. **Определение порядка AR-модели**: Метод определяет порядок p AR-модели как целую часть от деления длины ряда N на 3. Это значение порядка модели p будет использовано для построения авторегрессионной модели на дифференцированном временном ряде;

2. **Построение дифференцированного ряда**: Временной ряд X сначала преобразуется в дифференцированный ряд $d\mathsf{X}$, чтобы удалить трендовые компоненты;

3. **Построение AR-модели**: После этого для дифференцированного ряда вычисляются коэффициенты авторегрессионной модели A с использованием метода Юла-Уокера, основываясь на определенном ранее порядке p;

4. **Правое расширение ряда**: С помощью AR-модели ряд $d\mathsf{X}$ прогнозируется на H шагов вправо. Затем возвращается к своему изначальному состоянию путем интегрирования $d\mathsf{X}$. Получается расширение исходного ряда X на H шагов вправо;

5. **Левое расширение ряда**: Аналогично предыдущему пункту, ряд прогнозируется на H шагов влево;

6. **Возвращение расширенного ряда**: В конце метод возвращает расширенный временной ряд $\mathsf{X}_{\text{extended}}$, который содержит как левое, так и правое расширение на H шагов от исходного ряда X.

Таким образом, алгоритм расширения ряда позволяет выполнять предсказания временного ряда по обе стороны от его границ, основываясь на авторегрессионной модели, построенной на дифференцированном ряде, что полезно для выделения тренда. Однако поскольку мы рассматриваем расширенный ряд, то и периодические компоненты будут строиться по нему. Поэтому в угоду лучшего выделения трендовой составляющей, будет несколько жертвоваться точность разделения периодических компонентов.

3.2 Свойства

3.2.1 Асимптотическая эквивалентность методов

В статье [1] говорится, что асимптотически методы **SSA** и **CiSSA** эквивалентны и в доказательство приводится теорема.

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 $\lim\limits_{L\to\infty} \frac{\|S_1-S_2\|_F}{\sqrt{L}} = 0$, где $\|\cdot\|_F$ — норма Фробениуса. Тогда $M_1 \sim M_2$, $S_1 \sim S_2$.*

Теорема 2. *Дана $L \times K$ траекторная матрица $\mathbf{X}$, определенная в* (1). *Пусть $S_B = \mathbf{X}\mathbf{X}^T/K$, S_T — матрица, определенная в* (4), *S_C — матрица, определенная в* (5). *Тогда $S_B \sim S_T \sim S_C$.*

Доказательство. Доказательство в источнике [1]. □

Теорема 2 дает понимание похожих практических результатов при применении разных методов.

3.2.2 Связь CiSSA с разложением Фурье

Для описания конечных, но достаточно длинных рядов можно использовать разложение Фурье. Пусть $\mathsf{X} = (x_1, \dots, x_n, \dots)$ — временной ряд

Определение 7. *Разложение*

$$x_n = c_0 + \sum_{k=1}^{\lfloor \frac{N+1}{2} \rfloor} \left(c_k \cos(2\pi nk/N) + s_k \sin(2\pi nk/N) \right), \tag{6}$$

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

Таким образом, можно выделить компоненту ряда, отвечающую за частоту $w_k = \frac{k-1}{L}$, $k = 1 : \lfloor \frac{N+1}{2} \rfloor$;

Алгоритм **CiSSA** тесно связан с разложением Фурье. А именно, **CiSSA** можно представить так:

1. Вычисляем разложение Фурье для каждого вектора вложения L-траекторной матрицы $\mathbf{X}$, состоящей из $K = N - L + 1$ векторов. Получается K разложений Фурье по частотам $w_k = \frac{k-1}{L}$, $1 : \lfloor \frac{L+1}{2} \rfloor$;
2. По получившимся разложениям Фурье усредняем значения для соответствующих x_i и частот w_k.

3.2.3 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. Поэтому до применения алгоритма необходимо выделить интересующие частоты, то есть знать их заранее, и, исходя из них, выбирать значение L.

3.2.4 Асимптотическая разделимость

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.

4 Сравнение алгоритмов разложения Фурье, SSA и CiSSA

Все вычисления, а также код **CiSSA** можно найти в github репозитории [4].

4.1 Преимущества и недостатки методов

В данной секции проводится сравнение четырех различных методов: **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 3, где строки соответствуют методам, а столбцы — условиям (особым видам компонент ряда). 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 $\cos(2\pi\omega x + \varphi)$;
- X_{np1} — одна непериодическая компонента в ряде, остальные имеют период;
- X_{np} — несколько непериодических компонент в ряде, остальные имеют период, интересует разделение между непериодическими компонентами;

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, $Lw = k \notin \mathbb{N}$, $Kw = k \notin \mathbb{N}$	X_{np1}	X_{np}
SSA EOSSA	+	$\to$	$\to$	$\to$	$\to$
Fourier	+	+	$\to$	–	–
CiSSA	+	+	$\to$	–	–
CiSSA extended	+	+	$\to$	$\to$	–

Таблица 3: Преимущества и недостатки четырех методов

На основе таблицы 3 были выбраны примеры, следующие ниже.

Данные методы разложения временного ряда должны совпадать, если ряд состоит только из периодических компонент. Например, пусть $\mathsf{X} = \mathsf{X}_{\sin} + \mathsf{X}_{\cos} = \sin\frac{2\pi}{12}x + \cos\frac{2\pi}{8}x$, $L = 96$, $N = 96 \cdot 2$. Сравним результаты по среднеквадратичной ошибке:

Метод/Компонента	$\mathsf{X}_{\sin}$	$\mathsf{X}_{\cos}$
SSA EOSSA	1.4e-29	4.0e-30
Fourier	2.1e-28	3.5e-28
CiSSA	1.0e-29	4.7e-30
CiSSA extended	8.6e-04	1.0e-03

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

Таблица 4 показывает, что первые три разложения сделали правильное (с точностью до вычислений с помощью компьютера) разделение компонент ряда. Однако расширение в методе **CiSSA** ухудшило разделимость периодических частей.

Теперь добавим к этому ряду шум: $\mathsf{X} = \mathsf{X}_{\sin} + \mathsf{X}_{\cos} + \mathsf{X}_{\text{noise}} = \sin\frac{2\pi}{12}x + \cos\frac{2\pi}{8}x + \varepsilon_n$, где $\varepsilon_n \sim \mathrm{N}(0, 0.1)$, $L = 96$, $N = 96 \cdot 2$. Результаты должны ухудшиться. Проводилось 100 тестов, в таблице 5 указаны средние значения ошибки для одних и тех же реализаций шума.

Метод/Компонента	$\mathsf{X}_{\sin}$	$\mathsf{X}_{\cos}$
SSA EOSSA	3.1e-04	3.0e-04
Fourier	9.7e-05	1.1e-04
CiSSA	1.8e-04	1.8e-04
CiSSA extended	1.6e-03	1.7e-03

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

По таблице 5 видно, что зашумление ряда дало негативный эффект на ошибку. Также 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 оказалось меньше 0.05, что позволяет отвергнуть нулевую гипотезу.

Попробуем добавить к ряду непериодическую компоненту. $\mathsf{X} = \mathsf{X}_{\sin} + \mathsf{X}_{\cos} + \mathsf{X}_c + \mathsf{X}_e = \sin\frac{2\pi}{12}x + \cos\frac{2\pi}{8}x + 1 + e^{\frac{x}{100}}$, $L = 96$, $N = 96 \cdot 2$. 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 3 лучше всех должен справиться **SSA** с улучшением разделимости EOSSA. Хуже всех — разложение Фурье, поскольку он никаким образом не сможет вычленить из ряда экспоненту.

Метод/Компонента	$\mathsf{X}_c + \mathsf{X}_e$	$\mathsf{X}_{\sin}$	$\mathsf{X}_{\cos}$
SSA EOSSA	3.1e-28	2.3e-27	2.7e-27
Fourier	1.1e-01	3.1e-03	6.8e-03
CiSSA	5.3e-02	1.6e-04	4.9e-04
CiSSA extended	5.7e-04	7.5e-04	1.3e-03

Таблица 6: MSE разложений ряда $\mathsf{X} = \mathsf{X}_{\sin} + \mathsf{X}_{\cos} + \mathsf{X}_c + \mathsf{X}_e$ четырех методов

Результаты таблицы 6 повторяют вышеизложенные рассуждения. Также заметно, что периодические компоненты лучше выделились с помощью **CiSSA** без процедуры расширения ряда в сравнении с **CiSSA** с расширением.

Теперь добавим шум в предыдущий пример. Результаты всех разложений должны ухудшиться. $\mathsf{X} = \mathsf{X}_{\sin} + \mathsf{X}_{\cos} + \mathsf{X}_c + \mathsf{X}_e + \mathsf{X}_{\text{noise}} = \sin\frac{2\pi}{12}x + \cos\frac{2\pi}{8}x + 1 + e^{\frac{x}{100}} + \mathrm{N}(0, 0.1)$, $L = 96$, $N = 96 \cdot 2$. Было проведено 100 тестов, в таблице 7 указаны средние значения ошибки.

Метод/Компонента	X_{sin}	X_{cos}	$\mathsf{X}_c + \mathsf{X}_e$
SSA EOSSA	3.1e-04	3.0e-04	9.4e-04
Fourier	3.3e-03	7.2e-03	1.2e-01
CiSSA	3.4e-04	7.0e-04	5.5e-02
CiSSA extended	1.5e-03	2.0e-03	2.7e-03

Таблица 7: MSE разложений ряда $\mathsf{X} = \mathsf{X}_{\text{sin}} + \mathsf{X}_{\text{cos}} + \mathsf{X}_c + \mathsf{X}_e + \mathsf{X}_{\text{noise}}$ четырех методов

Как видно из таблицы 7, 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.

По результатам данных примеров и таблицы 3, можно понять, что **CiSSA** работает лучше, чем разложение Фурье. Однако это не удивительно, ведь разложение Фурье это частный случай **CiSSA** при $L = N$. Таким образом, далее не будем рассматривать разложение Фурье, остановимся на остальных трех методах. Кроме того, по умолчанию будет использоваться **CiSSA** с расширением, если есть непериодичность, и обычный **CiSSA**, если все компоненты периодичны. Также при написании **SSA** будет подразумеваться использование **SSA** с EOSSA, если нет конкретных указаний

4.2 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.

4.3 Точная разделимость

Как удалось выяснить, классов точной разделимости больше в базовом алгоритме **SSA**, однако в случае разделения cos, условия менее жесткие при использовании **CiSSA**.

Проверим на примерах. Возьмем временной ряд, с разложением которого оба алгоритма должны справиться: $\mathsf{X} = \mathsf{X}_C + \mathsf{X}_{cos} = 1 + \cos(\frac{2\pi}{12}x)$, $L = 96 \mid 12$, $N = 96 \cdot 2 - 1$, $K = 96 \mid 12$. Будем считать MSE между настоящими компонентами ряда и вычисленными. В случае **SSA** получилась ошибка при вычислении $C = 1$: 2.1e−30, а при вычислении $\cos(\frac{2\pi}{12}x)$: 4.9e−30. Если применить алгоритм **CiSSA**, получатся ошибки при $C = 1$: 3.6e−31, при $\cos(\frac{2\pi}{12}x)$: 5.2e−30. Эти ошибки можно посчитать за погрешность вычислений на компьютере.

Теперь возьмем временной ряд, при котором **SSA** должен отработать хуже **CiSSA**: $\mathsf{X} = \mathsf{X}_C + \mathsf{X}_{\text{cos}} = 1 + cos(\frac{2\pi}{12}x)$, $L = 96 \mid 12$, $N = 96 \cdot 2 + 5$, $K = 102 \nmid 12$. Поскольку K не делится на частоту косинуса, условия точной разделимости в **SSA** не выполняются. Будем считать MSE между настоящими компонентами ряда и вычисленными. В случае **SSA** получилась ошибка при вычислении $C = 1$: 9.5e−5, а при вычислении $\cos(\frac{2\pi}{12}x)$: 9.6e−5. Если применить алгоритм **CiSSA**, получатся ошибки при $C = 1$: 3.2e−31, при $\cos(\frac{2\pi}{12}x)$: 5.1e−30.

Таким образом, с разделением косинуса от константы лучше справился алгоритм **CiSSA**, поскольку в нем требуется меньше условий на параметры алгоритма.

4.4 Асимптотическая разделимость

Как было сказано, асимптотически разделимы в методе **SSA** полиномы, гармонические функции (косинус, косинус помноженный на экспоненту, экспонента) [3]. В алгоритме **CiSSA** при увеличении длины окна L меняется сетка разбиения частот. Из-за этого, даже если не удастся выбрать подходящее L, при котором будет точно отделим косинус, но постоянно его увеличивать, в конечном счете получится снизить ошибку выделения нужной компоненты косинуса, если брать соседние частоты с частотой компоненты. Однако в этом случае нужно выбирать диапазон частот, которые стоит объединить.

Непериодические компоненты повлияют на ошибку разложений всего временного ряда, они смешаются и их уже никак не получится отделить методом **CiSSA**. Рассмотрим более детально пример с экспонентой: $\mathsf{X} = \mathsf{X}_c + \mathsf{X}_e + \mathsf{X}_{\cos} + \mathsf{X}_{\sin} = 1 + e^{\frac{x}{100}} + \cos(\frac{2\pi}{12}x) + \sin(\frac{2\pi}{24}x)$, $N = 96 \cdot 2 - 1$, $L = 96$, можно получить следующие результаты:

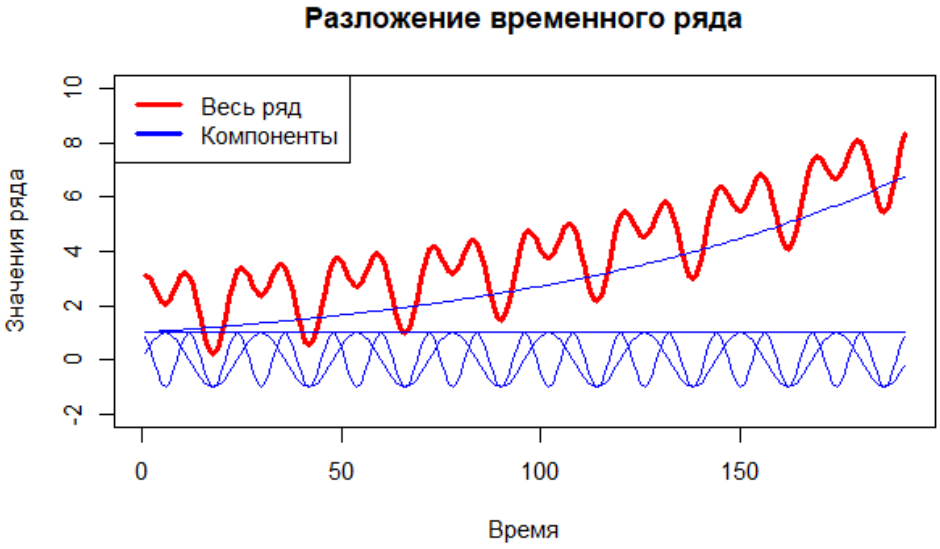

Рис. 1: Правильное разложение ряда $\mathsf{X} = \mathsf{X}_c + \mathsf{X}_e + \mathsf{X}_{\cos} + \mathsf{X}_{\sin}$

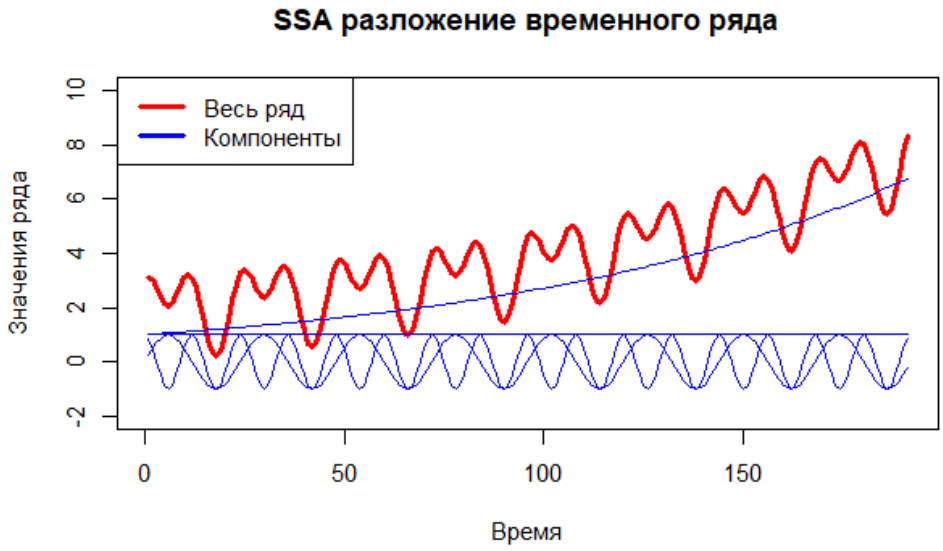

Рис. 2: Разложение ряда $\mathsf{X} = \mathsf{X}_c + \mathsf{X}_e + \mathsf{X}_{\cos} + \mathsf{X}_{\sin}$ методом **SSA**

Метод **SSA** разделил правильно все компоненты друг от друга.

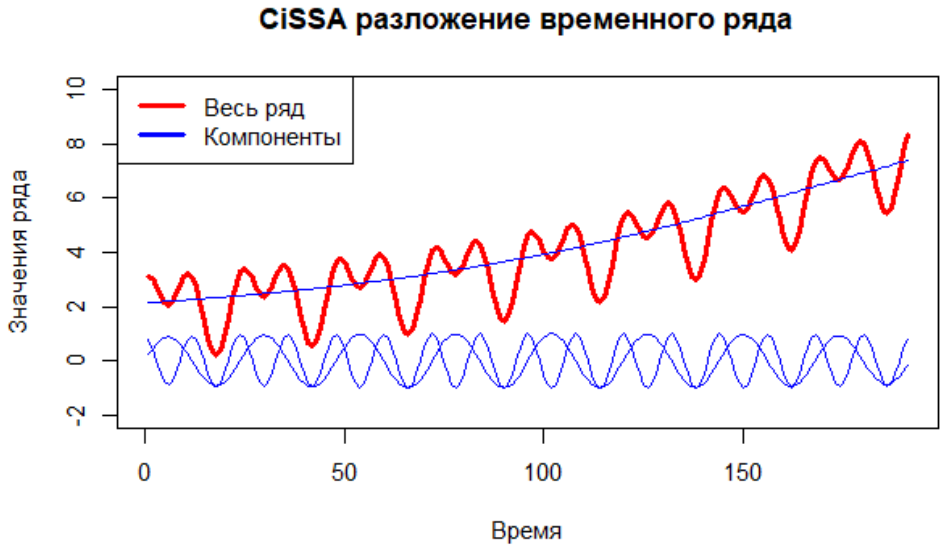

Рис. 3: Разложение ряда $\mathsf{X} = \mathsf{X}_c + \mathsf{X}_e + \mathsf{X}_{\cos} + \mathsf{X}_{\sin}$ методом **CiSSA**

В случае **CiSSA** получилось так, что экспонента и константа смешались в одну компоненту. Как и в примере сравнении разделения ряда Фурье и **CiSSA**, одни и те же частоты отвечают одновременно и за константу, и за экспоненту.

Метод/Компонента	X_e	X_c	$\mathsf{X}_c + \mathsf{X}_e$	$\mathsf{X}_{\sin}$	$\mathsf{X}_{\cos}$
SSA	2.2e-25	2.2e-25	4.2e-28	3.8e-29	1.6e-29
CiSSA	none	none	3.5e-02	1.4e-04	1.9e-03

Таблица 8: MSE разложений ряда $\mathsf{X} = \mathsf{X}_c + \mathsf{X}_e + \mathsf{X}_{\cos} + \mathsf{X}_{\sin}$ методов **SSA** и **CiSSA**

Таблица 8 и рисунки 2, 3 показывают, что метод **SSA** справился лучше в сравнении с **CiSSA**, причем как по разделимости, так и по ошибке. В алгоритме **CiSSA** трендовая составляющая также смешалась с сезонной, поэтому увеличилась ошибка при косинусе. Стоит отметить, что в данном примере использовался алгоритм улучшения разделимости EOSSA [2] для метода **SSA**. Без него не получились бы такие результаты.

Или же, если заменить X_e на $\mathsf{X}_{e\cdot\cos}$, то есть теперь ряд $\mathsf{X} = \mathsf{X}_c + \mathsf{X}_{e\cdot\cos} + \mathsf{X}_{\cos} + \mathsf{X}_{\sin} = 1 + e^{\frac{x}{100}}\cos(\frac{2\pi}{48}x) + \cos(\frac{2\pi}{12}x) + \sin(\frac{2\pi}{24}x)$, то получится следующая таблица ошибок:

Метод/Компонента	$\mathsf{X}_{e\cdot\cos}$	$\mathsf{X}_{\sin}$	$\mathsf{X}_{\cos}$
SSA	4.7e-29	1.1e-29	8.4e-30
CiSSA	3.2e-02	2.6e-04	5.8e-03

Таблица 9: MSE разложений ряда $\mathsf{X} = \mathsf{X}_c + \mathsf{X}_{e\cdot\cos} + \mathsf{X}_{\cos} + \mathsf{X}_{\sin}$ методов **SSA** и **CiSSA**

Таким образом, таблица 9 показывает тот же недостаток у метода **CiSSA**, что и таблица 8.

4.5 Отделение сигнала от шума

Рассматривая ряд из предыдущего пункта, добавим к нему гауссовский шум с стандартным отклонением 0.1: $\mathsf{X} = \mathsf{X}_c + \mathsf{X}_e + \mathsf{X}_{\cos} + \mathsf{X}_{\sin} + \mathsf{X}_{\text{noise}} = 1 + e^{\frac{x}{100}} + \cos(\frac{2\pi}{12}x) + \sin(\frac{2\pi}{24}x) + \mathrm{N}(0, 0.1)$, $N = 96 \cdot 2 - 1$, $L = 96$. Сделав такой тест 10000 раз, получим следующий результат по ошибке MSE между настоящим сигналом и его оценкой:

Метод/Статистики	min	median	mean	max	sd
SSA	5.8e-04	2.0e-03	2.1e-03	4.9e-03	6.2e-04
CiSSA	2.5e-02	3.4e-02	3.4e-02	4.9e-02	3.7e-03

Таблица 10: Данные по распределению ошибки восстановления сигнала разложений методов **SSA** и **CiSSA**

По таблице 10 можно увидеть что метод **SSA** отработал лучше **CiSSA**.

4.6 Автоматическая группировка и проверка на реальных данных

Авторы статьи [1] выделяют главным преимуществом то, что **CiSSA** автоматически разделяет компоненты ряда по частотам. Однако есть метод, позволяющий сделать автоматическое объединение частот по периодограмме в методе **SSA** [2]. При этом, прежде чем применять его, стоит выполнить процедуру улучшения разделимости. В данной работе будут использоваться методы EOSSA и FOSSA [2].

Сравним работы этих алгоритмов сначала на модельных примерах, затем на реальных данных.

Используем те же данные, что и в прошлом примере: $\mathsf{X} = \mathsf{X}_c + \mathsf{X}_e + \mathsf{X}_{cos} = 1 + e^{\frac{x}{100}} + cos(\frac{2\pi}{12})$, $N = 96 \cdot 2 - 1$, $L = 96$. Применяем алгоритм EOSSA [2] для лучшей разделимости и выбираем в качестве интересующих частот диапазоны $\left(\frac{1}{24} - \varepsilon, \frac{1}{24} + \varepsilon\right), \left(\frac{1}{12} - \varepsilon, \frac{1}{12} + \varepsilon\right), \varepsilon = \frac{1}{97}$. Результаты остаются теми же, как и в таблице 8 и рисунках 2, 3, однако теперь группировка ряда произошла по интересующим частотам.

Теперь рассмотрим реальные данные — месячные ряды промышленного производства (Industrial Production, IP), index 2010 = 100, в США. 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. Выборка охватывает период с января 1970 года по сентябрь 2014 года, поэтому размер выборки составляет $N = 537$. Источником данных является база данных IMF. Эти показатели демонстрируют различные тенденции, сезонность и цикличность (периодические компоненты, которые соответствуют циклам бизнеса). Данные IP также рассматривались в статье [1]. 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: $\left[0, \frac{1}{192}\right]$;
2. Циклы бизнеса по диапазонам: $\left[\frac{2}{192}, \frac{10}{192}\right]$;
3. Сезонность по частотам $\omega_k = 1/12, 1/6, 1/4, 1/3, 5/12, 1/2$;

На основе предыдущих требований взято $L = 192$.

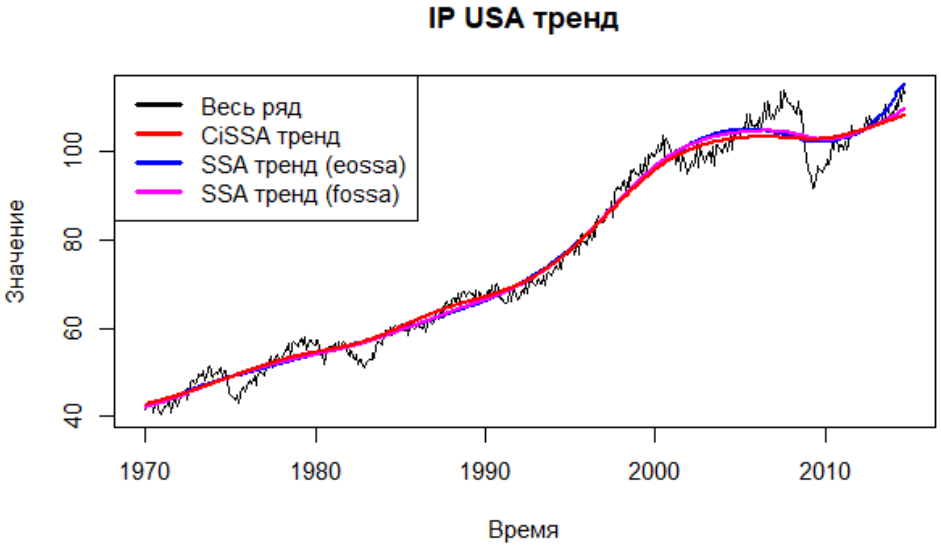

Рис. 4: Трендовая составляющая данных IP USA

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.

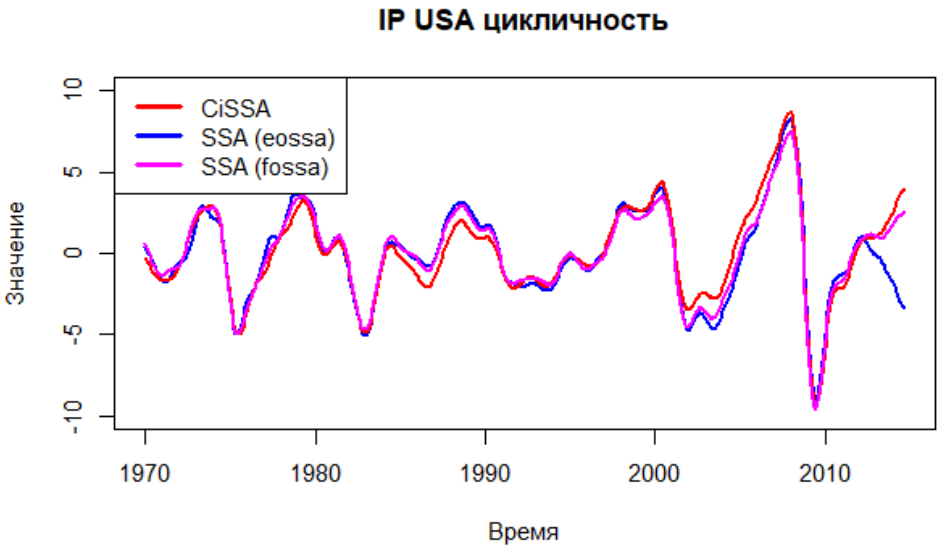

Рис. 5: Циклическая составляющая данных IP USA

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

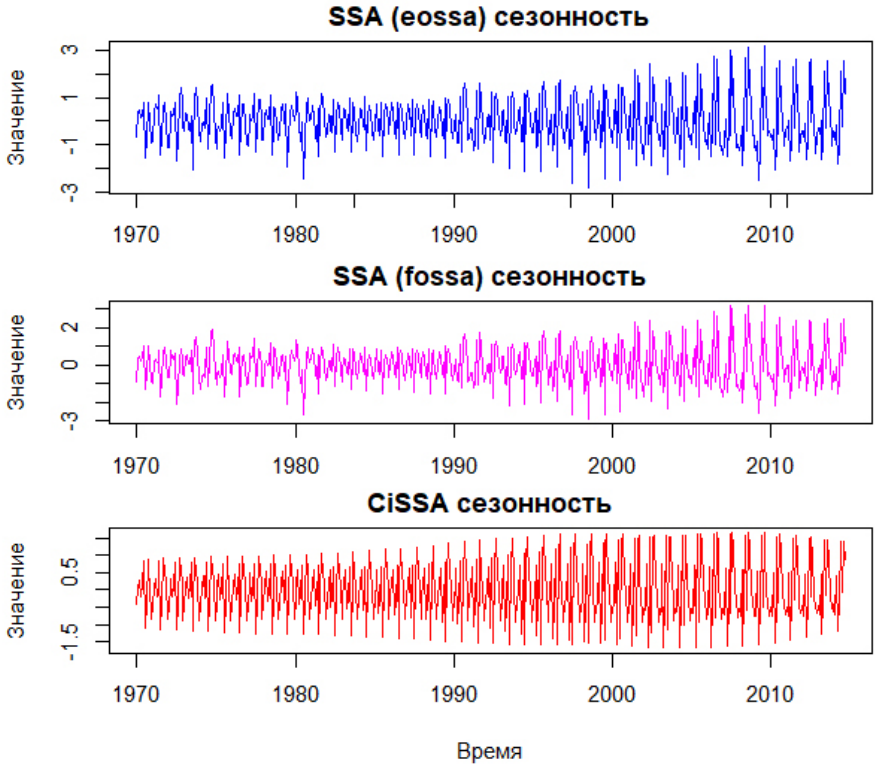

Рис. 6: Сезонная составляющая данных IP USA

Поскольку в базовом **SSA** адаптивный базис, сезонность является менее систематичной, разброс значений выше по сравнению с **CiSSA**.

Шум же является нормальным во всех случаях.

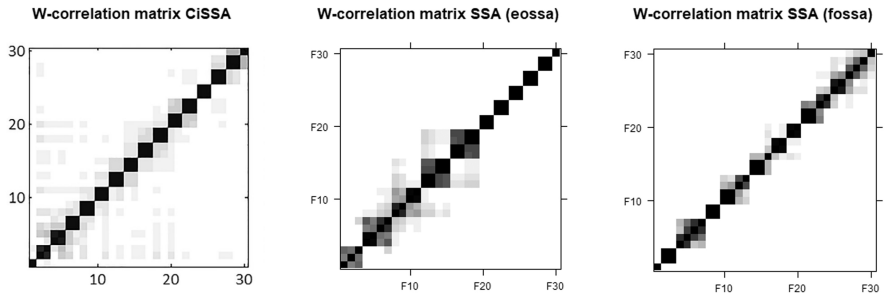

Рис. 7: Матрицы корреляций IP USA

По матрицам корреляции заметно, что при использовании **SSA** с улучшением разделимости EOSSA, сильно смешиваются первые по значимости компоненты ряда (они и являются трендовыми и циклическими).

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**.

4.7 Выводы

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.

5 Заключение

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**.

Список литературы

[1] Juan Bogalo, Pilar Poncela, and Eva Senra. Circulant singular spectrum analysis: A new automated procedure for signal extraction. *Signal Processing*, 177, September 2020. Received 24 March 2020; Revised 19 September 2020; Accepted 22 September 2020; Available online 24 September 2020.

[2] Nina Golyandina, Pavel Dudnik, and Alex Shlemov. Intelligent identification of trend components in singular spectrum analysis. *Algorithms*, 16(7):353, 2023. Submission received: 31 May 2023; Revised: 4 July 2023; Accepted: 14 July 2023; Published: 24 July 2023. This article belongs to the Special Issue Machine Learning for Time Series Analysis. Author to whom correspondence should be addressed: Nina Golyandina.

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

[4] Nikolay Pogrebnikov. SPbSU CISSA coursework: Time series analysis. `https://github.com/xSICHx/spbu_cissa_coursework`, 2024. Accessed: 2024-10-23.