Санкт-Петербургский государственный университет

Третьякова Александра Леонидовна

Выпускная квалификационная работа

Робастные варианты метода анализа сингулярного спектра

Уровень образования: магистратура

Направление 01.04.02 «Прикладная математика и информатика»

Основная образовательная программа ВМ.5688.2018 «Прикладная математика и информатика»

Профессиональная траектория «Статистическое моделирование»

Научный руководитель:

Доцент, кафедра статистического моделирования

к. ф.-м. н., доцент Н. Э. Голяндина

Рецензент:

Лектор, Университет Кардиффа (Великобритания)

к. ф.-м. н. А. Н. Пепелышев

Санкт-Петербург

2020

Saint Petersburg State University
Applied Mathematics and Computer Science
Statistical Modelling

Tretyakova Aleksandra Leonidovna

Graduation Project

Robust versions of the singular spectrum analysis method

Scientific Supervisor:
Associate Professor, Department of
Statistical Modelling N. E. Golyandina

Reviewer:
Lecturer, Cardiff University A. N. Pepelyshev

Saint Petersburg
2020

Оглавление

Введение

В реальной жизни часто возникают задачи исследования различных процессов с течением времени. Пусть имеется $x(t)$ — функция, описывающая некоторый процесс во времени. Если произвести измерения через одинаковые промежутки времени t_i, где $i = 1, \ldots, N$, тогда $x_i = x(t_i)$ представляют собой временной ряд $\mathsf{X} = (x_1, \ldots, x_N)$.

Для решения многих задач, к примеру, экономических, таких как планирование производства или инвестиций, оказывается полезным на основе данных за предшествующий период выделить основную динамику и тенденции, а также спрогнозировать развитие процесса. В данной работе для исследования временных рядов будет применен метод «Гусеница»-SSA (Singular Spectrum Analysis) [1, 2], который позволяет анализировать ряд без задания его параметрической модели. Метод нашел свое применение в задачах исследования климатических явлений [3], динамических систем [4, 5] и во многих других областях. Пусть имеется временной ряд $\mathsf{X} = (x_1, \ldots, x_N)$ длины N, который представляет собой сумму сигнала и шума: $x_i = s_i + r_i$, $i = 1, \ldots, N$. Данный метод позволяет получить разложение интересующего нас временного ряда X на интерпретируемые аддитивные составляющие:

$$\mathsf{X} = \mathsf{S} + \mathsf{R},$$

где S — сигнал, R — шум, например, некоторый стационарный процесс.

На практике часто возникают выделяющиеся наблюдения или выбросы, которые можно интерпретировать как ошибки в данных или сбои измерительного прибора, значительно большие, чем размер шума. Отфильтровать их оказывается непростой задачей, необходимо сначала разобраться со структурой ряда, чтобы понять, что данное значение является выбросом. Поэтому разработка исходно устойчивых к выбросам методов представляет значительный интерес.

Ранее в работе [6] уже были предложены несколько устойчивых к выбросам вариантов метода, но они оказались слишком трудоемкими. Поэтому необходимо найти методы, которые бы оставались устойчивыми, но время работы алгоритмов было бы меньше. В данной работе стоит задача предложить менее трудоемкие варианты метода, сравнить робастные модификации между собой и с базовым SSA.

В методе SSA при выделении сигнала используются два проектора, которые могут строиться по различным нормам. Один из проекторов — это проектор на пространство

ганкелевых матриц, второй — проектор на множество матриц ранга, не превосходящего r. В стандартном методе SSA используются проекторы в пространстве матриц по норме $\mathbb{L}_2$ (норма Фробениуса).

В качестве модификаций в работе [6] рассматривался стандартный прием использования аппроксимации (проекции) по норме в $\mathbb{L}_1$ вместо $\mathbb{L}_2$. Если построение проектора на ганкелевы матрицы по норме $\mathbb{L}_1$ не представляет трудности, то вычисление проектора на матрицы ранга, не превосходящего r, по норме $\mathbb{L}_1$ не имеет решения в замкнутой форме. Имеются методы, численно решающие приближенные задачи, но не известно достаточно хороших методов для задачи, которую требуется решить при построении проектора на матрицы ранга, не превосходящего r.

Еще одной идеей для достижения устойчивости метода к выделяющимся наблюдениям является присвоение значениям в точках, содержащих выбросы, меньший вес. В данной работе рассмотрим алгоритм, описанный в статье [7], включающий в себя итеративное обновление весов. Метод имеет параметры: $\{\sigma_{ij},\ i = 1, \dots, L, j = 1, \dots, K\}$ (нормировка для остатков) и α (значение, начиная с которого точку считать выбросом). В оригинальном методе из статьи [7] параметры σ_{ij} полагаются одинаковыми для всех элементов траекторной матрицы, то есть $\sigma_{ij} = \sigma\ \forall i, j$. Данный метод оказывается неподходящим для рядов, где шум имеет непостоянную дисперсию. Поэтому стоит задача предложить его модификацию, которая бы оставалась устойчивой как в случае рядов с шумом постоянной дисперсии, так и в случае рядов с гетероскедастичным шумом. В модификации метода предполагается замена параметра σ_{ij} на элементы траекторной матрицы тренда (оценки математического ожидания) ряда из модулей остатков. Так как в таком случае матрица $\mathbf{\Sigma} = \{\sigma_{ij}\}_{i,j=1}^{L,K}$ ганкелева, то это соответствует присваиванию весов элементам ряда. Поэтому возникает идея вычислять проекцию не траекторной матрицы ряда, а проектировать сам ряд на множество рядов ранга, не превосходящего r. Поэтому рассмотрим вариант с использованием модифицированного метода Гаусса-Ньютона из статьи [8] для решения задачи построения проекции исходного ряда на множество рядов ранга, не превосходящего r, по взвешенной норме в $\mathbb{L}_2$.

Структура работы следующая. В главе 1 опишем базовый алгоритм метода SSA, введем необходимые понятия и обозначения, обсудим выбор параметров метода на основе теории метода SSA.

В начале главы 2 рассмотрим общую схему методов без указания конкретной нор-

мы. Приведем вид проекторов на пространство ганкелевых матриц по трем нормам: по норме в пространствах $\mathbb{L}_2$, $\mathbb{L}_1$ и взвешенной норме в $\mathbb{L}_2$. В данной главе представим нахождение $\mathbb{L}_1$-проектора на множество матриц ранга, не превосходящего r. В работе идет речь о последовательном методе проектирования. Этот метод рассматривается в R-пакете в рамках построения устойчивого к выбросам анализа главных компонент [9]. Далее опишем метод с итеративным обновлением весов и его модификацию, подходящую для рядов с гетероскедастичным шумом. Также опишем модификацию этого метода, где веса присваиваются не траекторной матрице, а самому ряду. Приведем алгоритмы для каждого из методов. Произведем подсчет и сравнение теоретической трудоемкости описанных методов.

Глава 3 будет содержать численные сравнения с исследованием влияния выброса на результат восстановления сигнала. В данной главе представим сравнение рассмотренных методов между собой и с классическим SSA. Сравнение будем проводить при отсутствии выделяющихся наблюдений, при 1% выбросов в случайных точках ряда и при 5% выбросов.

Проведем сравнение на нескольких модельных примерах. Один из них уже был представлен в работе [6], но мы добавим большее количество выбросов. Второй пример — ряд с растущей амплитудой с шумом, имеющим непостоянную дисперсию. Третий пример — ряд с сильно растущей амплитудой и большим разбросом значений, но с шумом постоянной дисперсии. В конце проведем сравнение нескольких методов на реальном ряде с выбросами.

В заключении опишем основные результаты работы, подведем итоги. На основе проведенных исследований дадим рекомендации, какой из рассмотренных методов следует использовать в том или ином случае.

Глава 1

Стандартный метод SSA и его свойства

1.1. Алгоритм метода SSA

Кратко опишем базовый алгоритм метода «Гусеница»-SSA, следуя [2].

1.1.1. Вложение

На первом шаге алгоритма выбирается некоторое целое число L: $1 < L < N$, называемое *длиной окна*. Исходный временной ряд переводится в последовательность многомерных векторов длины L. В результате образуются $K = N - L + 1$ векторов вложения

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

Траекторной матрицей ряда X называется матрица

$$\mathbf{X} = [X_1 : \ldots : X_K] = \begin{pmatrix} x_1 & x_2 & x_3 & \ldots & x_K \\ x_2 & x_3 & x_4 & \ldots & x_{K+1} \\ x_3 & x_4 & x_5 & \ldots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \ldots & x_N \end{pmatrix}.$$

Заметим, что построенная таким образом траекторная матрица $\mathbf{X}$ является *ганкелевой*, т.е. элементы, находящиеся на диагоналях $i + j = \text{const}$, равны между собой.

1.1.2. Сингулярное разложение

Пусть $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$, обозначим $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_d > 0$ — ненулевые собственные числа матрицы $\mathbf{S}$, $U_1, \ldots, U_d$ — ортонормированная система собственных векторов матрицы $\mathbf{S}$, соответствующих собственным числам. *Сингулярным разложением* матрицы $\mathbf{X}$ называется разложение

$$\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}},$$

где $\sqrt{\lambda_i}$ — *сингулярные числа*, U_i — *левые сингулярные вектора*, $V_i = \frac{1}{\sqrt{\lambda_i}} X^{\mathrm{T}} U_i$ — *правые сингулярные вектора*.

Набор $(\sqrt{\lambda_i}, U_i, V_i)$ назовем *i-ой собственной тройкой* сингулярного разложения.

1.1.3. Группировка

Разделим множество индексов $\{1, \ldots, d\}$ на m дизъюнктных подмножеств $I_1, \ldots, I_m$. Пусть $I = \{i_1, \ldots, i_p\}$. Тогда *результирующая матрица* $\mathbf{X}_I$ имеет вид

$$\mathbf{X}_I = \mathbf{X}_{i_1} + \ldots + \mathbf{X}_{i_p}.$$

Таким образом, получаем разложение матрицы $\mathbf{X}$ в сгруппированном виде

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

1.1.4. Диагональное усреднение

На последнем шаге каждая матрица $\mathbf{X}_{I_i}$ переводится в новый ряд с помощью усреднения элементов матрицы вдоль антидиагоналей $i + j = k + 1$. Применяя диагональное усреднение к результирующим матрицам, получаем ряды $\tilde{\mathsf{X}}^{(k)} = (\tilde{x}_1^k, \ldots, \tilde{x}_N^k)$.

В результате получаем разложение исходного ряда $(x_1, \ldots, x_N)$ в сумму m рядов:

$$x_n = \sum_{k=1}^{m} \tilde{x}_n^{(k)}.$$

1.2. Разделимость

Введем понятие разделимости, следуя [2].

Пусть X_1 и X_2 — временные ряды длины N и $\mathsf{X} = \mathsf{X}_1 + \mathsf{X}_2$. Пусть выбрана длина окна L, тогда каждый из рядов порождает L-траекторную матрицу: $\mathbf{X}_1$, $\mathbf{X}_2$ и $\mathbf{X}$. Обозначим $\mathcal{L}_1^L$ и $\mathcal{L}_2^L$ — линейные пространства, порожденные столбцами траекторных матриц $\mathbf{X}_1$ и $\mathbf{X}_2$. Аналогично обозначим $\mathcal{L}_1^K$ и $\mathcal{L}_2^K$ — линейные пространства, порожденные строками траекторных матриц $\mathbf{X}_1$ и $\mathbf{X}_2$, $K = N - L + 1$.

Определение 1. *Ряды X_1 и X_2 называются слабо L-разделимыми, если $\mathcal{L}_1^L \perp \mathcal{L}_2^L$ и $\mathcal{L}_1^K \perp \mathcal{L}_2^K$.*

В результате выполнения этапа разложения в алгоритме SSA получаем некоторое разложение траекторной матрицы ряда, которое не обязательно соответствует разделимости двух рядов. Поэтому необходимо усилить понятие разделимости.

Определение 2. *Если ряды* X_1 *и* X_2 *слабо* L*-разделимы и множество собственных чисел сингулярного разложения траекторной матрицы одного ряда не пересекается с множеством собственных чисел разложения второго ряда* ($\lambda_{1k} \neq \lambda_{2m}\ \forall k, m$)*, то ряды* X_1 *и* X_2 *сильно* L*-разделимы.*

Утверждение 1. *Пусть ряды* X_1 *и* X_2 *сильно* L*-разделимы. Тогда любое сингулярное разложение траекторной матрицы* $\mathbf{X}$ *ряда* X *можно разбить на две части, являющиеся сингулярными разложениями траекторных матриц рядов* X_1 *и* X_2.

Однако условия разделимости являются слишком жесткими и редко выполнены в реальных задачах. Поэтому в пособии [2] введено также понятие асимптотической разделимости.

Асимптотическая раделимость выполняется для более широкого класса рядов, чем точная разделимость. К примеру, $e^{\alpha n}$ и $\sin(2\pi\omega n)$, где $\alpha \neq 0,\ \omega \in (0, 0.5]$, асимптотически разделимы.

Для достижения лучшей разделимости необходимо выбирать большую длину окна ($L \sim N/2$). Большая длина окна позволяет выделить сигнал из зашумленного ряда, отделить тренд от периодических компонент. Не имеет смысла брать длину окна, большую чем половина длины ряда, а маленькая длина окна может привести к смешиванию компонент ряда.

1.3. Ранг ряда

Пусть $\mathsf{X}_N = \mathsf{X}_N^{(1)} + \mathsf{X}_N^{(2)}$ и ряды $\mathsf{X}_N^{(1)}$ и $\mathsf{X}_N^{(2)}$ разделимы. Тогда в сингулярном разложении ряда X_N часть слагаемых относится к сингулярному разложению ряда $\mathsf{X}_N^{(1)}$, а другая часть — к сингулярному разложению ряда $\mathsf{X}_N^{(2)}$. Необходимо выяснить, сколько слагаемых относится к первому ряду и как их идентифицировать. На этапе группировки индексы слагаемых, относящихся к первому ряду, образуют подмножество I_1, остальные — подмножество I_2. Для того, чтобы понимать, сколько компонент относить к первому подмножеству, введем понятие ранга.

Рассмотрим ряд $\mathsf{X}_N = (x_1, \ldots, x_N)$, пусть L – длина окна.

Обозначим $\mathcal{L}^{(L)} = \mathrm{span}(X_1, \ldots, X_K)$ — траекторное пространство ряда X_N, где $X_i = (x_i, \ldots, x_{i+L-1})^T$ — векторы вложения, $1 \leq i \leq K$.

Определение 3. *Пусть* $0 < d \leq \min(L, K)$. *Будем говорить, что ряд* X_N *имеет* L*-ранг* d, *если* $\dim \mathcal{L}^{(L)} = d$.

Например, в случае экспоненциального ряда $e^{\alpha n}$ для любых N и L ранг ряда равен 1, а ранг гармонического ряда $\sin(2\pi\omega n + \phi)$ равен 2 при $\omega < 1/2$ и 1 при $\omega = 1/2$, $\phi \in [0, 2\pi)$.

Глава 2

Модификации метода SSA с проекторами по некоторой норме

Будем рассматривать вариант метода SSA для выделения сигнала, когда группировка заключается в выборе первых r компонент. Для стандартного метода SSA это эквивалентно проекции по норме Фробениуса траекторной матрицы ряда на множество матриц ранга, не превосходящего r.

2.1. Схема методов

Пусть имеется временной ряд $\mathsf{X} = (x_1, \ldots, x_N)$.

Выбирается длина окна L, и исходный временной ряд переводится в последовательность многомерных векторов длины L. В результате образуются $K = N - L + 1$ векторов вложения

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

Обозначим $\mathcal{M}$ — пространство матриц $L \times K$,

$\mathcal{M}_{\mathcal{H}}$ — пространство ганкелевых матриц $L \times K$,

$\mathcal{M}_r$ — пространство матриц ранга, не превосходящего r.

Определим следующие операторы:

- Оператор вложения $\mathcal{T} : \mathbb{R}^N \to \mathcal{M}_{\mathcal{H}} : \mathcal{T}(\mathsf{X}) = \mathbf{X}$.
- $\Pi_r : \mathcal{M} \to \mathcal{M}_r$ — проектор на множество матриц ранга, не превосходящего r, по некоторой норме в пространстве матриц.
- $\Pi_{\mathcal{H}} : \mathcal{M} \to \mathcal{M}_{\mathcal{H}}$ — проектор на пространство ганкелевых матриц по некоторой норме в пространстве матриц.

В результате применения данных операторов получаем оценку сигнала:

$$\tilde{\mathsf{S}} = \mathcal{T}^{-1}\Pi_{\mathcal{H}}\Pi_r\mathcal{T}(\mathsf{X}).$$

Это соответствует алгоритму SSA, описанному в разделе 1.1, для случая, когда восстановление производится по одной группе, состоящей из первых r компонент.

Проекторы Π_r и $\Pi_{\mathcal{H}}$ можно строить по различным нормам. С точки зрения вычислений, удобно выбирать $\mathbb{L}_2$-норму для построения проекторов на пространство ганкелевых матриц и матриц ранга, не превосходящего r, поскольку целевая функция является гладкой и выпуклой, и решить задачу минимизации довольно просто, можно даже говорить о задании решения в явной форме. Однако при наличии выбросов норма Фробениуса оказывается недостаточно устойчивой. Норма в пространстве $\mathbb{L}_1$ является более устойчивой к выделяющимся наблюдениям, однако сложность в ее использовании состоит в негладкой и невыпуклой строго целевой функции, поэтому возникает проблема в применении стандартных методов оптимизации. Существует также вариант использования взвешенной нормы в $\mathbb{L}_2$ с присваиванием меньших весов точкам, содержащим выделяющиеся наблюдения.

В работе будут рассмотрены проекторы по нормам в пространствах $\mathbb{L}_2$, $\mathbb{L}_1$ и взвешенной норме в пространстве $\mathbb{L}_2$. Будем рассматривать следующие методы:

- Стандартный метод SSA, где оба проектора Π_r и $\Pi_{\mathcal{H}}$ строятся по норме в пространстве $\mathbb{L}_2$,
- Метод с проекторами Π_r и $\Pi_{\mathcal{H}}$ по норме в пространстве $\mathbb{L}_1$,
- Метод с проекторами Π_r и $\Pi_{\mathcal{H}}$ по взвешенной норме в пространстве $\mathbb{L}_2$.

2.2. Вид проекторов на пространство ганкелевых матриц по различным нормам

Рассмотрим, как выглядят проекторы на пространство ганкелевых матриц по нормам в пространствах $\mathbb{L}_2$, $\mathbb{L}_1$ и взвешенной норме в $\mathbb{L}_2$.

Определение 4. *Пусть* $\mathbf{A}$ *— матрица* $L \times K$.

Норма в пространстве $\mathbb{L}_2$ *(норма Фробениуса):* $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum\limits_{i=1}^{L}\sum\limits_{j=1}^{K} a_{ij}^2}$.

Лемма 1 (Известный результат) $\operatorname*{argmin}\limits_a \mathbb{E}(\xi - a)^2 = \mathbb{E}\xi$.

Доказательство.

$$\mathbb{E}(\xi - a)^2 = \mathbb{E}((\xi - \mathbb{E}\xi) + (\mathbb{E}\xi - a))^2 = \mathbb{E}((\xi - \mathbb{E}\xi)^2 + 2(\xi - \mathbb{E}\xi)(\mathbb{E}\xi - a) + (\mathbb{E}\xi - a)^2) =$$

$$= \mathbb{E}(\xi - \mathbb{E}\xi)^2 + 2(\mathbb{E}\xi - a)\mathbb{E}(\xi - \mathbb{E}\xi) + (\mathbb{E}\xi - a)^2 = \mathbb{E}(\xi - \mathbb{E}\xi)^2 + (\mathbb{E}\xi - a)^2.$$

Следовательно, $\operatorname*{argmin}_a \mathbb{E}(\xi - a)^2 = \mathbb{E}\xi$. $\square$

Необходимо построить проектор на множество ганкелевых матриц по норме Фробениуса. Заметим, что

$$\|\mathbf{X} - \mathbf{A}\|_{\mathrm{F}}^2 = \sum_{i=1}^{L}\sum_{j=1}^{K}(x_{ij} - a_{ij})^2 = \sum_{k=1}^{L+K-1}\sum_{i+j=k+1}(x_{ij} - a_k)^2 \longrightarrow \min_{\mathbf{A}\in\mathcal{M}_{\mathcal{H}}}.$$

Тогда если взять в качестве ξ случайные величины, принимающие значения на побочной диагонали с равными вероятностями, то из Леммы 1 следует, что проектор на множество ганкелевых матриц по норме Фробениуса строится посредством усреднения элементов на диагоналях $i + j = \mathrm{const}$.

Опишем теперь вид данного проектора по норме в пространстве $\mathbb{L}_1$.

Определение 5. *Пусть* $\mathbf{A}$ *— матрица* $L \times K$.

Норма в пространстве $\mathbb{L}_1$ *:* $\|\mathbf{A}\|_1 = \sum\limits_{i,j}|a_{ij}|$.

Лемма 2 (Известный результат)

В случае непрерывного распределения $\operatorname*{argmin}_a \mathbb{E}|\xi - a| = \operatorname{med}\xi$.

Доказательство этой леммы приведено в работе [6].

Используя лемму 2, получаем, что $\Pi_{\mathcal{H}}$ в пространстве $\mathbb{L}_1$ строится посредством взятия выборочной медианы значений на побочных диагоналях $i + j = \mathrm{const}$.

Определение 6. *Пусть* $\mathbf{A}$ *— матрица* $L \times K$, $\mathbf{W}$ *— матрица весов* $L \times K$.

Норма в пространстве $\mathbb{L}_2$ *с весами* $\mathbf{W}$ *:* $\|\mathbf{A}\|_{\mathrm{W}} = \sqrt{\sum\limits_{i=1}^{L}\sum\limits_{j=1}^{K} w_{ij}a_{ij}^2}$.

Известно следующее утверждение о построении проектора на пространство ганкелевых матриц по данной норме.

Утверждение 2 ([10]) *Для построения проекции* $\widehat{\mathbf{A}} = \Pi_{\mathcal{H}}\mathbf{A} = \{\widehat{a}_{ij}\}_{i,j=1}^{L,K}$ *необходимо суммировать элементы на диагоналях* $i + j = const$ *с весами и нормировать на сумму весов:*

$$\widehat{a}_{ij} = \frac{\sum\limits_{l,k:l+k=i+j} w_{lk}a_{lk}}{\sum\limits_{l,k:l+k=i+j} w_{lk}}.$$

Замечание 1. *В случае ганкелевой матрицы весов* $\mathbf{W}$ *проектор на пространство ганкелевых матриц по взвешенной норме в* $\mathbb{L}_2$ *совпадает с проектором на пространство ганкелевых матриц по норме в* $\mathbb{L}_2$.

2.3. Построение проектора по норме в $\mathbb{L}_1$ на множество матриц ранга, не превосходящего r. Последовательный метод

В отличие от проектора на множество матриц ранга r по норме Фробениуса, построение данного проектора в пространстве $\mathbb{L}_1$ является вычислительно сложной задачей. Рассмотрим один из методов построения проектора на множество матриц ранга, не превосходящего r, по норме в пространстве $\mathbb{L}_1$.

Стоит задача проектирования матрицы $\mathbf{X}$ на множество матриц ранга, не превосходящего r. Задачу оптимизации можно представить в виде

$$\min_{\mathbf{V},\mathbf{U}} \|\mathbf{X}^{\mathrm{T}} - \mathbf{V}\mathbf{U}^{\mathrm{T}}\|_1 = \sum_{i=1}^{L} \|X_i - \mathbf{V}U_i\|_1,$$

где $\mathbf{V}$ — матрица $K \times r$, $\mathbf{U}$ — матрица $L \times r$. Столбцы матрицы $\mathbf{V}$ определяют главные компоненты. Матрица $\mathbf{E} = \mathbf{U}\mathbf{V}^{\mathrm{T}}$ — проекция $\mathbf{X}$ на множество матриц ранга, не превосходящего r, которую необходимо найти.

В пакете pcaL1 [11] имеется метод l1pca, позволяющий вычислить проекцию в $\mathbb{L}_1$ на множество матриц ранга, не превосходящего r. Подробнее метод описан в статье [9].

Приведем алгоритм в наших обозначениях. Пусть $\mathbf{Y} \in \mathbb{R}^{L\times K}$. Задача выглядит следующим образом:

$$\|\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}\|_1 \to \min_{\mathbf{U},\mathbf{V}}.$$

Далее представлен алгоритм последовательного метода решения данной задачи.

Алгоритм 1: Последовательный метод построения $\mathbb{L}_1$-проектора на множество матриц ранга, не превосходящего r

Входные данные: $\mathbf{Y} \in \mathbb{R}^{L\times K}$ — траекторная матрица ряда, r — ранг сигнала; параметры критерия остановки: $\varepsilon = 10^{-4}$, максимальное число итераций $N_{iter} = 10$

Выходные данные: $\widehat{\mathbf{Y}} = \mathbf{U}\mathbf{V}^{\mathrm{T}}$ — проекция траекторной матрицы на множество матриц ранга, не превосходящего r

1. Инициализация $\mathbf{U}(0) \in \mathbb{R}^{L\times r}$, нормировка столбцов $\mathbf{U}(0)$;

2. $t := 0$;

3. **повторять**

   a. $t := t + 1$;

   b. $\mathbf{V}(t) = \underset{\mathbf{V}\in\mathbb{R}^{K\times r}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{U}(t-1)\mathbf{V}^{\mathrm{T}}\|_1$;

   c. $\mathbf{U}(t) = \underset{\mathbf{U}\in\mathbb{R}^{L\times r}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}(t)\|_1$;

   d. Нормировка столбцов $\mathbf{U}(t)$;

до тех пор, пока $\max\limits_{i=1,\ldots,L,j=1,\ldots,r} |u_{ij}(t) - u_{ij}(t-1)| > \varepsilon$ u $t < N_{iter}$;

4. $\mathbf{U} := \mathbf{U}(t)$; $\mathbf{V} := \mathbf{V}(t)$

Решаем задачу, меняя на каждой итерации $\mathbf{U}$ и $\mathbf{V}$ и разбивая исходную задачу на линейные подзадачи. $\mathbf{U}(0)$ можно инициализировать с помощью сингулярного разложения траекторной матрицы $\mathbf{Y}$ в пространстве $\mathbb{L}_2$.

Рассмотрим подробнее решение задачи

$$\mathbf{V}(t) = \underset{\mathbf{V}\in\mathbb{R}^{K\times r}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{U}(t-1)\mathbf{V}^{\mathrm{T}}\|_1. \tag{2.1}$$

Целевую функцию можно представить в виде

$$\|\mathbf{Y} - \mathbf{U}(t-1)\mathbf{V}^{\mathrm{T}}\|_1 = \sum_{i=1}^{K} \|Y_i - \mathbf{U}(t-1)\mathbf{v}_i\|_1,$$

где $Y_i \in \mathbb{R}^L$ — столбцы $\mathbf{Y}$, $\mathbf{v}_i \in \mathbb{R}^r$ — строки $\mathbf{V}$. Согласно [12], задача (2.1) может быть разбита на K независимых небольших подзадач

$$\mathbf{v}_i = \underset{\mathbf{x}}{\operatorname{argmin}} \|Y_i - \mathbf{U}(t-1)\mathbf{x}\|_1. \tag{2.2}$$

С помощью решения каждой такой подзадачи получаем оценку $\mathbf{v}_i$, $i = 1, \ldots, K$. Глобальное решение задачи (2.2) находится с помощью задачи линейного программирования

$$\min_{\boldsymbol{\delta}} \mathbf{1}^{\mathrm{T}} \boldsymbol{\delta}$$

с ограничениями

$$-\boldsymbol{\delta} \leq Y_i - \mathbf{U}(t-1)\mathbf{x} \leq \boldsymbol{\delta},$$

где $\mathbf{1} \in \mathbb{R}^L$ — вектор-столбец, состоящий из единиц. Другими словами, задача линейного программирования находит минимальную границу $\boldsymbol{\delta}$ такую, что область

$$-\boldsymbol{\delta} \leq Y_i - \mathbf{U}(t-1)\mathbf{x} \leq \boldsymbol{\delta}$$

является непустой.

Задача $\mathbf{U}(t) = \underset{\mathbf{U} \in \mathbb{R}^{L \times r}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}(t)\|_1$ решается аналогичным образом.

2.4. Построение проектора по взвешенной норме в $\mathbb{L}_2$ на множество матриц ранга, не превосходящего r

В данном разделе опишем метод с итеративным обновлением весов из статьи [7], а также модификацию этого метода, подходящую для рядов с гетероскедастичным шумом.

2.4.1. Метод с итеративным обновлением весов

Пусть $\mathbf{Y} \in \mathbb{R}^{L \times K}$ — траекторная матрица ряда. Необходимо решить задачу

$$\left\| \mathbf{W}^{1/2} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}) \right\|_F^2 \longrightarrow \min_{\mathbf{U},\mathbf{V}},$$

где $\odot$ — поэлементное умножение, $\mathbf{W}^{1/2}$ — поэлементное взятие корня, веса $w_{ij} = w(\frac{y_{ij} - \hat{y}_{ij}}{\sigma_{ij}})$ вычисляются по формуле, как и в известном методе локальной регрессии loess,

$$w(x) = \begin{cases} (1 - (\frac{|x|}{\alpha})^2)^2, & |x| \leq \alpha \\ 0, & |x| > \alpha \end{cases}.$$

Значения α и $\{\sigma_{ij},\ i = 1, \ldots, L,\ j = 1, \ldots, K\}$ — параметры. График весовой функции $w(x)$ представлен на рисунке 2.1. Значение веса зависит от нормированных остатков:

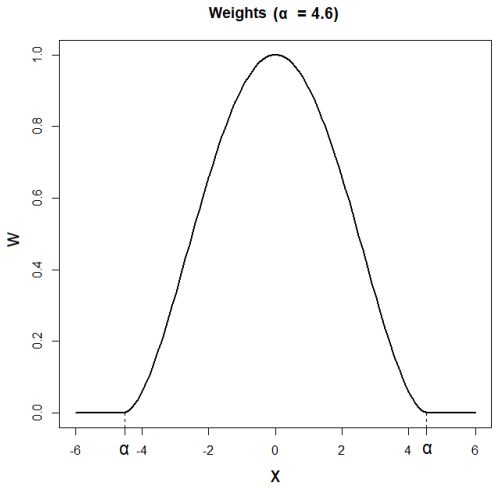

Рис. 2.1. График весовой функции $w(x)$.

если остаток маленький по модулю, то вес максимальный, если остаток по модулю превосходит заданный параметр α, то вес обнуляется.

Алгоритм решения этой задачи представлен в статье [7]. Для начала опишем алгоритм решения задачи аппроксимации для фиксированной матрицы весов $\mathbf{W} \in \mathbb{R}^{L \times K}$. Алгоритм 2 решает задачу

$$\left\| \mathbf{W}^{1/2} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}) \right\|_F^2 \longrightarrow \min_{\mathbf{U},\mathbf{V}},$$

при фиксированной $\mathbf{W}$.

Алгоритм 2: Алгоритм решения задачи взвешенной аппроксимации для фиксированной матрицы весов $\mathbf{W}$

Входные данные: $\mathbf{Y} \in \mathbb{R}^{L\times K}$ — траекторная матрица ряда, r — ранг сигнала, $\mathbf{W} \in \mathbb{R}^{L\times K}$ — матрица весов;
параметры критерия остановки: $\varepsilon = 10^{-4}$,
максимальное число итераций $N_\alpha = 5$

Выходные данные: $\widehat{\mathbf{Y}} = \mathbf{U}\mathbf{V}^{\mathrm{T}}$ — решение задачи взвешенной аппроксимации при фиксированной матрице весов $\mathbf{W}$

1. $t := 0$;
2. **повторять**
   a. Вычисление матрицы $\mathbf{U} \in \mathbb{R}^{L\times r}$ с помощью решения задачи

$$(y_i - \mathbf{V}u_i)^{\mathrm{T}}\mathbf{W}_i(y_i - \mathbf{V}u_i) \to \min_{u_i}, \quad i = 1, \dots L, \tag{2.3}$$

   где $\mathbf{W}_i = \mathrm{diag}(w_i) \in \mathbb{R}^{K\times K}$ — матрица,составленная из i-ой строки $\mathbf{W}$;
   b. Вычисление матрицы $\mathbf{V} \in \mathbb{R}^{K\times r}$ с помощью решения задачи

$$(Y_j - \mathbf{U}v_j)^{\mathrm{T}}\mathbf{W}^j(Y_j - \mathbf{U}v_j) \to \min_{v_j}, , \quad j = 1, \dots K, \tag{2.4}$$

   где $\mathbf{W}^j = \mathrm{diag}(W_j) \in \mathbb{R}^{L\times L}$ — матрица, составленная из j-го столбца $\mathbf{W}$;
   c. $t := t + 1$.

до тех пор, пока $\left\|\mathbf{W}^{1/2} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T})}\right\|_F^2 > \varepsilon$ *u* $t < N_\alpha$;

Замечание 2. *1. Задача 2.3 решается с помощью QR-разложения матрицы* $\mathbf{V}^{\mathrm{T}}\mathbf{W}_i\mathbf{V}$. *2. Задача 2.4 решается с помощью QR-разложения матрицы* $\mathbf{U}^{\mathrm{T}}\mathbf{W}^j\mathbf{U}$.

Далее представлен алгоритм решения задачи построения проектора на множество матриц ранга, не превосходящего r, методом с итеративным обновлением весов. Алгоритм содержит вычисление вспомогательной матрицы $\mathbf{\Sigma} = \{\sigma_{ij}\}_{i,j=1}^{L,K}$, которое обсудим далее в разделе 2.4.2.

Алгоритм 3: Метод с итеративным обновлением весов для нахождения проекции на множество матриц ранга, не превосходящего r

Входные данные: $\mathbf{Y} \in \mathbb{R}^{L \times K}$ — траекторная матрица ряда, r — ранг сигнала; параметры критерия остановки: $\varepsilon = 10^{-4}$, максимальное число итераций $N_{IRLS} = 10$

Выходные данные: $\widehat{\mathbf{Y}} = \mathbf{U}\mathbf{V}^{\mathrm{T}}$ — проекция траекторной матрицы на множество матриц ранга, не превосходящего r

1. Инициализация $\mathbf{U} \in \mathbb{R}^{n \times r}$ и $\mathbf{V} \in \mathbb{R}^{p \times r}$ (например, с помощью сингулярного разложения матрицы $\mathbf{Y}$);
2. Выбор параметра α;
3. $t := 0$;
4. **повторять**
   a. Вычисление матрицы остатков $\mathbf{R} = \{r_{ij}\}_{i,j=1}^{n,p} = \mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}$;
   b. Обновление матрицы $\mathbf{\Sigma} = \{\sigma_{ij}\}_{i,j=1}^{L,K}$;
   c. Вычисление матрицы весов $\mathbf{W} = \{w_{ij}\}_{i,j=1}^{L,K} = \{w(\frac{r_{ij}}{\sigma_{ij}})\}_{i,j=1}^{L,K}$, используя

$$w(x) = \begin{cases} (1 - (\frac{|x|}{\alpha})^2)^2, & |x| \le \alpha \\ 0, & |x| > \alpha \end{cases},$$

   d. Решение задачи взвешенной аппроксимации (обновление матриц $\mathbf{U}$, $\mathbf{V}$)

$$\left\| \mathbf{W}^{1/2} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}) \right\|_F^2 \longrightarrow \min_{\mathbf{U},\mathbf{V}}$$

   с помощью алгоритма 2;
   e. $t := t + 1$.

до тех пор, пока $\left\| \mathbf{W}^{1/2} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}) \right\|_F^2 > \varepsilon$ u $t < N_{IRLS}$;

Авторы статьи [7], ссылаясь на проведенные численные эксперименты, предлагают $\forall\ i, j$ выбрать $\sigma_{ij} = \sigma = 1.4826 \operatorname{med} |\mathsf{R} - \operatorname{med} |\mathsf{R}||$, где R — это вектор, составленный из всех элементов матрицы остатков $\mathbf{R} = \{r_{ij}\}_{i,j=1}^{L,K}$, то есть

$$\mathsf{R} \ = \ (r_{11}, \ldots, r_{1K}; r_{21}, \ldots r_{2K}; \ldots; r_{L1}, \ldots, r_{LK}).$$

Параметр α предлагается взять равным 4.685. Также говорится, что максимальное ко-

личество итераций N_α и N_{IRLS}, необходимых для сходимости, достаточно взять 5 и 10 для достижения приемлемой точности.

Инициализировать матрицы $\mathbf{U}$ и $\mathbf{V}$ можно с помощью первых r компонент сингулярного разложения матрицы $\mathbf{Y}$. Будем использовать один из вариантов truncated SVD из пакета svd [13], который находит только заданное число компонент, не вычисляя полное разложение. Подробнее о методе говорится в статье [14].

2.4.2. Выбор параметра σ_{ij}

У выбора σ_{ij} не зависящими от i, j присутствуют существенные недостатки. Описанный алгоритм из статьи не подходит, к примеру, для рядов с гетероскедастичным шумом. По умолчанию остатки нормировались на σ_{ij}, которая задавалась константой. Нормировка остатков на константный параметр $\sigma_{ij} = \sigma \quad \forall i, j$ в случае шума с непостоянной дисперсией приводит к неправильной идентификации точек с выбросами. Если шум растет к концу ряда, то веса у всех значений на конце ряда некорректно занижаются, и точки, не содержащие выбросов, могут получить вес, меньший, чем у выбросов в начале ряда. Поэтому приходим к выводу, что нормирующий параметр необходимо задавать динамически.

Будем рассматривать матрицу $\mathbf{\Sigma} = \{\sigma_{ij}\}_{i,j=1}^{L,K}$ ганкелевой, что соответствует приписыванию весов элементам ряда. Тогда элементы на диагоналях $i + j = \text{const}$ матрицы $\mathbf{\Sigma}$ равны между собой, и можем обозначить $\sigma_{i+j-1} = \sigma_{lk} \quad \forall l, k : \; l + k = i + j$, где $i = 1, \ldots, L, \; j = 1, \ldots, K$. Обозначим новый параметр $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)^{\mathrm{T}}$. Будем задавать параметр $\boldsymbol{\sigma}$ как тренд (математическое ожидание) ряда, состоящего из модулей остатков. Для оценки математического ожидания как тренд из ряда модулей остатков будем предполагать медленную зависимость модуля остатков от индекса. Выделять тренд будем следующими способами: локальной регрессией loess, скользящей медианой или взвешенной локальной регрессией lowess.

2.4.3. Выбор параметра α

У метода с итеративным обновлением весов есть параметр α, который влияет на то, какие точки будем считать выбросами, а какие — нет. Для того, чтобы понять, какое значение следует взять в качестве α, выведем вероятностную формулу для этого параметра. Будем задавать вероятность γ и подбирать параметр α так, чтобы нормиро-

ванные модули остатков попадали в промежуток $(0;\alpha)$ с вероятностью γ. Для вывода вероятностной формулы для параметра α введем следующее определение.

Определение 7. *Если* $r \sim N(0,\sigma^2)$, *то* $|r| \sim N_H(\sigma^2)$ — *полунормальное распределение с параметром* σ^2, *функция распределения:*

$$F_H(x;\sigma^2) = \frac{2}{\sqrt{\pi}} \int\limits_0^{x/\sqrt{2}\sigma^2} e^{-z^2} dz = \operatorname{erf}(\frac{x}{\sqrt{2}\sigma^2}). \tag{2.5}$$

Среднее и дисперсия: $\mathbb{E}|r| = \sigma\sqrt{\frac{2}{\pi}}, \quad \mathbb{D}|r| = \sigma^2(1-\frac{2}{\pi})$.

После ганкелизации матрицы остатков $\mathbf{R} = \mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}$, получаем ряд из остатков $\mathsf{R} = \{r_i\}_{i=1}^N$. Напомним, что мы рассматриваем ряд вида

$$x_i = s_i + \varepsilon_i, \quad i = 1,\dots,N,$$

где s_i — сигнал, ε_i — шум. Если предположить отделимость сигнала от шума, то матрица $\mathbf{R}$ соответствует траекторной матрице шума $\{\varepsilon_i\}_{i=1}^N$. Для случайного шума точная отделимость от сигнала невозможна, однако будем предполагать разделимость в дальнейших рассуждениях. Тогда в предположении точной отделимости сигнала от шума ряд из остатков соответствует шуму, то есть $\mathsf{R} = \{r_i\}_{i=1}^N = \{\varepsilon_i\}_{i=1}^N$. Обозначим $r^* = \frac{|\varepsilon|}{\sigma}$, где $\boldsymbol{\sigma} = (\sigma_1,\dots,\sigma_N)$ — тренд из ряда $|\mathsf{R}|$. В предположении точной отделимости это означает, что $\sigma_i = \mathbb{E}|\varepsilon_i|$.

Зададим вероятность γ: $\mathrm{P}(r^* \in (0,\alpha)) = \gamma$. Распишем:

$$\mathrm{P}(r^* \in (0,\alpha)) = \mathrm{P}(0 \le \frac{|\varepsilon|}{\sigma} \le \alpha) = \mathrm{P}(\frac{|\varepsilon|}{\sigma} \le \alpha) = \gamma.$$

Для того, чтобы получить выражение для α, нам необходимо знать распределение $\frac{|\varepsilon|}{\sigma}$.

Утверждение 3. *Пусть* $\varepsilon_i \sim N(0,\sigma_\varepsilon^2)$. *Пусть* $\sigma = \mathbb{E}|\varepsilon|$. *Тогда* $r^* = \frac{|\varepsilon|}{\sigma}$ *имеет полунормальное распределение* $N_h(\frac{\pi}{2})$, *среднее* $\mathbb{E}r^* = 1$, *дисперсия* $\mathbb{D}r^* = \frac{\pi}{2} - 1$.

Доказательство. Если $\varepsilon_i \sim N(0,\sigma_\varepsilon^2)$, то

$$\frac{\varepsilon}{\sigma} = \frac{\varepsilon}{\mathbb{E}|\varepsilon|} \sim N(0, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 \frac{2}{\pi}}) = N(0,\frac{\pi}{2}).$$

Тогда $\frac{|\varepsilon|}{\sigma} \sim N_h(\frac{\pi}{2})$ по определению полунормального распределения.

Посчитаем среднее и дисперсию r^*.

$$r^* = \frac{|\varepsilon|}{\sigma} = \frac{|\varepsilon|}{\mathbb{E}|\varepsilon|} = \frac{|\varepsilon|}{\sigma_\varepsilon\sqrt{\frac{2}{\pi}}}.$$

Тогда

$$\mathbb{E}r^* = \frac{\mathbb{E}|\varepsilon|}{\sigma_\varepsilon\sqrt{\frac{2}{\pi}}} = 1, \qquad \mathbb{D}r^* = \frac{\mathbb{D}|\varepsilon|}{\sigma_\varepsilon^2\frac{2}{\pi}} = \frac{\sigma_\varepsilon^2(1-\frac{2}{\pi})}{\sigma_\varepsilon^2\frac{2}{\pi}} = \frac{\pi}{2} - 1.$$

□

Из утверждения 3 следует, что уравнение $\mathrm{P}(\frac{|\varepsilon|}{\sigma} \leq \alpha) = \gamma$ перепишется в виде $F_h(\alpha) = \gamma$, где F_h — функция распределения $N_h(\frac{\pi}{2})$.

Используя формулу (2.5) для функции распределения полунормального распределения, получаем уравнение

$$\mathrm{erf}(\frac{\alpha}{\sqrt{2}\frac{\pi}{2}}) = \gamma,$$

где erf — функция ошибок, которая имеется во многих пакетах в R. Отсюда получаем вероятностную формулу для параметра α

$$\alpha = \frac{\sqrt{2}\pi}{2}\mathrm{erf}^{-1}(\gamma). \tag{2.6}$$

К примеру, для $\gamma = 0.95$ получим $\alpha \approx 3.079$. Для $\gamma = 0.99$ получим $\alpha \approx 4.046$.

На основе утверждения 3 можно можно провести следующие рассуждения для временных рядов:

Замечание 3. *1. Пусть дисперсия шума постоянна, $x_i = s_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, сигнал s_i точно отделим от шума. Пусть $\sigma = \mathbb{E}|\varepsilon|$. Тогда $r^* = \frac{|\varepsilon|}{\sigma}$ имеет полунормальное распределение $N_h(\frac{\pi}{2})$ среднее $\mathbb{E}r^* = 1$, дисперсия $\mathbb{D}r^* = \frac{\pi}{2} - 1$.*
2. Пусть шум гетероскедастичный, $x_i = s_i + t_i\varepsilon_i$, $t_i > 0$, $\varepsilon_i \sim N(0,1)$, сигнал s_i точно отделим от шума. Пусть $\mathsf{R} = (r_1, \ldots, r_N)^{\mathrm{T}}$ — вектор из остатков, $\mathsf{R}_+ = (|r_1|, \ldots, |r_N|)^{\mathrm{T}}$ — вектор из модулей остатков, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)^{\mathrm{T}} = \mathbb{E}(\mathsf{R}_+)$. Тогда каждая компонента вектора $\boldsymbol{r}^* = (\frac{|r_1|}{\sigma_1}, \ldots, \frac{|r_N|}{\sigma_N})^{\mathrm{T}}$ имеет полунормальное распределение: $r_i^* = \frac{|r_i|}{\sigma_i} \sim N_H(\frac{\pi}{2})$. В таком случае $\mathbb{E}r_i^* = 1$, $\mathbb{D}r_i^* = \frac{\pi}{2} - 1$.

Другими словами, получаем, что и в случае шума постоянной дисперсии, и в случае гетероскедастичного шума, в предположении нормальности шума r^* имеет полунормальное распределение $N_H(\frac{\pi}{2})$.

2.4.4. Модификация метода с итеративным обновлением весов

Далее представлена модификация алгоритма, подходящая для рядов с гетероскедастичным шумом.

Алгоритм 4: Модификация метода с итеративным обновлением весов для нахождения проекции на множество матриц ранга, не превосходящего r

Входные данные: $\mathbf{Y} \in \mathbb{R}^{L \times K}$ — траекторная матрица ряда, r — ранг сигнала; параметры критерия остановки: $\varepsilon = 10^{-4}$, максимальное число итераций $N_{IRLS} = 10$

Выходные данные: $\widehat{\mathbf{Y}} = \mathbf{U}\mathbf{V}^{\mathrm{T}}$ — проекция траекторной матрицы на множество матриц ранга, не превосходящего r

1. Инициализация $\mathbf{U} \in \mathbb{R}^{n \times r}$ и $\mathbf{V} \in \mathbb{R}^{p \times r}$ (например, с помощью сингулярного разложения матрицы $\mathbf{Y}$);
2. Выбор параметра α;
3. $t := 0$;
4. **повторять**
   a. Вычисление матрицы остатков $\mathbf{R} = \{r_{ij}\}_{i,j=1}^{n,p} = \mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}$;
   b. Ганкелизация матрицы $\mathbf{R}$ и получение ряда длины N из остатков: $\mathsf{R} = \mathcal{T}^{-1}\Pi_{\mathcal{H}}(\mathbf{R}) = (r_1, \ldots, r_N)^{\mathrm{T}}$;
   c. Пусть $\mathsf{R}_+ = (|r_1|, \ldots, |r_N|)^{\mathrm{T}}$ — вектор из модулей остатков. Вычисление $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)^{\mathrm{T}}$ как оценки мат. ожидания $\mathbb{E}(\mathsf{R}_+)$ некоторым методом: локальной регрессией loess, скользящей медианой или взвешенной локальной регрессией lowess (подробнее оценка мат. ожидания обсуждалась в пункте 2.4.2);
   d. Вычисление ряда $|\boldsymbol{\sigma}^{-1}\mathsf{R}| = (\frac{|r_1|}{\sigma_1}, \ldots, \frac{|r_N|}{\sigma_N})^{\mathrm{T}}$ и получение матрицы $\mathbf{R}^* = \{r_{ij}^*\}_{i,j=1}^{L,K} = \mathcal{T}(|\boldsymbol{\sigma}^{-1}\mathsf{R}|)$;
   e. Вычисление матрицы весов $\mathbf{W} = \{w_{ij}\}_{i,j=1}^{L,K} = \{w(r_{ij}^*)\}_{i,j=1}^{L,K}$, используя

$$w(x) = \begin{cases} (1 - (\frac{|x|}{\alpha})^2)^2, & |x| \le \alpha \\ 0, & |x| > \alpha \end{cases}$$

   f. Решение задачи взвешенной аппроксимации (обновление матриц $\mathbf{U}$ и $\mathbf{V}$)

$$\left\|\mathbf{W}^{1/2} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}})\right\|_F^2 \longrightarrow \min_{\mathbf{U},\mathbf{V}}$$

   g. $t := t + 1$.

до тех пор, пока $\left\|\mathbf{W}^{1/2} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}})\right\|_F^2 > \varepsilon$ *и* $t < N_{IRLS}$;

Посмотрим на график весов, чтобы убедиться, что только точки, содержащие выбросы, получили маленькие веса. На рисунках 2.2 и 2.3 изображен график ряда с 5% выбросов и веса, получившиеся в результате применения модификации метода. На графике весов видно, что точки, в которых содержались выделяющиеся наблюдения, получили нулевые веса. В остальных точках веса колеблются от 0.8 до 1.

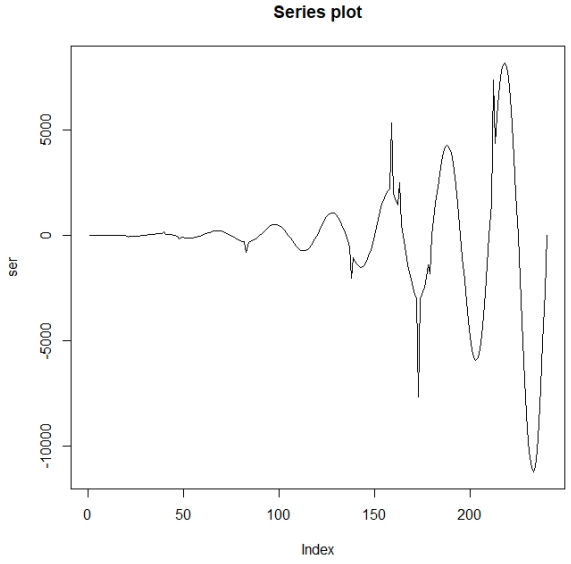

Рис. 2.2. График ряда с 5% выбросов.

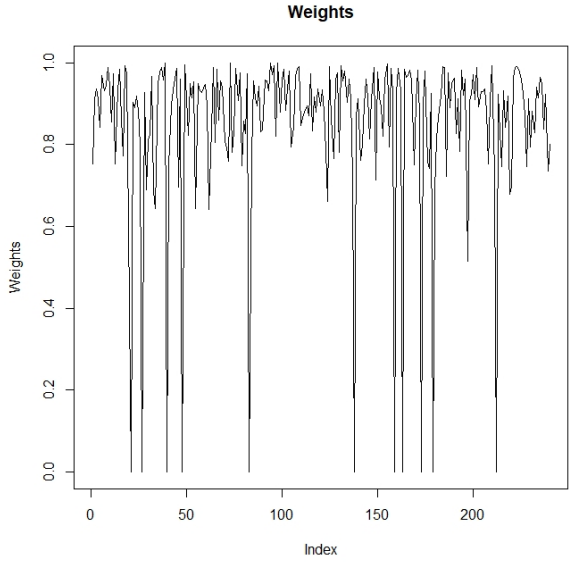

Рис. 2.3. Веса.

2.4.5. Модификация метода с итеративным обновлением весов с использованием метода Гаусса-Ньютона

Заметим, что использование ганкелевой матрицы $\{\sigma_{ij}\}_{i,j=1}^{L,K}$ и ее замена на параметр $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)^{\mathrm{T}}$ соответствует приписыванию весов элементам ряда. Однако в рассмотренной модификации после нормировки модуля остатков на параметр $\boldsymbol{\sigma}$ и вычисления матрицы весов, мы приписываем веса элементам траекторной матрицы ряда и вычисляем проекцию на множество матриц ранга, не превосходящего r. Возникает идея приписывать веса самому ряду, а не матрице, и находить оценку сигнала с этими весами.

Для нахождения проекции на множество рядов ранга, не превосходящего r, по взвешенной норме можно использовать модифицированный метод Гаусса-Ньютона, опи-

санный в статье [8], который итеративно решает задачу

$$\mathsf{X}^* = \operatorname*{argmin}_{\mathsf{X} \in \bar{\mathcal{D}}_r} \|\mathsf{Y} - \mathsf{X}\|_{\mathbf{W}},$$

где Y — ряд длины N, $\bar{\mathcal{D}}_r$ — замыкание множества рядов ранга, не превосходящего r, $\mathbf{W} \in \mathbb{R}^{N \times N}$ — матрица весов, $\|\mathsf{Y}\|_{\mathbf{W}} = \mathsf{Y}^{\mathrm{T}}\mathbf{W}\mathsf{Y}$.

Ниже представлен алгоритм модификации метода с итеративным обновлением весов с использованием метода Гаусса-Ньютона.

Алгоритм 5: Модификация метода с итеративным обновлением весов с использованием метода Гаусса-Ньютона

Входные данные: Y — ряд, $\mathbf{Y} \in \mathbb{R}^{L \times K}$ — траекторная матрица ряда Y, r — ранг сигнала; параметры критерия остановки: $\varepsilon = 10^{-4}$, максимальное число итераций $N_{IRLS} = 10$

Выходные данные: $\widehat{\mathsf{Y}}$ — проекция ряда Y на множество рядов ранга, не превосходящего r

1. Инициализация $\mathbf{U} \in \mathbb{R}^{n \times r}$ и $\mathbf{V} \in \mathbb{R}^{p \times r}$ (например, с помощью сингулярного разложения матрицы $\mathbf{Y}$);
2. Выбор параметра α;
3. Вычисление матрицы остатков $\mathbf{R} = \{r_{ij}\}_{i,j=1}^{n,p} = \mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}$;
4. $t := 0$;
5. **повторять**
   a. Ганкелизация матрицы $\mathbf{R}$ и получение ряда длины N из остатков: $\mathsf{R} = \mathcal{T}^{-1}\Pi_{\mathcal{H}}(\mathbf{R}) = (r_1, \ldots, r_N)^{\mathrm{T}}$;
   b. Пусть $\mathsf{R}_+ = (|r_1|, \ldots, |r_N|)^{\mathrm{T}}$ — вектор из модулей остатков. Вычисление $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)^{\mathrm{T}}$ как оценки мат. ожидания $\mathbb{E}(\mathsf{R}_+)$ некоторым выбранным методом (подробнее оценка мат. ожидания обсуждалась в пункте 2.4.2);
   c. Вычисление ряда $|\boldsymbol{\sigma}^{-1}\mathsf{R}| = (\frac{|r_1|}{\sigma_1}, \ldots, \frac{|r_N|}{\sigma_N})^{\mathrm{T}} = (r_1^*, \ldots, r_N^*)^{\mathrm{T}}$;
   d. Вычисление вектора весов $\mathsf{W} = (w_1, \ldots, w_N)^{\mathrm{T}} = (w(r_1^*), \ldots, w(r_N^*))^{\mathrm{T}}$, используя
   $$w(x) = \begin{cases} (1 - (\frac{|x|}{\alpha})^2)^2, & |x| \le \alpha \\ 0, & |x| > \alpha \end{cases}$$
   e. Решение задачи взвешенной аппроксимации с помощью модифицированного метода Гаусса-Ньютона с матрицей весов $\mathbf{W} = \mathrm{diag}(\mathsf{W}) \in \mathbb{R}^{N \times N}$:
   $$\mathsf{X}^* = \operatorname*{argmin}_{\mathsf{X} \in \bar{\mathcal{D}}_r} \|\mathsf{Y} - \mathsf{X}\|_{\mathbf{W}}.$$
   f. Пусть $\widehat{\mathbf{Y}}$ — траекторная матрица ряда X^*. Обновление матрицы остатков $\mathbf{R} = \{r_{ij}\}_{i,j=1}^{n,p} = \mathbf{Y} - \widehat{\mathbf{Y}}$ и ряда $\mathsf{Y} = \mathsf{X}^*$;
   g. $t := t + 1$.

до тех пор, пока $\left\|\mathbf{Y} - \widehat{\mathbf{Y}}\right\|_{\mathbf{W}} > \varepsilon$ *и* $t < N_{IRLS}$;

2.5. Оценка трудоемкости методов

Вопросу трудоемкости L1-SSA уделялось большое внимание во многих работах, посвященных построению $\mathbb{L}_1$-проекции на множество матриц ранга, не превосходящего r. В статье [15] приведен алгоритм, решающий точно эту задачу. Пусть имеется вещественная матрица $\mathbf{X} \in \mathbb{R}^{L \times K}$, необходимо вычислить проекцию на множество матриц ранга, не превосходящего r. Трудоемкость алгоритма, решающего задачу в явном виде, составляет $O(L^{\mathrm{rank}(\mathbf{X})r-r+1})$. В данной работе мы рассматривали методы, решающие эту задачу приближенно, но более эффективно.

Сравним теоретические трудоемкости описанных алгоритмов.

Последовательный метод

Вычислим трудоемкость последовательного алгоритма из раздела 2.3. Трудоемкость составляет $O((KP_1 + P_2)N_{iter})$, где P_1 и P_2 — трудоемкость решения задач линейного программирования, а N_{iter} — общее количество итераций для сходимости метода, которое также считаем не зависящим от L, K, r. Согласно [16], сложность вычисления задачи линейного программирования с v переменными и c ограничениями составляет $O(c \log v)$. В статье [9], содержащей описание алгоритма последовательного метода, вычислено количество переменных и ограничений в решаемых задачах, и получено, что трудоемкость может быть оценена как

$$\mathrm{T}_{\mathrm{l1pca}} = O(LK \log(2LK + Lr)N_{iter}), \tag{2.7}$$

Метод с итеративным обновлением весов

Теоретическая трудоемкость метода из раздела 2.4 составляет, согласно статье [7],

$$\mathrm{T}_{\mathrm{IRLS}} = O(LKr^2 N_\alpha N_{IRLS}), \tag{2.8}$$

где N_α и N_{IRLS} — общее количество итераций для решения задач (2.3), (2.4) и сходимости взвешенного метода наименьших квадратов с обновлением весов. Количество итераций мы брали постоянными и не зависящими от L, K и r. При подсчете трудоемкости авторы статьи [7] используют книгу [17], в которой приводятся эффективные алгоритмы QR-разложения матрицы.

Сравнение теоретических трудоемкостей

Сравним теоретические трудоемкости последовательного метода и взвешенного метода наименьших квадратов. Необходимо сравнить (2.7) и (2.8). Задача сводится к сравнению $\log(L(2K+r))N_{iter}$ и $r^2 N_\alpha N_{IRLS}$. Авторы статьи [7] утверждают, что максимальное число итераций N_α и N_{IRLS} для метода с обновлением весов достаточно взять 5 и 10 соответственно. Проведенные нами вычислительные эксперименты, описанные в разделе 3.5, показали, что максимальное число итераций для последовательного метода N_{iter} достаточно взять равным 5 (эксперименты проводились на рядах ранга 3 и 2 при $L = 120,\ K = 121$). Отличия в ошибках восстановления сигнала при увеличении числа итераций незначительны — ошибка уменьшается не более чем на 0.2%. Рассмотрим 2 случая.

Пусть L фиксировано, маленькое, а $K = N - L + 1 \sim N$. Это соответствует случаю, когда длина окна маленькая, и траекторная матрица вытянута. Трудоемкость метода с обновлением весов оказывается меньше.

Пусть траекторная матрица ряда близка к квадратной, то есть $L \sim N/2,\ K = N - L + 1 \sim N/2$. Тогда, если поводить сравнение, то получаем, что надо сравнить $\log(\frac{N}{2}(N+r))$ и r^2. Трудоемкость метода с обновлением весов снова оказывается меньше трудоемкости последовательного метода.

Для того, чтобы иметь возможность сравнить время работы методов, необходимо критерии остановки сделать такими, чтобы методы выдавали примерно одинаковые по точности результаты. Однако это оказывается нетривиальной задачей, поэтому поставим максимальное количество итераций для каждого метода такие, чтобы точность оказывалась приемлемой для каждого из методов. Сравним время работы, учитывая количество итераций.

В таблице 2.1 приведено время работы методов для 10 реализаций ряда

$$f_n = e^{4n/N}\sin(2\pi n/30) + Ae^{4n/N}\varepsilon_n,\ \varepsilon_n \sim N(0,1).$$

Ранг такого ряда равен 2, длина окна выбрана $L = 120$. Длину ряда будем постепенно увеличивать. Число выбросов положим равным 3 (это соответствует 1% выбросов при длине ряда $N = 240$), они будут находиться в случайных точках ряда.

При длине ряда $N > 240$ траекторная матрицы становится вытянутой, $K \sim N$, $L = 120$ — фиксировано. Исходя из формулы (2.8), время работы, разделенное на соот-

Таблица 2.1. Время работы программы и число итераций для последовательного метода и метода с итеративным обновлением весов (для $M = 10$ реализаций ряда) в зависимости от длины ряда N.

	$N = 240$	$N = 360$	$N = 480$	$N = 600$	$N = 720$	N_{iter}
IRLS	23 sec.	44 sec.	60 sec.	67 sec.	89 sec.	5*10
l1pca	54 sec.	136 sec.	403 sec.	721 sec.	1080 sec.	5

ветствующую длину ряда N, должно не зависеть от N для метода с обновлением весов. Для последовательного метода, пользуясь формулой (2.7), можно сделать вывод, что полученная величина должна логарифмически зависеть от N. Графики представлены на рисунках 2.4 и 2.5.

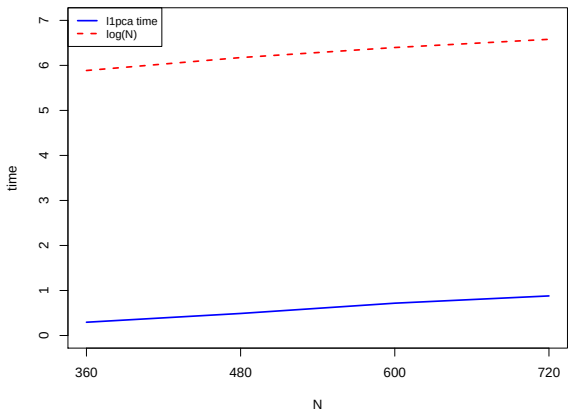

Рис. 2.4. Время работы метода l1pca

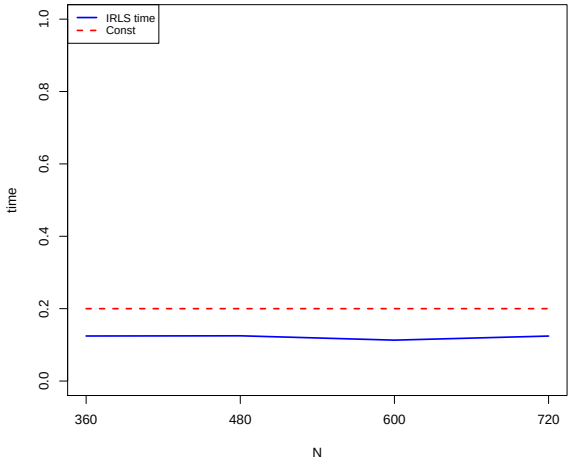

Рис. 2.5. Время работы метода IRLS

Глава 3

Вычислительные эксперименты

Сравним результаты работы описанных методов на нескольких примерах. Для начала возьмем ряд с экспоненциальным трендом и гауссовским шумом, который уже был рассмотрен в работе [6]. Затем рассмотрим ряд с растущей амплитудой и два случая: случай с гетероскедастичным шумом и с шумом с постоянной дисперсией, проведем сравнения для таких рядов. Код со сравнением различных методов, а также реализация метода с обновлением весов и его модификаций опубликованы на [18].

3.1. Модельный пример №1

Для начала рассмотрим пример из работы [6], но добавим большее количество выбросов (1% и 5%) в случайных точках ряда. Проверим, какой из приведенных алгоритмов окажется наиболее устойчивым.

Длина ряда $N = 240$. Рассмотрим временной ряд

$$f_n = e^{n/N} + \sin\left(2\pi n/120 + \pi/6\right) + \varepsilon_n,\ \varepsilon_n \sim N(0,1).$$

На рис. 3.1 изображен график ряда при 1% выбросов с величиной выброса $5f_i$. В случайно выбранных точках ряда f_i значение заменяется на $f_i + 5f_i$.

Сравнение будет проводиться по величине среднеквадратичной ошибки, согласованной с $\mathbb{L}_2$, которая вычисляется по формуле

$$\mathrm{MSE} = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}(s_i - \hat{s}_i)^2\right), \tag{3.1}$$

где $\mathsf{S} = (s_1, \ldots, s_N)^{\mathrm{T}}$ — сигнал, $\hat{\mathsf{S}} = (\hat{s}_1, \ldots, \hat{s}_N)^{\mathrm{T}}$ — его оценка. Будем вычислять

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}},$$

а также будем сравнивать методы по величине ошибки, согласованной с $\mathbb{L}_1$, которая имеет вид

$$\mathrm{MAD} = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}|s_i - \hat{s}_i|\right). \tag{3.2}$$

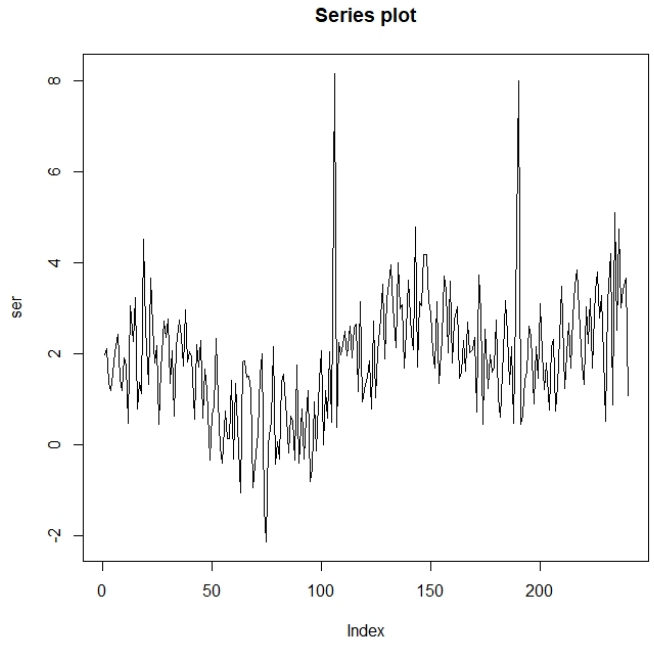

Рис. 3.1. График ряда при 1% выбросов с величиной выброса $5f_i$.

Возьмем количество реализаций ряда $M = 30$. Будем находить оценки математических ожиданий (3.1) и (3.2), а далее из оценки MSE будем извлекать корень, получая RMSE.

Ранг ряда равен 3. Во всех методах берется длина окна $L = 120$, равная половине длины ряда, и восстановление сигнала ведется по 3 компонентам.

Для метода с итеративным обновлением весов и его модификации выберем следующие параметры:

1. **Инициализация.** Возьмем в качестве начальных значений для обоих методов $\mathbf{U} = \mathbf{U}_r\mathbf{\Lambda}_r^{1/2}$, $\mathbf{V} = \mathbf{V}_r$, где $\mathbf{U}_r = [U_1 : \ldots : U_r]$, $\mathbf{V}_r = [V_1 : \ldots : V_r]$, $\mathbf{\Lambda}_r = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$ — первые r компонент сингулярного разложения траекторной матрицы. Для вычисления первых r компонент будем использовать один из вариантов truncated SVD из пакета svd [13].

2. **Параметр $\boldsymbol{\alpha}$.** Авторы статьи [7], содержащей описание исходного метода, предлагают выбрать $\alpha = 4.685$. Исходя из полученной формулы (2.6), связывающей параметр α и вероятность γ, получим для $\gamma = 0.99$ значение параметра $\alpha = 4.046$. Его и будем брать. Однако стоит заметить, что такая вероятностная интерпретация параметра α верна только для модификации метода.

Таблица 3.1. Оценки RMSE и MAD для различных методов для $M = 30$ реализаций ряда.

Method	0%	1%	5%
Оценки RMSE			
Basic SSA	**0.184**	0.256	0.653
l1pca	0.217	0.223	0.250
IRLS (orig.)	**0.184**	**0.189**	**0.206**
IRLS (loess)	0.196	0.199	**0.204**
IRLS (median)	0.210	0.221	0.223
IRLS (lowess)	0.206	0.207	**0.211**
IRLS (MGN)	0.398	0.222	0.798
Оценки MAD			
Basic SSA	**0.143**	0.197	0.505
l1pca	0.175	0.179	0.203
IRLS (orig.)	**0.145**	**0.147**	**0.161**
IRLS (loess)	0.155	0.158	**0.160**
IRLS (median)	0.170	0.178	0.178
IRLS (lowess)	0.165	0.166	**0.168**
IRLS (MGN)	0.259	0.178	0.601

В таблице 3.1 представлены результаты сравнения для четырех методов. Выброс добавлялся в случайных точках ряда заменой значения f_i на $f_i + 5f_i$.

При сравнении методов со стандартным использовался пакет Rssa [19]. Реализация модифицированного метода Гаусса-Ньютона находится в репозитории [20].

Первая строка таблиц соответствует стандартному методу SSA с большой длиной окна ($L = 120$). Вторая строка — метод l1pca из пакета pcaL1 [11] (соответствует последовательному методу из раздела 2.3). Третья строка соответствует стандартному методу с итеративным обновлением весов из раздела 2.4. Четвертая, пятая и шестая строки соответствуют модификации взвешенного метода наименьших квадратов с различными вариантами выделения тренда (локальная регрессия с параметром сглаживания 0.35, скользящая медиана с длиной окна 80 и взвешенная локальная регрессия с параметром сглаживания 0.35). Была использована реализация lowess в R из статьи [21]. Параметр сглаживания выбран 0.35, остальные параметры оставлены по умолчанию:

число итераций равно 3, параметр δ, требующийся для ускорения вычисления, равен $0.01(\max\limits_i f_i - \min\limits_i f_i)$. Седьмая строка посвящена модификации метода с обновлением весов с использованием метода Гаусса-Ньютона.

Из-за того, что дисперсия шума постоянная, можем предположить, что модификации IRLS с различными вариантами выделения тренда будут работать примерно одинаково. Поэтому подробнее об отличиях модификаций с разными вариантами выделения тренда пока что говорить не будем.

Жирным шрифтом в каждом столбце выделено лучшее значение и значение, которое незначимо отличается от лучшего. Проверка значимости сравнений представлена далее в таблицах 3.3 и 3.4.

Можно заметить, что при 5% выбросов модификация с использованием метода Гаусса-Ньютона оказывается неустойчивой. Можно предположить, что это происходит из-за того, что выбросы оказываются близко друг к другу, либо близко к началу или концу ряда. Исследуем, как ведут себя методы в зависимости от положения выброса. Результаты представлены в таблице 3.2.

Действительно, если выброс попадает близко к концу ряда (в точку x_{235}), то модификация с использованием метода Гаусса-Ньютона дает большую ошибку восстановления сигнала. Однако при наличии выброса в середине ряда ошибка восстановления сигнала с использованием этой модификации маленькая. Можно сделать вывод, что этот метод лучше использовать, если выбросов небольшое количество, либо при наличии информации, что выбросы не содержатся близко к началу или концу ряда. В наших исследованиях выбросы находятся в случайно выбранных точках ряда, чтобы максимально обобщить все варианты местоположения выделяющихся наблюдений. Поэтому в дальнейших сравнениях не будем рассматривать эту модификацию.

При отсутствии выбросов наиболее точным все так же остается классический метод SSA, а также метод с итеративным обновлением весов. В присутствии выделяющихся наблюдений наиболее устойчивым являются метод обновлением весов и его модификация с использованием локальной регрессии.

Проверка значимости сравнения

Опишем подробнее, как происходит сравнение метода, выдающего наименьшую ошибку, с остальными методами. Проверим значимость сравнения по критерию для

Таблица 3.2. RMSE в зависимости от положения выброса.

Method	x_{50}	x_{100}	x_{150}	x_{200}	x_{235}
Basic SSA	0.106	0.239	0.137	0.170	0.259
l1pca	0.213	0.251	0.195	0.165	0.278
IRLS (orig.)	0.111	0.201	0.121	0.163	0.222
IRLS (loess)	0.135	0.217	0.163	0.178	0.243
IRLS (median)	0.223	0.255	0.227	0.179	0.283
IRLS (lowess)	0.168	0.230	0.188	0.187	0.255
IRLS (MGN)	0.092	0.220	0.134	0.156	0.655

Таблица 3.3. P-value для сравнения различных методов с наилучшим без выбросов.

0%	l1pca	IRLS (orig.)	IRLS (loess)	IRLS (median)	IRLS (lowess)
Basic SSA	4.7e-5	**0.08**	0.022	0.001	0.009

зависимых выборок. Проверим гипотезу, что MSE для некоторых методов равны между собой.

$\mathrm{H}_0 : \mathbb{E}(\xi_1 - \xi_2) = 0$. Имеем две выборки $X = (x_1, \ldots, x_M)$ и $Y = (y_1, \ldots, y_M)$ объема M. Обозначим $\bar{X}$ и $\bar{Y}$ — их выборочные средние, s_x^2 и s_y^2 — выборочные дисперсии, $\hat{\rho}$ — коэффициент корреляции. Статистика критерия

$$t = \frac{\sqrt{M}(\bar{X} - \bar{Y})}{\sqrt{s_x^2 + s_y^2 - 2 s_x s_y \hat{\rho}}}$$

имеет асимптотически нормальное распределение. Критерий является двухсторонним.

Проверим, является ли отличие между этими методами значимым. В таблице 3.3 приведены p-value для сравнения среднеквадратичных ошибок для стандартного SSA и остальных методов без выделяющихся наблюдений. Все сравнения оказываются значимыми при уровне значимости 0.05. В таблице 3.4 приведены p-value для сравнения ошибок для метода взвешенных наименьших квадратов и остальных методов при 5% выбросов. При уровне значимости 0.05 при без выбросов сравнение стандартного SSA с оригинальным методом с обновлением весов оказывается незначимым. При 5% выбросов сравнения модификации IRLS (loess) с оригинальным IRLS и IRLS (lowess) оказываются незначимыми.

Таблица 3.4. P-value для сравнения различных методов с наилучшим в присутствии выбросов.

5%	Basic SSA	l1pca	IRLS (orig.)	IRLS (median)	IRLS (lowess)
IRLS (loess)	3.1e-13	1.1e-5	**0.589**	0.019	**0.185**

3.2. Модельный пример №2

Попробуем рассмотреть не похожий на предыдущий пример ряд, добавив растущую амплитуду и шум непостоянной дисперсии, и исследуем устойчивость методов.

Длина ряда $N = 240$. Рассмотрим ряд с гетероскедастичным шумом

$$f_n = e^{4n/N} \sin(2\pi n/30) + Ae^{4n/N}\varepsilon_n,\ \varepsilon_n \sim N(0,1).$$

График ряда представлен на рисунке 3.2. Ранг ряда равен 2.

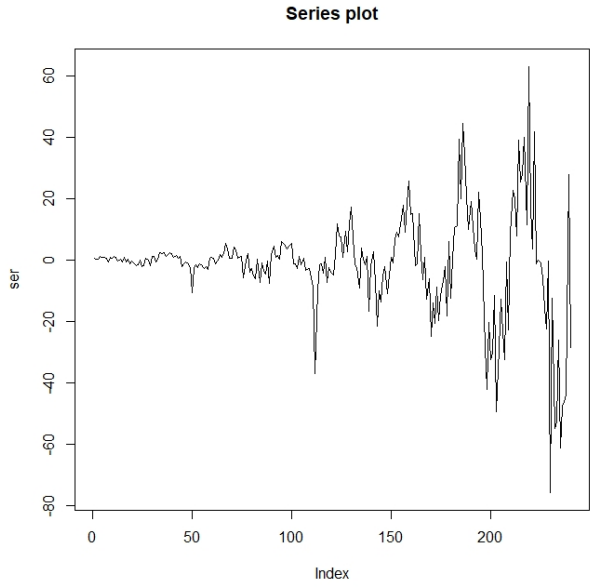

Рис. 3.2. График ряда при 1% выбросов с величиной выброса $5f_i$, $A = 0.5$.

Из-за наличия шума непостоянной дисперсии можно предположить, что модификация метода с итеративным обновлением весов даст хороший результат. Тогда необходимо выбрать наиболее подходящий метод оценки математического ожидания ряда из модулей остатков для задания параметра $\boldsymbol{\sigma}$. На рисунке 3.3 представлено выделение тренда из модуля остатков различными способами: локальной регрессией с параметром сглаживания 0.35, скользящей медианой с длиной окна 80 и взвешенной локальной ре-

грессией с параметром сглаживания 0.35. Вычислим реальный тренд из ряда из модулей остатков:

$$\mathbb{E}|\mathsf{R}| = Ae^{4n/N}\mathbb{E}|\varepsilon| = Ae^{4n/N}\sqrt{\frac{2}{\pi}}.$$

Можно сделать следующие выводы. Прежде всего, локальная регрессия сильнее реагирует на выбросы, тренд чуть завышен. Скользящая медиана плохо работает на конце ряда, даже если брать длину окна в скользящей медиане большой. Взвешенная локальная регрессия хорошо справляется с выбросами и показывает результат, близкий к реальному тренду.

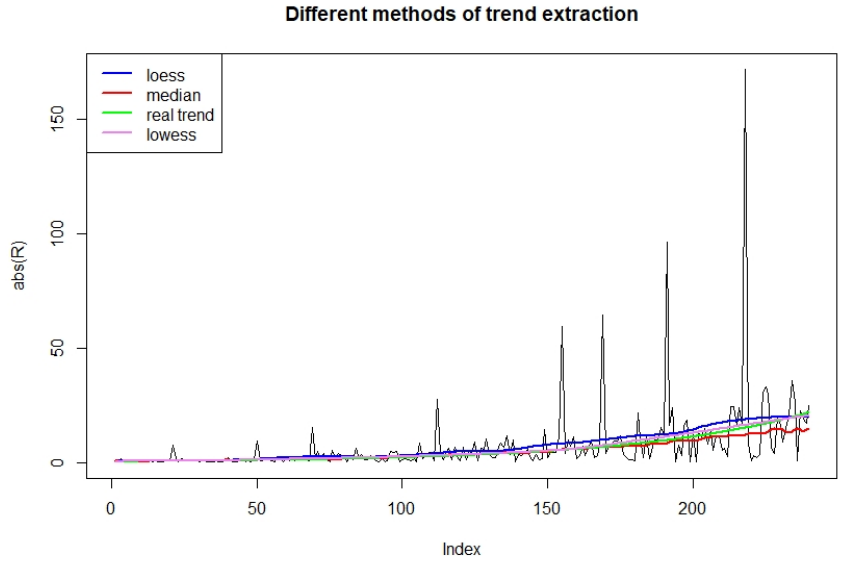

Рис. 3.3. Выделение тренда из ряда $|\mathsf{R}|$ несколькими способами.

Если сравнить точки, получившие нулевые веса различными методами, можно сделать следующие выводы: наибольшие выбросы все методы идентифицировали одинаково хорошо, скользящая медиана обнуляет наибольшее количество точек, нулевые веса получили также точки, не являющиеся выбросами. Loess и lowess присваивают нулевые веса одинаковым точки и не обнуляют ничего лишнего. Есть предположение, что с использованием этих двух методов ошибка восстановления сигнала будет наименьшая. Отличие в этих методах состоит в том, что выделенный с помощью lowess тренд лежит ниже, чем тренд, полученный с помощью loess. Поэтому lowess чуть сильнее занижает веса в точках, не являющихся выбросами. Возможно, из-за этого loess окажется лучше.

Результаты сравнения методов представлены в таблице 3.5. Сразу можно отметить, что для ряда с гетероскедастичным шумом метод с обновлением весов дает большую

Таблица 3.5. Оценки RMSE для различных методов для $M = 30$ реализаций ряда.

Method	0%	1%	5%
Оценки RMSE			
Basic SSA	**2.16**	2.78	5.96
l1pca	**2.45**	**2.47**	2.87
IRLS (orig.)	3.52	3.59	3.61
IRLS (loess)	**2.31**	**2.36**	**2.39**
IRLS (median)	2.84	2.84	2.86
IRLS (lowess)	2.59	2.60	2.63
Оценки MAD			
Basic SSA	**1.08**	1.33	2.91
l1pca	**1.18**	**1.26**	1.36
IRLS (orig.)	1.63	1.65	1.66
IRLS (loess)	**1.16**	**1.19**	**1.21**
IRLS (median)	1.37	1.38	1.38
IRLS (lowess)	1.26	1.28	1.29

Таблица 3.6. P-value для сравнения различных методов с наилучшим без выбросов.

0%	l1pca	IRLS (orig.)	IRLS (loess)	IRLS (median)	IRLS (lowess)
Basic SSA	**0.57**	6.7e-9	**0.69**	0.001	0.021

ошибку даже в отсутствии выбросов. Без выбросов модификация метода IRLS с выделением тренда с помощью локальной регрессии, а также последовательный метод дают маленькую ошибку, как и стандартный SSA. В присутствии выбросов (1%) оба эти метода показывают хорошие результаты, а при 5% выбросов наилучшим оказывается метод с обновлением весов с использованием локальной регрессии.

В таблицах 3.6 и 3.7 представлены p-value для проверки значимости сравнения наилучших методов с остальными при 0% и 5% выбросов.

Таблица 3.7. P-value для сравнения различных методов с наилучшим в присутствии выбросов.

5%	Basic SSA	l1pca	IRLS (orig.)	IRLS (median)	IRLS (lowess)
IRLS (loess)	1.7e-5	0.020	8.7e-4	0.018	0.026

3.3. Модельный пример №3

Возьмем похожий ряд, но с шумом, имеющим постоянную дисперсию. Рассмотрим пример, предложенный в статье [22], и проведем для этого примера вычислительный эксперимент.

Пусть длина ряда $N = 240$. Рассмотрим временной ряд

$$f_n = ne^{4n/N} \sin(2\pi n/30) + \varepsilon_n,\ \varepsilon_n \sim N(0,1).$$

Ранг ряда равен 4. У такого ряда разброс собственных значений очень велик. Это может приводить к тому, что некоторые компоненты сигнала могут смешиваться с шумом. Однако шум рассматриваемого размера не портит отделимость сигнала от шума. Выбросы будут находиться в случайно выбранных точках ряда. Сравнение будем проводить при 1% и 5% выбросов, а также без выделяющихся наблюдений. В случайно выбранных точках f_i значение будет заменяться на $f_i + 1.5f_i$.

На рис. 3.4 изображен график ряда при 1% выбросов с величиной выброса $1.5f_i$.

Результаты сравнения методов при различном проценте выделяющихся наблюдений представлены в таблице 3.8.

Без выделяющихся наблюдений оригинальный IRLS и его модификация с использованием loess незначимо отличаются от стандартного SSA. В присутствии выделяющихся наблюдений только модификация IRLS с использованием взвешенной локальной регрессии работает хорошо.

Проверка значимости сравнения наилучшего метода с остальными при отсутствии выбросов представлена в таблице 3.9.

3.4. Выводы

Результаты исследования для трех рассмотренных примеров представлены в таблице 3.10. На основе проведенного исследования можно сделать следующие выводы.

Таблица 3.8. Оценки RMSE и MAD для различных методов для $M = 30$ реализаций ряда.

Method	0%	1%	5%
Оценки RMSE			
Basic SSA	**0.215**	123.7	459.6
l1pca	0.256	8.65	21.11
IRLS (orig.)	**0.216**	220.4	398.2
IRLS (loess)	**0.227**	115.25	303.2
IRLS (median)	0.256	20.21	38.21
IRLS (lowess)	0.243	**0.260**	**0.301**
Оценки MAD			
Basic SSA	**0.164**	33.32	168.2
l1pca	0.197	1.117	3.145
IRLS (orig.)	**0.165**	27.311	88.54
IRLS (loess)	**0.174**	22.18	31.180
IRLS (median)	0.187	8.182	12.20
IRLS (lowess)	0.180	**0.179**	**0.187**

Таблица 3.9. P-value для сравнения различных методов с наилучшим без выбросов.

0%	l1pca	IRLS (orig.)	IRLS (loess)	IRLS (median)	IRLS (lowess)
Basic SSA	4.1e-4	**0.962**	**0.107**	0.004	0.024

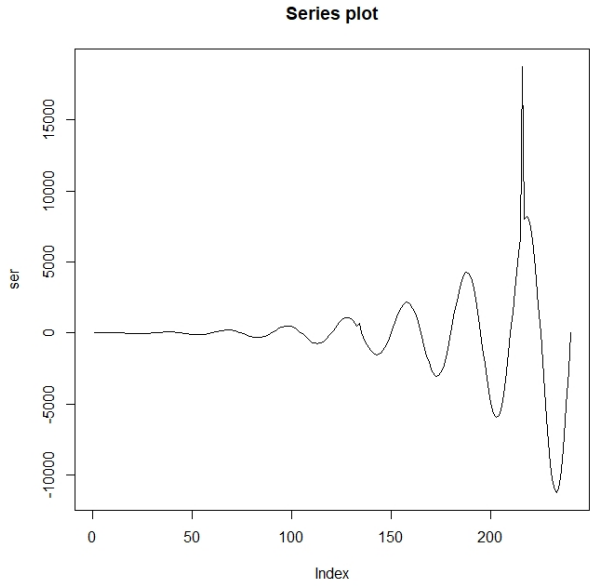

Рис. 3.4. График ряда при 1% выбросов с величиной выброса $1.5f_i$.

Для первого примера без растущей амплитуды ряда в случае гауссовского шума наиболее точным методом без выбросов оказывается стандартный SSA. Стандартный метод с обновлением весов также является точной при отсутствии выбросов. Самыми устойчивыми к выбросам являются стандартный IRLS, а также его модификации IRLS (loess) и IRLS (lowess).

В случае шума с непостоянной дисперсией преимущество стандартного метода с итеративным обновлением весов пропадает, однако его модификация с использованием локальной регрессии при выделении тренда из остатков оказывается устойчивой к выбросам и точной без выделяющихся наблюдений.

Если же разброс значений ряда очень велик, то при отсутствии выбросов точными являются стандартный SSA, оригинальный IRLS и модификация IRLS (loess). Однако устойчивой является IRLS (lowess).

Исходя из проведенного исследования можно сказать, что если у ряда нет растущей амплитуды и разброс значений небольшой, то можно использовать метод с обновлением весов. Он достаточно точный без выбросов и устойчивый к выделяющимся наблюдениям. В случае появления растущей амплитуды ряда и шума с непостоянной дисперсией, преимущество метода с обновлением весов пропадает. В таком случае следует использовать его модификацию с использованием локальной регрессии. Если же разброс значений у ряда большой,то следует использовать модификацию с выделени-

Таблица 3.10. Оценки RMSE и MAD для трех рассмотренных примеров для $M = 30$ реализаций ряда.

Оценки RMSE						
Method	0%	5%	0%	5%	0%	5%
Basic SSA	**0.184**	0.653	**2.16**	5.96	**0.215**	459.6
l1pca	0.217	0.250	**2.45**	2.87	0.256	21.11
IRLS (orig.)	**0.184**	**0.206**	3.52	3.61	**0.216**	398.2
IRLS (loess)	0.196	**0.204**	**2.31**	**2.39**	**0.227**	303.2
IRLS (median)	0.210	0.223	2.84	2.86	0.256	38.21
IRLS (lowess)	0.206	**0.211**	2.59	2.63	0.243	**0.301**
Оценки MAD						
Method	0%	5%	0%	5%	0%	5%
Basic SSA	**0.143**	0.505	**1.08**	2.91	**0.164**	168.2
l1pca	0.175	0.203	**1.18**	1.36	0.197	3.145
IRLS (orig.)	**0.145**	**0.161**	1.63	1.66	**0.165**	88.54
IRLS (loess)	0.155	**0.160**	**1.16**	**1.21**	**0.174**	31.180
IRLS (median)	0.170	0.178	1.37	1.38	0.187	12.20
IRLS (lowess)	0.165	**0.188**	1.26	1.29	0.180	**0.187**

ем тренда с помощью взвешенной локальной регрессии, которая хорошо справляется с выбросами.

3.5. Исследование числа итераций

В последовательном методе l1pca и методе IRLS есть дополнительный параметр: максимальное число итераций в цикле. В методе с итеративным обновлением весов задается максимальное число итераций для внешнего цикла и для внутреннего. Исследуем, какое количество итераций требуется для сходимости этих методов. На рисунках 3.5 и 3.6 показана зависимость ошибки RMSE от числа итераций для двух примеров (первый пример с экспоненциальным трендом, второй пример с быстрорастущей амплитудой ряда) при 5% выбросов в случайных точках. Длина ряда в обоих случаях полагалась равной $N = 240$. На рисунках 3.7 и 3.8 представлена зависимость RMSE от числа ите-

раций во внешнем цикле для метода IRLS. Сравнения для различного числа итераций проводились на одинаковых реализациях ряда.

Из графиков видно, что отличия в ошибках при увеличении итераций совсем незначительные, и в последовательном методе можно было бы ограничиться и 5 итерациями.

Для метода IRLS по графикам видно, что ошибка практически перестает убывать при 5 итерациях внешнего цикла для первого примера и 20 итерациях для второго. Число итераций во внутреннем цикле можно оставить по умолчанию равным 5, так как отличия в таком случае совсем незначительные.

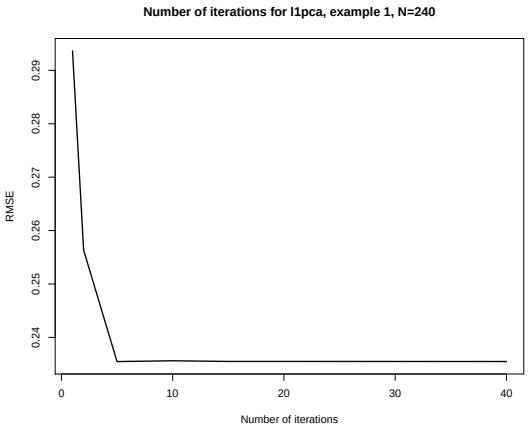

Рис. 3.5. Пример 1: Зависимость RMSE от числа итераций для последовательного метода l1pca.

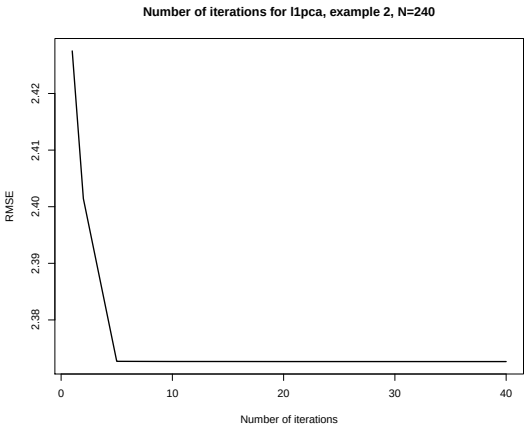

Рис. 3.6. Пример 2: Зависимость RMSE от числа итераций для последовательного метода l1pca.

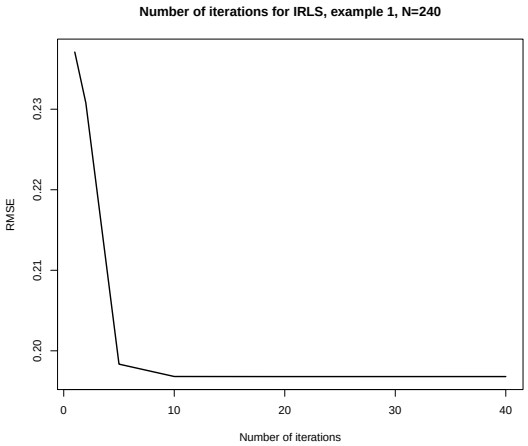

Рис. 3.7. Пример 1: Зависимость RMSE от числа итераций для метода IRLS.

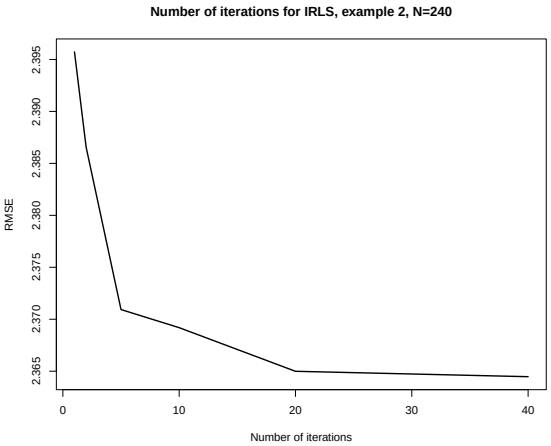

Рис. 3.8. Пример 2: Зависимость RMSE от числа итераций для метода IRLS.

Попробуем взять меньшую длину окна и проверим, сколько итераций понадобится

в таком случае. На рисунке 3.9 изображена зависимость RMSE от числа итераций во внешнем цикле для метода IRLS при выбранной длине окна $L = 84$. Требуется около 15 итераций для сходимости. На рисунке 3.10 изображена аналогичная зависимость для длины окна $L = 60$. Видно, что необходимо уже порядка 40 итераций.

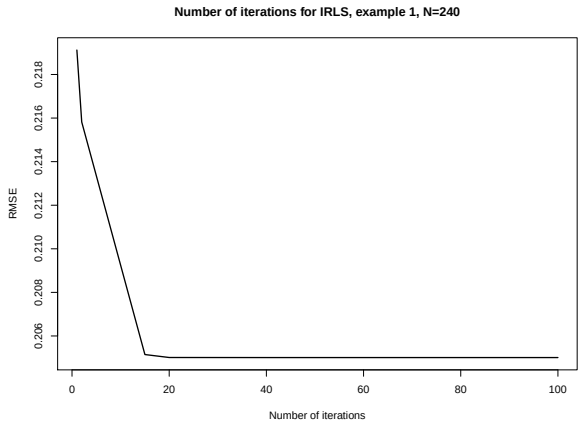

Рис. 3.9. Пример 1: Зависимость RMSE от числа итераций для метода IRLS, $L = 84$.

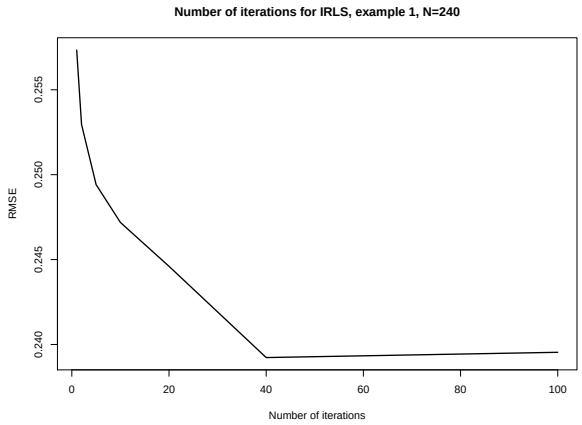

Рис. 3.10. Пример 1: Зависимость RMSE от числа итераций для метода IRLS, $L = 60$.

Проанализируем, почему при меньшей длине окна требуется большее количество итераций. Напомним, что изначально мы инициализируем матрицы $\mathbf{U}$ и $\mathbf{V}$ с помощью первых r компонент сингулярного разложения. Сравним разделимость сигнала от шума в присутствии выбросов и без выбросов. В таблице 3.11 представлены ошибки выделения сигнала по первым 3 компонентам для ряда без выбросов и для ряда с выбросами для разных длин окна. Ошибки для выделения сигнала без выброса и с выбросом считаются на одних и тех же реализациях ряда (шум одинаковый, во втором случае добавлены выбросы). В последнем столбце представлена ошибка между восстановлением сигнала с выбросом и без выброса. Видим, что для меньшей длины окна отделимость от шума оказывается хуже.

Можем сделать вывод, что нельзя говорить, что достаточно определенного числа итераций для сходимости. Число итераций зависит от того, насколько хорошо сигнал отделяется от шума в присутствии выделяющихся наблюдений.

Таблица 3.11. Ошибка RMSE выделения сигнала для ряда без выбросов и для ряда с 5% выбросов (для $M = 10$ реализаций ряда).

L	Without outliers	With outliers	Difference
120	0.146	0.678	0.673
84	0.159	0.684	0.677
60	0.190	0.785	0.779

3.6. Реальный ряд

Продемонстрируем работу методов на реальном примере. Рассмотрим ряд — импорт товаров в США из Кувейта [23] с ноября 1993 года по ноябрь 2012 года. Имеются данные за каждый месяц. Длина ряда $N = 229$. Увеличим размер выбросов в двух точках: x_{83} и x_{222}.

Возьмем длину окна $L = 60$, посмотрим на матрицу взвешенных корреляций (3.11) между восстановленными компонентами ряда. Матрица w-корреляций является одним из средств идентификации компонент, относящихся к сигналу. Подробнее матрица взвешенных корреляций описана в пособии [2]. Равенство взвешенной корреляции нулю является необходимым условием разделимости компонент ряда.

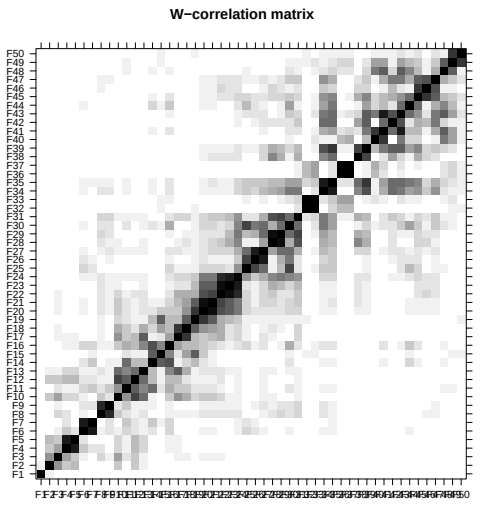

Рис. 3.11. Матрица взвешенных корреляций, $L = 60$.

Посмотрим также на элементарные восстановленные ряды (3.12). Будем восстанав-

ливать сигнал по первым 5 компонентам.

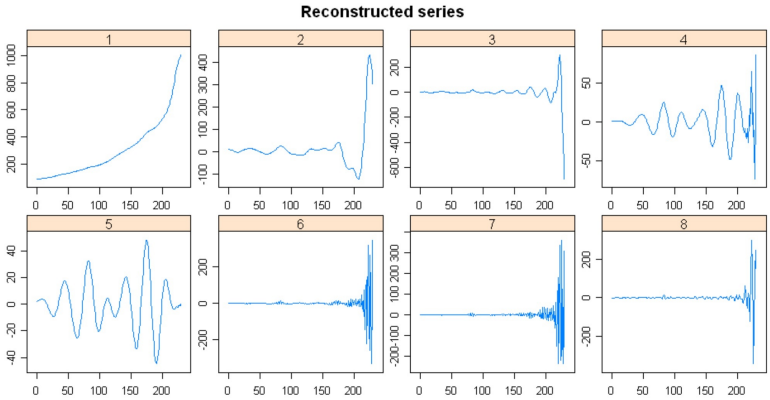

Рис. 3.12. Элементарные восстановленные ряды.

Так как настоящий сигнал нам неизвестен, то попробуем на месте выбросов поставить пропуски, заполнить пропущенные значения с помощью метода gapfill [24], который имеется в пакете Rssa [19], а затем выделить сигнал с помощью классического SSA. Полученный сигнал будем считать истинным. Сравним стандартный SSA и метод с обновлением весов из статьи [7] с предложенными модификациями с выделением тренда с помощью локальной регрессии loess и взвешенной локальной регрессии lowess. Результат восстановления сигнала различными методами представлен на рисунке 3.13.

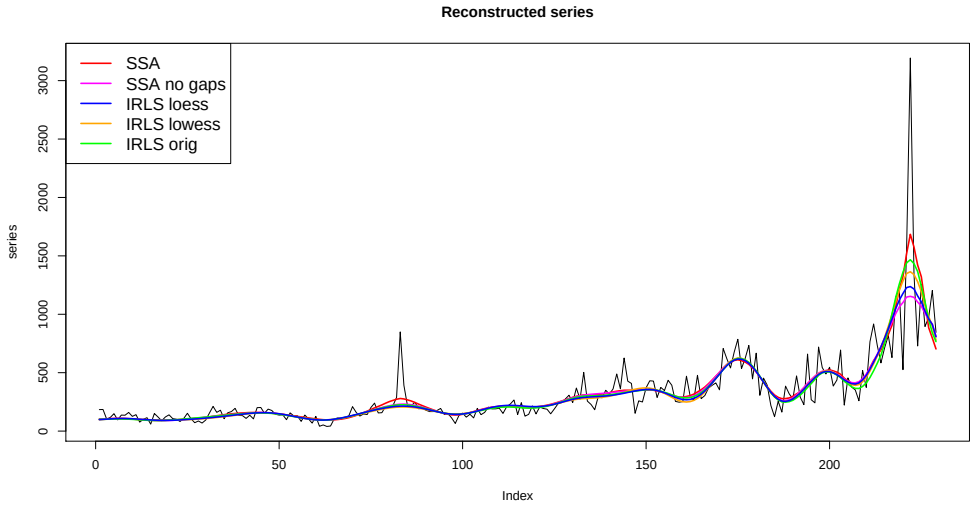

Рис. 3.13. Результат восстановления сигнала различными методами.

Можно заметить, что стандартный метод с обновлением весов плохо справляется с выбросом на конце ряда, где дисперсия шума увеличивается. Наиболее близким к "истинному" сигналу оказывается модификация метода с использованием loess.

Заключение

В работе были приведены и исследованы некоторые варианты модификации метода SSA с целью повышения устойчивости к выбросам.

Был проведен обзор двух известных подходов к построению устойчивых к выбросам вариантов SSA: замена проекторов по норме в $\mathbb{L}_2$ на проекторы по норме в $\mathbb{L}_1$ и взвешенной норме в $\mathbb{L}_2$. Для построения $\mathbb{L}_1$-проектора на множество матриц ранга, не превосходящего r, был использован последовательный метод, который уже реализован в R-пакете [11]. Второй подход соответствует методу с итеративным обновлением весов из статьи [7], где точкам, содержащим выбросы, присваивается меньший вес. Устойчивые модификации SSA были систематизированы и изложены в едином стиле.

Однако, метод из статьи [7] оказался неподходящим для рядов с шумом непостоянной дисперсии. Была предложена модификация этого метода, которая расширяет его применимость на случай нестационарного шума. В модификации метода предполагается замена параметра σ_{ij}, который в оригинальном методе полагается равным константе, на элементы траекторной матрицы тренда ряда из модулей остатков. Были рассмотрены несколько вариантов выделения тренда: скользящая медиана, локальная регрессия loess, а также взвешенная локальная регрессия lowess. Также была выведена формула для второго параметра метода с обновлением весов α, зависящая от вероятности γ (разделы 2.4.2 и 2.4.3). Благодаря полученной формуле стала понятна интерпретация этого параметра, а также даны рекомендации по его выбору.

Для рассмотренных методов было проведено их сравнение по трудоемкости. В трудоемкость входит число итераций, которое в статьях предполагается фиксированным. В работе показано, что предположение о достаточности фиксированного числа итераций не верно. Теоретически вывести достаточное число итераций не удалось. Однако, в предположении, что число итераций не растет с увеличением длины ряда, так как зависит от разделимости, которая только улучшается, удалось теоретически сравнить трудоемкости. Метод с итеративным обновлением весов оказался менее трудоемким.

Все рассматриваемые устойчивые модификации SSA были реализованы на R, опубликованы R-скрипты [18]. Последняя часть работы посвящена численным экспериментам. Сравнение методов проводилось на модельных примерах, сначала без выделяющихся наблюдений, а затем при 1% и 5% выбросов в случайных точках ряда. Сравнение

по точности проводилось по величине оценок ошибок RMSE и MAD. На основе проведенных экспериментов были выработаны некоторые рекомендации по применению методов, однако остается открытым вопрос, какой метод выделения тренда лучше использовать для выбора параметра σ_{ij}. Численные примеры подтвердили полученные формулы для порядка трудоемкости, а именно, было подтверждено, что трудоемкость используемой реализации метода L1-SSA имеет порядок $O(LK\log(2LK+Lr))$, а метод с итеративным обновлением весов — $O(LKr^2)$, где r — ранг сигнала, L — длина окна, $K=N-L+1$, N — длина временного ряда.

Таким образом, в работе были исследованы два варианта метода, которые являются более устойчивыми к выделяющимся наблюдениям, чем стандартный SSA, а также предложено несколько устойчивых модификаций, подходящих для рядов с гетероскедастичным шумом.

Список литературы

1. Golyandina N., Nekrutkin V., Zhigljavsky A. Analysis of Time Series Structure: SSA and Related Techniques. — Boca Raton, Fla. : Chapman & Hall/CRC, 2001.
2. Голяндина Н. Э. Метод «Гусеница»-SSA: анализ временных рядов: Учеб. пособие. — Санкт-Петербург : ВВМ, 2004.
3. Advanced spectral methods for climatic time series / M. Ghil, R. M. Allen, M. D. Dettinger et al. // Reviews of Geophysics. — 2002. — Vol. 40, no. 1. — P. 1–41.
4. Broomhead D., King G. Extracting qualitative dynamics from experimental data // Physica D. — 1986. — Vol. 20. — P. 217–236.
5. Vautard M., Ghil M. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series // Physica D. — 1989. — Vol. 35. — P. 395–424.
6. Третьякова А. Л. Устойчивые варианты метода SSA для анализа временных рядов: выпускная квалификационная работа, науч.рук. к.ф.-м.н., доцент Голяндина Н.Э. — 2018.
7. Chen K., Sacchi M. Robust reduced-rank filtering for erratic seismic noise attenuation // GEOPHYSICS. — 2015. — 01. — Vol. 80. — P. V1–V11.
8. Zvonarev N., Golyandina N. Image space projection for low-rank signal estimation: Modified gauss-newton method // arXiv preprint arXiv:1803.01419. — 2018.
9. Brooks J. P., Jot S. pcaL1: An implementation in R of three methods for L1-norm principal component analysis. — 2013.
10. Zvonarev N., Golyandina N. Iterative algorithms for weighted and unweighted finite-rank time-series approximations // Statistics and Its Interface. — 2017. — Vol. 10, no. 1. — P. 5–18.
11. Jot S., Brooks J. P., Visentin A., Park Y. W. — pcaL1: L1-Norm PCA Methods, 2017. — R package version 1.5.2.
12. Ke Q., Kanade T. Robust l1 norm factorization in the presence of outliers and missing data by alternative convex programming // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005). — 2005. — June.
13. Korobeynikov A., Larsen R. M., Lawrence Berkeley National Laboratory. — svd: Interfaces to Various State-of-Art SVD and Eigensolvers, 2017. — R package version 0.4.1. Access mode: `https://CRAN.R-project.org/package=svd`.

14. Korobeynikov A. Computation- and space-efficient implementation of SSA // Stat. Interface. — 2009. — 11. — Vol. 3.
15. Markopoulos P., Karystinos G., Pados D. Optimal algorithms for L1-subspace signal processing // IEEE Transactions on Signal Processing. — 2014. — 10. — Vol. 62. — P. 5046–5058.
16. Chvátal V. Linear Programming. Series of books in the mathematical sciences. — New York : W.H. Freeman Company, 1983. — ISBN: 9780716711957.
17. Golub G. H., Van Loan C. F. Matrix Computations. — Third edition. — The Johns Hopkins University Press, 1996.
18. Tretyakova A. Robust SSA. — 2020. — Jun. — Access mode: `https://doi.org/10.5281/zenodo.3871743`.
19. Korobeynikov A., Shlemov A., Usevich K., Golyandina N. — Rssa: A collection of methods for singular spectrum analysis, 2016. — R package version 1.0. Access mode: `http://CRAN.R-project.org/package=Rssa`.
20. Zvonarev N. R code for Modified Gauss-Newton algorithm. — `https://github.com/neg99/MGN`. — 2019.
21. Cleveland W. S. Robust locally weighted regression and smoothing scatterplots // Journal of the American Statistical Association. — 1979. — Vol. 74, no. 368. — P. 829–836.
22. Rodrigues P., Lourenco V., Mahmoudvand R. A robust approach to singular spectrum analysis // Quality and Reliability Engineering International. — 2018. — 06. — Vol. 34.
23. U.S. Bureau of Economic Analysis, U.S. Census Bureau. U.S. imports of goods by customs basis from Kuwait. — `https://fred.stlouisfed.org/series/IMP5130`. — 2020.
24. Golyandina N., Osipov E. The "Caterpillar"-SSA method for analysis of time series with missing values // Journal of Statistical Planning and Inference. — 2007. — Vol. 137, no. 8. — P. 2642–2653.