Робастные варианты метода анализа сингулярного спектра

Третьякова Александра Леонидовна

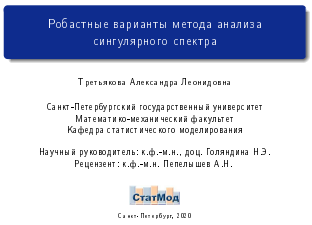
Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., доц. Голяндина Н.Э.
Рецензент: к.ф.-м.н. Пепелышев А.Н.

Санкт-Петербург, 2020

2020-06-12

Робастные варианты метода SSA


Робастные варианты метода анализа сингулярного спектра

Третьякова Александра Леонидовна

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., доц. Голяндина Н.Э.
Рецензент: к.ф.-м.н. Пепелышев А.Н.

СтатМод

Санкт-Петербург, 2020

Выпускная работа посвящена разработке и исследованию модификаций метода анализа сингулярного спектра, устойчивых к выбросам.
Работа выполнена на кафедре статистического моделирования, руководитель к.ф.-м.н., доц. Голяндина Н.Э.

Постановка задачи

Рассмотрим вещественнозначный временной ряд $\mathsf{X} = (x_1, \ldots, x_N)$, где N — длина ряда.
Предполагаем, что $x_i = s_i + r_i$, $i = 1, \ldots, N$, где r_i — шум.

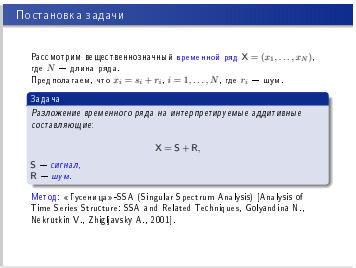
Задача

Разложение временного ряда на интерпретируемые аддитивные составляющие:

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

S — *сигнал,*
R — *шум.*

Метод: «Гусеница»-SSA (Singular Spectrum Analysis) [Analysis of Time Series Structure: SSA and Related Techniques, Golyandina N., Nekrutkin V., Zhigljavsky A., 2001].



В реальной жизни часто возникают задачи исследования различных процессов с течением времени. Работа посвящена одному из методов исследования временных рядов — методу анализа сингулярного спектра (Singular Spectrum Analysis, SSA), который позволяет анализировать ряд без задания его параметрической модели. Пусть имеется временной ряд X длины N, который представляет собой сумму сигнала и шума. 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.

Постановка задачи

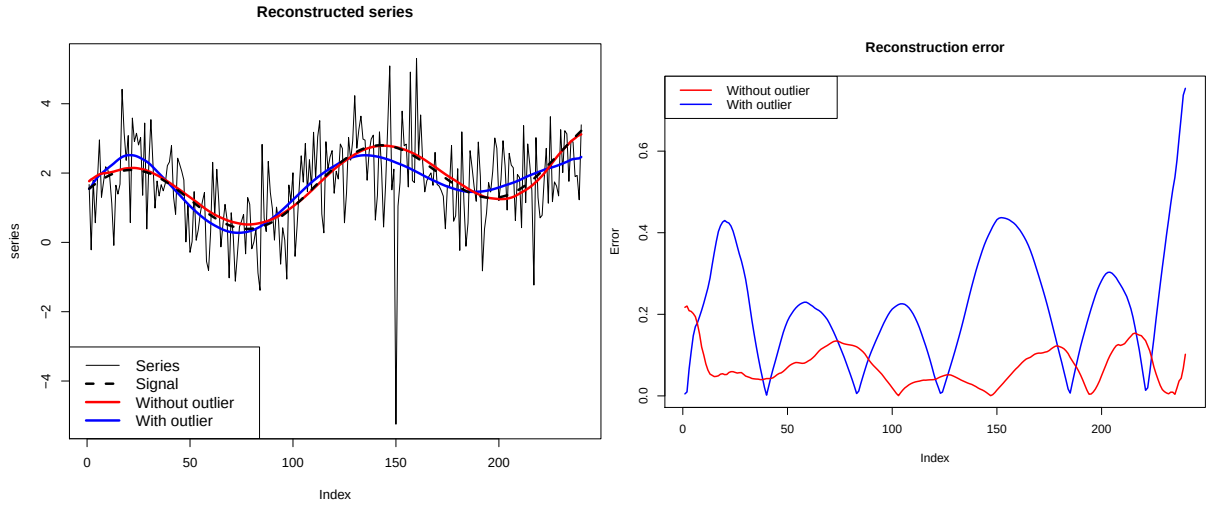
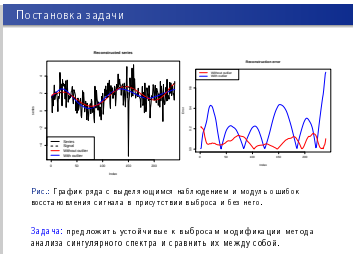


Рис.: График ряда с выделяющимся наблюдением и модуль ошибок восстановления сигнала в присутствии выброса и без него.

Задача: предложить устойчивые к выбросам модификации метода анализа сингулярного спектра и сравнить их между собой.



Однако на практике часто возникают выделяющиеся наблюдения или выбросы. Это могут быть ошибки в данных или сбои в работе измерительного прибора, которые могут сильно повлиять на восстановление сигнала методом SSA. На рисунках представлен график ряда с выбросом и восстановление сигнала в присутствии выделяющегося наблюдения и без него, а также модуль ошибки выделения сигнала. Видно, что выброс сильно портит восстановление сигнала. Поэтому значительный интерес представляет разработка устойчивых к выбросам модификаций метода SSA. Задачей работы является предложить робастные модификации метода анализа сингулярного спектра и сравнить их между собой, а также с классическим методом SSA.

Метод SSA для выделения сигнала ранга, не превосходящего r

Ряд X $= (x_1, \ldots, x_N)$.
Пусть $0 < L < N$ — длина окна. $K = N - L + 1$.
Обозначим $\mathcal{M}$ — пространство матриц $L \times K$,
$\mathcal{M}_{\mathcal{H}}$ — пространство ганкелевых матриц $L \times K$,
$\mathcal{M}_r$ — множество матриц ранга, не превосходящего r.

Ряд $\mapsto$ траекторная матрица $\mathbf{X}$:

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

Метод SSA для выделения сигнала ранга, не превосходящего r
Ряд X $= (x_1, \ldots, x_N)$.
Пусть $0 < L < N$ — длина окна. $K = N - L + 1$.
Обозначим $\mathcal{M}$ — пространство матриц $L \times K$,
$\mathcal{M}_{\mathcal{H}}$ — пространство ганкелевых матриц $L \times K$,
$\mathcal{M}_r$ — множество матриц ранга, не превосходящего r.
Ряд $\mapsto$ траекторная матрица $\mathbf{X}$:
$\mathbf{X} = [X_1 : \ldots : X_K] = \begin{pmatrix} x_1 & x_2 & x_3 & \ldots & x_K \\ x_2 & x_3 & x_4 & \ldots & x_{K+1} \\ x_3 & x_4 & x_5 & \ldots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \ldots & x_N \end{pmatrix}$.

Для начала опишем алгоритм метода SSA для выделения сигнала ранга, не превосходящего r.
Пусть имеется ряд длины N. Выберем целое число L — длина окна, K полагаем равным $N - L + 1$. Введем пространство ганкелевых матриц (матрица называется ганкелевой, если ее элементы на антидиагоналях равны между собой), а также множество матриц ранга, не превосходящего r. Введем понятие траекторной матрицы, состоящей из векторов вложения $(x_1, \ldots, x_L)^{\mathrm{T}}$, $(x_2, \ldots, x_{L+1})^{\mathrm{T}}$ и так далее.

- 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.
- $\Pi_{\mathcal{H}} : \mathcal{M} \to \mathcal{M}_{\mathcal{H}}$ — проектор на пространство ганкелевых матриц.

Структура сигнала S: ранг траекторной (ганкелевой) матрицы $\mathcal{T}(\mathsf{S})$ равен r.

В результате получаем оценку сигнала:

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

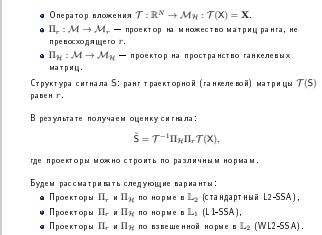
где проекторы можно строить по различным нормам.

Будем рассматривать следующие варианты:

- Проекторы Π_r и $\Pi_{\mathcal{H}}$ по норме в $\mathbb{L}_2$ (стандартный L2-SSA),
- Проекторы Π_r и $\Pi_{\mathcal{H}}$ по норме в $\mathbb{L}_1$ (L1-SSA),
- Проекторы Π_r и $\Pi_{\mathcal{H}}$ по взвешенной норме в $\mathbb{L}_2$ (WL2-SSA).

2020-06-12

Робастные варианты метода SSA



Введем оператор вложения $\mathcal{T}$, который переводит ряд в траекторную матрицу. А также два проектора: Π_r — проектор на множество матриц ранга, не превосходящего r, и $\Pi_{\mathcal{H}}$ — проектор на пространство ганкелевых матриц. Предполагаем, что ранг траекторной матрицы сигнала равен r. Для того, чтобы получить оценку сигнала, необходимо сначала применить оператор вложения, получить из ряда траекторную матрицу. Затем сделать проекцию на множество матриц ранга, не превосходящего r, затем проекцию на пространство ганкелевых матриц, и снова превратить в ряд. Проекторы можно строить по различным нормам. Для построения устойчивых модификаций рассмотрим два подхода. Первый подход состоит в использовании нормы в пространстве $\mathbb{L}_1$, которая является более устойчивой к выбросам. Второй подход — использование взвешенной нормы в $\mathbb{L}_2$, где точкам, содержащим выбросы, присваивается меньший вес. Таким образом, будем рассматривать следующие варианты: обычный SSA, где проекторы строятся по нормам в пространстве $\mathbb{L}_2$, вариант с проекторами по норме в пространстве $\mathbb{L}_1$ и вариант с проекторами по взвешенной норме в $\mathbb{L}_2$.

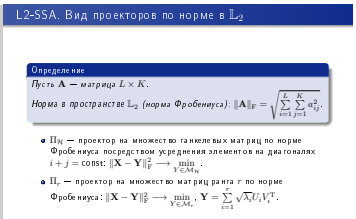
L2-SSA. Вид проекторов по норме в $\mathbb{L}_2$

Определение

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

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

- $\Pi_{\mathcal{H}}$ — проектор на множество ганкелевых матриц по норме Фробениуса посредством усреднения элементов на диагоналях $i+j = \text{const}$: $\|\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2 \longrightarrow \min\limits_{Y \in \mathcal{M}_{\mathcal{H}}}$.
- Π_r — проектор на множество матриц ранга r по норме Фробениуса: $\|\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2 \longrightarrow \min\limits_{Y \in \mathcal{M}_r}$, $\mathbf{Y} = \sum\limits_{i=1}^{r} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$.



Опишем, как строятся проекторы по норме в пространстве $\mathbb{L}_2$ (норма Фробениуса). Проектор на множество ганкелевых матриц строится посредством усреднения элементов на побочных диагоналях.
Для того, чтобы получить проекцию на множество матриц ранга, не превосходящего r, необходимо взять первые r компонент сингулярного разложения траекторной матрицы ряда.

L1-SSA. Вид проекторов по норме в $\mathbb{L}_1$

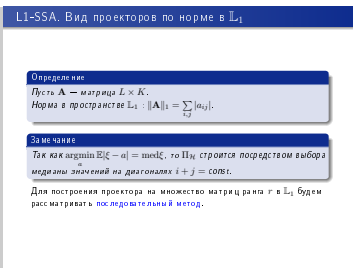
Определение

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

Замечание

Так как $\operatorname{argmin}\limits_a \mathbb{E}|\xi - a| = \operatorname{med}\xi$, *то* $\Pi_{\mathcal{H}}$ *строится посредством выбора медианы значений на диагоналях* $i + j = const$.

Для построения проектора на множество матриц ранга r в $\mathbb{L}_1$ будем рассматривать последовательный метод.



Для получения проекции на множество ганкелевых матриц по норме в пространстве $\mathbb{L}_1$, необходимо взять медиану значений на побочных диагоналях. Однако решения в явном виде задачи построения проектора на множество матриц ранга, не превосходящего r, в пространстве $\mathbb{L}_1$ нет. Мы будем рассматривать последовательный метод для решения этой задачи, который находит решение задачи итеративно.

L1-SSA. Реализация. Последовательный метод

В R-пакете pcaL1 [Jot et al., 2017] имеется реализация последовательного метода решения задачи $\|\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}\|_1 \longrightarrow \min\limits_{\mathbf{U},\mathbf{V}}$.

Алгоритм l1pca [Brooks J. P., Jot S., 2013] :

1. Инициализация $\mathbf{U}(0) \in \mathbb{R}^{L\times r}$, нормировка столбцов $\mathbf{U}(0)$,
2. $t := t + 1$,
3. $\mathbf{V}(t) = \underset{\mathbf{V}\in\mathbb{R}^{K\times r}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{U}(t-1)\mathbf{V}^{\mathrm{T}}\|_1$

   Задача разбивается на K независимых подзадач вида $\mathbf{v}_i = \underset{\mathbf{x}}{\operatorname{argmin}} \|Y_i - \mathbf{U}(t-1)\mathbf{x}\|_1$, где $Y_i \in \mathbb{R}^L$ — столбцы $\mathbf{Y}$, $\mathbf{v}_i \in \mathbb{R}^r$ — строки $\mathbf{V}$, $i = 1, \ldots, K$,
4. $\mathbf{U}(t) = \underset{\mathbf{U}\in\mathbb{R}^{L\times r}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}(t)\|_1$ (решается аналогично п.3)
5. Нормировка столбцов $\mathbf{U}(t)$,
6. if $\mathbf{U}(t) \neq \mathbf{U}(t-1)$ (по крит. остановки) then Go to Step 2
   else $\mathbf{U} := \mathbf{U}(t)$; $\mathbf{V} := \mathbf{V}(t)$.

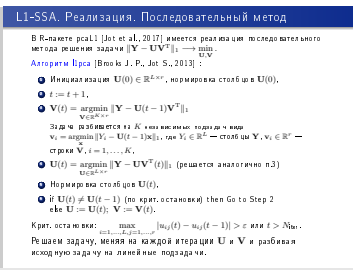
Крит. остановки: $\max\limits_{i=1,\ldots,L,j=1,\ldots,r} |u_{ij}(t) - u_{ij}(t-1)| > \varepsilon$ или $t > N_{\text{iter}}$.

Решаем задачу, меняя на каждой итерации $\mathbf{U}$ и $\mathbf{V}$ и разбивая исходную задачу на линейные подзадачи.

2020-06-12

Робастные варианты метода SSA

└L1-SSA. Реализация. Последовательный метод



Сам алгоритм описан в статье 2013 года, в R-пакете pcaL1 имеется его реализация. Задача представляется в виде $\|\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}\|_1 \longrightarrow \min\limits_{\mathbf{U},\mathbf{V}}$.

Изначально инициализируем матрицу $\mathbf{U}$, затем на каждой итерации фиксируем $\mathbf{U}$ и решаем задачу относительно $\mathbf{V}$. Затем фиксируем $\mathbf{V}$ и минимизируем по $\mathbf{U}$, пока не выполнен критерий остановки.

Задачу из пункта 3 алгоритма можно разбить на K независимых подзадач. С помощью решения каждой такой подзадачи получаем оценку $\mathbf{v}_i$, $i = 1, \ldots, K$ — строк матрицы $\mathbf{V}$. Согласно [Ke Q., Kanade T., 2005], каждая подзадача сводится к задаче линейного программирования с ограничениями. Задача из пункта 4 алгоритма решается аналогичным образом.

Таким образом, мы решаем задачу, фиксируя и меняя на каждой итерации $\mathbf{U}$ и $\mathbf{V}$, разбивая исходную задачу на линейные подзадачи.

WL2-SSA. Вид проекторов по взвешенной норме в $\mathbb{L}_2$

Определение

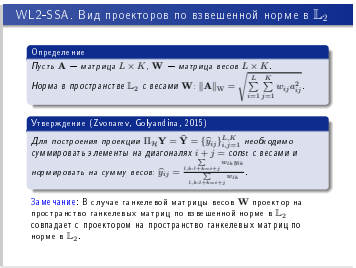
Пусть $\mathbf{A}$ *— матрица* $L \times K$, $\mathbf{W}$ *— матрица весов* $L \times K$.

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

Утверждение (Zvonarev, Golyandina, 2015)

Для построения проекции $\Pi_{\mathcal{H}}\mathbf{Y} = \widehat{\mathbf{Y}} = \{\widehat{y}_{ij}\}_{i,j=1}^{L,K}$ *необходимо суммировать элементы на диагоналях* $i+j = const$ *с весами и нормировать на сумму весов:* $\widehat{y}_{ij} = \dfrac{\sum\limits_{l,k:l+k=i+j} w_{lk} y_{lk}}{\sum\limits_{l,k:l+k=i+j} w_{lk}}$.

Замечание: 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 $\mathbb{L}_2$.



Введем норму в пространстве $\mathbb{L}_2$ с весами $\mathbf{W}$. Для построения проекции на пространство ганкелевых матриц необходимо суммировать элементы на побочных диагоналях с весами и нормировать на сумму весов. Можно заметить, что если матрица весов ганкелева, то проектор на пространство ганкелевых матриц по взвешенной норме в $\mathbb{L}_2$ совпадает с проектором на пространство ганкелевых матриц по норме в $\mathbb{L}_2$.

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

WL2-SSA. Метод с итеративным обновлением весов

Пусть $\mathbf{Y} \in \mathbb{R}^{L \times K}$ — траекторная матрица ряда, $\widehat{\mathbf{Y}} = \mathbf{U}\mathbf{V}^{\mathrm{T}} = \{\widehat{y}_{ij}\}_{i,j=1}^{L,K}$.
Решаем 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

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

Значения α и $\{\sigma_{ij},\ i = 1, \ldots, L,\ j = 1, \ldots, K\}$ — параметры.

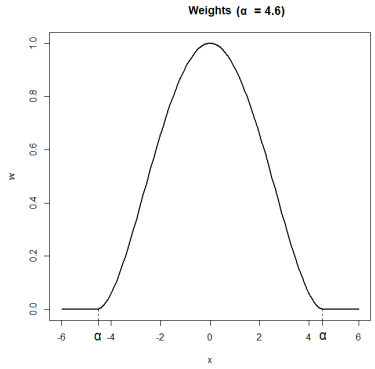
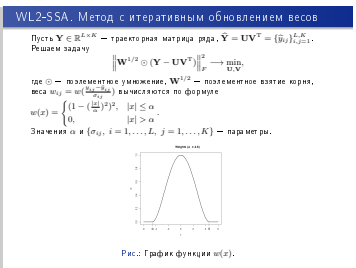


Рис.: График функции $w(x)$.



Пусть $\mathbf{Y} \in \mathbb{R}^{L \times K}$ — траекторная матрица ряда. Решаем 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 веса вычисляются по определенной формуле, как в известном методе локальной регрессии loess. Имеются параметры α и σ_{ij}. Их выбор мы обсудим далее. График весовой функции $w(x)$ представлен на слайде. Значение веса зависит от нормированных остатков: если остаток маленький по модулю, то вес максимальный, если остаток по модулю превосходит заданный параметр α, то вес обнуляется.

WL2-SSA. Метод с итеративным обновлением весов. Реализация

Задача:

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

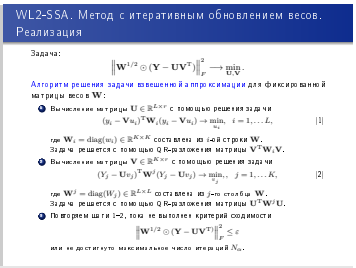
Алгоритм решения задачи взвешенной аппроксимации для фиксированной матрицы весов $\mathbf{W}$:

1. Вычисление матрицы $\mathbf{U} \in \mathbb{R}^{L\times r}$ с помощью решения задачи
$$(y_i - \mathbf{V}u_i)^{\mathrm{T}}\mathbf{W}_i(y_i - \mathbf{V}u_i) \to \min_{u_i}, \quad i = 1, \dots L, \tag{1}$$
где $\mathbf{W}_i = \mathrm{diag}(w_i) \in \mathbb{R}^{K\times K}$ составлена из i-ой строки $\mathbf{W}$.
Задача решается с помощью QR-разложения матрицы $\mathbf{V}^{\mathrm{T}}\mathbf{W}_i\mathbf{V}$.
2. Вычисление матрицы $\mathbf{V} \in \mathbb{R}^{K\times r}$ с помощью решения задачи
$$(Y_j - \mathbf{U}v_j)^{\mathrm{T}}\mathbf{W}^j(Y_j - \mathbf{U}v_j) \to \min_{v_j}, , \quad j = 1, \dots K, \tag{2}$$
где $\mathbf{W}^j = \mathrm{diag}(W_j) \in \mathbb{R}^{L\times L}$ составлена из j-го столбца $\mathbf{W}$.
Задача решается с помощью QR-разложения матрицы $\mathbf{U}^{\mathrm{T}}\mathbf{W}^j\mathbf{U}$.
3. Повторяем шаги 1–2, пока не выполнен критерий сходимости
$$\left\|\mathbf{W}^{1/2} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T})}\right\|_F^2 \le \varepsilon$$
или не достигнуто максимальное число итераций N_α.

2020-06-12

Робастные варианты метода SSA

└WL2-SSA. Метод с итеративным обновлением весов. Реализация

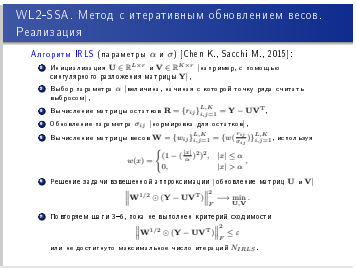


Для начала разберем алгоритм решения этой задачи при фиксированной матрице весов. На каждой итерации обновляем матрицы $\mathbf{U}$ и $\mathbf{V}$ с помощью решения задач (1) и (2). Решения находятся путем QR-разложения матриц $\mathbf{V}^{\mathrm{T}}\mathbf{W}_i\mathbf{V}$ и $\mathbf{U}^{\mathrm{T}}\mathbf{W}^j\mathbf{U}$ соответственно. Повторяем итерации, пока не выполнен критерий остановки или не достигнуто максимальное число итераций N_α.

WL2-SSA. Метод с итеративным обновлением весов. Реализация

Алгоритм IRLS (параметры α и σ) [Chen K., Sacchi M., 2015]:

1. Инициализация $\mathbf{U} \in \mathbb{R}^{L \times r}$ и $\mathbf{V} \in \mathbb{R}^{K \times r}$ (например, с помощью сингулярного разложения матрицы $\mathbf{Y}$),
2. Выбор параметра α (величина, начиная с которой точку ряда считать выбросом),
3. Вычисление матрицы остатков $\mathbf{R} = \{r_{ij}\}_{i,j=1}^{L,K} = \mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}$,
4. Обновление параметра σ_{ij} (нормировка для остатков),
5. Вычисление матрицы весов $\mathbf{W} = \{w_{ij}\}_{i,j=1}^{L,K} = \{w(\frac{r_{ij}}{\sigma_{ij}})\}_{i,j=1}^{L,K}$, используя
$$w(x) = \begin{cases} (1 - (\frac{|x|}{\alpha})^2)^2, & |x| \le \alpha \\ 0, & |x| > \alpha \end{cases},$$
6. Решение задачи взвешенной аппроксимации (обновление матриц $\mathbf{U}$ и $\mathbf{V}$)
$$\left\| \mathbf{W}^{1/2} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}) \right\|_F^2 \longrightarrow \min_{\mathbf{U},\mathbf{V}}.$$
7. Повторяем шаги 3–6, пока не выполнен критерий сходимости
$$\left\| \mathbf{W}^{1/2} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}) \right\|_F^2 \le \varepsilon$$
или не достигнуто максимальное число итераций N_{IRLS}.



На слайде представлен алгоритм решения задачи построения проектора на множество матриц ранга, не превосходящего r, методом с итеративным обновлением весов. На первом шаге инициализируем матрицы $\mathbf{U}$ и $\mathbf{V}$, например, взяв первые r компонент сингулярного разложения траекторной матрицы ряда. Затем фиксируем параметр α — величину, начиная с которой точку ряда будем считать выбросом. Далее вычисляем матрицу остатков и обновляем параметры σ_{ij}. Алгоритм содержит вычисление матрицы параметров $\mathbf{\Sigma} = \{\sigma_{ij}\}_{i,j=1}^{L,K}$, которое обсудим далее. Затем вычисляем матрицу весов $\mathbf{W}$ и решаем задачу аппроксимации при фиксированной матрице весов. Повторяем итерации, пока не выполнен критерий остановки или не достигнуто максимальное число итераций N_{IRLS}.

Далее обсудим выбор параметров σ_{ij} и α для метода с итеративным обновлением весов.

WL2-SSA. Метод с итеративным обновлением весов. Выбор параметров. Параметр σ

Проблема 1: Нормировка остатков на константный параметр $\sigma_{ij} = \sigma \ \ \forall i, j$ 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.

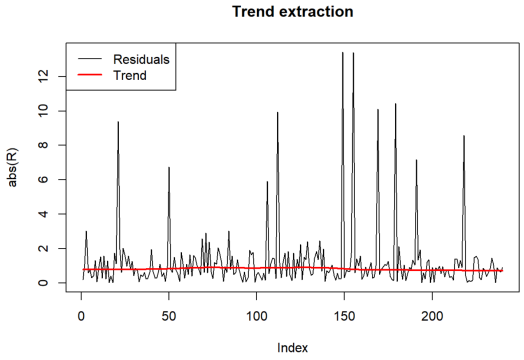


Рис.: График модуля остатков. Постоянная дисперсия шума.

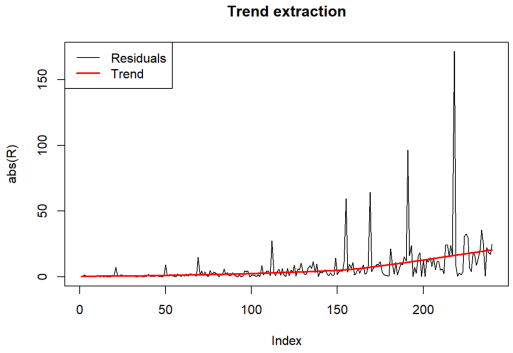
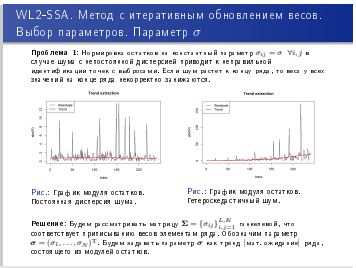


Рис.: График модуля остатков. Гетероскедастичный шум.

Решение: Будем рассматривать матрицу $\mathbf{\Sigma} = \{\sigma_{ij}\}_{i,j=1}^{L,K}$ ганкелевой, что соответствует приписыванию весов элементам ряда. Обозначим параметр $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)^{\mathrm{T}}$. Будем задавать параметр $\boldsymbol{\sigma}$ как тренд (мат. ожидание) ряда, состоящего из модулей остатков.



Авторы алгоритма предлагают выбрать σ_{ij} не зависящими от i, j. Однако такой выбор не подходит, к примеру, для рядов с гетероскедастичным шумом. Нормировка остатков на константный параметр $\sigma_{ij} = \sigma \ \forall i, j$ 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. В модификации метода предполагается замена параметра σ_{ij} на элементы траекторной матрицы тренда (оценки математического ожидания) ряда, состоящего из модулей остатков. Выделять тренд будем следующими способами: локальной регрессией loess, скользящей медианой или взвешенной локальной регрессией lowess.

WL2-SSA. Выбор параметров. Параметр α

Проблема 2: Непонятно, как задавать параметр α, влияющий на то, какие точки считать выбросами, а какие — нет.
Решение: Выведем вероятностную формулу для параметра α.
Модель ряда: $x_i = s_i + \varepsilon_i$, $i = 1, \ldots, N$. Ганкелизуем матрицу остатков $\mathbf{R} = \mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}$, получим ряд $\mathsf{R} = \{r_i\}_{i=1}^N$. В предположении точной отделимости сигнала от шума, ряд $\mathsf{R} = \{\varepsilon_i\}_{i=1}^N$. Будем задавать вероятность γ:
$\mathrm{P}(r^* \in (0, \alpha)) = \gamma$, где $r^* = \frac{|\varepsilon|}{\sigma}$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$ — тренд из ряда $|\mathsf{R}|$.

Определение

Если $r \sim N(0, \sigma^2)$, *то* $|r| \sim N_H(\sigma^2)$ — *полунормальное распределение с параметром* σ^2, *ф.р.* $F_H(x; \sigma^2) = \frac{2}{\sqrt{\pi}} \int\limits_0^{x/\sqrt{2\sigma^2}} e^{-z^2} dz = \mathit{erf}(\frac{x}{\sqrt{2\sigma^2}})$.

Утверждение

Пусть $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, $\sigma = E|\varepsilon|$. *Тогда* $r^* = \frac{|\varepsilon|}{\sigma}$ *имеет полунормальное распределение* $N_H(\frac{\pi}{2})$, *среднее* $Er^* = 1$, *дисперсия* $Dr^* = \frac{\pi}{2} - 1$.

Получаем выражение для α:

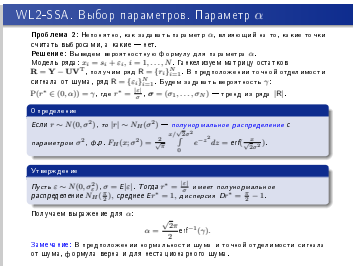
$$\alpha = \frac{\sqrt{2}\pi}{2} \mathrm{erf}^{-1}(\gamma).$$

Замечание: В предположении нормальности шума и точной отделимости сигнала от шума, формула верна и для нестационарного шума.

2020-06-12

Робастные варианты метода SSA

└WL2-SSA. Выбор параметров. Параметр α

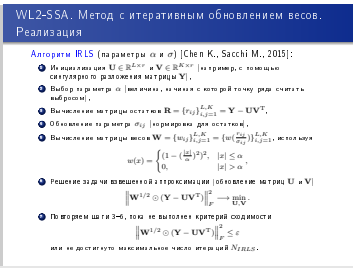


У метода с итеративным обновлением весов есть параметр α, который влияет на то, какие точки будем считать выбросами, а какие — нет. Для того, чтобы понять, какое значение следует взять в качестве α, выведем вероятностную формулу для этого параметра. Для вывода формулы введем определение полунормального распределения.
Напомним, что мы рассматриваем ряд, который является суммой сигнала и шума. Если предположить точную отделимость сигнала от шума, то ряд из остатков соответствует шуму. Обозначим $r^* = \frac{|\varepsilon|}{\sigma}$, где σ — компонента тренда ряда из модулей остатков. Мы хотим задавать вероятность γ так, чтобы нормированные модули остатков попадали в промежуток $(0; \alpha)$ с вероятностью γ. Для того, чтобы получить выражение для α, нам необходимо знать распределение r^*. Можно показать, что r^* имеет полунормальное распределение с параметром $\frac{\pi}{2}$. Тогда можем вывести формулу для параметра α, она представлена на слайде. Также можно заметить, что утверждение и выведенная формула верны не только для шума постоянной дисперсии, но и для гетероскедастичного шума.

WL2-SSA. Метод с итеративным обновлением весов. Реализация

Алгоритм IRLS (параметры α и σ) [Chen K., Sacchi M., 2015]:

1. Инициализация $\mathbf{U} \in \mathbb{R}^{L \times r}$ и $\mathbf{V} \in \mathbb{R}^{K \times r}$ (например, с помощью сингулярного разложения матрицы $\mathbf{Y}$),
2. Выбор параметра α (величина, начиная с которой точку ряда считать выбросом),
3. Вычисление матрицы остатков $\mathbf{R} = \{r_{ij}\}_{i,j=1}^{L,K} = \mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}$,
4. Обновление параметра σ_{ij} (нормировка для остатков),
5. Вычисление матрицы весов $\mathbf{W} = \{w_{ij}\}_{i,j=1}^{L,K} = \{w(\frac{r_{ij}}{\sigma_{ij}})\}_{i,j=1}^{L,K}$, используя
$$w(x) = \begin{cases} (1 - (\frac{|x|}{\alpha})^2)^2, & |x| \le \alpha \\ 0, & |x| > \alpha \end{cases},$$
6. Решение задачи взвешенной аппроксимации (обновление матриц $\mathbf{U}$ и $\mathbf{V}$)
$$\left\| \mathbf{W}^{1/2} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}) \right\|_F^2 \longrightarrow \min_{\mathbf{U},\mathbf{V}}.$$
7. Повторяем шаги 3–6, пока не выполнен критерий сходимости
$$\left\| \mathbf{W}^{1/2} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}) \right\|_F^2 \le \varepsilon$$
или не достигнуто максимальное число итераций N_{IRLS}.



Посмотрим еще раз на алгоритм оригинального метода с обновлением весов. Модификация отличается от оригинального метода пятым пунктом.

WL2-SSA. Метод с итеративным обновлением весов. Модификация

Модификация 5-ого пункта алгоритма IRLS:

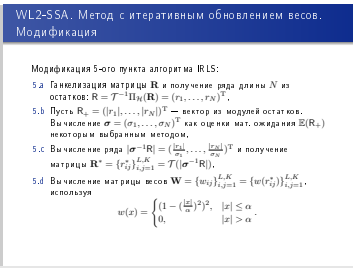
5.a Ганкелизация матрицы $\mathbf{R}$ и получение ряда длины N из остатков: $\mathrm{R} = \mathcal{T}^{-1}\Pi_{\mathcal{H}}(\mathbf{R}) = (r_1, \ldots, r_N)^{\mathrm{T}}$,

5.b Пусть $\mathrm{R}_+ = (|r_1|, \ldots, |r_N|)^{\mathrm{T}}$ — вектор из модулей остатков. Вычисление $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)^{\mathrm{T}}$ как оценки мат. ожидания $\mathbb{E}(\mathrm{R}_+)$ некоторым выбранным методом,

5.c Вычисление ряда $|\boldsymbol{\sigma}^{-1}\mathrm{R}| = (\frac{|r_1|}{\sigma_1}, \ldots, \frac{|r_N|}{\sigma_N})^{\mathrm{T}}$ и получение матрицы $\mathbf{R}^* = \{r^*_{ij}\}_{i,j=1}^{L,K} = \mathcal{T}(|\boldsymbol{\sigma}^{-1}\mathrm{R}|)$,

5.d Вычисление матрицы весов $\mathbf{W} = \{w_{ij}\}_{i,j=1}^{L,K} = \{w(r^*_{ij})\}_{i,j=1}^{L,K}$, используя

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



В модификации алгоритма после вычисления матрицы остатков мы должны ее ганкелизовать, получив ряд из остатков. Затем вычислить некоторым методом тренд (оценку мат. ожидания) из ряда, состоящего из модулей остатков, и нормировать модули на вычисленный тренд. Построив траекторную матрицу получившегося ряда, мы получаем новую матрицу остатков, и далее применяем функцию весов уже к этой матрице. Тренд будем выделять следующими способами: с помощью локальной регрессии loess, скользящей медианой или с помощью взвешенной локальной регрессии lowess.

Сравнение теоретических трудоемкостей

Ряд $\mathsf{X} = (x_1, \ldots, x_N)$ длины N, матрица $\mathbf{Y} \in \mathbb{R}^{L \times K}$ — траекторная матрица ряда X. Ранг траекторной матрицы сигнала равен r.

- Трудоемкость последовательного метода:

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

  где N_{iter} — общее кол-во итераций для сходимости метода (по выбранному критерию сходимости).

- Трудоемкость метода с обновлением весов:

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

  где N_α и N_{IRLS} — общее кол-во итераций для решения задач (1), (2) и сходимости метода (по выбранному критерию сходимости).

Число итераций в статьях предполагается фиксированным. Однако предположение о достаточности фиксированного числа итераций не верно. В предположении, что число итераций не растет с увеличением длины ряда, так как зависит от разделимости, метод с итеративным обновлением весов оказывается менее трудоемким.

Сравнение теоретических трудоемкостей
Ряд $\mathsf{X} = (x_1, \ldots, x_N)$ длины N, матрица $\mathbf{Y} \in \mathbb{R}^{L \times K}$ — траекторная матрица ряда X. Ранг траекторной матрицы сигнала равен r.
- Трудоемкость последовательного метода:
$$\mathrm{T}_{\mathrm{l1pca}} = O(LK \log(2LK + Lr) N_{iter}),$$
где N_{iter} — общее кол-во итераций для сходимости метода (по выбранному критерию сходимости).
- Трудоемкость метода с обновлением весов:
$$\mathrm{T}_{\mathrm{IRLS}} = O(LKr^2 N_\alpha N_{IRLS}),$$
где N_α и N_{IRLS} — общее кол-во итераций для решения задач (1), (2) и сходимости метода (по выбранному критерию сходимости).
Число итераций в статьях предполагается фиксированным. Однако предположение о достаточности фиксированного числа итераций не верно. В предположении, что число итераций не растет с увеличением длины ряда, так как зависит от разделимости, метод с итеративным обновлением весов оказывается менее трудоемким.

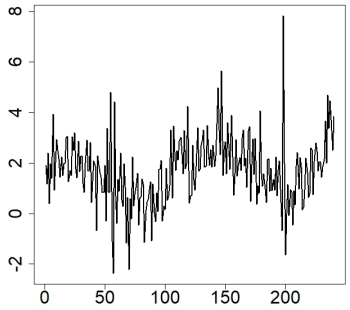
Сравним теоретические трудоемкости рассмотренных методов. Трудоемкости последовательного метода и метода с обновлением весов представлены на слайде. В трудоемкость входит число итераций, которое в статьях предполагается фиксированным. В работе показано, что предположение о достаточности фиксированного 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.

Вычислительный эксперимент. Структура исследования

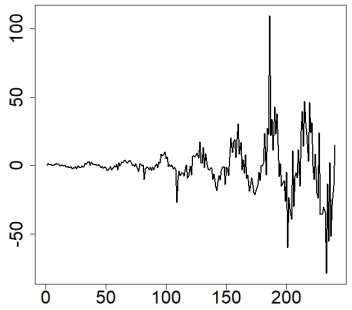
Пусть длина ряда $N = 240$. Рассмотрим следующие примеры:

1. $x_n = e^{n/N} + \sin(2\pi n/120 + \pi/6) + \varepsilon_n, \quad \varepsilon_n \sim N(0,1),\ r = 3,$
2. $x_n = e^{4n/N}\sin(2\pi n/30) + Ae^{4n/N}\varepsilon_n, \quad \varepsilon_n \sim N(0,1),\ r = 2,$
3. $x_n = ne^{4n/N}\sin(2\pi n/30) + \varepsilon_n, \quad \varepsilon_n \sim N(0,1),\ r = 4.$

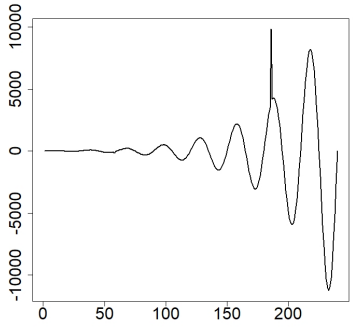
Выбросы: в случайно выбранных точках ряда x_i значение заменяется на $x_i + \delta x_i$, где δ — заданная константа.



Пр. 1: 1% выбросов, выброс размера $5x_i$.



Пр. 2: 1% выбросов, выброс размера $5x_i$.



Пр. 3: 1% выбросов, выброс размера $1.5x_i$.

2020-06-12

Робастные варианты метода SSA

└Вычислительный эксперимент. Структура исследования

Рассмотрим 3 модельных примера. Формулы для рядов и графики с выбросами представлены на слайде. Выбросы добавляются следующим образом: случайных точках ряда к значению ряда x_i добавляется δx_i, где δ — заранее заданная константа.

Вычислительный эксперимент. Структура исследования

Временной ряд $\mathsf{X} = (x_1, \ldots, x_N)$, $x_i = s_i + \varepsilon_i$, $i = 1, \ldots, N$.
Обозначим $\mathsf{S} = (s_1, \ldots, s_N)^{\mathrm{T}}$ — сигнал.

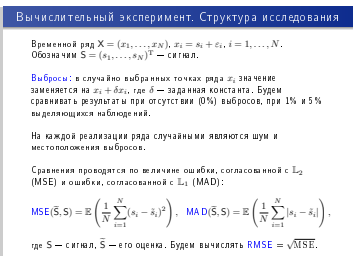
Выбросы: в случайно выбранных точках ряда x_i значение заменяется на $x_i + \delta x_i$, где δ — заданная константа. Будем сравнивать результаты при отсутствии (0%) выбросов, при 1% и 5% выделяющихся наблюдений.

На каждой реализации ряда случайными являются шум и местоположения выбросов.

Сравнения проводятся по величине ошибки, согласованной с $\mathbb{L}_2$ (MSE) и ошибки, согласованной с $\mathbb{L}_1$ (MAD):

$$\mathrm{MSE}(\widetilde{\mathsf{S}}, \mathsf{S}) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}(s_i - \tilde{s}_i)^2\right), \quad \mathrm{MAD}(\widetilde{\mathsf{S}}, \mathsf{S}) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}|s_i - \tilde{s}_i|\right),$$

где S — сигнал, $\widetilde{\mathsf{S}}$ — его оценка. Будем вычислять $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$.



Будем сравнивать результаты при отсутствии выбросов, при 1% и 5% выбросов, которые будут находиться в случайных точках ряда.
Сравнение будем проводить по величине ошибки MSE, согласованной с $\mathbb{L}_2$, и MAD, согласованной с $\mathbb{L}_1$. Из оценки ошибки MSE будем извлекать корень, получая RMSE.

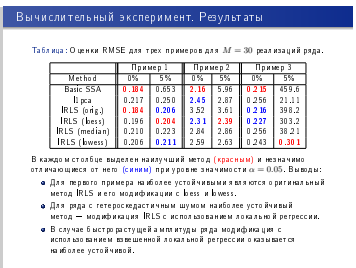
Вычислительный эксперимент. Результаты

Таблица: Оценки RMSE для трех примеров для $M = 30$ реализаций ряда.

	Пример 1		Пример 2		Пример 3	
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**

В каждом столбце выделен наилучший метод (красным) и незначимо отличающиеся от него (синим) при уровне значимости $\alpha = 0.05$. Выводы:

- Для первого примера наиболее устойчивыми являются оригинальный метод IRLS и его модификации с loess и lowess.
- Для ряда с гетероскедастичным шумом наиболее устойчивый метод — модификация IRLS с использованием локальной регрессии.
- В случае быстрорастущей амплитуды ряда модификация с использованием взвешенной локальной регрессии оказывается наиболее устойчивой.



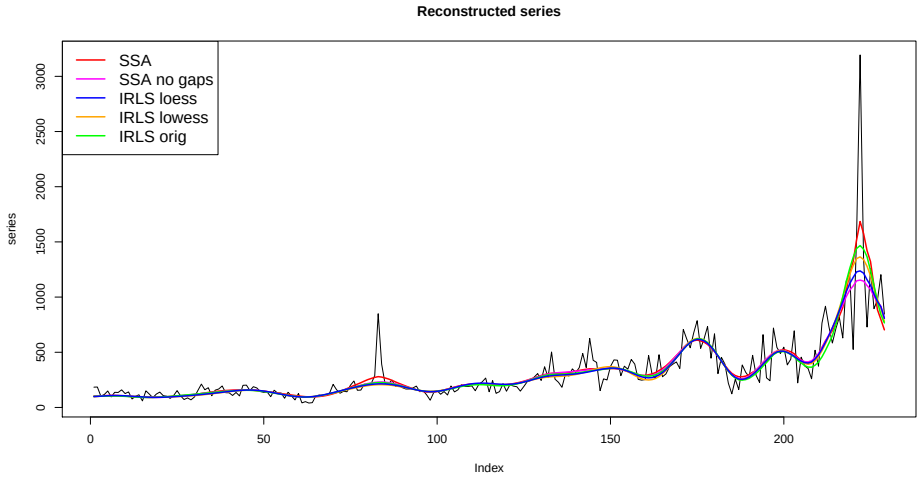
Результаты вычислительного эксперимента для модельных примеров представлены в таблице. В каждом столбце красным выделен метод, показавший наименьшую ошибку, а синим — незначимо отдичающиеся от него при уровне значимости 0.05. Проверка значимости проводилась по критерию для зависимых выборок, число реализаций ряда было взято равным 30.

Можно сделать следующие выводы. Для примера без растущей амплитуды ряда и с шумом постоянной дисперсии наиболее устойчивыми являются оригинальный метод с обновлением весов и его модификации с использованием локальной регрессии loess и взвешенной локальной регрессии lowess. Для ряда с гетероскедастичным шумом преимущество оригинального метода с обновлением весов IRLS пропадает, наиболее устойчивым методом оказывается модификация IRLS с использованием локальной регрессии. В случае быстрорастущей амплитуды ряда модификация с использованием взвешенной локальной регрессии оказывается наиболее устойчивой.

Реальный пример

Рассмотрим ряд — импорт товаров в США из Кувейта с ноября 1993 г. по ноябрь 2012 г.. Имеются данные за каждый месяц. Длина ряда $N = 229$. Возьмем длину окна $L = 60$, будем восстанавливать сигнал по 5 компонентам.

Выбросы находятся в точках x_{83} и x_{222}. В качестве истинного сигнала будем брать результат восстановления сигнала стандартным SSA для ряда с поставленными на места выбросов и впоследствии заполненными пропусками.



Продемонстрируем работу методов на реальном примере. Рассмотрим ряд — импорт товаров в США из Кувейта с ноября 1993 года по ноябрь 2012 года с данными за каждый месяц. Возьмем длину окна $L = 60$. Проанализировав графики элементарных восстановленных рядов и матрицу взвешенных корреляций, был сделан вывод, что восстанавливать сигнал необходимо по первым 5 компонентам. Так как настоящий сигнал нам неизвестен, то попробуем на месте выбросов поставить пропуски, заполнить пропущенные значения, а затем выделить сигнал с помощью классического SSA. Полученный сигнал будем считать истинным. Сравним стандартный SSA и оригинальный метод с обновлением весов с предложенными модификациями (с выделением тренда с помощью локальной регрессии loess и взвешенной локальной регрессии lowess). Результат восстановления сигнала различными методами представлен на рисунке. 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.

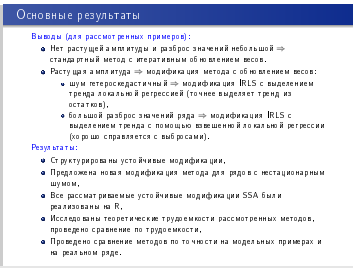
Основные результаты

Выводы (для рассмотренных примеров):

- Нет растущей амплитуды и разброс значений небольшой ⇒ стандартный метод с итеративным обновлением весов.
- Растущая амплитуда ⇒ модификация метода с обновлением весов:
  - шум гетероскедастичный ⇒ модификация IRLS с выделением тренда локальной регрессией (точнее выделяет тренд из остатков),
  - большой разброс значений ряда ⇒ модификация IRLS с выделением тренда с помощью взвешенной локальной регрессии (хорошо справляется с выбросами).

Результаты:

- Структурированы устойчивые модификации,
- Предложена новая модификация метода для рядов с нестационарным шумом,
- Все рассматриваемые устойчивые модификации SSA были реализованы на R,
- Исследованы теоретические трудоемкости рассмотренных методов, проведено сравнение по трудоемкости,
- Проведено сравнение методов по точности на модельных примерах и на реальном ряде.



Исходя из проведенного исследования можно сказать, что если у ряда нет растущей амплитуды и разброс значений небольшой, то можно использовать оригинальный 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 лок. регрессии. Если же разброс значений у ряда большой, то следует использовать модификацию с выделением тренда с помощью взвешенной лок. регрессии, которая хорошо справляется с выбросами.
Среди основных результатов работы можно выделить следующее: описаны и структурированы устойчивые модификации, предложена новая модификация метода, расширяющая его применимость на случай нестационарного шума. Все рассматриваемые устойчивые модификации SSA были реализованы на R, исследованы теоретические трудоемкости рассмотренных методов, проведено сравнение методов по трудоемкости. Также было проведено сравнение методов по точности на модельных примерах и на реальном ряде.