

Устойчивые варианты метода SSA для анализа временных рядов

Третьякова Александра Леонидовна, гр. 422

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

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



Санкт-Петербург
2018г.

Рассмотрим вещественнозначный **временной ряд**

$X = (x_1, \dots, x_N)$, где N — длина ряда.

Задача

Разложение временного ряда на интерпретируемые аддитивные составляющие:

$$X = S + R,$$

S — **сигнал**,

R — **шум**.

Метод: «Гусеница»-SSA (Singular Spectrum Analysis).

Постановка задачи

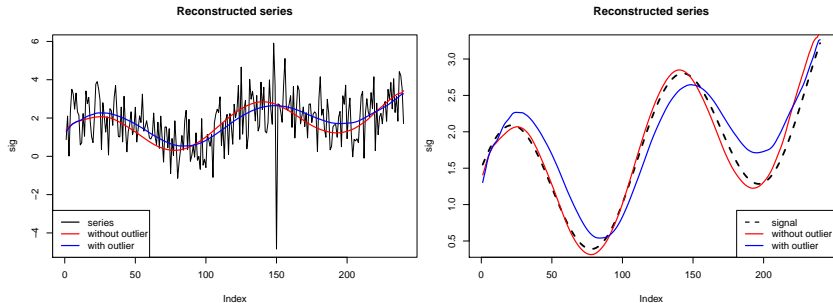


Рис.: График ряда с выделяющимся наблюдением и восстановление сигнала в присутствии выброса и без него.

Задача: необходимо исследовать устойчивость метода SSA к выделяющимся наблюдениям, а также предложить более устойчивые модификации метода.

Метод SSA для выделения сигнала ранга r

Ряд $X = (x_1, \dots, x_N)$.

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

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

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

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

Получаем оценку сигнала:

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

где проекторы можно строить по различным нормам.

Определение

Пусть \mathbf{A} — матрица $m \times n$.

Норма в пространстве \mathbb{L}_2 (норма Фробениуса):

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}.$$

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

Определение

Пусть \mathbf{A} — матрица $m \times n$.

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

Замечание

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

Для построения проектора на множество матриц ранга r в \mathbb{L}_1 имеется несколько методов:

- 1 Последовательный метод;
- 2 Метод с регуляризацией;
- 3 Регрессионный метод (шаг последовательного метода).

В R-пакете `rcaL1` имеется реализация последовательного метода решения задачи $\|\mathbf{X} - \mathbf{UV}^T\|_1^2 \rightarrow \min_{\mathbf{U}, \mathbf{V}}, \quad \mathbf{U}^T \mathbf{U} = \mathbf{I}_r$.

Алгоритм `l1pca` [Jot et al., 2017]:

- ❶ Инициализация $\mathbf{V}(0) \in \mathbb{R}^{r \times n}$.
- ❷ $t := t + 1$.
- ❸ $\mathbf{U}(t) = \underset{\mathbf{U} \in \mathbb{R}^{m \times r}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{UV}^T(t-1)\|_1$.
- ❹ $\mathbf{V}(t) = \underset{\mathbf{V} \in \mathbb{R}^{r \times n}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{U}(t)\mathbf{V}^T\|_1$.
- ❺ if $\mathbf{V}(t) \neq \mathbf{V}(t-1)$ then Go to Step 2
else $\mathbf{V} := \mathbf{V}(t); \mathbf{U} := \mathbf{U}(t)$.

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

Задача: $\|\mathbf{X} - \mathbf{UV}^T\|_1^2 \rightarrow \min_{\mathbf{U}, \mathbf{V}}, \quad \mathbf{U}^T \mathbf{U} = \mathbf{I}_r.$

Определение

Введем норму $\|\mathbf{E}\|_* = \sum_i \mu_i$, где μ_i — сингулярные числа матрицы \mathbf{E} .

Замечание

- 1 $\|\mathbf{E}\|_*$ — выпуклая.
- 2 Пусть $\mathbf{E} = \mathbf{UV}^T$. Тогда $\|\mathbf{E}\|_* = \|\mathbf{V}\|_*$, т.к. $\mathbf{U}^T \mathbf{U} = \mathbf{I}_r$.

Переходим к задаче:

$$\|\mathbf{X} - \mathbf{UV}^T\|_1^2 + \lambda \|\mathbf{V}\|_* \rightarrow \min_{\mathbf{U}, \mathbf{V}}, \quad \mathbf{U}^T \mathbf{U} = \mathbf{I}_r.$$

Метод был взят из статьи [Zheng, Liu, Sugimoto et al., 2012].

Реализация с исправлениями — собственная.

- Сравнение стандартного метода L2-SSA с различными параметрами, его модификации (SSA с центрированием) и треугольного фильтра (фильтра Бартлетта).
- Сравнение варианта L2SVD-L1H-SSA с проектором на пространство ганкелевых матриц в \mathbb{L}_1 с методом L2-SSA.
- Сравнение различных способов построения проектора в \mathbb{L}_1 на множество матриц ранга, не превосходящего r , между собой, а также сравнение их с базовым методом L2-SSA.

Известно (Шлемов, 2011; Golyandina, Zhigljavsky, 2013):

- Результат восстановления сигнала методом L2-SSA сводится к применению системы линейных фильтров.
- Модификация SSA с центрированием (совпадает с методом с проекцией на вектор из единиц) также представима в виде системы линейных фильтров.

Утверждение

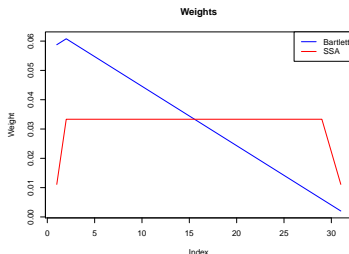
Коэффициенты фильтра, которому соответствует метод SSA с проекцией на вектор $(1, \dots, 1)^T$ с восстановлением по компоненте, соответствующей проекции, в точке \tilde{x}_m , где $1 \leq m \leq L - 1$:

$$\left(\frac{1}{(m+1)L}, \frac{2}{(m+1)L}, \dots, \frac{1}{L}, \dots, \frac{1}{L}, \dots, \frac{2}{(m+1)L}, \frac{1}{(m+1)L} \right).$$

Утверждение

Восстановление методом SSA с проекцией на подпространство, натянутое на вектор $(1, \dots, 1)^T$, по компоненте, соответствующей проекции, совпадает с применением фильтра Бартлетта с длиной окна, равной $2L - 1$ при $L \leq s \leq K$.

Коэффициенты метода SSA с проекцией не совпадают с коэффициентами фильтра Бартлетта при $1 \leq m \leq L - 1$ и $K + 1 \leq m \leq N$.



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

$$x_n = e^{n/N} + \sin(2\pi n/120 + \pi/6) + \varepsilon_n, \quad \varepsilon_n \sim N(0, 1),$$

$$y_{n_0} = x_{n_0} + \delta.$$

Выделяющиеся наблюдения находятся в точках x_5 или x_{130} .

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

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

где S — сигнал, \tilde{S} — оценка сигнала. Будем вычислять

$\text{RMSE} = \sqrt{\text{MSE}}$. Число повторов для оценок ошибок $M = 100$.

Сравнивались следующие методы:

- Полное выделение сигнала, r равно рангу сигнала
 - метод SSA с $L = 120, r = 3$ (лучшая отделимость от шума)
 - метод SSA с $L = 30, r = 3$ (отделимость хуже)
- Похожие между собой методы
 - метод SSA с $L = 30, r = 1$ (средний вариант между разделимостью и сглаживанием)
 - метод SSA с проекцией на вектор $(1, \dots, 1)^T$, $L = 30$
 - фильтр Бартлетта с длиной окна $2L - 1$, где $L = 30$
- Сглаживание
 - метод SSA с $L = 10, r = 1$

Ошибка L2-SSA в зависимости от размера выброса

δ (RMSE)	SSA120.3	SSA30.3	SSA30.1	SSAProj	Bartl	SSA10.1
0	0.018	0.027	0.139	0.143	0.136	0.034
9_{130}	0.886	0.760	0.547	0.544	0.542	0.804
9_5	0.366	1.810	0.298	0.310	0.552	0.921

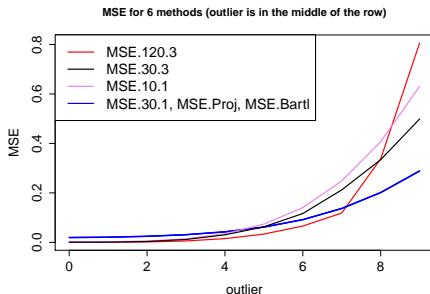


Рис.: Зависимость MSE от размера выброса, находящегося в точке x_{130} . Длина ряда $N = 240$.

Сравнение трех вариантов L1-SSA между собой и с методом L2-SSA

δ (RMSE)	L1-SSA			L2-SSA
	pcaL1	ALM	L1-regr	SSA120.3
0	0.021	0.019	0.022	0.018
9_{130}	0.022	0.020	0.064	0.886
9_5	0.024	0.023	0.031	0.366

δ (MAD)	pcaL1	ALM	L1-regr	SSA120.3
0	0.016	0.009	0.019	0.013
9_{130}	0.021	0.010	0.031	0.751
9_5	0.016	0.010	0.030	0.132

Лучшими методами оказываются:

При отсутствии выбросов, RMSE: метод L2-SSA,

MAD: L1-SSA с регуляризацией.

В присутствии выбросов, RMSE и MAD: метод L1-SSA с регуляризацией.

Разложение MSE в сумму дисперсии и квадрата смещения при разной длине окна

$$\text{MSE}(\tilde{S}, S) = \text{Var}(\tilde{S}) + (\text{Bias}(\tilde{S}, S))^2.$$

Без выделяющегося наблюдения основную часть MSE методов L2-SSA и L1-SSA составляет дисперсия.

L	Var (L2)	Bias ² (L2)	Var (L1)	Bias ² (L1)
10	0.0021	3.2e-05	0.0017	0.0003
120	0.0003	1.9e-06	0.0004	0.0001

В присутствии выброса в точке x_{130} основную часть MSE восстановления методом L2-SSA составляет смещение. Однако для L1-SSA основную часть ошибки составляет дисперсия.

L	Var ₁₃₀ (L2)	Bias ₁₃₀ ² (L2)	Var ₁₃₀ (L1)	Bias ₁₃₀ ² (L1)
10	0.0061	0.0255	0.0027	0.0003
120	0.0003	0.0054	0.0003	0.0001

- Проведено теоретическое сравнение метода SSA с проекцией и фильтра Бартлетта.
- Проведено сравнение метода L2-SSA и его модификаций. Построены рекомендации по выбору параметров метода при разной величине выброса.
- Рассмотрен промежуточный вариант L2SVD-L1H-SSA, однако эффекта устойчивости не выявлено.
- Рассмотрен метод L1-SSA, устойчивый к выбросам. Найдены и частично самостоятельно реализованы методы проекции в \mathbb{L}_1 на множество матриц ранга, не превосходящего r .
- Проведены численные эксперименты, в которых выявлена самая точная и устойчивая реализация метода L1-SSA (проблема: медленная реализация).