Singular spectrum analysis

Чуйкин Никита
Научный руководитель: Егорова Людмила Геннадьевна

October 2023

Содержание

Алгоритм SSA: агрегирование

Пусть дан одномерный временной ряд $(f_1, f_2 \dots f_n), f_i \in \mathbb{R}$. Наша цель - разложить его на компоненты $f(x) = g(x) + p(x) + \epsilon$, где $g(x)$ - тренд, $p(x)$ - сезонность, ϵ - случайная компонента.
Выберем длинну окна (*window length*) l и составим матрицу

$$S = \begin{pmatrix} f_1 & f_2 & f_3 & \dots & f_{n-l} \\ f_2 & f_3 & f_4 & \dots & f_{n-l+1} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ f_l & f_{l+1} & f_{l+2} & \dots & f_n \end{pmatrix}$$

Определение

Так построенную матрицу будем называть траекторной (*trajectory*).

$$S \in \mathbb{R}^{l \times (n-l+1)}$$

Алгоритм SSA: SVD приближение

Разложим траекторную матрицу S с помощью малорангового SVD: $S = U\Sigma V = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, u_i, v_i - векторы столбцы.

Теперь сгруппируем эти тройки (*eigentriplets*) в массивы $I_1, I_2 \dots I_k$ и обозначим $X_j = \sum_{i \in I_j} \sigma_i u_i v_i^T$. Каждая такая траекторная матрица представляет собой компоненту временного ряда (?).

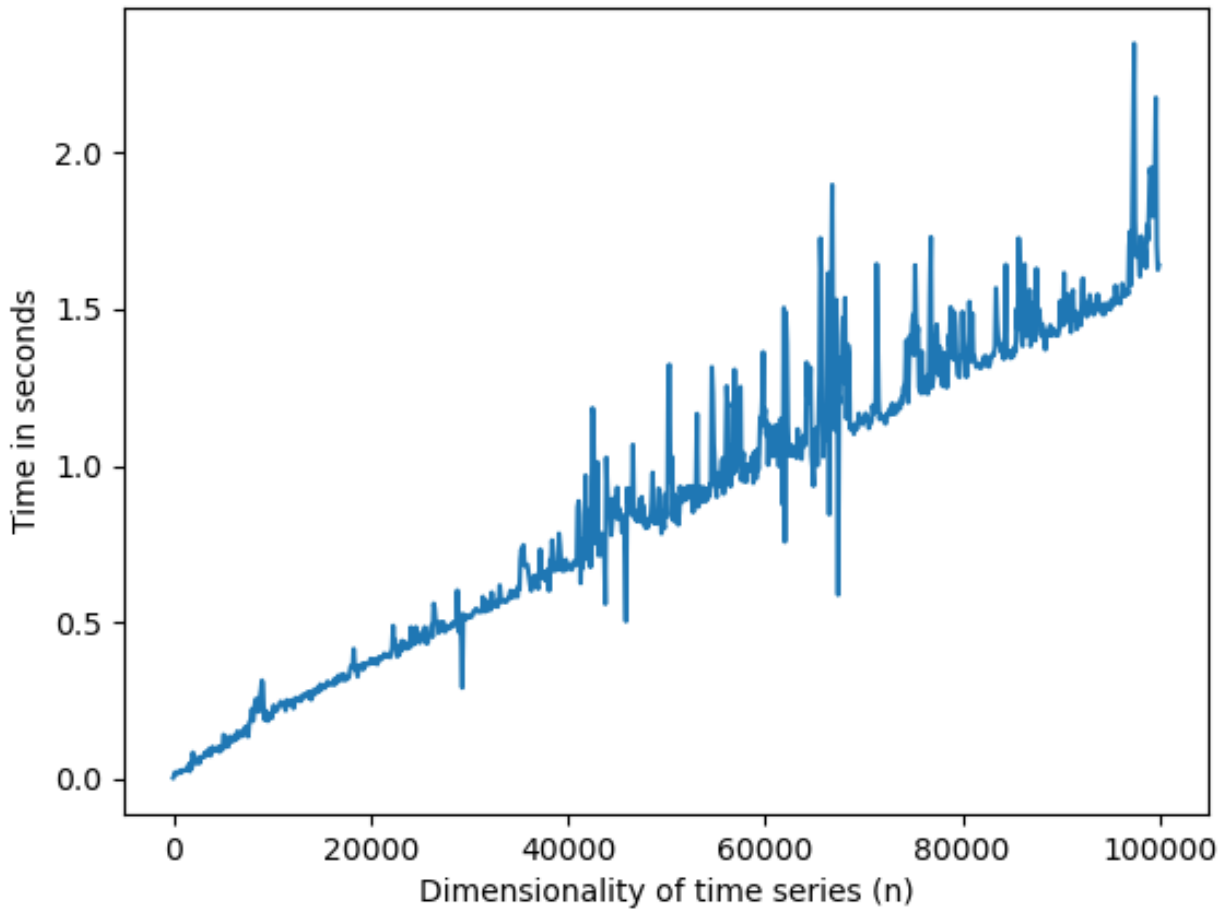
Утверждение

Скорость вычисления SVD разложения $= O(In^2)$, если $I \geq (n - I + 1)$, и $O(nI^2)$ - иначе. *(Trefethen N., Bau D. Numeric linear algebra)*

Как получить исходный ряд из матрицы S? Достаточно найти среднее по всем антидиагоналям матрицы (*henkelization*). **Свойство оптимальности**: траекторная матрица для ряда, полученного из Y генкелизацией является ближайшей с точки зрения нормы Фробениуса к Y из всех ганкелевых (*Н.Голяндина,В.Некруткин, Д.Степанов 2003*)

Настоящее время вычислений

Временной ряд длины n генерировался из нормального распределения с $l = 4$.



Обобщение для двумерного ряда

Есть несколько способов для анализа двумерных рядов:

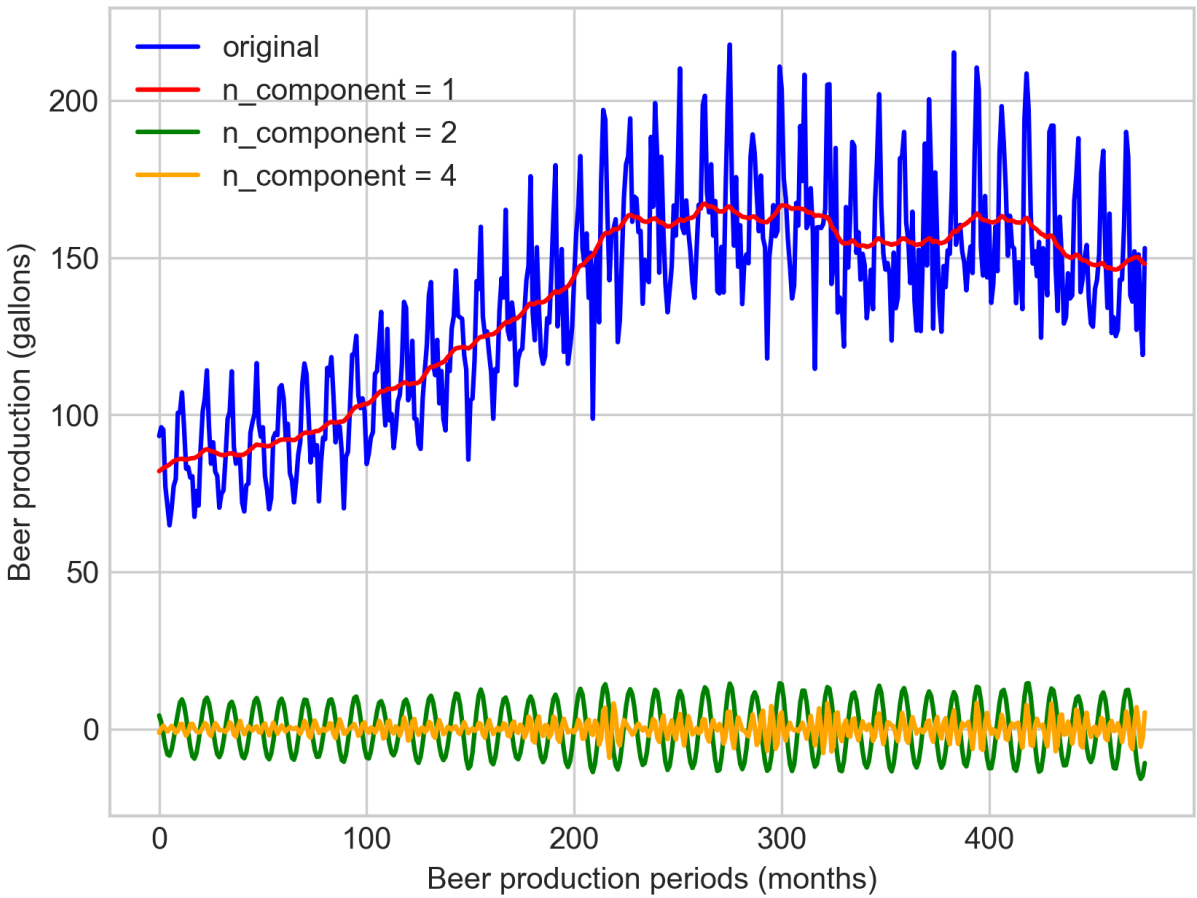
- Раскладывать ряды по отдельности.
- Составить траекторную матрицу как $S = [S_1 : S_2], S_{1,2}-$ траекторная матрица соответствующего ряда.
- Комплексное SSA. Конструируется новый временной ряд $c = (f_1 + ig_1, \ldots, f_n + ig_n)$. Для него аналогично считается SVD.

	FORT, SSA	DRY, SSA	FORT+DRY, MSSA	FORT+DRY, CSSA
1	тренд	тренд	тренд*+12*	тренд
2	12	12	12	12
3	12	12	12	4
4	4	4	4	4
5	4	4	4	6
6	6	3+6*+2.4*	тренд*+6+4+3+2.4*	3+12*+2.4*
7	6	3+6	6	12+2.4*+3*
8	2.4	2.4+3+6	6	2.4+3+12
9	2.4	3+6	3+2.4	2.4
10	3	2.4+3+6	3+2.4	3
11	3	2.4	2.4+3	6
12	шум	тренд	2.4+3	тренд
13	шум	тренд	шум	шум
14	шум	2	шум	шум

Все три способа значительно отличаются по распределению компонент рядов. **Почему?**

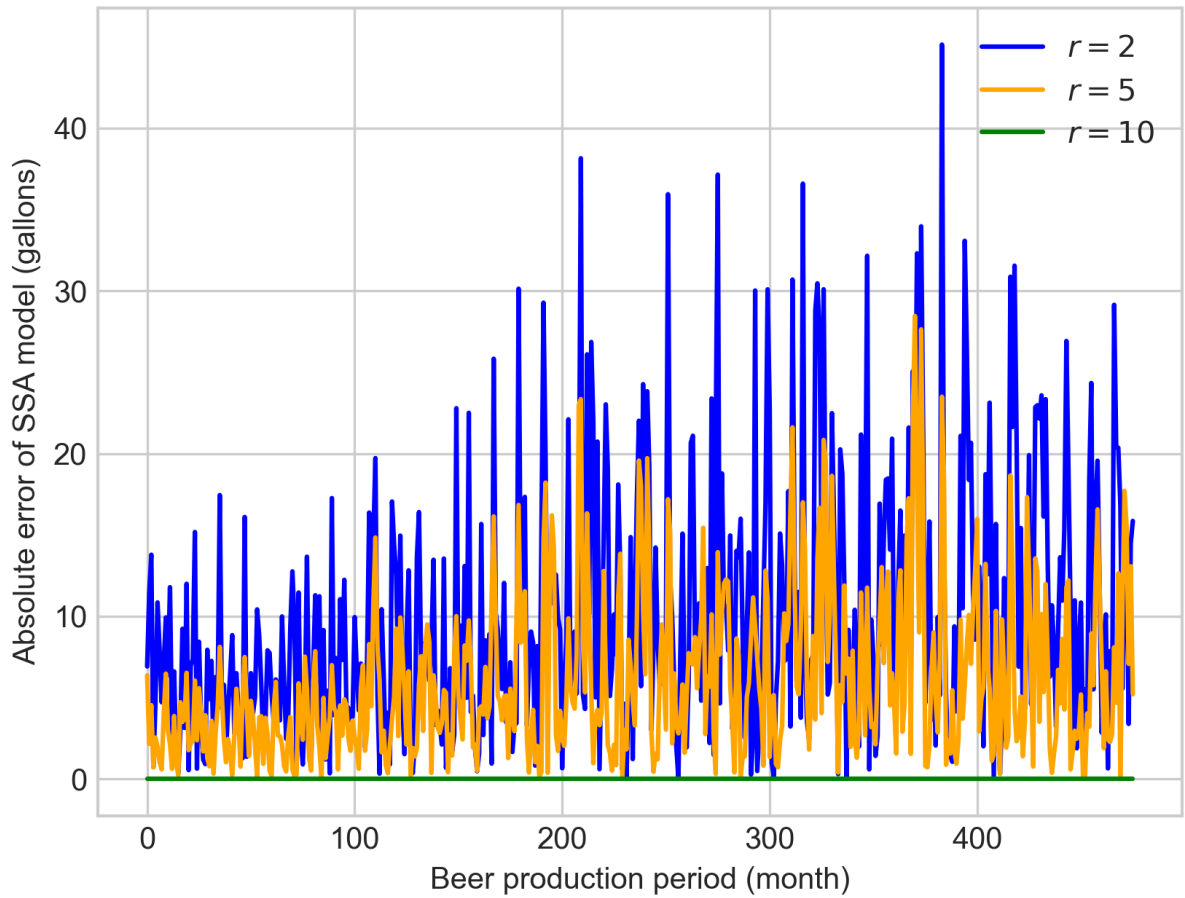
Пример на данных: как ведут себя компоненты

Данные о производстве пива в США: kaggle.com. График показывает, как ведут себя различные компоненты соответствующие 1,2 и 4 собственным тройкам для SSA с параметром $l = 10$.



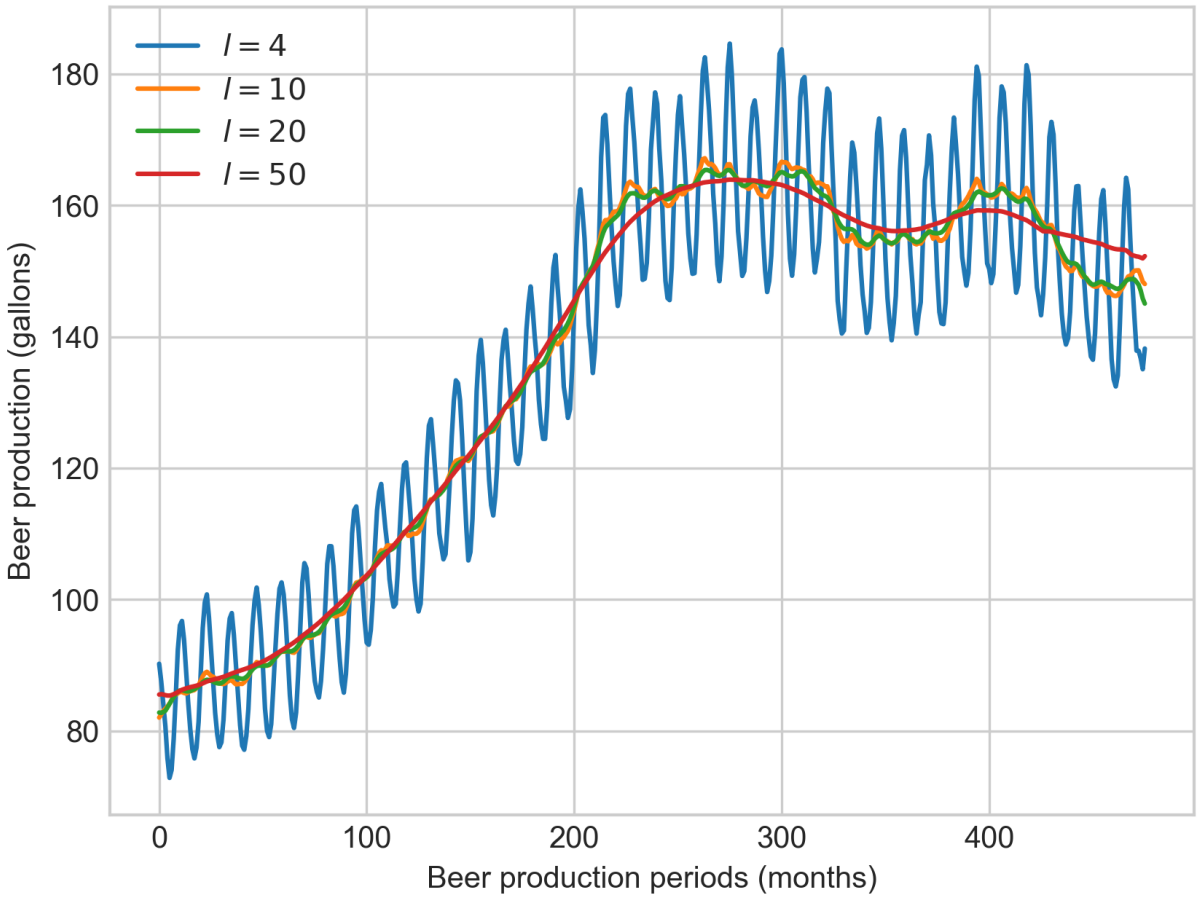
Пример на данных: как ведет себя ошибка

Теперь посмотрим на ошибку аппроксимации (r - ранг разложения).



Пример на данных: выбор длины окна

Посмотрим на то, как изменяется первая компонента разложения при изменении длинны окна.



Пример на данных: зависимость от будущего

Что если мы знаем будущее и умеем весь временной ряд. Тогда насколько будут отличаться компоненты разложения для ряда $(f_1, \ldots, f_k)$, где k - заданная точка отсечки?
Здесь должна была быть гифка
Оказывается, что по крайней мере первая компонента меняется лишь на первых и последних наблюдениях из временного ряда. **Почему так происходит?**

Как группировать траекторные матрицы?

Определение

Временные ряды $x_1, x_2 \in \mathbb{R}^n$, полученные из траекторных матриц $X_1, X_2 \in \mathbb{R}^{l \times n-l+1}$ слабо L-разделимы, если пространства столбцов (*column spaces*) матриц X_1 и X_2 ортогональны.

Утверждение

Пусть $x_1, x_2 \in \mathbb{R}^n, x = x_1 + x_2$ и x_1, x_2 слабо L-разделимы. Тогда для ряда x существует такое синглуряное разложение траекторной матрицы X, что его можно разбить на две части, являющиеся траекторными матрицами x_1, x_2

Но $L-$разделимость слишком сильное требование, которое нуждается в ослаблениях. Другие возможные варианты: автокорреляция, близость сингулярных значений, прокси-показатель ортогональности (например, скалярное произведение), ассимптотическая разделимость.

Области применения SSA

- Очистка данных от шума
- Интерполяция
- Предсказания на основе линейной рекуррентной формулы $f_k = \sum_{i=n-l}^{k-1} a_i f_i$
- Выделение отдельных компонент ряда

Проблемы для исследования

1. Вопросы разделимости компонент ряда и траекторных матриц
2. Подбор параметра l
3. Применение метода для рядов различной частоты на финансовых данных
4. Сравнение предсказательной силы с другими подходами

Источники

- Сайт, посвященный SSA
- Broomhead D.S., Ging G.P. *Extracting qualitative dynamics from experimental data, 1986*
- Golyandina N., Nekrutkin V., Zhigljavsky A. *Analysis of Time Series Structure, 2001*
- Голяндина Н., Некруткин В., Степанов Д. *Варианты метода «Гусеница»-SSA для анализа многомерных временных рядов, 2003*
- Вохмянин С. *Метод "Гусеница-SSA"как инструмент прогнозирования состояния финансового рынка, 2010*
- Голяндина Н. *Метод "ГусеницаSSA: анализ временных рядов, 2004*

Спасибо за внимание!