Singular spectrum analysis (Метод Гусеницы)

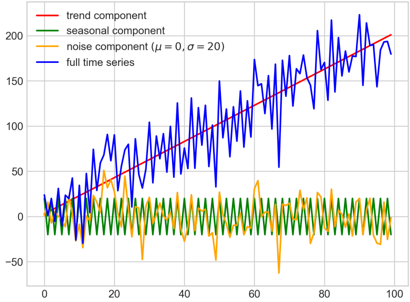
Чуйкин Никита
Научный руководитель: Егорова Людмила Геннадьевна

November 2023

Содержание

Алгоритм SSA: введение

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.



Попробуем получить информацию о корреляции частей временного ряда друг с другом.

Алгоритм SSA: агрегирование

Выберем длину окна (*window length*) l, будем двигать этот вектор по временному ряду 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*).

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

Все траекторную матрицы имеют **ганкелеву форму**, т.е. содержат одинаковые элементы на побочных диагоналях.

Иллюстративный пример

Пусть зависимость описывается простой линейной формулой $y = 2x, x \in 0, \dots 4$. Тогда временной ряд $y = (0, 2, 4, 6, 8)$. Составим траекторную матрицу для длины окна $l = 2$:

$$S = \begin{pmatrix} 0 & 2 & 4 & 6 \\ 2 & 4 & 6 & 8 \end{pmatrix}$$

SVD факторизация

Определение

SVD (singular value decomposition) матрицы $A \in \mathbb{R}^{n \times m}$ - это представление матрицы в форме $A = U \Sigma V$, где $U \in \mathbb{R}^{n \times n}$ и $V \in \mathbb{R}^{m \times m}$ - ортогональные матрицы, а $\Sigma \in \mathbb{R}^{n \times m}$ - диагональная матрица.

На диагонали матрицы Σ в невозрастающем порядке стоят сингулярные числа $\sigma_1, \sigma_2 ... \sigma_r$, где r - ранг матрицы A. Впредь мы будем использовать только низкоранговое SVD для заданного r, i.e. "отрезать"последние столбцы матрицы U и строки матрицы V.

SVD факторизация: свойства

Определение

Пусть $A \in \mathbb{R}^{n \times n}$. Нормой Фробениуса будем называть

$$||A||_2 = \sqrt{\sum_{i,j} a_{ij}^2}$$

Теорема Эккарта-Янга

Матрица, построенная через малоранговое SVD является ближайшей среди матриц ранга r к матрице A по норме Фробениуса.

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^T - 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 (?).

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

Иллюстративный пример: продолжение

$$S = \begin{pmatrix} 0 & 2 & 4 & 6 \\ 2 & 4 & 6 & 8 \end{pmatrix}$$

SVD полного ранга будет иметь вид

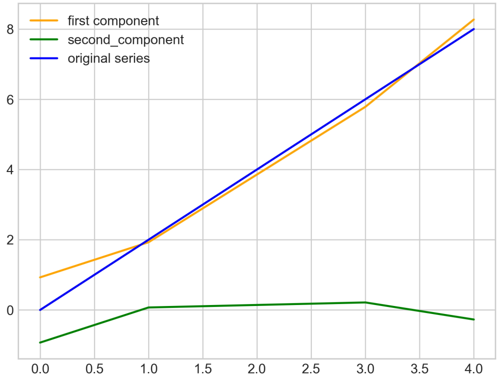
$$S \approx \begin{pmatrix} -0.58 & 0.82 \\ 0.82 & -0.58 \end{pmatrix} \begin{pmatrix} 12.8 & 0 \\ 0 & 1.42 \end{pmatrix} \begin{pmatrix} -0.13 & 0.35 & -0.5 & -0.78 \\ -0.82 & -0.49 & 0.26 & 0.18 \end{pmatrix}$$

$$u_1 = \begin{pmatrix} -0.58 \\ 0.82 \end{pmatrix}, v_1 = \begin{pmatrix} -0.13 & 0.35 & -0.5 & -0.78 \end{pmatrix}$$

$$u_2 = \begin{pmatrix} 0.82 \\ -0.58 \end{pmatrix}, v_2 = \begin{pmatrix} -0.82 & -0.49 & 0.26 & 0.18 \end{pmatrix}$$

Иллюстративный пример: сборка

$$y_1 = \text{Hankelization}(\sigma_1 u_1 v_1) = (0.94, 1.94, 3.65, 5.5, 8.15)$$
$$y_2 = \text{Hankelization}(\sigma_2 u_2 v_2) = (-0.94, -0.06, 0.35, 0.5, -0.15)$$



Алгоритм SSA: резюме

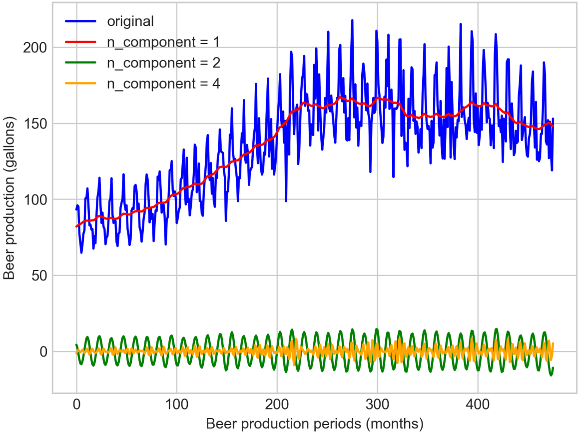
1. Составляем траекторную матрицу исходного временного ряда $(f_1, f_2, \ldots, f_n)$
2. Получаем SVD разложение и группируем полученные компоненты сингулярного разложения
3. В каждой группе складываем траекторные матрицы и проводим ганкелизацию для получения временного ряда

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

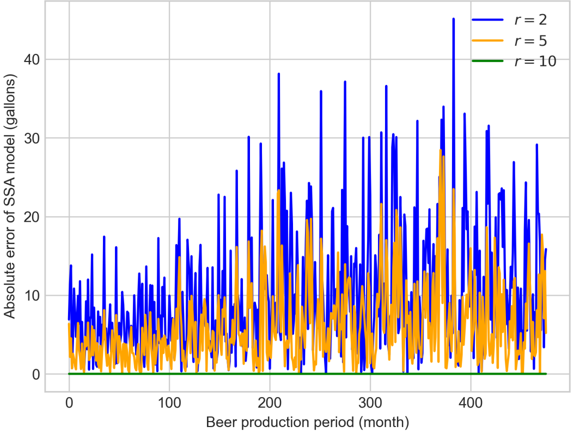
Пример на данных: как ведут себя компоненты

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 $l = 10$.



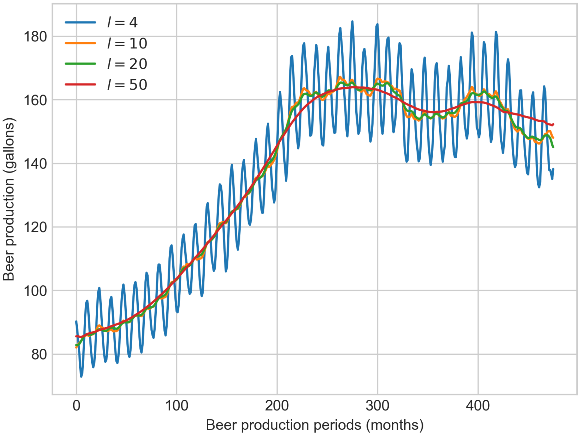
Пример на данных: как ведет себя ошибка

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?**

Скорость вычислений

Посчитаем наихудшее время исполнения ($l = \frac{n}{2}$) с помощью прямых вычислений. Подбробнее - (*Anton Korobeynikov,2010*).

- **Агрегирование:** незначительно
- **SVD:** $O(n^3)$
- **Группировка:** $O(n^3)$
- **Генкелизация:** $O(n^3)$ сложений и $O(n^2)$ умножений

Утверждение (*Anton Korobeynikov,2010*)

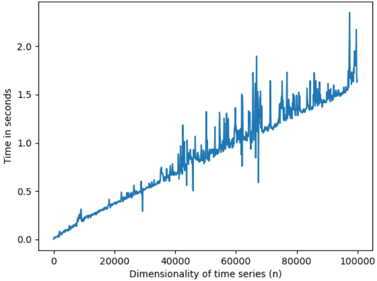
Если требуется найти r собственных троек, то худшее время исполнения можно сократить до $O(rn \log n)$. Такой алгоритм даёт прирост скорости при достаточно больших $n \gg 0$

Скорость вычислений: пример

Утверждение

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

Временной ряд длины n генерировался из нормального распределения с $l = 4$. Скорость всего алгоритма SSA так же оказалась линейна для случая $l \leq (n - l + 1)$.



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*) и строк (*row 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.

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 "Гусеница SSA": анализ временных рядов, 2004*
- Korobeynikov A. *Computation- and Space-Efficient Implementation of SSA, 2010*

Спасибо за внимание!