Метод анализа временных рядов «Гусеница» или SSA (анализ сингулярного спектра)

Метод основан на преобразовании одномерного временного ряда в многомерный ряд с последующим применением к полученному многомерному временному ряду метода главных компонент.

«Гусеница»-SSA – универсальный метод для решения задач общего назначения, таких как выделение тренда, обнаружение периодичностей, корректировка на сезонность, сглаживание, подавление шума (Golyandina et al, 2001)

Singular spectrum analysis (SSA) – метод спектрального анализа стационарных временных рядов (Vautard and Ghil, 1989).

В зависимости от специфики временных рядов и выбора параметров, многие проблемы, связанные с аддитивным расширением временных рядов, могут быть решены с помощью метода «Caterpillar»-SSA. Среди прочего можно отметить:

- Поиск трендов разного разрешения;
- Сглаживание;
- Извлечение сезонных составляющих;
- Одновременное извлечение циклов с малыми и большими периодами;
- Выделение периодичностей с разной амплитудой;
- Одновременное извлечение сложных трендов и периодичностей;
- Поиск структуры в коротких временных рядах.

1. Базовый алгоритм метода «Гусеница»-SSA

Излагается по Голяндина Н.Э. Метод «Гусеница»-SSA: анализ временных рядов: Учеб. пособие. СПб., 2004. – 76 с.

Пусть $N > 2$. Рассмотрим вещественнозначный временной ряд $F = (f_0, \ldots, f_{N-1})$ длины N. Будем предполагать, что ряд F — ненулевой, т. е. существует, по крайней мере, одно i, такое что $f_i \neq 0$. Обычно считается, что $f_i = f(i\Delta)$ для некоторой функции $f(t)$, где t — время, а Δ — некоторый временной интервал, однако это не будет играть особой роли в дальнейшем. Более того, числа $0, \ldots, N-1$ могут быть интерпретированы не только как дискретные моменты времени, но и как некоторые метки, имеющие линейно-упорядоченную структуру.

Нумерация значений временного ряда начинается с $i = 0$, а не стандартно с $i = 1$ только из-за удобства обозначений.

Базовый алгоритм состоит из двух дополняющих друг друга этапов, разложения и восстановления.

1.1. Первый этап: разложение

Шаг 1. Вложение

Процедура вложения переводит исходный временной ряд в последовательность многомерных векторов.

Пусть L — некоторое целое число (*длина окна*), $1 < L < N$. Процедура вложения образует $K = N - L + 1$ *векторов вложения*

$$X_i = (f_{i-1}, \ldots, f_{i+L-2})^{\mathrm{T}}, \quad 1 \le i \le K,$$

имеющих размерность L. Если нам нужно будет подчеркнуть размерность X_i, то мы будем называть их векторами L-вложения.

L-Траекторная матрица (или просто *траекторная матрица*) ряда F

$$\mathbf{X} = [X_1 : \ldots : X_K]$$

состоит из векторов вложения в качестве столбцов.

Другими словами, траекторная матрица — это матрица

$$\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_K \\ f_2 & f_3 & f_4 & \cdots & f_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1} \end{pmatrix}. \qquad (1)$$

Очевидно, что $x_{ij} = f_{i+j-2}$ и матрица $\mathbf{X}$ имеет одинаковые элементы на «диагоналях» $i + j = \text{const}$. Таким образом, траекторная матрица является *ганкелевой*. Существует взаимно-однозначное соответствие между ганкелевыми матрицами размерности $L \times K$ и рядами длины $N = L + K - 1$.

Параметр L выбирается произвольно.

Однако, дадим несколько рекомендаций по выбору длины гусеницы.

- Сингулярные разложения одного и того же ряда длины n, соответствующие выбору длины гусеницы l и $n - l + 1$ эквивалентны. Следовательно, для анализа структуры временного ряда не имеет смысла брать длину гусеницы, большую чем половина длины ряда.
- Чем больше длина гусеницы, тем более детальным получается разложение исходного ряда. Таким образом, наиболее детальное разложение достигается при выборе длины гусеницы, приблизительно равной половине длины ряда ($l \sim n/2$). Причем, чем больше длина гусеницы, тем более детальным получается разложение исходного ряда.
- Маленькая длина гусеницы может привести к смешиванию интерпретируемых компонент ряда.
- При решении задачи выделения периодической компоненты с периодом τ следует выбирать длину гусеницы l кратной τ.
- В общем метод гусеницы устойчив относительно изменения длины гусеницы. Эффект проявляется не столько в количественном, сколько в качественном смысле.

Предположим, что исходный временной ряд является суммой нескольких рядов.

Следующий шаг – сингулярное разложение траекторной матрицы в сумму элементарных матриц. Каждая элементарная матрица задается набором из собственного числа и двух сингулярных векторов.

Шаг 2. Сингулярное разложение

Результатом этого шага является сингулярное разложение (SVD = Singular Value Decomposition) траекторной матрицы ряда.

Пусть $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$. Обозначим $\lambda_1, \ldots, \lambda_L$ *собственные числа* матрицы $\mathbf{S}$, взятые в неубывающем порядке ($\lambda_1 \geq \ldots \geq \lambda_L \geq 0$) и $U_1, \ldots, U_L$ — ортонормированную систему *собственных векторов* матрицы $\mathbf{S}$, соответствующих собственным числам.

Пусть $d = \max\{i : \lambda_i > 0\}$. Если обозначить $V_i = \mathbf{X}^{\mathrm{T}} U_i / \sqrt{\lambda_i}$, $i = 1, \ldots, d$, то сингулярное разложение матрицы $\mathbf{X}$ может быть записано как

$$\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d, \tag{2}$$

где $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$. Каждая из матриц $\mathbf{X}_i$ имеет ранг 1. Поэтому их можно назвать *элементарными матрицами*.

Набор $(\sqrt{\lambda_i}, U_i, V_i)$ мы будем называть i-й *собственной тройкой* сингулярного разложения (2).

1.2. Второй этап: восстановление

Шаг 3. Группировка

На основе разложения (2) процедура группировки делит все множество индексов $\{1, \ldots, d\}$ на m непересекающихся подмножеств $I_1, \ldots, I_m$.

Пусть $I = \{i_1, \ldots, i_p\}$. Тогда *результирующая матрица* $\mathbf{X}_I$, соответствующая группе I, определяется как

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

Такие матрицы вычисляются для $I = I_1, \ldots, I_m$, тем самым разложение (2) может быть записано в сгруппированном виде

$$\mathbf{X} = \mathbf{X}_{I_1} + \ldots + \mathbf{X}_{I_m}. \tag{3}$$

Процедура выбора множеств $I_1, \ldots, I_m$ и называется *группировкой собственных троек*.

Шаг 4. Диагональное усреднение

На последнем шаге базового алгоритма каждая матрица сгруппированного разложения (3) переводится в новый ряд длины N.

Пусть $\mathbf{Y}$ — некоторая $L \times K$ матрица с элементами y_{ij}, где $1 \le i \le L$, $1 \le j \le K$. Положим $L^* = \min(L, K)$, $K^* = \max(L, K)$ и $N = L + K - 1$. Пусть $y^*_{ij} = y_{ij}$, если $L < K$, и $y^*_{ij} = y_{ji}$ иначе. *Диагональное усреднение* переводит матрицу $\mathbf{Y}$ в ряд $g_0, \ldots, g_{N-1}$ по формуле

$$g_k = \begin{cases} \dfrac{1}{k+1} \displaystyle\sum_{m=1}^{k+1} y^*_{m,k-m+2} & \text{для } 0 \le k < L^* - 1, \\ \dfrac{1}{L^*} \displaystyle\sum_{m=1}^{L^*} y^*_{m,k-m+2} & \text{для } L^* - 1 \le k < K^*, \\ \dfrac{1}{N-k} \displaystyle\sum_{m=k-K^*+2}^{N-K^*+1} y^*_{m,k-m+2} & \text{для } K^* \le k < N. \end{cases} \qquad (4)$$

Выражение (4) соответствует усреднению элементов матрицы вдоль «диагоналей» $i + j = k + 2$: выбор $k = 0$ дает $g_0 = y_{11}$, для $k = 1$ получаем $g_1 = (y_{12} + y_{21})/2$ и т. д. Заметим, что если матрица $\mathbf{Y}$ является траекторной матрицей некоторого ряда $(h_0, \ldots, h_{N-1})$ (другими словами, если матрица $\mathbf{Y}$ является ганкелевой), то $g_i = h_i$ для всех i.

Применяя диагональное усреднение (4) к результирующим матрицам $\mathbf{X}_{I_k}$, мы получаем ряды $\widetilde{F}^{(k)} = (\widetilde{f}^{(k)}_0, \ldots, \widetilde{f}^{(k)}_{N-1})$, и, следовательно, исходный ряд $(f_0, \ldots, f_{N-1})$ раскладывается в сумму m рядов:

$$f_n = \sum_{k=1}^{m} \widetilde{f}^{(k)}_n. \qquad (5)$$

На практике матрицу X можно представить в виде суммы двух матриц, одна из которых – сглаженная матрица, а другая является шумовой составляющей.

То есть множество индексов разбивают на два подмножества I_1 и I_2. Так как компоненты упорядочены в порядке убывания, компоненты с меньшим номером вносят больший вклад в разложение ряда.