Тензорная декомпозиция и прогноз для набора временных рядов

Сёмкин Кирилл

Московский Физико-Технический Институт

2024

Мотивация

Проблема

Известные модели многомерных временных рядов (e.g. RNN, VAR) не позволяют строить их аддитивные декомпозиции. Методы разложения одномерных сигналов (e.g. Trend-Cycle Estimation, Seasonal Adjustment Methods) не учитывают взаимозависимость рядов.

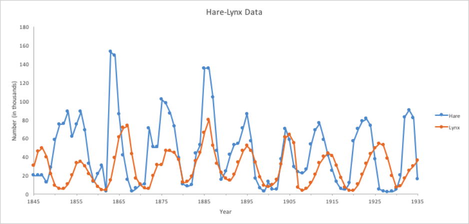

Рис. 1: График количества особей хищников и травоядных от времени

Цель исследования

Предложить метод, позволяющий выделить общую для набора сигналов структуру. На её основании произвести разложение каждого сигнала на компоненты. Найти способ построения прогноза.

Решение

tSSA = модель собственного пространства сигнала + тензорное представление данных + CPD

Данный подход опирается на гипотезу порождения динамической системой и является расширением метода SSA.

Постановка задачи

Динамическая система в дискретном времени:

$$\begin{cases} \mathbf{y}(t+1) = f(\mathbf{y}(t)), \ \ t \in \mathbb{N} \\ \mathbf{y}(0) = \mathbf{y}_0 \end{cases}$$

$\mathbf{y} \in X$, где X — гладкое многообразие большой размерности.

Конкретная траектория этой системы *порождает* наблюдаемые ряды через некое многомерное отображение $\varphi : X \to \mathbb{R}^m$:

$$\varphi(\mathbf{y}(t)) = \mathbf{x}_t \Leftrightarrow \begin{cases} \varphi_1(\mathbf{y}(t)) = x_1(t) \\ \dots \\ \varphi_m(\mathbf{y}(t)) = x_m(t) \end{cases}$$

Постановка задачи

Выдвигается гипотеза, что траектории $\mathbf{y}(t)$ лежат в многообразии $M \subset X$ размерности меньшей, чем у X. Ставится задача поиска вложения (embedding) M в $\mathbb{R}^L$ для некоторого L и нахождения базиса в образе этого вложения.

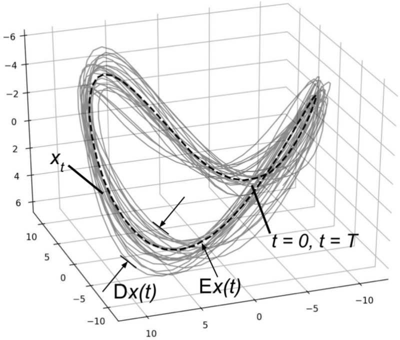

Рис. 2: Динамическая система собственной размерности < 3

Поиск решения

Theorem (Такенс)

Искомое вложение даётся построением соответствующего вектора задержки

$$\mathbf{y}(t) \xrightarrow{emb} \begin{pmatrix} \varphi \circ f^{t-L+1}(\mathbf{y}(t)) \\ \vdots \\ \varphi \circ f(\mathbf{y}(t)) \\ \varphi \circ \mathbf{y}(t) \end{pmatrix} = \begin{pmatrix} x(t-L+1) \\ \vdots \\ x(t-1) \\ x(t) \end{pmatrix} = \overleftarrow{\mathbf{x}}_t \in \mathbb{R}^L$$

Полученное пространство вложения $\mathrm{Lin}(\{\overleftarrow{\mathbf{x}}_t\})$ и есть *собственное пространство сигнала* $x(t)$.

С помощью разложения *траекторной матрицы* $\mathrm{H}_x = [\overleftarrow{\mathbf{x}}_1 \dots \overleftarrow{\mathbf{x}}_{N-L+1}]$. выделяем в нём базис. Далее можем декомпозировать и строить прогноз (SSA).

Метод tSSA

В случае нескольких рядов упакуем их вектора задержек в *матрицы задержек* $(\overleftarrow{\mathbf{x}}_{1_t} \dots \overleftarrow{\mathbf{x}}_{m_t}) := \overleftarrow{\mathrm{X}}_t$. Состыкуем их в *траекторный тензор* $\mathbf{T}$.

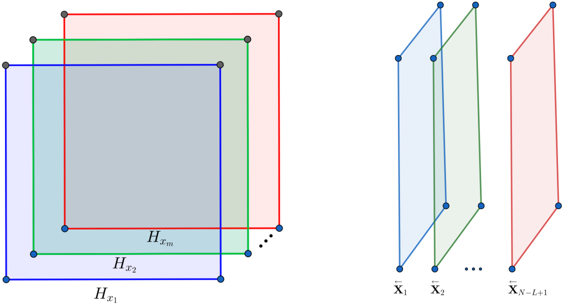

Рис. 3: Два вида на $\mathbf{T}$. Слева — в виде набора траекторных матриц сигналов. Справа — в виде набора матриц задержки.

Метод tSSA

Применим *CPD-разложение* к траекторному тензору и рассмотрим его вид для каждого сечения по третьему измерению:

$$\mathbf{T} = \sum_{i=1}^{r} \mathrm{a}_i \otimes \mathrm{b}_i \otimes \mathrm{c}_i \Leftrightarrow \begin{cases} \mathrm{H}_{x_1} = \sum^{r} \boldsymbol{\sigma}_{x_1}(i) \cdot \mathrm{a}_i \mathrm{b}_i^{\mathrm{T}} \\ \mathrm{H}_{x_2} = \sum^{r} \boldsymbol{\sigma}_{x_2}(i) \cdot \mathrm{a}_i \mathrm{b}_i^{\mathrm{T}} \\ \dots \\ \mathrm{H}_{x_m} = \sum^{r} \boldsymbol{\sigma}_{x_m}(i) \cdot \mathrm{a}_i \mathrm{b}_i^{\mathrm{T}} \end{cases} \tag{1}$$

Получили разложение траекторных матриц сигналов по *общему базису*, что выражает предположение взаимосвязанности рядов.

Декомпозиция сигналов

Для получения декомпозиции факторы разложения H_{x_k} разбиваем на группы и *ганкелизуем* — усредняем матрицы по антидиагоналям.

$$\mathrm{H}_{x_k} = \sum_{i=1}^{r} \boldsymbol{\sigma}_{x_k}(i) \cdot \mathrm{a}_i \mathrm{b}_i^{\mathrm{T}} = \sum_{i \in \mathbb{I}_1} \boldsymbol{\sigma}_{x_k}(i) \cdot \mathrm{a}_i \mathrm{b}_i^{\mathrm{T}} + \ldots + \sum_{i \in \mathbb{I}_s} \boldsymbol{\sigma}_{x_k}(i) \cdot \mathrm{a}_i \mathrm{b}_i^{\mathrm{T}} =$$
$$= C_1 + \ldots + C_s = \mathrm{Hankel}(C_1) + \ldots + \mathrm{Hankel}(C_s) \Leftrightarrow x_k(t) = c_1(t) + \ldots c_s(t)$$

Проблема

Хочется группировать факторы так, что каждая матрица C_i была как можно более 'ганкелевой'.
Мера 'ганкелевости' — *невязка ганкелизации*:

$$\mathrm{r}(\mathrm{H}) = \|\mathrm{H} - \mathrm{Hankel}(\mathrm{H})\|_F^2$$

Декомпозиция сигналов

Будем искать разбиение факторов, наилучшее в плане средней ганкелевой невязки: $\frac{1}{s}\sum\limits_{j=1}^{s} \mathrm{r}(C_s) \to \min$.

Полученная задача дискретной оптимизации не имеет быстрого алгоритма поиска решения. В работе предложена эвристическая процедура на основе дихотомического разделения групп.

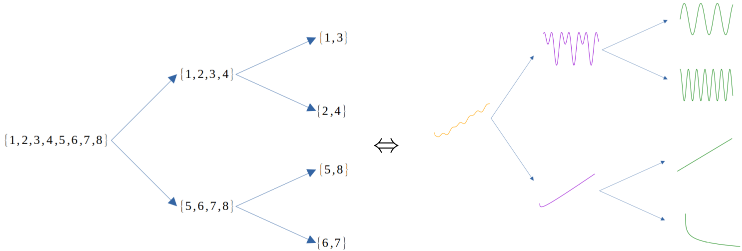

Рис. 4: Пример рекурсивного разбиения факторов (в виде индексов) на 4 группы и соответствующая декомпозиция сигнала.

Прогнозирование

Базис в пространстве векторов задержек сигнала даётся как $\mathrm{Lin}(\{\mathrm{a}_i\}) \Leftrightarrow A = [\mathrm{a}_1 \ldots \mathrm{a}_r]$.

Прогноз на один шаг вперёд сводится к решению частично неизвестной СЛАУ:

$$\overleftarrow{\mathbf{x}}_{N+1} = A\boldsymbol{\lambda} \Leftrightarrow \begin{cases} \mathrm{x}_{kn} = A_{kn}\boldsymbol{\lambda} \\ x(N+1) = \mathrm{a}_{pr}^{\mathrm{T}}\boldsymbol{\lambda} \end{cases}, \text{ где}$$

$$A = \begin{pmatrix} A_{kn} \\ \hline \mathrm{a}_{pr}^{\mathrm{T}} \end{pmatrix}$$

$$\overleftarrow{\mathbf{x}}_{N+1} = (\mathrm{x}_{kn},\ x(N+1))^{\mathrm{T}}$$

Из-за переопределённости системы, решение выражается в смысле наименьших квадратов:

$$x(N+1) = \mathrm{a}_{pr}^{\mathrm{T}}(A_{kn}^T A_{kn})^{-1} A_{kn}^T \mathrm{x}_{kn}$$

Эксперимент

Цель:

- сравнить качество разложения набора сигналов по *невязке ганкелизации* с похожим методом mSSA.
- сравнить качество построенного прогноза по метрикам *MSE*, *MAPE* с моделями RNN, VAR, mSSA

Рассматриваемые данные:

- план выработки электричества на день и его цена производства на МВт
- погодные условия в TODO

Электричество. Декомпозиция

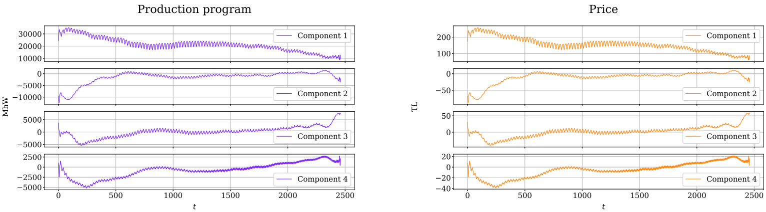

Рис. 5: Декомпозия методом tSSA на 4 компоненты через дихотомию. Вид компонент для сигналов идентичный. Наблюдается основной тренд убывания, и три тренда на возрастание. Шумовую часть метод не извлёк.

Электричество. Декомпозиция

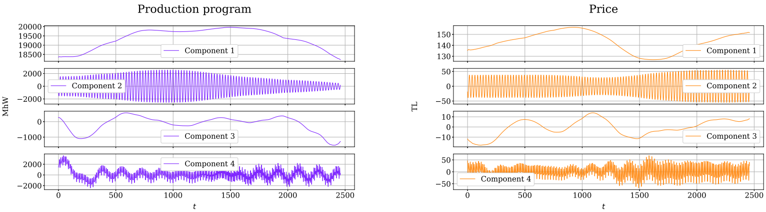

Рис. 6: Декомпозия методом mSSA на 4 компоненты подбором групп на основе близости сингулярных чисел. Полученные разложения имеют различный вид для сигналов. Выделены основные тренды и три низкоамплитудных осциллирующих сигнала. Последняя компонента содержит в себе шум, остальные от него очищены.

Электричество. Прогноз

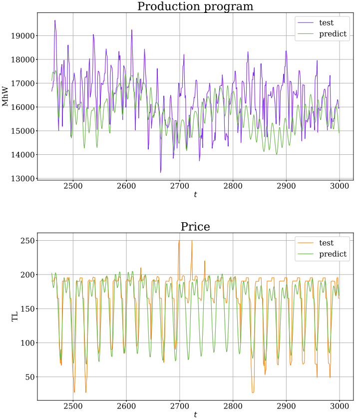

Рис. 7: Прогноз tSSA на тестовой части рядов

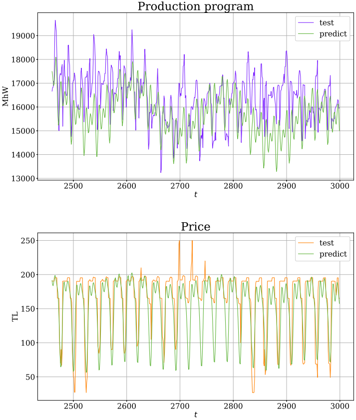

Рис. 8: Прогноз mSSA на тестовой части рядов

Электричество. Прогноз

Рис. 9: Прогноз VAR на тестовой части рядов

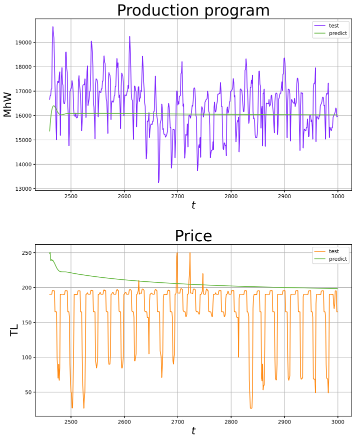

Рис. 10: Прогноз RNN на тестовой части рядов

Электричество. Итоги

Таблица 1: Сравнение методов по качеству декомпозиции сигналов.

Метод	tSSA	mSSA
Средняя невязка ганкелизации	0.802	0.309

Таблица 2: Сравнение методов по качеству прогноза производства электричества.

Метод	tSSA	mSSA	VAR	RNN
MSE	$1.56 \cdot 10^6$	$1.50 \cdot 10^6$	$7.81 \cdot 10^6$	-
MAPE	0.059	0.059	0.13	-

Таблица 3: Сравнение методов по качеству прогноза цены электричества.

Метод	tSSA	mSSA	VAR	RNN
MSE	$1.03 \cdot 10^3$	$1.03 \cdot 10^3$	$4.85 \cdot 10^3$	-
MAPE	0.18	0.17	0.36	-