

Тензорная декомпозиция и прогноз для набора временных рядов

Сёмкин Кирилл // Московский физико-технический институт //
Кафедра интеллектуальных систем // Научный руководитель:
д.ф.-м.н. Стрижов Вадим

2024

Мотивация

Проблема

Обработка многомерных временных рядов влечёт необходимость учёта взаимосвязей сигналов. Нейросетевые и статистические методы имеют большое количество параметров, и требуют изощрённых процедур обучения. При этом они не предлагают способа построения декомпозиции.

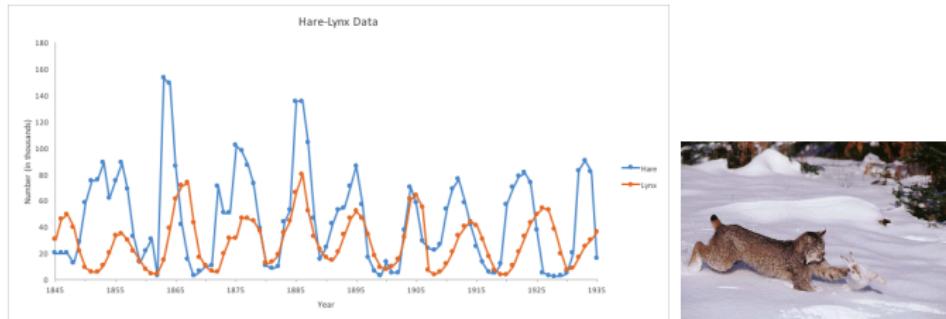


Рис. 1: График количества особей хищников и жертв от времени

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

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 $\varphi : X \rightarrow \mathbb{R}^m$:

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

Постановка задачи

Выдвигается гипотеза, что траектории $y(t)$ 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.

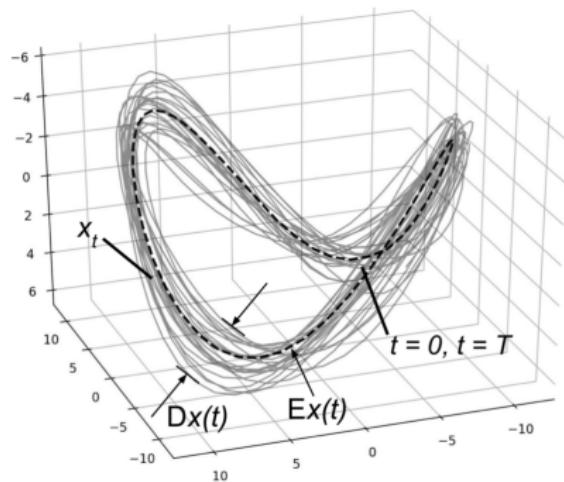


Рис. 2: Динамическая система собственной размерности < 3

Поиск решения

Theorem (Такенс)

Искомое вложение даётся построением соответствующего вектора задержки

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

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

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

Метод tSSA

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

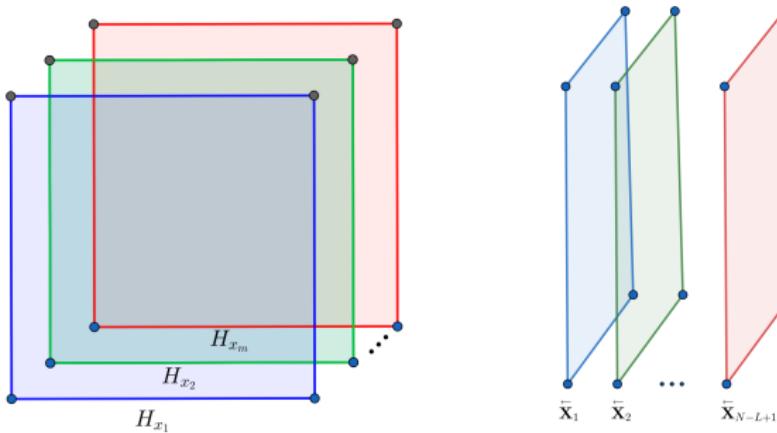


Рис. 3: Два вида на \mathbf{T} . Слева — в виде набора траекторных матриц сигналов. Справа — в виде набора матриц задержки.

Метод tSSA

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

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

Получили разложение траекторных матриц сигналов по *общему базису*, что и выражает взаимосвязанность рядов.

Декомпозиция сигналов

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

$$\begin{aligned} H_{x_k} &= \sum_{i=1}^r \sigma_{x_k}(i) \cdot a_i b_i^T = \sum_{i \in \mathbb{I}_1} \sigma_{x_k}(i) \cdot a_i b_i^T + \dots + \sum_{i \in \mathbb{I}_s} \sigma_{x_k}(i) \cdot a_i b_i^T = \\ &= C_1 + \dots + C_s = \text{Hankel}(C_1) + \dots + \text{Hankel}(C_s) \Leftrightarrow x_k(t) = c_1(t) + \dots + c_s(t) \end{aligned}$$

Проблема

Хочется группировать факторы так, что каждая матрица C_i была как можно более «ганкелевой». Это бы усилило связь разложения траекторной матрицы и самого сигнала.

Оптимальная декомпозиция

Обозначим невязку $H_i = \sigma_{x_k}(i) \cdot a_i b_i^T - Hankel(\sigma_{x_k}(i) \cdot a_i b_i^T)$, тогда из предыдущего выражения следует:

$$H_1 + \dots + H_r = 0 \Leftrightarrow H_r = - \sum_{j=1}^{r-1} H_j$$

Раскладываем на две группы. В каждой хотя бы две невязки, вместе суммирующиеся в ноль. Т.о. условие оптимальной декомпозиции:

$$\begin{cases} \sum_{j=1}^{r-1} \beta_j H_j = 0 \\ \beta_j \in \{0, 1\}, \forall j \in 1, \dots, r \\ \sum_{i=1}^{r-1} \beta_j \geq 2 \end{cases} \Rightarrow \begin{cases} \|\Lambda \beta\| \rightarrow \min_{\beta} \\ \beta_j \in \{0, 1\}, \forall j \in 1, \dots, r \\ \sum_{i=1}^{r-1} \beta_j \geq 2 \end{cases}$$

Оптимальная декомпозиция

Раскладываем на две группы. В каждой хотя бы две невязки, вместе суммирующиеся в ноль. Т.о. условие оптимальной декомпозиции:

$$\begin{cases} \|\Lambda\beta\| \rightarrow \min_{\beta} \\ \beta_j \in \{0, 1\}, \forall j \in 1, \dots, r \\ \sum_{i=1}^{r-1} \beta_j \geq 2 \end{cases}$$

Имеем задачу наименьших квадратов с целочисленными ограничениями (ILS), которая является NP-сложной.

Прогнозирование

Базис в пространстве векторов задержек сигнала даётся как
 $\text{Lin}(\{a_i\}) \Leftrightarrow A = [a_1 \dots a_r]$.

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

$$\overset{\leftarrow}{x}_{N+1} = A\lambda \Leftrightarrow \begin{cases} x_{kn} = A_{kn}\lambda \\ x(N+1) = a_{pr}^T\lambda \end{cases}, \text{ где}$$

$$A = \left(\frac{A_{kn}}{a_{pr}^T} \right)$$

$$\overset{\leftarrow}{x}_{N+1} = (x_{kn}, x(N+1))^T$$

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

$$x(N+1) = a_{pr}^T (A_{kn}^T A_{kn})^{-1} A_{kn}^T x_{kn}$$

Вычислительный эксперимент

Цель:

- применить метод tSSA для декомпозиции и прогноза многомерных временных рядов, а также сравнить его с другими моделями: mSSA, VAR, RNN

Рассматриваемые ряды:

- план выработки электричества на день и цена его производства
- метеорологические данные (температура, осадки, атмосферное давление)

Метрики:

- MSE, MAPE для прогнозирования
- AHE, RHE для декомпозиции

Вычислительный эксперимент

Определение

Абсолютной ошибкой ганкелизации матрицы M назовём

$$\text{AHE} = \|M - \text{Hankel}(M)\|_2$$

Определение

Относительной ошибкой ганкелизации матрицы M назовём

$$\text{RHE} = \frac{\text{AHE}}{\|M\|_2}$$

AHE можно представлять как суммарное стандартное отклонение антидиагоналей матрицы. RHE её более интерпретируемая модификация, которая и будет использоваться далее.

Данные метрики будут применяться к матрицам группировки факторов C_i .

Данные

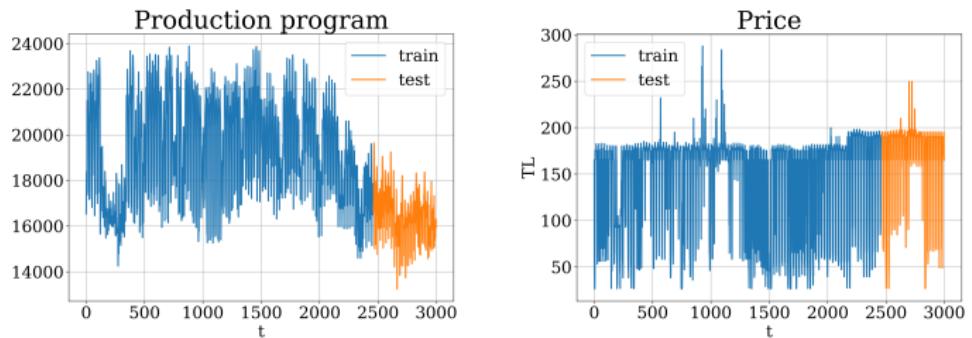
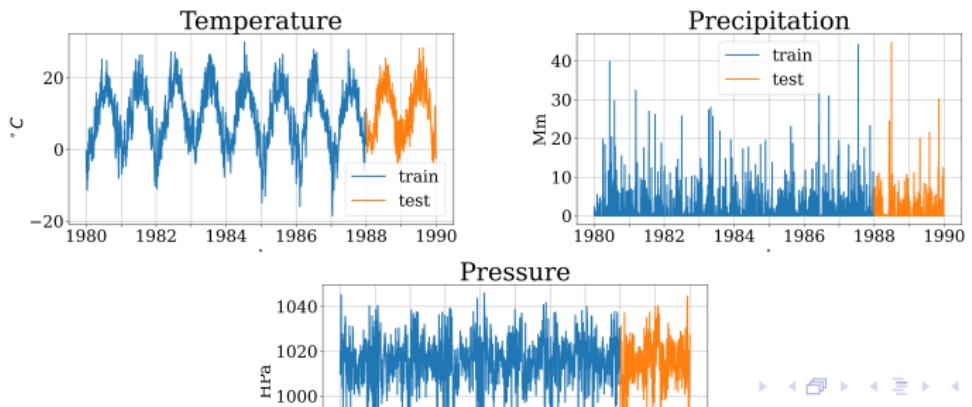


Рис. 4: Потребление электричества и его цены за единицу мощности



Электроэнергия. Прогноз

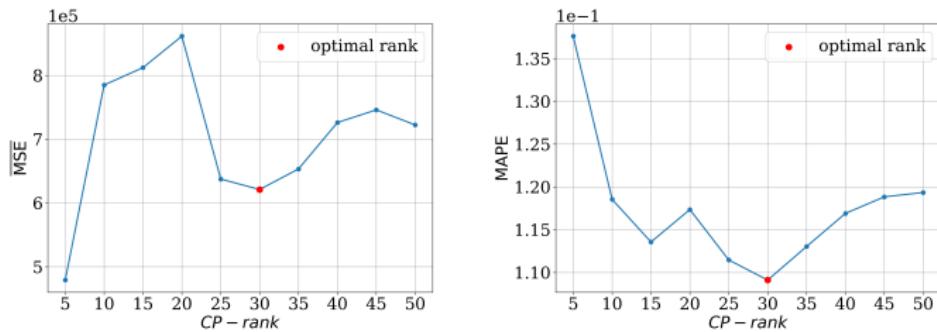
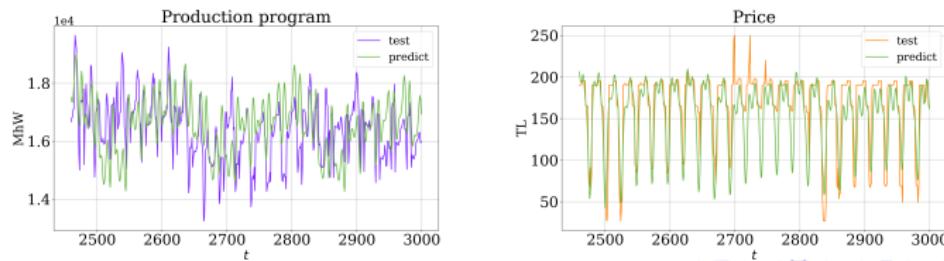


Рис. 6: Значения метрик \overline{MSE} и \overline{MAPE} предсказания tSSA для разных рангов СР-разложения. Отдельно выделен оптимальный ранг.

Наблюдается эффект переобучения при повышении значения ранга.



Электроэнергия. Прогноз

Таблица 1: Метрики моделей на прогнозировании данных электроэнергии

	$tSSA$	$mSSA$	VAR	RNN
$\overline{MSE}_{Production}, 10^6$	1.24	1.51	7.81	2.70
$\overline{MSE}_{Price}, 10^3$	0.88	1.03	4.85	30.0
$\overline{MSE}, 10^6$	0.62	0.75	3.91	135.00
$\overline{MAPE}_{Production}$	0.054	0.060	0.137	0.999
\overline{MAPE}_{Price}	0.164	0.170	0.360	1.004
\overline{MAPE}	0.109	0.115	0.249	1.002

Наш метод показал наилучшие результаты, хотя $mSSA$ был близок по точности. Метод VAR оказался нестабилен на выбранном горизонте прогнозирования, а RNN обучался в константную функцию.

Электроэнергия. Декомпозиция

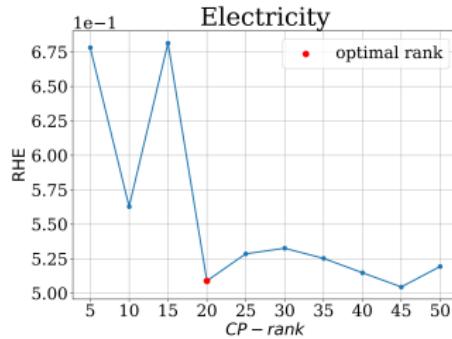


Рис. 6: Зависимость метрики \overline{RHE} от ранга СР-разложения. Данные электроэнергии. Отдельно выделен оптимальный ранг.

Метод tSSA достигает наибольшей обобщающей способности на небольших канонических рангах траекторных тензоров рядов, что подтверждает гипотезу о низкой собственной размерности данных.

Электроэнергия. Декомпозиция

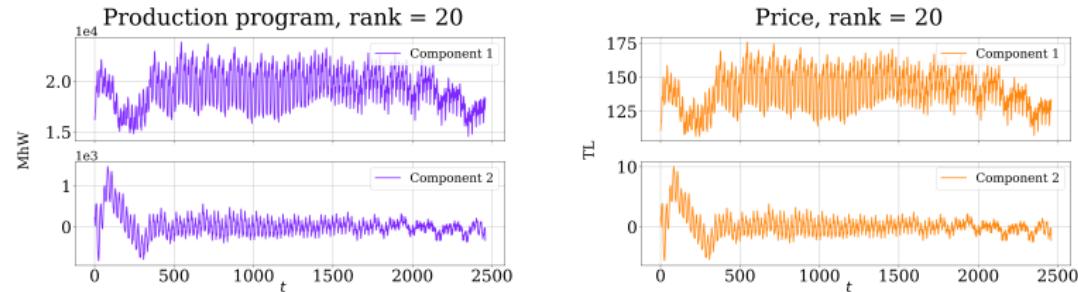


Рис. 7: Разложение рядов на две компоненты методом tSSA. Данные электроэнергии. CPD-ранг = 20

В силу вычислительной сложности задачи ILS, разложение на большее количество компонент не производится.

Из графиков видно, как сильно общий базис сигналов влияет на их декомпозицию: компоненты рядов получились почти идентичными, с точностью до смещения и масштаба. Метод выдели две гармоники разных амплитуд.

Электроэнергия. Декомпозиция

Таблица 2: Метрики моделей на декомпозиции данных электроэнергии

	tSSA	mSSA
RHE _{Production}	0.507	0.308
RHE _{Price}	0.511	0.31
RHE	0.509	0.309

Метод tSSA немного проигрывает по точности разложения. Кроме того, для него пока отсутствуют эвристики поиска оптимальных группировок, в отличие от mSSA.

Т.о. для tSSA необходимо или открывать собственные эмпирические закономерности, или менять принцип построения декомпозиции.

Выносится на защиту

Тензорный метод tSSA был разработан для прогноза и декомпозиции набора временных рядов, обработка которых требует учёта фактора взаимосвязанности. Его главное достоинство — проста в использовании: имеется всего два настраиваемых параметра, для которых не требуется изощрённых процедур обучения. Вместе с тем, теоретическое обоснование на основе довольно общей теории динамических систем. Главным вызовом метода является результат о NP-сложности поиска оптимальной декомпозиции рядов.

Литература

- D. Stepanov и N. Golyandina. «SSA-based approaches to analysis and forecast of multidimensional time series». В: Proceedings of the 5th St.Petersburg Workshop on Simulation. 2005, с. 293—298.
- Floris Takens. «Detecting Strange Attractors in Turbulence». В: Dynamical Systems and Turbulence, Warwick 1980. Под ред. David Rand и Lai-Sang Young. Т. 898. Lecture Notes in Mathematics. Berlin: Springer, 1981. Гл. 21, с. 366—381.
- Tamara Kolda и Brett Bader. «Tensor Decompositions and Applications». В: SIAM Review 51 (авг. 2009), с. 455—500.