

Модели обнаружения связей между временными рядами в задачах построения прогностических моделей

Карина Равилевна Усманова

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Научный руководитель: д.ф.-м.н. В.В.Стрижов

Выпускная квалификационная работа бакалавра

Москва 2018

Обнаружение связей во временных рядах

Задача

Повысить качество прогноза временного ряда путем учета экзогенных факторов. Установить связи между временными рядами, с целью их включения в прогностическую модель.

Решение

Для установления связей предлагается использовать ССМ. Считается, что два ряда связаны, если существуют их траекторные подпространства, проекции в которые связаны. Проекции связаны, если окрестность фазовой траектории одного ряда отображается в окрестность фазовой траектории другого ряда.

Требуется

Отыскать траекторные подпространства, обнаруживающие эту связь.

- Sugihara G., May R., Ye H., Hsieh C., Deyle E., Fogarty M., and Munch S // Detecting causality in complex ecosystems. 2012.
- Sugihara G., May R. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series // Nature. 1990.
- Motrenko A. and Strijov V. Extracting fundamental periods to segment biomedical signals. 2015.
- Golyandina N. and Stepanov D. SSA-based approaches to analysis and forecast of multidimensional time series // 5th St. Petersburg workshop on simulation. 2005.

Постановка задачи установления связи

Для временных рядов $\mathbf{x} = [x_1, \dots, x_N]$ и $\mathbf{y} = [y_1, \dots, y_N]$ установить наличие связи между ними.

Решение

Считаем, что ряд \mathbf{y} зависит от ряда \mathbf{x} , если существует липшицево отображение $\varphi : \mathbb{H}_x \rightarrow \mathbb{H}_y$

$$\rho_{\mathbb{H}_y}(\varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j)) \leq L \cdot \rho_{\mathbb{H}_x}(\mathbf{x}_i, \mathbf{x}_j), \quad \forall \mathbf{x}_i, \mathbf{x}_j \in \mathbb{H}_x$$

Траекторная матрица \mathbf{H}_x :

$$\mathbf{H}_x = \begin{pmatrix} x_1 & \dots & x_{L-1} & x_L \\ x_2 & \dots & x_L & x_{L+1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N-L+1} & \dots & x_{N-1} & x_N \end{pmatrix} = \begin{pmatrix} \mathbf{x}_L \\ \mathbf{x}_{L+1} \\ \vdots \\ \mathbf{x}_N \end{pmatrix}$$

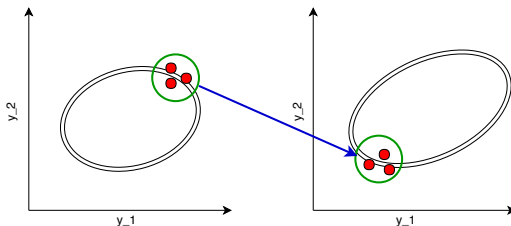
где L – ширина окна, $\mathbf{x}_i = (x_{i-L+1}, \dots, x_{i-1}, x_i) \in \mathbb{H}_x$,
 \mathbb{H}_x – траекторное пространство.

Построение сходящегося перекрестного отображения

- Выбираем $\mathbf{x}_{t^*} = (x_{t^*-L+1}, \dots, x_{t^*-1}, x_{t^*})$.
- Пусть $\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_k}$ – k ближайших соседей вектора \mathbf{x}_{t^*} в пространстве $\mathbb{H}_{\mathbf{x}}$. Тогда $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \dots, \mathbf{y}_{t_k}$ – строки матрицы $\mathbf{H}_{\mathbf{y}}$, соответствующие индексам t_1, \dots, t_k .

$$S(\mathbf{x}, \mathbf{y}) = \frac{\text{dist}(\mathbf{x})}{\text{dist}(\mathbf{y})}, \quad \text{dist}(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^k \|\mathbf{x}_{t^*} - \mathbf{x}_{t_i}\|_2$$

Если $S(\mathbf{x}, \mathbf{y})$ больше некоторого порога s , то ряд \mathbf{y} зависит от ряда \mathbf{x} .



Построение проекции

- Сингулярное разложение траекторной матрицы:

$$\mathbf{H}_x = \mathbf{U}_x \mathbf{\Lambda}_x \mathbf{V}_x$$

- Выберем \mathcal{T}_x – некоторый набор индексов компонент ряда \mathbf{x}
- $\mathbb{H}_{\mathcal{T}_x} \subset \mathbb{H}_x$ – траекторное подпространство
- Проекция ряда \mathbf{x} в подпространство $\mathbb{H}_{\mathcal{T}_x}$, описывается траекторной матрицей

$$\mathbf{P}_{\mathcal{T}_x} = \mathbf{U}_x \tilde{\mathbf{\Lambda}}_x \mathbf{V}_x$$

$$S(\mathbf{x}, \mathbf{y}, \mathcal{T}_x, \mathcal{T}_y) = \frac{\text{dist}(\mathbf{x}, \mathcal{T}_x)}{\text{dist}(\mathbf{y}, \mathcal{T}_y)}, \quad \text{dist}(\mathbf{x}, \mathcal{T}_x) = \frac{1}{k} \sum_{i=1}^k \|\mathbf{x}_{t^*} - \mathbf{x}_{t_i}\|_2$$

Задача поиска подпространств $\mathbb{H}_{\mathcal{T}_x}$ и $\mathbb{H}_{\mathcal{T}_y}$ эквивалентна поиску номеров главных компонент $(\mathcal{T}_x, \mathcal{T}_y)$

$$(\mathcal{T}_x, \mathcal{T}_y) = \arg \max_{\mathcal{T}_x, \mathcal{T}_y} S(\mathbf{x}, \mathbf{y}, \mathcal{T}_x, \mathcal{T}_y),$$

$$|\mathcal{T}_x| \rightarrow \min$$

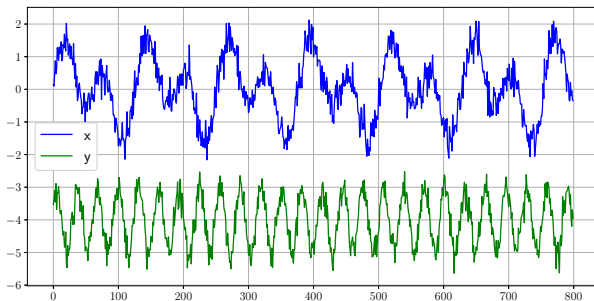
$$|\mathcal{T}_y| \rightarrow \min$$

Эксперимент на искусственных данных

Эксперимент проводился на двух зашумленных синусах с разным периодом.

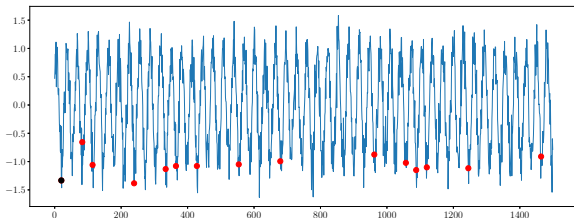
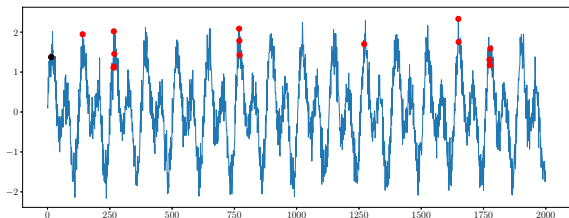
$$\mathbf{x} = \sin t + 2 \sin \frac{t}{2} + \sigma_x^2 \varepsilon, \quad \sigma_x^2 = 0.3, \quad \varepsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\mathbf{y} = \sin(2t + 5) + \sigma_y^2 \varepsilon, \quad \sigma_y^2 = 0.25, \quad \varepsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I})$$

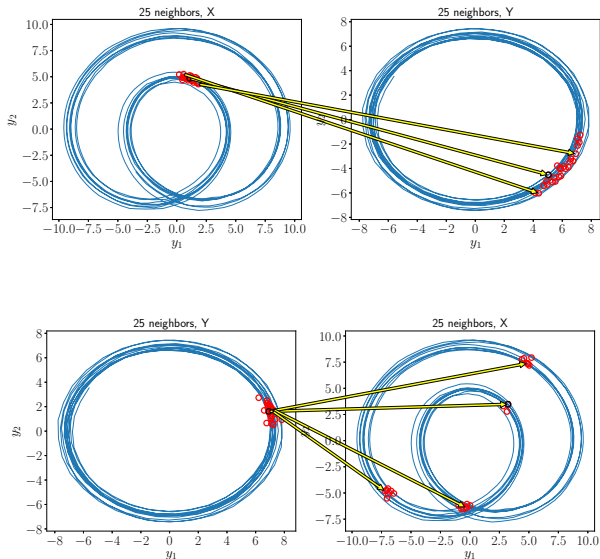


Найденные ближайшие соседи

Для $t^* = 15$ найдем ближайших соседей векторов \mathbf{x}_{t^*} и \mathbf{y}_{t^*} в траекторных пространствах \mathbb{H}_x и \mathbb{H}_y . Изобразим соответствующие им моменты времени на рядах \mathbf{x} и \mathbf{y} .

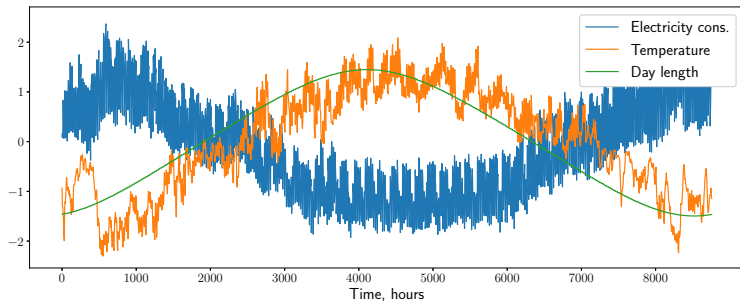


Ближайшие соседи на фазовых траекториях



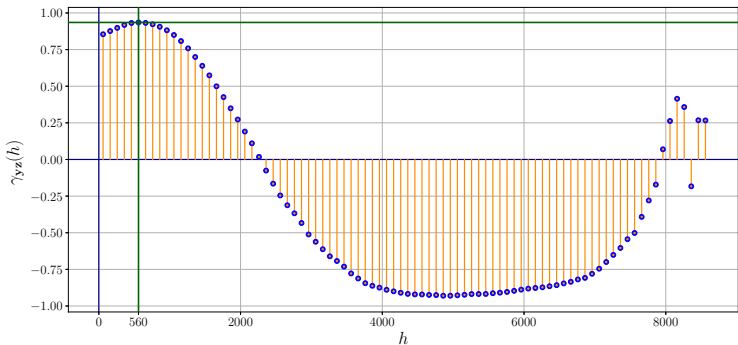
Эксперимент на данных потребления электроэнергии

- Эксперимент проводился на данных потребления электроэнергии x и температуры y в течение года.
- Для приведения ряда температуры к стационарной форме используется ряд длины светового дня z .
- Нормированные ряды потребления электроэнергии, температуры и длины светового дня:



Кросс-корреляция температуры и длины светового дня

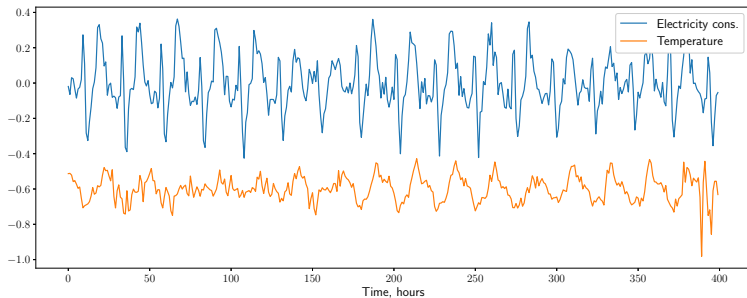
Найдем сдвиг h ряда длины светового дня z относительно ряда температуры y .



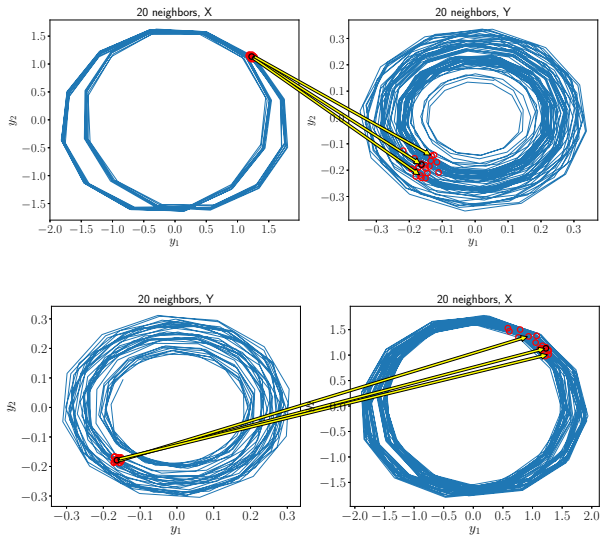
Максимум кросс-корреляции достигается при $h = 560$. Ряд температуры, приведенный к стационарной форме имеет вид:

$$y(t) := y(t) - z(t - 560)$$

Продифференцированные ряды

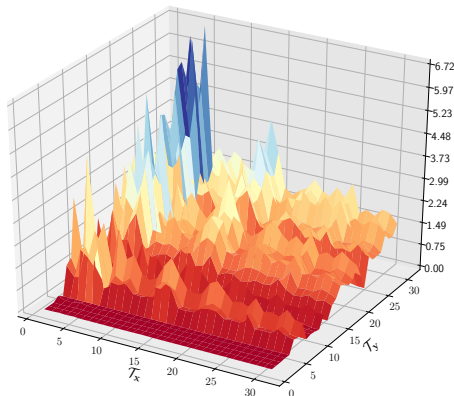


Ближайшие соседи на фазовых траекториях



Выбор траекторных подпространств

Рассмотрим различные подпространства \mathbb{H}_x и \mathbb{H}_y . Для каждой пары компонент $(\mathcal{T}_x, \mathcal{T}_y)$ найдем значение $S(\mathbf{x}, \mathbf{y}, \mathcal{T}_x, \mathcal{T}_y)$



Оптимальные подпространства задаются номерами компонент
 $\mathcal{T}_x = \{4\}, \quad \mathcal{T}_y = \{0, 1\}$

- Проведен вычислительный эксперимент по обнаружению связей между временными рядами.
- С помощью метода ССМ исследованы связи между рядами потребления электроэнергии и температуры.
- Исследована связь между проекциями этих рядов в различные подпространства.

К. Р. Усманова, С. П. Кудияров, Р. В. Мартышкин, А. А. Замковой, В. В. Стрижов // Анализ зависимостей между показателями при прогнозировании объема грузоперевозок // Системы и средства информатики, 2018.