

# Модели обнаружения зависимостей во временных рядах (проекция в латентные пространства)

## 1 Цель

Работа посвящена обнаружению причинно-следственных связей между разнородными временными рядами. Примеры зависимых разнородных временных рядов:

1. Эконометрические временные ряды.
2. Связь показателей ЭКГ и пульса (<http://smartlab.ws/component/content/article?id=60>)

Если прогноз временного ряда  $\mathbf{x}$  строится с использованием группы временных рядов  $\mathbf{y}_1, \dots, \mathbf{y}_k$ , то установление зависимостей ряда  $\mathbf{x}$  от  $\mathbf{y}_1, \dots, \mathbf{y}_k$  может повысить качество прогноза и упростить прогностическую модель. Если установлено, что ряд  $\mathbf{x}$  не зависит от ряда  $\mathbf{y}_i$ , то  $\mathbf{y}_i$  можно исключить из прогностической модели. В данной работе для обнаружения зависимостей между рядами в работе применяется два подхода: тест Гренджера и метод снижения размерности PLS.

### 1.1 Тест Гренджера

В основе теста Гренджера лежит следующий подход. Считаем, что ряд  $\mathbf{x}$  зависит от ряда  $\mathbf{y}$  (или следует из ряда  $\mathbf{y}$ ), если использование истории ряда  $\mathbf{y}$  при построении прогностической модели улучшает прогноз ряда  $\mathbf{x}$ . Тест Гренджера позволяет установить причинно-следственные связи между рядами и основан на сравнении качества прогноза, в котором используется история только прогнозируемого ряда, и прогноза, который дополнительно использует историю других рядов. Если улучшение качества прогноза подтверждается статистически, то говорят, что прогнозируемый ряд следует из использовавшихся во втором прогнозе рядов. Более формально используемый в этой работе тест Гренджера описан в разделе 4. Тест Гренджера применим к стационарным временным рядам, поэтому в случае нестационарных рядов их необходимо продифференцировать перед проведением теста Гренджера. Тест Гренджера используется в различных задачах, в которых необходимо исследовать взаимосвязь между развивающимися во времени процессами

В данной работе для построения прогноза одного временного ряда по нескольким используется алгоритм многомерной гусеницы (MSSA-L). Этот алгоритм является обобщением на многомерный случай алгоритма анализа спектральных компонент SSA.

Метод SSA основан на разложении временного ряда в сумму интерпретируемых компонент. Он делится на четыре основных шага: запись ряда в виде траекторной матрицы, ее сингулярное разложение, группировка компонент полученных при сингулярном разложении, по каждой сгруппированной матрице восстанавливается временной ряд. Таким образом исходный временной ряд представляется в виде суммы временных рядов. Метод SSA применяется в таких задачах, как выявление трендов во временных рядах, подавления шума во временных рядах, прогнозирование временных рядов.

Может оказаться, что у исследуемых рядов зависимость существует только между определенными компонентами. Поэтому важно правильно выбрать главные компоненты при разложении в методе SSA. Снижение размерности (проекция в фазовое подпространство) позволяет более детально изучить связь между главными компонентами рядов и найти подпространство, в котором наблюдается зависимость между рядами.

## 2 Постановка задачи прогнозирования

Поставим задачу прогноза многомерного временного ряда.

Обозначим  $\mathbf{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(s)})^\top$  – заданный  $s$ -мерный временной ряд. Построим матрицу плана из сегментов ряда:

$$\begin{pmatrix} x_0^{(1)} & \dots & x_{n-1}^{(1)} \\ \vdots & & \\ x_0^{(s)} & \dots & x_{n-1}^{(s)} \end{pmatrix} = \mathbf{X}_{0:(n-1)}. \quad (1)$$

Пусть  $\mathbf{x}_n = (x_n^{(1)}, \dots, x_n^{(s)})^\top$  – значение ряда  $\mathbf{X}$  в момент времени  $n$ . Построим прогноз  $\hat{\mathbf{x}}$  ряда  $\mathbf{X}$  в точке  $\mathbf{x}_n$ . Прделаем это  $k$  раз для различных обучающих выборок  $\mathbf{X}_{\text{train}}^i = \mathbf{X}_{i:(n+i-1)}$ ,  $i = 0, \dots, (k-1)$ . Получим  $k$  прогнозов  $\hat{\mathbf{X}} = (\hat{\mathbf{x}}_n, \hat{\mathbf{x}}_{n+1}, \dots, \hat{\mathbf{x}}_{n+k-1})$  ряда  $\mathbf{X}$  в точках  $\mathbf{x}_n, \mathbf{x}_{n+1}, \dots, \mathbf{x}_{n+k-1}$ .

Прогностическая модель имеет вид

$$\hat{\mathbf{x}}_{t+1} = \mathbf{f}(\hat{\mathbf{w}}, \mathbf{x}_t, \mathbf{x}_{t-1}, \dots, \mathbf{x}_{t-L+2}),$$

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} S(\mathbf{w}, \mathbf{X}, \hat{\mathbf{x}}_n, \hat{\mathbf{x}}_{n+1}, \dots, \hat{\mathbf{x}}_{n+k-1}) = S(\mathbf{w}, \mathbf{X}, \hat{\mathbf{X}}),$$

где функция потерь

$$S(\mathbf{w}, \mathbf{X}, \hat{\mathbf{X}}) = \sum_{i=0}^{k-1} \mathcal{L}(\mathbf{x}_{n+i}^{(1)}, \hat{\mathbf{x}}_{n+i}^{(1)}).$$

В данной работе в качестве прогностической модели  $\mathbf{f}$  используется алгоритм многомерной гусеницы (MSSA-L). Функция  $\mathbf{f}$  имеет вид:

$$\mathbf{f}(\hat{\mathbf{w}}, \mathbf{x}_t, \mathbf{x}_{t-1}, \dots, \mathbf{x}_{t-L+2}) = \begin{pmatrix} x_{t-L+2}^{(1)} & \dots & x_t^{(1)} \\ x_{t-L+2}^{(2)} & \dots & x_t^{(2)} \\ \vdots & & \\ x_{t-L+2}^{(s)} & \dots & x_t^{(s)} \end{pmatrix} \cdot \mathbf{p}.$$

вектор коэффициентов  $\mathbf{p}$  определяется алгоритмом многомерной гусеницы MSSA-L. Алгоритм MSSA-L подробнее описан в следующем разделе.

## 3 Алгоритм многомерной гусеницы (MSSA-L)

Алгоритм MSSA-L является обобщением на многомерный случай алгоритма гусеницы (SSA). Задача алгоритма MSSA-L состоит в представлении временного ряда в виде суммы интерпретируемых компонент. Это осуществляется в четыре шага: запись ряда в виде траекторной матрицы, сингулярное разложение этой матрицы, группировка компонент, полученных при сингулярном разложении, в интерпретируемые компоненты и восстановление временного ряда по каждой из интерпретируемых компонент.

По ряду (1) построим матрицу Ганкеля  $\mathbf{H} \in \mathbb{R}^{L \times sK}$ ,  $K = N - L + 1$ :

$$\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_s],$$

где  $L$  – ширина окна,  $\mathbf{H}_i \in \mathbb{R}^{L \times K}$  – матрица Ганкеля для ряда  $\mathbf{x}^{(i)}$ ,

$$\mathbf{H}^{(i)} = \begin{pmatrix} x_0^{(i)} & x_1^{(i)} & \dots & x_{N-L}^{(i)} \\ x_1^{(i)} & x_2^{(i)} & \dots & x_{N-L+1}^{(i)} \\ \vdots & & & \\ x_{L-1}^{(i)} & x_L^{(i)} & \dots & x_{N-1}^{(i)} \end{pmatrix}.$$

По матрице Ганкеля  $\mathbf{H}$  восстановим временной ряд  $\mathbf{X}$ . Метод многомерной гусеницы строит приближение  $\hat{\mathbf{H}}$  матрицы  $\mathbf{H}$  меньшего ранга с помощью сингулярного разложения этой матрицы и восстанавливает ряд по матрице  $\hat{\mathbf{H}}$ . Сингулярное разложение матрицы  $\mathbf{H}$  имеет вид

$$\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V} = \sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{v}_i^T$$

где  $\lambda_1, \dots, \lambda_d > 0$  – сингулярные числа матрицы  $\mathbf{H}$ ,  $\mathbf{u}_i$  и  $\mathbf{v}_i$  – столбцы матриц  $\mathbf{U}$  и  $\mathbf{V}$ . Тогда наилучшее приближение матрицы  $\mathbf{H}$  матрицей ранга  $r < d$  имеет вид :

$$\hat{\mathbf{H}} = \sum_{i=1}^r \lambda_i \mathbf{u}_i \mathbf{v}_i^T$$

По матрице  $\hat{\mathbf{H}}$  восстанавливается временной ряд  $\mathbf{X}$  путем усреднения элементов, стоящих на анти-диагоналях.

Алгоритм многомерной гусеницы также позволяет построить прогноз временного ряда в момент  $N$  по  $(L - 1)$  предыдущим значениям ряда. Алгоритм находит такой вектор коэффициентов  $\mathbf{p} \in \mathbb{R}^{(L-1)}$ , что значения ряда  $\mathbf{X}$  в момент  $N$ :

$$\mathbf{x}_N = \begin{pmatrix} x_{N-L+1}^{(1)} & \cdots & x_{N-1}^{(1)} \\ x_{N-L+1}^{(2)} & \cdots & x_{N-1}^{(2)} \\ \vdots & & \\ x_{N-L+1}^{(s)} & \cdots & x_{N-1}^{(s)} \end{pmatrix} \cdot \mathbf{p} = \mathbf{Y} \cdot \mathbf{p} \quad (2)$$

Заметим, что коэффициенты  $\mathbf{p}$  оказываются общими для всех компонент ряда  $\mathbf{X}$ .

Для каждого  $i \in [1, r]$  обозначим  $\tilde{\mathbf{u}}_i$  первые  $(L - 1)$  компонент столбца  $\mathbf{u}_i$ ,  $\pi_i$  – последнюю компоненту столбца  $\mathbf{u}_i$  и  $\nu = \sum_{i=1}^r \pi_i^2$ . Тогда вектор коэффициентов  $\mathbf{p}$  вычисляется по формуле:

$$\mathbf{p} = \frac{1}{1 - \nu^2} \sum_{i=1}^r \pi_i \tilde{\mathbf{u}}_i \quad (3)$$

Заметим, что для одномерного временного ряда справедливы все приведенные соотношения при  $s = 1$ .

## 4 Тест Гренджера

В работе для установления причинно-следственных связей предлагается использовать статистический тест Гренджера. Ниже приведен алгоритм теста Гренджера для проверки наличия зависимости одного временного ряда от другого. Пусть требуется проверить, зависит ли ряд  $\mathbf{x}$  от ряда  $\mathbf{y}$ . Выдвинем гипотезу о независимости ряда  $\mathbf{x}$  от ряда  $\mathbf{y}$  и проверим ее. Делаем это следующим образом.

1. Строим прогноз ряда  $\mathbf{x}$  без использования ряда  $\mathbf{y}$  и находим значение функции потерь

$$S_{\mathbf{x}} = \sum_{i=1}^n \mathcal{L}(\mathbf{x}_i, \hat{\mathbf{x}}_i),$$

где  $n$  – длина тестовой выборки.

Функцию  $\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}})$  выбираем в зависимости от распределения ошибок прогноза на тестовой выборке (??).

2. Строим прогноз ряда  $\mathbf{x}$  с использованием ряда  $\mathbf{y}$ . Вычисляем для него значение функции потерь

$$S_{\mathbf{xy}} = \sum_{i=1}^n \mathcal{L}(\mathbf{x}_i, \hat{\mathbf{x}}_i).$$

3. Рассмотрим статистику

$$T(\mathbf{x}, \mathbf{y}) = \frac{N - 2k}{k} \cdot \frac{S_{\mathbf{x}} - S_{\mathbf{xy}}}{S_{\mathbf{xy}}},$$

где  $N$  – длина обучающей выборки,  $k$  – размерность регрессионной модели. Статистика  $T$  имеет распределение  $F(k, N - 2k)$  (распределение Фишера с параметрами  $(k, N - 2k)$ ).

4. Если ряд  $\mathbf{x}$  не зависит от ряда  $\mathbf{y}$ , то значения  $S_{\mathbf{x}}$  и  $S_{\mathbf{xy}}$  будут близки, а статистика  $T(\mathbf{x}, \mathbf{y})$  – незначима. Поэтому в случае больших значений статистики  $T(\mathbf{x}, \mathbf{y})$  отвергаем гипотезу о независимости ряда  $\mathbf{x}$  от  $\mathbf{y}$ . Выберем некоторое критическое значение  $t$  статистики  $T(\mathbf{x}, \mathbf{y})$ . Тогда критерий зависимости ряда  $\mathbf{x}$  от ряда  $\mathbf{y}$  выглядит следующим образом:

Из  $T(\mathbf{x}, \mathbf{y}) > t$  следует, что ряд  $\mathbf{x}$  зависит от ряда  $\mathbf{y}$

5. Аналогично проверим зависимость ряда  $\mathbf{x}$  от восстановленного (с помощью алгоритма MSSA-L) ряда  $\hat{\mathbf{y}}$ . Для этого используем статистику

$$T(\mathbf{x}, \hat{\mathbf{y}}) = \frac{N - 2k}{k} \cdot \frac{S_{\mathbf{x}} - S_{\mathbf{x}\hat{\mathbf{y}}}}{S_{\mathbf{x}\hat{\mathbf{y}}}}.$$

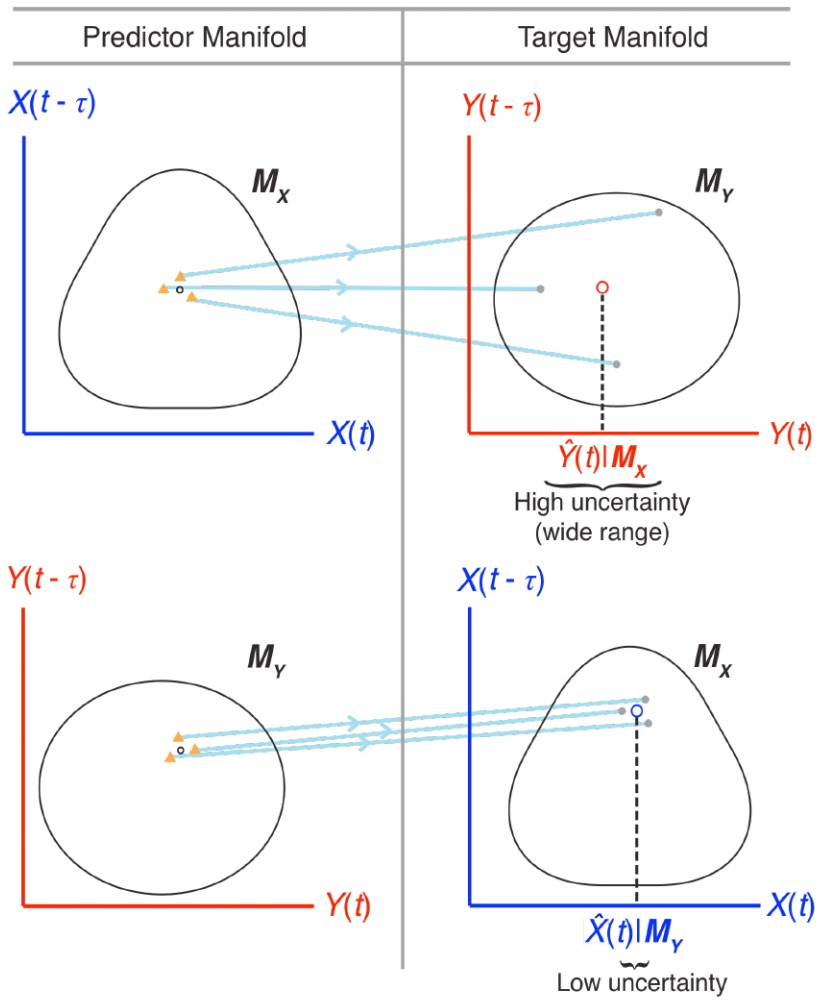
Для более подробного изучения связи между временными рядами  $\mathbf{x}$  и  $\mathbf{y}$  вычисляем кросс-корреляционную функцию  $\gamma_{\mathbf{xy}}(h)$

$$\gamma_{\mathbf{xy}}(h) = \frac{\mathbb{E}[(\mathbf{x}_t - \mu_{\mathbf{x}})(\mathbf{y}_{t+h} - \mu_{\mathbf{y}})]}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}},$$

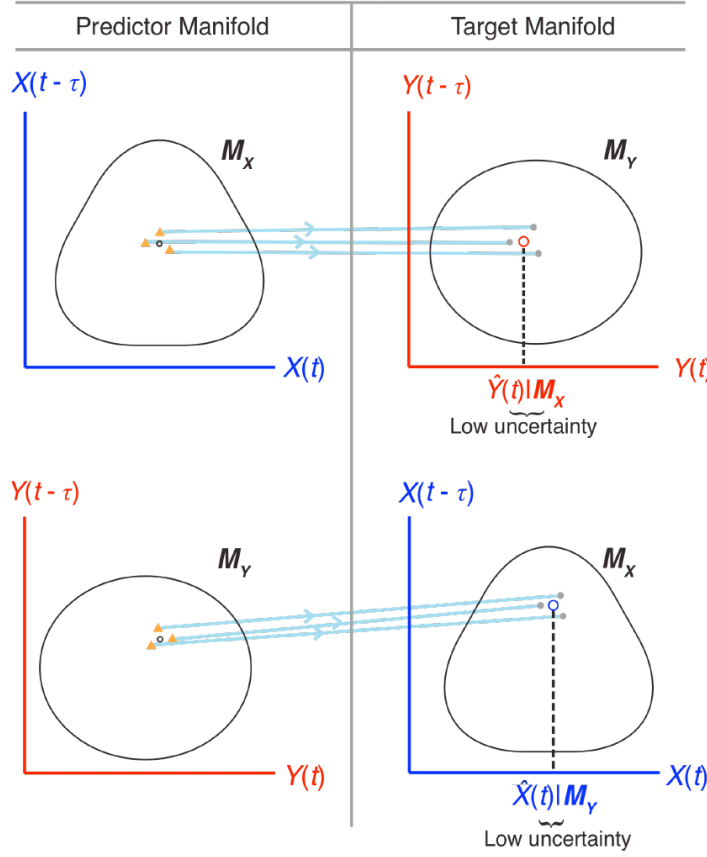
где  $\mathbb{E}$  – математическое ожидание,  $\mu$  – выборочное среднее,  $\sigma$  – выборочная дисперсия.

Если  $h^*$  соответствует максимальному значению кросс-корреляции, то говорят, что ряд  $\mathbf{y}$  сдвинут на  $h^*$  относительно  $\mathbf{x}$ . Заметим, что если ряд  $\mathbf{x}$  сдвинут на  $h_1$  относительно ряда  $\mathbf{y}$ , а ряд  $\mathbf{y}$  сдвинут на  $h_2$  относительно ряда  $\mathbf{z}$ . То ряд  $\mathbf{x}$  сдвинут на  $h_3 = h_1 + h_2$  относительно ряда  $\mathbf{z}$ .

Пусть прогноз ряда  $\mathbf{x}$  строится с использованием истории ряда  $\mathbf{y}$  и пусть с помощью вычисления кросс-корреляции рядов  $\mathbf{x}$  и  $\mathbf{y}$  получено, что ряд  $\mathbf{x}$  отстает от ряда  $\mathbf{y}$  на  $h$  отсчетов времени. Тогда использование при прогнозе ряда  $\mathbf{y}$ , сдвинутого на  $h$  отсчетов назад, может повысить качество прогноза.

**B****Asymmetric Causality,  $X \Rightarrow Y$** 

## A Bidirectional Causality (generic case), $X \Leftrightarrow Y$



## 5 CCSM

Пусть  $\mathbf{X}$  и  $\mathbf{Y}$  – временные ряды длины  $N$ .

Прогноз ряда  $\mathbf{Y}$  с помощью ряда  $\mathbf{X}$  строится следующим образом.

Строим матрицу Ганкеля

$$\mathbf{H}_x = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \dots & \mathbf{X}_{L-1} & \mathbf{X}_L \\ \mathbf{X}_2 & \mathbf{X}_3 & \dots & \mathbf{X}_L & \mathbf{X}_{L+1} \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{X}_{N-L+1} & \mathbf{X}_{N-L+2} & \dots & \mathbf{X}_{N-1} & \mathbf{X}_N \end{pmatrix},$$

где  $L$  – ширина окна, длина истории ряда, используемая при нахождении главных компонент. Аналогично строим матрицу  $\mathbf{H}_y$

Обозначим  $t$ -ю строку матрицы  $\mathbf{H}_x$  через  $\mathbf{x}_t$ . Найдем  $k$  ближайших соседей вектора  $\mathbf{x}_t$ . Обозначим их индексы через  $t_1, \dots, t_k$ . Тогда ближайшие соседи  $\mathbf{x}_t$  это строки матрицы  $\mathbf{H}_x$  с номерами  $t_1, \dots, t_k$ :

$$\mathbf{x}_{t_i} = (\mathbf{x}_{t_i}, \mathbf{x}_{t_i-1}, \dots, \mathbf{x}_{t_i-(L-1)}), \quad i = 1, \dots, k$$

Прогноз  $\hat{\mathbf{Y}}_t$  строится следующим образом

$$\hat{\mathbf{Y}}_t = \sum_{i=1}^k w_i Y_{t_i}, \quad \text{где } t_i \text{ – индексы ближайших соседей } \mathbf{x}_t$$

$$w_i = \frac{u_i}{\sum_i u_i}, \quad u_i = \exp - \left( \frac{\|\mathbf{x}_t - \mathbf{x}_{t_i}\|_2}{\|\mathbf{x}_t - \mathbf{x}_{t_{L+1}}\|_2} \right)$$

Аналогично строится прогноз ряда  $\mathbf{X}$  с использованием ряда  $\mathbf{Y}$ .

Этот подход можно применять для обнаружения зависимости между рядами следующим образом. Пусть выбран момент времени  $t^*$  и вектор  $\mathbf{x}_{t^*}$ . И пусть  $\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_k}$  – ближайшие соседи вектора  $\mathbf{x}_{t^*}$ . Тогда вектора  $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \dots, \mathbf{y}_{t_k}$  – строки матрицы  $\mathbf{Y}$ , соответствующие строки матрицы  $\mathbf{H}_y$ . Тогда, если вектора  $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \dots, \mathbf{y}_{t_k}$  расположены достаточно близко, то утверждается, что ряд  $\mathbf{Y}$  зависит от ряда  $\mathbf{X}$ .

Заметим, что можно искать ближайших соседей не во всем фазовом пространстве, а только в некотором его подпространстве, натянутом на первые главные компоненты. Пусть сингулярное разложение матрицы  $\mathbf{H}_x$  имеет вид

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

Пусть  $\mathbf{\Lambda}_x^l$  – минор матрицы  $\mathbf{\Lambda}_x$  размера  $l$ . Тогда проекция ряда  $\mathbf{X}$  в подпространство, натянутое на  $l$  главных компонент имеет вид.

$$\mathbf{P}_x^l = \mathbf{U}_x \mathbf{\Lambda}_x^l.$$

Аналогично строится  $\mathbf{P}_y^l$ . Далее предлагается искать ближайших соседей не в полных фазовых пространствах, задающихся матрицами Ганкеля  $\mathbf{H}_x$  и  $\mathbf{H}_y$ , а в подпространствах, задающихся матрицами  $\mathbf{P}_x^l$  и  $\mathbf{P}_y^l$ .

Рассмотрев различные подпространства, можно выбрать то, которое будет наилучшим образом описывать исследуемый временной ряд и будет иметь минимальную размерность. Перебор различных подпространств также позволяет установить зависимости между подпространствами ряда  $\mathbf{X}$  и ряда  $\mathbf{Y}$ .

## 6 ССМ, эксперимент

### 6.1 Сгенерированные данные

Эксперимент проводился на двух сгенерированных рядах  $\mathbf{X}$  и  $\mathbf{Y}$ :

$$\mathbf{X} = \sin t + 2 \sin \frac{t}{2} + \sigma_x^2 \boldsymbol{\varepsilon}, \quad \sigma_x^2 = 0.3$$

$$\mathbf{Y} = \sin(2t + 5) + \sigma_y^2 \boldsymbol{\varepsilon}, \quad \sigma_y^2 = 0.25,$$

где  $\boldsymbol{\varepsilon} \in \mathcal{N}(\mathbf{0}, \mathbf{I})$

Строим матрицу Ганкеля  $\mathbf{H}_x$  по ряду  $\mathbf{X}$ , взяв ширину окна  $L = 250$ . Для некоторого момента времени  $t^*$  рассмотрим вектор  $\mathbf{x}_{t^*}$ , равный  $t^*$ -й строке матрицы  $\mathbf{H}_x$ . Выберем  $k$  и найдем среди строк матрицы  $\mathbf{H}_x$   $k$  ближайших (в смысле евклидовой нормы) соседей вектора  $\mathbf{x}_{t^*}$ . Обозначим индексы найденных векторов  $t_1, \dots, t_k$ , а сами найденные вектора –  $\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_k}$ .

На рисунке изображен ряд  $\mathbf{X}$  и  $k = 25$  ближайших соседей для момента  $t^* = 15$ . Моменты времени  $t_1, \dots, t_k$  выделены красным, момент  $t^*$  – черным.

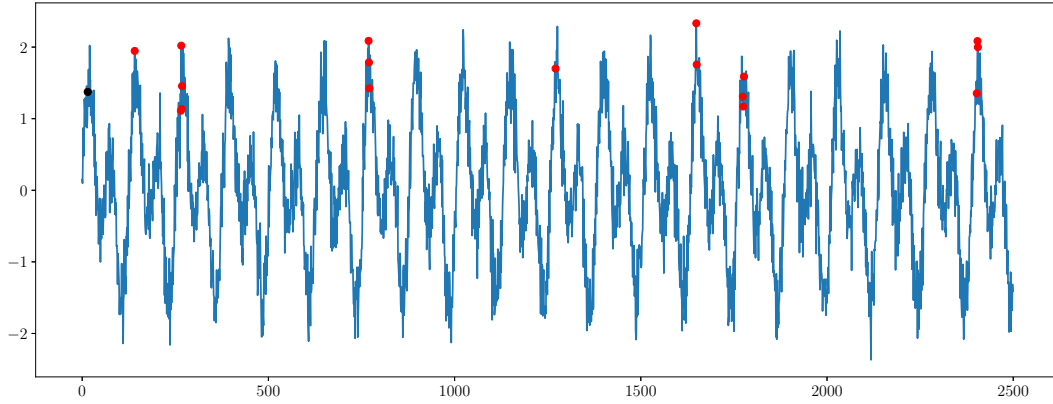


Рис. 1: Ближайшие соседи точки  $\mathbf{x}_{15}$

Строим матрицу Ганкеля  $\mathbf{H}_y$  по ряду  $\mathbf{Y}$ . Обозначим  $i$ -ю строку  $\mathbf{H}_y$  через  $\mathbf{y}_i$ . Тогда по найденным индексам  $t_1, \dots, t_k$  можно отобрать соответствующие  $\mathbf{y}_{t_1}, \dots, \mathbf{y}_{t_k}$ . Если ряд  $\mathbf{Y}$  зависит от ряда  $\mathbf{X}$ , то вектора  $\mathbf{y}_{t_1}, \dots, \mathbf{y}_{t_k}$ , как и  $\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_k}$ , будут находиться рядом в фазовом пространстве. Изобразим  $\mathbf{x}_{t^*}, \mathbf{x}_{t_1}, \dots, \mathbf{y}_{t_k}$  и  $\mathbf{y}_{t^*}, \mathbf{x}_{t_1}, \dots, \mathbf{y}_{t_k}$  на фазовых траекториях.

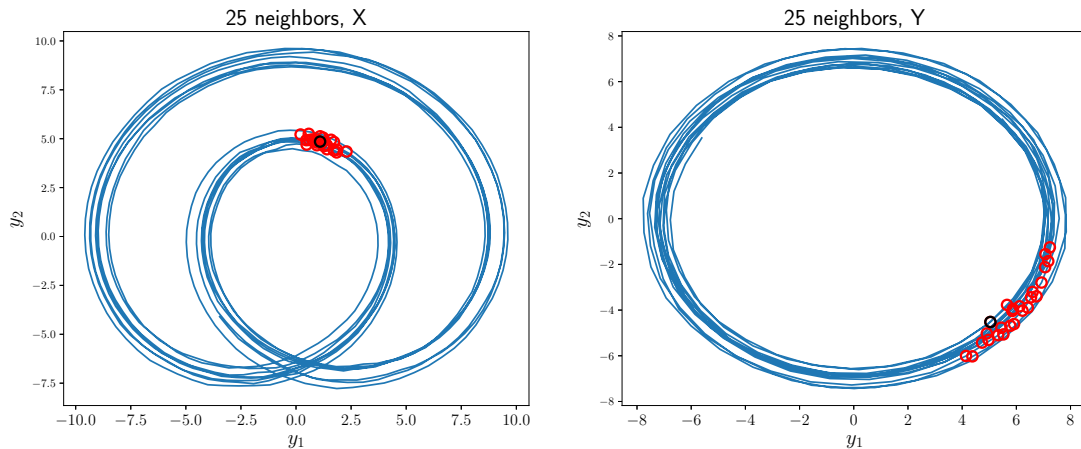
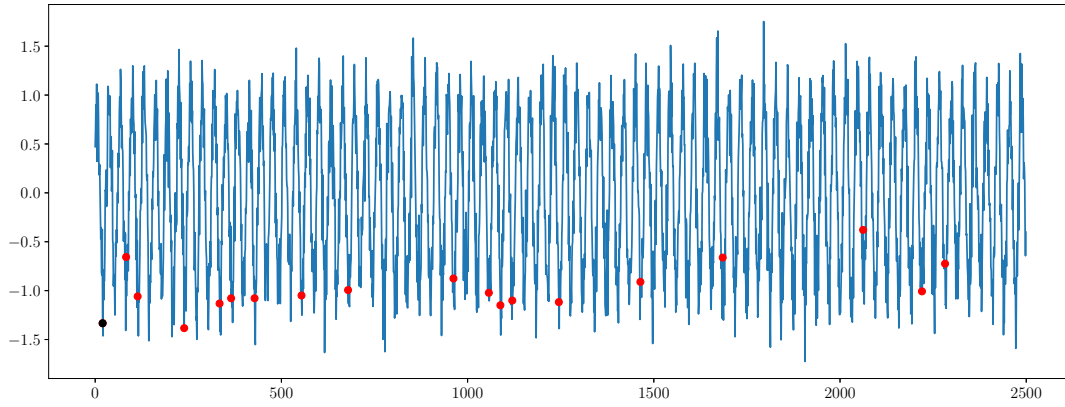


Рис. 2: Точки на фазовых траекториях рядов  $\mathbf{X}$  и  $\mathbf{Y}$ , соответствующие 15-ти ближайшим соседям  $\mathbf{x}_{15}$



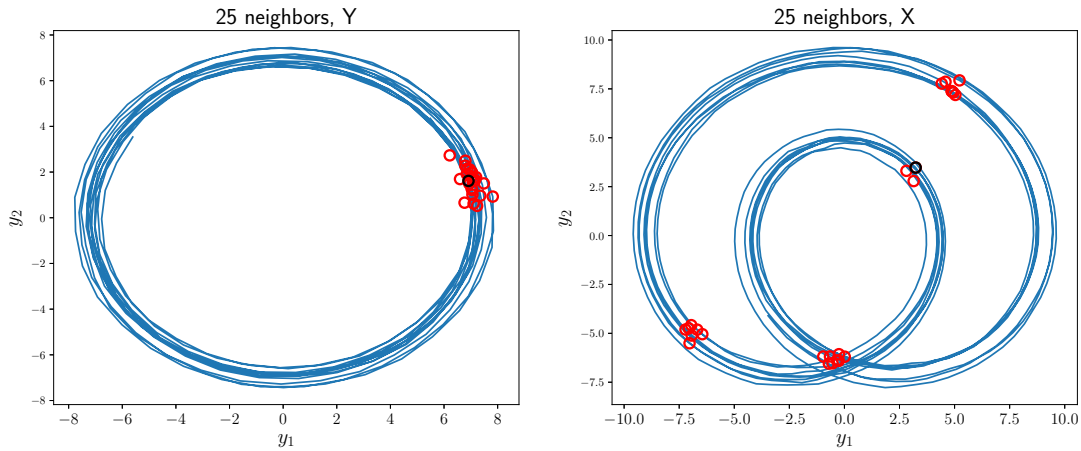
Видно, что точки обеих фазовых траекторий расположены близко друг другу. Значит, ряд  $\mathbf{Y}$  зависит от ряда  $\mathbf{X}$ .

Аналогично для некоторого  $t^*$  находим ближайших соседей вектора  $\mathbf{y}_{t^*}$ . Обозначим их  $\mathbf{y}_{t_1}, \dots, \mathbf{y}_{t_k}$ . На рисунке изображен ряд  $\mathbf{Y}$  и  $k = 25$  ближайших соседей вектора  $\mathbf{y}_{20}$ .



**Рис. 3:** Ближайшие соседи точки  $\mathbf{y}_{20}$

Изобразим  $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \dots, \mathbf{y}_{t_k}$  и соответствующие им  $\mathbf{x}_{t^*}, \mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_k}$  на фазовых траекториях.



**Рис. 4:** Точки на фазовых траекториях рядов  $\mathbf{X}$  и  $\mathbf{Y}$ , соответствующие 15-ти ближайшим соседям  $\mathbf{y}_{20}$

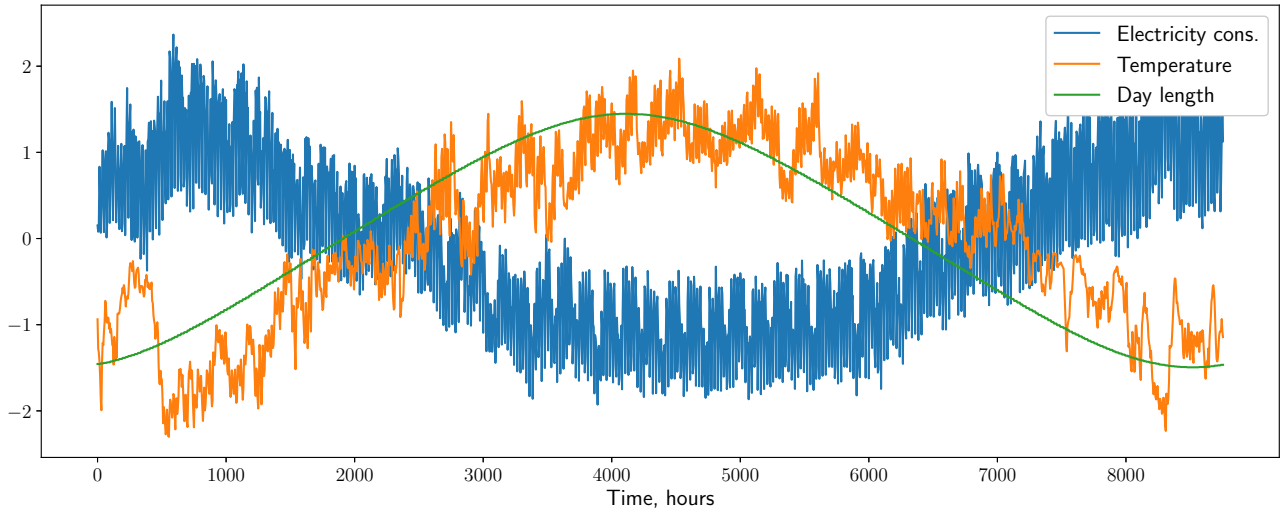
Видим, что точки  $\mathbf{x}_{t^*}, \mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_{15}}$  расположены на фазовых траекториях близко друг к другу. При этом они распадаются на четыре плотные группы. Это связано с тем, что период ряда  $\mathbf{X}$  в четыре раза меньше периода ряда  $\mathbf{Y}$ .

## 6.2 Реальные данные

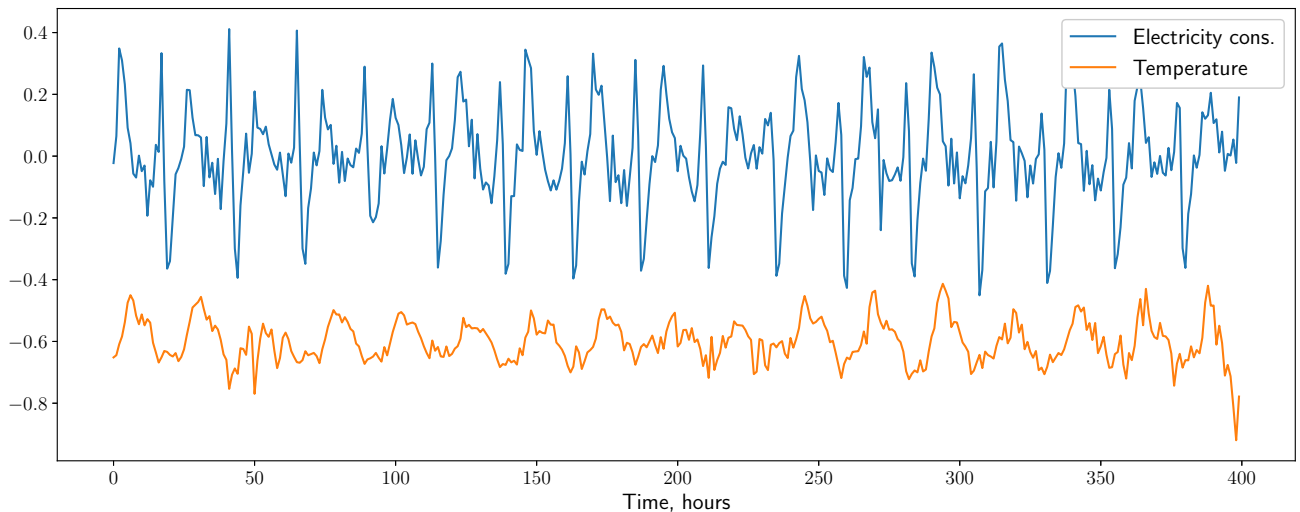
В эксперимент исследуются ряд объема потребления электроэнергии  $\mathbf{X}$  и ряд значений температуры  $\mathbf{Y}$  в течение года. Так как эти ряды не являются стационарными, их необходимо продифференцировать и отнормировать перед тем, как исследовать зависимости между ними. Ряд температуры будем приводить к стационарной форме следующим образом. Рассмотрим ряд длины светового дня в течение года  $\mathbf{Z}$ .

Определим, насколько ряд  $\mathbf{Z}$  опережает ряд  $\mathbf{Y}$ . То есть найдем такое  $h$ , что  $\mathbf{Y}(t + h) = \mathbf{Z}(t)$ . Вычтем из ряда  $\mathbf{Y}$  ряд  $\mathbf{Z}$  с учетом сдвига. Полученный ряд будет стационарным рядом температуры.

Исходные ряды потребления электроэнергии, температуры и длины светового дня изображены на рис. 5. Продифференцированные и нормированные ряды потребления электроэнергии и температуры изображены на рис. 6.



**Рис. 5:** Продифференцированные и нормированные ряды потребления электроэнергии и температуры



**Рис. 6:** Продифференцированные и нормированные ряды потребления электроэнергии и температуры

Исследуем зависимость ряда температуры  $\mathbf{Y}$  от ряда потребления электроэнергии  $\mathbf{X}$ . Делаем это аналогично эксперименту на искусственных данных. Выбираем ширину окна  $L$  и некоторый момент времени  $t^*$ . Находим  $k$  ближайших соседей векторов  $\mathbf{x}_{t^*}$  и  $\mathbf{y}_{t^*}$  и их расположение в фазовом пространстве.

Возьмем  $L = 170$ , что соответствует периоду в семь дней. Возьмем  $t^* = 400$ . На рис. 7 красным показаны ближайшие соседи вектора  $\mathbf{x}_{t^*}$

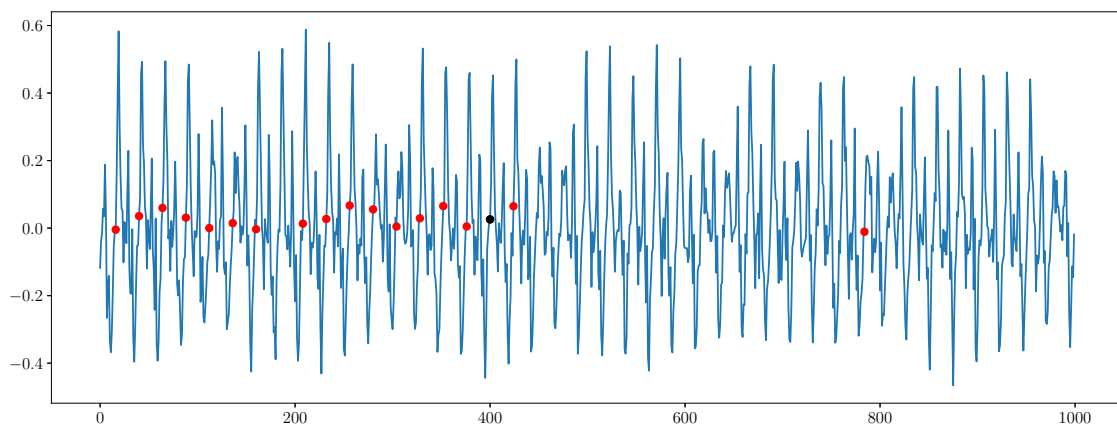


Рис. 7: Ближайшие соседи вектора  $\mathbf{x}_{t^*}$ , ширина окна  $L = 170$

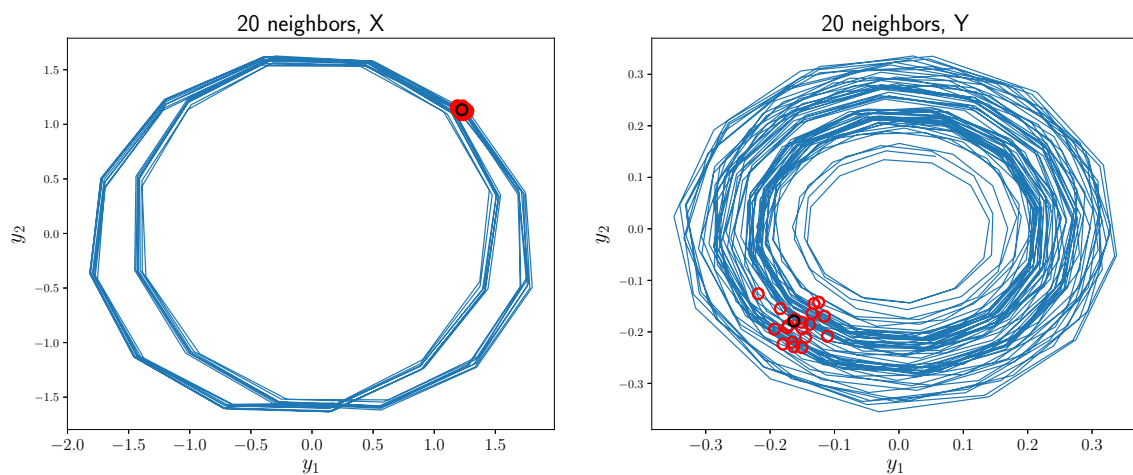


Рис. 8: Ближайшие соседи векторов  $\mathbf{x}_{t^*}$  и  $\mathbf{y}_{t^*}$  на фазовых диаграммах с периодом 12 часов

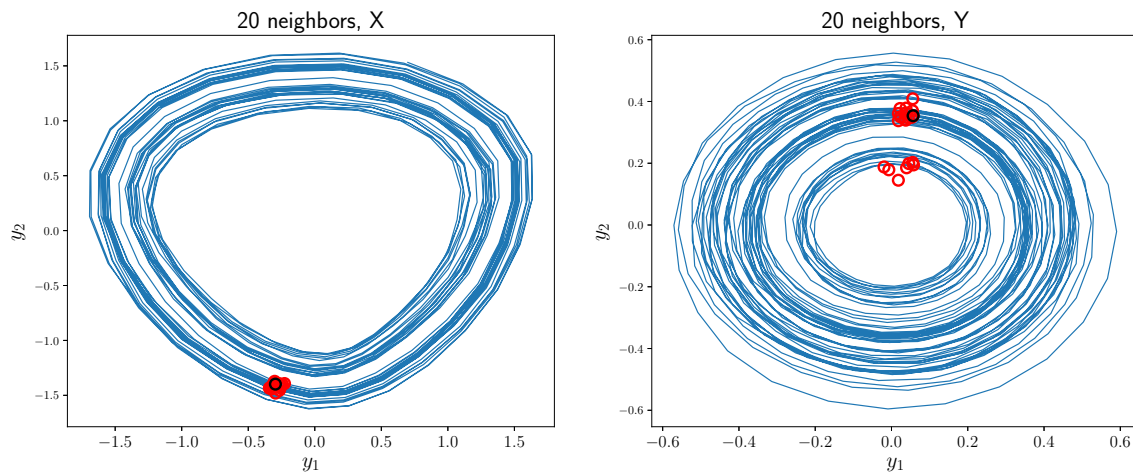
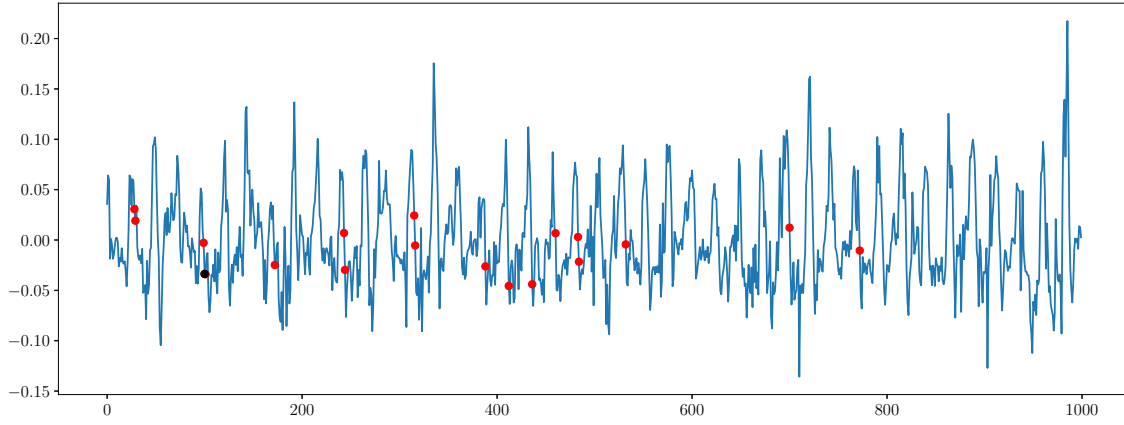
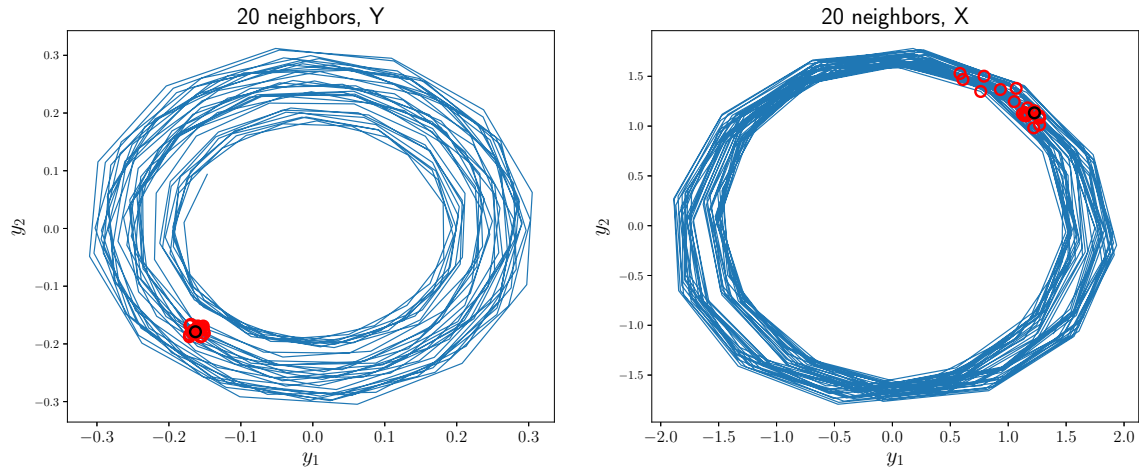


Рис. 9: Ближайшие соседи вектора  $\mathbf{x}_{t^*}$  и  $\mathbf{y}_{t^*}$  на фазовых диаграммах с периодом 24 часа

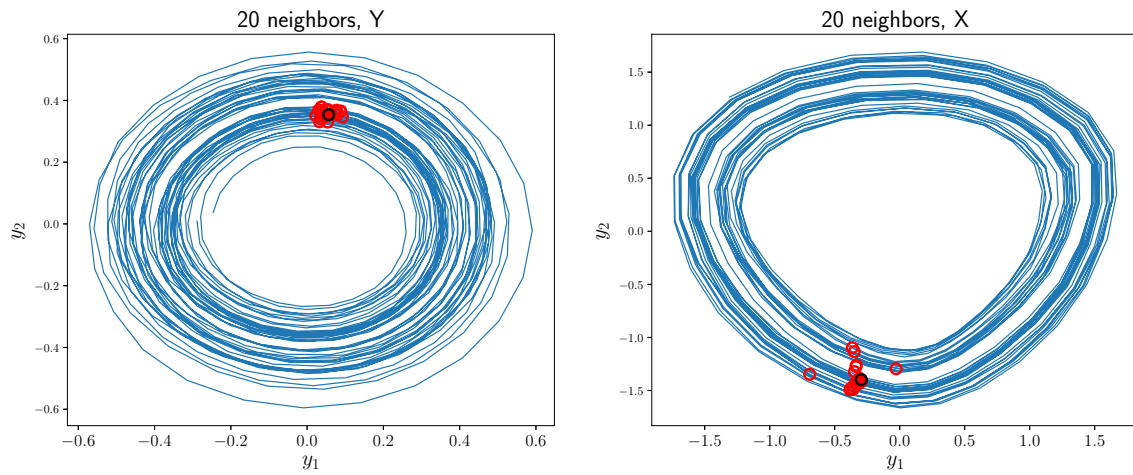
Исследуем зависимость ряда  $\mathbf{X}$  от ряда  $\mathbf{Y}$ . Возьмем  $t^* = 400, L = 170$ . На рис. 10 красным показаны ближайшие соседи вектора  $\mathbf{y}_{t^*}$



**Рис. 10:** Ближайшие соседи вектора  $\mathbf{y}_{t^*}$ , ширина окна  $L = 100$



**Рис. 11:** Ближайшие соседи векторов  $\mathbf{x}_{t^*}$  и  $\mathbf{y}_{t^*}$  на фазовых диаграммах с периодом 12 часов



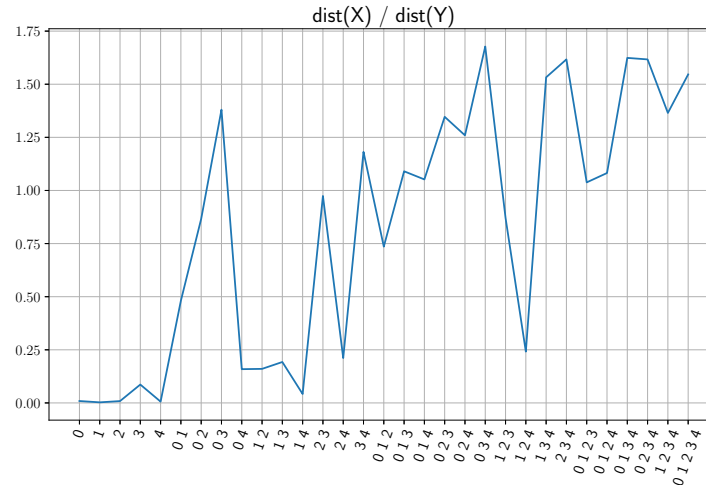
**Рис. 12:** Ближайшие соседи векторов  $\mathbf{x}_{t^*}$  и  $\mathbf{y}_{t^*}$  на фазовых диаграммах с периодом 24 часа

### 6.3 Перебор подпространств

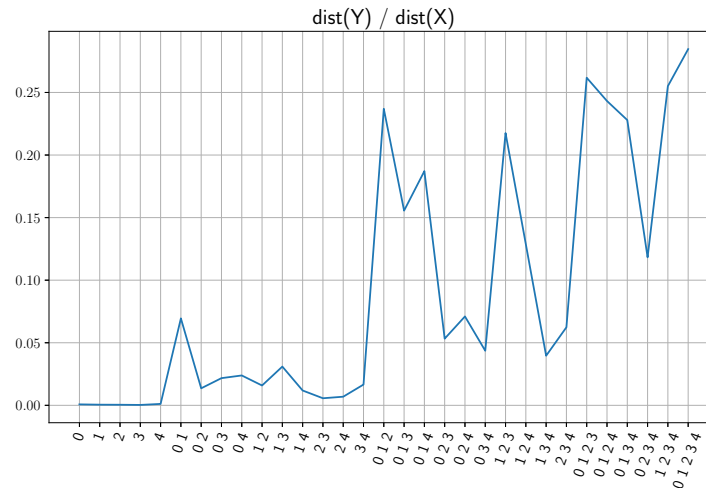
Пусть  $\mathbf{P}_x$  и  $\mathbf{P}_y$  – проекции рядов  $\mathbf{X}$  и  $\mathbf{Y}$  в некоторые фазовые подпространства. Для фиксированного  $t^*$  находим ближайших соседей  $\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_k}$  и соответствующие им  $\mathbf{y}_{t_1}, \mathbf{y}_{t_1}, \dots, \mathbf{y}_{t_k}$ . Здесь  $\mathbf{x}_t$  и  $\mathbf{y}_t$  – строки матриц  $\mathbf{P}_x$  и  $\mathbf{P}_y$  соответственно.

Будем перебирать различные комбинации из первых  $l$  главных компонент и соответствующие им подпространства  $\mathbf{P}_x$  и  $\mathbf{P}_y$ . Для каждого набора  $\mathcal{T}$  главных компонент будем находить среднее расстояние между  $k$  ближайшими соседями для ряда  $\mathbf{X}$  и между ближайшими соседями для ряда  $\mathbf{Y}$ .

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



**Рис. 13:** Отношение расстояния между ближайшими соседями ряда  $\mathbf{X}$  к расстоянию между соседями ряда  $\mathbf{Y}$ . Ближайшие соседи определяются по ряду  $\mathbf{X}$



**Рис. 14:** Отношение расстояния между ближайшими соседями ряда  $\mathbf{Y}$  к расстоянию между соседями ряда  $\mathbf{X}$ . Ближайшие соседи определяются по ряду  $\mathbf{Y}$