Погружение временных рядов с высокой волатильностью в метрическое пространство

Алтай Эйнуллаев Эльшан оглы

Московский физико-технический институт

Курс: Автоматизация научных исследований (практика, В. В. Стрижов)/Группа 105

Эксперт: д.ф-м.н. В.В.Стрижов

Консультант: К.Яковлев

2024

Цель исследования

Решается задача прогнозирования набора сильно коррелированных временных рядов с высокой волатильностью. Предлагается прогнозировать с помощью матрицы попарных расстояний между временными рядами набора. Исследуются влияние различных способов вычисления попарного расстояния на точность прогноза.

Постановка задачи

Пусть $X = \{\mathbf{x} = [x_1, \ldots, x_m]^T | x_i \in R\}$ — множество временных рядов, заданных своей реализацией. Обозначим через $\mathbf{Y} \in R^{n \times m}$ заданный набор из n временных рядов:

$$\mathbf{Y} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}]^T. \tag{1}$$

Через $\mathbf{Y}_t \in R^{n \times t}$ обозначим $t < m$ первых столбцов $\mathbf{Y}$:

$$\mathbf{Y}_t = [\mathbf{x}^{(1)}_{1:t}, \ldots, \mathbf{x}^{(n)}_{1:t}]^T. \tag{2}$$

.

Таким образом, по известной $\mathbf{Y}_t$ выполняется прогноз значений набора временных рядов в момент времени $t + 1$:

$$\mathbf{Y}_t \to \mathbf{x}_{t+1} = [x^{(1)}_{t+1}, \ldots, x^{(n)}_{t+1}]^T. \tag{3}$$

.

Решение

Определим функцию расстояния между временными рядами: $d : X \times X \to R$, удовлетворяющую условиям Мерсера.

$$d_t(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = d_t(i, j) \tag{4}$$

Таким образом, в каждый момент времени t набору временных рядов $\mathbf{Y}_t$ поставлена в соответствие матрица попарных расстояний $\Sigma_t \in \mathcal{S}_n^+$ (симметричная, неотрицательно определенная матрица). С помощью Score-Based Generative Models и LSTM прогнозируем матрицу попарных расстояний $\hat{\Sigma}_{t+1}$. Далее возвращаем прогноз в исходное пространство:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in R^n} \|\Sigma_{t+1} - \hat{\Sigma}_{t+1}\|. \tag{5}$$

Выбор функции попарных расстояний

Набор временных рядов в момент времени t:

$$\mathbf{X}_t = \begin{pmatrix} x^{(1)}_{t-\mathrm{L}+1} & x^{(1)}_{t-\mathrm{L}+2} & \cdots & x^{(1)}_t \\ x^{(2)}_{t-\mathrm{L}+1} & x^{(2)}_{t-\mathrm{L}+2} & \cdots & x^{(2)}_t \\ \cdots & \cdots & \cdots & \cdots \\ x^{(n)}_{t-\mathrm{L}+1} & x^{(n)}_{t-\mathrm{L}+2} & \cdots & x^{(n)}_t \end{pmatrix} \tag{6}$$

i-й столбец матрицы $\mathbf{X}_t$ — $\mathbf{y}^i$, i-я строка — $\mathbf{x}^{(i)}$.

$$\Sigma^1_t = \frac{1}{L}\sum_{i=1}^{L} \mathbf{y}^i \mathbf{y}^{iT} \tag{7}$$

Введем вектор $\mathbf{m} \in R^{\mathrm{n}}$ следующим образом:

$$\mathbf{m}_j = \frac{1}{L}\sum_{i=1}^{L} \mathbf{y}^i_j. \tag{8}$$

Выбор функции попарных расстояний

Второй способ:

$$\Sigma_t^2 = \frac{1}{L}\sum_{i=1}^{L}(\mathbf{y}^i - \mathbf{m})(\mathbf{y}^i - \mathbf{m})^T \tag{9}$$

Для каждой строки $\mathbf{x}^{(j)}$, $j \in 1, \ldots, n$ матрицы $\mathbf{X}$ введем величину:

$$\sigma_j = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(\mathbf{x}_i^{(j)} - \mathbf{m}_i)^2}. \tag{10}$$

Теперь введем матрицу Σ_t^3 следующим образом:

$$(\Sigma_t^3)_{ij} = \frac{(\Sigma_t^2)_{ij}}{\sigma_i \sigma_j}. \tag{11}$$

Вычислительный эксперимент

Эксперимент состоял в прогнозировании матрицы попарных расстояний для каждого способа ее вычисления.

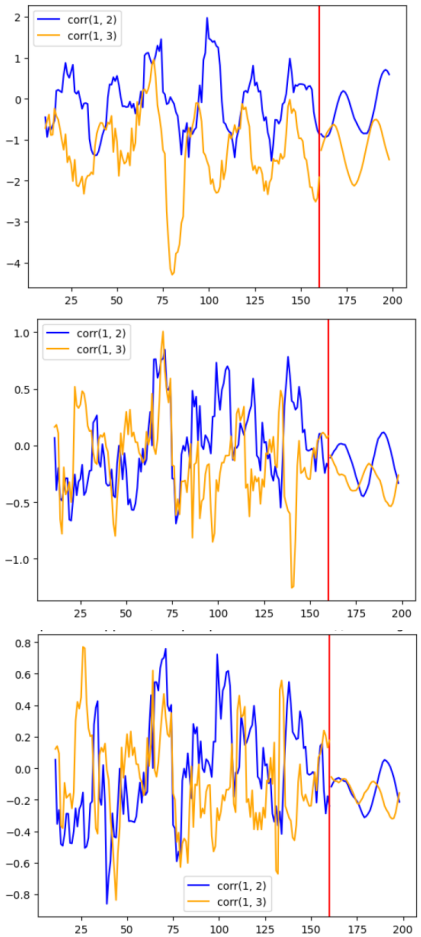

Рис.: Прогнозирование Σ_t^i на синтетическом наборе синусов

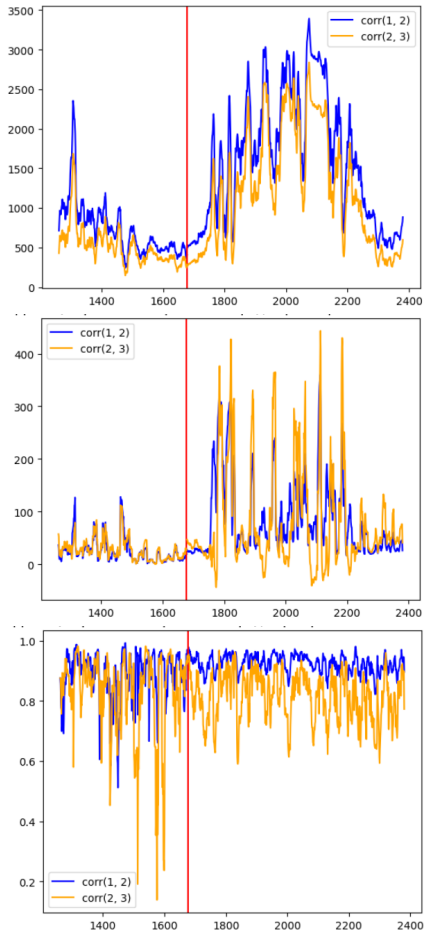

Рис.: Прогнозирование Σ_t^i на наборе цен на электричество

Анализ ошибки

Вычислим ошибку MAE каждого из прогнозов. В случае синтетического набора: $MAE_1 = 0.55$, $MAE_2 = 0.17$, $MAE_3 = 0.17$. На реальном наборе: $MAE_1 = 895.2$, $MAE_2 = 152.8$, $MAE_3 = 0.11$.

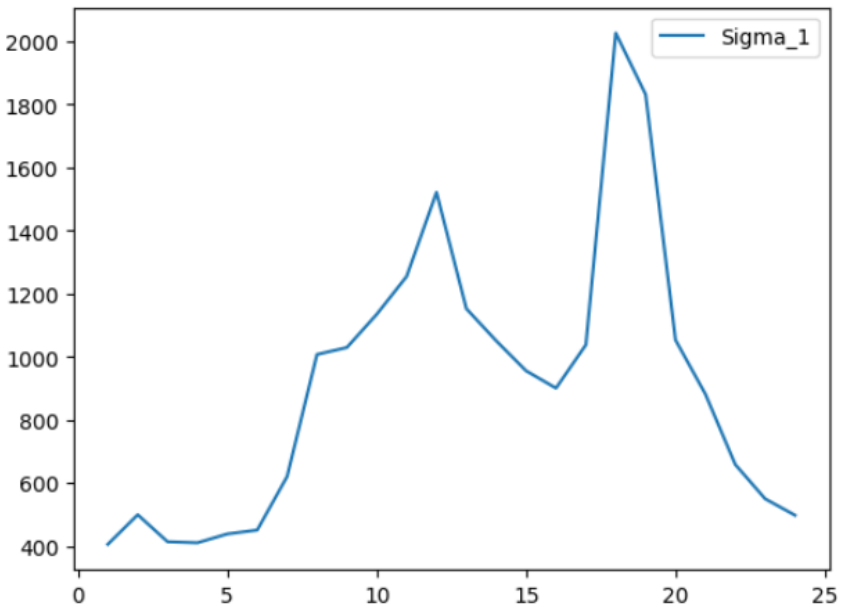

Рис.: Средняя MAE прогноза попарных расстояний для каждого из рядов при 1-ом способе подсчета Σ

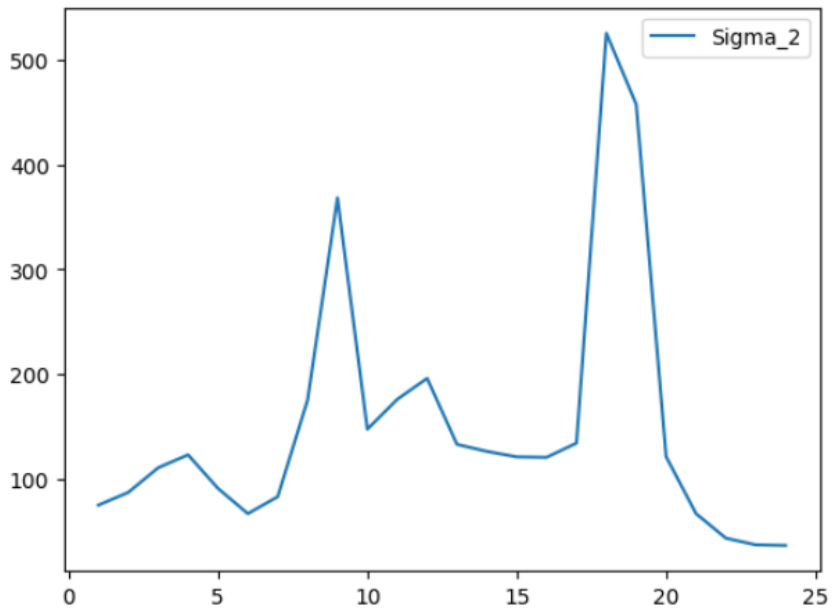

Рис.: Средняя MAE прогноза попарных расстояний для каждого из рядов при 2-ом способе подсчета Σ

Анализ ошибки

Наибольшая ошибка на временных рядах под номером $18, 19$ связанас высокой дисперсией соответствующих рядов. Отказываясь от их прогнозирования, получаем существенное улучшение прогноза:

$$\mathrm{MAE}_1 = 785.7 \tag{12}$$

$$\mathrm{MAE}_2 = 94.1 \tag{13}$$

Заключение

Перечислите ваши результаты

- Прогнозирования набора высоко коррелированных временных рядов с высокой волатильностью.
- Проведен анализ различных способов вычисления матрицы попарных расстояний и их прогноза.
- Осталось провести эксперимент с Score-Based Gen Models и оценить точность возвращения прогноза в исходное пространство.