Порождающие модели для прогнозирования (наборов временных рядов) в метрическом вероятностном пространстве [1]

A Preprint

Карпеев Глеб
Кафедра интеллектуальных систем
МФТИ
karpeev.ga@phystech.edu

Vadim Strijov
FRC CSC of the RAS
Moscow, Russia
strijov@phystech.edu

Яковлев Константин
Кафедра интеллектуальных систем
МФТИ
iakovlev.kd@phystech.edu

Abstract

Исследование посвящено проблеме прогнозирования временных рядов с высокой ковариацией. Задача решается для наборов временных рядов с высокой дисперсией, проявляющейся, например, в сигналах головного мозга или ценах финансовых активов. Для решения данной задачи предлагается построение пространства парных расстояний, представляющего метрическую конфигурацию временных рядов. Прогноз осуществляется в данном пространстве, а затем результат возвращается в исходное пространство с использованием метода многомерного шкалирования. В данной работе изучаются порождающие модели для прогнозирования наборов временных рядов в метрическом вероятностном пространстве. Новизна работы заключается в применении Римановых моделей для регрессии и использовании Римановых генеративных диффузных моделей.

Keywords Riemannian Generative Models · Trades

1 Introduction

Развитие технологий, которое мы наблюдаем в последние годы, открывает перед нами широкие возможности для получения данных, передаваемых человеком с помощью различных устройств. Анализ данных систем, предназначенных для мониторинга состояния человека, позволяет решать задачи в сфере здравоохранения, которые включают в себя анализ сигналов головного мозга [1], мониторинга физической активности [2]. Задача прогнозирования временного ряда является важной частью анализа сигналов и может использоваться во многих биомедицинских приложениях.

В данной работе предлагается метод прогнозирования временных рядов с высокой ковариацией и высокой дисперсией. Предлагаемое решение задачи прогнозирования состоит из трех этапов. Во-первых, осуществляется построение пространства парных расстояний, где используемая метрика удовлетворяет условию Мерсера. Во-вторых, выполняется прогноз матрицы попарных расстояний. В-третьих, результат возвращается в исходное пространство. В данной работе изучаются Римановы генеративные диффузные модели (RSGMs) [3] для выполнения прогнозирования матрицы попарных расстояний.

Классическими алгоритмами прогнозирования временных рядов являются метод SSA (Singular Spectrum Analysis) [4], LSTM [5], State space model [6]. Новизна предложенного метода заключается в том, что выполняется кодирование временных рядов с помощью матрицы расстояний, выполняется прогноз, а затем декодирование полученной матрицы.

Анализ предлагаемого метода прогнозирования проводится на синтетических и реальных данных. Синтетический набор данных построен на основе синусоидальных сигналах со случайной амплитудой и частотой. Реальные данные были получены с помощью акселерометра, а так же на основе финансовых временных рядов. Целью эксперимента является нахождение оптимальной модели для прогнозирования временных рядов.

2 Problem Statement

Даны временные ряды с высокой ковариацией и высокой дисперсией

$$x_1, x_2, \ldots, x_T \in \mathbb{R}^d, \tag{1}$$

где d — количество временных рядов.

Нужно спрогнозировать x_{T+1}.

Алгоритм:

1. Построить матрицу расстояний.

$$\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T}(x_t - \mu_T)(x_t - \mu_T)^T \tag{2}$$

$$\mu_T = \frac{1}{T}\sum_{t=1}^{T} x_t \tag{3}$$

2. Спрогнозировать матрицу $\hat{\Sigma}^s_{T+1} \approx \hat{\Sigma}_{T+1}|\hat{\Sigma}_T$

Базовый алгоритм - линейная регресиия:

$$\hat{\Sigma}^s_{T+1} = W \cdot \hat{\Sigma}_T \tag{4}$$

3. Найти такой оптимальный x_{T+1}, что ошибка прогнозирования временных рядов

$$\|\hat{\Sigma}^s_{T+1} - \hat{\Sigma}_{T+1}\|_2 \to \min_{x_{T+1}} \tag{5}$$

3 Computation experiment

Эксперимент проводился на реальных и синтетических данных. Синтетический набор данных построен на основе синусоидальных сигналах со случайной амплитудой и частотой. Реальные данные были получены с энергетической биржи Nord Pool [8] и представляют собой временной ряд цены на электроэнергию. Временной ряд энергии состоит из почасовых записей (всего 50826 наблюдений).

Для оценивания качества аппроксимации вычисляется значение среднеквадратичной ошибки.

$$MSE(y_{pred}, y_{true}) = \frac{1}{n}\sum_{i=1}^{n}(y_{pred} - y_{true})^2 \tag{6}$$

Сравним два базовых алгоритма предсказания временных рядов - SSA и MSSA.

3.1 Синтетические данные

Сгенерируем выборку из двух синусоидальных сигналов размера $N = 200$, с количеством периодов 2 и 4, соответственно. Предскажем последние 40 значений выборки, для предсказания будем использовать последние 60 сгенерированных значений.

6

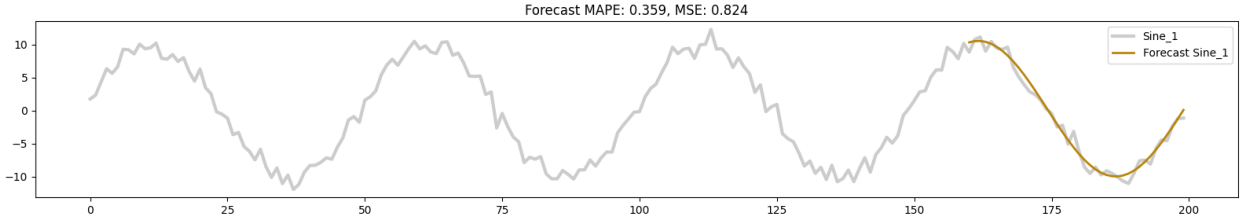

Рис. 1: Прогноз синуса с 4 периодами алгоритмом SSA

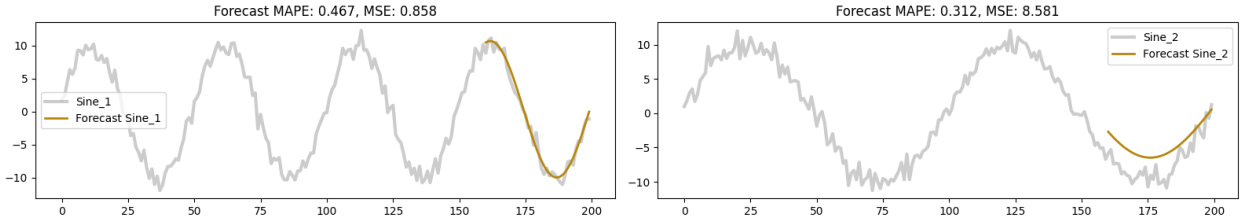

Рис. 2: Прогноз синуса с 2 и 4 периодами алгоритмом MSSA

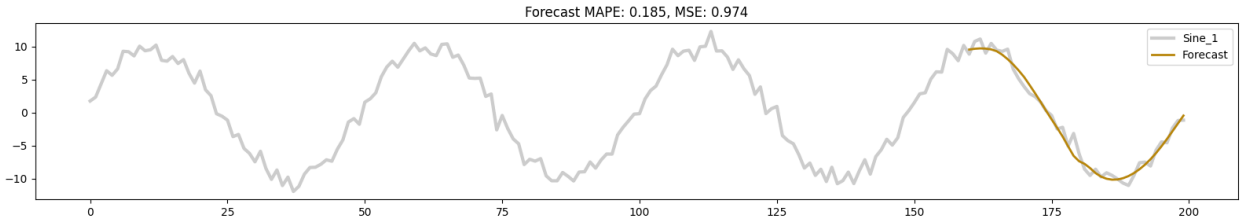

Рис. 3: Прогноз синуса с 4 периодами алгоритмом LSTM

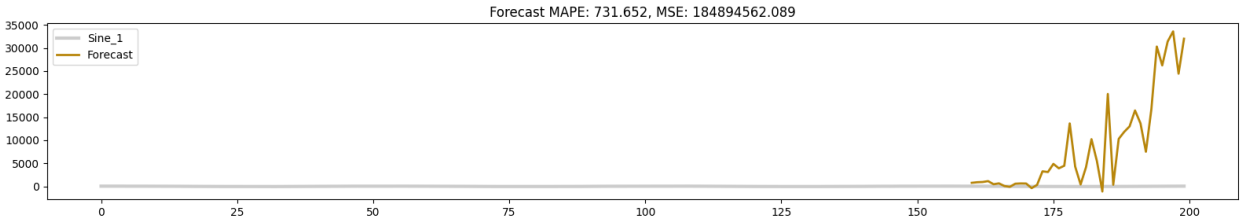

Рис. 4: Прогноз сильно зашумленного синуса с 4 периодами алгоритмом LSTM

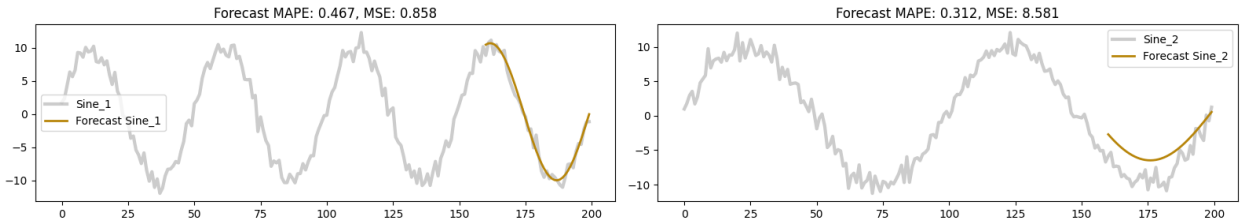

Рис. 5: Прогноз сильно зашумленного синуса с 2 и 4 периодами алгоритмом MSSA

3.2 Данные цен на электроэнергию [7]

Строка матрицы X — локальная история сигнала за одну неделю $n = 24 \times 7$. Строка матрицы Y — локальный прогноз потребления электроэнергии в следующие 24 часа. Прогноз выполняется с помощью алгоритма SSA.

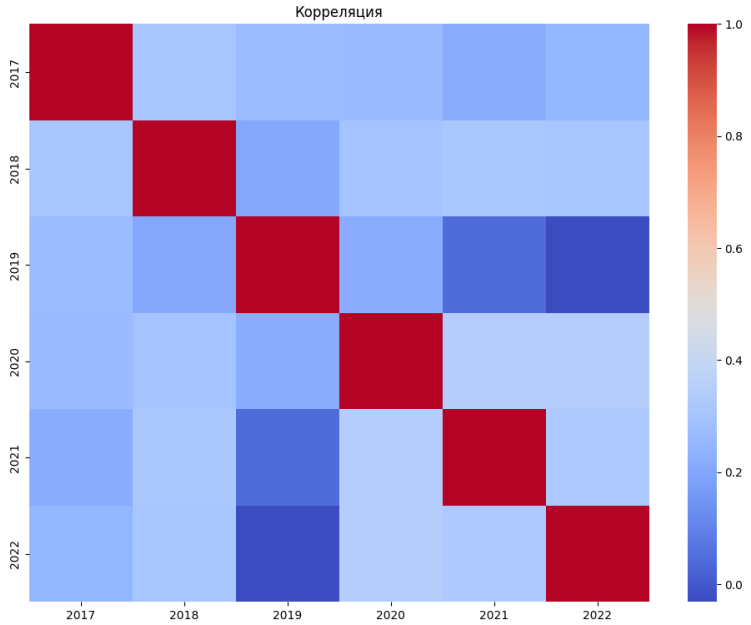

Рис. 6: Корреляция между временными рядами

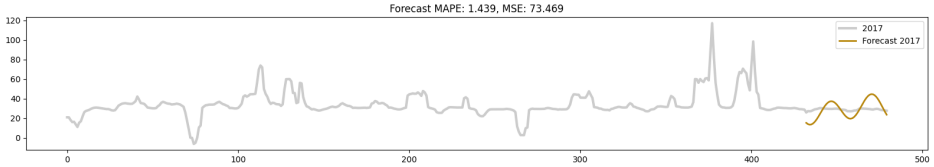

Рис. 7: Прогноз спотовых цен на электроэнергию алгоритмом SSA

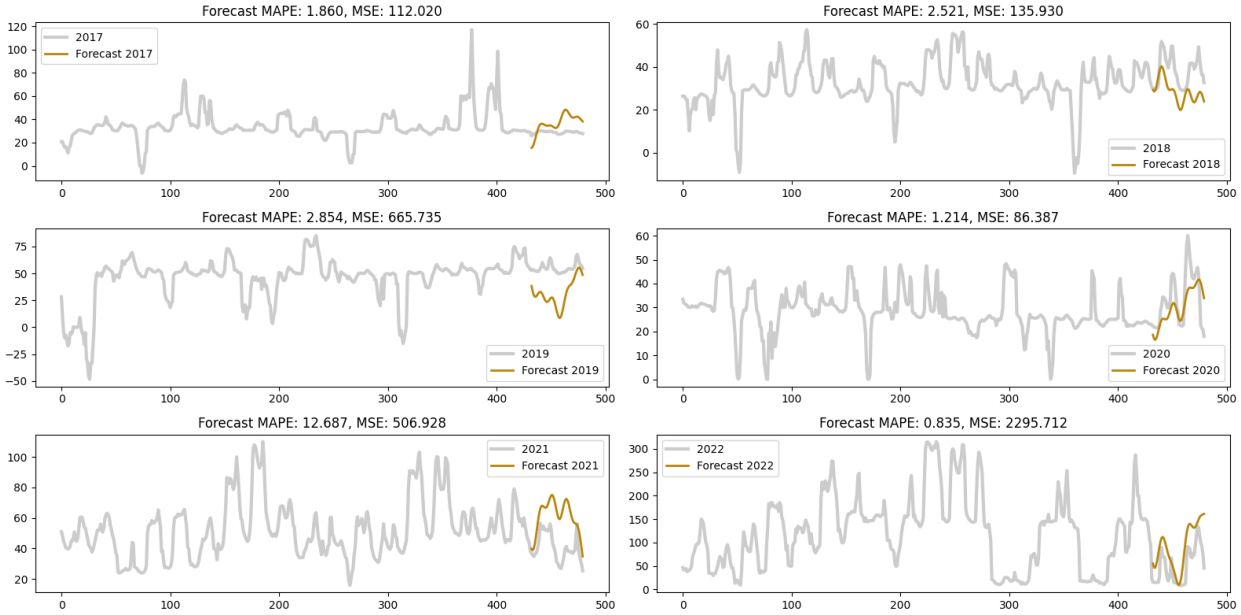

Рис. 8: Прогноз спотовых цен по годам на электроэнергию алгоритмом MSSA

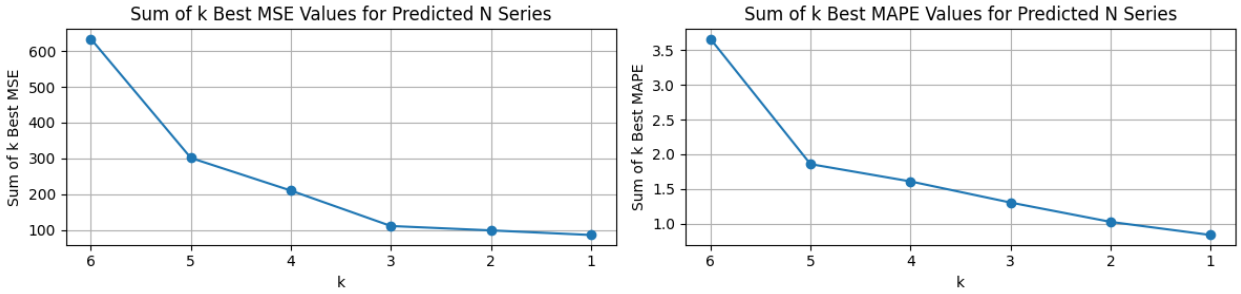

Рис. 9: Pareto front для MSSA прогноза

8

Список литературы

[1] S. Vaid, P. Singh and C. Kaur, "EEG Signal Analysis for BCI Interface: A Review, 2015 Fifth International Conference on Advanced Computing & Communication Technologies, Haryana, India, 2015

[2] Amit Purwar, Do Un Jeong and Wan Young Chung, "Activity monitoring from real-time triaxial accelerometer data using sensor network,"2007 International Conference on Control, Automation and Systems, Seoul, Korea (South), 2007

[3] Riemannian Score-Based Generative Modelling. Valentin De Bortoli, Émile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, Arnaud Doucet, 2022

[4] Elsner, J.B. and Tsonis, A.A. (1996): Singular Spectrum Analysis. A New Tool in Time Series Analysis, Plenum Press.

[5] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation.

[6] Koller D, Friedman N. (2009) Probabilistic Graphical Models. Cambridge, MA: MIT Press.

[7] Dataset for "Trades Quotes and Prices"

[8] Electricity Spot Price Data. https://www.kaggle.com/datasets/arashnic/electricity-spot-price

1. Поговорите с руководителями, это название явно рабочее. Лучше со скобками в заголовке не играться (по крайней мере, для первой статьи)

2. Посмотрите нотацию, которую я реомендовал на занятии (есть в презентации в канале).

3. В постановке задачи не хватает задачи оптимизации.
"нужно спрогнозировать" требуется переписать в формате формулы.

Алгоритм - это видимо предлагаемый метод.
Опишите его отдельной секции, опишите более подробно что происходит с выборкой,
почему используется такой вид матрицы расстояния и т.п.
Сейчас очень куце.

4. После формул ставят знаки препинания как после обычных частей предложения

5. Лучше "Эксперимент на синтетических данных"

6. Графики - ок, но лучше отрендерить с лучшим разрешением (через matplotlib.pylab.savefig)
и с бОльшим размером шрифта

7. См. пункт 5

8. Нужны выводы и заключение