Восстановление прогноза, сделанного в метрическом вероятностном пространстве, в исходное пространство (временных рядов)

A Preprint

Maxim Divilkovskiy
Chair of Data Analysis
MIPT
divilkovskii.mm@phystech.edu

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

Yakovlev Konstantin
Chair of Intellectual Systems
MIPT
iakovlev.kd@phystech.edu

Abstract

Решается задача поточечного прогнозирования набора временных рядов с высокой ковариацией и высокой дисперсией. Для решения данной задачи предлагается построить пространство парных расстояний. В этом пространстве прогнозируется матрица попарных расстояний, а затем по известной матрице восстанавливаются значения временных рядов. В данной статье изучается способ восстановления прогноза в пространстве временных рядов по известной матрице попарных расстояний. Показывается недостаточность одной матрицы попарных расстояний. Предлагается несколько алгоритмов, основанных на использовании нескольких матриц, построенных по различным временным интервалам с использованием попарной корреляции. Так же, в статье выводится общий вид восстановленных значений. Помимо этого, приводится оценка качества восстановления при добавлении шума в матрицы попарных расстояний.

Keywords Metric · Trades · Correlation · Time Series Forecasting

1 Введение

Цель данной работы заключается в представлении нового метода прогнозирования временных рядов, характеризующихся высокой попарной ковариацией и высокой дисперсией. Задача заключается в прогнозировании значений временных рядов в следующий момент по имеющимся историческим данным. Задача разбивается на три этапа: сначала исходное пространство временных рядов некоторым образом трансформируется в метрическое пространство при помощи построения матрицы попарных расстояний, затем в этом пространстве осуществляются прогнозы, необходимые для получения информации о временных рядах. Необходимые данные описаны в теоретической части статьи. Последним этапом, результат возвращается в исходное пространство. В статье исследуется последний этап данного метода.

Существующие способы предсказания временных рядов, такие как LSTM [3], SSA [2] и другие [6], [1] основаны на предсказании значения одного ряда. При этом, данные методы могут быть изменены для прогноза в том числе набора временных рядов, если рассматривать набор рядов как один многомерный ряд. Однако, такой подход не моделирует в явном виде зависимости между различными рядами. В данной работе предлагается анализировать изменение набора временных рядов, явно используя связи между ними. Подобное исследование проводится в статье [4], однако в ней делается упор на задаче feature selection. Данная задача заключается в выборе такого поднабора из исходных временных рядов, для которых возможно делать прогноз достаточного качества.

Новизна работы заключается в том, что прогнозирование делается не в исходном пространстве, а в пространстве попарных расстояний, способ построения которого предложен в статье. Преимущество данного метода заключается в том, что в реальных наборах временных рядов (природных, физиче-

ских, финансовых и т.д.), часто наблюдается зависимость, близкая к линейной. Эта дополнительная информация способна улучшить качество итогового прогноза. Помимо этого, прогнозируемую матрицу можно рассматривать как набор временных рядов. В этом случае размерность данных возрастает до $O(n^2)$ против n рядов, что увеличивает информативность входных данных.

Далее рассматриваются условия на функцию расстояния между рядами при которых существует способ восстановления значений временных рядов. Доказывается недостаточность одной матрицы для восстановления ответа. Предлагается два метода, использующие несколько матриц, для случая точного прогноза и для случая прогноза с погрешностью. Так же предлагается сам алгоритм восстановления ответов.

2 Постановка задачи

Ниже приведена постановка задачи поточечного прогнозирования набора временных рядов в общем виде.

Набор из d временных рядов задан t векторами:

$$[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t], \text{для всех } k : \mathbf{x}_k \in \mathbb{R}^d,$$

$\mathbf{x}_{t_i,k}$ задаёт собой значение ряда с индексом k в момент времени t_i.

Задача заключается в прогнозе $\mathbf{x}_{t+1}$.

$\mathbf{x}$ рассматривается как *многомерный* временной ряд, рассматривая значение в точке как элемент пространства $\mathbb{R}^d$.

В качестве критериев качества используются MSE и MAE. В статье [5] показано, что они являются наиболее подходящими для задачи прогнозирования временных рядов.

Общий вид алгоритма при прогнозе одной матрицы расстояний

1. Строятся матрицы расстояний по предыдущим шагам. Способ построения описан далее.

$$\begin{gathered}[\mathbf{x}_1, \ldots, \mathbf{x}_s] \to \Sigma_s \\ [\mathbf{x}_2, \ldots, \mathbf{x}_{s+1}] \to \Sigma_{s+1} \\ \vdots \\ [\mathbf{x}_{t-s}, \ldots, \mathbf{x}_t] \to \Sigma_t\end{gathered}$$

2. По этим матрицам прогнозируется матрица $\hat{\Sigma}_{t+1}$

3. Найти такой $\hat{x}_{t+1}$, что

$$||\hat{\Sigma}_{t+1} - \bar{\Sigma}_{t+1}||_2^2$$

минимальна, где $\bar{\Sigma}_{t+1}$ — матрица расстояний, построенная по набору $[\vec{x}_{t-s+1}, \ldots, \hat{x}_{t+1}]$. Достижение минимума этой функцией будет означать равенство $\hat{\Sigma}_{t+1}$ и $\bar{\Sigma}_{t+1}$. В свою очередь, это означает что найденное продолжение ряда на момент времени $t+1$ имеет матрицу расстояний, равную прогнозу. В общем случае, данная функция невыпуклая и минимумов может быть несколько.

3 Метрика

Остутствие единственности решения задачи минимизации, описанной выше, является центральной проблемой, рассматриваемой в данной статье. В данной секции показывается, что только по матрице, построенной по произвольной метрике, невозможно однозначно восстановить значение рядов в следующий момент времени.

В данном параграфе рассматривается возвращение прогноза из матрицы Σ_{t+1} в пространство временных рядов в предположении, что Σ_{t+1} есть матрица попарных расстояний, отвечающая многомерному ряду $\mathbf{x} = [\mathbf{x_1}, \ldots, \mathbf{x_{t+1}}]$.

Недостаточность одной матрицы попарных расстояний

Пусть дана предсказанная матрица попарных расстояний Σ размера $d \times d$ для многомерного временного ряда $\overline{X} \in \mathbb{R}^{d \times (t+1)}$. Предсказывается $y \in \mathbb{R}^d$. Так же, известна метрика $d : \mathbb{R}^{t+1} \times \mathbb{R}^{t+1} \to \mathbb{R}$, введённая на временных рядах, обладающая свойствами метрики. То есть, $\Sigma_{i,j} = d(X_i \circ y_i, X_j \circ y_j)$, где $\circ$ означает конкатенацию векторов.

В качестве примера рассмотрим евклидову метрику:

$$d(X_i \circ y_i, X_j \circ y_j) = \sqrt{\left(\sum_{k=1}^{t}(X_{ik} - X_{jk})^2\right) + (y_i - y_j)^2}.$$

Использование данной метрики приводит к тому, что прибавление ко всем y_i некоторой константы C не изменяет ответ. В случае задачи предсказывания временных рядов это свойство критично, поскольку даже в случае точного предсказания матрицы Σ существует бесконечно много значений временных рядов в момент времени $t + 1$, отвечающих этой матрице.

Это приводит к невозможности использования алгоритма MDS для восстановления ответа в исходное пространство временных рядов.

Однако, даже использование других метрик не позволяет избавиться от проблемы.

Теорема 1. *Для любой метрики, введённой в пространстве временных рядов $\mathbb{R}^t$, существует более одного способа восстановить исходные временные ряды из построенной по данной метрике матрице попарных расстояний.*

Доказательство. Достаточно показать, что метрика не является биекцией. Это будет означать, что существуют несколько различных пар рядов, расстояние между которыми одинаковое.

Покажем, что метрика — непрерывная функция. Возьмём последовательность

$$\{(x_n, y_n)\} \subset \mathbb{R}^t \times \mathbb{R}^t, (x_n, y_n) \to (x, y).$$

Тогда,

$$x_n \to x, y_n \to y \Rightarrow d(x_n, x) \to 0, d(y_n, y) \to 0,$$

$n \to \infty$. Воспользовавшись неравенством треугольника для метрики, получаем

$$d(x_n, y_n) \leq d(x_n, x) + d(x, y) + d(y_n, y) \to d(x, y),$$

следовательно, $d(x_n, y_n) \to d(x, y)$.

То есть метрика — непрерывное отображение из $\mathbb{R}^t \times \mathbb{R}^t$ в $\mathbb{R}$. Покажем, что такое отображение не может быть гомеоморфизмом. Предположим, что $f : \mathbb{R} \to \mathbb{R}^t \times \mathbb{R}^t$ — искомый гомеоморфизм. Возьмём некоторую точку $a \in \mathbb{R}$ и $f(a)$. Выкинув точку a, $\mathbb{R}$ перестаёт быть связным, а $\mathbb{R}^t \times \mathbb{R}^t$ нет. Значит, это не гомеоморфизм. Противоречие. ■

Замечание. Существенно, в доказательстве используется только непрерывность функции. Это означает, что даже не метрические функции не дадут единственность ответа. Например, попарная корреляция рядов тоже является непрерывной функцией.

Таким образом, зная только матрицу расстояний невозможно однозначно восстановить исходные ряды.

Рассмотрим ту же задачу, помимо матрицы Σ_{t+1} воспользовавшись значением ряда до момента времени t: $\mathbf{x} = [\mathbf{x_1}, \ldots, \mathbf{x_t}]$. Задача переформулируется следующим образом:

Имеется n объектов в $\mathbb{R}^{t+1}$, известны их первые t координат. Так же известна матрица расстояний $\Sigma_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$. Требуется восстановить $t + 1$ координату каждого из объектов. В терминах временных рядов, $t + 1$-я координата является значением каждого из рядов в этот момент времени.

4 Попарная корреляция

В данной секции исследуется восстановление ответа при помощи матрицы попарной корреляции. Такая функция расстояния используется, поскольку в статье [7] показано, что оценка попарной корреляции выборки аппроксимирует своё математическое ожидание.

Матрица попарных расстояний строится следующим образом:

$$\Sigma_T = \frac{1}{T}\sum_{t=1}^{T}(x_t - \mu_T)(x_t - \mu_T)^\mathsf{T}$$

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

Теорема 2. *В случае, если мы точно спрогнозировали матрицу расстойний, функция* $||\hat{\Sigma}_{t+1} - \bar{\Sigma}_{t+1}||_2^2$ *будет иметь два минимума, задающихся явно следующим образом:*

$$\hat{y_i} = y_i$$

$$\hat{y_i} = \frac{2}{T-1}\sum_{k=1}^{T-1} a_{ik} - y_i,$$

где $\hat{y_i}$ *— i-я координата предсказываемого значения ряда в момент* $T+1$*,* $A = (a_{ik})$ *— исходный многомерных временной ряд,* y_i *— истинные значения ряда в момент* $T+1$*.*

Доказательство. Обозначим Σ — истинную матрицу в момент времени T, а $\hat{\Sigma}$ — спрогнозированную. По построению, $\Sigma = \frac{1}{T}\sum_{k=1}^{T}(a_k^T - \mu_T)(a_k^T - \mu_T)^T$. Матрица A представляет собой транспонированную матрицу X временных рядов, первая размерность — номер ряда, а не момент времени как в случае с X. Тогда, рассмотрим чему равны элементы матриц Σ и $\hat{\Sigma}$.

$$\Sigma_{ij} = \frac{1}{T}\sum_{k=1}^{T}(a_{ik} - \mu_i)(a_{jk} - \mu_j)$$

$$\hat{\Sigma}_{ij} = \frac{1}{T}\sum_{k=1}^{T-1}(a_{ik} - \hat{\mu}_i)(a_{jk} - \hat{\mu}_j) + (y_i - \hat{\mu}_i)(y_j - \hat{\mu}_j)$$

Поскольку мы минимизируем норму разности, обе матрицы равны друг другу. Рассмотрим равенство диагональных элементов.

$$\forall i, j : i = j : \Sigma_{ii} = \hat{\Sigma}_{ii}$$

$$\sum_{k=1}^{T}(a_{ik} - \mu_i)(a_{ik} - \mu_i) = \sum_{k=1}^{T-1}(a_{ik} - \hat{\mu}_i)(a_{ik} - \hat{\mu}_i) + (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)$$

$$(a_{iT} - \mu_i)^2 = (y_i - \hat{\mu}_i)^2$$

Распишем $\hat{\mu}_i$ и $\hat{\mu}$:

$$\hat{\mu}_i = \frac{1}{T}\sum_{k=1}^{T-1} a_{ik} + \frac{1}{T}y_i$$

$$\mu_i = \frac{1}{T}\sum_{k=1}^{T} a_{ik}$$

$$\left[\begin{array}{l} \frac{T-1}{T}y_i - \frac{1}{T}\sum_{k=1}^{T-1} a_{ik} = a_{iT} - \mu_i \\ \frac{T-1}{T}y_i - \frac{1}{T}\sum_{k=1}^{T-1} a_{ik} = \mu_i - a_{iT} \end{array}\right.$$

$$\left[\begin{array}{l} \frac{T-1}{T}y_i = \frac{T-1}{T}a_{iT} \\ \frac{T_1}{T}y_i = \frac{2}{T}\sum_{k=1}^{T-1} a_{ik} - \frac{T-1}{T}a_{iT} \end{array}\right.$$

$$\left[\begin{array}{l} y_i = a_{iT} \\ y_i = \frac{2}{T-1}\sum_{k=1}^{T-1} a_{ik} - a_{iT} \end{array}\right.$$

Переобозначив, получаем

$$\hat{y_i} = y_i,$$
$$\hat{y_i} = \frac{2}{T-1} \sum_{k=1}^{T-1} a_{ik} - y_i.$$

■

Используя эту формулу, достаточно найти только один из минимумов, второй ищется через первый.

4.1 Алгоритм прогноза

Представленные выше алгоритм возвращает два ответа, из которых невозможно выбрать нужный.

Предлагается следующий алгоритм:

1. Зафиксируем T и $T' : T \neq T'$.

2. Для T и T' произведем полученный выше алгоритм и получим наборы ответов: $[ans_1, ans_2], [ans'_1, ans'_2]$.

3. Найдём тот ответ, который лежит в пересечении.

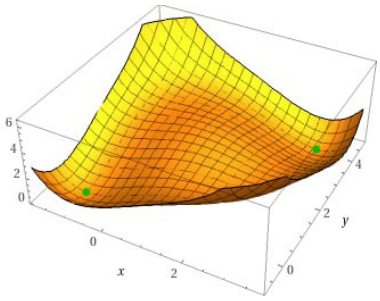

Рис. 1: Вид функции для двух рядов: (13) и (24). Точки минимума: (3; 4) — искомая

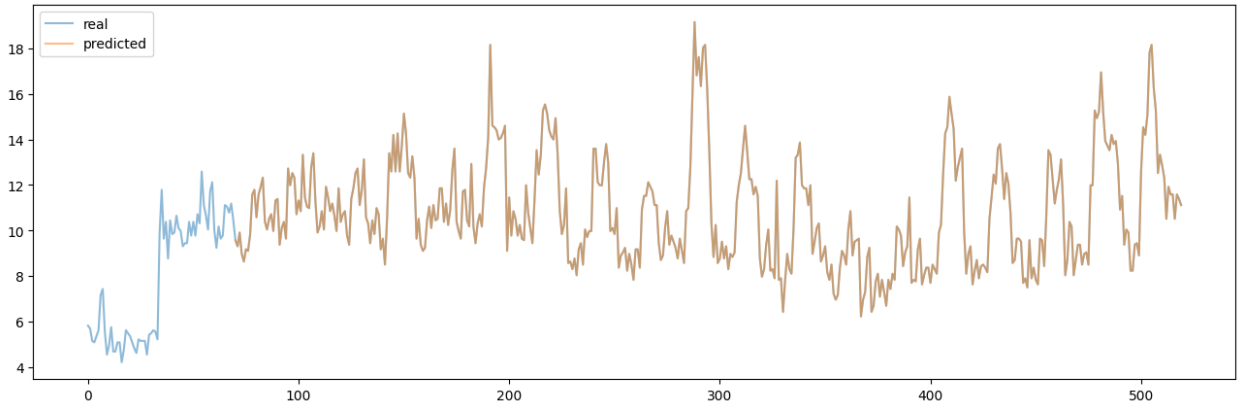

Рис. 2: Возвращение прогноза при идеальном прогнозе Sigma. $T = 20$, $T' = 10$

4.2 Алгоритм при неидеальном прогнозе

Проблема вышеописанного алгоритма заключается в том, что при неидеальном прогнозе, пересечения может не быть. Более того, если брать два ответа с минимальным расстоянием, мы можем взять не ту пару ответов.

Вместо двух значений T и T' предлагается брать K значений.

Тогда мы молучим K наборов ответов:

$$[ans_{11}, ans_{12}],$$
$$[ans_{21}, ans_{22}],$$
$$\vdots$$
$$[ans_{K1}, ans_{K2}]$$

Далее предлагается перебрать 2^K наборов ответов и выбрать тот набор, в котором диаметр минимален. Асимптотическая сложность данного восстановления будет $O(2^K \times K \times N)$ + сложность используемого алгоритма поиска минимума.

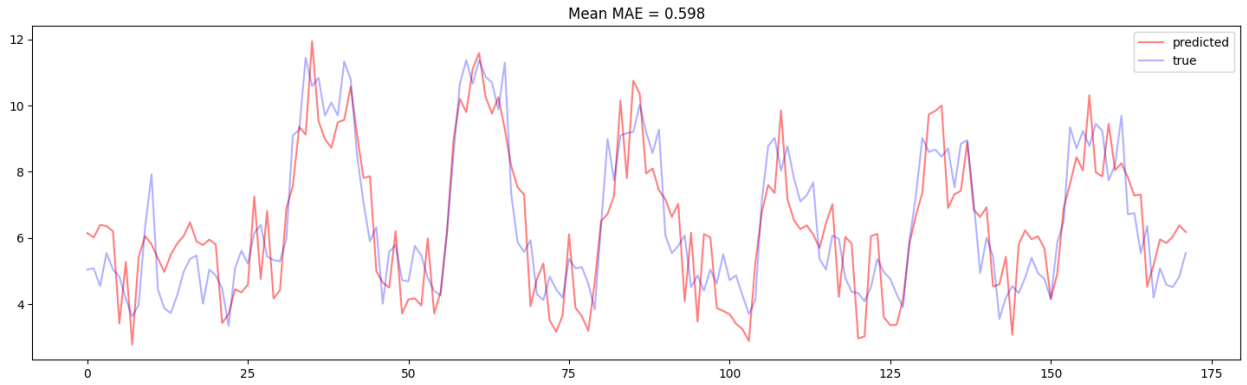

Рис. 3: Возвращение прогноза при неидеальном прогнозе Sigma при помощи Bidirectional LSTM

5 Computational Experiment

Исследуются следующие алгоритмы прогнозирования:

- LSTM [3]
- SARIMA [8]
- LSTM на матрице расстояний
- Bidirectional LSTM на матрице расстояний

5.1 LSTM

LSTM, в отличии от обыкновенной RNN позволяет выделять как кратковременные, так и долговременные зависимости, что позволяет с довольно высокой точностью прогнозировать временные ряды.

В качестве теста используется зашумленный временной ряд длины T, состоящий из суммы синусов и косинусов разных амплитуд и сдвигов. Из этого временного ряда генерируется выборка следующим алгоритмом:
1. Выбирается размер окна W.
2. Ряд разбивается на $T - W - 1$ окон размера $W + 1$ со сдвигом 1. Эти окна будут семплами
3. В каждом из полученных окон первые W будут аргументами на данном семпле, а последнее — результатом.

Ряд восстанавливается неплохо, однако минусом является то, что при усложнении данных сильно растёт сложность модели. Так же, LSTM не может работать с многомерными рядами.

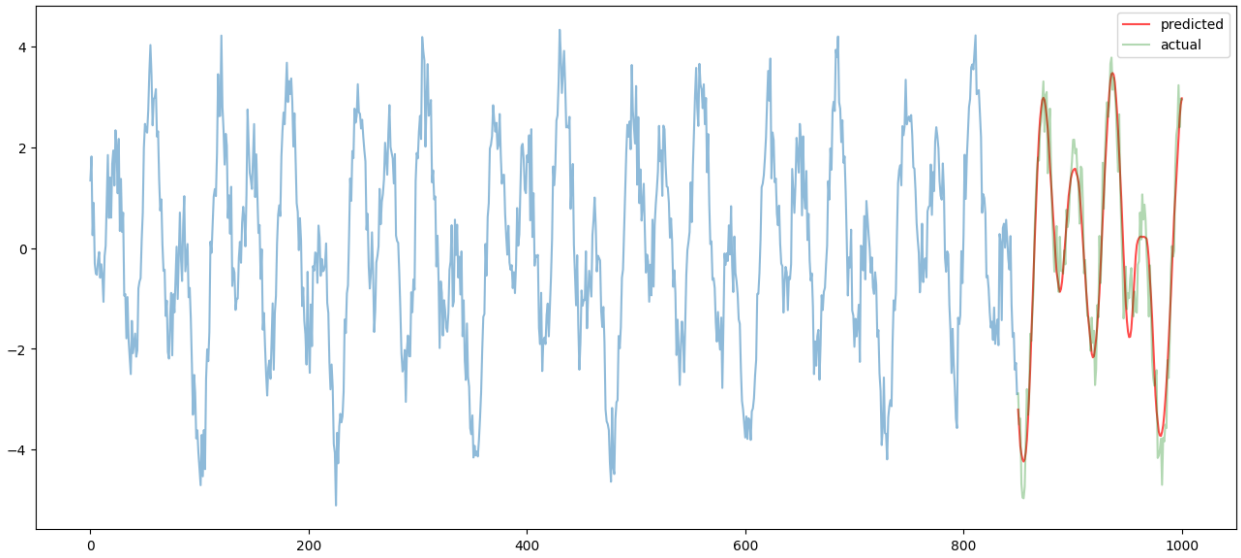

Рис. 4: Прогноз с использованием LSTM

5.2 SARIMA

ARIMA позволяет находить авторегрессионные зависимости. SARIMA (Seasonal ARIMA) учитывает так же сезонность данных. Это может быть полезным в случае с данными природного характера, как например, температура воздуха или выработка электричества.

Ряд прогнозируется довольно плохо, в случае если он имеет достаточно нетривиальную структуру. Так же, в данных может не быть явной сезонности, что ухудшает точность данного метода.

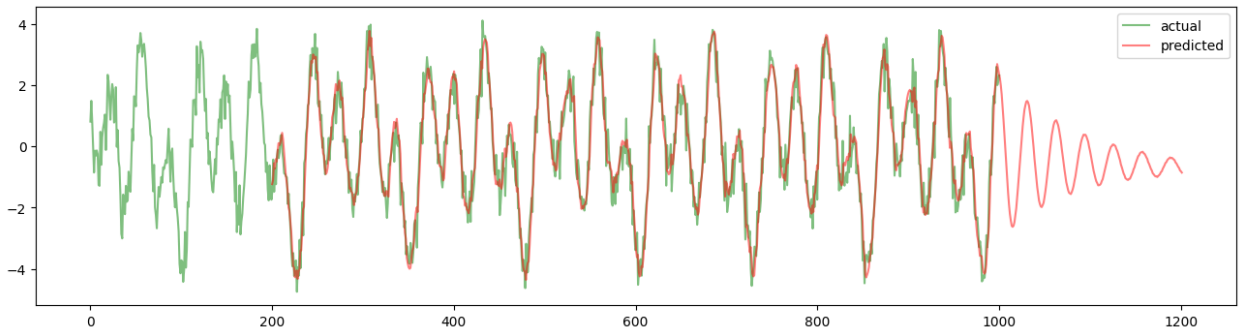

Рис. 5: Прогноз с использованием SARIMA

5.3 LSTM на матрице попарных корреляций

5.4 Bidirectional LSTM на матрице попарных корреляций

Список литературы

[1] Stephen Boyd, Enzo Busseti, Steven Diamond, Ronald N. Kahn, Kwangmoo Koh, Peter Nystrup, and Jan Speth. Multi-period trading via convex optimization, 2017.

[2] James B. Elsner and Anastasios A. Tsonis. Singular spectrum analysis: A new tool in time series analysis. 1996.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[4] Roman Isachenko and Vadim Strijov. Quadratic programming feature selection for multicorrelated signal decoding with partial least squares. *Expert Systems with Applications*, 207:117967, 11 2022.

[5] Aryan Jadon, Avinash Patil, and Shruti Jadon. A comprehensive survey of regression based loss functions for time series forecasting, 2022.

[6] Anastasia Motrenko and Vadim Strijov. Extracting fundamental periods to segment biomedical signals. *IEEE journal of biomedical and health informatics*, 20, 08 2015.

[7] Nikita Puchkin, Fedor Noskov, and Vladimir Spokoiny. Sharper dimension-free bounds on the frobenius distance between sample covariance and its expectation, 2023.

[8] Sima Siami-Namini and Akbar Siami Namin. Forecasting economics and financial time series: Arima vs. lstm, 2018.