Санкт-Петербургский государственный университет

Прикладная математика и информатика

Отчет по учебной практике 1 (научно-исследовательской работе) (семестр 1)

Использование метода SSA в машинном обучении для прогноза временных рядов

Выполнил:

Ежов Федор Валерьевич

группа 20.Б03-мм

Научный руководитель:

к.физ.-мат.н., доцент

Голяндина Нина Эдуардовна

Кафедра Статистического Моделирования

Санкт-Петербург

2020

Оглавление

Введение 3

Глава 1. Singular Spectrum Analysis 4

1.1. Базовый алгоритм SSA 4

1.1.1. Этап 1. Построение траекторной матрицы (Вложение) 4

1.1.2. Этап 2. Singular Value Decomposition (SVD) 5

1.1.3. Этап 3. Группировка 5

1.1.4. Этап 4. Диагональное усреднение 5

1.2. Пример разложения ряда 6

Глава 2. Использование SSA в машинном обучении 10

2.1. Статья №1 10

2.2. Статья №2 12

2.3. Статья №3 14

2.4. Статья №4 16

Заключение 18

Список литературы 19

Введение

Метод Singular Spectrum Analysis (SSA) — хорошо развитая методология анализа и прогнозирования временных рядов, которая включает в себя множество различных, но взаимосвязанных методов. Область применения SSA очень широка — от непараметрической декомпозиции и фильтрации временных рядов до оценки параметров и прогнозирования.

В этой работе были поставлены следующие задачи: разобраться в методике SSA, в частности в базовом алгоритме SSA. Провести самостоятельное разложения временного ряда методом SSA с помощью библиотеки Rssa на языке R. Рассмотреть статьи, где метод SSA применяется вместе с алгоритмами machine learning для расширения знаний практического применения метода SSA.

Глава 1

Singular Spectrum Analysis

Метод SSA используется для разложение исходного ряда в сумму рядов, которые легко интерпретировать и понять их поведение. Обычно исходный ряд раскладывается в сумму трех рядов: тренд — медленно меняющаяся компонента, сезонность — циклическая компонента с фиксированным периодом и шум.

1.1. Базовый алгоритм SSA

Базовый SSA состоит из четырех этапов:

1. Построение траекторной матрицы (Вложение).
2. SVD.
3. Группировка.
4. Диагональное усреднение.

Рассмотрим каждый этап подробнее.

Пусть $F = (f_0, \ldots, f_{N-1})$ — временной ряд, где $N > 2$. Также будем предполагать, что найдется хоть одно $f_i \neq 0$, то есть ряд не нулевой. Обычно считается, что $f_i = f(i\Delta)$ для некоторой функции $f(t)$, где t — время, а Δ — некоторый временной интервал.

1.1.1. Этап 1. Построение траекторной матрицы (Вложение)

Выберем целое L — длина окна, такое что $1 < L < N$. Тогда $K = N - L - 1$. Построим вектора $X_i = (f_{i-1}, \ldots, f_{i+L-2})^T$, для $1 \leqslant i \leqslant K$. Составим из векторов X_i траекторную матрицу:

$$\mathbf{X} = [X_1 : \ldots : X_K] = \begin{pmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_K \\ f_2 & f_3 & f_4 & \cdots & f_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1} \end{pmatrix}.$$

Получили матрицу $\mathbf{X}$ размерностью $L \times K$, составленную из пересекающихся частей исходного временного ряда. Можно заметить, что на побочных диагоналях стоят одинаковые числа, такая матрица называется ганкелевой. Существует взаимно-однозначное соответствие между ганкелевыми матрицами $L \times K$ и рядами длиной $N = L + K - 1$.

1.1.2. Этап 2. Singular Value Decomposition (SVD)

На данном этапе применяется метод SVD к траекторной матрице $\mathbf{X}$. Пусть $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ и $\lambda_1 > \ldots > \lambda_L$ — собственные числа матрицы $\mathbf{S}$, $U_1, \ldots, U_L$ — ортонормированная система базисных векторов, соответствующих собственным числам. Обозначим $V_i = \frac{X^T U_i}{\sqrt{\lambda_i}}$ и $d = max\{i : \lambda_i > 0\}$. Тогда сингулярное разложение матрицы $\mathbf{X}$ запишется следующим образом:

$$\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d, \text{ где } \mathbf{X}_i = \sqrt{\lambda_i} U V^T,$$

Набор $(\sqrt{\lambda_i}, U_i, V_i^T)$ будем называть i-й собственной тройкой.

1.1.3. Этап 3. Группировка

На этапе группировки все значения $1 \ldots d$ делятся на m непересекающихся групп $I_1, \ldots, I_m$. Пусть, $I_j = \{i_1, \ldots, i_p\}$, тогда результирующая матрица соответствующая группе I_j имеет вид: $\mathbf{X}_{I_j} = \mathbf{X}_{i_1} + \ldots + \mathbf{X}_{i_p}$.

Такие матрицы вычисляются для каждой группы, тем самым можно записать разложение для матрицы в сгруппированном виде:

$$\mathbf{X} = \sum_{j=1}^{m} \mathbf{X}_{I_j},$$

Процедура составление групп I_j называется группировкой собственных троек. Она подробно описана в книге Analysis of Time Series Structure: SSA and Related Techniques 2001 [1].

1.1.4. Этап 4. Диагональное усреднение

Пусть $\mathbf{Y}$ — матрица $L \times K, L < K$. y_{ij} - элементы матрицы, где $1 \leqslant i \leqslant L, 1 \leqslant j \leqslant K$. Также пусть $N = L + K - 1$. Диагональное усреднение преобразует матрицу $\mathbf{Y}$ в ряд $g_0, \ldots, g_{N-1}$ по формуле:

$$g_k = \begin{cases} \frac{1}{k+1}\sum_{m=1}^{k+1} y_{m,k-m+2} & \text{, для } 0 \leqslant k < L-1 \\ \frac{1}{L}\sum_{m=1}^{L} y_{m,k-m+2} & \text{, для } L-1 \leqslant k < K \\ \frac{1}{N-k}\sum_{m=k-K+2}^{N-K+1} y_{m,k-m+2} & \text{, для } K \leqslant k < N \end{cases}$$

Применяя диагональное усреднение к каждой результирующей матрицы, получаем m рядов $F^{(k)} = (f_1^{(k)} \dots f_{N-1}^{(k)})$. Тогда исходные ряд F раскладывается в сумму рядов:

$$F = \sum_{k=1}^{m} F^{(k)}.$$

1.2. Пример разложения ряда

Продемонстрируем работу метода SSA. Пользоваться будем библиотекой *Rssa* из языка R. В качестве исходного ряда возьмем простую функцию (рис. 1) $f(t) = f_1(t) + f_2(t) + \epsilon(t)$, где $1 \leqslant t < 250$, $f_1(t) = 0.05t + 5$ – тренд, $f_2(t) = \sin(2\pi\frac{t}{T})$ – сезонность, период $T = 25$ и $\epsilon(t)$ - гауссовский шум в точке t.

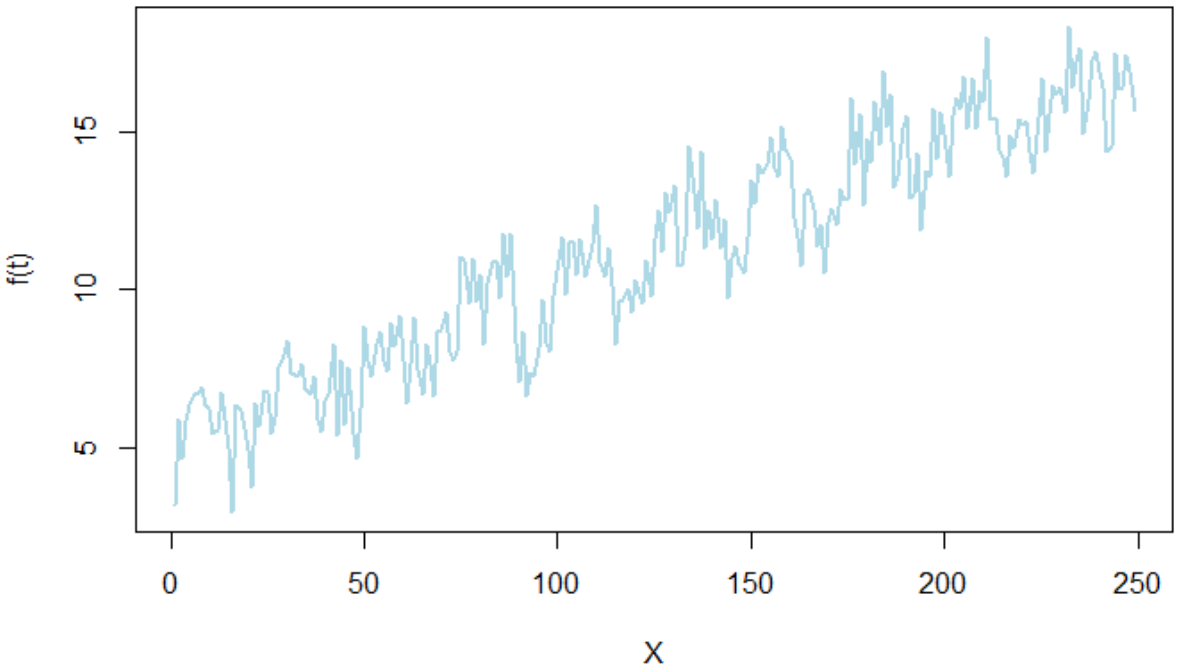

Рис. 1

По теории рекомендуется брать длину окна $L = N/2$. Также $f_2(t)$ – гармонический ряд и $\omega < 1/2$, если K и L делятся нацело на T, то элементы фазовых векторов этого

ряда в этом случае будут представлены *sin* и *cos*. Зададим $L = 125$. Применим метод *ssa* из библиотеки *Rssa* к исходному ряду и посмотрим на нормы компонент (рис. 2).

```
s <- ssa(F, L=125)
plot(s)
```

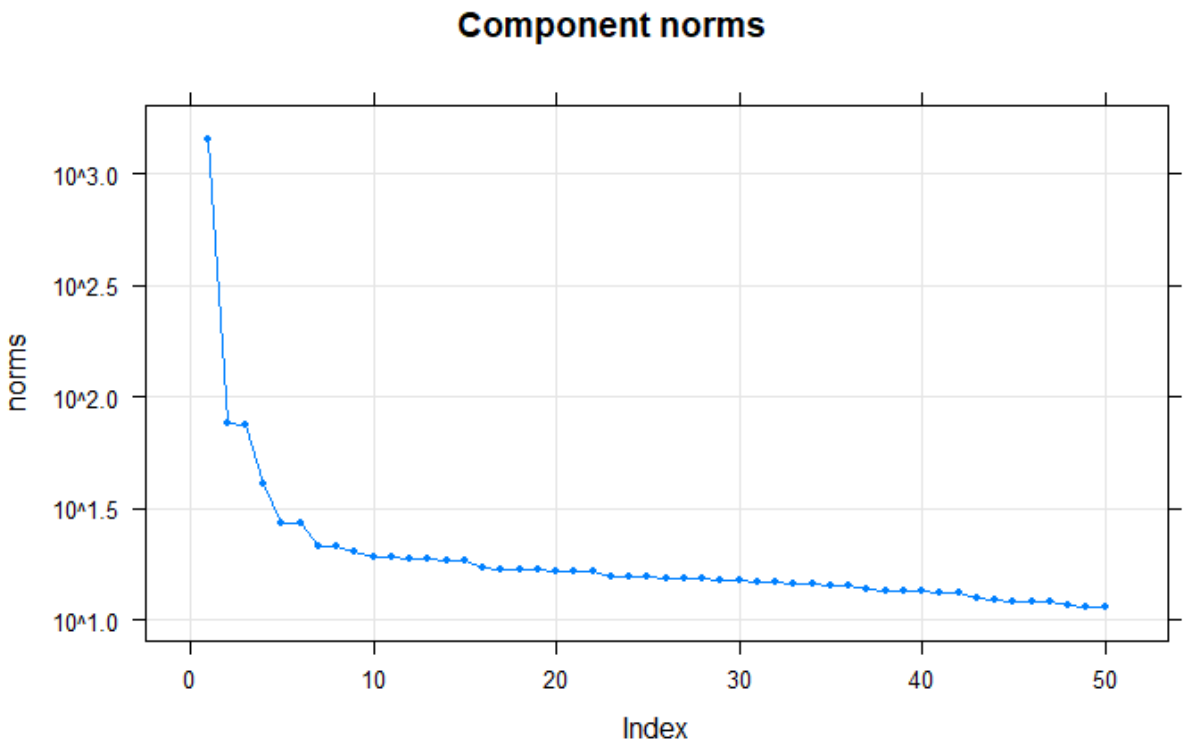

Рис. 2

Следуя теории, нам нужно найти 4 компоненты, так как наш ряд состоит из линейного ряда и гармонического ряда с частотой $\omega < 1/2$. Ранги обоих этих рядов равны двум. Теперь займемся группировкой компонент. Построим график собственных векторов (рис. 3).

Посмотрим на рисунок 3. На нем можно заметить, что компоненты 1 и 4 — медленно-меняющиеся, значит их можно отнести к тренду. Компоненты 2 и 3 больше похожи на *sin* и *cos*, зная что изначальная сезонная была гармоническим рядом, относим их к сезонности. Компоненты 5 и 6 из-за своей непостоянной амплитуды больше похожи на шум. Также их вклад в сумму довольно мал, поэтому их можно отнести к шуму.

Далее применяем функцию *reconstruct* и раскладываем ряд в сумму трех рядов: F1 – тренд, F2 – сезонность и Residuals – остатки.

```
res = reconstruct(s, groups = list(c(1, 4), c(2,3)))
```

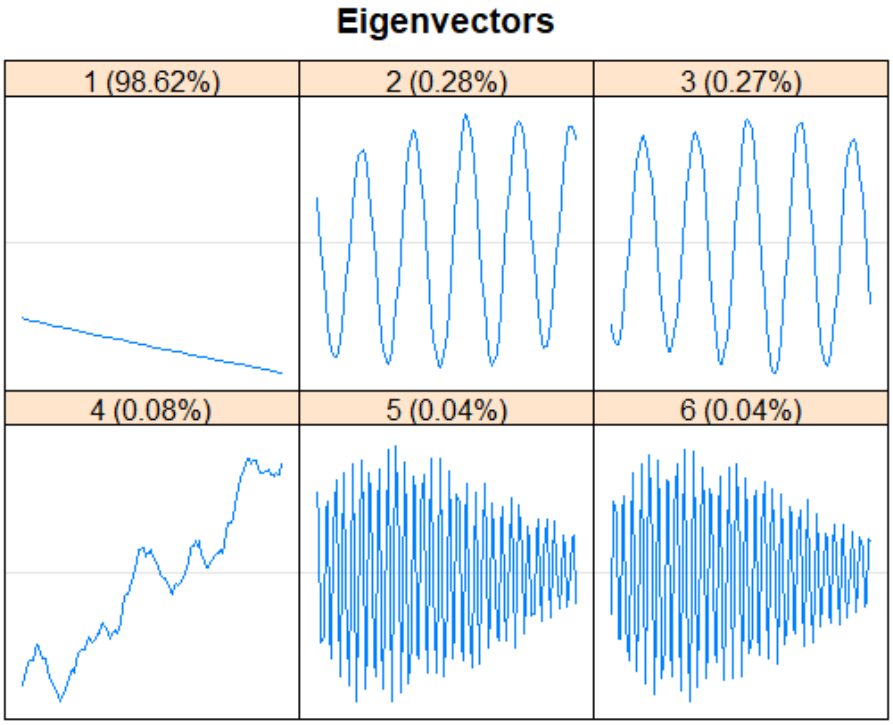

Рис. 3

```
plot(res)
```

Посмотрим на рисунок 4. На графике *Original* мы можем наблюдать наш исходный ряд. На графике *F1* изображена медленно-меняющаяся монотонно возрастающая функцию. Как уже говорилось, график функции в ячейке *F1* описывает тренд исходного ряда. Теперь посмотрим на график *F2*, на нем изображен *sin*, данная функция описывает сезонность исходного ряда. Зная, что в нашем исходном ряду тренд задавался как $f_1(t) = 0.05t + 5$, а сезонность $f_2(t) = \sin(2\pi\frac{t}{T})$ можно сказать, что метод успешно выделил тренд и сезонность из исходного ряда. Наконец, посмотрим на график Residual. Обратим внимание, что почти все значение лежат в интервале от -2 до 2. Зная, что шум пришел из распределения $N(0, 1)$, можно сказать, что 95 процентов всех значений лежат в интервале $[-2\sigma, 2\sigma]$. Следовательно можно предположить, что третье слагаемое в полученной после разложения сумме: остатки – это и есть шум.

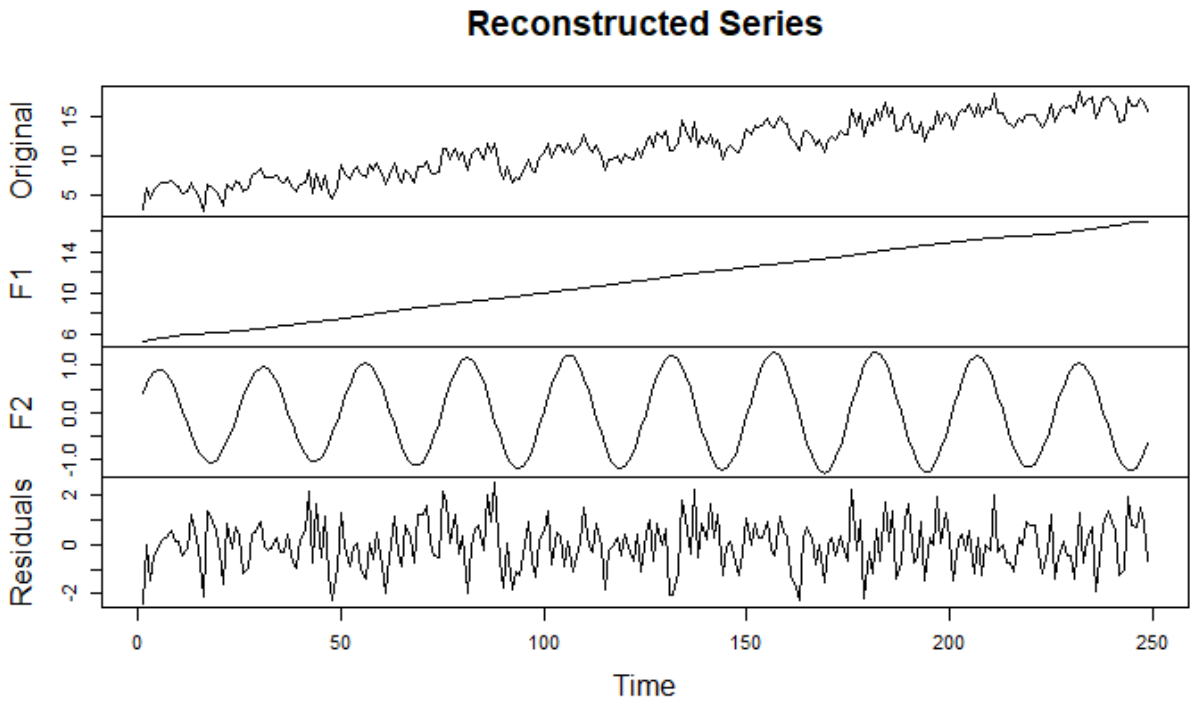
Reconstructed Series
Original
F1
F2
Residuals
Time
0
50
100
150
200
250

Рис. 4

Глава 2

Использование SSA в машинном обучении

Метод SSA — мощный инструмент для обработки данных. С его помощью можно уменьшать шум в данных или выделять полезные признаки из данных. В этой главе обозреваются разные способы использования SSA в методах машинного обучения. В качестве примеров были взяты четыре статьи, в которых разными методами решалась задача прогнозирования. Ниже представлен обзор каждой статьи. Ссылки на статьи доступны в списке литературы.

2.1. Статья №1

Статья «The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series» [2] была написана в 2017 учеными Kongchang Du, Ying Zhao, Jiaqiang Lei из Синьцзянского института географии и экологии. В ней обозревается некорректность использования SSA и метода DWT в качестве предобработки данных для дальнейшего их использования в artificial neural network (ANN) или support vector machine (SVM). Сравнение двух данных подходов проводилось на основе задачи прогнозирования рядов в гидрологической области.

Данные В качестве данных использовался усредненное количество месячных осадков в Индии в промежутке от января 1871 года до декабря 2013 года. Длина временного ряда равна 1716.

Данные разбиваются для следующим образом: выборка *train* содержит в себе первые 750 значений, выборка *validation* содержит значения 751-900, значения 901-1000 используется для «inner» теста, значения 1001-1100 – для «outer» теста.

Тесты «inner» и «outer» В статье используются два типа тестирования. Изначальный ряд имел длину 1100 значений. Авторы задали границу $M = 1000$, первые M значений были исходного ряда предобрабатывались с помощью SSA и используются для построения модели ANN и дальше выполнялся тест на выборке «inner» (то есть

модель обучалась на выборке *train*, валидация проводилась на выборке *validation*, тестировалась на выборке «inner» теста). Авторы статьи отмечают, что при таком подходе и обучающая выборка и метки были предобработанны с помощью SSA.
«outer» тест, был выбран авторами для приближения к реальности модели, чтобы проверить предсказательные возможности модели за границей M. После того как модель ANN уже была построена, она пытается предсказать значение $M+K$, используя предшествующие значения. После значение $M+K$ используется для предсказания следующих значений (то есть начиная со значений $M-11,\ldots,M$ предсказывается значение $M+1$, потом используя значения $M-10,\ldots,M+1$ предсказывается $M+2$ и т.д.). Авторы статьи никак не отмечают в статье изменения стратегии обучения модели для теста «outer», из чего можно сделать предположения, что результаты на тесте «outer» получатся неудовлетворительными (используются значения 1-900 для построения модели, а тестируются на 1000-1100).

Artificial Neural Network В статье модель ANN для прогнозирования данных описывают следующей формулой:

$$x_t = f(X_t, w, \theta, m, h) = \theta_0 + \sum_{j=1}^{h} w_j^{out} \phi(\sum_{i=1}^{m} w_{ji} x_{t-i} + \theta_j),$$

где x – интересующий нас гидрологический временной ряд, m – размер скользящего окна, ϕ – функция активации, w_{ij} – вес между i-й элементом входного вектора и j-м элементом скрытого слоя, θ_j – смещения, связанные с j-м элементом скрытого слоя, w_j^{out} – вес между j-м элементом скрытого слоя и выходным слоем, θ_0 – смещение, связанное с выходным слоем. Все параметры оптимизируются с помощью метода back-propagation. Модель использует значения $[t-m,\ldots,t-1]$ ряда для прогнозирования значения t. Также авторы статьи утверждают, что для достижения достаточной сложности модели достаточно иметь один скрытый слой.

Использование SSA SSA используется для предобработки данных. В статье авторы задают L – длина окна в методе SSA равна 12, а также используют 7 первых собственных троек для реконструкции ряда методом SSA. Далее последние 12 значений реконструированного ряда используются в качестве входного слоя в ANN, в качестве меток используется текущее ground truth значение количества месячных осадков.

Метрики Для сравнения моделей использовались следующие метрики: root mean square errors (RMSE), mean absolute errors (MAE), Nash–Sutcliffe coefficient (NS).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_{i_{observed}} - \hat{Y}_i)^2},$$

$$MAE = \frac{1}{N}|\sum_{i=1}^{N}(Y_{i_{observed}} - \hat{Y}_i)|,$$

$$NS = 1 - \frac{\sum_{i=1}^{N}(\hat{Y}_i - Y_{i_{observed}})^2}{\sum_{i=1}^{N}(Y_{i_{observed}} - \bar{Y}_{i_{observed}})^2}.$$

Результаты Авторы заявляют, что сравнение показало обычных и гибридных моделей показало наличие малой точности гибридных моделей на «outer» тесте и, напротив, большую на «inner» тесте. Обычные модели показали примерно одинаковые результаты в обоих тестах. Авторы обуславливают такой исход тем, что метод SSA для реконструкций ранних значений ряда использует «будущие», что приводит к высокому показателю точности на «inner» тесте.

2.2. Статья №2

Статья «Comparison of ARIMA, SSA, and ARIMA – SSA Hybrid Model Performance in Indonesian Economic Growth Forecasting» [3] была написана в июле 2020 года учеными Muhammad Fajar и Sri Hartini Rachmad из «BPS-Statistics Indonesia». В работе ставится задача сравнение точности трех подходов в задаче прогнозирования экономического роста Индонезии. В статье рассматриваются следующие методы прогнозирования: ARIMA, SSA и ARIMA-SSA.

Данные В статье используются данные экономического роста (квартал к кварталу) начиная с 1983 Q2 до 2018 Q2 взятые из «Badan Pusat Statistik» (BPS). Данные для тестов были разделены следующим образом: 20% наблюдений(28 прогнозов на будущее), 10% наблюдений(14 прогнозов на будущее), 5% наблюдений(7 прогнозов на будущее) и 3% наблюдений(4 прогнозов на будущее). Таким образом получилось 4 теста (то есть в первом тесте используется 80% данных для тренировки модели, а 20% для тестирования. Во втором тесте используется 90% для тренировки, 10% для тестирования модели

и также по аналогии с другими двумя наборами данных). В статье явно не обговаривается, используется ли SSA только на тренировочных данных.

ARIMA Autoregressive-Moving Average. В общем модель $ARIMA(p,d,q)(P,D,Q)^S$ для временного ряда x_t выглядит так:

$$\Phi_P B^S \phi_p (1-B)^d (1-B^S)^D x_t = \theta_q(B)\Theta_Q(B^S)\epsilon_t.$$

С обозначениями можно подробнее ознакомиться в статье в разделе 2.2.1. Параметры модели вычисляются с помощью метода правдоподобия. Лучший вариант модели ARIMA выбирается на основе информационного критерия AIC.

Использование SSA В этой статье метод SSA использовался для прогнозирования данных (подробнее о методе SSA в следующем параграфе). ARIMA-SSA — гибридный метод совмещающий в себе модель ARIMA и метод SSA. Авторы предполагают, что значения временных рядов состоят из линейных и нелинейных компонент, поэтому их можно представить в виде:

$$x_t = P_t + N_t,$$

где P_t – линейная компонента, а N_t – нелинейная компонента. В гибридном методе ARIMA используется для предсказания линейной компоненты, тогда остаток – есть нелинейная компонента. Для предсказаний нелинейных компонент используется SSA.

$$\hat{x}_{T+h} = \hat{P}_{T+h} + \hat{N}_{T+h},$$

где $\hat{x}_{T+h}$ является результатом прогнозирования ряда x на периоде $T+h$, P_{T+h} является результатом прогнозирования P на периоде $T+h$, N_{T+h} является результатом прогнозирования N на периоде $T+h$ и h – предстоящий период.

Прогнозирование SSA В этой статье использовался рекуррентный метод SSA с оценкой коэффициента min-norm LRR (Linear Recurrence Relation). Коэффициент LRR вычислялся следующим образом.

1. Входные данные: $\mathbf{P} = [P_1, \ldots, P_r]$ – матрица, состоящая из собственных векторов U_i с шага SVD. Определим $\bar{\mathbf{P}}$ – матрица $\mathbf{P}$ без последнего ряда.

2. Из каждого вектора-столбца P_i возьмем последнюю компоненты, обозначим ее π_i.

3. Посчитаем: $v^2 = \sum_{i=1}^{r} \pi^2$.

4. Посчитаем коэффициент min-norm LRR:

$$R = \frac{1}{1 - v^2} \sum_{i=1}^{r} \pi_i \bar{\mathbf{P}}_i.$$

5. Из пункта 4 получаем: $R = (\alpha_{r-1}, \ldots, \alpha_1)$.

6. Считаем прогнозируемое значение:

$$\hat{x}_n = \sum_{i=1}^{r-1} \alpha_i \widetilde{x}_{n-1}, \quad n = T+1, \ldots, T+h.$$

Метрики В статье для оценки качества моделей используется метрика RMSE.

Результаты Сравнение показал, что гибридная модель ARIMA-SSA показывает большую точность в задачи прогнозирования роста экономики, чем методы SSA и ARIMA по отдельности.

2.3. Статья №3

Статья «SSA-based hybrid forecasting models and applications» [4] написанна в октябре 2020 года учеными Winita Sulandari, Subanar, Suhartono, Herni Utami, Muhammad Hisyam Lee и Paulo Canas Rodrigues. В статье рассматривается способ комбинирования SSA с другими методами для улучшения точности в задачи прогнозирования временных рядов со сложными паттернами. В работе рассматривается две модификации модели TLSAR(Two-Level Seasonal Autoregressive) – TLSNN (Two-Level Seasonal Neural Network) и TLCSNN (Two-Level Complex Seasonal Neural Network).

Данные В работе использовалось два набора данных. Первый датасета «Monthly accidental deaths in USA» авторы статьи делят данные следующим образом: от января 1973 года до декабря 1978 года – тренировочный датасет, начиная с января 1979 года по июнь 1979 года – тестовый датасет. Второй датасет «Daily electricity load of Jawa-Bali in the specific hours» собранный за период первого января 2009 года по 31 декабря 2011. Однако, данные - это только электрическая нагрузка в 01.00, 02.00, 03.00

и 04.00 до полудня. Таким образом получаются 4 разных временных ряда, в связи с влиянием привычек индонезийских граждан в месяц Рамадана. Каждый из 4ех временных рядов были разбиты на выборки следующим образом: первые 1088 наблюдений – тренировочный датасет, последние 41 наблюдение – тестовый датасет. В статье явно не обговаривается, используется ли SSA только на тренировочных данных.

TLSNN и TLCSNN Две предложенные модели называемые two-level seasonal neural networks (TLSNN) и the two-level complex seasonal neural networks (TLCSNN). Обе модели являются модификациями TLSAR. В детерминированную компоненту были включены полиномиальный тренд и изменяющаяся во времени синусоидальная функция, чтобы захватить более сложную картину во временном ряду. В этом случае результаты декомпозиции SSA облегчают идентификацию и определение правильной детерминированной функции для каждого компонента TLSNN и TLCSNN. Как правило, TLSNN и TLCSNN выражаются в следующей формуле:

$$Y_t = S_t + Z_t,$$

где Y_t наблюдение в момент времени t, S_t – детерминированная компонента, а Z_t – стохастическая компонента. Разница в двух моделях заключается в разной форму S_t. Z_t аппроксимируется с помощью нейронной сети (NN). Подробнее о этих моделях можно прочесть в статье в параграфе 2.1.

Использование SSA В статье SSA используется в качестве feature-extraction. Метод SSA используется на изначальном временом ряде, раскладывая его на тренд, несколько колебательных составляющих и шум. Используется алгоритм описанный в главе 1.1 «Базовый алгоритм SSA». Далее, полученные ряды идут на вход в детерминированная модели TLSAR и TLSNN. Основывая на w-матрицах корреляции временной ряд для данных «Monthly accidental deaths in USA» был разбит на 4 компоненты (см. в статье Figure 1(b)). Окно для метода SSA – L было выбрано равным 24, пропорционально периоду. В ввиду схожести результатов для четырех рядов в данных «Daily electricity load of Jawa-Bali in the specific hours» авторы приводят выбранные параметры для метода SSA только для четвертого ряда. Авторы выбрали длину окна L равной 490, ряд был разбит на 3 компоненты (см. в статье Figure 3)

Метрики В статье используются метрики RMSE и MAPE.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_{i_{observed}} - \hat{Y}_i}{Y_{i_{observed}}}\right|.$$

Результаты В итоге гибридные модели показали результаты лучше, чем модель TLSAR. В статье отмечается значительный прирост в точности. Авторы связывают это с тем, что детерминированные компоненты моделей TLSNN и TLCSNN, построенных на основе результатов декомпозиции SSA, учитывали не только линейный тренд, но и квадратичный или другой полином более высокого порядка для захвата других трендовых поведений. Также стохастическая компонента TLSNN и TLCSNN была моделирована с помощью NN, что помогло преодолеть проблему нелинейной взаимосвязи в данных.

2.4. Статья №4

Статья «Linking Singular Spectrum Analysis and Machine Learning for Monthly Rainfall Forecasting» [5] была написана в мае 2020 года учеными Pa Ousman Bojang, Tao-Chang Yang, Quoc Bao Pham и Pao-Shan Yu из национального университета Ченг-Кунг. В статье рассматривается способ предобработки данных с помощью метода SSA для улучшения точности в задачи прогнозирования в гидрологической области. В статье сравниваются две гибридные модели SSA-LSSVR и SSA-RF.

Данные Данные были собраны из Тайваньского бюро водных ресурсов, которое включает ежемесячные данные об осадках за 1958–2018 годы для каждого измерительного прибора осадков в водоразделе Шихмена и за 1981–2017 годы для каждого измерительного прибора осадков в водоразделе водохранилища Деджи. Для каждого водораздела водохранилища первые 70% всего сбора данных использовались в качестве выборки *train*, а оставшиеся 30% - для валидации. Описательную статистику данных можно посмотреть в статье в таблице 3 (Table 3). Авторы отмечают, что подают на вход SSA подпоследовательность ряда из тренировочной выборки.

LS-SVM и Random Forest Прогнозирования после предобработки данных с помощью SSA проводилось с помощью хорошо известного Random Forest (RF) и least-squares support vector machine (LS-SVM). LS-SVM – новый тип модели SVM. Вместо решения

задачи выпуклого квадратичного программирования решения LS-SVM достигаются путем решения серии линейных уравнений. Это изменение снижает вычислительную сложность и делает LS-SVM более привлекательным. Более подробно про LS-SVM написано в статье в главе 3.1. Стандартные модели LS-SVM и RF обучались на тренировочном датасете, валидационная выборка использовалась для кросс-валидации.

Использование SSA SSA использовался в качестве feature-extraction. Выбирается подпоследовательность из изначального ряда в качестве учителя, далее методом SSA выделяется из подпоследовательности тренд и периодические компоненты. Длина окна L определяется заранее. Таким образом авторы статьи выделяют новые компоненты из изначального ряда. Далее, модели LS-SVR и RF обучались на всех компонентах. В статье использовался базовый алгоритм SSA описанный в главе 1.1. Схему работы гибридных моделей SSA-LSSVR и SSA-RF можно посмотреть в статье на изображении 2 (Figure 2, с. 10).

Метрики Для оценки качества работы моделей авторами статьи использовались метрики RMSE и NS.

Результаты Одним из основных выводов является то, что гибридные модели (SSA-LSSVR и SSA-RF) имеют лучшие показатели точности, чем стандартные модели (LS-SVR и RF) для обоих наборов данных. Можно сделать вывод, что гибридные модели представляют собой перспективный подход к моделированию, который может применяться для прогнозирования месячных осадков в исследуемом регионе. Однако две гибридные модели работают по-разному в двух водоразделах (SSA-LSSVR работает лучше, чем SSA-RF в одном водоразделе, но хуже - в другом).

Заключение

В работе был рассмотрен базовый алгоритм SSA. Приведен пример разложения простого ряда на тренд и сезонность методом SSA с помощью библиотеки Rssa на языке R. Также в работе были рассмотрены 4 статьи, в которых различным образом применялся метод SSA вместе с алгоритмами machine learning. Статьи были детально изучены и будут использоваться в дальнейшем для повторения экспериментов, описанных в статьях, или в качестве основы для проведения собственных исследований с использованием SSA и алгоритмами machine learning.

Список литературы

1. Golyandina, N., Nekrutkin, V., & Zhigljavsky, A. (2001). Analysis of time series structure: SSA and related techniques. Chapman & Hall/CRC.
2. Kongchang Du, Ying Zhao, Jiaqiang Lei (2017). The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series.
3. Muhammad Fajar, Sri Hartini Rachmad (2020). Comparison of ARIMA, SSA, and ARIMA – SSA Hybrid Model Performance in Indonesia Economic Growth Forecasting.
4. Winita Sulandari, Subanar, Suhartono, Herni Utami, Muhammad Hisyam Lee, Paulo Canas Rodrigues (2020). SSA-based hybrid forecasting models and applications.
5. Pa Ousman Bojang, Tao-Chang Yang, Quoc Bao Pham, Pao-Shan Yu (2020). Linking Singular Spectrum Analysis and Machine Learning for Monthly Rainfall Forecasting.