

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра Статистического Моделирования

«Научно-исследовательская работа» (семестр 6)

ПОДДЕРЖИВАЮЩИЕ ВРЕМЕННЫЕ РЯДЫ В MSSA

Выполнил:

Ткаченко Егор Андреевич

группа 19.Б04-мм

Научный руководитель:

к. ф.-м. н., доцент

Голяндина Нина Эдуардовна

Оглавление

Введение	3
1. Определения	3
2. Применение SSA и MSSA	4
3. ЛРФ	4
Глава 1. Постановка задачи	6
1.1. Примеры	7
Глава 2. Численные эксперименты	8
2.1. Выбор компонент для MSSA	8
2.2. старое	11
2.3. старые Результаты	11
Заключение	12
Список литературы	13

Введение

Полезность умения строить прогнозы не нуждается в доказательстве. Прогноз временных рядов может использоваться в прогнозе погоды, приливов, спроса на товары и многом другом.

С помощью книги [1] был изучен базовый SSA, разложение рядов, заполнение пробелов в данных, прогноз и базовый MSSA. Для работы с временными рядами и их прогнозом использовался пакет Rssa. Проведены эксперименты с простейшими моделями сигналов для изучения связи между согласованностью сигналов и поддерживающими рядами.

1. Определения

Вещественным временным рядом длины N называется вектор

$$\mathbf{F} = (f_0, \dots, f_{N-1}), \quad f_j \in \mathbb{R}.$$

Многомерным временным рядом $\vec{\mathbf{F}}$ называется набор s временных рядов $\mathbf{F}^{(p)}$ длин N_p :

$$\vec{\mathbf{F}} = \{\mathbf{F}^{(p)} = (f_0^{(p)}, \dots, f_{N_p-1}^{(p)}), \quad p = 1, \dots, s\}.$$

L -траекторная матрица (или просто траекторная матрица) ряда \mathbf{F} имеет структуру ганкелевой матрицы, а ее столбцами являются отрезки длины L ряда \mathbf{F}

$$\mathcal{T}_{SSA}(\mathbf{F}) = \begin{pmatrix} f_0 & f_1 & \dots & f_{K-1} \\ f_1 & f_2 & \dots & f_K \\ \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & \dots & f_{N-1} \end{pmatrix}.$$

L -траекторная матрица многомерного ряда $\vec{\mathbf{F}}$ состоит из горизонтально склеенных траекторных матриц рядов $\mathbf{F}^{(p)} \in \vec{\mathbf{F}}$:

$$\mathcal{T}_{MSSA}(\vec{\mathbf{F}}) = [\mathcal{T}_{SSA}(\mathbf{F}^{(1)}) : \dots : \mathcal{T}_{SSA}(\mathbf{F}^{(s)})].$$

Из траекторной матрицы можно восстановить ряд. Из любой матрицы подходящего размера можно получить траекторную матрицу проектированием на пространство ганкелевых матриц (или склеенных горизонтально ганкелевых матриц для многомерного случая).

Ранг ряда равен рангу его траекторной матрицы:

$$\text{rank } \mathbf{F} = \text{rank } \mathcal{T}_{SSA}(\mathbf{F}), \quad \text{rank } \vec{\mathbf{F}} = \text{rank } \mathcal{T}_{MSSA}(\vec{\mathbf{F}}).$$

2. Применение SSA и MSSA

Алгоритмы SSA и MSSA могут быть применены для аппроксимации временного ряда рядом конечного ранга.

Алгоритм.

Вход: Ряд F для SSA или многомерный ряд \vec{F} для MSSA, длина окна $L \leq N$ для SSA или $L \leq N_p, \forall N_p$ для MSSA, ранг аппроксимирующего ряда r .

1 Вложение. Временной ряд переводится в L -траекторную матрицу X

$$X = \mathcal{T}_{SSA}(F) \text{ для SSA, } \quad X = \mathcal{T}_{MSSA}(\vec{F}) \text{ для MSSA.}$$

2 Сингулярное разложение. Методом SVD матрица X раскладывается на сумму d матриц $X_i = \sqrt{\lambda_i} U_i V_i^T$, где $d = \text{rank } X = \text{rank } XX^T \leq L$, λ_i — собственные числа матрицы XX^T ($\lambda_1 \geq \dots \geq \lambda_L \geq 0$), U_i — собственные вектора матрицы XX^T , $V_i = X^T U_i / \sqrt{\lambda_i}$ — факторные вектора матрицы X .

3 Группировка. Множество индексов $\{1, \dots, d\}$ делится на m непересекающихся множеств I_1, \dots, I_m . Далее для каждого множества индексов (пусть $I = i_1, \dots, i_t$) получается матрица $X_I = X_{i_1} + \dots + X_{i_t}$.

Для аппроксимации рядом конечного ранга r , понадобится множество из первых r индексов $\{1, \dots, r\}$, соответствующую ему матрицу обозначу $\hat{X}_r = X_1 + \dots + X_r$.

4 Сгруппированные матрицы X_{I_j} восстанавливаются в ряды (SSA) или многомерные ряды (MSSA). Для получения аппроксимирующего ряда нужно восстановить его из матрицы \hat{X}_r .

Выход: Аппроксимирующий ряд \hat{F}_r конечного ранга r .

3. ЛРФ

Линейная рекуррентная формула (ЛРФ) выражает каждый член последовательности через линейную комбинацию предыдущих членов.

Ряд F длины N — управляемый ЛРФ, если существуют такие a_1, \dots, a_d , что:

$$f_{i+d} = \sum_{k=1}^d a_k f_{i+d-k}, \quad 0 \leq i \leq N-1-d, \quad a_d \neq 0, \quad t < N-1.$$

Важно отметить, что ряд конечного ранга является управляемым ЛРФ [1, 2.1.2.2, стр. 35].

Вещественный временной ряд F , управляемый ЛРФ, естественным образом прогнозируется на одну точку:

$$\tilde{f}_N = \sum_{k=1}^{L-1} a_k f_{N-k}.$$

Но тогда его можно прогнозировать и на любое количество точек.

Глава 1

Постановка задачи

Пусть имеется временной ряд $F^{(1)} = S^{(1)} + R^{(1)}$, где сигнал $S^{(1)}$ — ряд управляемый ЛРФ, шум $R^{(1)}$ — ряд без структуры. Рассмотрим задачу прогнозирования $S^{(1)}$. Эта задача уже решается методом SSA, но как можно улучшить прогноз?

Пусть помимо ряда $F^{(1)}$ имеется временной ряд $F^{(2)} = S^{(2)} + R^{(2)}$. Если структура сигналов $S^{(1)}$ и $S^{(2)}$ похожа, то использование ряда $F^{(2)}$ может улучшить прогноз сигнала $S^{(1)}$, потому что второй ряд дает алгоритму больше данных, которые могут улучшить ЛРФ. Возможность такого улучшения прогноза подтверждена [1, 4.3.3.3, стр. 216]. Но второй ряд может сделать прогноз хуже, если структура его сигнала отличается от первого.

Простейший пример похожих по структуре сигналов — гармонические колебания с одинаковыми периодом. Они даже могут быть смещены по фазе или иметь разную амплитуду.

Для объективной оценки качества прогноза буду использовать среднюю квадратичную ошибку.

$$\text{MSE}(\tilde{S}, S) = \frac{1}{N_f} \sum_{i=N}^{N+N_f-1} (\tilde{s}_i - s_i)^2,$$

где \tilde{S} — прогноз сигнала S на N_f точек, N — длина прогнозируемого сигнала S .

Ряд $F^{(2)}$ называется поддерживающим для прогноза, если прогноз с его использованием лучше чем без него, т.е.

$$\text{MSE}(\tilde{S}_{SSA}, S^{(1)}) < \text{MSE}(\tilde{S}_{SSA}, S^{(1)}).$$

Такое определение можно применить только в экспериментах с известным продолжением ряда. На практике не с чем сравнить прогноз, поэтому появляется вопрос. Как понять, что ряд поддерживающий не зная продолжения прогнозируемого ряда? Помочь ответить на этот вопрос может понятие согласованности.

Сигналы $S^{(1)}$, $S^{(2)}$ называются полностью согласованными, если ранг $r_{1,2} = r_1 = r_2$ и полностью несогласованными, если $r_{1,2} = r_1 + r_2$, где $r_1 = \text{rank } S^{(1)}$, $r_2 = \text{rank } S^{(2)}$, $r_{1,2} = \text{rank } \vec{S} = \text{rank } \{S^{(1)}, S^{(2)}\}$.

1.1. Примеры

когда следует ожидать, что MSSA лучше

Глава 2

Численные эксперименты

2.1. Выбор компонент для MSSA

Гипотеза: Когда сигналы похожи, их можно считать согласованными и лучше использовать (при прогнозе или восстановлении сигнала) ранг равный рангу одного сигнала. Когда сигналы отличаются, их следует считать не согласованными и использовать ранг равный сумме рангов сигналов.

Выберем первый ряд как простой сигнал, зависящий от параметра (например, период) с аддитивным гауссовым шумом. Второй ряд будет простым сигналом того же вида, с несколько отличающимся параметром и без шума. Восстановим и спрогнозируем первый ряд с помощью *SSA*, *MSSA* считая ряды согласованными и *MSSA* считая ряды несогласованными.

Для устойчивости результатов, повторим это 50 раз и возьмем среднее ошибок.

Назовем относительной ошибкой прогноза (восстановления) значение

$$\frac{error_{SSA} - error_{MSSA}}{error_{SSA} + error_{MSSA}},$$

где $error_{SSA}, error_{MSSA}$ — ошибки прогноза (восстановления) методами *SSA* и *MSSA* соответственно.

Значения относительной ошибки легко расположить на графике (она принимает значения от -1 до 1). По значению относительной ошибки легче понять, какой метод лучше (не надо сравнивать два значения ошибок, которые просто положительны и могут быть любых порядков).

Как интерпретировать значения относительной ошибки?

- Значения > 0 значат, что что *MSSA* лучше *SSA*;
- Значения < 0 значат, что что *MSSA* хуже *SSA*;
- значения около 0 значат что ошибки примерно равны.
- значения далеко от 0 значат, что ошибки сильно отличаются.

Так же все сигналы будут нормироваться, чтобы амплитуда сигнала не влияла на ошибки прогноза и восстановления. Например, для косинуса: $s_j^{(i)} = A \cos(\frac{2\pi j}{T_i})$, где $A = \text{mean}(|s_j^{(i)}|)^{-1}$.

2.1.1. Сигнал косинус

Сигнал $S^{(1)}$ — косинус с периодом 8. Сигналы $S^{(2)}$ — косинусы с периодами 8, 8.02, 8.04, 8.06, 8.08, 8.1, 8.15.

Ранг косинуса равен 2, поэтому для $MSSA$ используются первые 2 или первые 4 компоненты разложения, а для SSA только 2.

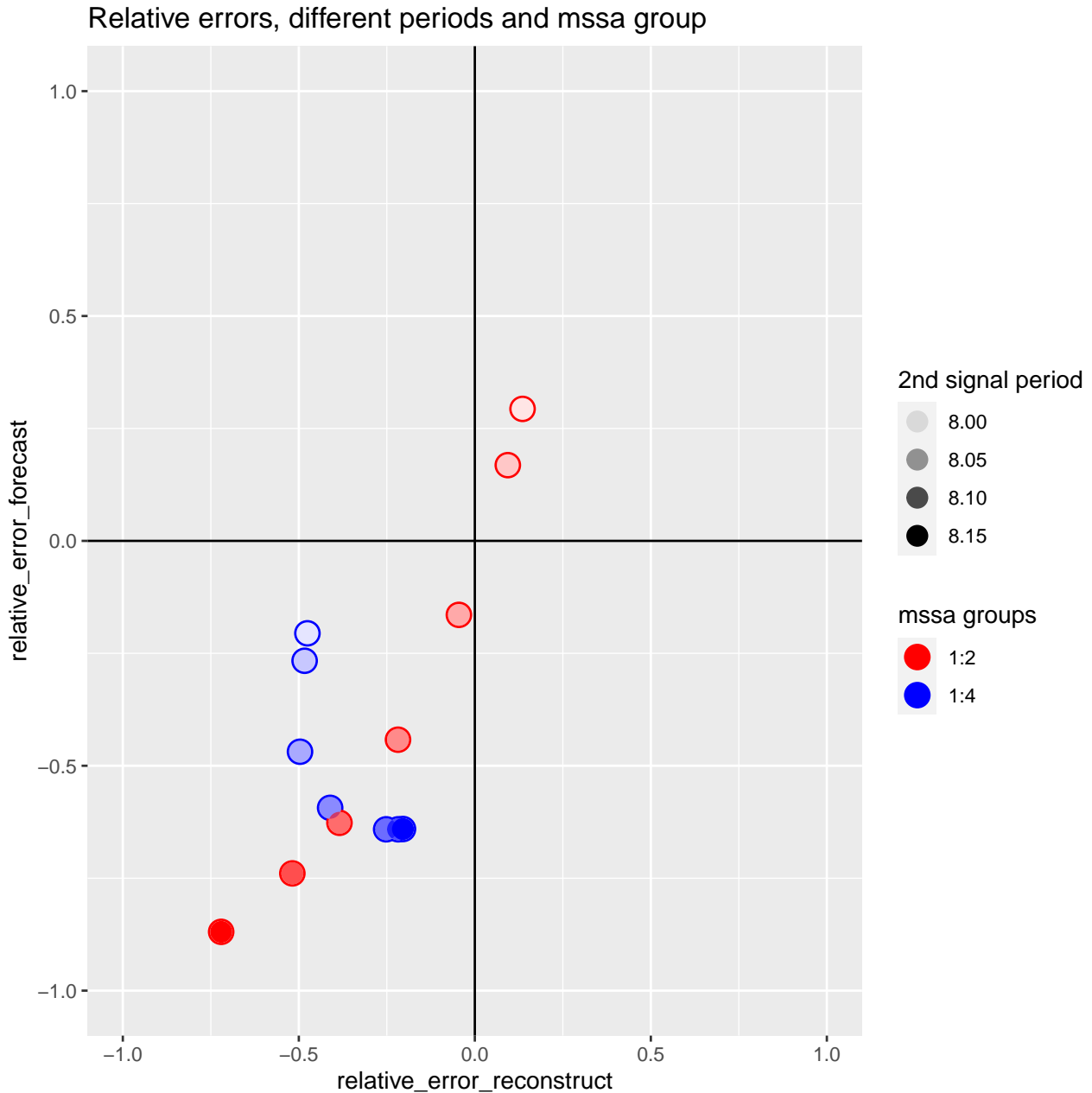


Рис. 2.1. Зависимость относительных ошибок от разницы сигналов и выбора ранга для $MSSA$

На рис. 2.1 видим подтверждение гипотезы: с увеличением разницы рядов использование четырех компонент становится лучше.

2.1.2. Сигнал экспонента

Функция для сигналов — $s_j^{(i)} = A \exp(j\lambda_i)$. Сигнал $S^{(1)}$ — экспоненциальная функция с $\lambda =$. Сигналы $S^{(2)}$ — косинусы с периодами 8, 8.02, 8.04, 8.06, 8.08, 8.1, 8.15.

Ранг косинуса равен 2, поэтому для *MSSA* используются первые 2 или первые 4 компоненты разложения, а для *SSA* только 2.

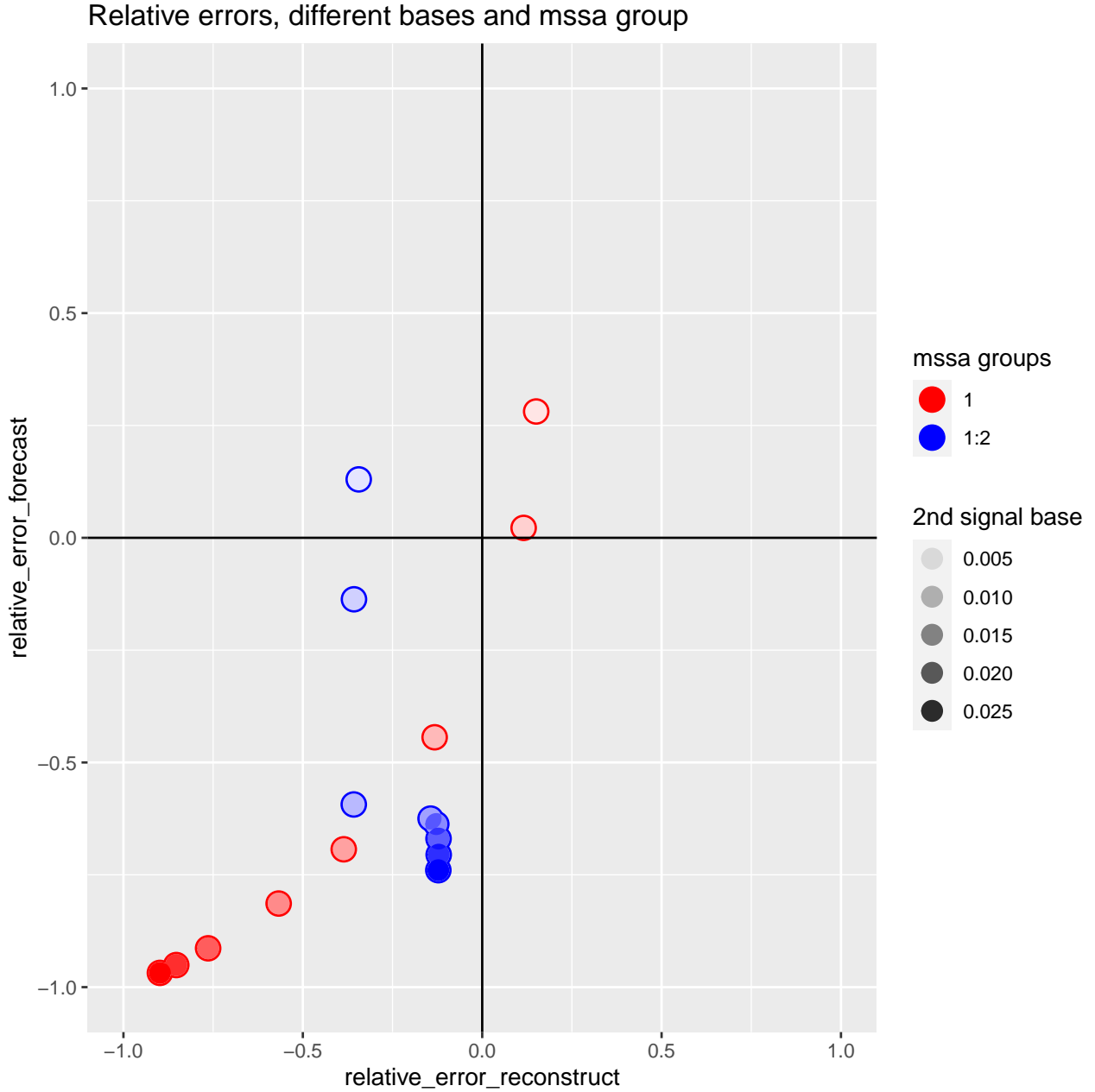


Рис. 2.2. Зависимость относительных ошибок от разницы сигналов и выбора ранга для *MSSA*

На рис. 2.2 видим подтверждение гипотезы: с увеличением разницы рядов использование четырех компонент становится лучше.

2.2. старое

Выдвинута гипотеза, что ряд $F^{(2)}$ поддерживающий для ряда $F^{(1)}$, если их сигналы $S^{(1)}$, $S^{(2)}$ согласованны и шум $R^{(2)}$ небольшой.

Так как на практике могут встречаться отклонения от полностью согласованных сигналов, хочется узнать, на сколько можно исказить сигнал второго ряда, прежде чем он перестанет быть поддерживающим?

Для подтверждения гипотезы и ответа на вопрос была проведена серия описанных ниже экспериментов.

Временные ряды строились как сумма сигнала и шума. В качестве моделей сигнала использовались функции из списка:

1. $s_j^{(i)} = \exp(j\lambda_i)$;
2. $s_j^{(i)} = \cos(\frac{2\pi j}{12}) \exp(j\lambda_i)$;
3. $s_j^{(i)} = \cos(\frac{2\pi j}{T_i})$;
4. $s_j^{(i)} = \cos(\frac{2\pi j}{12}) + \exp(j\lambda_i)$.

Параметр первого сигнала (период косинуса T_1 или база показательной функции λ_1) фиксировался, параметр второго сигнала менялся, начиная с такого же значения как у первого, до момента, когда второй ряд переставал быть поддерживающим. В качестве шума для обоих рядов использовались независимые белые гауссовские шумы со средними, равными 0, и дисперсиями σ_1^2, σ_2^2 , соответственно.

2.3. старые Результаты

Гипотеза была подтверждена для всех четырех моделей сигнала. Ряд, сигнал которого полностью согласован с сигналом прогнозируемого, перестает быть поддерживающим, если его шум слишком большой.

Показательные функции, умноженные на косинус с общим периодом, оставались поддерживающими дольше, чем показательные функции без косинусов. Показательные функции, к которым был прибавлен косинус, тоже были поддерживающими дольше, чем показательные функции без косинусов.

Заключение

В будущем планируется дальнейшее изучение зависимости согласованности от структуры рядов, рассмотрение случаев, когда ряды имеют разные ранги.

Использование SSA и MSSA сложно автоматизировать, потому что алгоритму требуются дополнительные параметры, лучший выбор которых не очевиден. Возможно, будет разработан метод поиска оптимальных входных параметров.

Список литературы

1. Golyandina N, Korobeynikov A, Zhigljavsky A. Singular Spectrum Analysis with R. — Springer, 2018. — P. 272.