Некоторые задачи, связанные с методом анализа сингулярного спектра и его многомерным обобщением

Грязнов Святослав Игоревич
Научный руководитель: к. ф.-м. н., доцент Н. Э. Голяндина
Рецензент: к. ф.-м. н., доцент А. И. Коробейников

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

2016

План

- Введение
- Задача 1: Factor-MSSA
- Задача 2: $\{Q, R\}$-разложения

Введение

Введение: Постановка задачи

$\mathbb{X}$ – многомерный временной ряд (система из s временных рядов).

$$\mathbb{X} = (\mathbb{X}^{(1)}, \ldots, \mathbb{X}^{(s)}), \mathbb{X}^{(k)} = (x_j^{(k)})_{j=1}^{N},$$

$$\mathbb{X} = \mathbb{F} + \mathbb{G} + \mathbb{N},$$

$$\mathbb{F} = (\mathbb{F}^{(1)}, \ldots, \mathbb{F}^{(s)}), \mathbb{F}^{(k)} = (f_j^{(k)})_{j=1}^{N},$$

$$\mathbb{G} = (\mathbb{G}^{(1)}, \ldots, \mathbb{G}^{(s)}), \mathbb{G}^{(k)} = (g_j^{(k)})_{j=1}^{N},$$

$$\mathbb{N} = (\mathbb{N}^{(1)}, \ldots, \mathbb{N}^{(s)}), \mathbb{N}^{(k)} = (n_j^{(k)})_{j=1}^{N},$$

где $\mathbb{F}$ и $\mathbb{G}$ – компоненты сигнала, $\mathbb{N}$ – шум, s – размерность рядов.

Задача: Выделить компоненты $\mathbb{F}$ и $\mathbb{G}$.

Метод: Многомерный анализ сингулярного спектра (Multivariate Singular Spectrum Analysis, MSSA, [Broomhead D.S., King G.P., 1986]).

Введение: Постановка задачи

$\mathbb{X}$ – многомерный временной ряд (система из s временных рядов).

$$\mathbb{X} = (\mathbb{X}^{(1)}, \ldots, \mathbb{X}^{(s)}), \mathbb{X}^{(k)} = (x_j^{(k)})_{j=1}^{N},$$

$$\mathbb{X} = \mathbb{F} + \mathbb{G} + \mathbb{N},$$

$$\mathbb{F} = (\mathbb{F}^{(1)}, \ldots, \mathbb{F}^{(s)}), \mathbb{F}^{(k)} = (f_j^{(k)})_{j=1}^{N},$$

$$\mathbb{G} = (\mathbb{G}^{(1)}, \ldots, \mathbb{G}^{(s)}), \mathbb{G}^{(k)} = (g_j^{(k)})_{j=1}^{N},$$

$$\mathbb{N} = (\mathbb{N}^{(1)}, \ldots, \mathbb{N}^{(s)}), \mathbb{N}^{(k)} = (n_j^{(k)})_{j=1}^{N},$$

где $\mathbb{F}$ и $\mathbb{G}$ – компоненты сигнала, $\mathbb{N}$ – шум, s – размерность рядов.

Модель: $\mathbb{F}^{(k)}$ и $\mathbb{G}^{(k)}$ удовлетворяют линейным рекуррентным соотношениям (ЛРС).

$\mathbb{S} = (s_1, \ldots, s_N)$ удовлетворяет линейному рекуррентному соотношению порядка r, если $s_n = a_1 s_{n-1} + \ldots + a_r s_{n-r}$, $n = r+1, \ldots, N$; $a_r \neq 0$.

Задача: Factor-MSSA

Factor-MSSA: SSA-вложение

$\mathbb{X} = (x_j)_{j=1}^N$ – временной ряд.
Выберем L (длину окна), $1 \leqslant L \leqslant N$ и положим $K = N - L + 1$.
$\mathcal{M}_{L,K}$ – пространство матриц размера $L \times K$.
$\mathcal{M}_{L,K}^{(H)}$ – пространство ганкелевых матриц размера $L \times K$.
$X_j = (x_j, \ldots, x_{j+L-1})^{\mathrm{T}}$, $1 \leqslant j \leqslant K$.
$\mathbf{X} = [X_1 : \ldots : X_K]$ – **траекторная матрица** ряда $\mathbb{X}$.

Оператор вложения

Оператор $\mathcal{T} : \mathbb{R}^N \mapsto \mathcal{M}_{L,K}^{(H)}$, действующий как $\mathcal{T}(\mathbb{X}) = \mathbf{X}$.

Оператор ганкелизации

$\mathcal{H} : \mathcal{M}_{L,K} \to \mathcal{M}_{L,K}^{(H)}$ – проектор на $\mathcal{M}_{L,K}^{(H)}$ по норме Фробениуса посредством усреднения элементов на диагоналях $i + j = \mathrm{const}$.

Factor-MSSA: Метод Factor-MSSA

Алгоритм метода Factor-MSSA [Groth A., Ghil M., 2011]. Первые два шага совпадают с MSSA. Для простоты $s = 2$.

1. Вложение: Траекторная матрица $\mathbf{X} = [\mathbf{X}^{(1)} : \mathbf{X}^{(2)}]$, $\mathbf{X}^{(k)} = \mathcal{T}\mathbb{X}^{(k)}$.
2. Сингулярное разложение: $\mathbf{X} = \sum\limits_{k=1}^{d} \sqrt{\lambda_k} U_k V_k^{\mathrm{T}}$, $\{U_k\}$, $\{V_k\}$ ортонормированы, $\lambda_1 \geqslant \lambda_2 \geqslant \ldots$
3. Поворот некоторых факторных векторов:
   - Выберем группу индексов $J = (j_1, \ldots, j_\ell) \subset \{1, \ldots, d\}$ и рассмотрим $\mathbf{X}_J = \sum\limits_{k \in J} \sqrt{\lambda_k} U_k V_k^{\mathrm{T}}$, $\mathbf{V}_J = [V_{j_1} : \cdots : V_{j_\ell}]$.
   - Посредством метода MVarimax найдем матрицу поворота $\mathbf{R}$ матрицы $\mathbf{V}_J$.
   - Повернем матрицу $\mathbf{V}_J$ с помощью $\mathbf{R}$ и найдем новые векторы $\widetilde{\mathbf{U}}_J$:

$$\widetilde{\mathbf{V}}_J = [\widetilde{V}_{J_1} : \cdots : \widetilde{V}_{J_\ell}] = \mathbf{V}_J \mathbf{R},$$

$$\widetilde{\mathbf{U}}_J = [\widetilde{U}_{J_1} : \cdots : \widetilde{U}_{J_\ell}] = \mathbf{X}_J \widetilde{\mathbf{V}}_J \mathbf{\Lambda}_J^{-1}, \text{ где } \mathbf{\Lambda}_J = \mathrm{diag}(\lambda_{j_1}, \ldots, \lambda_{j_\ell}).$$

Новое разложение: $\mathbf{X} = \sum\limits_{k \in J} \sqrt{\lambda_k} \widetilde{U}_k \widetilde{V}_k + \sum\limits_{k \in \{1,\ldots,d\} \setminus J} \sqrt{\lambda_k} U_k V_k.$

Factor-MSSA: Метод Factor-MSSA

4. Группировка: Как и в MSSA, выберем m дизъюнктных групп, просуммируем внутри каждой элементарные матрицы.

$$\mathbf{X} = \mathbf{X}_{I_1} + \ldots + \mathbf{X}_{I_m}.$$

5. Диагональное усреднение: В результате этого шага получим m восстановленных временных рядов.

$$\widetilde{\mathbf{X}}_I^{(k)} = \mathcal{H}\mathbf{X}_I^{(k)},\ \widetilde{\mathbb{X}}_I^{(k)} = \mathcal{T}^{-1}\widetilde{\mathbf{X}}_I^{(k)}, k = 1, 2,$$
$$\widetilde{\mathbb{X}}_I = [\widetilde{\mathbb{X}}_I^{(1)} : \widetilde{\mathbb{X}}_I^{(2)}].$$

Результат алгоритма: разложение $\mathbb{X}$ на интерпретируемые аддитивные компоненты.

$$\mathbb{X} = \mathbb{X}_{I_1} + \ldots + \mathbb{X}_{I_m}.$$

Factor-MSSA: Метод MVarimax

Метод MVarimax: Поиска поворота $\mathbf{R}$ на третьем шаге алгоритма. Является обобщением обычного метода многомерной статистики Varimax в факторном анализе, [Groth A., Ghil M., 2011].

s-мерный временной ряд длины N, L – длина окна, $K = N - L + 1$, ℓ – число выбранных для поворота компонент.

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 \\ \cdot\cdot \\ \vdots \\ \cdot\cdot \\ \mathbf{V}_s \end{bmatrix}, \text{ где } \mathbf{V}_j = \begin{pmatrix} v_{j,1}(1) & v_{j,2}(1) & \cdots & v_{j,\ell}(1) \\ \vdots & \vdots & \ddots & \vdots \\ v_{j,1}(K) & v_{j,2}(K) & \cdots & v_{j,\ell}(K) \end{pmatrix}.$$

$$\mathrm{Var_K}(\mathbf{V}, \mathbf{R}) = \sum_{k=1}^{\ell} \left(\frac{1}{s} \sum_{d=1}^{s} \left(\tilde{\tilde{v}}^2_{d,k}\right)^2 - \left(\frac{1}{s} \sum_{d=1}^{s} \tilde{\tilde{v}}^2_{d,k}\right)^2 \right) \to \max_{\mathbf{R} \in \mathrm{SO}(\ell)},$$

где $\tilde{\tilde{v}}^2_{d,k} = \sum\limits_{m=1}^{K} \tilde{v}^2_{d,k}(m)$, а $\tilde{v}_{d,k}(m)$ – элементы матрицы $\widetilde{\mathbf{V}}(\mathbf{R}) = \mathbf{VR}$.

Factor-MSSA: Разделимость

Многомерные ряды $\mathbb{F}$ и $\mathbb{G}$, $\mathbb{X} = \mathbb{F} + \mathbb{G}$.
$\mathbf{X} = \mathbf{F} + \mathbf{G}$, где $\mathbf{X}$, $\mathbf{F}$ и $\mathbf{G}$ – траекторные матрицы рядов $\mathbb{X}$, $\mathbb{F}$ и $\mathbb{G}$.

Слабая разделимость для MSSA:

Ряды $\mathbb{F}$ и $\mathbb{G}$ **слабо MSSA-разделимы**, если существует такое сингулярное разложение $\mathbf{X}$ на элементарные матрицы, что их можно разбить на две части: первая из которых в сумме составляет траекторную матрицу $\mathbf{F}$, а вторая – $\mathbf{G}$.

Слабая разделимость для Factor-MSSA:

Ряды $\mathbb{F}$ и $\mathbb{G}$ **слабо Factor-MSSA-разделимы**, если существует такое сингулярное разложение, такая группа J и такой поворот $\mathbf{R}$, что результирующее разложение $\mathbf{X}$ на элементарные матрицы можно разбить на две части: первая из которых в сумме составляет траекторную матрицу $\mathbf{F}$, а вторая – $\mathbf{G}$.

Factor-MSSA: Условия разделимости

Условия достижения слабой MSSA-разделимости [Голяндина Н., Некруткин В., Степанов Д., 2003]

$\mathbb{F}$ и $\mathbb{G}$ **слабо MSSA-разделимы**, тогда и только тогда, когда

1. Строковые пространства rowspan($\mathbf{F}$) и rowspan($\mathbf{G}$) ортогональны.
2. Столбцовые пространства colspan($\mathbf{F}$) и colspan($\mathbf{G}$) ортогональны.

Теорема: Условия достижения слабой Factor-MSSA-разделимости

$\mathbb{F}$ и $\mathbb{G}$ **слабо Factor-MSSA-разделимы**, если

1. Строковое пространство каждого одномерного ряда $\mathbf{F}^{(p)}$ из системы рядов первого ряда $\mathbf{F}$ (rowspan($\mathbf{F}^{(p)}$)) ортогонально строковому пространству rowspan($\mathbf{G}^{(q)}$), $\forall p, q$.
2. Столбцовое пространство каждого одномерного ряда $\mathbf{F}^{(p)}$ из системы рядов первого ряда $\mathbf{F}$ (colspan($\mathbf{F}^{(p)}$)) не совпадает со столбцовым пространством colspan($\mathbf{G}^{(q)}$), $\forall p, q$.

Factor-MSSA: Сравнение

$$\mathbb{X} = \mathbb{S} + \mathbb{N},$$
$$s_k = \begin{pmatrix} 10\sin(2\pi\frac{1}{8}k) \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 10\sin(2\pi\frac{1}{9}k) \end{pmatrix},$$
$$L = 51, 61, 71,\ N = 110,\ K = N - L + 1.$$

$\mathbb{N}$ – гауссовский белый шум с σ^2, 100 испытаний, мера качества – среднее значение MSE. Для проверки значимости paired t-test.

Метод	L=51 (σ=0)	L=61 (σ=0)	L=71 (σ=0)	L=51 (σ=5)	L=61 (σ=5)	L=71 (σ=5)
Factor-MSSA	0	0	0	2.648	2.344	2.803
MSSA	25.020	21.979	8.631	23.481	18.351	13.355
1d-SSA	0	0	0	2.617	2.368	2.792

MSE: Factor-MSSA $\approx$ 1d-SSA $<$ MSSA

Factor-MSSA: Сравнение

$$\mathbb{X} = \mathbb{S} + \mathbb{N},$$

$$s_k = \begin{pmatrix} 10\sin(2\pi\frac{1}{6}k) \\ 12\sin(2\pi\frac{1}{6}k) \end{pmatrix} + \begin{pmatrix} 14\sin(2\pi\frac{1}{10}k) \\ 16\sin(2\pi\frac{1}{10}k) \end{pmatrix}$$

$$L = 51, 61, 71,\ N = 110,\ K = N - L + 1.$$

$\mathbb{N}$ – гауссовский белый шум с σ^2, 100 испытаний, мера качества – среднее значение MSE. Для проверки значимости paired t-test.

Метод	L=51 (σ=0)	L=61 (σ=0)	L=71 (σ=0)	L=51 (σ=5)	L=61 (σ=5)	L=71 (σ=5)
Factor-MSSA	2.196	2.915	1.974	6.603	7.098	5.535
MSSA	2.196	2.915	1.974	6.603	7.098	5.560
1d-SSA	2.234	2.971	2.003	8.083	9.226	7.157

MSE: Factor-MSSA $\approx$ MSSA $<$ 1d-SSA

Factor-MSSA: Результаты

- Реализован алгоритм метода Factor-MSSA, основная сложность – алгоритм MVarimax.
- Получены условия разделимости для Factor-MSSA и проведено их сравнение с условиями разделимости для MSSA.
- Рассмотрена модификация метода, более устойчивая в случае приближенной MSSA-разделимости.
- Результаты проиллюстрированы численными примерами.
- Проведено численное сравнение Factor-MSSA, MSSA и одномерного SSA.

Задача: $\{Q, R\}$-разложения

$\{Q, R\}$-разложения: Постановка задачи

$$\mathbb{X} = \mathbb{S} + \mathbb{N}, \ \mathbf{X} = \mathcal{T}(\mathbb{X}).$$

Задача: выделение сигнала $\mathbb{S}$ конечного ранга.

Решение: метод HSLRA аппроксимации $\mathbf{S}$ (траекторной матрицы $\mathbb{S}$) с весами $Q \in \mathbb{R}^L$, $R \in \mathbb{R}^K$; в итерационном алгоритме метода одна итерация совпадает с SSA.

Результат: аппроксимация $\mathbb{S}$ по взвешенному МНК с весами W, где W – свертка Q и R, то есть $W = Q \star R$.

Обычно Q и R берутся единичными, и поэтому W неравномерные.

Проблема: найти такие Q и R, что $Q \star R = (1, \ldots, 1)$, $q_i \geqslant 0, r_i \geqslant 0$.

$\{Q, R\}$-разложения: Свойства

Множество решений:

Множество всех решений для заданного N:
$\mathcal{S}_N = \{\{Q, R\} : Q \star R = \underbrace{(1, \ldots, 1)}_{N}, |Q| \geqslant |R|\}$.

Свойства $\{Q, R\}$-разложений [Zhigljavsky A., Golyandina N., Gryaznov S., 2016]:

1. Q, R – состоят из нулей и единиц, но не из всех единиц (кроме случая $\{1, \underbrace{1 \ldots 1}_{N}\}$). Можно сопоставлять Q и R двоичные числа.
2. Q, R – палиндромы.

Bin – оператор представления вектора из разложения как двоичного числа.
Доказано, что $\text{Bin}(Q) \mid (2^N - 1)$ и $\text{Bin}(R) \mid (2^N - 1)$. Поэтому будем искать Q и R с помощью делителей $2^N - 1$.

$\{Q, R\}$-разложения: Класс решений

Класс разложений $\mathcal{C}$:

1. $\{1, 1\} \in \mathcal{C}$.
2. Если $\{Q, R\} \in \mathcal{C}$, то $\{(R\underbrace{0\ldots0}_{|Q|-1})^p R, Q\} \in \mathcal{C}$, $p = 1, 2, \ldots$

Разложения для заданного N, лежащие в классе $\mathcal{C}$:
$\mathcal{C}_N = \mathcal{S}_N \cap \mathcal{C}$, $\mathcal{S}_N$ – множество всех $\{Q, R\}$-разложений.

Свойства класса $\mathcal{C}$:

- Теорема: $\mathcal{C} \subset \bigcup\limits_N \mathcal{S}_N$.
- Проверено перебором: $\mathcal{C}_N = \mathcal{S}_N$, $N \leqslant 600$.

Гипотеза: $\mathcal{C}_N = \mathcal{S}_N, \forall N$.

$\{Q, R\}$-разложения: Упорядоченные факторизации

$\mathcal{C}_N$ – множество $\{Q, R\}$-разложений из класса $\mathcal{C}$.
$\mathcal{OF}_N$ – множество упорядоченных факторизаций [OEIS A074206].

Упорядоченная факторизация числа N – упорядоченный набор чисел $(m_1, m_2, \ldots, m_k)$, такой что $N = \prod_{i=1}^{k} m_i$.

Например, для $N = 12$:
$(12), (6, 2), (2, 6), (3, 4), (4, 3), (3, 2, 2), (2, 3, 2), (2, 2, 3)$.

Теорема: Между множествами $\mathcal{C}_N$ и $\mathcal{OF}_N$ есть биекция.

Результат: Алгоритм нахождения $\{Q, R\}$-разложений с использованием данной биекции.

$\{Q, R\}$-разложения: Алгоритм

Алгоритм нахождения разложений из $\mathcal{C}_N$ по заданному N.

- Нахождение упорядоченных факторизаций $\mathcal{OF}_N$.
  - Нахождение делителей N.
  - Перебор подмножеств мультимножества делителей [Knuth D.E., 2009].
- Построение $\mathcal{C}_N$ по $\mathcal{OF}_N$ с помощью доказанной биекции между $\mathcal{C}_N$ и $\mathcal{OF}_N$.

N	Тривиальный алгоритм (с.)	Данный алгоритм (с.)
256	35.463	0.253
10400	>10000	11.517

Результат: Web-приложение (реализация данного алгоритма)
`http://0101-nightuser.rhcloud.com/`

$\{Q, R\}$-разложения: Результаты

1. Описан класс $\mathcal{C}$, и выдвинута гипотеза о том, что класс содержит все разложения.
2. Доказано, что определение класса корректно, то есть все его элементы являются разложениями.
3. Предложен и реализован алгоритм, находящий по заданному N все разложения из класса $\mathcal{C}$, доказана его корректность.
4. Показано, что для небольших N ($N \leqslant 600$) алгоритм находит все разложения, и при этом его производительность превосходит производительность наивного перебора.
5. Создан web-сервис, позволяющий найти разложения для любого N:
   `http://0101-nightuser.rhcloud.com/`