Идентификация компонент в анализе сингулярного спектра

Жорникова Полина Георгиевна, гр. 16.М03-мм

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Статистическое моделирование

Научный руководитель: к.ф.-м.н., доцент Голяндина Н.Э.
Рецензент: к.ф.-м.н., лектор Пепелышев А.Н.

Санкт-Петербург
2018г.

Введение: Общая постановка задачи

$\mathbb{X}$ — объект: временной ряд (вещественный или комплексный), система временных рядов, прямоугольное цифровое изображение.

Проблема: найти составляющие разложения $\mathbb{X} = \mathbb{T} + \mathbb{P} + \mathbb{N}$, где

$\mathbb{T}$ — **тренд** (медленно меняющаяся составляющаяся);
$\mathbb{P}$ — **колебательная составляющая** (различные регулярные колебания);
$\mathbb{N}$ — **шум** (случайный процесс с нулевым средним).

Метод: SSA (Singular Spectrum Analysis) [Golyandina N., Nekrutkin V., Zhigljavsky A., 2001].

Вход: объект $\mathbb{X}$: временный ряд, система рядов, изображение; оператор $\mathcal{T}$.

Алгоритм:

1. $\mathbf{X} = \mathcal{T}(\mathbb{X})$, $\mathbf{X} \in \mathcal{M}_{L,K}^{(H)}$ — *траекторная матрица*.
2. $\mathbf{X} = \sum_{j=1}^{d} \mathbf{X}_j$, $\mathbf{X}_j = \mu_j U_j V_j^*$ — сингулярное разложение.
   Каждая компонента $\mathbf{X}_j$ характеризуется тремя значениями:
   U_j — левый и V_j — правый синг. вектора, μ_j — сингулярное число.
3. $\mathbf{X} = \sum_{j=1}^{d} \mathbf{X}_j \longrightarrow \mathbf{X} = \mathbf{X}_{I_\mathbb{T}} + \mathbf{X}_{I_\mathbb{P}} + \mathbf{X}_{I_\mathbb{N}}$, $\mathbf{X}_I = \sum_{j \in I} \mathbf{X}_j$.
4. $\mathbb{X} = \widetilde{\mathbb{T}} + \widetilde{\mathbb{P}} + \widetilde{\mathbb{N}} = \mathcal{T}^{-1}\left(\Pi^{(H)}\mathbf{X}_{I_\mathbb{T}}\right) + \mathcal{T}^{-1}\left(\Pi^{(H)}\mathbf{X}_{I_\mathbb{P}}\right) + \mathcal{T}^{-1}\left(\Pi^{(H)}\mathbf{X}_{I_\mathbb{N}}\right)$,
   $\Pi^{(H)}$ — ортогональный проектор на $\mathcal{M}_{L,K}^{(H)}$ по норме Фробениуса.

Выход: разложение $\mathbb{X} = \widetilde{\mathbb{T}} + \widetilde{\mathbb{P}} + \widetilde{\mathbb{N}}$.

Как осуществить группировку на этапе 3?

Введение: Постановка задачи работы

Метод	Данные	Обозначение
1D-SSA	временной ряд	$\mathbb{X} = (x_1, \ldots, x_N)$
CSSA	комплексный временной ряд	$\mathbb{X}^{(1)} + \mathrm{i}\mathbb{X}^{(2)}$
MSSA	система временных рядов	$\mathbb{X}^{(p)}$, $p = 1, \ldots, s$
2D-SSA	прямоугольное изображение	$\mathbb{X} = (x_{ij})_{i,j=1}^{N_x, N_y}$

Хотим сгруппировать компоненты разложения $\mathbf{X} = \sum_{j=1}^{d} \mu_j U_j V_j^*$ в группы $\mathbf{X} = \mathbf{X}_{I_\mathbb{T}} + \mathbf{X}_{I_\mathbb{P}} + \mathbf{X}_{I_\mathbb{N}}$, чтобы получить $\mathbb{X} = \widetilde{\mathbb{T}} + \widetilde{\mathbb{P}} + \widetilde{\mathbb{N}}$.

Задача работы: разработать алгоритмы автоматической группировки (идентификации) индексов $\{1, \ldots, d\}$ в три группы $I_\mathbb{T}$, $I_\mathbb{P}$, $I_\mathbb{N}$ для всех вариантов SSA.

Метод	Данные	Обозначение
1D-SSA	временной ряд	$\mathbb{X} = (x_1, \ldots, x_N)$
CSSA	комплексный временной ряд	$\mathbb{X}^{(1)} + \mathrm{i}\mathbb{X}^{(2)}$
MSSA	система временных рядов	$\mathbb{X}^{(p)}$, $p = 1, \ldots, s$
2D-SSA	прямоугольное изображение	$\mathbb{X} = (x_{ij})_{i,j=1}^{N_x, N_y}$

Хотим сгруппировать компоненты разложения $\mathbf{X} = \sum_{j=1}^{d} \mu_j U_j V_j^*$ в группы $\mathbf{X} = \mathbf{X}_{I_\mathbb{T}} + \mathbf{X}_{I_\mathbb{P}} + \mathbf{X}_{I_\mathbb{N}}$, чтобы получить $\mathbb{X} = \widetilde{\mathbb{T}} + \widetilde{\mathbb{P}} + \widetilde{\mathbb{N}}$.

Задача работы: разработать алгоритмы автоматической группировки (идентификации) индексов $\{1, \ldots, d\}$ в три группы $I_\mathbb{T}$, $I_\mathbb{P}$, $I_\mathbb{N}$ для всех вариантов SSA.

Было сделано в работе:

1. Обзор существующих методов автоматической группировки, автоматической идентификации тренда и колебательной составляющей для одномерного вещественного ряда и 1D-SSA.
2. Создание нового метода автоматической идентификации колебательной составляющей для 1D-SSA.
3. Обобщения методов для расширений SSA (CSSA, MSSA, 2D-SSA).
4. Реализация ранее не реализованных методов для 1D-SSA и всех обобщений на языке программирования R.
5. Примеры реальных приложений методов.

План рассказа:

1. Описание существующих и нового методов автоматической идентификации тренда и колебательной составляющей для 1D-SSA.
2. Описание обобщений методов идентификации тренда и колебательной составляющей для CSSA, MSSA, 2D-SSA.
3. Примеры реальных приложений методов.

- **Тренд** $\mathbb{T}$ — составляющая с низкими частотами: в разложении Фурье ряда $\mathbb{T} = (t_1, \ldots, t_N)$ длины N

$$t_n = c_0 + \sum_{k=1}^{\lfloor (N-1)/2 \rfloor} \sqrt{c_k^2 + s_k^2} \cos(2\pi nk/N + \phi_k) + c_{N/2}(-1)^n$$

  большое значение имеют только $\sqrt{c_k^2 + s_k^2}$ с маленькими k.
- **Колебательная составляющая** $\mathbb{P}$ — сумма э-м гармоник вида $a\, e^{\alpha n} \cos(2\pi\omega n + \phi)$, $\omega < 0.5$, $0 \leqslant \phi < 2\pi$, $a \neq 0$.

- Сингулярные вектора U_j, V_j разложения $\mathbf{X} = \sum_{j=1}^{d} \mu_j U_j V_j^*$ имеют такую же структуру, что и компонента $\mathbb{X}_j$, которой они соответствуют.
- Э-м гармонике соответствует 2 члена разложения $\mathbf{X} = \sum_{j=1}^{d} \mu_j U_j V_j^*$.

1. Для выделения **тренда** каждую компоненту рассматриваем по отдельности и используем частотный подход.
2. В случае **колебательной составляющей** учитываем соотношения между компонентами (например, между сингулярными векторами).

Вещественный ряд $\mathbb{X} = (x_1, \ldots, x_N)$. **Периодограмма** ряда

$$\Pi_{\mathbb{X}}^N(k/N) = \frac{N}{2} \begin{cases} 2c_0^2 & \text{для } k = 0, \\ c_k^2 + s_k^2 & \text{для } 0 < k < N/2, \\ 2c_{N/2}^2 & \text{для } k = N/2, \text{ если } N \text{ четное}, \end{cases}$$

c_k и s_k — коэффициенты разложения Фурье ряда $\mathbb{X}$.

Параметры метода: частота $0 \leqslant \omega_0 \leqslant 0.5$, порог $0 \leqslant T_0 \leqslant 1$.

Мера $T(\mathbb{X}; \omega_0) = T\left(\sum_{k:0 \leqslant k/N < \omega_0} \Pi_{\mathbb{X}}^N(k/N)\right)$ **— вклад частот**, содержащихся в интервале $[0, \omega_0)$.

Вход метода: $\mathbb{Y}_i$, $i = 1, \ldots, d$, — либо ряд, восстановленный по i-ой компоненте, либо левый U_i или правый V_i сингулярный вектор.

Алгоритм: отбираем компоненты, для которых $T(\mathbb{Y}_i; \omega_0) \geqslant T_0$.

Э-м гармоника: $a\,e^{\alpha n}\cos(2\pi\omega n+\phi)$, $0\leqslant\phi<2\pi$, $a\neq 0$; $\mathbf{X}=\sum_{j=1}^{d}\mu_j U_j V_j^*$.
Имеет два сингулярных вектора U_1 и U_2, L — длина векторов.

Вход метода: $\{U_j\}_{j=1}^{r}$ — левые сингулярные вектора разложения ряда $\mathbb{X}$.
Параметры метода: порог $0\leqslant\rho_0\leqslant 1$.

Для немодулированной гармоники ($\alpha=0$) с $L\omega\in\mathsf{Z}_+$ максимумы периодограмм $\Pi_{U_1}^{L}(k/L)$ и $\Pi_{U_2}^{L}(k/L)$ достигаются в одной точке и равны 1.

Схема метода: поверяем условие про периодограммы и отбираем нужные компоненты с помощью порога ρ_0.

Проблема метода: нет универсальных рекомендаций по выбору порога ρ_0, и есть обоснование только для немодулированных гармоник.

Э-м гармоника: $a\,e^{\alpha n}\cos(2\pi\omega n+\phi)$, $0\leqslant\phi<2\pi$, $a\neq 0$; $\mathbf{X}=\sum_{j=1}^{d}\mu_j U_j V_j^*$.
Имеет два сингулярных вектора U_1 и U_2, L — длина векторов.

Специальная мера:

- $\tau(U_1,U_2):=\frac{1}{L-1}\sum_{k=1}^{L-1}\left(\theta_k-\bar{\theta}\right)^2$, где

  θ_k — угол между $\left(u_k^{(1)},u_k^{(2)}\right)^{\mathrm{T}}$ и $\left(u_{k+1}^{(1)},u_{k+1}^{(2)}\right)^{\mathrm{T}}$, $\bar{\theta}=\frac{1}{L-1}\sum_{k=1}^{L-1}\theta_k$.
- $\lim_{L\to\infty}\tau(U_1,U_2)=0$ при некоторых условиях.

Вход: $\{U_j\}_{j=1}^{r}$ — сингулярные вектора разложения ряда $\mathbb{X}$.

Параметры: порог $\tau_0>0$.

Алгоритм: отбираем пары индексов $j,j+1$, для которых $\tau(U_j,U_{j+1})<\tau_0$.

Преимущество метода: обоснован для модулированных гармоник.

Проблема метода: нет универсальных рекомендаций по выбору порога τ_0.

1D-SSA: Колебательная составляющая: Сравнение частотного метода и метода идентификации по регулярности углов

Ряд: $e^{0.02k}\cos(2\pi k/7) + e^{0.02k}\sigma\varepsilon_k$, $\varepsilon_k \sim N(0,1)$; $N = 99$, $L = 50$.

- Параметры методов подбирались специальным алгоритмом, одинаковым для обоих методов, и считалась ошибка идентификации.
- Методы сравнивались на одних и тех же реализациях рядов.
- Проводилось 1000 моделирований.

Таблица: Средняя ошибка идентификации.

	Для метода по регулярности углов	Для частотного метода
$\sigma = 0.2$	0.0003	0.050
$\sigma = 0.4$	0.0010	0.080
$\sigma = 0.6$	0.0082	0.103
$\sigma = 0.8$	0.0549	0.185
$\sigma = 1$	0.1917	0.239

$x_k = e^{0.01k} + 2\cos(2\pi k/10) + e^{0.009k}\cos(2\pi k/4) + \varepsilon_k$, $\varepsilon_k \sim N(0, 2)$, $N = 199$.

Верная группировка компонент: 1 — тренд, 2–5 — колеб. составляющая.

Метод низких частот для тренда: частота $\omega_0 = 0.01$, порог $T_0 = 0.9$.

Номер компоненты	1	26	...
Значение меры T	0.93	0.07	...

Частотный метод для колеб. составляющей: порог $\rho_0 = 0.9$.

Номера компонент	4–5	2–3	10–11	...
Значение меры ρ	0.96	0.95	0.73	...

Метод по регулярности углов для колеб. составляющей: порог $\tau_0 = 0.01$.

Номера компонент	4–5	2–3	6–7	...
Значение меры τ	0.0009	0.0040	0.0566	...

$$x_k = e^{0.01k} + 2\cos(2\pi k/10) + e^{0.009k}\cos(2\pi k/4) + \varepsilon_k,\ \varepsilon_k \sim N(0,2),\ N = 199.$$

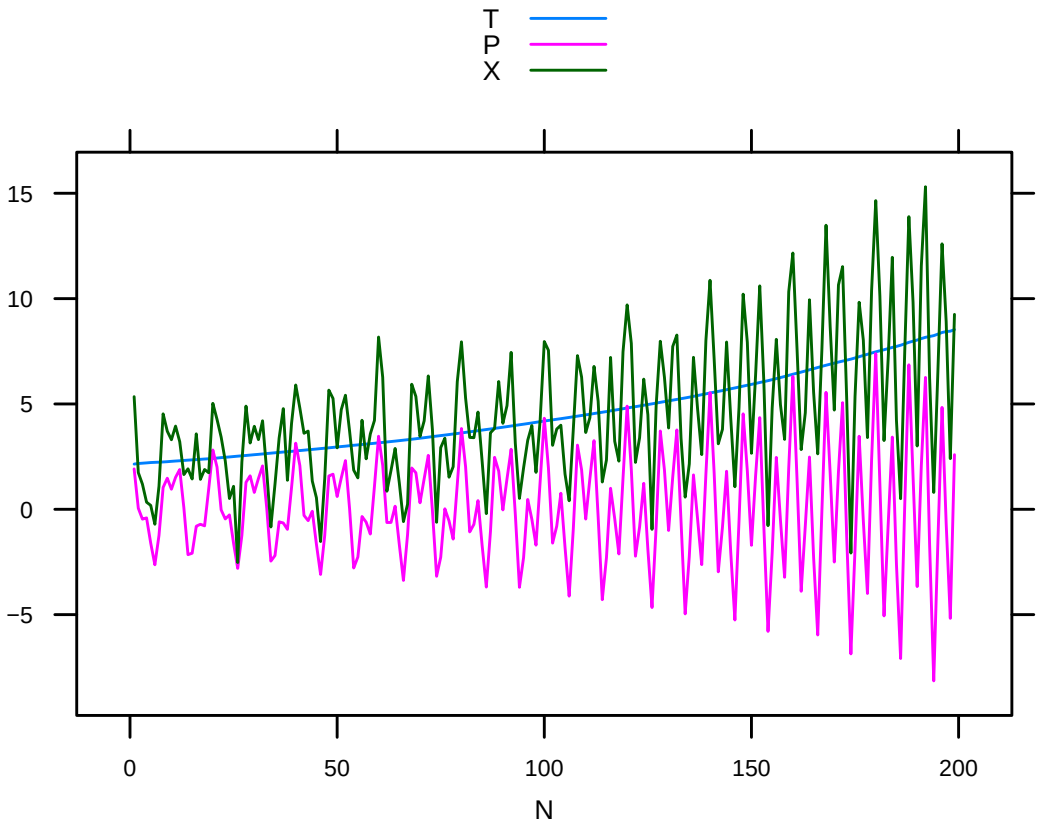


Рис.: Восстановленные тренд (по компоненте 1) и колебательная составляющая (по компонентам 2–3, 4–5) на фоне исходного ряда.

Можно ли идеи методов автоматической идентификации для 1D-SSA обобщить на другие варианты метода: 2D-SSA, MSSA, CSSA?

Да = обобщили метод; Нет = пока непонятно, как обобщить.

	CSSA	MSSA	2D-SSA
Метод низких частот для тренда	Да	Да	Да
Частотный метод для колебательной составляющей	Да	Да	Нет
Метод идентификации колебательной составляющей по регулярности углов	Да	Да	Нет

Комплексный ряд $\mathbb{X} = \mathbb{X}^{(1)} + \mathrm{i}\mathbb{X}^{(2)}$.

Сингулярное разложение: $\mathbf{X} = \sum_{j=1}^{d} \mu_j U_j V_j^*$.

Особенности CSSA:

- И левые, и правые сингулярные вектора, и восстановленные ряды являются комплексными.
- Сингулярные вектора определяются с точностью до комплексного поворота, т.е. умножения на $e^{2\mathrm{i}\pi t}$, $0 \leqslant t < 1$.
- В случае колебательной составляющей идентифицируем комплексные экспоненты: $e^{\mathrm{i}(2\pi\omega k+\phi)+\alpha k}$. В сингулярном разложении комплексной экспоненте соответствует один сингулярный вектор.

С учетом особенностей были обобщены методы:

- метод низких частот для тренда;
- частотный метод для колебательной составляющей;
- метод идентификации колеб. составляющей по регулярности углов.

Система временных рядов $\mathbb{X}^{(p)}$, $p = 1, \ldots, s$.

Сингулярное разложение: $\mathbf{X} = \sum_{j=1}^{d} \mu_j U_j V_j^*$.

Особенности MSSA:

- И левые, и правые сингулярные вектора, и восстановленные ряды имеют разную структуру.
- Применение алгоритмов к разным видам входных данных может давать разные результаты.
- Для разных входных данных нужны разные варианты алгоритма.

С учетом особенностей были обобщены методы:

- метод низких частот для тренда;
- частотный метод для колебательной составляющей;
- метод идентификации колеб. составляющей по регулярности углов.

Поле $\mathbb{X} = (x_{ij})_{i,j=1}^{N_x,N_y}$. **Двумерная периодограмма**

$$\Pi_{\mathbb{X}}^{N_x N_y}\left(\frac{k}{N_x}, \frac{l}{N_y}\right) = N_x N_y |G_{kl}|^2,$$

где $0 \leqslant k \leqslant N_x$, $0 \leqslant l \leqslant N_y$,

G_{kl} — коэффициент двумерного разложения Фурье ряда $\mathbb{X}$.

Параметры метода: частоты $0 \leqslant \omega_0^{(1)}, \omega_0^{(2)} \leqslant 0.5$, порог $0 \leqslant T_0 \leqslant 1$.

$$T(\mathbb{X}; \omega_0^{(1)}, \omega_0^{(2)}) = T\left(\sum_{k:0\leqslant k/N_x<\omega_0^{(1)}} \sum_{l:0\leqslant l/N_y<\omega_0^{(2)}} \Pi_{\mathbb{X}}^{N_x N_y}\left(\tfrac{k}{N_x}, \tfrac{l}{N_y}\right)\right) -$$

вклад частот, содержащихся в $\left\{\left(-\omega_0^{(1)}, \omega_0^{(1)}\right) \times \left(-\omega_0^{(2)}, \omega_0^{(2)}\right)\right\}$.

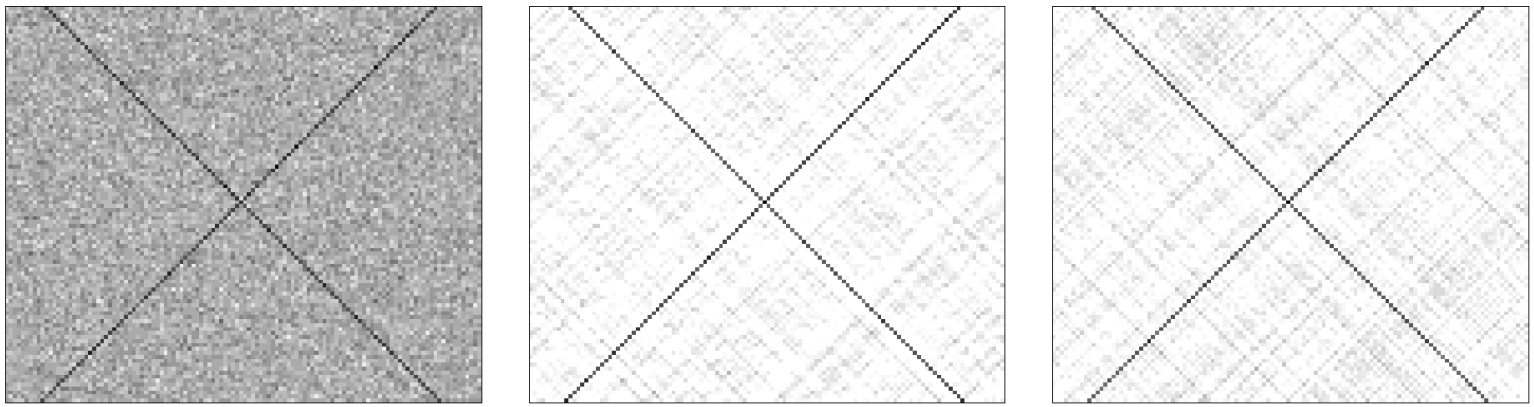
Далее алгоритм такой же, как в 1D-SSA.

Реальные примеры применения: Выделение линий с изображения

Алгоритм (Trickett, 2003) выделяет линии с зашумленного изображения.

На одном из этапов применяется преобразование Фурье, и задача сводится к выделению комплексных экспонент. Используется метод CSSA.

Применим два метода автоматической идентификации колебательной составляющей: метод по регулярности углов и частотный метод.



(a) Исходное изображение.

(b) Алгоритм с методом по регулярности углов.

(c) Алгоритм с частотным методом.

Рис.: Изображение с двумя линиями.

Реальные примеры применения: Анализ данных активности генов

Двумерные данные: значения активности гена, измеренные на неравномерной двумерной решетке.

Предполагается, что данные зашумлены, и шум имеет дисперсию σ^2.

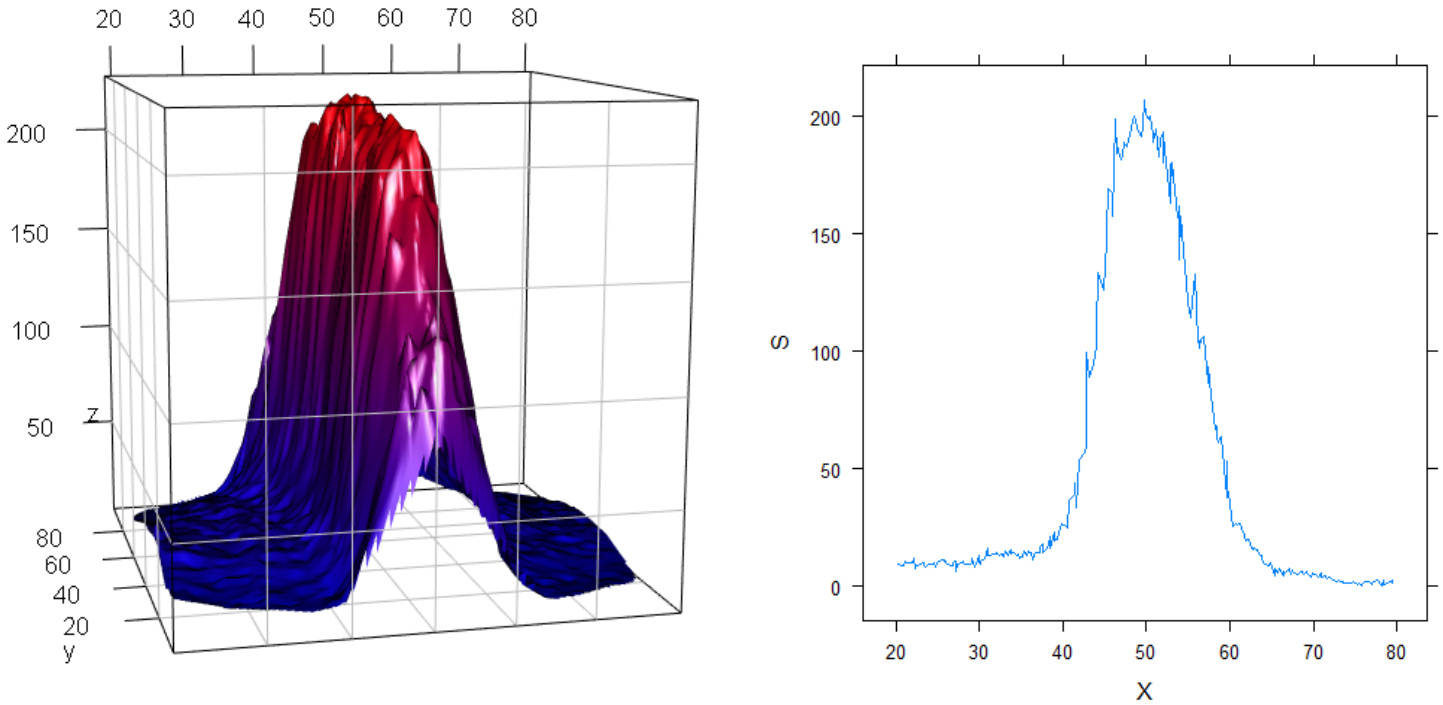
Исходная задача: оценить σ^2.
Промежуточная задача: оценить тренд.

Качество идентификации тренда можно оценивать по ошибке оценки σ.

Два подхода к решению:

- Рассмотреть одномерный срез данных и решать задачу дня них.
- Решать задачу для исходных $2D$-данных.

Реальные примеры применения: Анализ данных активности генов



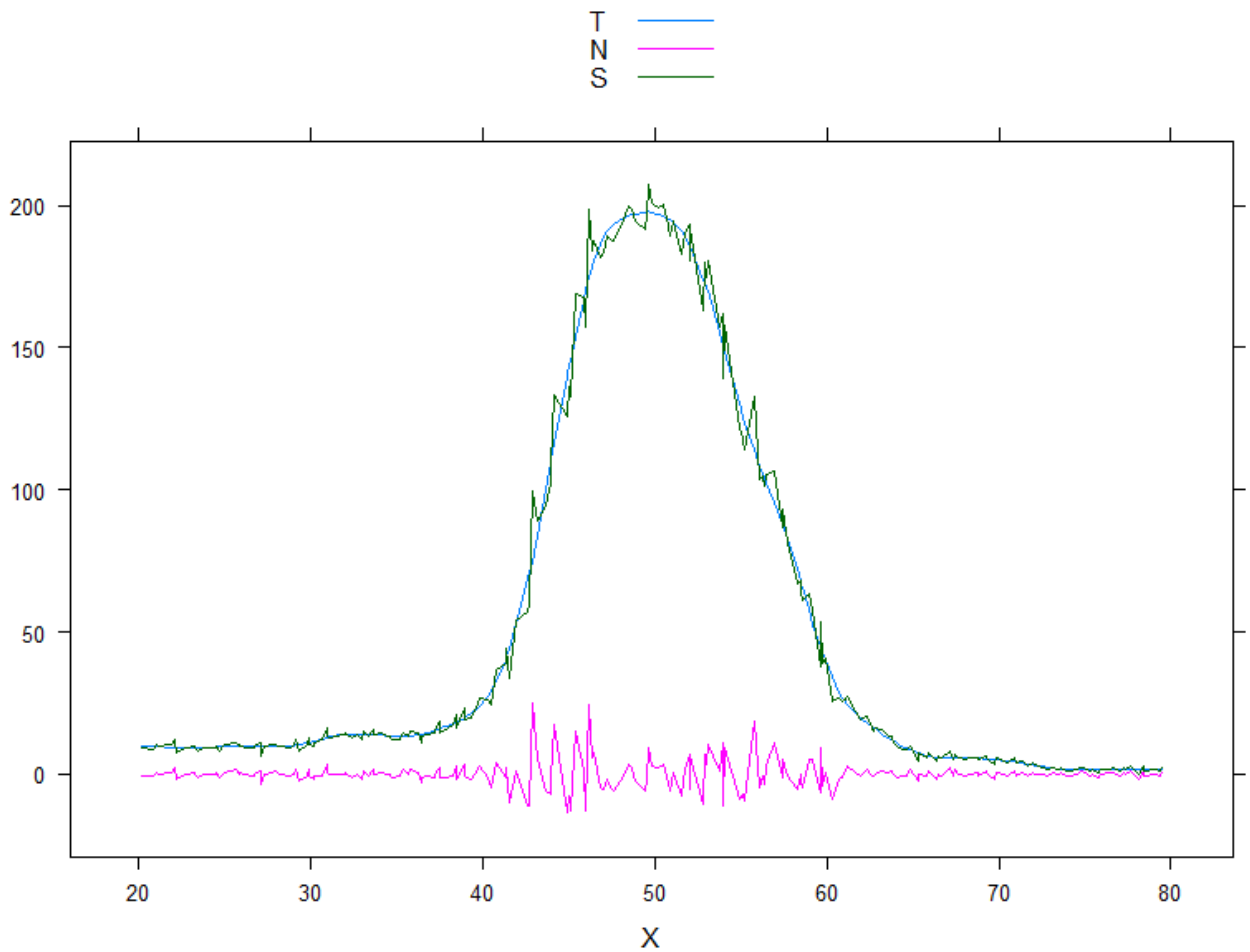
(a) Исходные данные

(b) Одномерный срез

Рис.: Изображение данных.

Идея: подбираем параметры ω_0 и T_0 для автоматической идентификации тренда на модельных данных, имитирующих поведение реальных. Используем полученные параметры для реальных данных.

Значения подобранных параметров: $\omega_0 = 0.04$, $T_0 = 0.4$.

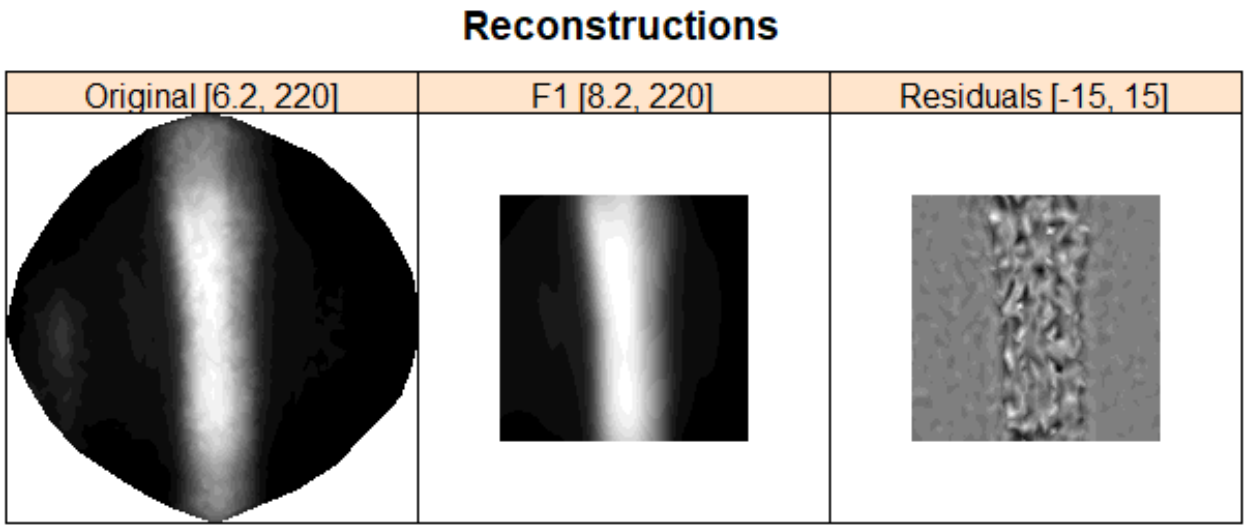


Реальные примеры применения: анализ данных активности генов

Перед 2D-SSA процедурой данные были интерполированы на регулярную решетку с шагом $\Delta = 0.5$.

Идея: подбираем параметры $\omega_0^{(1)}, \omega_0^{(2)}$ и T_0 для автоматической идентификации тренда на модельных данных, имитирующих поведение реальных. Используем полученные параметры для реальных данных.

Значения подобранных параметров: $\omega_0^{(1)} = 0.08\Delta$, $\omega_0^{(2)} = 0.1\Delta$, $T_0 = 0.2$.



- 2D-SSA дал более высокую точность оценок.
- Автоматическая идентификация практически такого же качества, как идентификация с фиксированным числом компонент.

Таблица: Модельные данные, истинное значение параметра $\sigma^2 = 0.03$.

	1D	2D	1D auto	2D auto
Среднее значение оценки	0.0285	0.0311	0.0285	0.0318
sd	0.0037	0.0006	0.0037	0.0008

Таблица: Реальные данные.

	1D	2D	1D auto	2D auto
Среднее значение оценки	0.0392	0.0347	0.0390	0.0341
sd	0.0123	0.0063	0.0120	0.0054

Результаты

- Для метода SSA сделан обзор существующих методов автоматической идентификации тренда, колебательной составляющей, методов группировки компонент для одномерного временного ряда.
- Предложен новый метод идентификации колебательной составляющей для одномерного ряда, который работает эффективнее, чем ранее известный.
- Многие описанные методы обобщены для случаев комплексного ряда, системы временных рядов, цифрового изображения.
- Методы для одномерного ряда, нереализованные ранее в R-пакете Rssa, и все обобщенные варианты методов реализованы на языке R.
- Приведены реальные примеры применения разработанных методов для разных вариантов входного объекта.