Проекции в методе анализа сингулярного спектра

Сазыкин Дмитрий Сергеевич, 422 гр.

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

Научный руководитель: к.ф.–м.н., доц. Голяндина Н.Э.
Рецензент: м.н.с. Шлемов А.Ю.

СтатМод

5 июня 2017 г.

Введение: Постановка задачи

Временной ряд $F = (f_0, \ldots, f_{N-1})$, $f_i \in \mathbb{R}$.

Рассматриваемая модель: $F = F^{(\mathrm{tr})} + F^{(\mathrm{s})} + F^{(\mathrm{n})}$, где

$f_i^{(\mathrm{tr})} = f^{(\mathrm{tr})}(x_i) = ax_i + b$ – тренд

$f_i^{(\mathrm{s})} = f^{(\mathrm{s})}(x_i) = \sum\limits_{j=1}^{J} c_j \sin(2\pi\omega_j x_i + \varphi_j)$ – неслучайная помеха

$f_i^{(\mathrm{n})} = \varepsilon_i$ – гауссовский белый шум

в равноотстоящих узлах $x_i = i$.

Задача

- Оценка линейного тренда
- Прогноз линейного тренда

Введение: Оценка линейного тренда

Рассматриваемые методы

1. Метод наименьших квадратов (МНК, или OLS, ordinary least squares) [Линник, 1962].
2. Метод анализа сингулярного спектра (АСС, или SSA, singular spectrum analysis) [Golyandina N., Nekrutkin V., Zhigljavsky A., 2001].

Таким образом, имеется два метода оценки тренда временного ряда.

Задачи

- Исследовать влияние параметров ряда и параметров методов на ошибку оценивания линейного тренда.
- С учетом полученных результатов рассмотреть комбинированные методы с улучшенной точностью.
- Численно сравнить ошибки методов.

МНК: Алгоритм

Постановка задачи МНК. Пусть $F = \{f_i\}_{i=0}^{N-1}$ — некоторые измерения в точках x_i. Найти $\hat{a}$ и $\hat{b}$ как аргумент минимума:

$$\operatorname*{arg\,min}_{a',b' \in \mathbb{R}} \sum_{i=0}^{N-1} \Big(f_i - (a'x_i + b')\Big)^2.$$

Замечание. Пусть $F = F^{(\mathrm{tr})} + G$. Тогда ошибки оценки тренда $\widetilde{f}_i^{(\mathrm{tr})} - f_i^{(\mathrm{tr})}$ не зависят от коэффициентов линейного тренда $F^{(\mathrm{tr})}$.

Утверждение. В модели $f_i = ai + b + \sum_{j=1}^{J} c_j \sin(2\pi\omega_j i + \varphi_j)$, $0 < \omega_j \leq 0.5$ при длине ряда $N \to \infty$ ошибка МНК-оценки стремится к нулю:

$$\frac{1}{N} \sum_{i=0}^{N-1} \Big(ai + b - \widehat{a}_N i - \widehat{b}_N\Big)^2 \xrightarrow[N\to\infty]{} 0.$$

Замечание. Будем предполагать, что $\omega = \omega_1$ — фундаментальная частота, т.е. $\omega_j = k_j\omega$, $0 < \omega_j < 0.5$ для некоторых целых $k_j > 1$.

МНК: Поведение ошибки в непрерывном случае

Рассмотрим непрерывный аналог дискретной модели, положив $N_{\mathrm{per}}^{(j)} = \omega_j N$, $A = a/N$, $B = b$, $C_j = c_j$:

$$f(t) = At + B + \sum_{j=1}^{J} C_j \sin(2\pi N_{\mathrm{per}}^{(j)} t + \varphi_j), \quad t \in [0,1].$$

Тогда $N_{\mathrm{per}} = N_{\mathrm{per}}^{(1)} = N_{\mathrm{per}}^{(j)}/k_j$.

Задача МНК: $\int\limits_0^1 \big(f(t) - (A't + B')\big)^2 dt \to \min\limits_{A',B' \in \mathbb{R}}$.

Утверждение. Пусть $N_{\mathrm{per}} = N_{\mathrm{per}}^{(1)}$ целое. Тогда существует такой сдвиг $0 \le \delta < 1/N_{\mathrm{per}}$, что оценка по МНК для ряда $f(t+\delta)$ точно выделяет линейный тренд.

МНК: Поведение ошибки в непрерывном случае

Рассмотрим частный случай:

$$f(t) = At + B + C\sin(2\pi N_{\mathrm{per}} t + \varphi), \quad t \in [0, 1].$$

Утверждение. MSE ошибка МНК-оценки тренда равна $\dfrac{\widehat{A}^2}{12} + S_{\mathrm{f}}^2$, где

$$\widehat{A} = 6C\cos(\pi N_{\mathrm{per}} + \varphi)\left(\frac{\sin(\pi N_{\mathrm{per}})}{\pi^2 N_{\mathrm{per}}^2} - \frac{\cos(\pi N_{\mathrm{per}})}{\pi N_{\mathrm{per}}}\right),$$

$$S_{\mathrm{f}} = C\frac{\sin(\pi N_{\mathrm{per}})\sin(\pi N_{\mathrm{per}} + \varphi)}{\pi N_{\mathrm{per}}}.$$

Следствие. Если N_{per} — целое, то ошибка равна $\dfrac{4C^2\cos^2(\varphi)}{\pi^2 N_{\mathrm{per}}^2}$, то есть минимум достигается при $\varphi = \pi/2$, а максимум при $\varphi = 0$.

SSA with double centering: Матрицы центрирования

Пусть $\mathbf{X} \in \mathbb{R}^{L \times K}$.

Матрица однократного центрирования:

По строкам: $\mathcal{A}(\mathbf{X})$ — каждый элемент заменяется на усредненное значение элементов строки.

По столбцам: $\mathcal{B}(\mathbf{X})$ — аналогично.

Замечание. Матрица однократного центрирования — результат проектирования строк/столбцов матрицы на пространство, порожденное вектором из единиц соответствующей размерности $(1, \ldots, 1)^{\mathrm{T}}$.

Матрица двойного центрирования:
$\mathbf{C}(\mathbf{X}) = \mathcal{A}(\mathbf{X}) + \mathcal{B}(\mathbf{X} - \mathcal{A}(\mathbf{X}))$ — последовательное построение матриц центрирования по строкам и по столбцам.

Замечание. Центрирование — линейная операция.

Пусть имеется ряд $F = (f_0, \ldots, f_{N-1})$.
Алгоритм метода SSA с двойным центрированием (SSAwDC):

1. Выбор длины окна L (по умолчанию $\lfloor (N+1)/2 \rfloor$)
2. Траекторная матрица:
$$(F, L) \Rightarrow \mathbf{X}$$
$$\mathbf{X} = \begin{bmatrix} f_0 & f_1 & \ldots & f_{K-1} \\ f_1 & f_2 & \ldots & f_K \\ \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & \ldots & f_{N-1} \end{bmatrix}$$
3. Построение матрицы двойного центрирования:
$$\mathbf{X} \Rightarrow \mathbf{C}(\mathbf{X})$$
4. Диагональное усреднение:
$$\mathbf{C}(\mathbf{X}) \Rightarrow (\widetilde{f}_0^{(\mathrm{tr})}, \ldots, \widetilde{f}_{N-1}^{(\mathrm{tr})})$$
$$\begin{bmatrix} c_{1,1} & c_{1,2} & c_{1,3} & \cdots & c_{1,K} \\ c_{2,1} & c_{2,2} & c_{2,3} & \cdots & c_{2,K} \\ c_{3,1} & c_{3,2} & c_{3,3} & \cdots & c_{2,K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{L,1} & c_{L,2} & c_{L,3} & \cdots & c_{L,K} \end{bmatrix}$$

Таким образом, построение SSAwDC-оценки тренда — линейное преобразование ряда с параметром L.

Замечание. Полученная оценка линейного тренда обычно не является линейной функцией.

SSAwDC: Свойства оценки тренда

Замечание [Golyandina, Shlemov, 2017].
SSA with double centering — частный случай SSA with projection.

Утверждение [Golyandina N., Nekrutkin V., Zhigljavsky A., 2001].
Если $\mathbf{X}^{(\mathrm{tr})}$ — траекторная матрица линейного тренда $F^{(\mathrm{tr})}$, то ее матрица двойного центрирования совпадает с ней: $\mathbf{C}(\mathbf{X}^{(\mathrm{tr})}) = \mathbf{X}^{(\mathrm{tr})}$.

Утверждение. Пусть дан ряд $F = F^{(\mathrm{tr})} + G$ длины N. Тогда ошибки SSAwDC-оценки тренда $\widetilde{f}_i^{(\mathrm{tr})} - f_i^{(\mathrm{tr})}$ не зависят от коэффициентов линейного тренда $F^{(\mathrm{tr})}$.

SSAwDC: Свойства оценки тренда

Модель: $f_i = ai + b + \sum_{j=1}^{J} c_j \sin(2\pi\omega_j i + \varphi_j),\ i = 0, \ldots, N-1.$

Утверждение [Golyandina N., Nekrutkin V., Zhigljavsky A., 2001]. При фиксированном N и выбранной длине окна L если все произведения $\omega_j L$ и $\omega_j(N+1)$ — целые числа, то SSAwDC точно выделяет линейный тренд.

Замечание. В отличие от МНК условия точного выделения линейного тренда не зависят от сдвига ряда.

Утверждение. Пусть $N \to \infty$ и $L(N) = [\alpha N]$, $0 < \alpha < 1$. Тогда ошибка оценки тренда с помощью SSAwDC в данной модели стремится к нулю

$$\frac{1}{N} \sum_{i=0}^{N-1} \left(ai + b - \widetilde{f}_i^{(\mathrm{tr})}(N) \right)^2 \xrightarrow[N \to \infty]{} 0.$$

Краткий итог: Сравнение методов

Модель без шума: $f_i = ai + b + c\sin(2\pi\omega i + \varphi),\ i = 0, \ldots, (N-1)$.

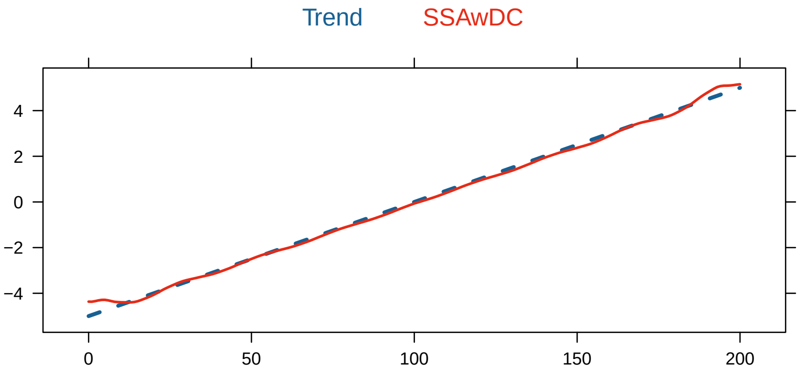
- OLS. Если ωN — целое, то есть N кратно периоду синуса, то при $\varphi = \pi/2$ ошибка оценивания линейного тренда будет наименьшей, а при $\varphi = 0$ — наибольшей.
- SSAwDC с длиной окна L. Если $L\omega$ и $(N+1)\omega$ — целые, то есть L и $N+1$ кратны периоду синуса, то ошибка SSAwDC-оценки линейного тренда равна 0.

Ошибка OLS-оценки зависит от сдвига периодической компоненты, однако в методе нет никаких дополнительных параметров.

Условие нулевой ошибки SSAwDC-оценки при правильном выборе длины окна L не зависит от сдвига периодической компоненты.

Ошибки обоих методов не зависят от a и b.

Комбинированные методы: SSAwDC+OLS



Пример SSAwDC-оценки линейного тренда.

SSAwDC+OLS. Результат работы SSAwDC не всегда является линейной функцией, поэтому применим к результатам OLS. Этот метод можно рассматривать как использование SSAwDC в качестве препроцессинга для OLS.

Комбинированные методы: cut-Методы

$$F = F^{(\mathrm{tr})} + F^{(\mathrm{s})} + F^{(\mathrm{n})}$$

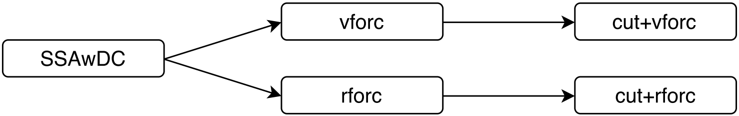
Схема (cut-методы)

1. Оценка периодической компоненты $F^{(\mathrm{s})}$ и ее периода T (с помощью пакета Rssa [Korobeynikov A., Shlemov A., Usevich K., Golyandina N., 2016]: `cran.r-project.org/package=Rssa`).
2. Обрезание ряда на основе теоретических свойств методов.

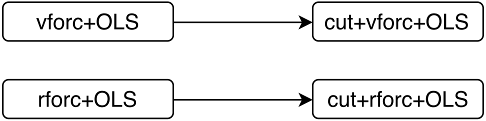
- cut+SSAwDC+OLS
    1. Рассмотрим последние R членов ряда F, такие, что $R+1$ кратно $\widehat{T}$, L возьмем ближайшее к $R/2$ и кратное $\widehat{T}$.
    2. Применим SSAwDC+OLS к обрезанному ряду.
- cut+OLS
    1. Рассмотрим отрезки ряда $\widehat{F}^{(\mathrm{s})}$ с длиной, кратной $\widehat{T}$ (просмотр сдвигов обрезанного ряда).
    2. Выберем отрезок, для которого OLS-оценка нулевого тренда дает наименьшую ошибку (поиск лучшего сдвига).
    3. Применим OLS к соответствующему отрезку исходного ряда F.

Прогноз линейного тренда

Рассмотрим алгоритмы rforecast и vforecast прогноза тренда на основе SSAwDC-оценки, которые реализованы в пакете Rssa.



К полученным оценкам для уменьшения ошибки применим OLS.



Оценки OLS, SSAwDC+OLS, cut+SSAwDC+OLS и cut+OLS являются линейными функциями, поэтому в качестве прогноза просто продолжим их вперед по формулам.

Практическая часть: Сравнение оценок линейного тренда

$$f_i = 0.1i - 10 + 7\sin(2\pi i/T_1 + \alpha_1) + 5\sin(2\pi i/T_2 + \alpha_2) + \varepsilon_i,$$
$$\varepsilon_i \in \mathrm{N}(0,1), \quad \alpha \in \mathrm{U}(0,\pi/2), \quad i = 0,\ldots,200,$$

Период T_1 — случайный от 16 до $N/2$, кратный четырем, $T_2 = T_1/2$.

Сравниваем ошибки оценивания тренда для методов OLS, SSAwDC, SSAwDC+OLS, cut+SSAwDC+OLS, cut+OLS.

Прогноз на 1 и на 100 шагов вперед сравнивается для OLS, SSAwDC+OLS, cut+SSAwDC+OLS, cut+OLS, vforc+OLS, rforc+OLS, cut+vforc+OLS и cut+rforc+OLS.

Сравниваем ошибку RMSE на основе 1000 реализаций модели:

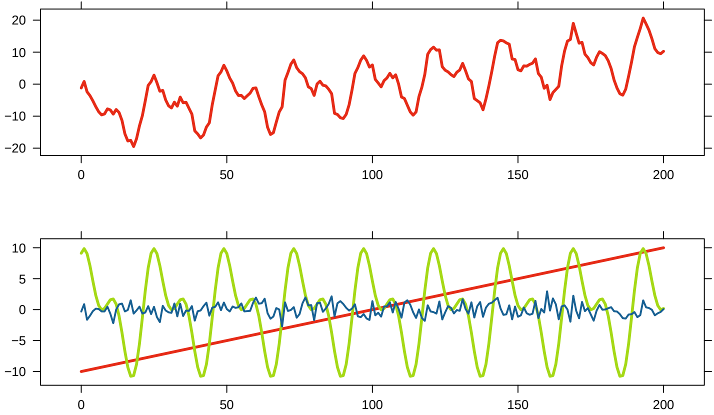
$$\mathsf{MSE}^{(j)} = \frac{1}{N}\sum_{k=0}^{N}\left(f_k^{(\mathrm{tr})} - \widehat{f}_k^{(j)}\right)^2, \quad j = 1,\ldots,1000;$$

$$\mathsf{RMSE} = \sqrt{\frac{1}{1000}\sum_{j=1}^{1000}\mathsf{MSE}^{(j)}}.$$

Ряд с синусами, имеющими общий период

$$f_i = 0.1i - 10 + 7\sin(2\pi i/24 + \pi/5) + 5\sin(2\pi i/12) + \varepsilon_i,$$

$$\varepsilon_i \in \mathrm{N}(0,1), \quad i = 0, \ldots, 200.$$



Временной ряд (сверху) и его составляющие по отдельности (снизу).

Рассматриваемая модель:

$$f_i = 0.1i - 10 + 7\sin(2\pi i/T_1 + \alpha_1) + 5\sin(2\pi i/T_2 + \alpha_2) + \varepsilon_i,$$

$$\varepsilon_i \in \mathrm{N}(0,1), \quad \alpha \in \mathrm{U}(0,\pi/2), \quad i = 0,\ldots,200.$$

Период T_1 — случайный от 16 до $N/2$, кратный четырем, $T_2 = T_1/2$.

Значение ошибки RMSE на основе 1000 реализаций модели:

Метод	Оценка тренда	Прогноз 1	Прогноз 100
OLS	0.831	1.162	2.237
cut+OLS	0.136	0.318	0.612
SSAwDC+OLS	0.389	0.537	1.015
cut+SSAwDC+OLS	**0.12**	0.177	0.353
SSAwDC	0.697		
vforc+OLS		0.869	2.105
rforc+OLS		0.576	1.61
cut+vforc+OLS		**0.17**	**0.335**
cut+rforc+OLS		0.177	0.346

Рассматриваемая модель с периодикой:

$$f_i = 0.1i - 10 + 7\sin(2\pi i/T_1 + \alpha_1) + 5\sin(2\pi i/T_2 + \alpha_2) + \varepsilon_i,$$

$$\varepsilon_i \in \mathrm{N}(0,1), \quad \alpha \in \mathrm{U}(0,\pi/2), \quad i = 0,\ldots,200.$$

Период T_1 — случайный от 16 до $N/2$, кратный четырем, $T_2 = T_1/2$.

Без периодики (тренд+шум):

$$f_i = 0.1i - 10 + \varepsilon_i, \quad \varepsilon_i \in \mathrm{N}(0,1), \quad i = 0,\ldots,200.$$

Значение RMSE оценки тренда на основе 1000 реализаций моделей:

Метод	С периодикой	Без периодики
OLS	0.831	**0.097**
cut+OLS	0.136	0.116
SSAwDC+OLS	0.389	0.114
cut+SSAwDC+OLS	**0.12**	0.12
SSAwDC	0.697	0.118

Заключение

Результаты

- Проведено теоретическое исследование свойств методов МНК и SSA с двойным центрированием в модели с неслучайной периодической ошибкой.
- На основе полученных результатов разработаны модификации алгоритмов для уменьшения ошибки оценки тренда.
- Проведено численное сравнение предложенных алгоритмов на языке программирования R с использованием пакета Rssa.
- Показано, что использование SSA как препроцессинга для МНК (SSAwDC+OLS) уменьшает ошибку в модели с периодикой.
- Показано, что при сильно выраженной периодической компоненте предложенные модификации с обрезанием ряда значительно уменьшают ошибку.

Открытые вопросы

- Оценка фундаментального периода периодической компоненты в общем случае.
- Исследование поведения ошибок при изменении соотношения дисперсии шума и амплитуды периодической компоненты.