Комбинация линейной регрессии и анализа сингулярного спектра для выделения линейного тренда при наличии периодической компоненты

Сазыкин Д.С., студент кафедры статистического моделирования СПбГУ, sazykin.ds@gmail.com

Голяндина Н.Э., к.ф.-м.н., доцент кафедры статистического моделирования СПбГУ, n.golyandina@spbu.ru

Аннотация

Рассматривается задача выделения линейного тренда при наличии в ряде выраженной периодической компоненты с неизвестным периодом. Для решения этой задачи используется сочетание параметрического метода линейной регрессии (она же метод наименьших квадратов) и непараметрического метода анализа сингулярного спектра с центрированием. Предлагаются несколько комбинаций методов, которые увеличивают точность оценивания тренда как по сравнению с линейной регрессий, так и по сравнению с анализом сингулярного спектра. Приведено численное сравнение предлагаемых комбинаций.

Введение

Задача выделения линейного тренда в задачах анализа и прогноза временных рядов широко известна с давних времен. Самым стандартным методом выделения линейного тренда является метод наименьших квадратов (МНК, или OLS, ordinary least squares). Его результатом является так называемая линейная регрессия. Известно, что этот метод дает наилучшую оценку линейного тренда, если временной ряд состоит из линейного тренда и белого шума. Для не белого шума существуют версии взвешенного метода наименьших квадратов. Однако, если во временном ряде присутствует некоторая периодичность, метод МНК уже не является оптимальным.

Если модель ряда известна, то возможно применение метода МНК для оценивания всех параметров, включая параметры линейного тренда. Однако, если период и форма периодической компоненты неизвестны, оценка параметров ряда по МНК может оказаться нереализуемой

на практике. Так как необходимо только выделить тренд, оценивать периоды всех гармоник, входящих в периодическую компоненту, необязательно.

Поэтому рассмотрим метод анализа сингулярного спектра (АСС, или SSA, singular spectrum analysis) [1]. Этот метод способен выделять тренды и периодики без предположения об исходной модели ряда, в частности, для него не надо задавать значение периода. У метода SSA есть параметр, называемый длина окна, а также для выделения линейного тренда разработана специальная модификация, SSA с двойным центрированием.

Будем рассматривать следующую модель ряда $F = (f_0, \ldots, f_{N-1})$, $f_i \in \mathbb{R}$: $F = F^{(tr)} + F^{(s)} + F^{(n)}$, где

$$\begin{aligned} f_i^{(tr)} &= f^{(tr)}(x_i) = ax_i + b \ - \text{ тренд} \\ f_i^{(s)} &= f^{(s)}(x_i) = \sum_{j=1}^{J} C_j \sin(2\pi\omega_j x_i + \varphi_j) \ - \text{ неслучайная ошибка} \\ f_i^{(n)} &= \varepsilon_i \ - \text{гауссовский белый шум} \end{aligned} \quad (1)$$

в равноотстоящих узлах $x_i = i$.

Задачей данной работы является анализ свойств методов линейной регрессии и SSA с двойным центрированием, исследование зависимости ошибки оценивания линейного тренда от параметров ряда, а для SSA и от длины окна, и разработка с учетом полученных результатов комбинированных методов с улучшенной точностью.

Сравнение методов будет произведено с помощью численных экспериментов на основе средне-квадратического отклонения (СКО, или MSE, mean squared error)

$$\sum_{i=0}^{N-1} \left(f_i^{(tr)} - \tilde{f}_i^{(tr)} \right)^2 / N, \quad (2)$$

где $\tilde{f}_i^{(tr)}$ — оценка тренда в точке x_i.

Базовые методы

Линейная регрессия

Постановка задачи МНК, с помощью которой находится линейная регрессия, следующая. Пусть $F = \{f_i\}_{i=0}^{N-1}$ — некоторые измерения в

точках x_i. Тогда оценки $\hat{a}$ и $\hat{b}$ параметров линейного тренда находятся по МНК как аргумент минимума

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

Оценкой тренда является ряд $\widetilde{F}^{(tr)} = \{\tilde{f}_i^{(tr)}\}_{i=0}^{N-1}$, где $\tilde{f}_i^{(tr)} = \hat{a}x_i + \hat{b}$. Метод хорошо известен и мы не будем приводить его решение.

Далее в этом разделе будем рассматривать дискретную модель (1) без шума, причем 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. Помимо дискретной модели будем рассматривать её непрерывный аналог (3), положив $N_{per}^{(j)} = \omega_j N$, $A = aN$, $B = B$:

$$f(t) = At + B + \sum_{j=1}^{J} C_j \sin\Big(2\pi N_{per}^{(j)} t + \varphi_j\Big), \quad t \in [0,1]. \tag{3}$$

Здесь N_{per} обозначает число периодов, проходимых синусом на данном промежутке. В этом случае оценки по МНК $\hat{A}$ и $\hat{B}$ находятся как аргумент минимума

$$\min_{A',B'\in\mathbb{R}} \int_0^1 \big(f(t) - (A't + B')\big)^2 dt.$$

Перечислим свойства метода, которые либо хорошо известны, либо их несложно доказать

- Ошибка МНК оценки линейного тренда в каждой точке не зависит от параметров тренда.
- Для непрерывной модели (3) с $J = 1$, если N_{per} — целое, то ошибка МНК-оценки имеет вид $\dfrac{4C^2\cos^2(\varphi)}{\pi^2 N_{per}^2}$. Отсюда следует, что в этих условиях в непрерывном модели наименьшую (нулевую) ошибку МНК дает в случае фазы $\varphi = \pi/2$, а наибольшую — при $\varphi = 0$.
- Для непрерывной модели (3) с целым числом периодов N_{per} всегда существует такой сдвиг δ, что для ряда $f(t+\delta)$ оценка МНК имеет нулевую ошибку.
- В модели (1) ошибка оценки тренда стремится к нулю при $N \to \infty$.

SSA с двойным центрированием

Пусть имеется ряд $F = (f_0, \ldots, f_{N-1})$. Параметром метода является длина окна L, целое число $1 < L < N$. Метод SSA с двойным центрированием (SSAwDC) описан в [1] и является частным случаем SSA с проекциями, предложенном в [2]. Алгоритм оценки линейного тренда состоит из трех шагов:

1. Вложение, результатом которого является траекторная матрица $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} f_0 & f_1 & \ldots & f_{K-1} \\ f_1 & f_2 & \ldots & f_K \\ \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & \ldots & f_{N-1} \end{bmatrix}$$

2. Вычисление матрицы двойного центрирования

$$\mathbf{C}(\mathbf{X}) = \mathbf{A}_1(\mathbf{X}) + \mathbf{A}_2(\mathbf{X} - \mathbf{A}_1(\mathbf{X})),$$

   где столбцы матрицы $\mathbf{A}_1(\mathbf{Y})$ одинаковые и состоят из средних по строкам матрицы $\mathbf{Y}$; аналогично, строки матрицы $\mathbf{A}_2(\mathbf{Y})$ одинаковые и состоят из средних по столбцам матрицы $\mathbf{Y}$.

3. Диагональное усреднение, состоящее в ганкелизации (усреднении по побочным диагоналям) матрицы $\mathbf{C}(\mathbf{X})$ и получении временного ряда $\widetilde{F}^{(tr)}$ операцией, обратной к вложению.

Перечислим свойства метода:

- Ошибка SSAwDC-оценки линейного тренда в каждой точке не зависит от параметров тренда.
- Результатом SSA с двойным центрированием является, вообще говоря, не линейная функция.
- В модели (1) без шума, если $L\omega$ и $(N+1)\omega$ — целые, т.е. L и $N+1$ кратны фундаментальному периоду синуса, то ошибка SSAwDC-оценки линейного тренда равна 0.
- В модели (1), при $N \to \infty$ и $\min(L, N - L + 1) \to \infty$, ошибка оценки тренда стремится к нулю.

Из свойств метода следуют рекомендации по выбору длины окна: если период неизвестен, то длина окна выбирается близкой к половине длины ряда. Если период известен, то длина окна выбирается кратной периоду и как можно ближе к половине длины ряда.

Комбинированные методы

Итак, у нас есть два метода, ошибки обоих не зависят от коэффициентов линейного тренда, но свойства разные. В частности, метод МНК не имеет параметров, но ошибка зависит от сдвига ряда, в то время как ошибка SSAwDC не зависит от сдвига, но зависит от параметра метода L.

Предложим три комбинации методов.

SSAwDC+OLS

В этой модификации метода SSAwDC выступает как препроцессинг для МНК, т.е. сначала к ряду применяется SSAwDC, а потом к его результату применяется МНК.

cut+SSAwDC+OLS

Данная модификация оценивает период и подает на вход SSAwDC такую часть исходного ряда, чтобы был возможен рекомендуемый способ выбора длины окна.

- Найдем с помощью SSAwDC+OLS метода предварительную оценку линейного тренда $\widehat{F}^{(tr)}$ и рассмотрим остаток $F^{(1)}$ с $f_i^{(1)} = f_i - \widehat{f}_i^{(tr)}$.
- Для ряда $F^{(1)}$ с помощью Basic SSA [1] выделим периодику $\widehat{F}^{(s)}$ и оценим ее фундаментальный период $\widehat{T}$. В наших предположениях, это будет округление до целого максимального из периодов, найденных в ряде.
- Применим SSAwDC+OLS к последним R членов ряда F, таким, что $R+1$ кратно $\widehat{T}$. В качестве длины окна L возьмем ближайшее к $R/2$ и кратное $\widehat{T}$.

cut+OLS

Для того чтобы уменьшить ошибку МНК, нужно подать на вход МНК такую часть ряда, чтобы число периодов было целым, а также чтобы сдвиг ряда (в случае ряда из одного синуса сдвиг ряда определяется фазой синуса) соответствовал минимальной ошибке.

- Так же, как в алгоритме метода cut+SSAwDC+OLS, построим сначала предварительную оценку тренда $\widehat{F}^{(tr)}$, затем выделим периодику $\widehat{F}^{(s)}$ и оценим ее фундаментальный период $\widehat{T}$.
- Рассмотрим отрезки ряда $\widehat{F}^{(s)}$ длины R такие, что R было бы кратно общему периоду $\widehat{T}$ и таких отрезков было не меньше $\widehat{T}$ штук (для просмотра всех возможных сдвигов ряда).
- Находим тот отрезок (т.е. нужный сдвиг), при котором восстановление нулевого тренда по МНК дает минимальную ошибку.
- Берем соответствующий отрезок исходного ряда F и по нему строим окончательную оценку линейного тренда по методу МНК.

Численное сравнение

Рассмотрим ряд длины $N = 201$ вида

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

где $i = 0, \ldots, 200$, $\varepsilon_i \in \mathrm{N}(0, 1)$ и независимы, α_1 и α_2 равномерно распределены на $[0, \pi/2]$, период T_1 случайно выбирается из чисел из интервала от 16 до $N/2$, кратных четырем, $T_2 = T_1/2$.

Полученные результаты $\mathrm{MSE}^{(total)}$ на основе 1000 реализаций (моделирование было выполнено с помощью программы, написанной на R, с использованием пакета Rssa [3]):

OLS: 0.690
SSAwDC: 0.485
SSAwDC+OLS: 0.151
cut+SSAwDC+OLS: 0.014
cut+OLS: 0.018

Как видно, обе модификации значимо уменьшают ошибку восстановления линейного тренда. Также, менее затратный SSAwDC+OLS оказывается существенно лучше по сравнению с базовыми методами.

Метод cut+SSAwDC+OLS в данном случае оказался значимо лучшим.

Заключение

В работе были рассмотрены методы линейной регрессии, SSA с двойным центрированием и предложены их комбинации. Оказалось, что если использовать сильные стороны каждого из методов, то можно получить существенное увеличение точности оценивания линейного тренда при наличии сильно-выраженной неслучайной периодической помехи.

Список литературы

[1] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky. *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman&Hall/CRC, 2001.

[2] N. Golyandina and A. Shlemov. Semi-nonparametric singular spectrum analysis with projection. *Stat Interface*, 10(1):47–57, 2017.

[3] A. Korobeynikov, A. Shlemov, K. Usevich, and N. Golyandina. Rssa*: A collection of methods for singular spectrum analysis*, 2016. R package version 0.14.