Выбор параметров при автоматическом выделении трендовых и периодических составляющих временного ряда в рамках подхода “Гусеница”-SSA

Ф.И.Александров, Н.Э.Голяндина

theo@pdmi.ras.ru, nina@ng1174.spb.edu

С.-Петербургский Государственный Университет

Аппроксимация сигнала

$F_N = (f_0, \ldots, f_{N-1}) : \quad f_n = s_n + \varepsilon_n,$

$S_N = (s_0, \ldots, s_{N-1})$ – детерминированный сигнал,
$(\varepsilon_0, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{N-1})$ – остаток (шум).

Аппроксимация сигнала – в среднеквадратичном.

Хотим аппроксимировать сигналы:

- нестационарные,
- не зная их параметрической модели,
- более того, не зная их структуры.

Метод “Гусеница”-SSA

Метод решает задачи:

- нахождение тренда различной степени детализации,
- сглаживание,
- выделение сезонности,
- выделение периодичностей с меняющимися амплитудами,
- прогноз,
- обнаружение точек разладки.

История:

- США, Великобритания – SSA (Singular Spectrum Analysis),
- Россия – “Гусеница”-SSA.

Преимущества:

- не требует знания о параметрической модели ряда,
- работает с широким спектром реальных временных рядов,
- подходит для нестационарных временных рядов,
- обрабатывает такие естественные составляющие, как модулированные гармоники.

“Гусеница”-SSA: базовый алгоритм

- Разложение ряда на составляющие: $F_N = F_N^{(1)} + \ldots + F_N^{(m)}$.
- Предоставляет информацию о каждой составляющей.

Алгоритм:

1. Построение траекторной матрицы: $F_N \to \mathbf{X} \in \mathbb{R}^{L \times K}$ (L – длина окна, параметр)

$$\mathbf{X} = \begin{bmatrix} f_0 & f_1 & \cdots & f_{N-L} \\ f_1 & f_2 & \cdots & f_{N-L+1} \\ \vdots & \ddots & \ddots & \vdots \\ f_{L-1} & f_L & \cdots & f_{N-1} \end{bmatrix}.$$

2. Сингулярное разложение (SVD): $\mathbf{X} = \sum \mathbf{X}_j$,

$\mathbf{X}_j = \sqrt{\lambda_j} U_j V_j^{\mathrm{T}}$,
λ_j – с.ч. $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$, U_j – с. в-р $\mathbf{S}$,
V_j – с. в-р $\mathbf{S}^{\mathrm{T}}$, $V_j = \mathbf{X}^{\mathrm{T}} U_j \sqrt{\lambda_j}$.

3. Группировка компонент SVD: $\{1, \ldots, d\} = \bigoplus I_k$,

$\mathbf{X}^{(k)} = \sum_{j \in I_k} \mathbf{X}_j$.

4. Восстановление диагональным усреднением: $\mathbf{X}^{(k)} \to \widetilde{F_N}^{(k)}$.

Группировка

Общий случай: $F_N = F_N^{(1)} + F_N^{(2)}$ $\qquad$ $I_1 : \mathbf{X}^{(1)} \leftrightarrow \widetilde{F_N}^{(1)}$.

Группировка возможна, если:

1. $F_N^{(1)}$ – ряд конечного ранга (конечное кол-во компонент),
2. $F_N^{(1)}$ отделим от остатка.

Случай аппроксимации:

$F_N = F_N^{(1)} + F_N^{(2)}$ $\qquad$ $I_1 : \mathbf{X}^{(1)} \leftrightarrow \widetilde{F_N}^{(1)}$ – аппроксимация сигнала.

сигнал, шум

1. Любая линейная комбинация произведений **экспонент**, **э.-м. гармоник** и **полиномов** является рядом конечного ранга.

2. Примеры асимптотической отделимости:
   - Детерм. сигнал асимптотически отделим от белого шума.
   - Периодика асимптотически отделима от тренда.

Идентификация

Идентификация – выбор компонент при группировке.

Экспоненциальный тренд: $f_n = Ae^{\alpha n}$.

- ранг равен 1, т. е. ему соответствует одна компонента SVD,
- собственный вектор:
  $U = (u_1, \ldots, u_L)^{\mathrm{T}}: \quad u_k = Ce^{\alpha k}$.
  ("экспоненциальный" вид с тем же α)

Экспоненциально-модулированная гармоника: $f_n = Ae^{\alpha n}\cos(2\pi\omega n)$.

- ранг равен 2, т. е. ей соответствуют две компоненты SVD,
- собственные вектора:

$U_1 = (u_1^{(1)}, \ldots, u_L^{(1)})^{\mathrm{T}}: \quad u_k^{(1)} = C_1 e^{\alpha k}\cos(2\pi\omega k)$.
$U_2 = (u_1^{(2)}, \ldots, u_L^{(2)})^{\mathrm{T}}: \quad u_k^{(2)} = C_2 e^{\alpha k}\sin(2\pi\omega k)$.

("экспоненциально-модулированный" вид с теми же α и ω)

Идентификация

Идентификация – выбор компонент при группировке.

Экспоненциальный тренд: $f_n = Ae^{\alpha n}$.

- ранг равен 1, т. е. ему соответствует одна компонента SVD,
- собственный вектор:
  $U = (u_1, \ldots, u_L)^{\mathrm{T}}$: $u_k = Ce^{\alpha k}$.
  ("экспоненциальный" вид с тем же α)

Экспоненциально-модулированная гармоника: $f_n = Ae^{\alpha n}\cos(2\pi\omega n)$.

- ранг равен 2, т. е. ей соответствуют две компоненты SVD,
- собственные вектора:

$U_1 = (u_1^{(1)}, \ldots, u_L^{(1)})^{\mathrm{T}}$: $u_k^{(1)} = C_1 e^{\alpha k}\cos(2\pi\omega k)$.

$U_2 = (u_1^{(2)}, \ldots, u_L^{(2)})^{\mathrm{T}}$: $u_k^{(2)} = C_2 e^{\alpha k}\sin(2\pi\omega k)$.

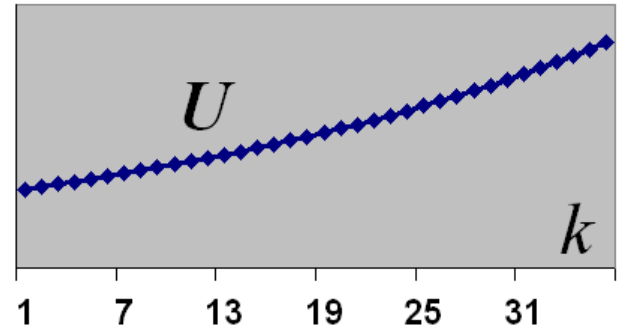
("экспоненциально-модулированный" вид с теми же α и ω)

Идентификация

Идентификация – выбор компонент при группировке.

Экспоненциальный тренд: $f_n = Ae^{\alpha n}$.

- ранг равен 1, т. е. ему соответствует одна компонента SVD,
- собственный вектор:
  $U = (u_1, \ldots, u_L)^{\mathrm{T}}: \quad u_k = Ce^{\alpha k}$.
  ("экспоненциальный" вид с тем же α)



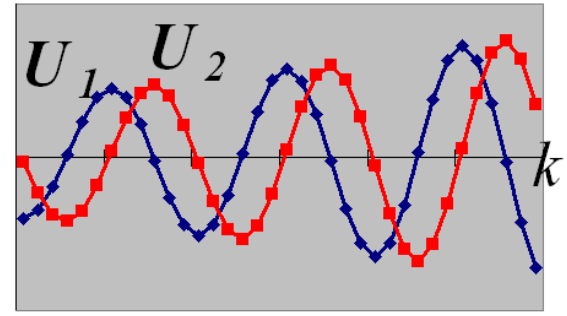
Экспоненциально-модулированная гармоника: $f_n = Ae^{\alpha n}\cos(2\pi\omega n)$.

- ранг равен 2, т. е. ей соответствуют две компоненты SVD,
- собственные вектора:

$U_1 = (u_1^{(1)}, \ldots, u_L^{(1)})^{\mathrm{T}}: \quad u_k^{(1)} = C_1 e^{\alpha k}\cos(2\pi\omega k)$.

$U_2 = (u_1^{(2)}, \ldots, u_L^{(2)})^{\mathrm{T}}: \quad u_k^{(2)} = C_2 e^{\alpha k}\sin(2\pi\omega k)$.

("экспоненциально-модулированный" вид с теми же α и ω)



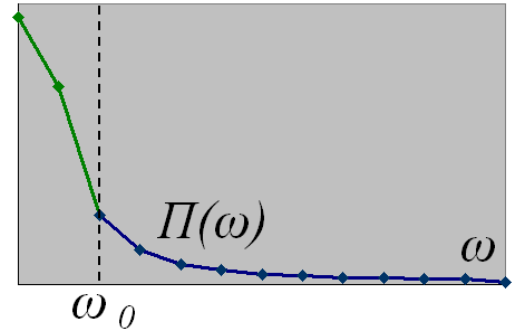
Тренд: метод низких частот

Исследуем каждый собственный вектор U_j. Опишем для $U = (u_1, \ldots, u_L)^{\mathrm{T}}$.

МЕТОД НИЗКИХ ЧАСТОТ

- $u_n = c_0 + \sum\limits_{1 \leqslant k \leqslant \frac{L-1}{2}} \left(c_k \cos(2\pi nk/L) + s_k \sin(2\pi nk/L)\right) + (-1)^n c_{L/2},$
- Периодограмма:

$$\Pi_U^L(k/L) = \frac{L}{4} \begin{cases} 2{c_0}^2, & k = 0, \\ {c_k}^2 + {s_k}^2, & 1 \leqslant k \leqslant \frac{L-1}{2}, \\ 2{c_{L/2}}^2, & L \text{ – чётное и } k = L/2. \end{cases}$$



$\Pi_U^L(\omega)$, $\omega \in \{k/L\}$, **отражает вклад в вид U гармоники с частотой ω.**

- Параметр: ω_0 – верхняя граница области низких частот

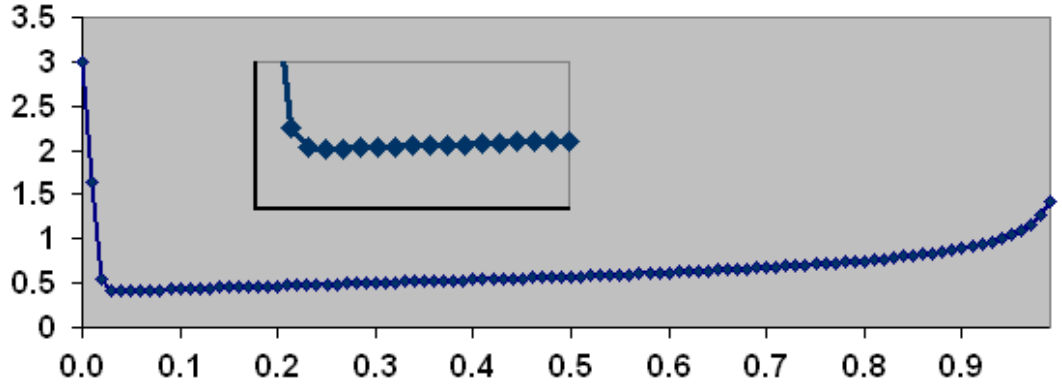
$$\mathcal{C}(U) = \frac{\sum_{\mathbf{L\omega_0 < k \leqslant L/2}} \mathbf{\Pi_U^L(k/L)}}{\sum_{\mathbf{0 \leqslant k \leqslant L/2}} \mathbf{\Pi_U^L(k/L)}}$$ – вклад "не-низких" частот.

$\boxed{\mathcal{C}(U) \leqslant \mathcal{C}_0 \Rightarrow \textbf{с. вектор } U \textbf{ соответствует тренду.}}$

($\mathcal{C}_0 \in (0,1)$ – пороговое значение)

Метод НЧ: оптимальные пороговые значения

Зависимость СКО от $\mathcal{C}_0$ (от аппроксимации, полученной с таким $\mathcal{C}_0$)



Оптимальные $\mathcal{C}_0$

	0	0.01	0.02
0	0	0.004	0.01
1	0.003	0.01	0.02
2	0.01	0.02	0.04

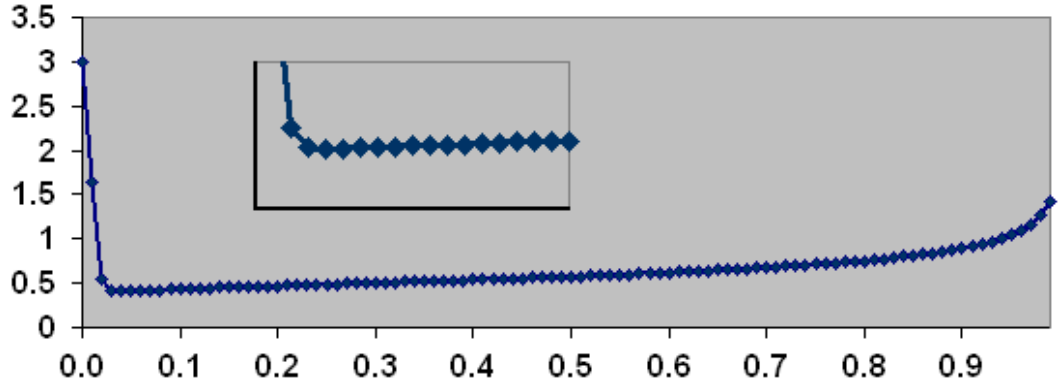
Соотв. им значения СКО

	0	0.01	0.02
0	0	0	0
1	0.18	0.19	0.20
2	0.36	0.37	0.40

Количество компонент, идентифицируемых при оптимальном $\mathcal{C}_0$, очень близко к рангу ряда (1 для экспоненты).

Метод НЧ: оптимальные пороговые значения

Зависимость СКО от $\mathcal{C}_0$ (от аппроксимации, полученной с таким $\mathcal{C}_0$)



Оптимальные $\mathcal{C}_0$

	0	0.01	0.02
0	0	0.004	0.01
1	0.003	0.01	0.02
2	0.01	0.02	0.04

Соотв. им значения СКО

	0	0.01	0.02
0	0	0	0
1	0.18	0.19	0.20
2	0.36	0.37	0.40

Количество компонент, идентифицируемых при оптимальном $\mathcal{C}_0$, очень близко к рангу ряда (1 для экспоненты).

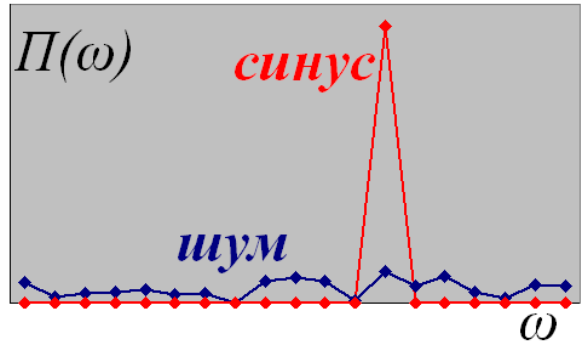
Периодика: метод Фурье

Исследуем последовательности элементов собственных векторов соседних компонент U_j,U_{j+1}

МЕТОД ФУРЬЕ

- **Часть 1.** Проверка на "максимальную" частоту: $\theta_j = \arg\min_k \Pi^M_{U_j}(k/M)$,
  $M|\theta_j - \theta_{j+1}| \leqslant s_0 \Rightarrow$ пара $(j, j+1)$ – "гармоническая".

- **Часть 2.** Проверка на форму периодограммы:
  $\rho_{(j,j+1)} = \frac{1}{2}\max_k \left(\Pi^M_{U_j}(k/M) + \Pi^M_{U_{j+1}}(k/M)\right)$, для гарм. пары $\rho_{(j,j+1)} = 1$.

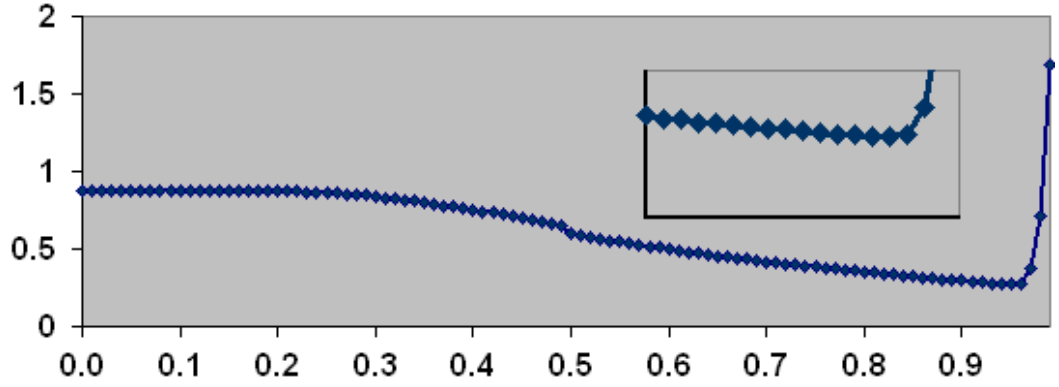


$\rho_{(j,j+1)} \geqslant \rho_0 \Rightarrow$ **пара компонент** $(j, j+1)$ **соответствует гармонике.**

($\rho_0 \in (0,1)$ – пороговое значение)

Метод Фурье: оптимальные пороговые значения

Зависимость СКО от ρ_0 (от аппроксимации, полученной с таким ρ_0)



Оптимальные ρ_0

	0	0.01	0.02
0	1.00	0.99	0.96
1	0.98	0.95	0.89
2	0.93	0.90	0.82

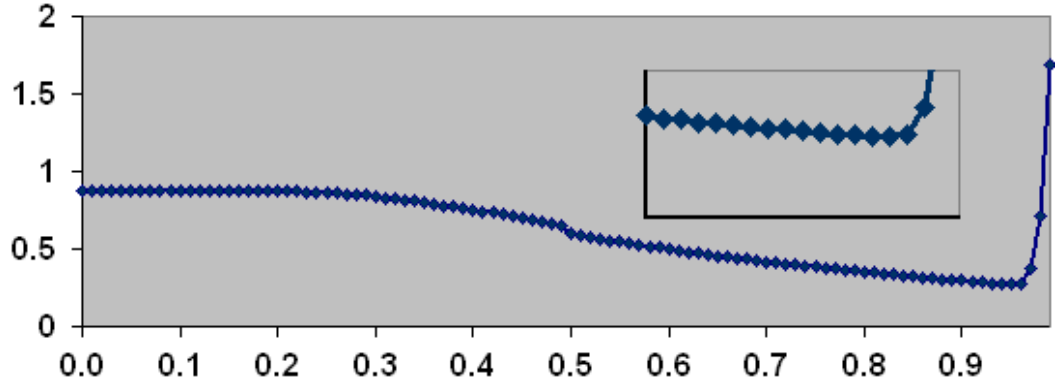
Соотв. им значения СКО

	0	0.01	0.02
0	0	0	0
1	0.27	0.28	0.31
2	0.58	0.62	0.69

Количество компонент, идентифицируемых при оптимальном ρ_0, очень близко к рангу ряда (2 для э.-м. гармоники).

Метод Фурье: оптимальные пороговые значения

Зависимость СКО от ρ_0 (от аппроксимации, полученной с таким ρ_0)



Оптимальные ρ_0

	0	0.01	0.02
0	1.00	0.99	0.96
1	0.98	0.95	0.89
2	0.93	0.90	0.82

Соотв. им значения СКО

	0	0.01	0.02
0	0	0	0
1	0.27	0.28	0.31
2	0.58	0.62	0.69

Количество компонент, идентифицируемых при оптимальном ρ_0, очень близко к рангу ряда (2 для э.-м. гармоники).

Ситуация, близкая к реальной

Расчёт оптимального порогового значения в случае, когда параметры модели известны с погрешностью.

Экспоненциальный тренд ($f_n = Ce^{\alpha n}$)

- $\alpha \in [0.01, 0.02]$,
- $\sigma_0 \in [1, 2]$.

То есть рассматривается класс рядов. Пороговое значение должно таким, чтобы выделять тренд любого ряда из этого класса.

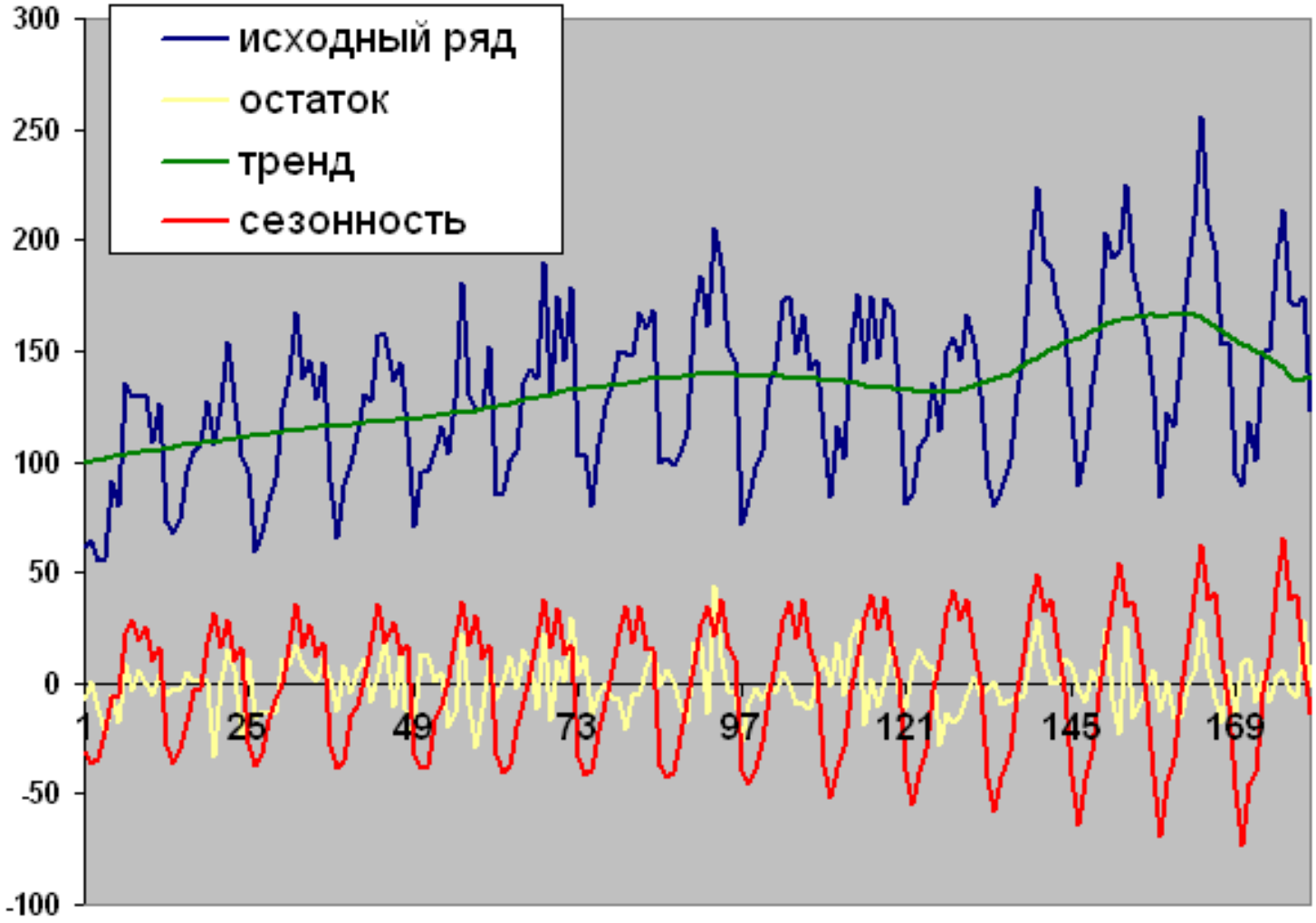
Увеличение $\mathcal{C}_0$ – ослабление ограничений (взятие большего количества компонент).

Наихудший случай в этом классе: ряд с $\alpha = 0.02$ и $\sigma_0 = 2$. Для такого ряда расчитанное оптимальное $\mathcal{C}_0 = 0.039$.

Пусть в реальности было: $\alpha = 0.13, \sigma_0 = 1.6$**.**

СКО при выбранном $\mathcal{C}_0 = 0.039$ больше всего на 0.05%, чем минимальное СКО для такого ряда, которое достигается при $\mathcal{C}_0 = 0.018$.

Заключение



Ежемесячные данные: аварии на дорогах, 1960-1974, Онтарио.

Номера компонент тренда: 1, 4, 5.

Номера компонент сезонности: 2, 3, 6-8, 11-14.