Анализ временных рядов. Лекция 4

SARIMAX

Костромина Алина

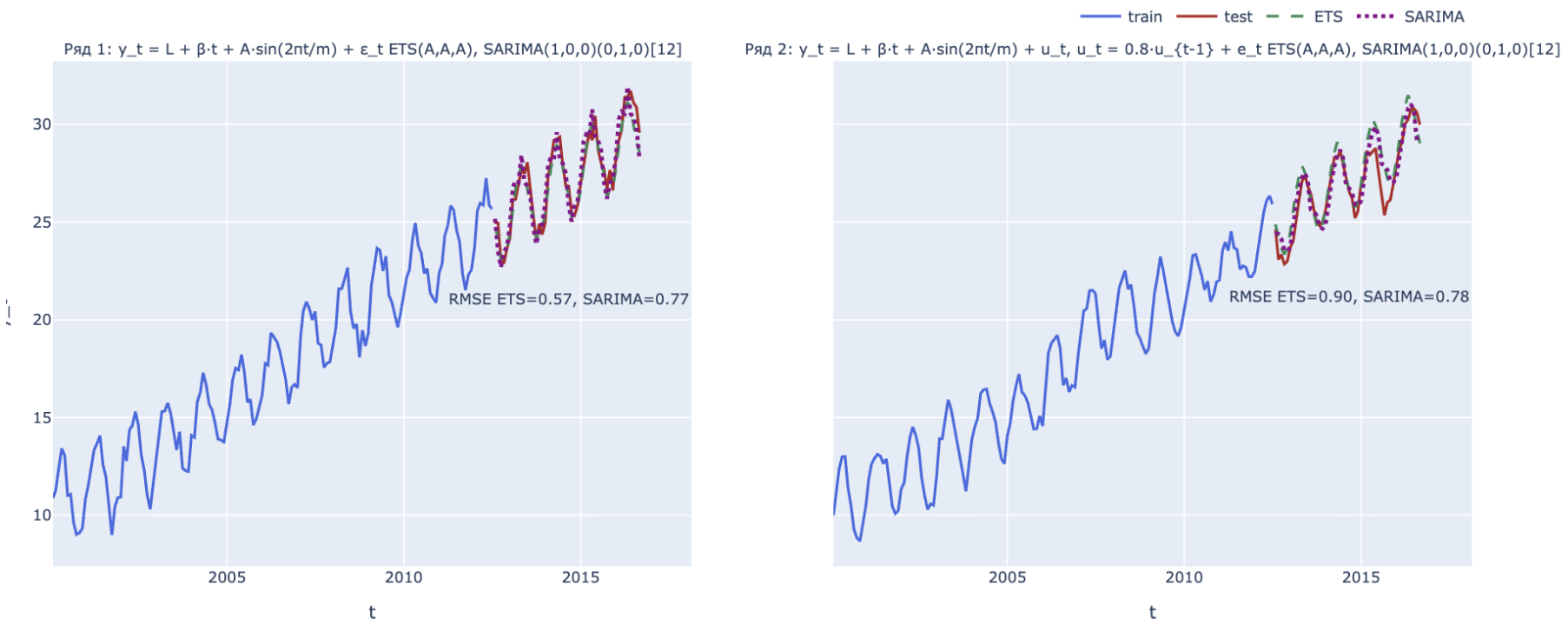
1.12.2025

Что сегодня в программе?

1. Где ETS модели не справляются — зачем нам SARIMA(X)?
2. Предпосылки. Еще раз говорим о стационарности.
3. ARMA модель. ACF / PACF.
4. Добавляем интегрирование — ARIMA.
5. Добавляем сезонное интегрирование — SARIMA.
6. Добавляем внешние признаки — SARIMAX.

Две парадигмы: компоненты vs автокорреляции

ETS vs. SARIMA на практике



ARIMA. Предпосылка — стационарность

AR(p) (Autoregression) - модель авторегрессии порядка *p*.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

MA(q) (Moving average) - модель скользящего среднего порядка *q*:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

Если объединим две модели, то получим **ARMA(p, q)** модель:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

Теоретическая мотивация: любой стационарный в широком смысле процесс можно аппроксимировать с заданной точностью, выбрав необходимые *p* и *q*.

ARIMA(p, d, q) является расширением модели ARMA на нестационарные ряды:

$$\Delta^d y_i = c + \sum_{j=1}^{p} \phi_j \Delta^d y_{i-j} + \varepsilon_i + \sum_{j=1}^{q} \theta_j \varepsilon_{i-j}$$

где *d* - порядок дифференцирования для приведения ряда к стационарному

Почему стационарность важна?

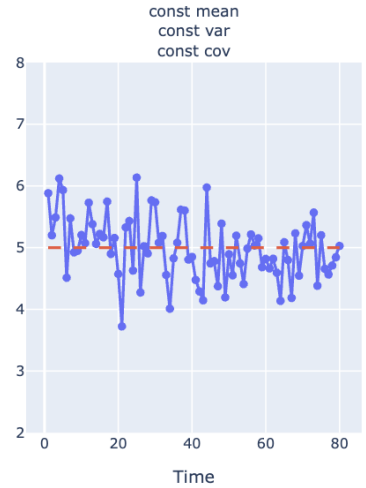
1. Хотим «хорошие» прогнозы и оцениваемые параметры, которые стабильны и адекватны и имеют удобные статистические свойства (можно проверять гипотезы, считать доверительные / прогнозные интервалы и т. д.).
2. Без стационарности не сможем подобрать гиперпараметры модели (p, q) — разберем позже на примерах.
3. Даже в ETS-модели ошибки (инновации) предполагаются стационарными (i.i.d шум) — так что это не новая вещь.

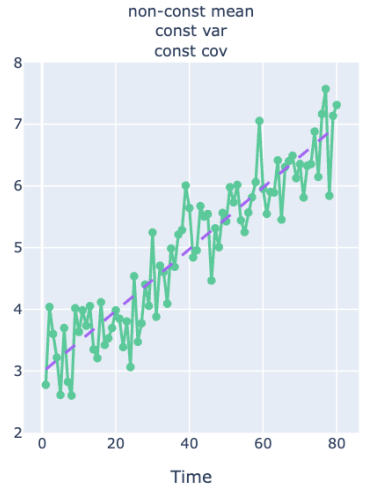
Повторим разговор про стационарность

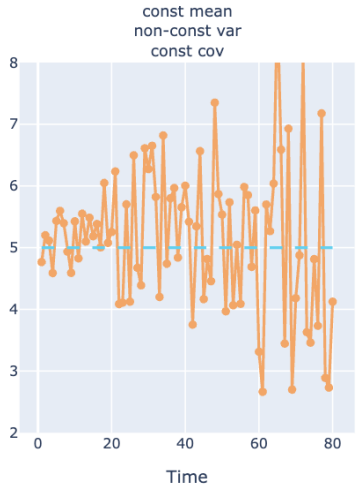
В узком смысле (сильная стационарность) *— временной ряд $y_1 \ldots, y_t$ стационарен, если для любого s совместное распределение $y_1 \ldots, y_{t+s}$ не зависит от t.*

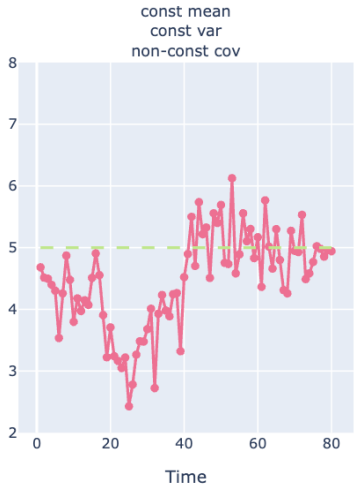
В широком смысле (слабая стационарность) *— временной ряд $y_1 \ldots, y_t$ стационарен, если:*

1. *Существуют конечные матожидание m(t) и дисперсия v(t) в каждой точке $t \in T$.*
2. *m(t) = const. v(t) = const.*
3. *Автоковариационная функция зависит от лага τ и не зависит от t — $Cov(y_t, y_{t+\tau}) = \gamma(\tau)$.*









Парочка упражнений

Упр. 1 — Стационарен ли ряд …?

$y_t = \mu + \varepsilon_t$, где μ — константа, а $\{\varepsilon_t\}$ — белый шум:
$E[\varepsilon_t] = 0, \quad \mathrm{Var}(\varepsilon_t) = \sigma^2, \quad \mathrm{Cov}(\varepsilon_t, \varepsilon_{t+h}) = 0$ при $h \neq 0$.

Упр. 2 — Стационарен ли ряд …?

$y_t = y_{t-1} + c + \varepsilon_t, \quad t \geq 1, \quad y_0 = 0$, где c — константа, а $\{\varepsilon_t\}$ — белый шум:
$E[\varepsilon_t] = 0, \quad \mathrm{Var}(\varepsilon_t) = \sigma^2, \quad \mathrm{Cov}(\varepsilon_t, \varepsilon_{t+h}) = 0$ при $h \neq 0$.

Парочка упражнений

Упр. 1 — Стационарен ли ряд ДА?

$y_t = \mu + \varepsilon_t$, где μ — константа, а $\{\varepsilon_t\}$ — белый шум:
$E[\varepsilon_t] = 0, \quad \mathrm{Var}(\varepsilon_t) = \sigma^2, \quad \mathrm{Cov}(\varepsilon_t, \varepsilon_{t+h}) = 0$ при $h \neq 0$.

Решение

1. $\mathbb{E}[y_t] = \mathbb{E}[\mu + \varepsilon_t] = \mu + \mathbb{E}[\varepsilon_t] = \mu + 0 = \mu = \text{const}$.

2. $\mathrm{Var}(y_t) = \mathrm{Var}(\mu + \varepsilon_t) = \mathrm{Var}(\varepsilon_t) = \sigma^2 = \text{const}$.

3. $\mathrm{Cov}(y_t, y_{t+h}) = \mathrm{Cov}(\mu + \varepsilon_t,\ \mu + \varepsilon_{t+h}) = \mathrm{Cov}(\mu, \mu) + \mathrm{Cov}(\mu, \varepsilon_{t+h}) = \mathrm{Cov}(\varepsilon_t, \varepsilon_{t+h})$

$$\mathrm{Cov}(\varepsilon_t, \varepsilon_{t+h}) = \begin{cases} \sigma^2, & h = 0, \\ 0, & h \neq 0. \end{cases}$$

Зависит только от лага h, а не от времени t.

Вывод.

Ряд $y_t = \mu + \varepsilon_t$ **стационарен в слабом смысле**.

Упр. 2 — Стационарен ли ряд НЕТ?

$y_t = y_{t-1} + c + \varepsilon_t, \quad t \geq 1, \quad y_0 = 0$, где c — константа, а $\{\varepsilon_t\}$ — белый шум:
$E[\varepsilon_t] = 0, \quad \mathrm{Var}(\varepsilon_t) = \sigma^2, \quad \mathrm{Cov}(\varepsilon_t, \varepsilon_{t+h}) = 0$ при $h \neq 0$.

Решение

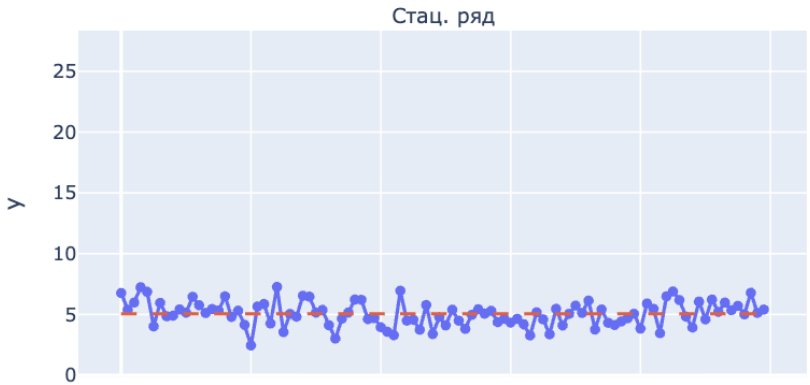
$$y_1 = y_0 + c + \varepsilon_1 = c + \varepsilon_1,$$
$$y_2 = y_1 + c + \varepsilon_2 = 2c + \varepsilon_1 + \varepsilon_2,$$
$$y_3 = y_2 + c + \varepsilon_3 = 3c + \varepsilon_1 + \varepsilon_2 + \varepsilon_3.$$

Отсюда по индукции: $y_t = tc + \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t$.

$$1. E[y_t] = E[tc + \varepsilon_1 + \cdots + \varepsilon_t] = tc + E[\varepsilon_1] + \cdots + E[\varepsilon_t] = tc.$$

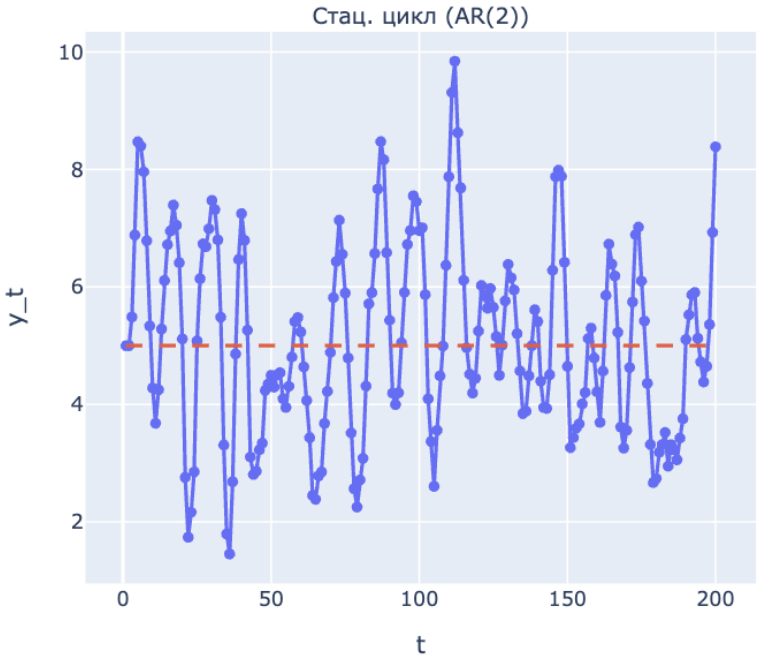
$E[y_t] = tc$ зависит от времени t.

```python
# Стационарный ряд: константа + белый шум
mu = 5.0
sigma = 1.0
eps_stat = np.random.normal(loc=0.0, scale=sigma, size=n)
y_stat = mu + eps_stat
```




```python
# Нестационарный ряд: случайное блуждание с дрейфом
c = 0.2
eps_ns = np.random.normal(loc=0.0, scale=1.0, size=n)
y_ns_raw = np.cumsum(c + eps_ns)
```

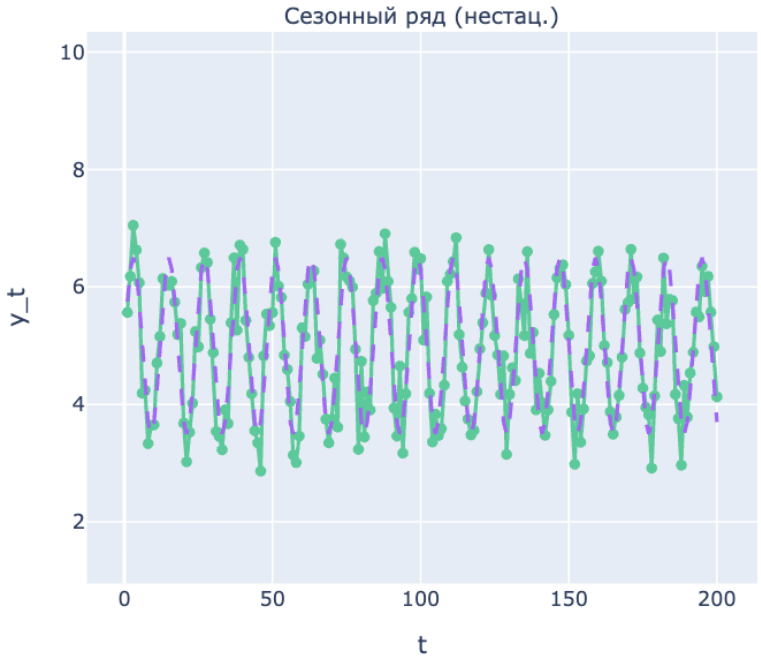
Примечание 1: Цикл vs. сезонность



$$y_t = \mu + u_t,$$
$$u_t = \varphi_1 u_{t-1} + \varphi_2 u_{t-2} + \varepsilon_t,$$

$$\varphi_1 = 2r\cos w, \qquad \varphi_2 = -r^2,$$

$$r < 1, \; w = 2\pi/20$$

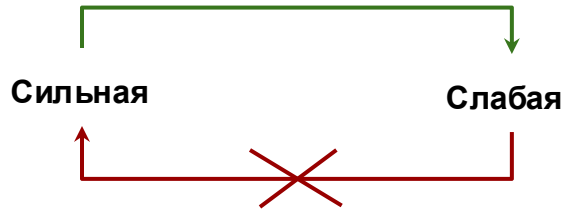


$$y_t = s_t + \varepsilon_t, \qquad s_{t+m} = s_t, \quad t = 1, 2, \ldots$$

Примечание 2: Сильная vs. слабая стационарности

Если существуют конечные $\mathrm{E}(y_t)$ и $Var(y_t)$.

Совместные распределения не меняются



mean/variance/autocov не меняются

https://en.wikipedia.org/wiki/Cauchy_process

Тесты на стационарность

Aspect	KPSS Test	Dickey-Fuller Test
Hypothesis Tested	Null: The series is stationary.	Null: The series is non-stationary.
Alternative Hypothesis	The series is non-stationary.	The series is stationary.
Sensitivity	Detects trend stationarity issues.	Detects unit root (random walk) stationarity.
Approach	Tests for stationarity by modeling residuals.	Checks for unit roots in the series.

ADF — оцениваем регрессию на приросты и смотрим статистику при y_{t-1} (тест на единичный корень).

KPSS — суммируем остатки от регрессии на детерминированные компоненты (константа + тренд) и смотрим, насколько «уносит» частичные суммы от нуля (дисперсия == 0?).

https://www.hse.ru/mirror/pubs/share/359311210.pdf — *подробнее про ADF & KPSS*
http://www.machinelearning.ru/wiki/index.php?title=Критерий_KPSS — *подробнее про KPSS*

Тесты на стационарность

ADF

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta y_{t-i} + \varepsilon_t$$

$H_0: \gamma = 0$ (ряд нестационарен)

KPSS

$$y_t = \alpha + \beta t + r_t + \varepsilon_t$$
$$r_t = r_{t-1} + u_t, \qquad u_t \sim i.i.d(0, \sigma_u^2)$$

$H_0: \sigma_u^2 = 0$ (ряд стационарен)

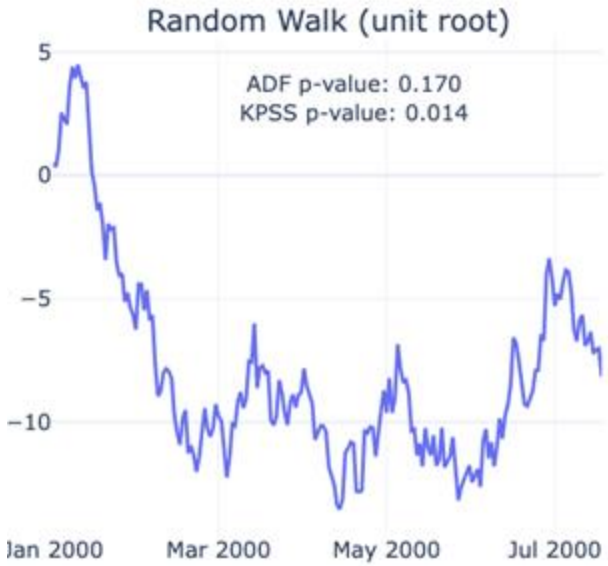
остатки $y_t = \alpha + \beta t$ (т. к. мы под H_0)

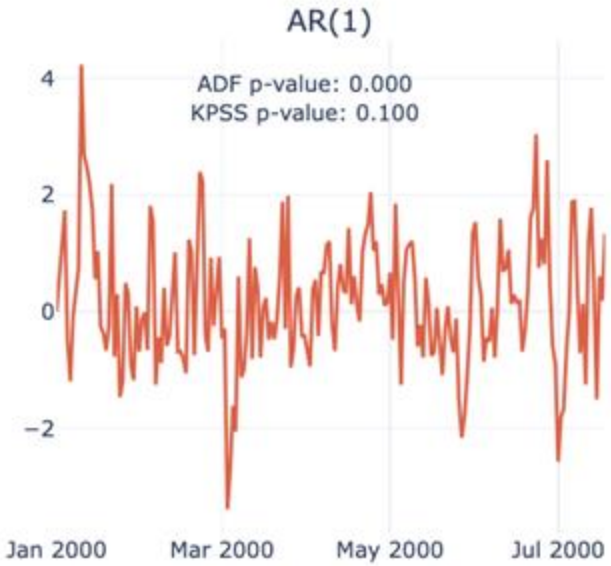
ADF — оцениваем регрессию на приросты и смотрим статистику при y_{t-1} (тест на единичный корень).

KPSS — суммируем остатки от регрессии на детерминированные компоненты (константа + тренд) и смотрим, насколько «уносит» частичные суммы от нуля (дисперсия == 0?).

https://www.hse.ru/mirror/pubs/share/359311210.pdf — подробнее про ADF & KPSS
http://www.machinelearning.ru/wiki/index.php?title=Критерий_KPSS — подробнее про KPSS

Тесты на стационарность





Вопросы?

Определение коэффициентов вручную — ACF, PACF

ACF $\rho(k)$: насколько в среднем похожи значения ряда на расстоянии k шагов.

$$\rho(k) = \mathrm{Corr}(x_t; x_{t+k}) = \frac{Cov(x_t, x_{t+k})}{\sqrt{D(x_t) \cdot D(x_t + k)}} = \frac{\gamma(k)}{\sqrt{\gamma(0) \cdot \gamma(0)}} = \frac{\gamma(k)}{\gamma(0)}, \quad k \in Z$$

$$\hat{\rho}(k) = \frac{\sum_{t=1}^{T-k}(x_t - \bar{x})\,(t_{t+k} - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})}, \quad \bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$$

Определение коэффициентов вручную — ACF, PACF

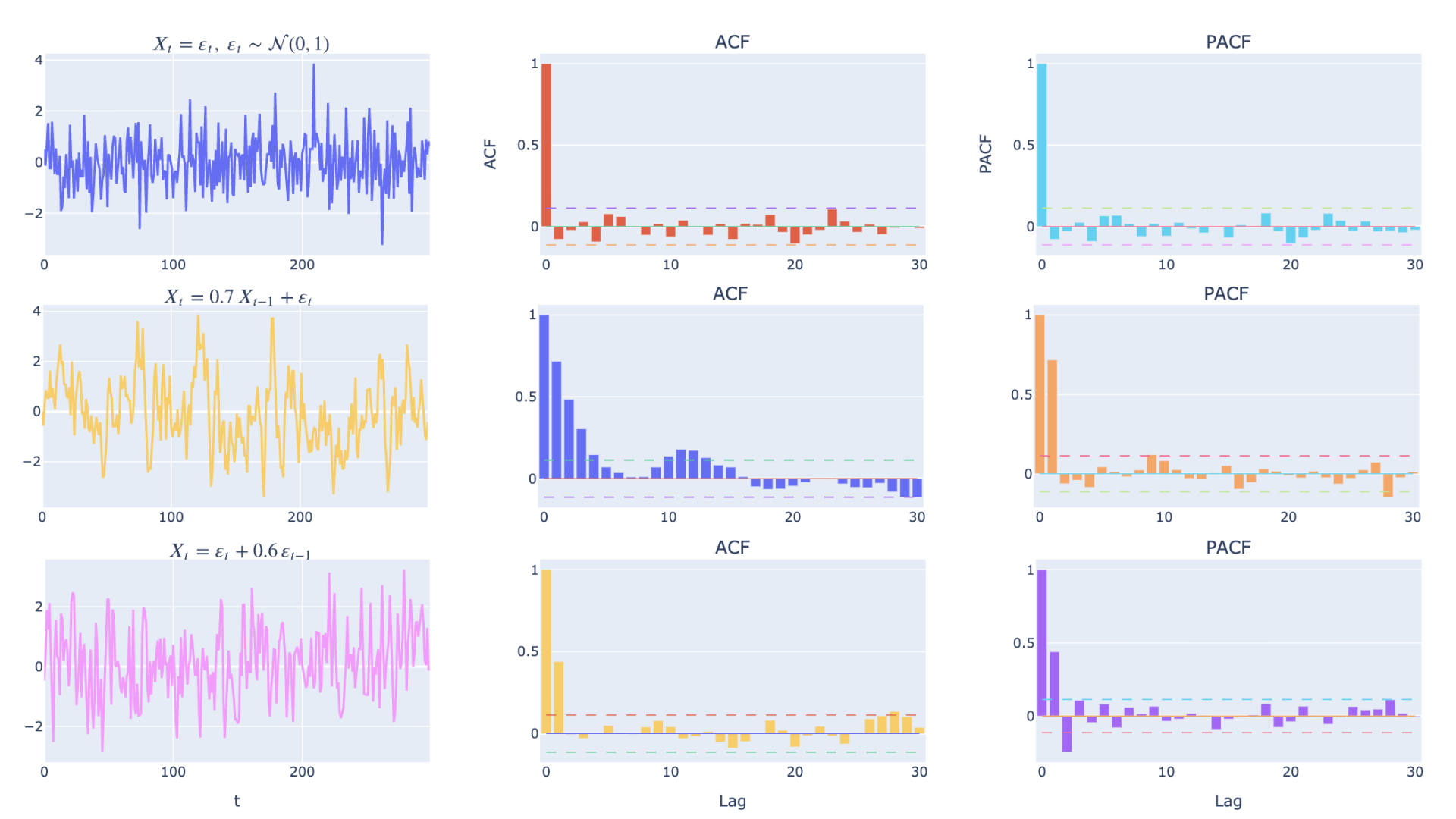
PACF $\rho_{PACF(\mathrm{k})}$: есть ли линейная информация на лаге k, которая не объясняется влиянием промежуточных лагов $x_{t+1}, \ \ldots, x_{t+\mathrm{k}-1}$.

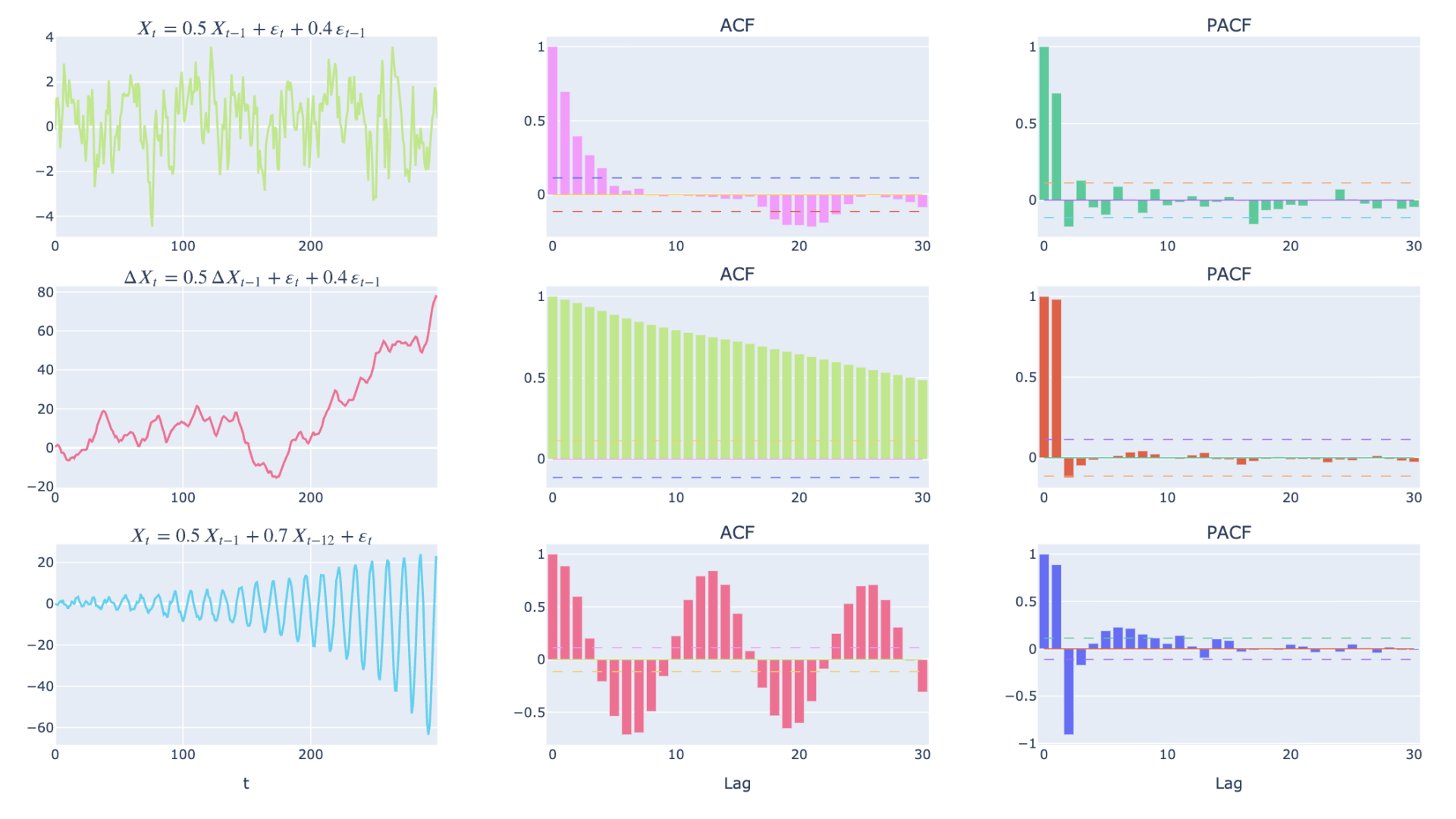
$$\rho_{PACF}(k) = \mathrm{Co}rr(x_t - \hat{\mathrm{x}}_{\mathrm{t}}; x_{t+\mathrm{k}} - \hat{\mathrm{x}}_{t+\mathrm{k}}),$$

$$\widehat{x_t} = \beta_1^{(1)} x_{t+1} + \cdots + \beta_{k-1}^{(1)} x_{t+k-1}, \quad \hat{\mathrm{x}}_{t+\mathrm{k}} = \beta_1^{(2)} x_{t+1} + \cdots + \beta_{k-1}^{(2)} x_{t+k-1}$$

Коэффициенты минимизируют MSE предсказаний x_t и x_{t+k}

Сначала «вычитаем» влияние промежуточных лагов из обеих переменных, а потом смотрим, насколько сильно связаны очищенные остатки.


$X_t = \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1)$
ACF
PACF
$X_t = 0.7\, X_{t-1} + \varepsilon_t$
ACF
PACF
$X_t = \varepsilon_t + 0.6\, \varepsilon_{t-1}$
ACF
PACF
t
Lag
Lag


$X_t = 0.5\,X_{t-1} + \varepsilon_t + 0.4\,\varepsilon_{t-1}$
ACF
PACF
$\Delta X_t = 0.5\,\Delta X_{t-1} + \varepsilon_t + 0.4\,\varepsilon_{t-1}$
ACF
PACF
$X_t = 0.5\,X_{t-1} + 0.7\,X_{t-12} + \varepsilon_t$
ACF
PACF
t
Lag
Lag

ARIMA

Как подбирать гиперпараметры:
https://docs.google.com/document/d/1wVsBkRIZbHdPMQIbUoXdznrmxkSZNEVFLq9D_OzCIdA/edit?usp=sharing

AR(p) (Autoregression) - модель авторегрессии порядка *p*.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

MA(q) (Moving average) - модель скользящего среднего порядка *q*:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

Если объединим две модели, то получим **ARMA(p, q)** модель:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

Теоретическая мотивация: любой стационарный в широком смысле процесс можно аппроксимировать с заданной точностью, выбрав необходимые *p* и *q*.

ARIMA(p, d, q) является расширением модели ARMA на нестационарные ряды:

$$\Delta^d y_i = c + \sum_{j=1}^{p} \phi_j \Delta^d y_{i-j} + \varepsilon_i + \sum_{j=1}^{q} \theta_j \varepsilon_{i-j}$$

где *d* - порядок дифференцирования для приведения ряда к стационарному

Как строится прогноз?

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

Пусть известны значения ряда y_t и возмущения ε_t до момента времени T включительно, а также получены веса $(\varphi_i, \mathrm{i} = \overline{1,p},\ \theta_\mathrm{j}, \mathrm{j} = \overline{1,q})$ модели ARMA(p,q).

Выражение для y_{T+1} в рамках модели:

$$\mathrm{y}_{T+1} = \mathrm{c} + \sum_{\mathrm{i}=1}^{p} \varphi_\mathrm{i} \mathrm{y}_{T+1-\mathrm{i}} + \varepsilon_{T+1} + \sum_{\mathrm{j}=1}^{q} \theta_\mathrm{j} \varepsilon_{T+1-\mathrm{j}}\,.$$

Неизвестно, $E(\varepsilon_{T+1}) = 0$

$\mathrm{Cov}(\varepsilon_{T+1}, y_t) = 0, \mathrm{t} \le \mathrm{T}$

Выражение для y_{T+2} в рамках модели:

$$\mathrm{y}_{T+2} = \mathrm{c} + \sum_{\mathrm{i}=1}^{p} \varphi_\mathrm{i} \mathrm{y}_{T+2-\mathrm{i}} + \varepsilon_{T+2} + \sum_{\mathrm{j}=1}^{q} \theta_\mathrm{j} \varepsilon_{T+2-\mathrm{j}}\,.$$

Неизвестно,

$E(\varepsilon_{T+1}) = 0$

$E(\varepsilon_{T+2}) = 0$

$E(\mathrm{y}_{T+1}) = \hat{\mathrm{y}}_{T+1}$

- Неизвестные в AR части заменяет на прогнозы
- Неизвестные в MA части зануляем

Как подбирать гиперпараметры:
https://docs.google.com/document/d/1wVsBkRIZbHdPMQIbUoXdznrmxkSZNEVFLq9D_OzCldA/edit?usp=sharing

SARIMA

$$y_t = c + \sum_{n=1}^{p} \alpha_n y_{t-n} + \sum_{n=1}^{q} \theta_n \epsilon_{t-n} + \sum_{n=1}^{P} \phi_n y_{t-sn} + \sum_{n=1}^{Q} \eta_n \epsilon_{t-sn} + \epsilon_t$$

SARIMA(p, d, q)(P, D, Q, s) - расширение модели ARIMA для учета сезонности:

- s - период сезонности
- P - порядок сезонной AR части
- D - порядок сезонного дифференцирования ряда
- Q - порядок сезонной MA части

Как подбирать гиперпараметры:
https://docs.google.com/document/d/1wVsBkRIZbHdPMQIbUoXdznrmxkSZNEVFLq9D_OzCldA/edit?usp=sharing

SARIMAX

$$d_t = c + \sum_{n=1}^{p} \alpha_n d_{t-n} + \sum_{n=1}^{q} \theta_n \epsilon_{t-n} + \sum_{n=1}^{r} \beta_n x_{n_t} + \sum_{n=1}^{P} \phi_n d_{t-sn} + \sum_{n=1}^{Q} \eta_n \epsilon_{t-sn} + \epsilon_t$$

SARIMAX(p, d, q)(P, D, Q, s):

- Работает с нестационарными рядами
- Учитывает дополнительные факторы
- Моделирует сезонные паттерны

Примечание: стационарность и инвертируемость

AR(p) (Autoregression) - модель авторегрессии порядка *p*.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

Говорят, что AR-часть стационарна, если все корни уравнения $\Phi(z) = 0$ лежат **вне единичного круга**: $| z |> 1$.

Существует и сходится MA(∞)-представление:

$$y_t = \mu + \sum_{k=0}^{\infty} \psi_k \, \varepsilon_{t-k}, \sum_{k=0}^{\infty} | \psi_k | < \infty.$$

MA(q) (Moving average) - модель скользящего среднего порядка *q*:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

Говорят, что MA-часть инвертируема, если все корни уравнения $\Theta(z) = 0$ лежат **вне единичного круга**: $| z |> 1$.

Существует и сходится AR(∞)-представление:

$$\varepsilon_t = \sum_{k=0}^{\infty} \pi_k \, y_{t-k}.$$

Примечание: стационарность и инвертируемость

Формула суммы бесконечной геометрической прогрессии при |q|<1

$$S = \frac{b_1}{1-q}$$

$$y_t = \varphi y_{t-1} + \varepsilon_t.$$

$$\begin{aligned} y_t &= \varphi y_{t-1} + \varepsilon_t \\ &= \varphi(\varphi y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \varphi^2 y_{t-2} + \varphi\varepsilon_{t-1} + \varepsilon_t \\ &\dots \\ &= \varphi^k y_{t-k} + \sum_{j=0}^{k-1} \varphi^j \varepsilon_{t-j}. \end{aligned}$$

Если $|\varphi| < 1$, то при $k \to \infty$ $\varphi^k y_{t-k} \to 0$,

$$y_t = \sum_{j=0}^{\infty} \varphi^j \varepsilon_{t-j}.$$

Получили MA(∞)-представление для y_t.

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1}.$$

$$y_t = (1 + \theta L)\varepsilon_t.$$

$$\varepsilon_t = (1 + \theta L)^{-1} y_t.$$

$$(1 + \theta L)^{-1} = 1 - \theta L + \theta^2 L^2 - \theta^3 L^3 + \cdots$$

Этот ряд сходится только при $|\theta| < 1$.

$$\varepsilon_t = y_t - \theta y_{t-1} + \theta^2 y_{t-2} - \theta^3 y_{t-3} + \cdots$$

Получили AR(∞)-представление для ε_t.

Вопросы?

Что почитать?

- Р. Хайдман, Дж. Атанасопулос — Прогнозирование: принципы и практика, гл. 9 (стр. 272-327) (рус.).
- https://education.yandex.ru/handbook/ml/article/modeli-vida-arima — учебник ШАД (глава 10.4) (рус.).
- https://mlgu.ru/909 — SARIMA в Python (рус.).

- https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html — ARIMA в Python (англ.).
- https://nixtlaverse.nixtla.io/statsforecast/docs/models/autoarima.html — AutoARIMA в Python (англ.).

Совсем дополнительное чтение:
https://faculty.washington.edu/ezivot/econ584/notes/unitroot.pdf — подробное описание и проблемы тестов на стационарность (англ.).