Анализ временных рядов. Лекция 5

Бонусные темы по SARIMA, диагностике ошибок и многомерным рядам

Кострома Алина

4.12.2025

Что сегодня в программе?

1. Несколько примечаний к SARIMA моделям:
   - Алгоритм Хандакара-Хайндмана
   - Случайные vs детерминированные тренд и сезонность?
   - А что про связь ETS и SARIMA?

2. Диагностика и коррекция ошибок

3. Королева моделей в 2000-м году. И чуток про сами конкурсы ☺

4. Что если у меня несколько временных рядов?

Алгоритм (Hyndman & Khandakar, 2008)

1) Сколько раз нужно брать сезонное дифференцирование?
   - 0-2 раза.
   - В помощь — график исходного ряда, ACF / PACF, STL.

2) Сколько раз нужно брать обычное дифференцирование?
   - 0-2 раза.
   - В помощь — KPSS, ADF.

3) Оцениваем набор SARMA моделей.
   - $P + q <= 5, P + Q <= 5.$
   - В помощь $\mathrm{AIC}, \mathrm{BIC}$, кросс-валидация.

При этом напоминаю нашу расширенную инструкцию:

https://docs.google.com/document/d/1wVsBkRIZbHdPMQIbUoXdznrmxkSZNEVFLq9D_OzCldA/edit?tab=t.0

Случайный или детерминированный?

И тренд, и сезонность могут быть случайными (stochastic) и детерминированными (deterministic).

Тренд

statsmodels.tsa.stattools.kpss

regression : `str` {“c”, “ct”}

The null hypothesis for the KPSS test.

- “c” : The data is stationary around a constant (default).
- “ct” : The data is stationary around a trend.

statsmodels.tsa.stattools.adfuller

regression : {“c”,”ct”,”ctt”,”n”}

Constant and trend order to include in regression.

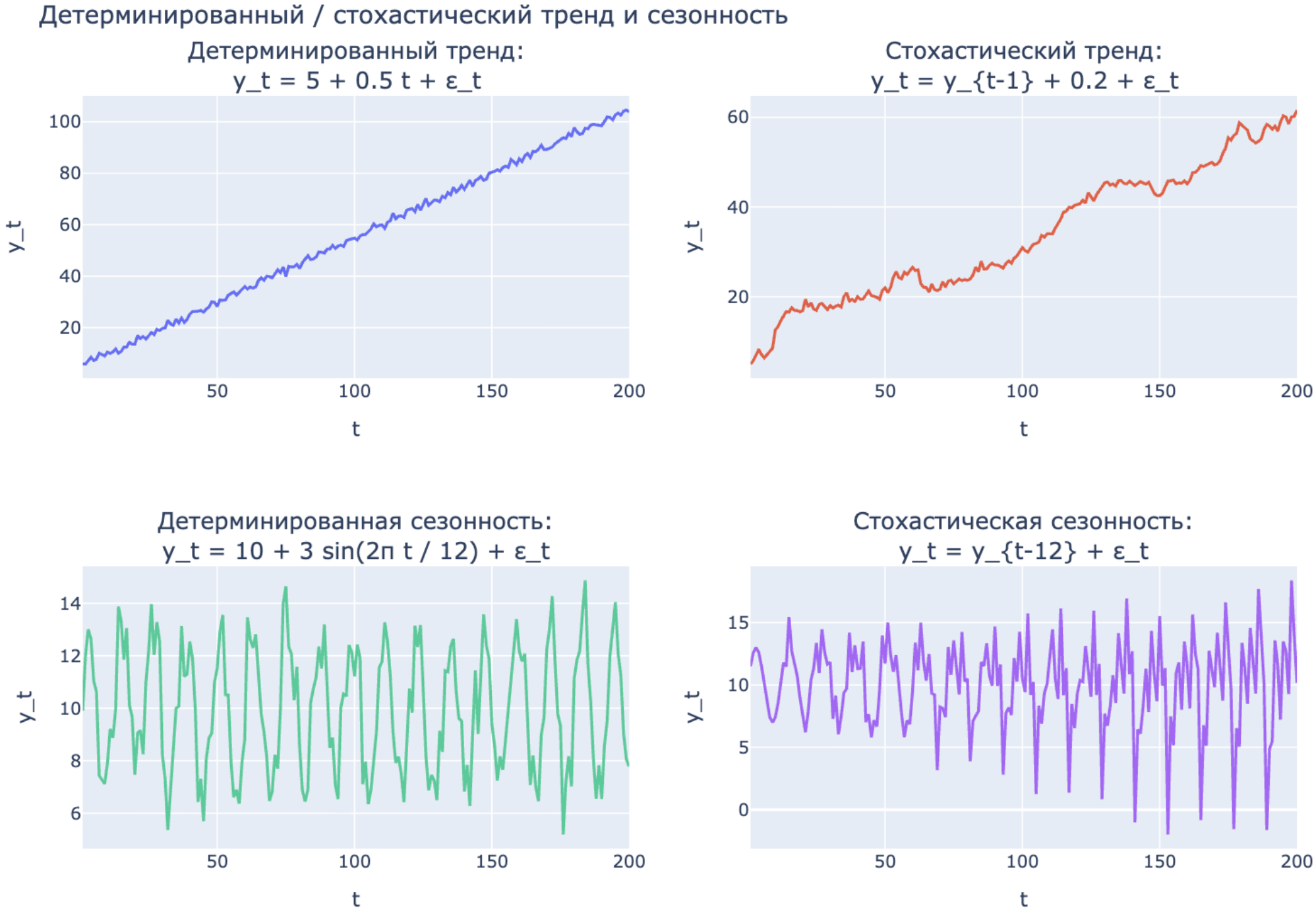
- “c” : constant only (default).
- “ct” : constant and trend.
- “ctt” : constant, and linear and quadratic trend.
- “n” : no constant, no trend.

Сезонность

pmdarima.arima.CHTest

H_0: сезонность **детерминированная** (нет сезонного unit root)
H_1: сезонность **случайная**

https://users.ssc.wisc.edu/~behansen/papers/jbes_95.pdf


Детерминированный / стохастический тренд и сезонность
Детерминированный тренд:
y_t = 5 + 0.5 t + ε_t
Стохастический тренд:
y_t = y_{t-1} + 0.2 + ε_t
Детерминированная сезонность:
y_t = 10 + 3 sin(2π t / 12) + ε_t
Стохастическая сезонность:
y_t = y_{t-12} + ε_t
y_t
t

Случайный или детерминированный?

Случайная компонента требует дифференцирования, детерминированная — параметризации.

Тренд

Случайный:
- Дифференцирование (d)

Детерминированный:
- bt (или другой полином) в SARIMAX

Сезонность

Случайная:
- Сезонное дифференцирование (D)

Детерминированная:
- Сезонные дамми в SARIMAX
- Сезонные AR/MA-члены (P, Q)

А что про связь ETS и SARIMA?

ETS model	ARIMA model	Parameters
ETS(A,N,N)	ARIMA(0,1,1)	$\theta_1 = \alpha - 1$
ETS(A,A,N)	ARIMA(0,2,2)	$\theta_1 = \alpha + \beta - 2$ $\theta_2 = 1 - \alpha$
ETS(A,A_d,N)	ARIMA(1,1,2)	$\phi_1 = \phi$ $\theta_1 = \alpha + \phi\beta - 1 - \phi$ $\theta_2 = (1 - \alpha)\phi$
ETS(A,N,A)	ARIMA(0,1,m)(0,1,0)$_m$	
ETS(A,A,A)	ARIMA(0,1,$m+1$)(0,1,0)$_m$	
ETS(A,A_d,A)	ARIMA(1,0,$m+1$)(0,1,0)$_m$	

https://otexts.com/fpp3/arima-ets.html

Вопросы?

Диагностика и коррекция ошибок

Построили модель → посмотрели ошибки → поняли, где лажаем → повторили

1) **В остатках не должно остаться автокорреляции**
   - ACF, тест Ljung-Box (первые m автокорреляции совместно равны 0).

statsmodels.stats.diagnostic.acorr_ljungbox

Вычисляем статистику:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\widehat{\rho}_k^2}{n-k}, \qquad H_0: \rho_1 = \rho_2 = \cdots = \rho_m = 0$$

где n — длина ряда, $\widehat{\rho}_k$ — автокорреляция k-го порядка, m — количество проверяемых лагов. Пусть α — уровень значимости, тогда при $Q > \chi^2_{1-\alpha,m}$, где $\chi^2_{1-\alpha,m}$ — α–квантиль распределения хи-квадрат с m степенями свободы, нулевая гипотеза отвергается и признается наличие автокорреляции до m-го порядка во временном ряду.

   - Как поправить: в зависимости от графика ACF (PACF) поподбирать еще p, d, q, P, D, Q. Может тренд или сезонность детерминированная, а мы ее делаем случайной (или наоборот)?

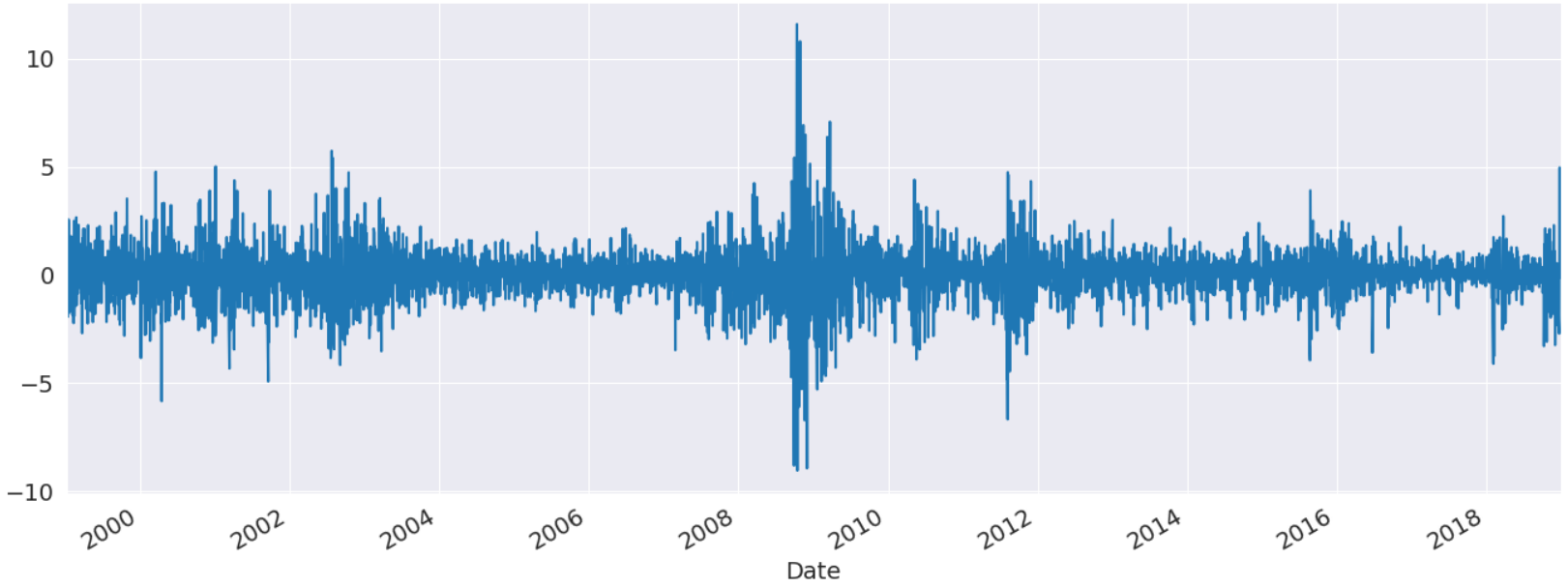
http://www.machinelearning.ru/wiki/index.php?title=Критерий_Льюнга-Бокса

Диагностика и коррекция ошибок

Построили модель → посмотрели ошибки → поняли, где лажаем → повторили

2) Остатки должны быть гетероскедастичны (иметь постоянную дисперсию)

- График остатков — не должно быть «кластеров» с разным значением дисперсии.
- Как поправить: логарифмирование/Box-Cox/Yeo-Johnson, ARCH/GARCH



GARCH (1, 1)

$$r_t = \mu + \epsilon_t$$

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

$$\epsilon_t = \sigma_t e_t, \quad e_t \sim N(0,1)$$

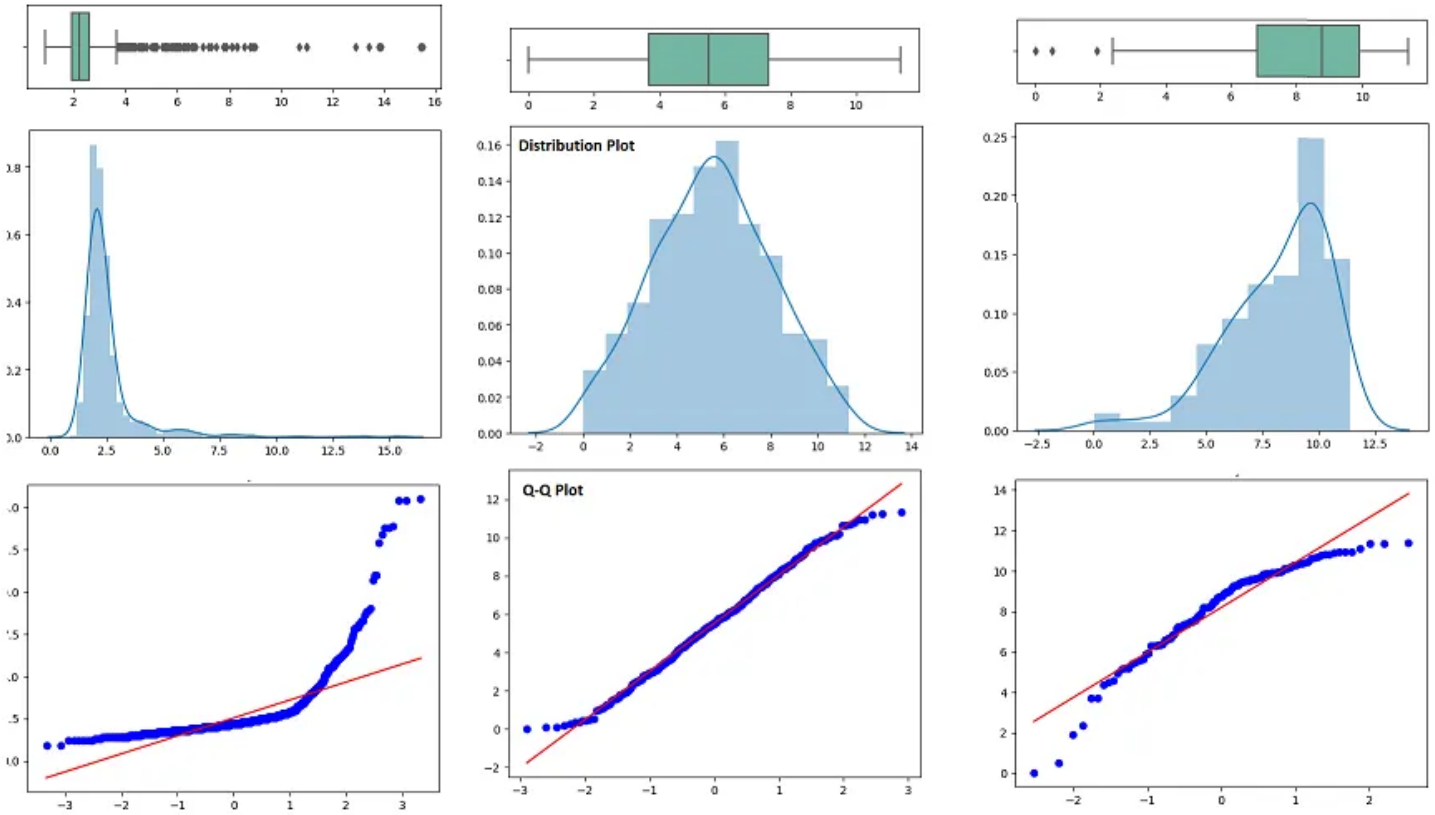
https://arch.readthedocs.io/en/latest/univariate/univariate_volatility_modeling.html

Диагностика и коррекция ошибок

Построили модель → посмотрели ошибки → поняли, где лажаем → повторили

3) Остатки должны быть нормальны

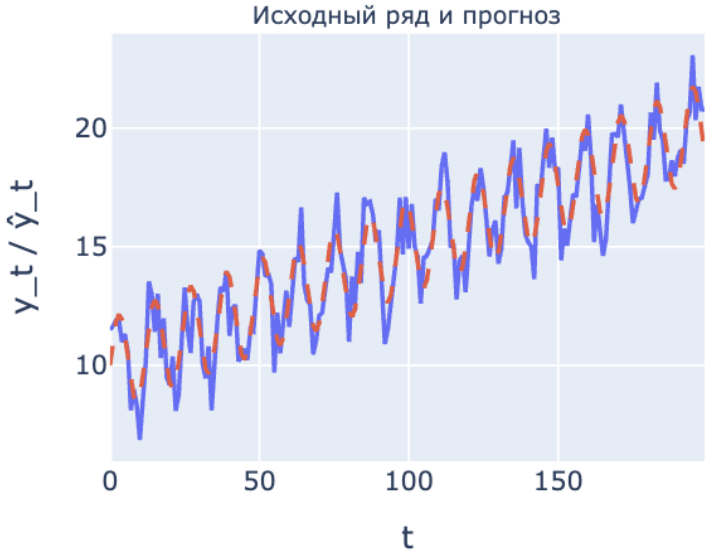
- Гистограмма + QQ-plot.
- Как поправить: как повезёт. Обработка выбросов, обработка через экзогенные признаки.

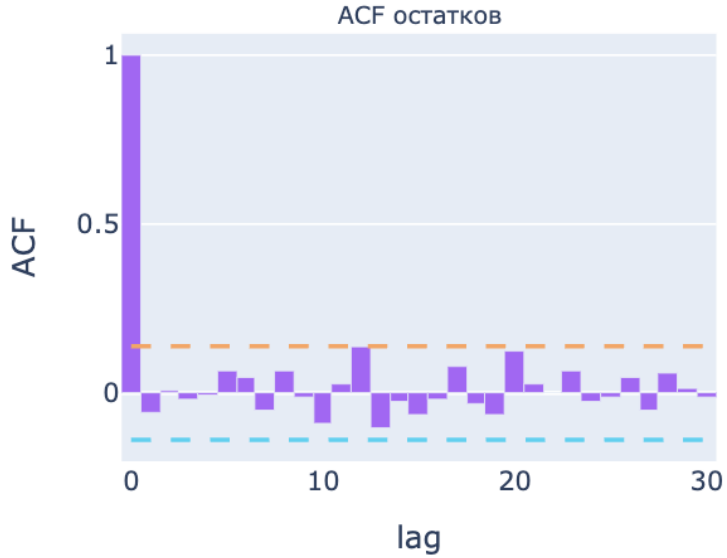


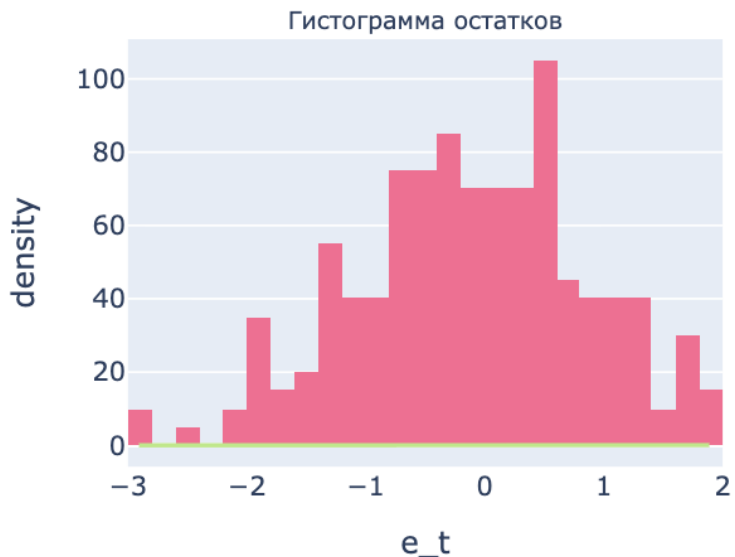
https://amitprius.medium.com/fully-understand-q-q-plot-for-probability-distribution-in-machine-learning-7ba16166cae6

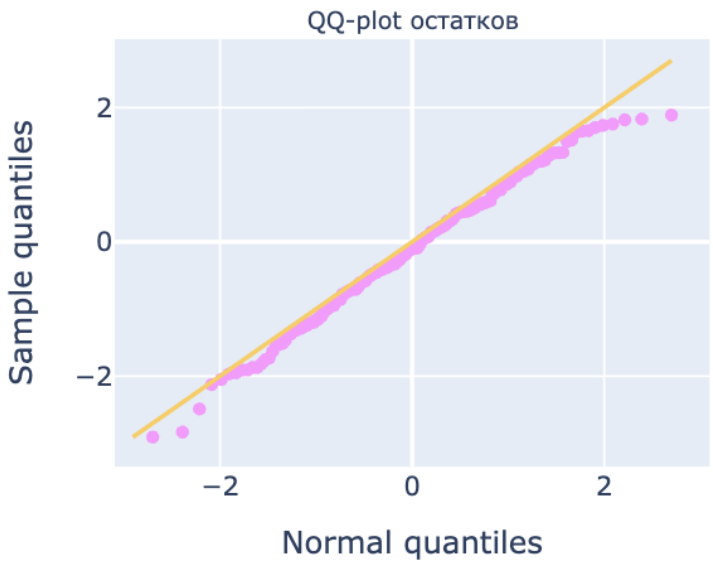
Пример 1: модель почти верна

$y_t = 10 + 0.05 \cdot t + 2 \cdot \sin(2\pi\, t / 12) + \varepsilon_t,\ \varepsilon_t \sim N(0,1);\ \hat{y}_t = 10 + 0.05 \cdot t + 2 \cdot \sin(2\pi\, t / 12)$





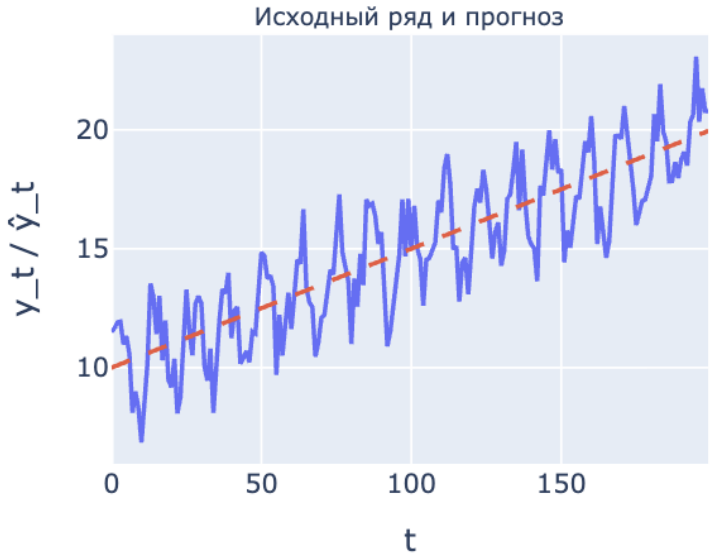


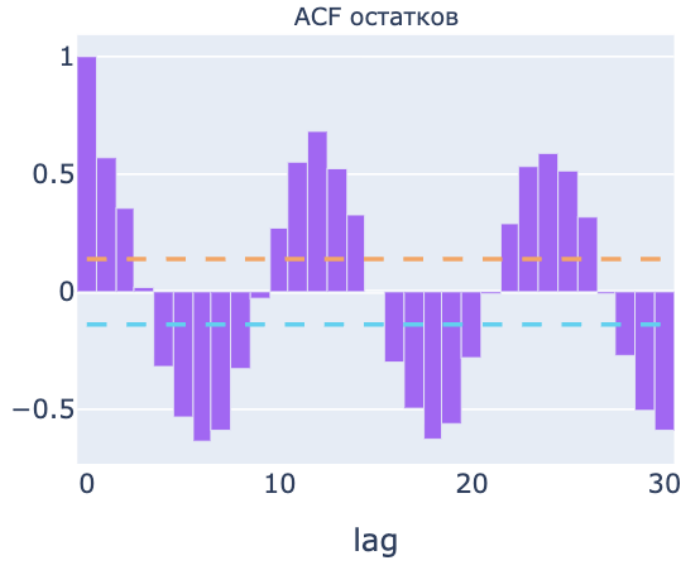


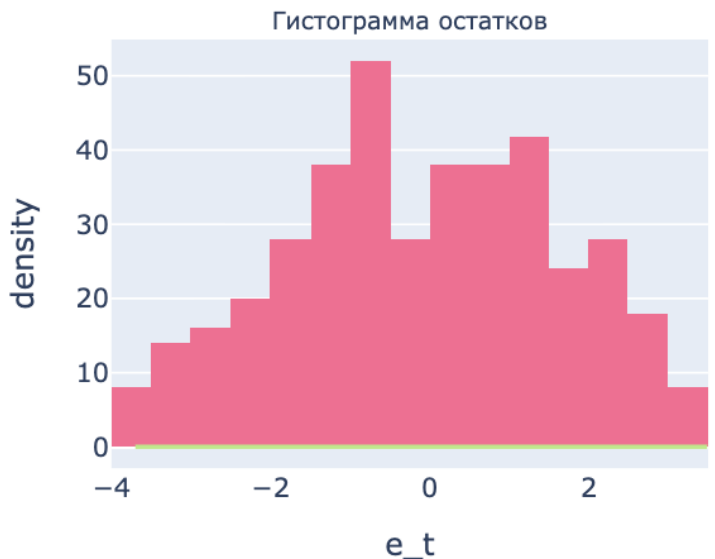
Ljung–Box p-value = 0.517 ≥ 0.05
⇒ нет автокорреляции остатков

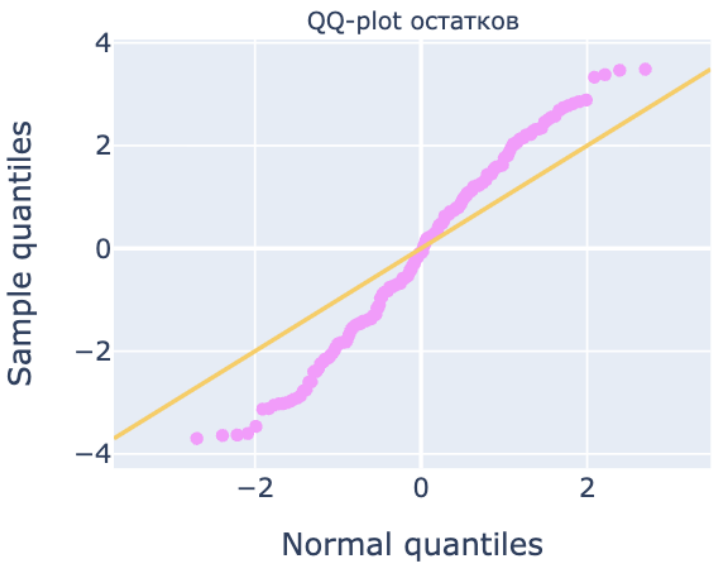
Пример 2: пропущенная сезонность

$y_t = 10 + 0.05 \cdot t + 2 \cdot \sin(2\pi\, t / 12) + \varepsilon_t$; $\hat{y}_t = 10 + 0.05 \cdot t$ (сезонность пропущена)





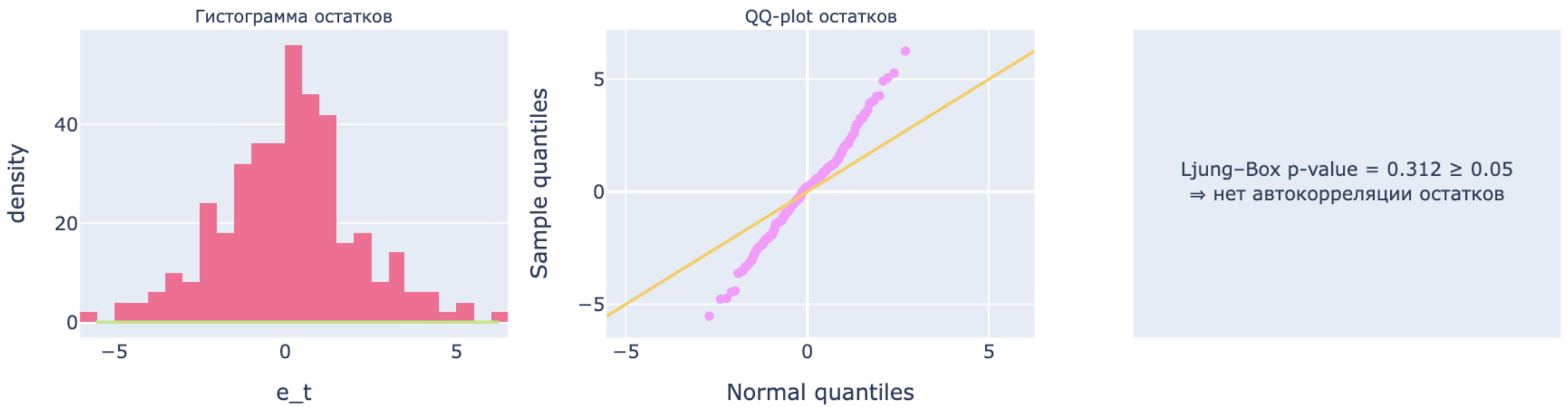




Ljung–Box p-value = 0.000 < 0.05
⇒ есть автокорреляция остатков

Пример 3: гетероскедастичность

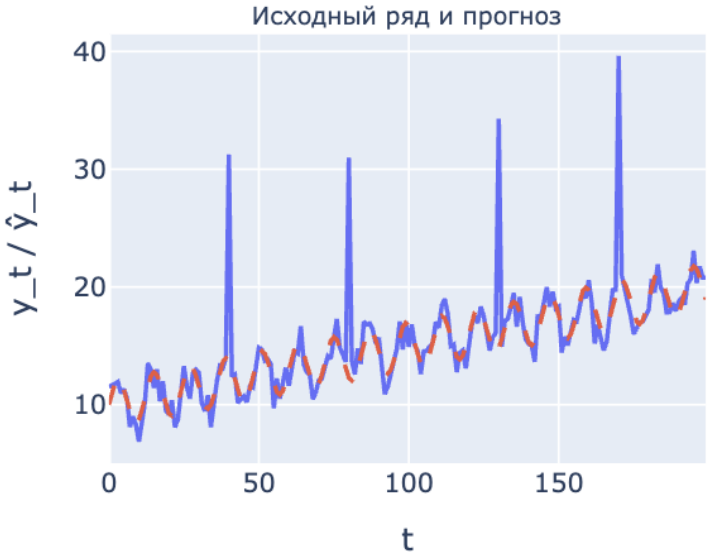
y_t = 10 + ε_t, ε_t ~ N(0, σ_t^2), σ_t↑; ŷ_t = 10 (гетероскедастичность не учтена)

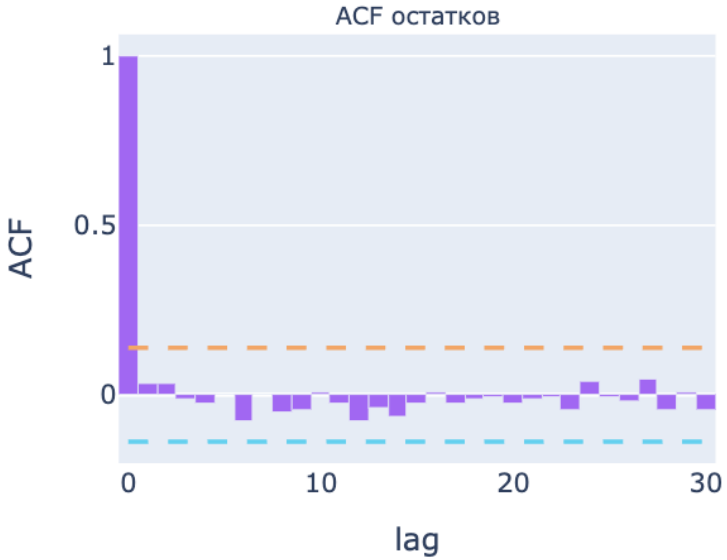


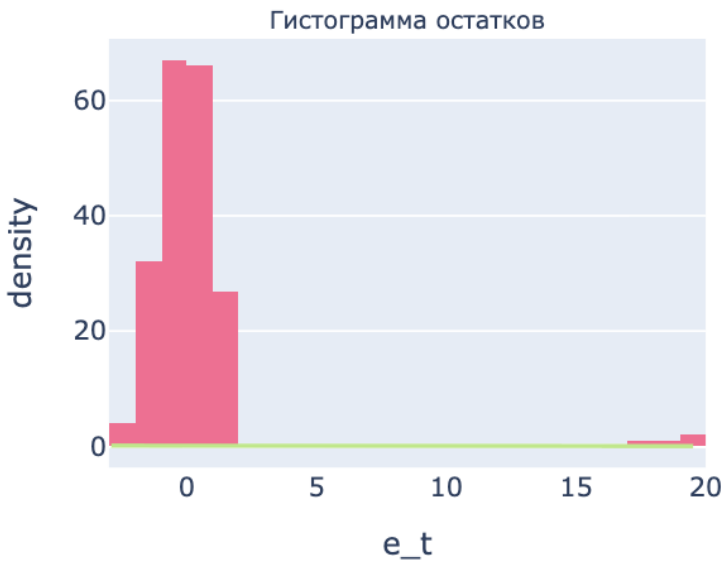
Ljung–Box p-value = 0.312 ≥ 0.05
⇒ нет автокорреляции остатков

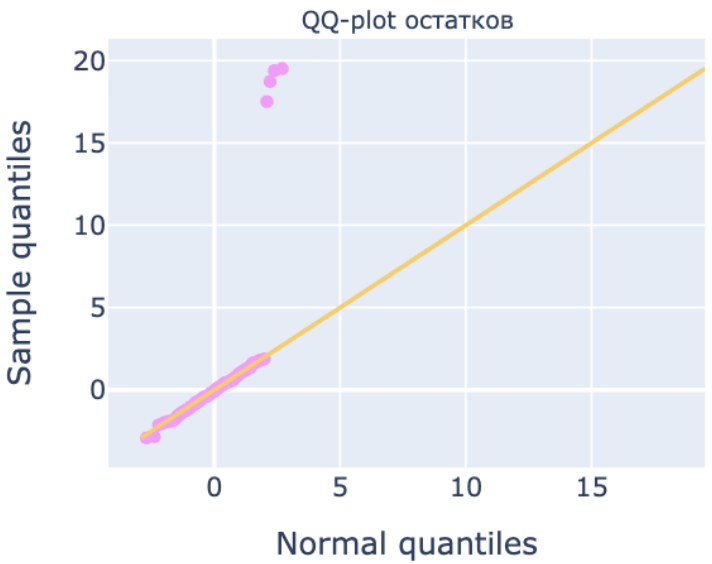
Пример 4: выбросы

$y_t = 10 + 0.05\cdot t + 2\cdot\sin(2\pi\, t / 12) + \varepsilon_t + 20\cdot I\{t \in \{40,80,130,170\}\};\ \hat{y}_t = 10 + 0.05\cdot t + 2\cdot\sin(2\pi\, t / 12)$





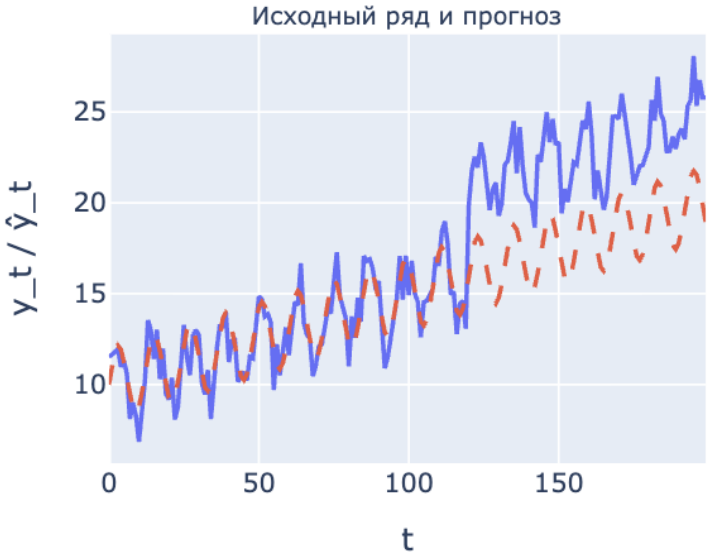


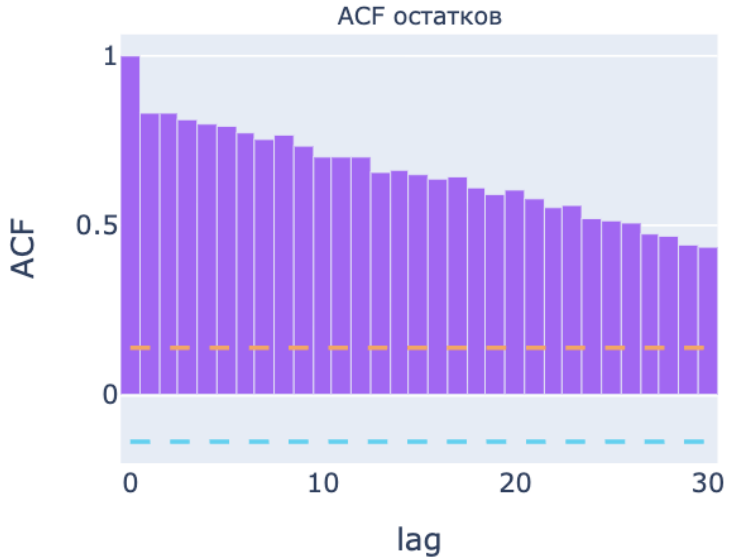


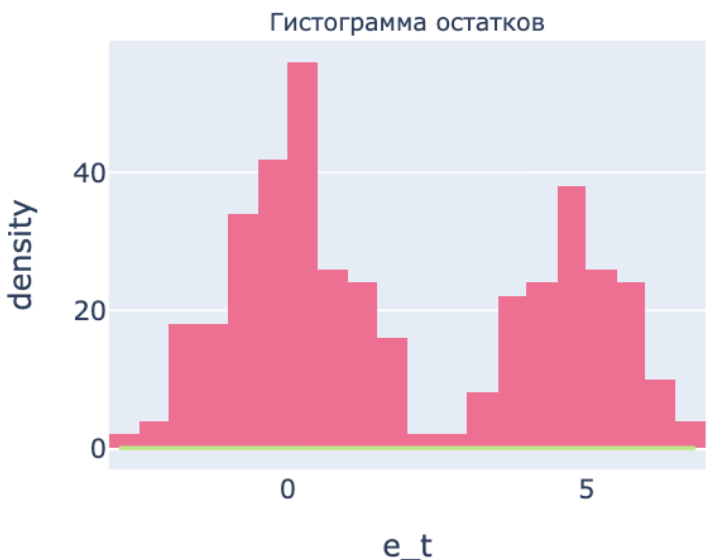
Ljung–Box p-value = 0.999 ≥ 0.05
⇒ нет автокорреляции остатков

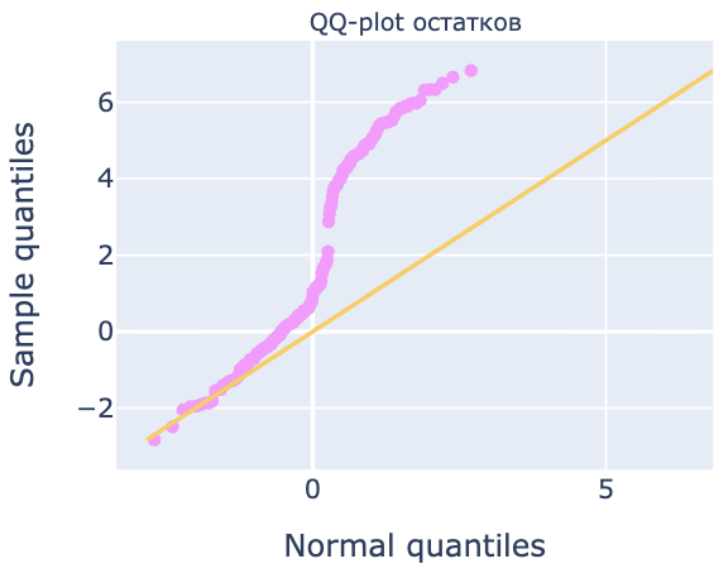
Пример 5: структурный сдвиг уровня

$y_t = 10 + 0.05 \cdot t + 2 \cdot \sin(2\pi\, t / 12) + \varepsilon_t + 5 \cdot I\{t \geq 120\};\ \hat{y}_t = 10 + 0.05 \cdot t + 2 \cdot \sin(2\pi\, t / 12)$









Ljung–Box p-value = 0.000 < 0.05
⇒ есть автокорреляция остатков

Вопросы?

Тета-метод

- Выделим из ряда несколько «тета»-рядов
- Спрогнозируем каждый из них отдельно
- Усредним прогнозы

«Тета»-ряды, соответствующие разным θ ловят разную информацию о локальной кривизне ряда.

- Извлечь сезонную компоненту из ряда
- Посчитать “тета”-ряды для $\theta = 0$ (линейный тренд) и $\theta = 2$
- Экстраполировать линейный тренд, а вторую линию продолжить экспоненциальным сглаживанием ETS(ANN).
- Усреднить прогнозы для “тета”-рядов
- Вернуть сезонную компоненту

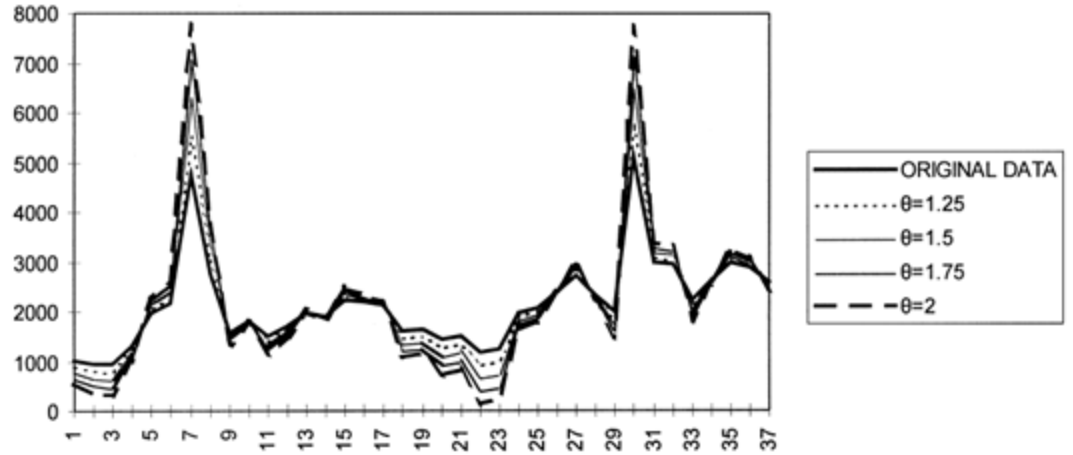
Определение второй разности

$$\Delta^2 y_t = y_t - 2y_{t-1} + y_{t-2}$$

Определение тета-ряда

$$\Delta^2 y_{t,\theta} = \theta\, \Delta^2 y_t.$$

Ускорение тета-линии в theta раз больше ускорения исходного ряда.



Источник:
Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. International journal of forecasting, 16(4), 521-530.

https://www.youtube.com/watch?v=lG5WXZ4dYVU

Тета-метод

Берём $\theta = 2$:

$$\Delta^2 y_{t,\theta=2} = 2\,\Delta^2 y_t.$$

$$y_{t,\theta=2} - 2y_{t-1,\theta=2} + y_{t-2,\theta=2} = 2(y_t - 2y_{t-1} + y_{t-2}).$$

$$\sum_{t=1}^{T} (y_t - y_{t,\theta=2})^2 \to \min_{y_{1,\theta=2},\, y_{2,\theta=2}}.$$

Определение второй разности

$$\Delta^2 y_t = y_t - 2y_{t-1} + y_{t-2}$$

Определение тета-ряда

$$\Delta^2 y_{t,\theta} = \theta\,\Delta^2 y_t.$$

Ускорение тета-линии в theta раз больше ускорения исходного ряда.

ETS(AAN):

$$\begin{cases} y_t = \ell_{t-1} + b_{t-1} + u_t, \\ \ell_t = \ell_{t-1} + b_{t-1} + \alpha u_t, \\ b_t = b_{t-1} + \beta u_t, \\ u_t \sim N(0, \sigma^2),\ u_t \text{ независимы}. \end{cases}$$

Положим $\beta = 0$ и $\ell_1 = y_1$:

$$\begin{cases} y_t = \ell_t + \mathrm{b} + u_t, \\ \ell_t = \ell_{t-1} + \mathrm{b} + \alpha u_t, \\ \ell_1 = y_1, \end{cases}$$

https://www.youtube.com/watch?v=lG5WXZ4dYVU

Название	Год	Описание данных	Метрики	# Моделей	Вывод
M1	1982	1001 реальный ряд (годовые, квартальные, месячные; экономика, финансы, демография и др.)	Процентные (MAPE и др. процентные ошибки)	15 (plus 9 variations)	Сложные методы не дают явного преимущества над простыми; комбинации методов улучшают точность; качество сильно зависит от горизонта прогнозирования.
M2	1993	29 реальных рядов (23 из 4 компаний + 6 макро); прогнозы в «реальном времени» с возможностью экспертных поправок	Те же процентные метрики (MAPE и др.)	16 (including 5 human forecasters and 11 automatic trend-based methods) plus 2 combined forecasts and 1 overall average	Результаты очень близки к M1: простые методы и их комбинации остаются конкурентоспособными.
M3	2000	3003 ряда разных частот (yearly, quarterly, monthly, other) и доменов (micro, industry, macro, finance, demography)	sMAPE + ранги и относительные ошибки (MdSAPE, MdRAE, % better)	24 метода	Подтверждены выводы M1–M2: простые методы и их комбинации очень сильны; новый метод Theta показывает лучшую среднюю точность.
M4	2018–2020	100 000 реальных временных рядов разных частот и предметных областей	sMAPE, MASE (точечные прогнозы), MSIS (интервальные), интегральный показатель OWA	All major ML and statistical methods	Все топ-методы — комбинации и/или гибриды; лучший метод (Smyl) сочетает статистику и ML; «чистый» ML без статистики в среднем слабее комбинаций статистических методов.
M5 (Walmart, Kaggle)	2020–2022	≈42 000 иерархических дневных рядов продаж Walmart + экзогенные фичи (цены, промо, календарь, запасы)	WRMSSE для точности; Weighted Scaled Pinball Loss / WSPL для неопределённости	All major forecasting methods, including Machine and Deep Learning, and Statistical ones	Наверху лидерборда — чистые ML-ансамбли (LightGBM, глубокие сети); ML особенно силён, когда много данных, иерархия и богатый набор признаков; ансамбли снова лучше, чем одиночные модели.
M6 (финансовый)	2022–2024	100 активов (50 акций S&P 500 + 50 ETF); участники сами собирают и выбирают данные, прогнозируют доходности и риск	Комбинация метрик точности прогнозов и доходности/риска портфелей (out-of-sample)	All major forecasting methods, including Machine and Deep Learning, and Statistical ones	Почти никто не смог стабильно обогнать простой равновзвешенный портфель.

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

Вопросы?

А что, если рядов несколько? А если много?

1) Можем моделировать ряды независимо

- ОК, если рядов немного И / ИЛИ если модели «лёгкие»

Computational time

N time series	Time (mins)	N cpus	CVWindows	Cost (Dollars)
10,000	0.32	128	7	$0.14
100,000	0.74	128	7	$0.33
1,000,000	4.81	128	7	$2.14
5,000,000	21.87	128	7	$9.73
10,000,000	44.12	128	7	$19.63

Performace (MSE)

N time series	Croston	SeasNaive	Naive	ADIDA	HistoricAverage	SeasWindowAverage	iMAPA	WindowAverage	SeasExpSmooth
10,000	4.1045	0.0414	8.0418	4.1366	4.0313	0.026	4.1366	4.0239	8.0377
100,000	4.1035	0.0418	8.0403	4.1373	4.0307	0.0261	4.1373	4.0233	8.0372
1,000,000	4.1046	0.0417	8.0417	4.1381	4.0314	0.026	4.1381	4.024	8.038
5,000,000	4.1042	0.0417	8.0416	4.1377	4.0311	0.026	4.1377	4.0237	8.038
10,000,000	4.1043	0.0417	8.0418	4.1379	4.0313	0.026	4.1379	4.0239	8.0381

https://github.com/Nixtla/statsforecast/blob/main/experiments/benchmarks_at_scale/README.md

А что, если рядов несколько? А если много?

1) Можем моделировать ряды независимо

- Не ОК, если модели хоть немного потяжелее

```
Train shape: (82928, 3)
Val shape: (21955, 3)
Test shape: (8784, 3)
Test targets shape: (13176, 3)
Number of series in train set: 366
Number of series in val set: 366
Number of series in test set: 366
Number of series in test targets set: 366
```

```
forecasts_df = sf.forecast(df=df_train_val_test, h=HORIZON)
forecasts_df.head()
```

```
Forecast: 100% ████████████████████████████████ 366/366 [Elapsed: 24:25]
```

:

	unique_id	ds	Naive	SeasonalNaive	AutoARIMA	AutoETS	AutoTheta
0	0	1991-08-01	6048.7421	6053.7042	6223.552704	6437.175095	6391.699647
1	0	1991-09-01	6048.7421	3878.1285	4066.305395	3922.360583	3897.060043
2	0	1991-10-01	6048.7421	2806.5149	2953.139918	2595.374443	2596.826689
3	0	1991-11-01	6048.7421	1735.5382	1893.951884	1655.352541	1657.382880
4	0	1991-12-01	6048.7421	2128.9200	2256.408123	1935.798299	1936.535375

https://www.kaggle.com/code/elineii/tsururu-baselines

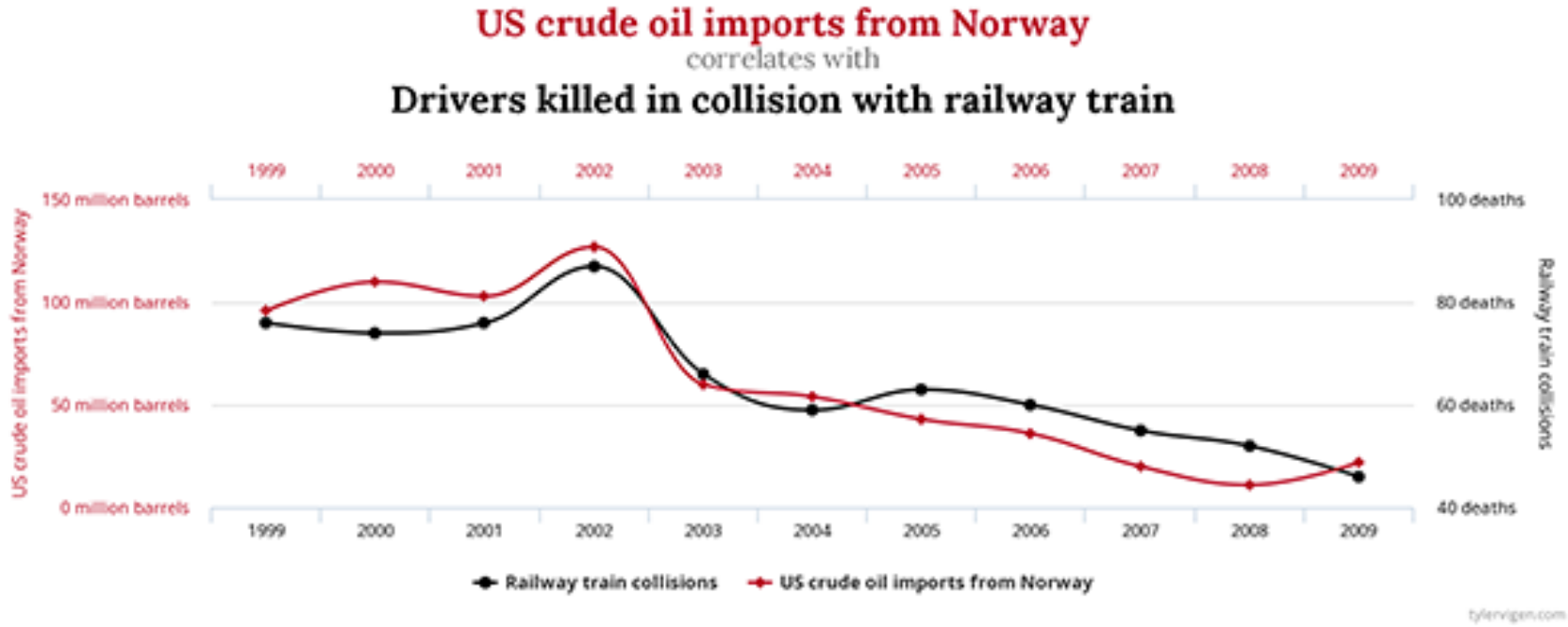
А что, если рядов несколько? А если много?

2) Можем моделировать зависимости между рядами через эконометрику

$$y_{1,t} = c_1 + \phi_{11,1} y_{1,t-1} + \phi_{12,1} y_{2,t-1} + \varepsilon_{1,t}$$
$$y_{2,t} = c_2 + \phi_{21,1} y_{1,t-1} + \phi_{22,1} y_{2,t-1} + \varepsilon_{2,t}$$

Могут быть взаимно коррелированы

- Если не приводить ряды к стационарному виду, можно получить бессмысленные регрессии (spurious regression).



https://statisticsbyjim.com/basics/spurious-correlation/

А что, если рядов несколько? А если много?

2) Можем моделировать зависимости между рядами через эконометрику

$$y_{1,t} = c_1 + \phi_{11,1} y_{1,t-1} + \phi_{12,1} y_{2,t-1} + \varepsilon_{1,t}$$
$$y_{2,t} = c_2 + \phi_{21,1} y_{1,t-1} + \phi_{22,1} y_{2,t-1} + \varepsilon_{2,t}$$

Могут быть взаимно коррелированы

- Если не приводить ряды к стационарному виду, можно получить бессмысленные регрессии (spurious regression).

1. Проверяем, что каждый их них **I(1)** (ADF/KPSS и т. п.).
2. Тестируем коинтеграцию (между рядами есть **долгосрочное стационарное соотношение**).
    - Engle–Granger:
        - оцениваем $y_t = \beta x_t + u_t$ проверяем стационарность u_t ;
    - Johansen — для системы из нескольких рядов.

3. **Коинтеграции нет** → модель на приведенных к стационарности рядах.
   Коинтеграция есть → ECM/VECM.

$$\triangle y_t = \beta_0 \triangle x_t - \gamma(y_{t-1} - \alpha_L - \beta_L x_{t-1}) + \varepsilon_t$$

https://help.fsight.ru/ru/mergedProjects/lib/02_time_series_analysis/uimodelling_errcorrmodel.htm

А что, если рядов несколько? А если много?

Можем ответить на вопрос «правда ли, что y_2 влияет на y_1»?

Granger-causality test

1) $y_{1,t} = c_1 + \phi_{11,1} y_{1,t-1} + \phi_{12,1} y_{2,t-1} + \varepsilon_{1,t}$ — full

2) $y_{1,t} = c_1 + \phi_{11,1} y_{1,t-1} + \varepsilon_{1,t}$ — restricted

$$F = \frac{(R_F^2 - R_R^2)/q}{(1 - R_F^2)/(n - k_{\mathrm{F}})} \sim F(q, n - k_{\mathrm{F}})$$

Объем выборки

Количество параметров

Кол-во ограничений ($k_F - k_R$)

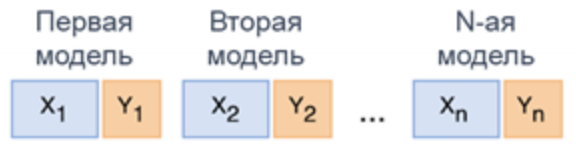
H_0: коэффициенты при ограничениях = 0 (нужно выбрать restricted модель)

H_1: коэффициенты при ограничениях != 0 (нужно выбрать full модель) — прошлые значения x_tсодержат информацию, полезную для прогноза y_t**поверх** той информации, которая есть в прошлых y_t(но это не значит, что есть «истинная причинность»)

А что, если рядов несколько? А если много?

2) Можем моделировать зависимости между рядами через ML / DL

Local



Много моделей, по одной для каждого временного ряда.

Multivariate

Одна модель для всех одномерных временных рядов.

Признаки наблюдений, относящихся к одной временной точке, **объединяются**

Global

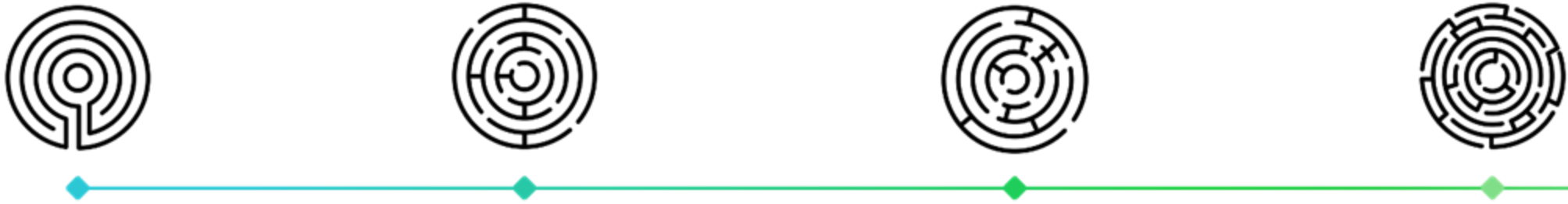
Одна модель для всех одномерных временных рядов.

Признаки отдельных наблюдений **не пересекаются** между рядами.

https://arxiv.org/abs/2008.00444

Как получить прогнозы?

Группы методов, которые используются на практике



Naive methods

- (Seasonal) Naive
- Mean, Median

Statistical methods

- ETS
- Theta
- ARIMA

ML methods

- LinearRegression
- GradientBoosting

DL methods

- DLinear
- NBEATS
- PatchTST
- GPT4TS

Начинаем всегда с них!

Как получить прогнозы?

Группы методов, которые используются на практике

Группа методов	Мало данных	Гибкость и адаптация	Многомерные ряды и внешние признаки
Naive methods	+	-	-
Statistical methods	+	+-	+-
ML methods	+-	+-	+
DL methods	-	+	+