Theory–First Tour of the Time-Series Pipeline

Below is a theory-first tour of every statistical object in our pipeline. For each we give:

(i) A definition or formula,

(ii) The property it measures or enforces,

(iii) Why it is needed here.

1. Box–Cox Power Transformation

Definition

For $y > 0$ and $\lambda \neq 0$:

$$y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda}, \quad y^{(0)} = \log y.$$

Role

Simultaneously stabilises variance and linearises exponential growth.[1]

Why here

All series exhibit level-dependent variance; the MLE $\hat{\lambda}$ decides additive vs. multiplicative form.

2. STL Decomposition

$$y_t = T_t + S_t + R_t,$$

with T_t trend, S_t seasonal of known period s, R_t remainder.[2]

- *Role*: Robustly separates trend/seasonality/noise.
- *Why*: Compare RMS of R_t on raw vs. log series to choose additivity.

[1] Box & Cox (1964).
[2] Cleveland et al. (1990).

3. Lomb–Scargle Periodogram

$$P(f) = \frac{1}{2\sigma^2}\Big[\frac{[\sum(y_i - \bar{y})\cos\omega(t_i - \tau)]^2}{\sum\cos^2\omega(t_i - \tau)} + \frac{[\sum(y_i - \bar{y})\sin\omega(t_i - \tau)]^2}{\sum\sin^2\omega(t_i - \tau)}\Big], \quad \omega = 2\pi f,$$

τ chosen to decorrelate sine/cos terms.[3]

- *Role*: Finds periodicity in gapped data.
- *Why*: No strong peak set $s = 1$ (no season).

4. State-Space Kalman Filter in ARIMA/SARIMAX

State-Space Formulation

Any ARIMA/SARIMA can be written as

$$\begin{cases} \boldsymbol{x}_{t+1} = \mathbf{F}\,\boldsymbol{x}_t + \mathbf{G}\,\varepsilon_t, \\ y_t = \mathbf{H}^\top \boldsymbol{x}_t + d_t, \end{cases}$$

where $\boldsymbol{x}_t$ is a latent state vector, ε_t white noise.

Kalman Filter

A recursive algorithm that computes the optimal (minimum-variance) estimate $\hat{\boldsymbol{x}}_t$ of the unobserved state given data up to t, and updates its error covariance.[4]

- *Role*:
  - *Interpolation*: fills missing observations via the filter's prediction step.
  - *Extrapolation*: yields multi-step forecasts from the last state.
- *Why*: The `statsmodels` `SARIMAX.fit()` method uses Kalman filtering/smoothing under the hood to handle NaNs and to compute both in-sample state estimates and out-of-sample forecasts.

5. ARIMA / SARIMA Family

ARIMA(p,d,q) $\Phi(B)\nabla^d y_t = \Theta(B)\,\varepsilon_t$, $\Phi(B) = 1 - \sum_{i=1}^{p}\phi_i B^i$, $\Theta(B) = 1 + \sum_{j=1}^{q}\theta_j B^j$.

SARIMA(p,d,q)×(P,D,Q)$_s$ Adds seasonal ∇_s^D and polynomials $\Phi_s(B^s)$, $\Theta_s(B^s)$.

- *Role*: Parametric serial dependence in mean.
- *Why*: Grid-search picks ARIMA(1,0,0) by AIC, confirming no seasonal block.

[3] Lomb (1976); Scargle (1982).
[4] Kalman (1960).

6. Akaike Information Criterion (AIC)

$$\text{AIC} = -2\,\ell_{\max} + 2k,$$

where $\ell_{\max}$=maximized log-likelihood, k=parameters.[5]

- *Role*: Penalises complexity vs. fit.
- *Why*: AIC88 in favor of ARIMA(1,0,0).

7. Ljung–Box Test

$$Q_{LB}(m) = n(n+2)\sum_{h=1}^{m}\frac{\hat{\rho}_h^2}{n-h}, \quad Q_{LB} \sim \chi^2_{m-k}.$$

- *Role*: Tests residual autocorrelation.
- *Why*: p1.0 residuals are white.

8. ARCH–LM Test

Regress ε_t^2 on its own q lags; $TR^2 \sim \chi^2_q$.[6]

- *Role*: Detects conditional heteroskedasticity.
- *Why*: p10 fit GARCH(1,1).

9. GARCH(1,1) Volatility Model

$$\begin{cases} \varepsilon_t = \sigma_t z_t,\ z_t \sim N(0,1), \\ \sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2,\ \omega > 0,\ \alpha,\beta \geq 0,\ \alpha+\beta < 1. \end{cases}$$

- *Role*: Time-varying conditional variance.
- *Why*: Combine with ARIMA variance for density forecasts.

10. Bias-Correct Back-Transform

If $\log Y \sim N(\mu, \sigma^2)$, then $\mathbb{E}[Y] = \exp(\mu + \frac{1}{2}\sigma^2)$.

- *Role*: Removes downward bias from $\exp(\hat{\mu})$.
- *Why*: Unbiased level forecasts.

[5] Akaike (1974).
[6] Engle (1982).

11. Workflow Summary

1. Log transform (Box–Cox $\hat{\lambda} \approx 0$).
2. Detrend and confirm no season ($s = 1$).
3. Fit ARIMA(1,0,0) $\xrightarrow{\text{Kalman filter}}$ fill gaps.
4. Residuals pass Ljung–Box; ARCH–LM triggers GARCH(1,1).
5. Forecast mean μ_t & variance σ_t^2; back-transform.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control*, 19(6):716–723.
- Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *JRSS B*, 26(2):211–252.
- Bollerslev, T. (1986). GARCH: Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, 31(3):307–327.
- Cleveland, R. B. et al. (1990). STL: Seasonal-trend decomposition using Loess. *J. Official Stat.*, 6(1):3–73.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity. *Econometrica*, 50(4):987–1007.
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng.*, 82(1):35–45.
- Lomb, N. R. (1976). Least-squares frequency analysis of unevenly spaced data. *Astrophys. Space Sci.*, 39:447–462.
- Ljung, G. M. & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.
- Scargle, J. D. (1982). Studies in astronomical time series analysis. *ApJ*, 263:835–853.