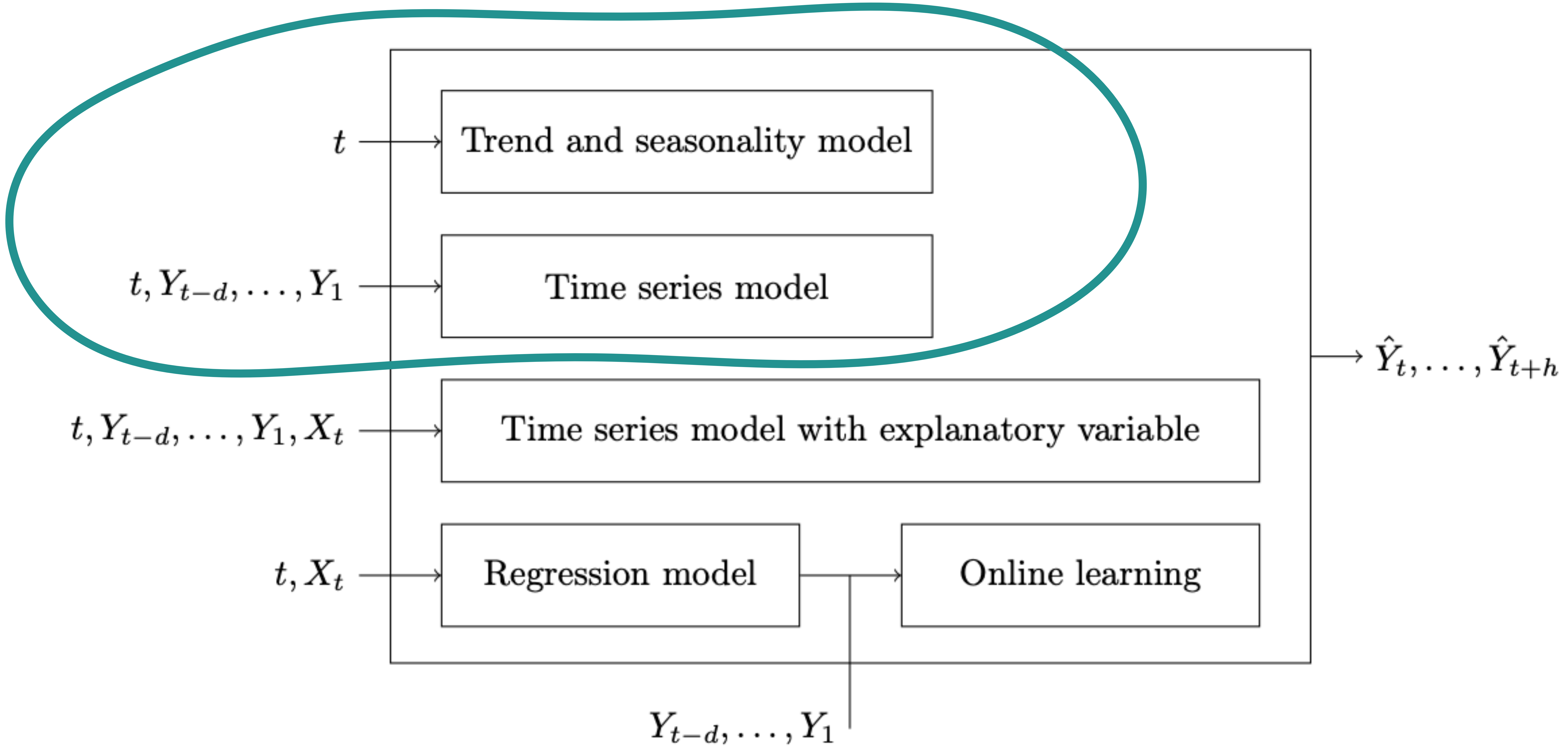
Statistical and Sequential Learning for Time Series Forecasting

Time Series Analysis

Margaux Brégère

Assumption

Let us assume that the random variable Y depends only on t and, possibly, on its past values



Time series decomposition

- Trend, seasonality and noise
- Stationarity

Modelling the deterministic part - Trend and seasonality estimation

- Moving average
- Parametric models
- Nonparametric models

Modelling the noisy part - Residuals analysis

- Stationarity check
- ARMA models

Other approaches

- ARIMA / SARIMA models
- Exponential smoothing

Time series decomposition

Trend, seasonality and noise

It is possible to decompose the time series in three components:

- The trend part $T_t = f(t)$ corresponds to the long-term evolution of the series

  polynomial $f(t) = a_k t^k + \ldots a_1 t + a_0$

  logarithmic $f(t) = \log t$, etc.

- The seasonal part S_t corresponds to periodic phenomena: $\exists \, \tau \in \mathbb{N}^\star \mid \forall t \in \mathbb{N}^\star, S_{t+\tau} = S_t$

  This is not restrictive since two or more periodic phenomenas of periods $\tau_1, \tau_2, \ldots$ can be gathered in a single on of period $\tau =$ smallest common multiple of $\tau_1, \tau_2, \ldots$

- The noise part ε_t whose expectation is zero and which is the only random part of the series; generally (or ideally) it is assumed to be strictly stationary

This decomposition can be

- additive: $Y_t = T_t + S_t + \varepsilon_t$
- multiplicative: $Y_t = T_t \times S_t \times \varepsilon_t$
- combination of the two: $Y_t = T_t + S_t \times \varepsilon_t$, e.g.

Stationarity

Definition

The time series $(\varepsilon_t)_t$ is strictly stationary if, for all $k \in \mathbb{N}$, the joint distribution of $(\varepsilon_t, \ldots, \varepsilon_{t+k})$ does not depend on t

The weak-sense stationarity of $(\varepsilon_t)_t$ only requires that its first moment (i.e. its expectation) and auto-covariances do not vary with respect to time and that the second moment is finite for all times:

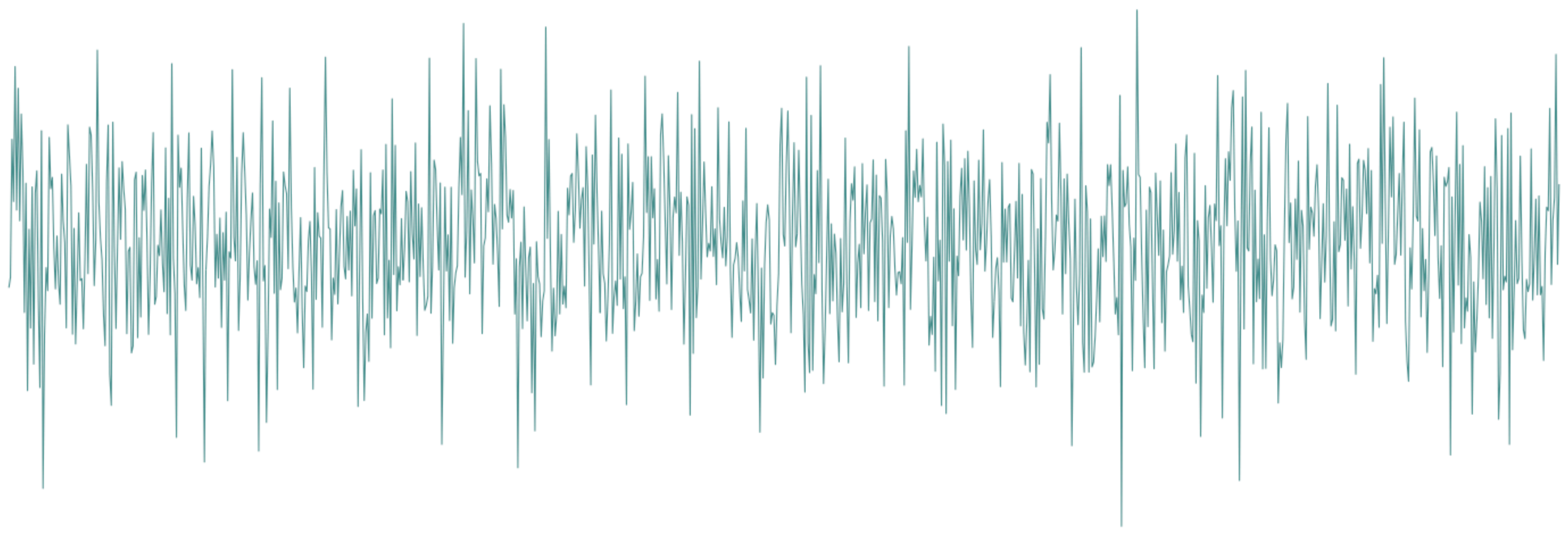
$\exists\, \mu \in \mathbb{R}$\$ and $\gamma : \mathbb{N}^\star \to \mathbb{R}$ such that:

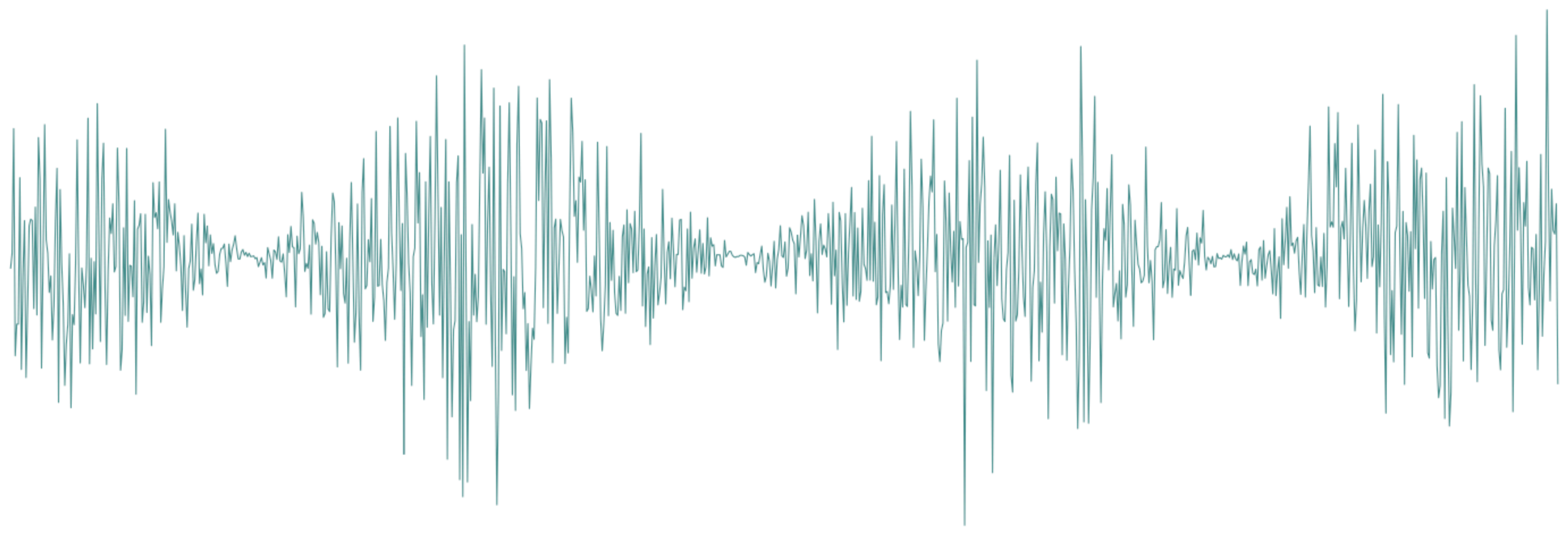
$$\forall\, t \in \mathbb{N}^\star, \mathbb{E}[\varepsilon_t] = \mu$$

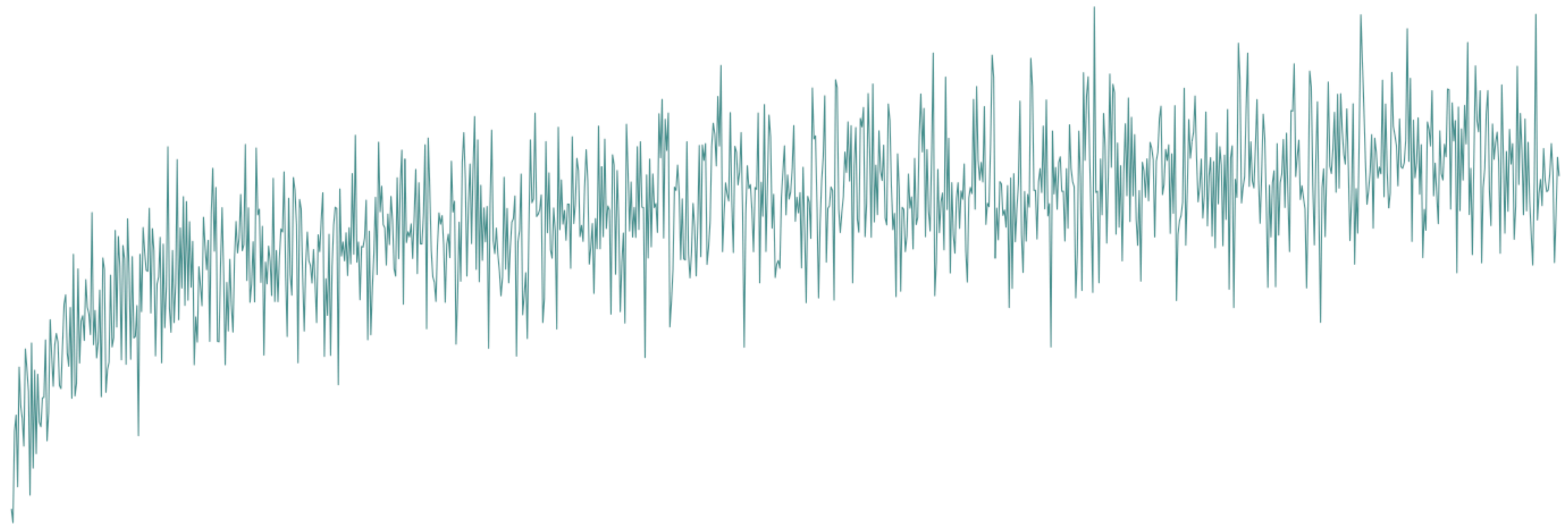
$$\forall\, t \in \mathbb{N}^\star, \forall\, h \in \mathbb{N},\ \mathrm{Cov}(\varepsilon_t, \varepsilon_{t+h}) = \mathbb{E}\left[(\varepsilon_t - \mu)(\varepsilon_{t+h} - \mu)\right] = \gamma(h)$$

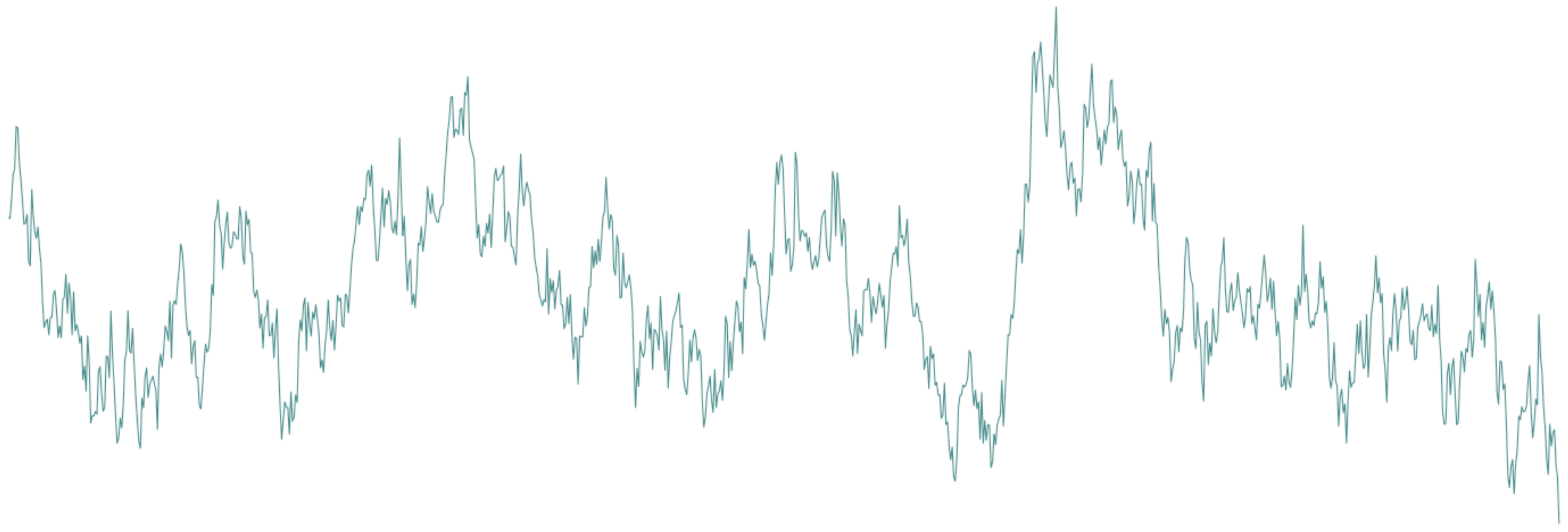
$$\forall\, t \in \mathbb{N}^\star, \mathbb{E}\left[|\varepsilon_t|^2\right] < +\infty$$

Examples

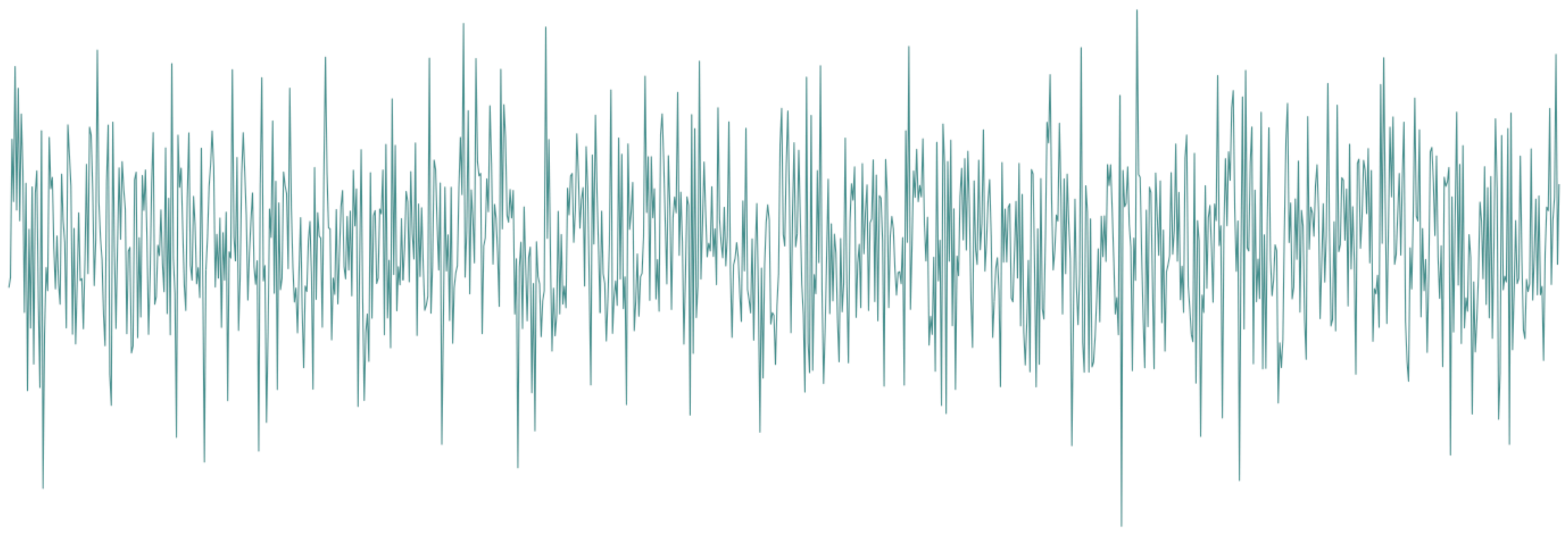




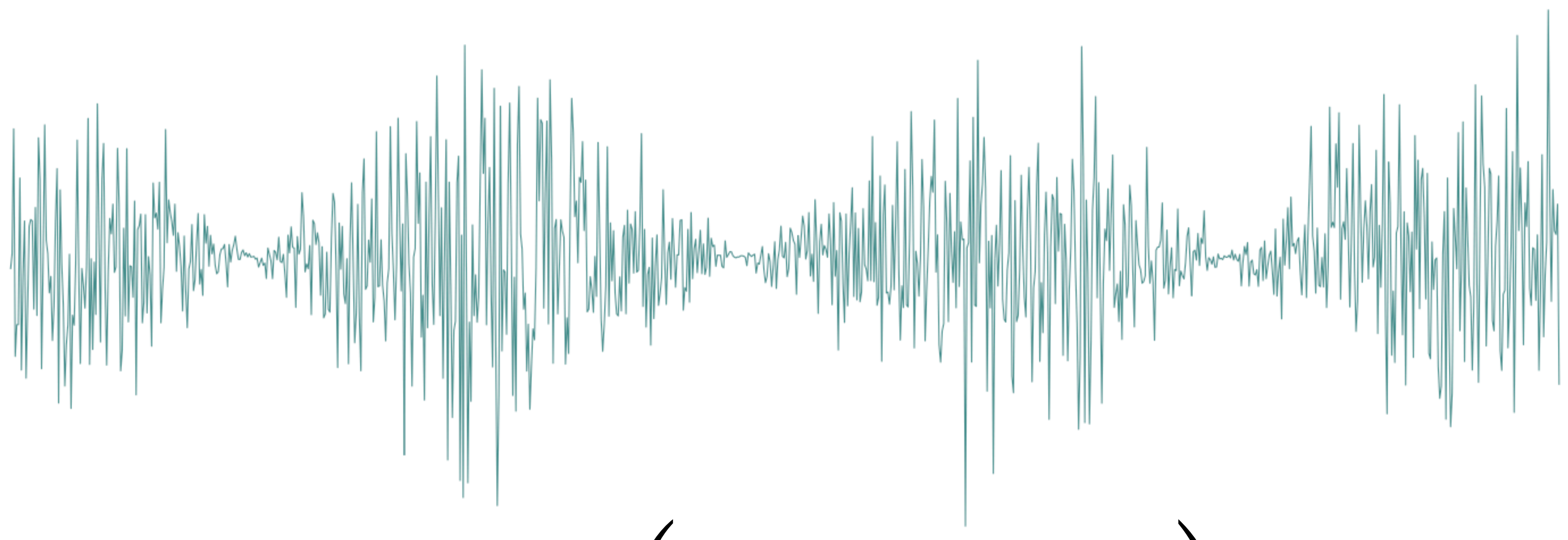




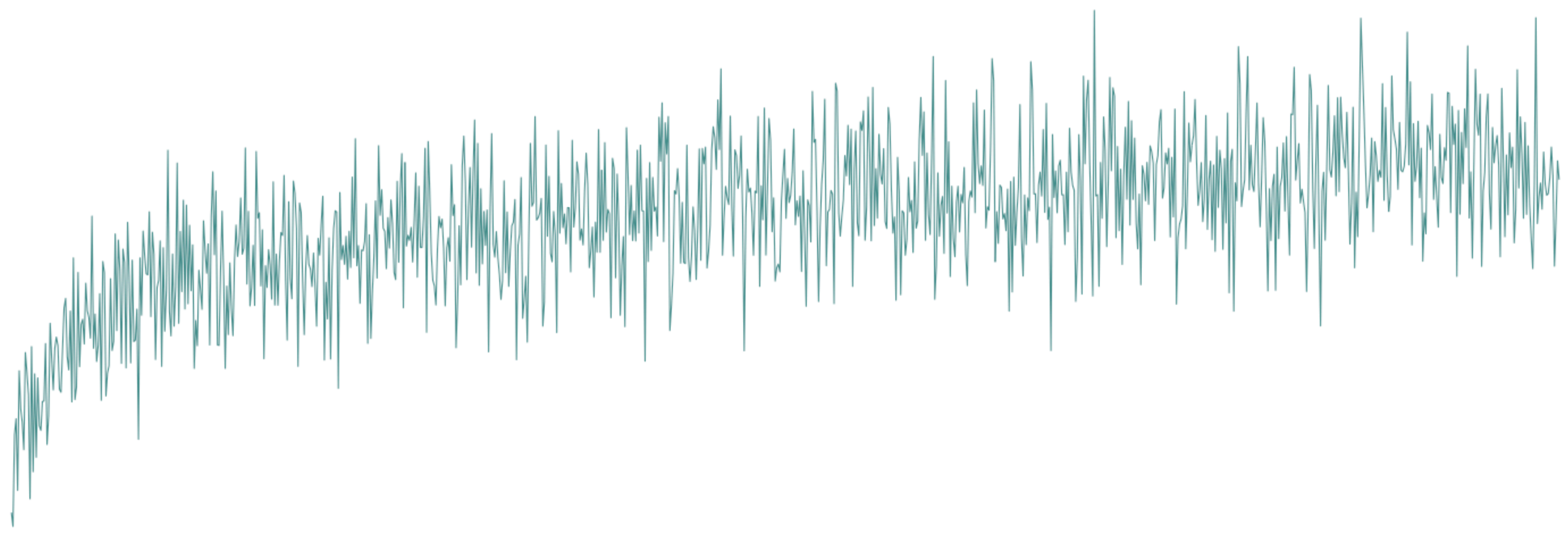
Examples



Gaussian white noise $\varepsilon_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$



$X_t \sim \mathcal{N}\left(0, \left|\cos \frac{t}{100}\right|\right)$

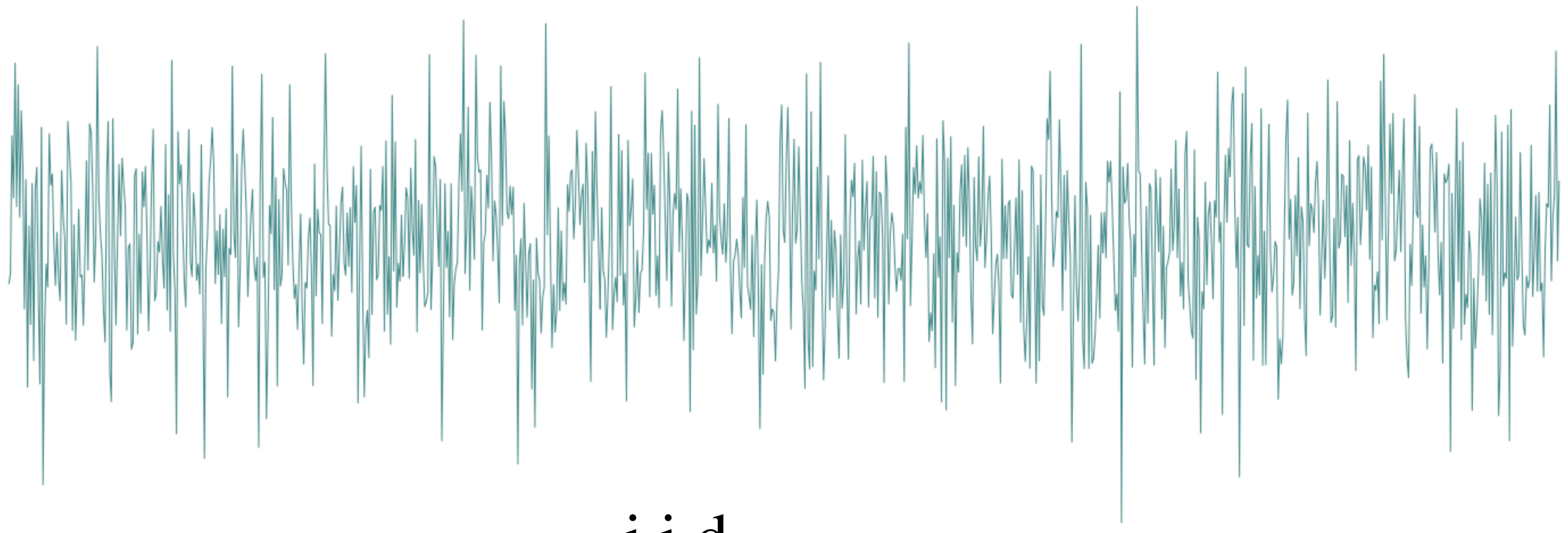


$Y_t = \log t + \varepsilon_t$

Auto-regressive process of order 1

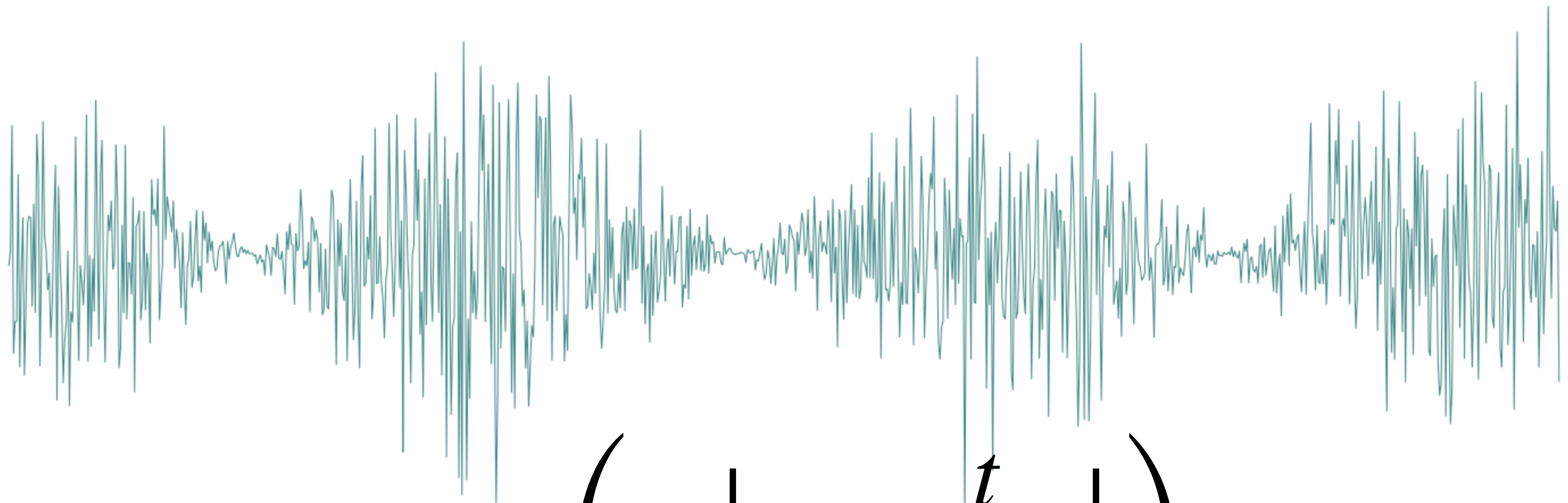
$Z_t \ = \ 0.95 \times Z_{t-1} + \varepsilon_t$

Examples



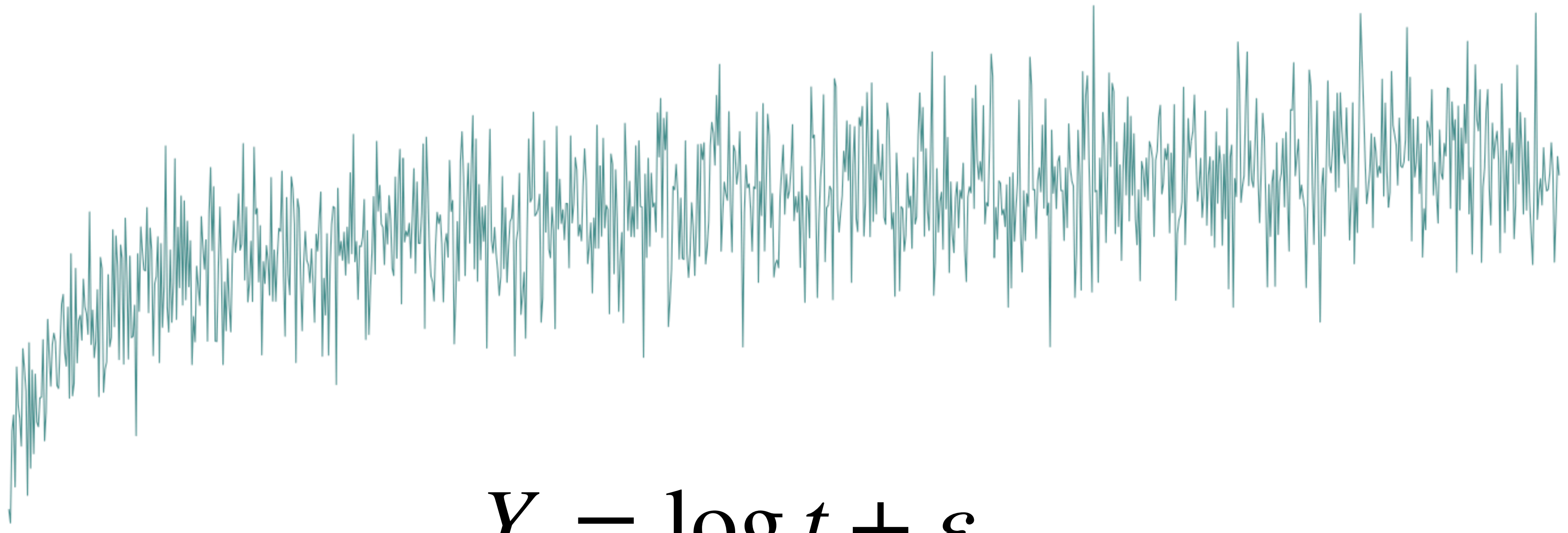
$$\varepsilon_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$$

stationary



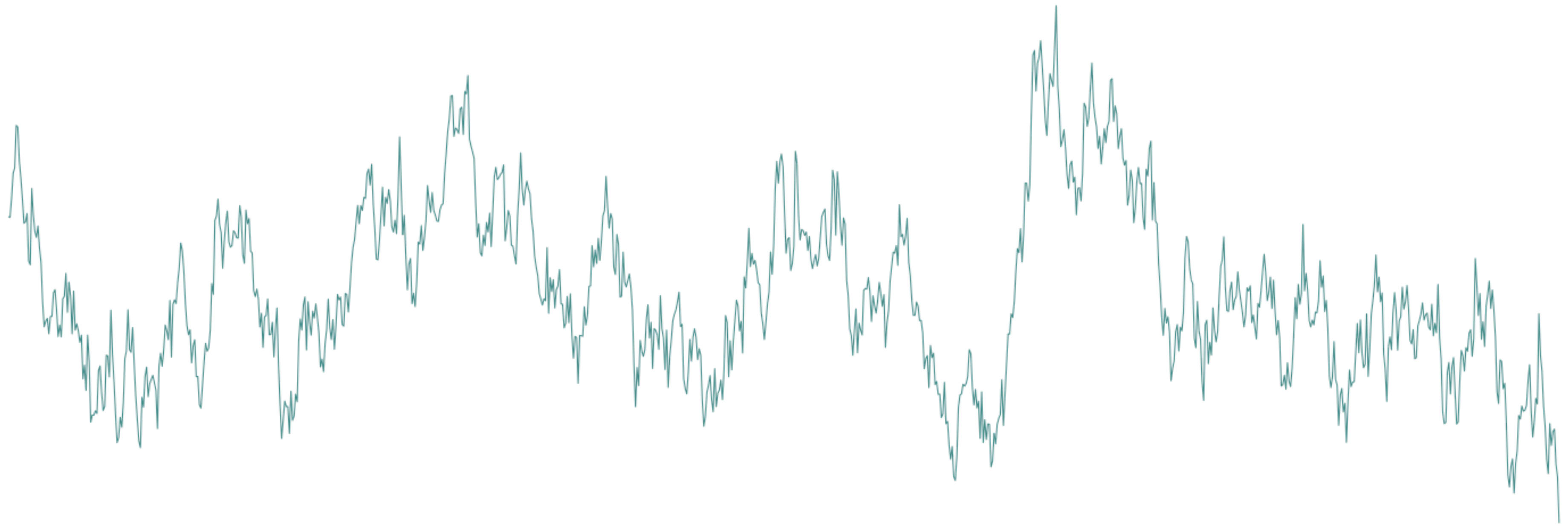
$$X_t \sim \mathcal{N}\left(0, \left|\cos\frac{t}{100}\right|\right)$$

non-stationary: $\mathbb{E}\left[X_t^2\right] = cos^2\frac{t}{100}$



$$Y_t = \log t + \varepsilon_t$$

non-stationary: $\mathbb{E}[Y_t] = \log t$



$$Z_t = 0.95 \times Z_{t-1} + \varepsilon_t$$

stationary

Week-sense stationarity of an AR(1)

Auto-regressive process of order 1 and parameter $|\phi| < 1$ and $\varepsilon_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,\sigma^2)$ a Gaussian white noise:

$$Z_t = \varepsilon_t + \phi Z_{t-1} = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2 Z_{t-2} = \ldots = \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i}$$

- Constant expectation:

$$\forall t \in \mathbb{N}^\star, \mathbb{E}[Z_t] = \sum_{i=0}^{\infty} \phi^i \mathbb{E}\left[\varepsilon_{t-i}\right] = 0$$

- Constant auto-covariance:

$$\forall t \in \mathbb{N}^\star, \forall h \in \mathbb{N}, \operatorname{Cov}(Z_t, Z_{t+h}) = \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i}\right)\left(\sum_{j=0}^{\infty} \phi^j \varepsilon_{t+h-j}\right)\right] = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \phi^{i+j}\mathbb{E}\left[\varepsilon_{t-i}\,\varepsilon_{t+h-j}\right]$$

As $\mathbb{E}\left[\varepsilon_{t-i}\,\varepsilon_{t+h-j}\right] = \sigma^2$ if $j = i + h$ and 0 otherwise and $\sum_{i=0}^{\infty} \phi^{2i+h} = \phi^h \frac{1}{1-\phi^2}$,

$$\operatorname{Cov}(Z_t, Z_{t+h}) = \frac{\phi^h \sigma^2}{1-\phi^2} = \gamma(h)$$

Modelling the deterministic part
-
Trend and seasonality estimation

Moving average

The moving average of bandwidth w (related to the number of observations included in the calculation) is:

$$\bar{Y}_w(t) = \frac{1}{2w+1} \sum_{s=t-w}^{t+w} Y_t$$

It extracts the low-frequency components (trend and if $2w + 1 = \tau$ seasonality)

The greater the window width, the greater the smoothing

Well known in signal theory: it acts like a low-pass filter that eliminates noise.

This estimator is non-parametric, since it assumes no a priori structure on the trend (e.g. linear or polynomial).

Parametric models

Once we have observed the time series well, it is often possible to infer a parametric representation of the trend and seasonality:

- Linear Regression
- Generalised additive models …

Example: we assume that $Y_t = at^2 + b\cos\frac{2\pi t}{\tau} + \varepsilon_t$ with a and b some unknown parameters

With the matrix notation $X = \begin{bmatrix} 1 & \cos\frac{2\pi}{\tau} \\ \vdots & \vdots \\ T^2 & \cos\frac{2\pi T}{\tau} \end{bmatrix}$ $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_T \end{bmatrix}$ $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{bmatrix}$, we get $Y = X\begin{bmatrix} a \\ b \end{bmatrix} + \varepsilon$

We estimate parameters a and b using Ordinary Least Squares (OLS) estimator: $\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (X^TX)^{-1}X^TY$

Trend - which parametric model?

To rid a series of its trend, we can proceed by differentiation: this works for series with polynomial trend

The differentiation operator Δ is defined as $\Delta(Y_t) = Y_t - Y_{t-1}$ and at an order k: $\Delta^k(Y_t) = \Delta\left(\Delta^{k-1}(Y_t)\right)$

Proposition:

Let Y be a time series with a polynomial trend of order k:

$$Y_t = \sum_{j=0}^{k} a_j t^j + \varepsilon_t$$

then the time series $\Delta(Y_t)$ has a polynomial trend of order $k - 1$

By induction, it is enough to apply k times the differentiation operator in order to obtain a stationary time series and this gives an idea of the parametric model to choose!

Differenciation

Proof:

Using Binomial theorem, we get

$$Y_{t-1} = \sum_{j=0}^{k} a_j (t-1)^j + \varepsilon_{t-1}$$

$$= \sum_{j=0}^{k} a_j \sum_{\ell=0}^{j} (-1)^{j-\ell} \binom{\ell}{j} t^\ell + \varepsilon_{t-1} = a_k t^k + \sum_{j=0}^{k-1} a_j \sum_{\ell=0}^{j} (-1)^{j-\ell} \binom{\ell}{j} t^\ell + \varepsilon_{t-1}$$

So the trend of $\Delta(Y_t) = Y_t - Y_{t-1}$ is polynomial of order $k-1$.

The noise term of the series is $\varepsilon_t - \varepsilon_{t-1}$ is stationary as soon as ε_t is:

$$\mathbb{E}[\varepsilon_t - \varepsilon_{t-1}] = \mu - \mu = 0$$

$$\forall\, h \in \mathbb{N},\ \mathrm{Cov}(\varepsilon_t - \varepsilon_{t-1}, \varepsilon_{t+h} - \varepsilon_{t+h-1}) = \mathbb{E}\left[\varepsilon_t \varepsilon_{t+h}\right] - \mathbb{E}\left[\varepsilon_t \varepsilon_{t+h-1}\right] - \mathbb{E}\left[\varepsilon_{t-1} \varepsilon_{t+h}\right] + \mathbb{E}\left[\varepsilon_{t-1} \varepsilon_{t+h-1}\right]$$

$$= 2\gamma(h) - \gamma(h-1) - \gamma(h+1)$$

Polynomial trend

Once $\hat{k}$ (number of times we applied the differentiation operator before getting a stationary process) has been estimated, we assume that

$$Y_t = \sum_{j=0}^{\hat{k}} a_j t^j + \varepsilon_t$$

With the matrix notation $X = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & 4 & \dots & 2^{\hat{k}} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & T & T^2 & \dots & T^{\hat{k}} \end{bmatrix}$ $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_T \end{bmatrix}$ $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{bmatrix}$,

we get $Y = [a_0 \ a_1 \ a_2 \ \dots, a_{\hat{k}}] \, X + \varepsilon$

We estimate parameters a_j using Ordinary Least Squares (OLS) estimator: $\begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \vdots \\ \hat{a}_{\hat{k}} \end{bmatrix} = \left(X^T X\right)^{-1} X^T Y$

Seasonality

To rid a series of an additive seasonality $Y_t = S_t + \varepsilon_t$, we can proceed by differentiation

With Δ_τ is defined as $\Delta_\tau(Y_t) = Y_t - Y_{t-\tau}$

Proposition:

Let Y be a time series with an additive seasonality of period τ, then the time series $\Delta_\tau(Y_t)$ is stationary

Proof:

$\Delta_\tau(Y_t) = Y_t - Y_{t-\tau} = \varepsilon_t - \varepsilon_{t-\tau}$ because by definition, $S_t = S_{t-\tau}$

Non-parametric models

An underlying parametric model is not always obvious and a classical assumption is:

$$Y_t = f(t) + \varepsilon_t \ ,$$

where f is a smooth function on which no parametric assumptions are made and ε is stationary

A classical approach uses kernel estimators:

Given a kernel $K : \mathbb{R} \to \mathbb{R}$, namely a non-negative symmetric integrable function with $\int_{-\infty}^{+\infty} K(x)\,\mathrm{d}x = 1$, and a bandwidth w, the kernel estimator is:

$$\hat{f}_{K,w}(t) = \frac{\sum_{s=1}^{T} Y_s K\left(\frac{t-s}{w}\right)}{\sum_{s=1}^{T} K\left(\frac{t-s}{w}\right)}$$

Kernel estimators

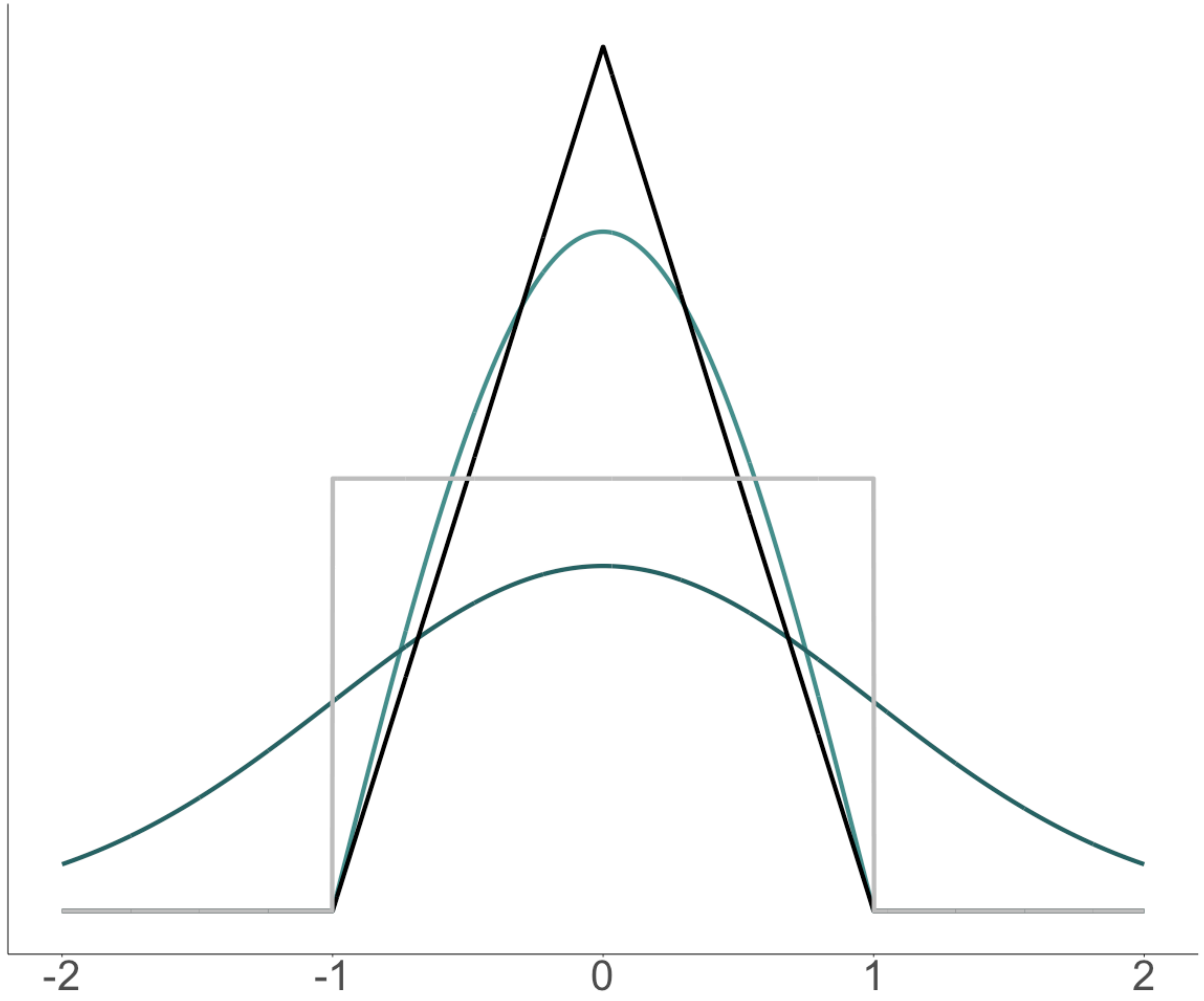
Examples:

Gaussian: $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$

Uniforme: $K(x) = \frac{1}{2}\mathbf{1}_{|x|\leq 1}$

Triangular: $K(x) = \left(1 - |x|\right)\mathbf{1}_{|x|\leq 1}$

Epanechnikov: $K(x) = \frac{3}{4}\left(1 - x^2\right)\mathbf{1}_{|x|\leq 1}$



Kernel estimators - various bandwidth

Kernel estimators

Note that the moving average is none other than the uniform kernel estimator:

$$\hat{f}_{\text{Uniform},w}(t) = \frac{\sum_{s=1}^{T} \frac{1}{2} Y_s \mathbf{1}_{\{|t-s| \leq w\}}}{\sum_{s=1}^{T} \frac{1}{2} \mathbf{1}_{\{|t-s| \leq w\}}} = \frac{1}{2w+1} \sum_{s=t-w}^{t+w} Y_t = \bar{Y}_w(t)$$

Thus, kernel estimators can be seen has weighted moving average.

Modelling the noisy part

-

Residuals analysis

Check stationarity and characterise the noise

Once we have estimated the trend $\hat{T}_t$ and the seasonality $\hat{S}_t$, we can an estimation of the noise part ε_t of the time series, which can be, depending on the times series decomposition:

- if additive: $Y_t = T_t + S_t + \varepsilon_t \rightarrow \hat{\varepsilon}_t = Y_t - \hat{S}_t - \hat{T}_t$
- if multiplicative: $Y_t = T_t \times S_t \times \varepsilon_t \rightarrow \hat{\varepsilon}_t = \frac{Y_t}{\hat{S}_t \times \hat{T}_t}$
- if combination of the two: $Y_t = T_t + S_t \times \varepsilon_t$, e.g. $\rightarrow \hat{\varepsilon}_t = \frac{Y_t - \hat{T}_t}{\hat{S}_t}$

Then, we must check that the times series $\hat{\varepsilon}_t$ is stationary:

- Check moving averages
- Check moving variances
- Fit an ARMA process to predict $\hat{\varepsilon}_t$ (because of Wold's representation theorem)

From now on, we denote by $\epsilon_t = \hat{\varepsilon}_t$ the time series rid of its seasonality and trend

Importance of the Wold's representation

AR(p): $\epsilon_t = \sum_{i=1}^{p} \varphi_i \epsilon_{t-i} + Z_t$, with Z_t a white noise process

MA(q): $\epsilon_t = Z_t + \sum_{i=1}^{q} \theta_i Z_{t-i}$

ARMA(p, q): $\epsilon_t = Z_t + \sum_{i=1}^{p} \varphi_i \epsilon_{t-i} + \sum_{i=1}^{q} \theta_i Z_{t-i}$

The Wold's representation theorem implies that, for any stationary process ε_t can be written as

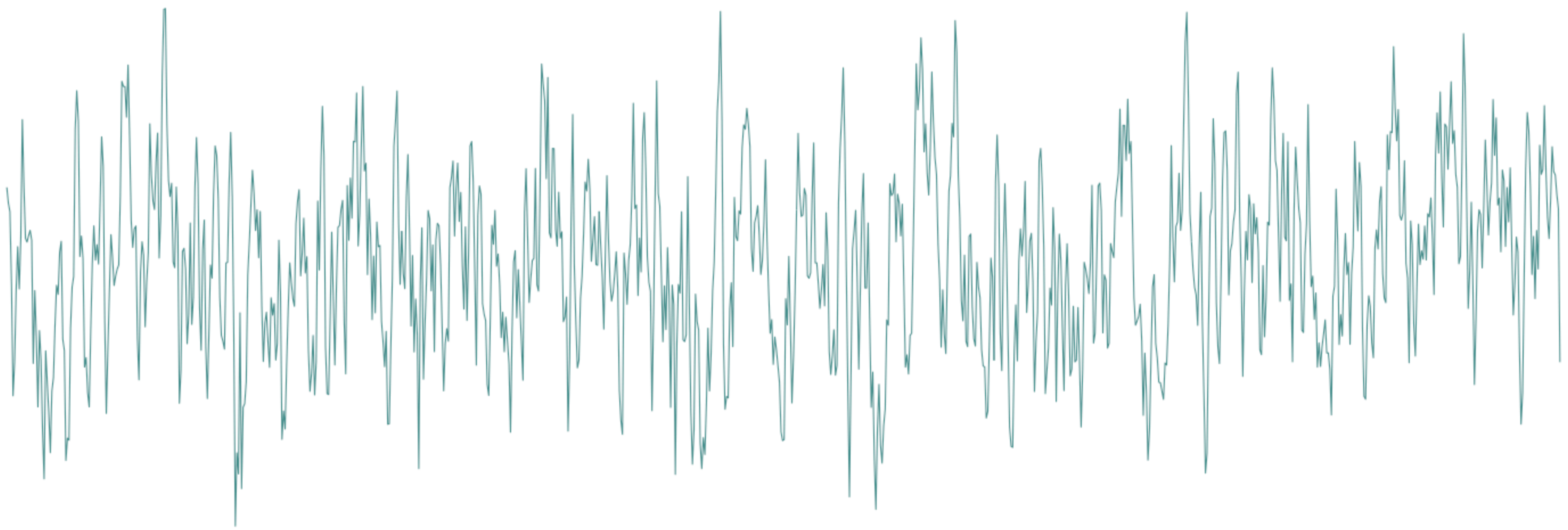
- as a linear combination of a lagged values of a white noise process = MA(∞) representation
- as a linear combination of the lagged values of the process = AR(∞) representation

→ Estimation of a lot of parameters… ARMA models are sparse representations (few no-zero parameters) to approximate the process

How to choose p and q and estimate ϵ_t ?

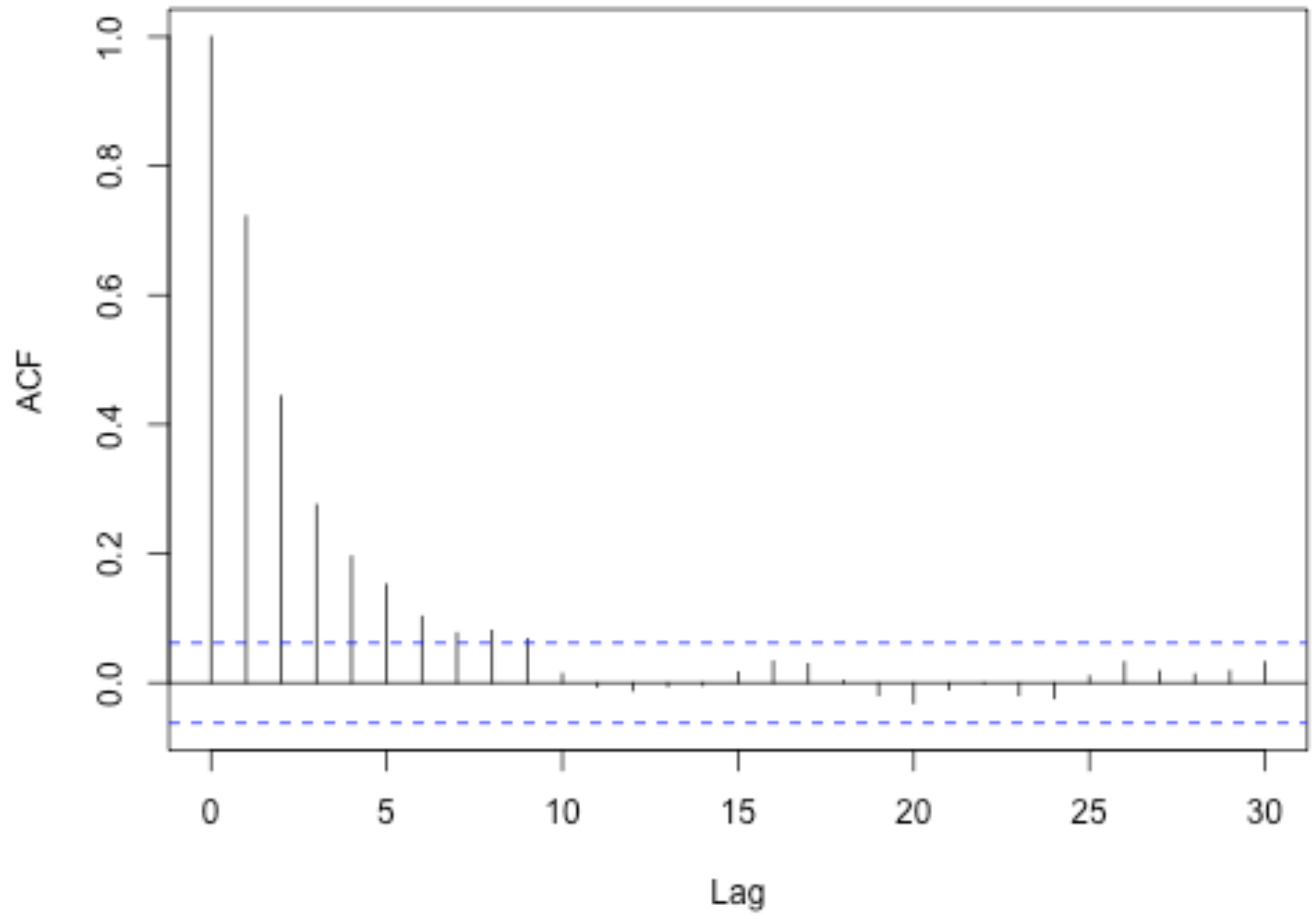
Auto-correlation function (ACF)

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\mathrm{Cov}(\epsilon_t, \epsilon_{t+h})}{\mathrm{Var}(\epsilon_t)} \approx \frac{\frac{1}{n-h}\sum_{t=h+1}^{n}(\epsilon_t - \overline{\epsilon})(\epsilon_{t+h} - \overline{\epsilon})}{\sum_{t=h+1}^{n}(\epsilon_t - \overline{\epsilon})^2}, \text{ with } \overline{\epsilon} = \frac{1}{n}\sum_{t=1}^{n}\epsilon_t$$



Auto-regressive process of order 2:

$$Z_t = 0.9 \times Z_{t-1} - 0.2 \times Z_{t-2} + \varepsilon_t$$



Auto-correlation function (ACF)

Example: MA(1): $\epsilon_t = Z_t + \theta_1 Z_{t-1}$, with Z_t a white noise process of variance σ^2

$$\begin{aligned}
\mathrm{Cov}(\epsilon_t, \epsilon_{t+h}) &= \mathbb{E}\left[(Z_t + \theta_1 Z_{t-1})(Z_{t+h} + \theta_1 Z_{t+h-1})\right] \\
&= \mathbb{E}[Z_t Z_{t+h}] + \theta_1 \mathbb{E}[Z_t Z_{t+h-1}] + \theta_1 \mathbb{E}[Z_{t-1} Z_{t+h}] + \theta_1^2 \mathbb{E}[Z_{t-1} Z_{t+h-1}] \\
&= \sigma^2 \mathbf{1}_{h=0} + \theta_1 \sigma^2 \mathbf{1}_{h=1} + \theta_1 \sigma^2 \mathbf{1}_{h=-1} + \theta_1^2 \sigma^2 \mathbf{1}_{h=0}
\end{aligned}$$

Therefore, $\rho(h) = \begin{cases} 1 & \text{if} \quad h = 0 \\ \frac{\theta_1}{1+\theta_1^2} & \text{if} \quad h \pm 1 \\ 0 & \text{else} \end{cases}$

Proposition

If the time series $(\epsilon_t)_t$ is a $MA(q)$ process, its auto-correlation function satisfies

$\forall h > q, \ \rho(h) = 0$

Partial auto-correlation function (PACF)

$$r(h) = \mathrm{Corr}\left(\epsilon_t - \mathrm{P}_{\epsilon_{t+1},\ldots,\epsilon_{t+h-1}}(\epsilon_t), \epsilon_{t+h} - \mathrm{P}_{\epsilon_{t+1},\ldots,\epsilon_{t+h-1}}(\epsilon_{t+h}) \right)$$

where $\mathrm{P}_{X_1,\ldots,X_h}(Y) \in \underset{X=\sum_{i=1}^{h} \alpha_i X_i \mid (\alpha_1,\ldots,\alpha_h)\mathbb{R}^h}{\mathrm{argmin}} \mathbb{E}\left[\left(Y - X\right)^2\right]$

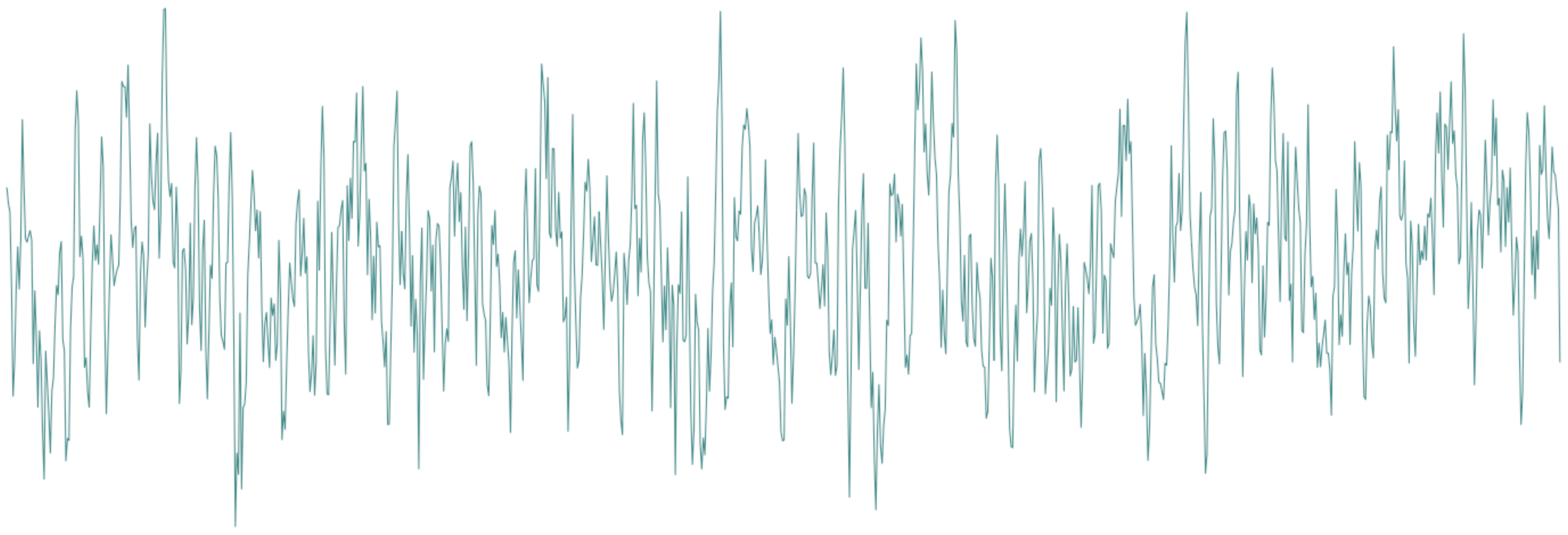
is the orthogonal projection of Y_t over the space generated by

$Y_{t+1}, \ldots, Y_{t+h-1}$ for the distance $d(X, Y) = \sqrt{\mathbb{E}[(X - Y)^2]}$

Other formulation: $r(h) = \mathrm{Corr}\left(\epsilon_t, \epsilon_{t+h} \mid \epsilon_{t+1}, \ldots, \epsilon_{t+h-1}\right)$

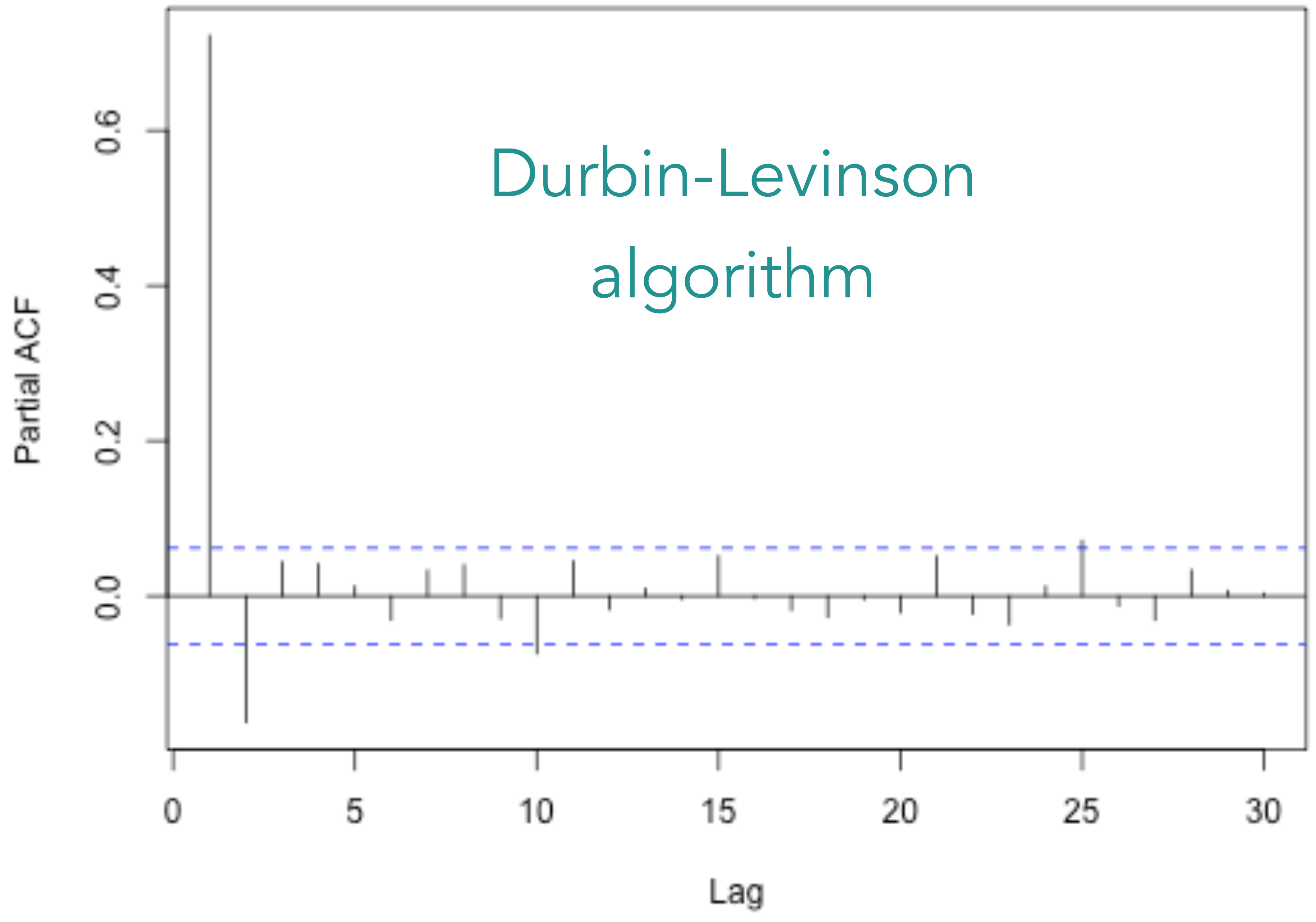
Partial auto-correlation function (PACF)

Idea: $Y_t - \mathrm{P}_{Y_{t+1},\ldots,Y_{t+h-1}}(Y_t)$ is the part of Y_t independent of the realisations of Y which occur between $t+1$ and $t+h-1$ so $r(h)$ measures the « pure » correlation between Y_t and Y_{t+h}, eliminating correlations with realisations that took place between these two observations



Auto-regressive process of order 2:

$$Z_t \;=\; 0.9 \times Z_{t-1} - 0.2 \times Z_{t-2} + \varepsilon_t$$



Partial auto-correlation function (PACF)

Example: $AR(1)$: $\epsilon_t = Z_t + \varphi_1 \epsilon_{t-1}$, with Z_t a white noise process of variance σ^2

- $r(0) = 1$
- $r(1) = \text{Corr}(\epsilon_1, \epsilon_{t+1}) = \varphi_1$
- For $h \geq 2$, $\mathrm{P}_{\epsilon_{t+1},\ldots,\epsilon_{t+h-1}}\left(\epsilon_t\right) = \frac{1}{\varphi_1}\epsilon_{t+1}$ and $\mathrm{P}_{\epsilon_{t+1},\ldots,\epsilon_{t+h-1}}\left(\epsilon_{t+h}\right) = \varphi_1 \epsilon_{t+h-1}$ so

$$r(h) = \text{Corr}\left(\frac{1}{\varphi_1} Z_{t+1}, Z_{t+h}\right) = 0$$

Proposition

If the time series $(\epsilon_t)_t$ is a $AR(p)$ process, its partial auto-correlation function satisfies $\forall h > p, \; r(h) = 0$

Estimation of the ARMA processes

Choosing p and q

	Auto-correlation function	Partial auto-correlation function
AR(p)	Decreases to 0	0 if h>p
MA(q)	0 if h>q	Decreases to 0
ARMA(p,q)	Decreases to 0 for h>q	Decreases to 0 for h>p

Estimating coefficients

- Yule-Walker equations for pure AR model
- Least squares regression
- Maximum likelihood estimation

Final prediction of the time series

Once the ARMA process has been estimated, if we observe $\epsilon_1, \ldots, \epsilon_{t-1}$, it is possible to predict ϵ_t with

$$\hat{\epsilon}_t = \sum_{i=1}^{\hat{p}} \hat{\varphi}_i \epsilon_{t-i} + \sum_{i=1}^{\hat{q}} \hat{\theta}_i Z_{t-i}$$

To access to $Z_1, \ldots, Z_{t-1}$ we may use the $AR(\infty)$ representation of the process and approximate them the start of the series

Once the trend $\hat{T}_t$ and the seasonality $\hat{S}_t$, and the ARMA process (i.e $\hat{\varphi}_1, \ldots, \hat{\varphi}_{\hat{p}}$ and $\hat{\theta}_1, \ldots \hat{\theta}_{\hat{q}}$)

- if additive: $Y_t = T_t + S_t + \varepsilon_t \; \rightarrow \hat{Y}_t = \hat{T}_t + \hat{S}_t + \hat{\epsilon}_t$
- if multiplicative: $Y_t = T_t \times S_t \times \varepsilon_t \rightarrow \hat{Y}_t = \hat{T}_t \times \hat{S}_t \times \hat{\epsilon}_t$
- if combination of the two: $Y_t = T_t + S_t \times \varepsilon_t$, e.g. $\rightarrow \hat{Y}_t = \hat{T}_t + \hat{S}_t \times \hat{\epsilon}_t$

Remark: offline predictions $\rightarrow \hat{\epsilon}_t = 0$

Validation

To validate the final modelling, it is crucial to analyse residuals $\hat{Z}_t = Y_t - \hat{Y}_t$

- White noise: portemanteau test (uses Ljung–Box statistic):

  Under the white noise hypotheses, with n is the sample size, $\hat{\rho}_k$ the autocorrelation at lag k, and h the number of lags being tested: $n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \sim \chi_h^2$

- Heteroskedasticity: Test the absence of autoregressive conditional heteroskedasticity (ARCH - model that describes the variance of the time series) components

- Normality: no skewness nor kurtosis: Jarque–Bera test

Other approaches

ARIMA and SARIMA

Autoregressive integrated moving average (ARIMA) models generalise ARMA models for non-stationarity in the sense of mean (but not variance) time series: a differencing step (« integrated » part of the model) can be applied one or more times to eliminate the non-stationarity of the trend

→ $\text{ARIMA}(p, d, q)$ is suitable for modelling a time series with a polynomial trend of degrees d

p = Trend autoregression order

d = Trend difference order

q = Trend moving average order

Seasonal Autoregressive Integrated Moving Average (SARIMA) extension of ARIMA models explicitly model the seasonality of the time series using four new parameters:

- P = Seasonal autoregressive order
- D = Seasonal difference order
- Q = Seasonal moving average order
- m = The number of time steps for a single seasonal period

Exponential smoothing

Back to 1940s (signal processing) /1950s (in statistics with Brown and Holt) - no theoretical guarantees)

The simplest exponential smoothing $\tilde{Y}_t$ of the time series Y_t is

$$\tilde{Y}_t = \alpha \tilde{Y}_{t-1} + (1-\alpha) Y_t \text{ , with } \alpha \in [0,1]$$

It may be use to predict Y_{t+1}:

$$\hat{Y}_{t+1} = \sum_{s=1}^{t} \alpha (1-\alpha)^s Y_{t-s} \quad \text{(nice benchmark!)}$$

The closer α is to 1, the more memory the smoothing has, conversely, if α is close to 0, the past values of the time series are quickly forgotten

→ Estimation of α on training data

Other approaches:

- Double exponential smoothing - Holt linear
- Triple exponential smoothing - Holt Winters

That's all folks!