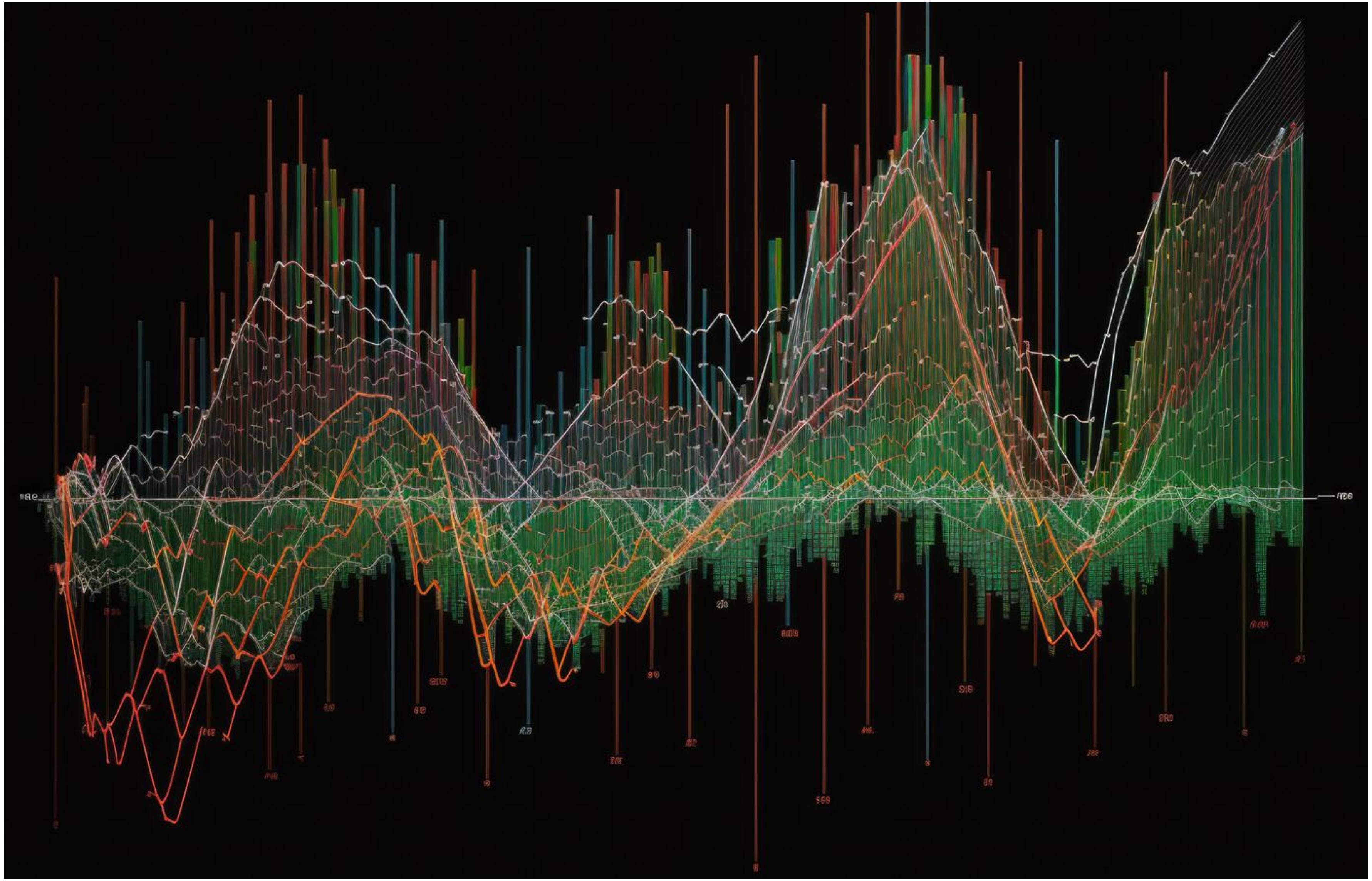
Part III: Time Series Models



Outline

- **ARIMA Models:**
  - AR Process
  - Backshift Operator
  - MA Process
  - ARMA Models

- **Stationarize Nonstationary Time Series**
  - Differencing and Deterending

- **Stationarity Testing:**
  - KPSS test for stationarity

ARIMA Models

- ARIMA is an acronym that stands for **A**uto-**R**egressive **I**ntegrated **M**oving **A**verage. Specifically,
  - **AR** *Autoregression.* A model that uses the dependent relationship between an observation and some number of lagged observations.
  - **I** *Integrated.* The use of differencing of raw observations in order to make the time series stationary.
  - **MA** *Moving Average.* A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.
- Each of these components are explicitly specified in the model as a parameter.
- Note that **AR** and **MA** are two widely used linear models that work on stationary time series, and **I** is a preprocessing procedure to "stationarize" time series if needed.

Notations

- A standard notation is used of $ARIMA(p, d, q)$ where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.
  - **p** The number of lag observations included in the model, also called the lag order.
  - **d** The number of times that the raw observations are differenced, also called the degree of differencing.
  - **q** The size of the moving average window, also called the order of moving average.
- A value of 0 can be used for a parameter, which indicates to not use that element of the model.
- In other words, ARIMA model can be configured to perform the function of an ARMA model, and even a simple AR, I, or MA model.

Autoregressive (AR) Models

- **Intuition**
  - Autoregressive models are based on the idea that current value of the series, X_t, can be explained as a linear combination of p past values, $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$, together with a random error in the same series.
- **Definition**
  - An autoregressive model of order p, abbreviated $AR(p)$, is of the form

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + w_t = \sum_{i=1}^{p} \phi_i X_{t-i} + w_t$$

where X_t is stationary, $w_t \sim wn(0, \sigma_w^2)$, and $\phi_1, \phi_2, \ldots, \phi_p$ ($\phi_p \neq 0$) are model parameters. The hyperparameter p represents the length of the "direct look back" in the series.

Backshift Operator

- Before we dive deeper into the AR process, we need some new notations to simplify the representations.
- **Backshift Operator**
  - The backshift operator is defined as

$$BX_t = X_{t-1}.$$

It can be extended, $B^2X_t = B(BX_t) = B(X_{t-1}) = X_{t-2}$, and so on. Thus,

$$B^kX_t = X_{t-k}$$

- We can also define an inverse operator (*forward-shift operator*) by enforcing $B^{-1}B = 1$, such that

$$X_t = B^{-1}BX_t = B^{-1}X_{t-1}.$$

Autoregressive Operator of AR Process

- Recall the definition for $AR(p)$ process:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + w_t$$

By using the backshift operator we can rewrite it as:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \cdots - \phi_p X_{t-p} = w_t$$
$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) X_t = w_t$$

- The autoregressive operator is defined as:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p = 1 - \sum_{j=1}^{p} \phi_j B^j,$$

then the $AR(p)$ can be rewritten more concisely as:

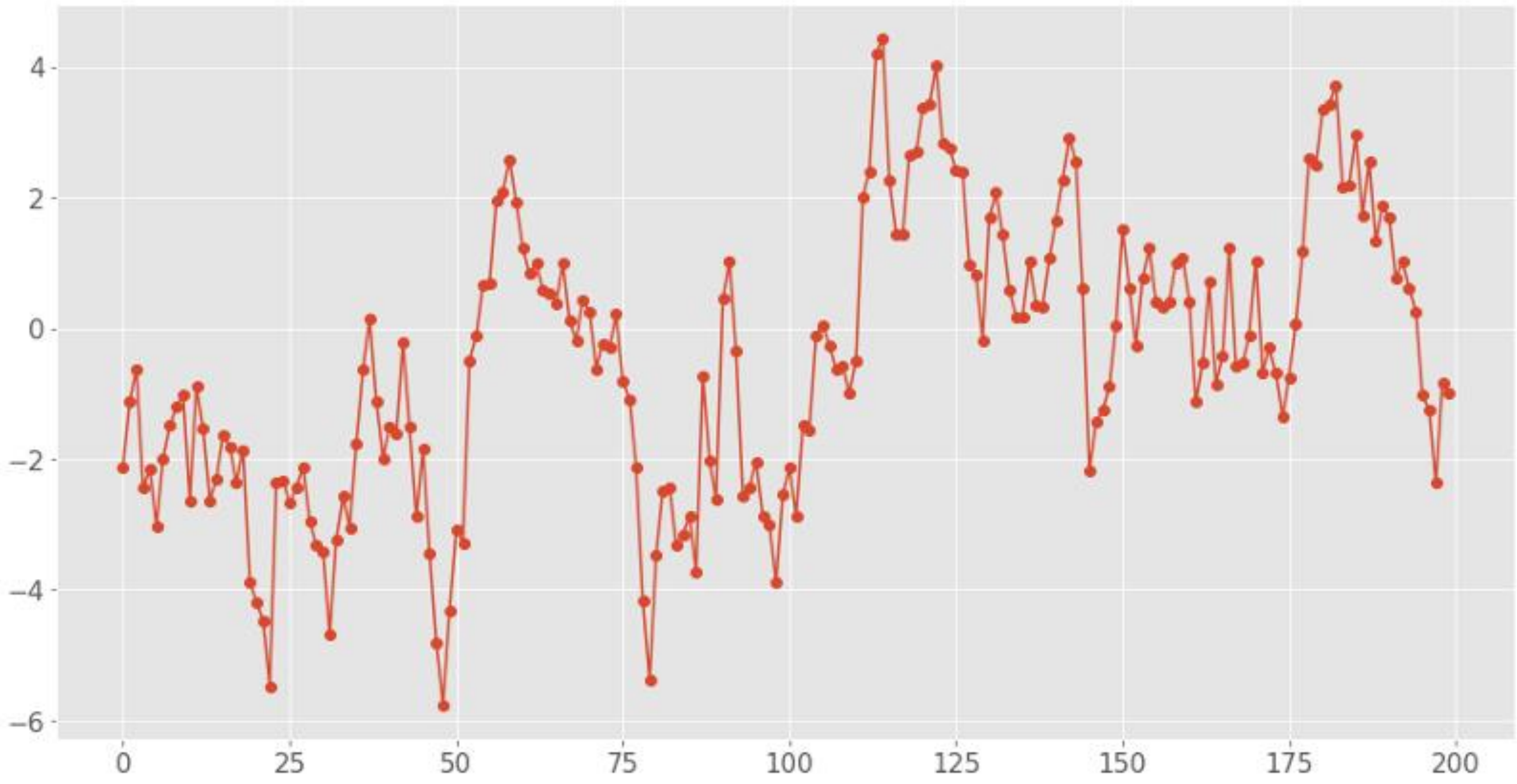
$$\boxed{\phi(B) X_t = w_t}$$

AR Example: AR(0) and AR(1)

- The simplest AR process is $AR(0)$, which has no dependence between the terms. In fact, $AR(0)$ is essentially white noise.
- $AR(1)$ can be given by $X_t = \phi_1 X_{t-1} + w_t$.
  - Only the previous term in the process and the noise term contribute to the output.
  - If $|\phi_1|$ is close to 0, then the process still looks like white noise.
  - If $\phi_1 < 0$, X_t tends to oscillate between positive and negative values.
  - If $\phi_1 = 1$ then the process is equivalent to random walk, which is not stationary as the variance is dependent on t (and infinite).

AR Examples: AR(1) Process

- Simulated $AR(1)$ Process $X_t = 0.9X_{t-1} + w_t$:

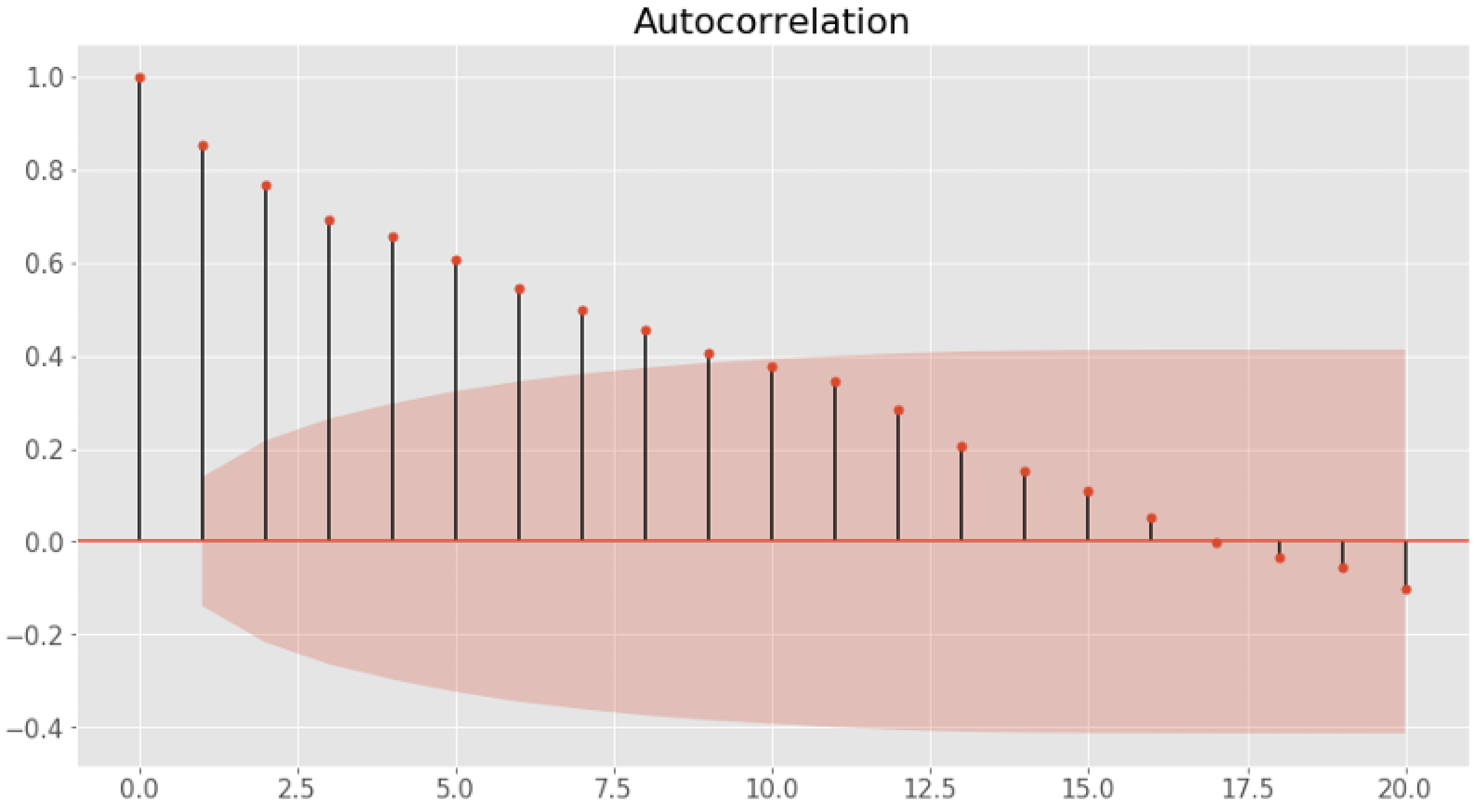


- **Mean** $E[X_t] = 0$
- **Variance** $\mathrm{Var}(X_t) = \dfrac{\sigma_w^2}{(1 - \phi_1^2)}$

AR Examples: AR(1) Process

- **Autocorrelation Function (ACF)**

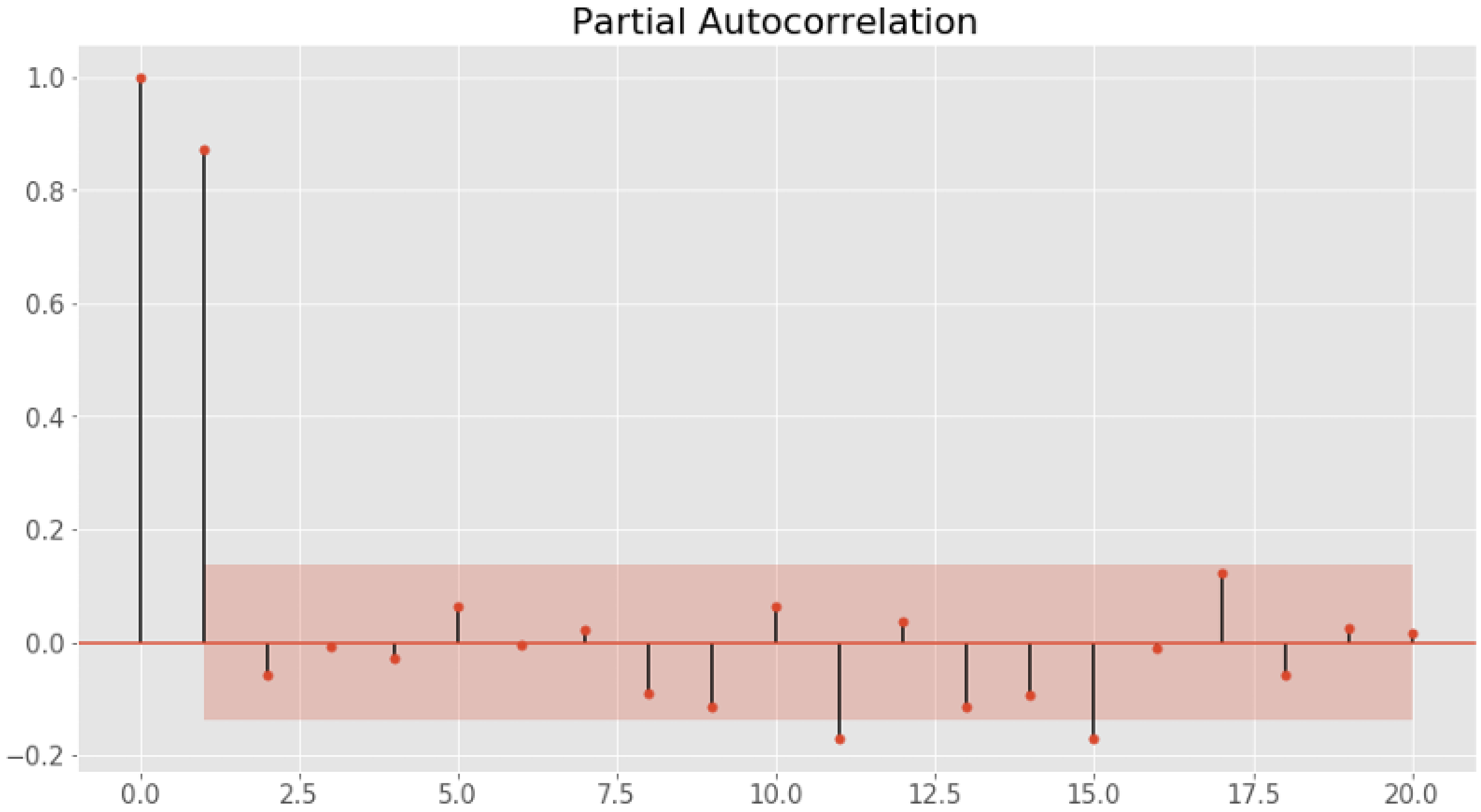
$$\rho_h = \phi_1^h$$



AR Examples: AR(1) Process

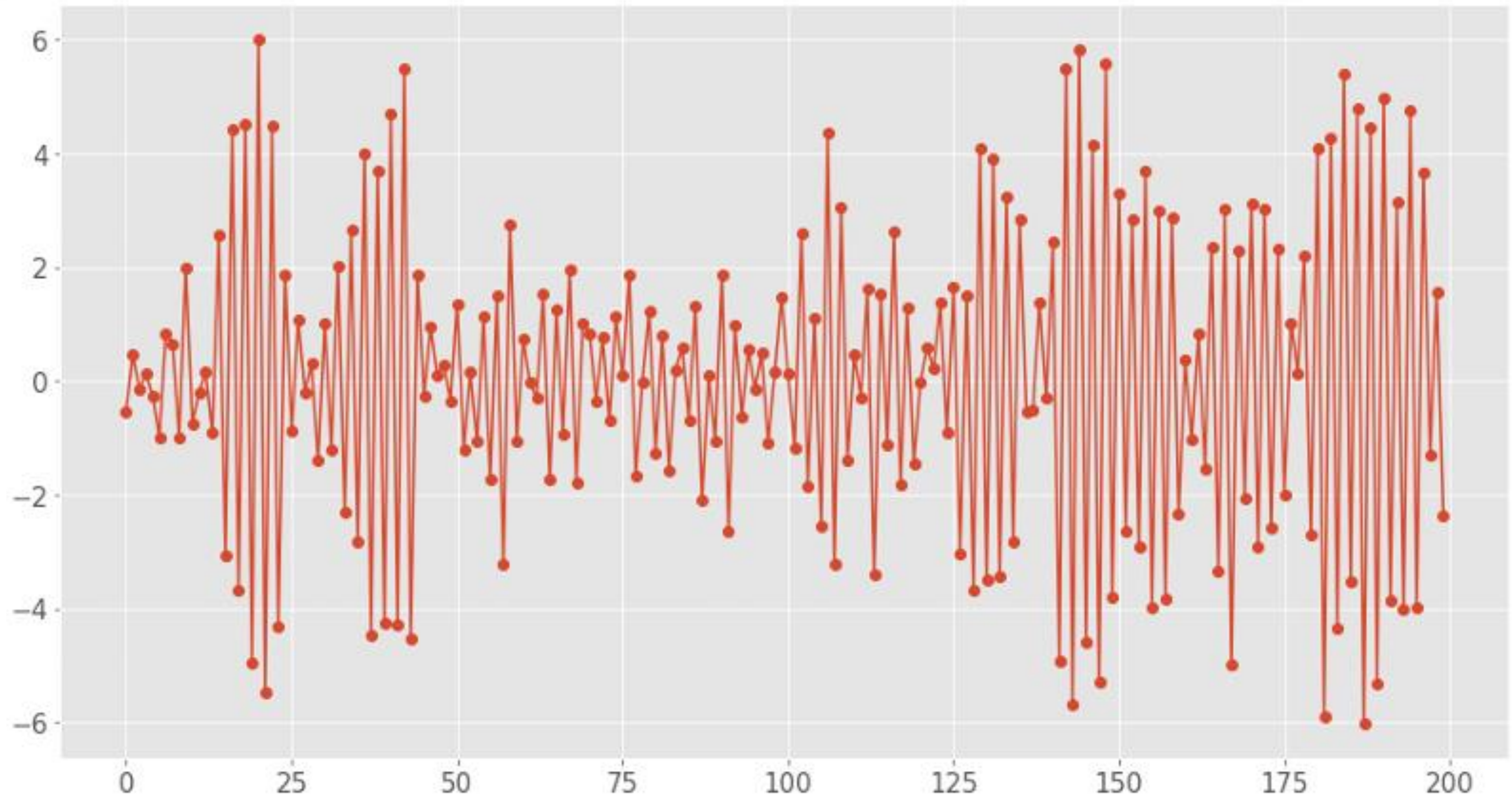
- **Partial Autocorrelation Function (PACF)**

$$\phi_{11} = \rho_1 = \phi_1 \qquad \phi_{hh} = 0, \forall h \geq 2$$



AR Examples: AR(1) Process

- Simulated $AR(1)$ Process $X_t = -0.9X_{t-1} + w_t$:

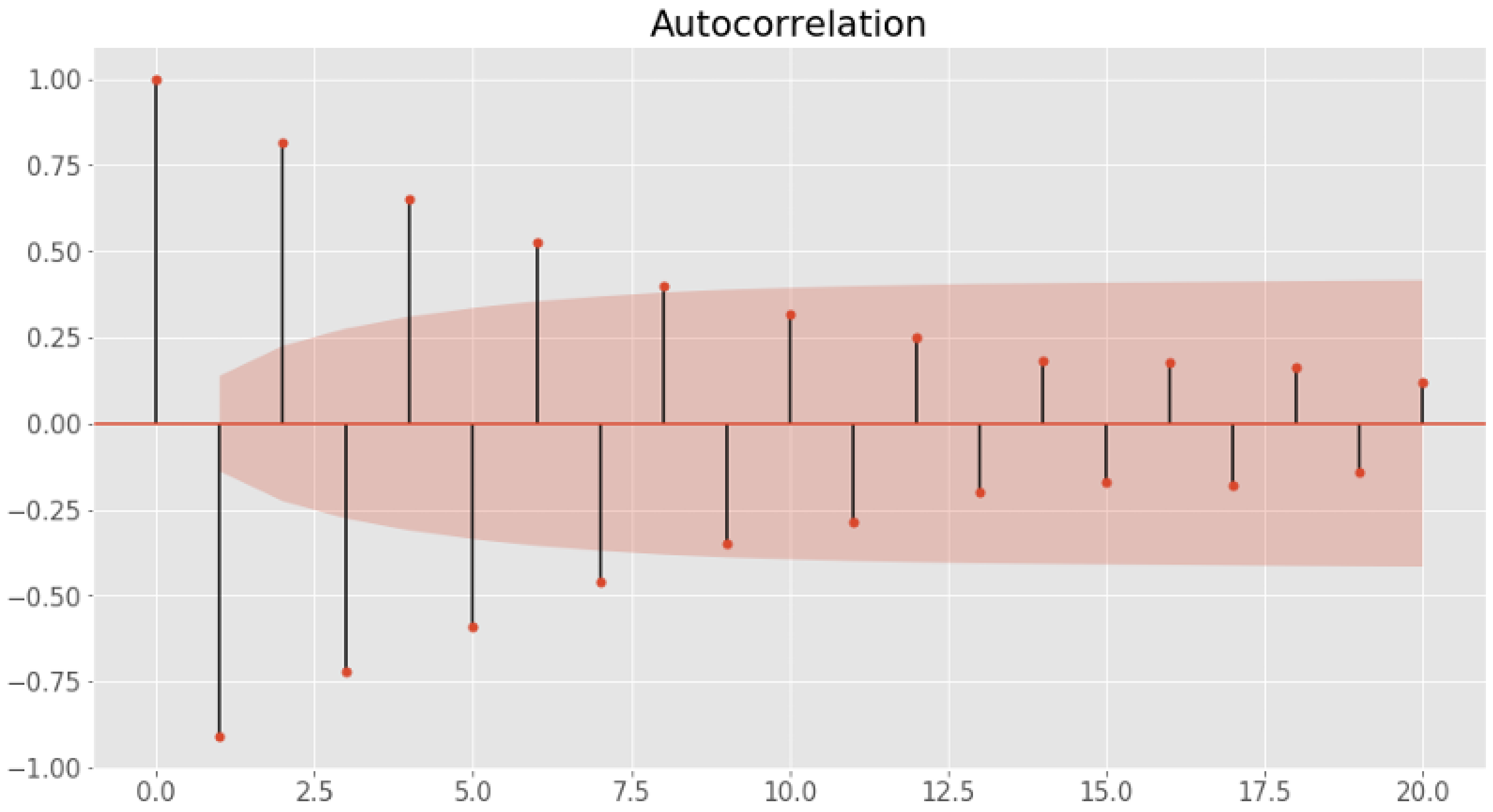


- **Mean** $E[X_t] = 0$
- **Variance** $\mathrm{Var}(X_t) = \dfrac{\sigma_w^2}{(1-\phi_1^2)}$

AR Examples: AR(1) Process

- **Autocorrelation Function (ACF)**

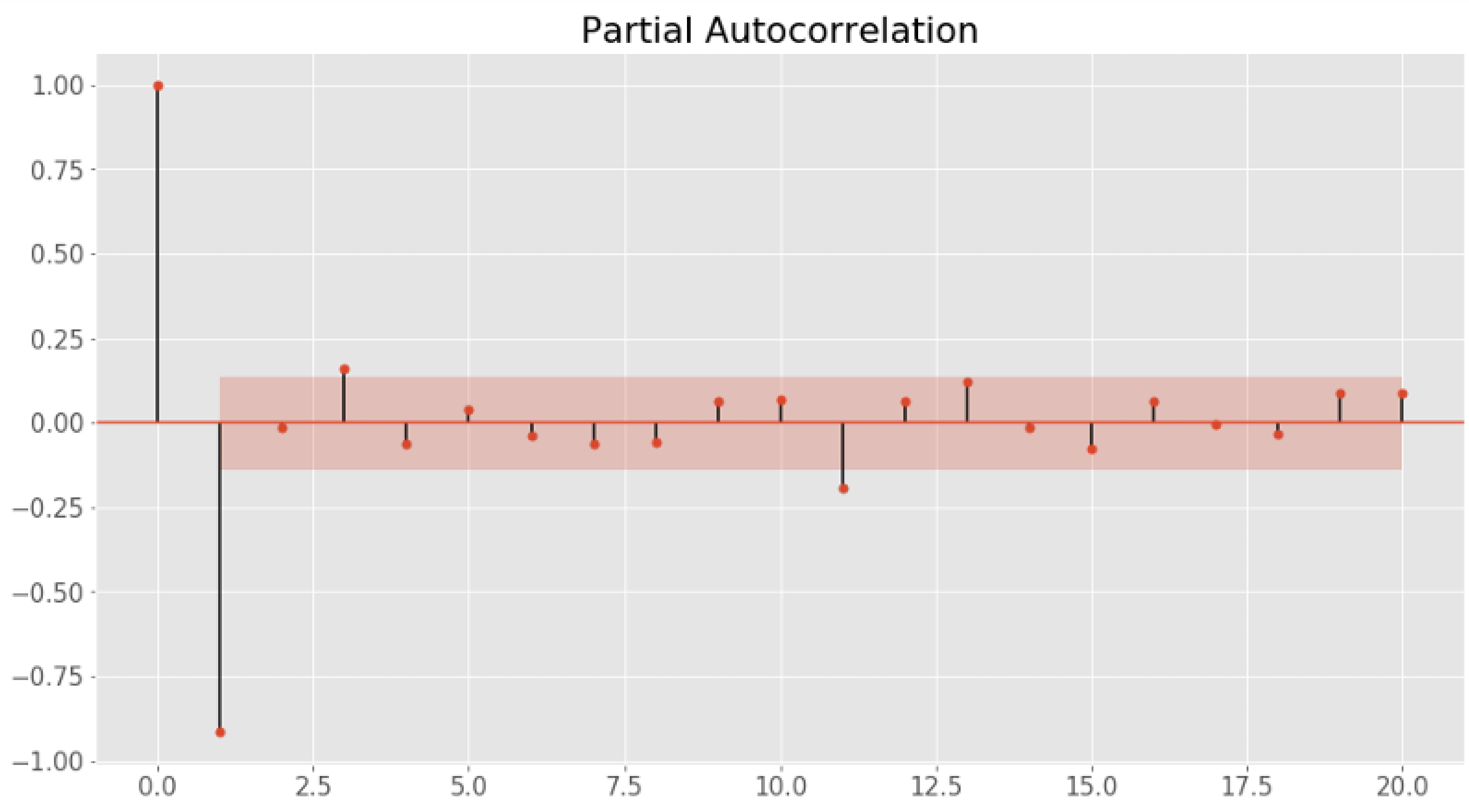
$$\rho_h = \phi_1^h$$



AR Examples: AR(1) Process

- **Partial Autocorrelation Function (PACF)**

$$\phi_{11} = \rho_1 = \phi_1 \qquad \phi_{hh} = 0, \forall h \geq 2$$



General AR(p) Process

- An important property of $AR(p)$ models in general is
  - When $h > p$, theoretical partial autocorrelation function is 0:

$$\phi_{hh} = \text{corr}(X_{t+h} - \hat{X}_{t+h}, X_t - \hat{X}_t) = \text{corr}(w_{t+h}, X_t - \hat{X}_t) = 0.$$

  - When $h \leq p$, ϕ_{pp} is not zero and $\phi_{11}, \phi_{22}, \ldots, \phi_{h-1,h-1}$ are not necessarily zero.
- In fact, identification of an AR model is often best done with the PACF.

AR Models: Parameters Estimation

- Note that p is like a hyperparameter for the $AR(p)$ process, thus fitting an $AR(p)$ model presumes p is known and only focusing on estimating coefficients, i.e. $\phi_1, \phi_2, \ldots, \phi_p$.
- There are many feasible approaches:
  - Method of moments estimator (e.g. Yule-Walker estimator)
  - Maximum Likelihood Estimation (MLE) estimator
  - Ordinary Least Squares (OLS) estimator
- If the observed series is short or the process is far from stationary, then substantial differences in the parameter estimations from various approaches are expected.

Moving Average Models (MA)

- The name might be misleading, but moving average models should not be confused with the moving average smoothing.
- **Motivation**
  - Recall that in AR models, current observation X_t is regressed using the previous observations $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$, plus an error term w_t at current time point.
  - One problem of AR model is the ignorance of correlated noise structures (which is unobservable) in the time series.
  - In other words, the imperfectly predictable terms in current time, w_t, and previous steps, $w_{t-1}, w_{t-2}, \ldots, w_{t-q}$, are also informative for predicting observations.

Moving Average Models (MA)

- **Definition**
  - A moving average model of order q, or $MA(q)$, is defined to be

$$X_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q} = w_t + \sum_{j=1}^{q} \theta_j w_{t-j}$$

    where $w_t \sim wn(0, \sigma_w^2)$, and $\theta_1, \theta_2, \ldots, \theta_q$ $(\theta_q \neq 0)$ are parameters.
  - Although it looks like a regression model, the difference is that the w_t is not observable.
- Contrary to AR model, finite MA model is always stationary, because the observation is just a weighted moving average over past forecast errors.

Moving Average Operator

- **Moving Average Operator**
  - Equivalent to autoregressive operator, we define moving average operator as:

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q,$$

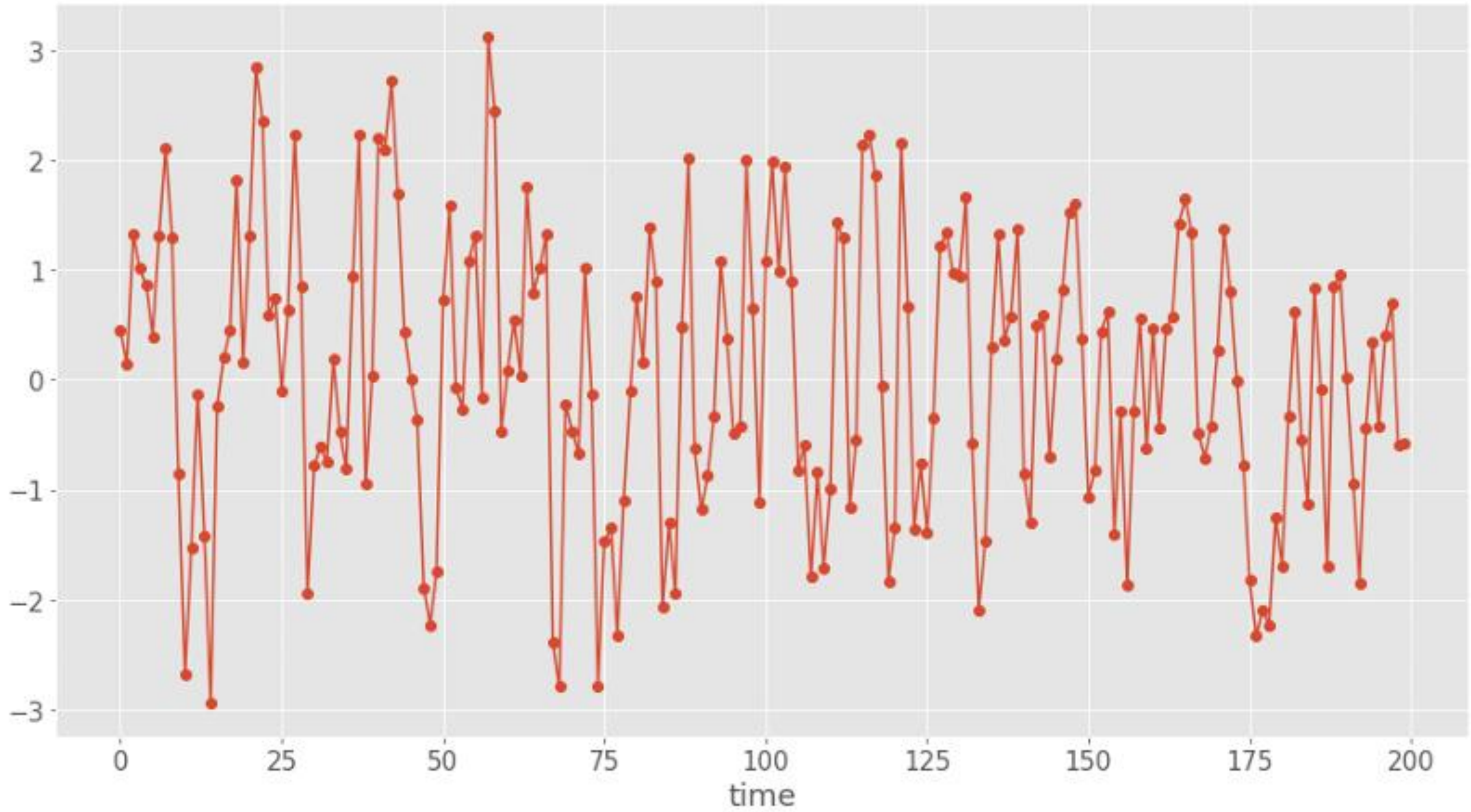
where B stands for backshift operator, thus $B(w_t) = w_{t-1}$.

- Therefore the moving average model can be rewritten as:

$$X_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}$$
$$X_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q) w_t$$
$$\boxed{X_t = \theta(B) w_t}$$

MA Examples: MA(1) Process

- Simulated $MA(1)$ Process $X_t = w_t + 0.8 \times w_{t-1}$:

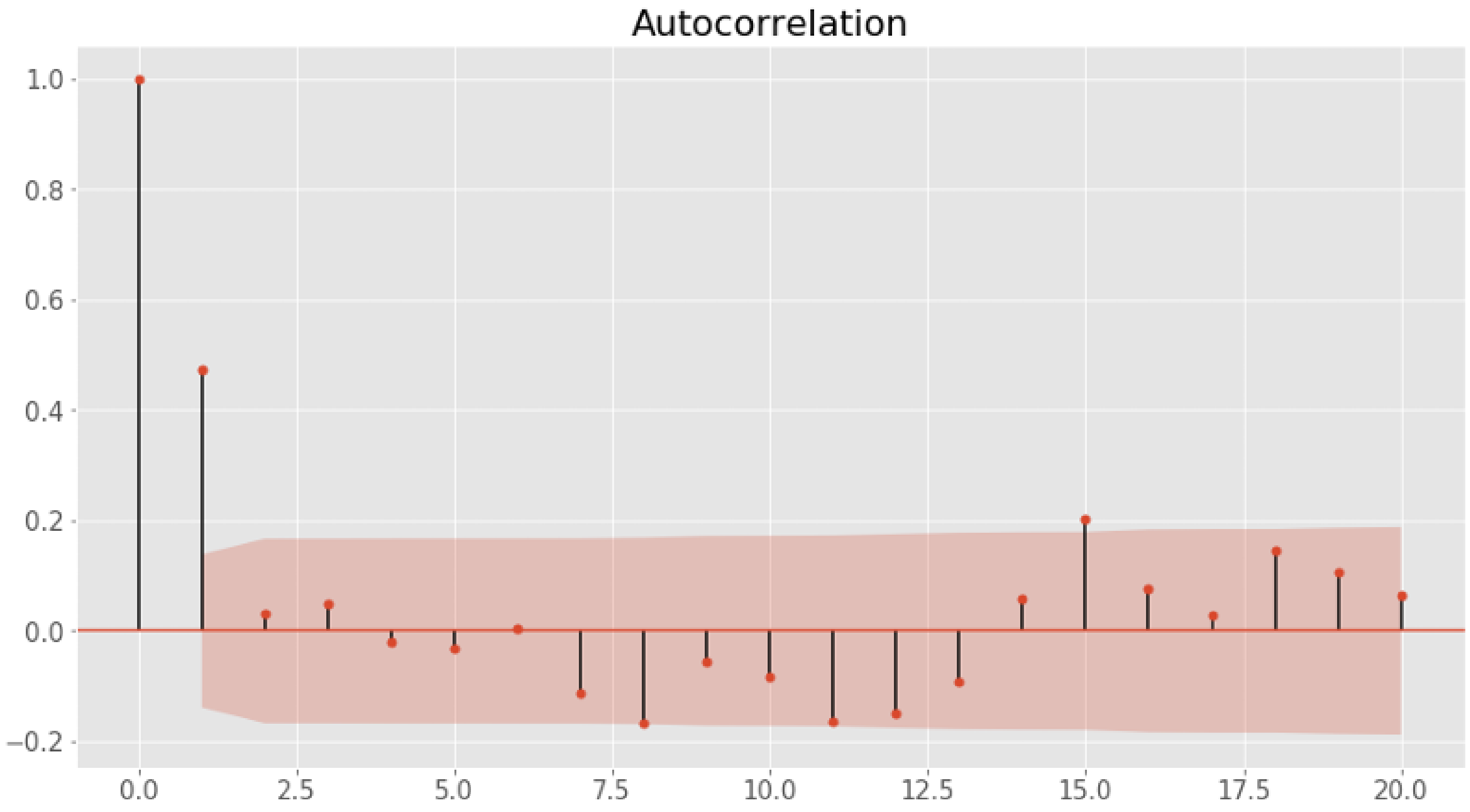


- **Mean** $E[X_t] = 0$
- **Variance** $\mathrm{Var}(X_t) = \sigma_w^2(1 + \theta_1^2)$

MA Examples: MA(1) Process

- **Autocorrelation Function (ACF)**

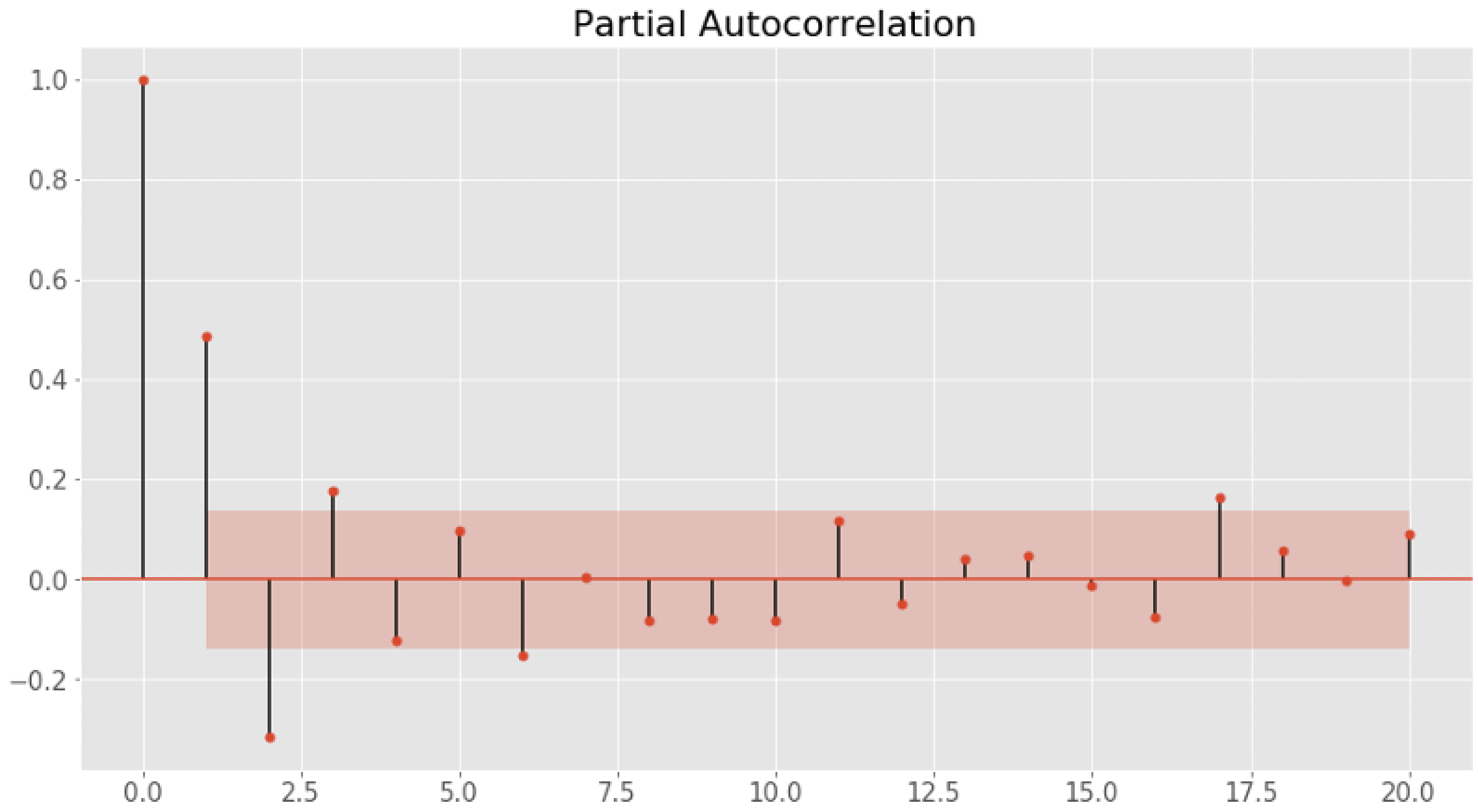
$$\rho_1 = \frac{\theta_1}{1+\theta_1^2} \qquad \rho_h = 0, \forall h \geq 2$$



MA Examples: MA(1) Process

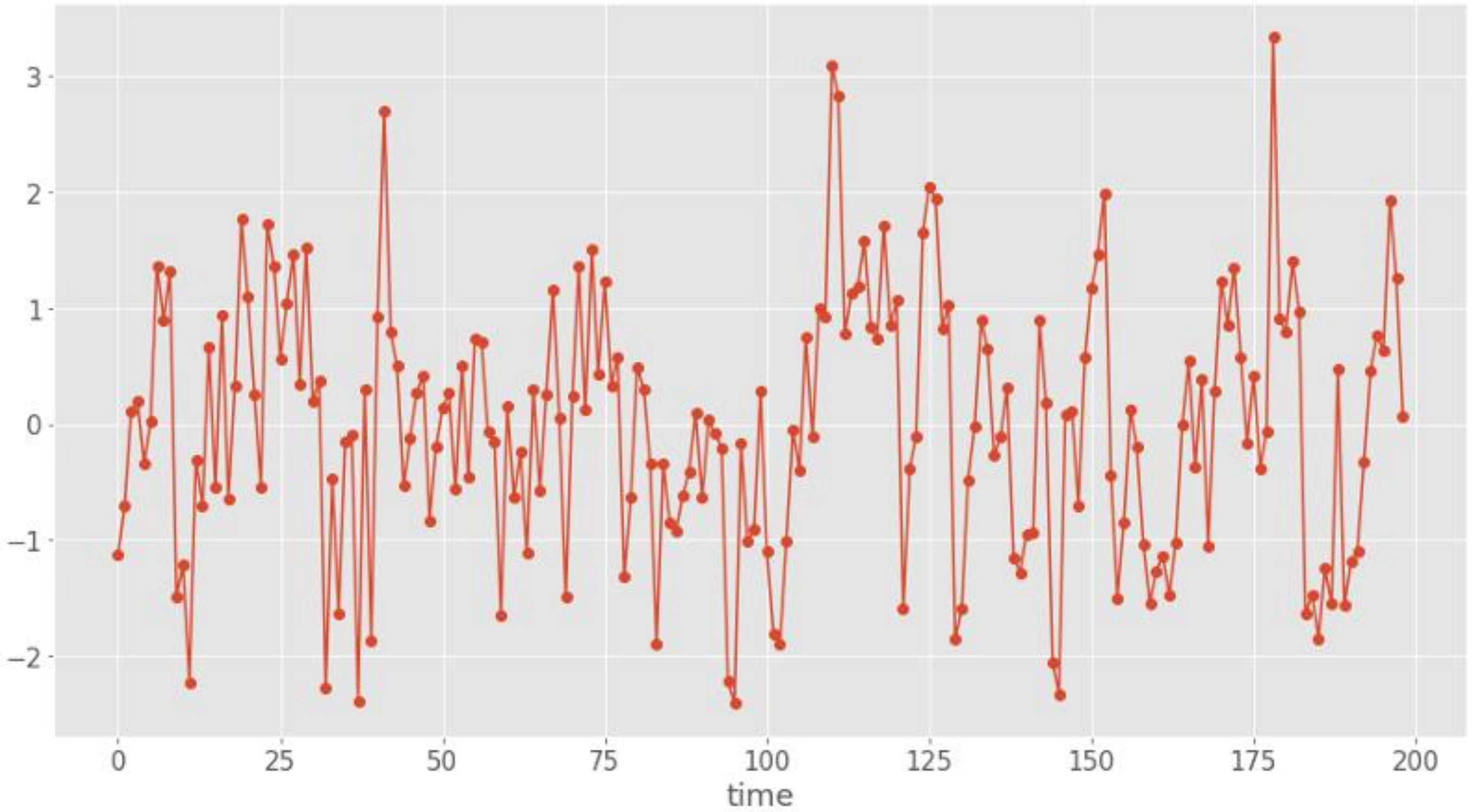
- **Partial Autocorrelation Function (PACF)**

$$\phi_{hh} = -\frac{(-\theta_1)^h(1-\theta_1^2)}{1-\theta_1^{2(h+1)}}, \qquad h \geq 1$$



MA Examples: MA(2) Process

- Simulated $MA(2)$ Process $X_t = w_t + 0.5 \times w_{t-1} + 0.3 \times w_{t-2}$:

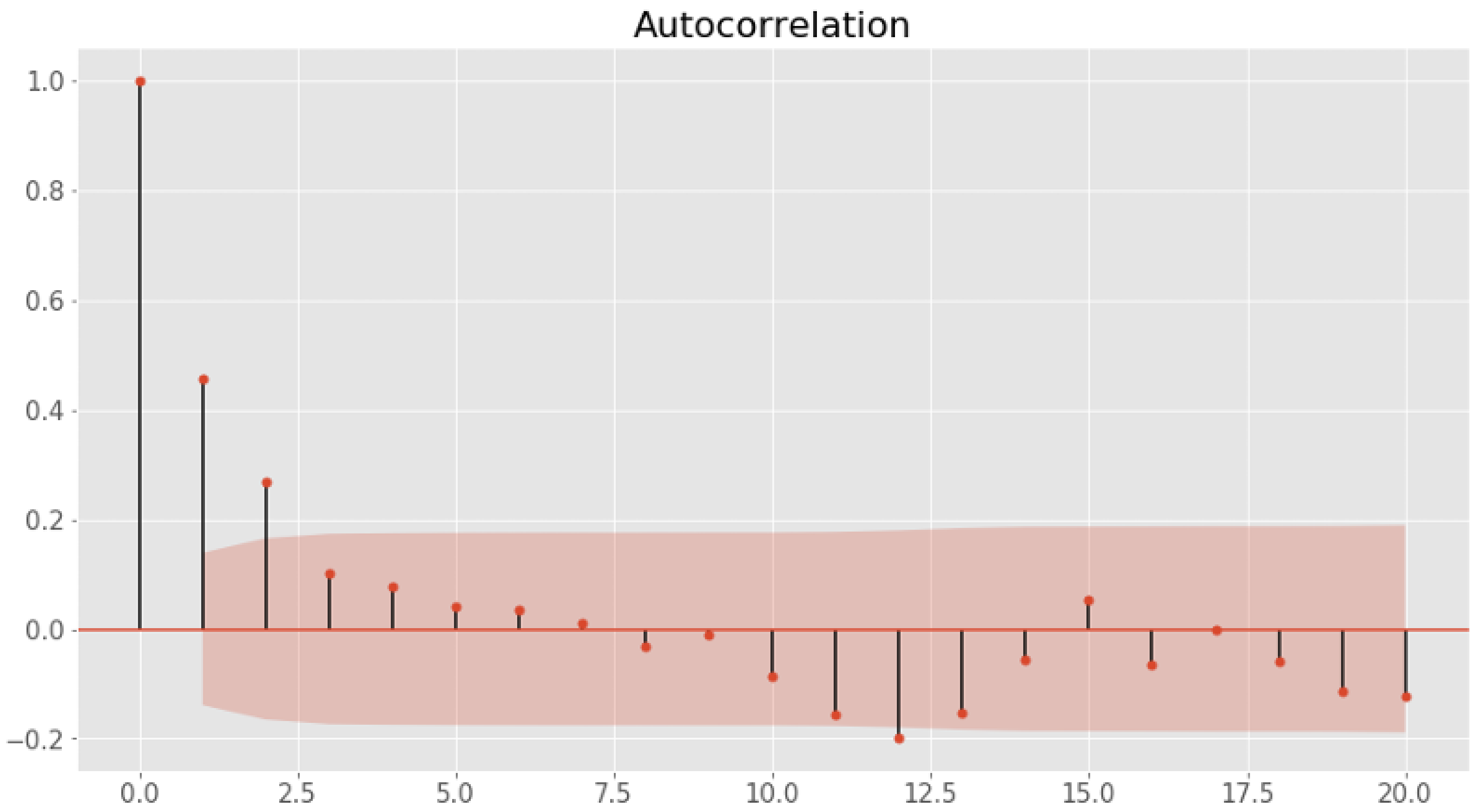


- **Mean** $E[X_t] = 0$
- **Variance** $\mathrm{Var}(X_t) = \sigma_w^2(1 + \theta_1^2 + \theta_2^2)$

MA Examples: MA(2) Process

- **Autocorrelation Function (ACF)**

$$\rho_1 = \frac{\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2} \qquad \rho_2 = \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2} \qquad \rho_h = 0, \forall h \geq 3$$



General MA(q) Process

- An important property of $MA(q)$ models in general is that there are nonzero autocorrelations for the first q lags, and $\rho_h = 0$ for all lags $h > q$.
- In other words, ACF provides a considerable amount of information about the order of the dependence q for $MA(q)$ process.
- Identification of an MA model is often best done with the ACF rather than the PACF.

MA Models: Parameters Estimation

- A well-known fact is that parameter estimation for MA model is more difficult than AR model.
  - One reason is that the lagged error terms are not observable.
- We can still use method of moments estimators for MA process, but we won't get the optimal estimators with Yule-Walker equations.
- In fact, since MA process is nonlinear in the parameters, we need iterative non-linear fitting instead of linear least squares.
- From a practical point of view, modern scientific computing software packages will handle most of the details after given the correct configurations.

Autoregressive Moving Average (ARMA) Models

Autoregressive Integrated Moving Average Models (ARIMA)

- ARIMA(p, d, q) model:
  - The model has autoregressive (AR) part of order p
  - The model has moving average (MA) part of order q
  - The data has been differenced d times
- Selected models
  - White noise = ARIMA (0, 0, 0)
  - Random walk = ARIMA(0, 1, 0) with no constant
  - Random walk with drift = ARIMA(0, 1, 0) with a constant
  - AR(p) = ARIMA(p, 0, 0); MA(q) = ARIMA(0, 0, q)

ARMA Models

- Autoregressive and moving average models can be combined together to form ARMA models.
- **Definition**
  - A time series $\{x_t; t = 0, \pm 1, \pm 2, \ldots\}$ is $ARMA(p, q)$ if it is stationary and

$$X_t = w_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j w_{t-j},$$

    where $\phi_p \neq 0$, $\theta_q \neq 0$, and $\sigma_w^2 > 0$, $w_t \sim wn(0, \sigma_w^2)$.
  - With the help of AR operator and MA operator we defined before, the model can be rewritten more concisely as:

$$\boxed{\phi(B) X_t = \theta(B) w_t}$$

Choosing Model Specification

- Recall we have discussed that ACF and PACF can be used for determining ARIMA model hyperparamters p and q.

	$AR(p)$	$MA(q)$	$ARMA(p, q)$
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

- Other criterions can be used for choosing q and q too, such as AIC (Akaike Information Criterion), AICc (corrected AIC) and BIC (Bayesian Information Criterion).
- Note that the selection for p and q is not unique.

Stationarize Nonstationary Time Series

- One limitation of ARMA models is the stationarity condition.
- In many situations, time series can be thought of as being composed of two components, a non-stationary trend series and a zero-mean stationary series, i.e. $X_t = \mu_t + Y_t$.
- **Strategies**
  - Detrending: Subtracting with an estimate for trend and deal with residuals.

$$\hat{Y}_t = X_t - \hat{\mu}_t$$

  - Differencing: Recall that random walk with drift is capable of representing trend, thus we can model trend as a stochastic component as well.

$$\mu_t = \delta + \mu_{t-1} + w_t$$

$$\nabla X_t = X_t - X_{t-1} = \delta + w_t + (Y_t - Y_{t-1}) = \delta + w_t + \nabla Y_t$$

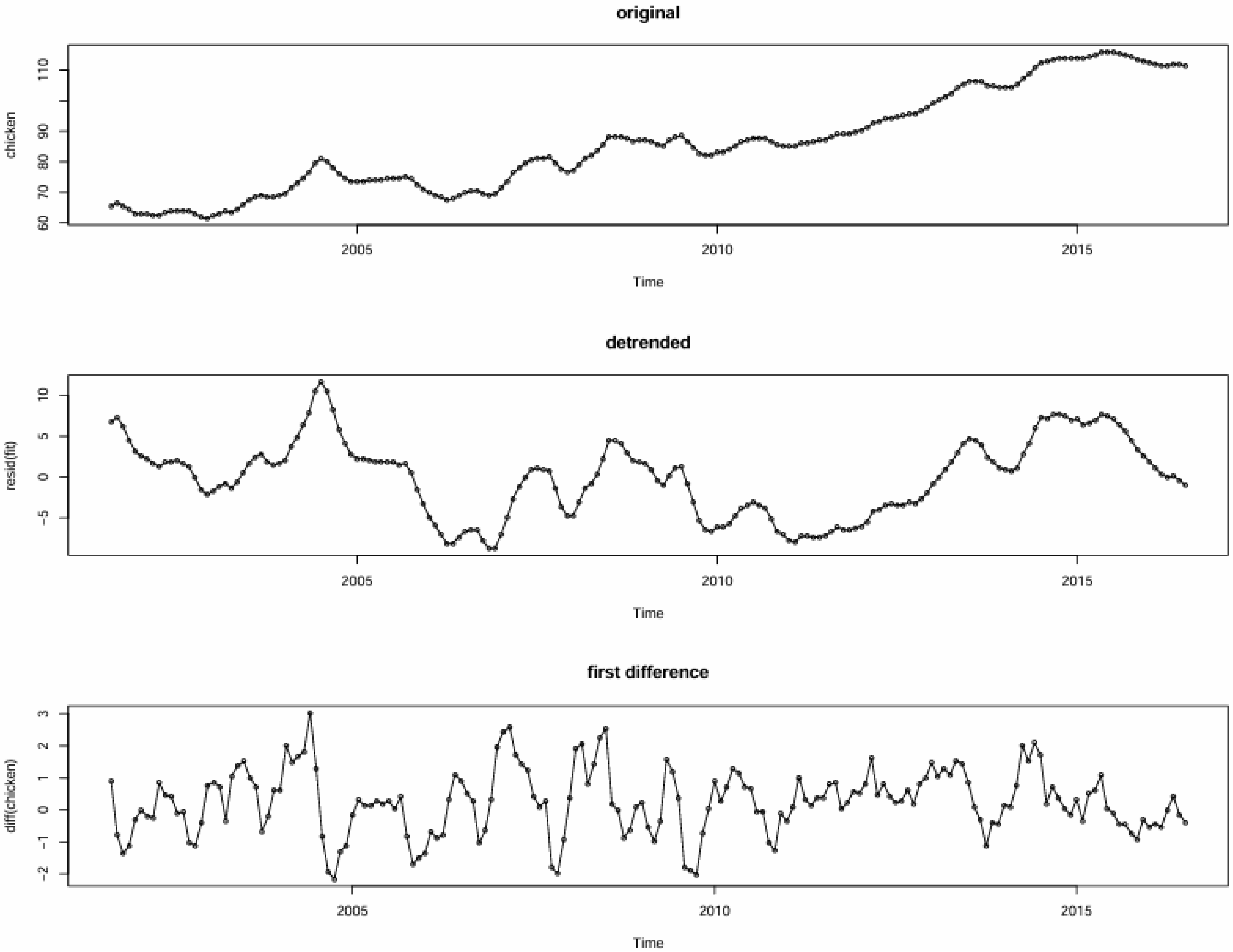
∇ is defined as the first difference and it can be extended to higher orders.

Diferencing

- One advantage of differencing over detrending for trend removal is that no parameter estimation is required.
- In fact, differencing operation can be repeated.
  - The first difference eliminates a linear trend.
  - A second difference, i.e. the difference of first difference, can eliminate a quadratic trend.
- Recall the backshift operator $X_t = BX_{t-1}$:

$$
\begin{aligned}
\nabla X_t &= X_t - X_{t-1} = X_t - BX_t = (1 - B)X_t \\
\nabla^2 X_t &= \nabla(\nabla X_t) = \nabla(X_t - X_{t-1}) \\
&= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\
&= X_t - 2X_{t-1} + X_{t-2} = X_t - 2BX_t + B^2 X_t \\
&= (1 - 2B + B^2)X_t = (1 - B)^2 X_t
\end{aligned}
$$

Detrending vs. Diferencing



Stationarize Time Series (cont.)

- Detrending
- Differencing
- Transformation: Applying arithmetic operations like log, square root, cube root, etc. to stationarize a time series.
- Aggregation: Taking average over a longer time period, like weekly/monthly.
- Smoothing: Removing rolling average from original time series.
- Decomposition: Modeling trend and seasonality explicitly and removing them from the time series.

KPSS Stationarity Test

How to interpret KPSS test results?

The output of the KPSS test contains 4 things:

- The KPSS statistic
- p-value
- Number of lags used by the test
- Critical values

Output Interpretation:

- The **p-value** helps decide whether to reject the null hypothesis. If the p-value is less than 0.05, the null hypothesis is rejected.
- The **KPSS statistic** is the test's computed value. If this statistic is greater than the critical value, the p-value will be low.

GO VS PYTHON